Functional genomic studies of the structure and regulation of eukaryotic transcriptomes

Thesis by

Georgi K. Marinov

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California
2014
(Defended May 1st, 2014)

To my mother, Ginka Desheva, for making it possible for me to reach this far, despite overwhelming odds against.

Acknowledgments

I have worked on numerous projects during my time at Caltech, collaborated with a large number of people both here and at other institutions and this is in large part what has carried me through the difficult moments, which I have been told that pretty much everyone experiences during their graduate school career but some were nevertheless unique to my situation. In addition, the research I have been doing, especially in my later and more analytical years has only been possible because it rests on a rather large experimental, computational and intellectual infrastructure provided by members of the Wold lab, other labs and associated facilities of campus, and it involved a great deal of collaboration. For these reasons, this section is going to be somewhat long as I owe at least a little (and sometimes quite a lot) to a long list of people:

I would like to first thank my advisor, Barbara Wold, for the continuous support, understanding and useful discussions throughout the years.

I would also like to thank my committee members for their useful input and the discussions I have had with them, even if we did not meet nearly as often as it might have been expected.

I wish to express enormous gratitude to Ali Mortazavi for introducing me to the computational aspects of the world of genomics, to Brian Williams for playing the same role on the experimental side, for generating much of the data on which this thesis is based and for serving as a role model for dedicated and committed hard work, and to Sreeram Balasubramanian for the many hours of insightful discussion about the world of science we have had over the last three years, a time when there were not many other opportunities for such activities.

As mentioned above, the collaborations I had got me through many difficult moments in my time at graduate school, by exposing me to the contagious enthusiasm for science of other people at times I had almost lost mine and by helping me find my way forward to completing my projects on my own. I will be forever grateful to Katalin Fejes Tóth and Alexei Aravin and members of their labs (Adrien LeThomas, Alicia Rogers, Sergei Manakov, Alexandre Webster, Ivan Olovnikov, Evelyn Stuwe, Dubravka Pezic), and to Yun Elizabeth Wang from the David Chan lab, without who this thesis would not have been possible.

I would also like to thank Gilberto DeSalvo, Antoni Kirilusha and Katherine Fisher-Aylor for initially introducing me to the lab, chromatin immunoprecipitation and working with muscle cells (an area the that has been the main focus of the lab and I learned a lot working on even though it features very little in this particular thesis); Daniel Kim for giving me a chance to explore the world of ES cells in more depth; Jost Vielmetter, Clarke Gasper, and Max Scott at the Protein Expression Center for their dedicated work on generating antibodies and automating ChIP-seq; Ken McCue for his statistical expertise; John Allman and members of the Allman lab; Anna Cecilia Therese Abelin, Shirley Pepke, Libera Berghella, Say-Tar Goh, Brian He and all other members of the Wold lab; and Gordon Kwan and Gordon Dan for their technical assistance throughout the year.

Much of the work described here would not have happened without the people working "at the other end of the hall", who not only made sure that the computing (Henry Amrhein, Diane Trout, Sean Upchurch, and Brandon King) and sequencing (Igor Antoshechkin and Lorian Schaeffer) infrastructures were working and in order, but also did their best to teach me how programming should be done professionally, and had significant intellectual input into a number of the projects I will discuss later.

Working within the ENCODE Consortium has to a great extent defined my graduate career, and has allowed me to interact with and learn from a great number of great people. They include our main collaborators in the Myers lab at the HudsonAlpha Institute for Biotechnology (Richard Myers, Florencia Pauli-Behn, Tim Reddy, Jason Gertz, Chris Partridge, Kimberly Newberry, and many others), with whom we have had many intellectually stimulating meetings and conference calls; Ross Hardison and members of his lab at Penn State; Anshul Kundaje (Stanford University) as well as Stephen Landt (Stanford University), Michael Snyder (Stanford University) and Peter Park (Harvard Medical School) who played a major role in the development and establishment of the ChIP-seq quality control metrics to which a part of this thesis is dedicated; Manolis Kellis (MIT) and Lucas Ward (MIT) with whom we collaborated on the work discussed in the second-to-last chapter of this thesis; and the members of the ENCODE RNA Working Group (Thomas Gingeras, Roderic Guigó, Carie Davis, Alex Dobin, Sarah Djebali, and many others).

I will finish this section by giving thanks to Yevgeni Raskatov, Fei Yang and Nick Nichols from the Dervan lab at Caltech for the several collaborative projects we worked on; Erich Schwartz for transferring his considerable knowledge on how to assemble genomes to me, Cole Trapnell for his input in the development of the RNA-seq analysis described in the first chapter of the thesis; Ellen Rothenberg, in whose lab I rotated in, Mark Zarnegar who I rotated with, and Jingli Zhang from the Rothenberg lab working with whom helped me clarify my thinking about analyzing functional genomics data and stimulated me to develop many of the tools that were the foundation of much of my later research; Arhav Mehta and Michael T. Bethune from the Baltimore lab; and Gary Schroth (Illumina) for all the stimulating meetings, discussions and invaluable first-hand knowledge straight from the source.

Abstract

The main focus of this thesis is the use of high-throughput sequencing technologies in functional genomics (in particular in the form of ChIP-seq, chromatin immunoprecipitation coupled with sequencing, and RNA-seq) and the study of the structure and regulation of transcriptomes. Some parts of it are of a more methodological nature while others describe the application of these functional genomic tools to address various biological problems. A significant part of the research presented here was conducted as part of the ENCODE (ENCyclopedia Of DNA Elements) Project.

The first part of the thesis focuses on the structure and diversity of the human transcriptome. Chapter 1 contains an analysis of the diversity of the human polyadenylated transcriptome based on RNA-seq data generated for the ENCODE Project. Chapter 2 presents a simulation-based examination of the performance of some of the most popular computational tools used to assemble and quantify transcriptomes. Chapter 3 includes a study of variation in gene expression, alternative splicing and allelic expression bias on the single-cell level and on a genome-wide scale in human lymphoblastoid cells; it also brings forward a number of critical to the practice of single-cell RNA-seq measurements methodological considerations.

The second part presents several studies applying functional genomic tools to the study of the regulatory biology of organellar genomes, primarily in mammals but also in plants. Chapter 5 contains an analysis of the occupancy of the human mitochondrial genome by TFAM, an important structural and regulatory protein in mitochondria, using ChIP-seq. In Chapter 6, the mitochondrial DNA occupancy of the TFB2M transcriptional regulator, the MTERF termination factor, and the mitochondrial RNA and DNA polymerases is characterized. Chapter 7 consists of an investigation into the curious phenomenon of the physical association of nuclear transcription factors with mitochondrial DNA, based on the diverse collections of transcription factor ChIP-seq datasets generated by the ENCODE, mouseENCODE and modENCODE consortia. In Chapter 8 this line of research is further extended to existing publicly available ChIP-seq datasets in plants and their mitochondrial and plastid genomes.

The third part is dedicated to the analytical and experimental practice of ChIP-seq. As part of the ENCODE Project, a set of metrics for assessing the quality of ChIP-seq experiments was developed, and the results of this activity are presented in Chapter 9. These metrics were later used to carry out a global analysis of ChIP-seq quality in the published literature (Chapter 10). In Chapter 11, the development and initial application of an automated robotic ChIP-seq (in which these metrics also played a major role) is presented.

The fourth part presents the results of some additional projects the author has been involved in, including the study of the role of the Piwi protein in the transcriptional regulation of transposon expression in *Drosophila* (Chapter 12), and the use of single-cell RNA-seq to characterize the heterogeneity of gene expression during cellular reprogramming (Chapter 13).

The last part of the thesis provides a review of the results of the ENCODE Project and the interpretation of the complexity of the biochemical activity exhibited by mammalian genomes that they have revealed (Chapters 15 and 16), an overview of the expected in the near future technical developments and their impact on the field of functional genomics (Chapter 14), and a discussion of some so far insufficiently explored

research areas, the future study of which will, in the opinion of the author, provide deep insights into many fundamental but not yet completely answered questions about the transcriptional biology of eukaryotes and its regulation.

Contents

Part I The Structure of Eukaryotic Transcriptomes

Part II Functional Genomics of Organelles

Part III Quality Assessment and Analysis of Chromatin Immunoprecipitation Data

Part IV Other Projects

Part V Conclusions and Towards the Future

Part VI Appendices

Part VII Bibliography

List of Figures

List of Tables

Preface

The path my graduate career took was somewhat unusual. I had the fortune to be able to work on a large number of diverse projects (especially as a result of being part of the ENCODE project). This means I have a correspondingly large number of at least somewhat interesting scientific stories to tell in my thesis. However, the flip side of this is that the common thread between all of them is not necessarily obvious and the "lack of focus" type of criticism towards it would not be entirely misplaced. For a long time, what that common thread was going to be was not obvious for me either, except for the rather trivial common denominator "High-throughput sequencing-based functional genomics" and the so-broad-as-to-be-almost-meaningless in the context of a graduate thesis "Understanding the mechanism of gene regulation and the structure and dynamics of transcriptomes eukaryotes". Yet, after some reflection, and especially after the response of the general scientific community to the presentation of ENCODE results and the subsequent activities I got involved in, I have come to think that the latter is not only not that useless after all, but I in fact have quite a lot to say on the subject and from a unique perspective and position shared by not many other people. Thus even if all I can offer is numerous very small compared to the magnitude of the general and very big task of understanding gene regulation contributions, they can nevertheless be brought under a common theme and put in their proper place in the bigger picture of where the field is circa 2013/2014 and what directions, in my humble opinion, it might not be a bad idea for at least a portion of it to move into in the near- and medium-term future.

My thoughts on the latter subject are presented in the chapters comprising the last part of this thesis, which also contain most of what would normally go into an introductory section. The rest of it is organized in four parts, each containing separate chapters. The first part is dedicated to the analysis of eukaryotic transcriptomes, using a variety of experimental techniques and data types, from bulk samples and on the single-cell level. The second grew in a completely unexpected way from a collaboration with Yun Elisabeth Wang in the Chan lab that initially focused on characterizing the binding of TFAM to the human mitochondrial genome but eventually grew into multiple studies applying functional genomic tools and data to organelles in both animals and plants. The third part concerns a number of technical issues having to do with the practice of carrying out chromatin immunoprecipitation (ChIP) experiments and their coupling with high-throughput sequencing (ChIP-seq), in particular the application of ChIP-seq quality control metrics to real-life data. It also includes a chapter on the development of a robotic ChIP assay in the Wold lab, something that will be a vital part of the future practice in the field. The fourth part includes chapters on some of the various other projects I have been involved in. The last part, as already mentioned, summarizes my work in the broader context of the current state of the field and defines what in my opinion would be fruitful directions for future research, both from the perspective of the current and expected near-future state of technology, and from the point of view of the general questions about the evolution of regulatory and genomic complexity arising from ENCODE results and their interpretation. Most of the individual chapters contained in each part were initially written as standalone papers, to which I later made (mostly slight) modifications in order to better fit the format of a thesis. Some of them have already been published, and a few of the ones that have not been will hopefully some day join them. The chapters can still be read independently of each other (this is especially true about those in the "Other Projects" part), although I hope an overarching team would become apparent to anyone reading the thesis from cover to cover, in its entirety.

Part I

The Structure of Eukaryotic Transcriptomes

This part contains four chapters dedicated to several functional genomic studies of the structure of eukaryotic transcriptomes that I have carried out. The first one describes the results of an early project aimed at characterizing the human polyadenylated transcriptome using some of the very first paired-end RNA-seq on multiple cell lines in existence (generated as part of the ENCODE Project). That work made it very clear that isoform assembly and isoform-level quantification are critical and potentially very weak points in the analysis of short-read RNA-seq data. To clarify the extent, impact and nature of these problems, I carried out an extensive simulation study on some of the most popular existing computational algorithms for carrying out these tasks, the results of which are described in the second chapter of this part. The third chapter contains a study of cell-to-cell variation in gene expression in human lymphoblastoid cell lines using single-cell RNA-seq, which also discusses in detail multiple key experimental and analytical issues with the practice of single-cell transcriptomics. Finally, I include a short chapter describing a proof-of-principle demonstration of a simple but elegant and robust approach to the analysis of mixed-species RNA-seq data.

1

The polyadenylated transcriptome of ENCODE cell lines

The material in this chapter (which consists of work done between 2010 and early 2012) was intended to form the core of an ENCODE companion paper to complement the main ENCODE transriptome paper (Djebali et al. 2012), and also present a somewhat different perspective of what the data is telling us:

Marinov GK*, Williams BA*, Trout D, Balasubramanian S, Fauli F, Reddy T, Gertz J, Murad R, Mortazavi A, Myers RM, Wold BJ. The polyadenylated transcriptome of ENCODE cell lines. 2012

This unfortunately never happened for various reasons I will not go into here. It is based on data generated primarily by Brian Williams in the Wold lab. The RNA Polymerase II and TAF1 ChIP-seq data from the Myers lab at the HudsonAlpha Institute for Biotechnology; the Nanostring miRNA data is courtesy of Rabi Murad in the Mortazavi lab at the University of California, Irvine.

Abstract

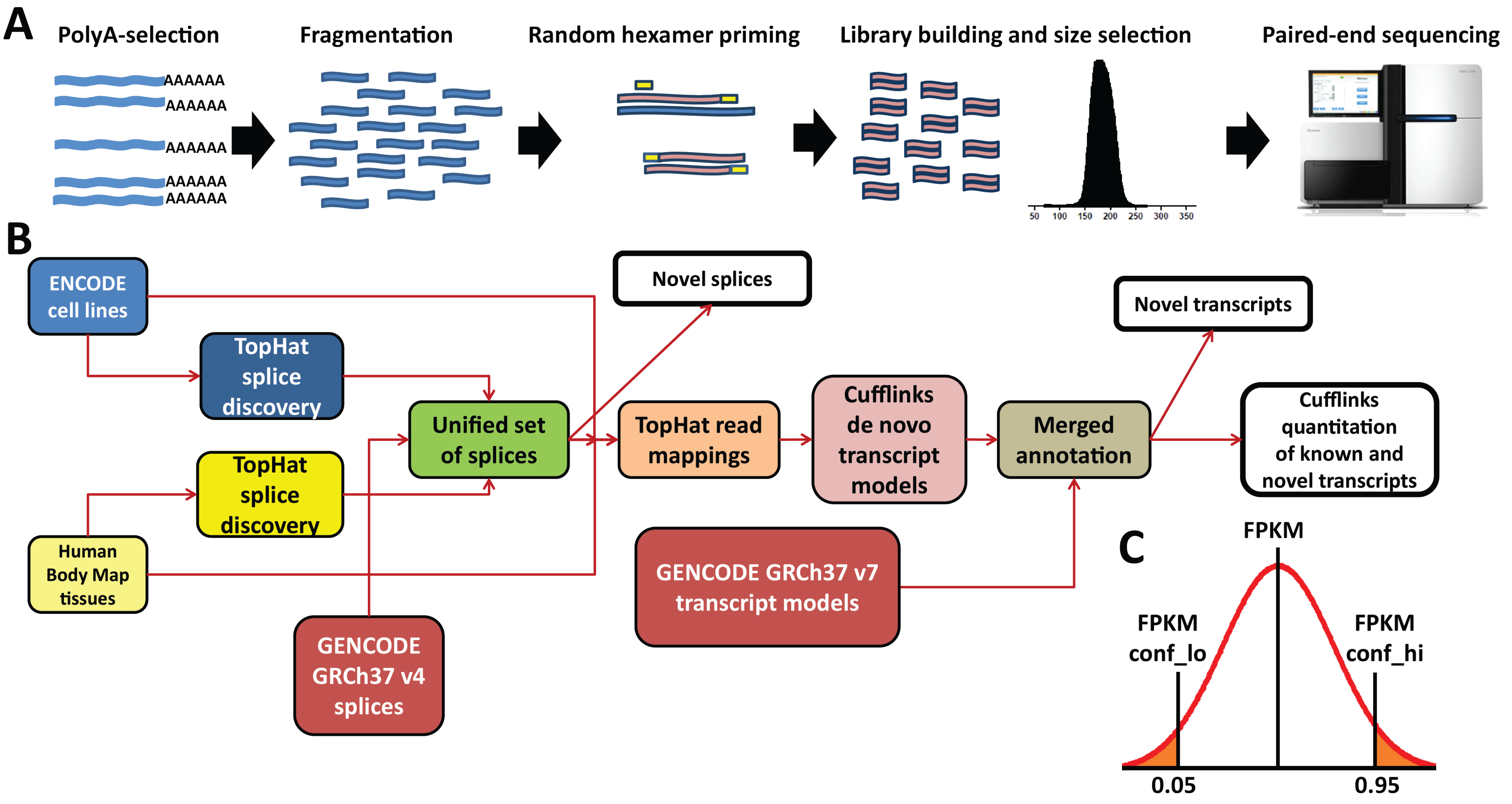
Multiple lines of evidence have previously suggest that the complexity of the transcript products generated by mammalian genomes is high. However, until the advent of RNA sequencing technology, it has not been possible to directly study this diversity at the resolution and depth provided by RNA-seq. In this study, we performed the first large-scale characterization of the human polyadenylated transcriptome using RNA-seq data from ENCODE cell lines and from a diverse collection of human tissues, as well as CAGE (Capped Analysis of Gene Expression) and ChIP-seq data for the TAF1 subunit of the transcription initiation complex. State-of-the-art analysis tools were then used to generate and quantify a conservative set of annotated and novel transcriptome elements, including splice junctions, exons, intergenic transcripts, isoforms of protein coding genes and alternative transcription initiation sites. The results reveal the high complexity of the transcriptome, but they also emphasize the interpretative challenges presented by the fact that much of the observed diversity is present at low absolute levels, meaning it is difficult to distinguish it from biochemical noise generated by the transcription and splicing machinery. Finally, I highlight the areas where future technical advances that should help resolve some of these issues are needed and expected.

1.1 Introduction

Contemporary polyA transcriptome measurements, made by deep sequencing of cDNA (RNA-seq), are remarkably information rich (Mortazavi and Williams et al. 2008; Nagalakshmi et al. 2008; Wang et al. 2008; Wilhelm et al. 2008; Pan et al. 2008; Sultan et al. 2008; Cloonan et al. 2008; Guttman et al. 2010; Cabili et al. 2011; Li et al. 2011). High-quality reference datasets can be mined, quantified, and analyzed in different ways, using different software and significance thresholds, to serve a wide range of biological investigations. For example, the majority of currently known mammalian genes were mapped by working backwards from knowledge of cloned RNA product(s) (Adams et al. 1991; Adams et al. 1995; Curwen et al. 2004). In principle, a deeply sequenced transcriptome can be used similarly to construct a more complete catalog of genes and their alternately processed RNA products, including both protein coding and long non-coding RNAs (lncRNAs; Guttman et al. 2009; Guttman et al. 2010; Cabili et al. 2011). This discovery mapping function has been a major motivation for ENCODE RNA-seq measurements (Myers et al. 2011; Djebali et al. 2012; this work), although both computational and biological complexities addressed below make this a challenging enterprise, especially for genes and isoforms expressed at relatively low levels. Reference RNA-seq data can also be used to quantify differential gene expression among cell types and tissues (Trapnell et al. 2012; Wang et al. 2010; Adams & Huber 2010); to quantify RNA splice use (Wang et al. 2008; Bradley et al. 2012); RNA editing (Li et al. 2011; Park et al. 2012), and other post-transcriptional processing (Jan et al. 2011; Kodzius et al. 2006; Hoskins et al. 2011; Affymetrix ENCODE Transcriptome Project 2009). Finally, since these transcriptome measurements reflect the steady state balance of RNA biogenesis and decay, RNA-seq data can be integrated with other genome-wide data-types such as RNA Polymerase II (RNA Pol2) occupancy and microRNA levels to gain insight into the specifics transcription initiation, and RNA processing, and turnover.

These diverse uses of mRNA-seq data are best and most efficiently served by sequencing to high depth, because greater depth increases sensitivity; by using longer sequence reads, typically in the paired-end format, because this increases the specificity of mapping reads to the correct gene and transcript isoform; and by using source RNA that is highly enriched for being in the polyA fraction, which reduces background from other RNA types and improves interpretability. As part of the ENCODE Porject, we therefore developed a community resource of human polyA RNA-seq transcriptomes (100–200 million sequence reads in each biological replicate) by applying a widely used polyA RNA-seq method (updated from Mortazavi et al. 2008), to diverse human cell lines (ENCODE tier 1 and Tier 2). The analysis of these cell-line and primary cell culture RNAs was substantially augmented by including and comparing RNA-seq data from 16 adult human tissues sequenced as part of the Human Body Map (HBM) project (primary data available from GEO, accession code GSE30611). The resulting data resource was analyzed using a computational Cufflinks-based pipeline (updated from Trapnell et al. 2010 and Roberts et al. 2011) to examine the structure and diversity of the human transcriptome, in particular focusing on: 1) known and novel splice junctions, protein coding transcripts and lncRNAs, and other elements of the transcriptome were analyzed as a function of expression level, confidence value and locus complexity; 2) global integrative mining was illustrated by using ChIP-seq data for TAF1 and RNA Polymerase II to determine the number and cell type specific usage of alternative promoters; 3) specific loci, including the protocadherin gene clusters and the transcription factor BHLHE40, were used to illustrate how the transcriptome data and models can be used, alone and in conjunction with other data-types to generate explicit new hypotheses.

A particular computational challenge presented by short-read RNA-seq data is accurately building and quantifying new gene models and new isoform models of existing genes. The sequence read lengths used in this study were 2x75 (ENCODE) and 2x50 or 1x100 bp (HBM) coming from on average ∼200bp-long RNA fragments, while essentially all mRNAs are much longer, with the median GENCODE V7 protein coding transcript being ∼1600bp long. This prevents the direct measurement of long-range contiguity, which is instead inferred, and this inference process becomes extremely challenging for genes with many exons and large number of coexpressed alternative isoforms. Another great challenge in analyzing and mining transcriptomic and other high-throughput data comes from our limited understanding of the levels and sources of biological noise in the underlying processes, including transcription initiation, splicing, and polyadenylation. Computational tools, such as Cufflinks (Trapnell et al. 2010; Roberts et al. 2011; used here) or Scripture (Guttman et al. 2010), address these issues with


A
PolyA-selection
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
Fragmentation
Random hexamer priming
Library building and size selection
50
100
150
200
250
300
350
Paired-end sequencing
B
ENCODE cell lines
TopHat splice discovery
Human Body Map tissues
TopHat splice discovery
GENCODE GRCh37 v4 splices
Unified set of splices
Novel splices
TopHat read mappings
Cufflinks de novo transcript models
GENCODE GRCh37 v7 transcript models
Merged annotation
Novel transcripts
Cufflinks quantitation of known and novel transcripts
C
FPKM
FPKM conf_lo
FPKM conf_hi
0.05
0.95

algorithms designed to balance sensitivity of detection with robustness and parsimony of transcript identification. It is expected that quantification on the final transcript model set will be significantly affected by uniformity of coverage over any given transcript, by its true level of expression, and by the number of models offered for each gene. Therefore the datasets were also used to explore how transcript models are affected by characteristics such as gene size, locus complexity, overall expression level, and strength of evidence for alternative splice junction use.

This analysis revealed, first, that the high sensitivity and resolution of RNA-seq provides evidence for the very high complexity of the human transcriptome, with large numbers of novel splice junctions, coding and noncoding transcripts, alternative splicing and alternative initiation events detectable in the data. Second, the majority of this diversity is rare in abundance, thus most of it likely represents biological noise rather than biologically functional transcriptional products. However, as there is no simple relationship between expression levels and functionality, it is at present not possible to determine in a straightforward way which of these transcriptional elements are functional and which are not. Third, a confounding factor that has becoming apparent during the course of the analysis, and one that has to feature prominently in the interpretation of all data of this kind, originates from the fact that the computational challenges posed by short-read RNA-seq are very difficult to solve thus making any results that solely depend on the performance of the tools used to carry out the analysis provisional at best in the absence of deeper investigation using orthogonal means. This topic is explored in more detail in the following chapter.

1.2 Results

We generated 2x75 bp paired-end RNA-seq data on polyadenylated RNA from a diverse set of 10 human cell lines (Figure 1.1A) that include primary cultures, immortalized lines, tumor-derived lines, and a pluripotent embryonic stem line. Derivatives of all three germ layers were included, although these lines represent only a small fraction of the hundreds of human cell types. Two biological replicates were sequenced for every cell line, to an average depth of 100–120×10^6 mapped reads each (Table 1.1). These sequencing depths are sufficient to reach saturation of gene and transcript detection. The data was of high quality as evidenced by the absence of 3' bias and robust coverage of all of the length of genes. In addition to these data, we added to our analysis polyadenylated RNA-seq data for 16 human tissue samples generated as part of the Human Body Map 2 project (HBM), sequenced to an average depth of 200–250×10^6 reads. In contrast to the ENCODE lines, each human tissue is composed of multiple cell types and none have experienced effects or artifacts of ex-vivo culture or growth transformation. For a subset of the ENCODE cell lines, we also generated ChIP-seq data for RNA Polymerase 2 and for the transcription initiation complex component TAF1, sequenced to a depth of at least 12×10^6 uniquely mappable reads per replicate

Figure 1.1 *(preceding page)*: Overview of data generation protocols and computational analysis. (A) PolyA-selected RNA-seq library generation. Libraries are built from PolyA-selected RNA from ENCODE cell lines using fragmentation and random hexamer priming. Libraries are size-selected so that the average fragment length is around 200bp and paired-end reads are generated on the Illumina GAIIx or HiSeq 2000. (B) Data analysis workflow. RNA-seq reads from ENCODE cell lines and from HBM tissues are individually mapped with TopHat in de novo splice junction discovery mode. Next, all newly discovered splice junctions are combined with splice junctions from the GENCODE annotation to create a consolidated set of junctions, which is supplied to TopHat for remapping of all reads. The TopHat alignments are used to run Cufflinks in de novo transcript discovery mode. The Cufflinks models for all cell lines and tissues are then merged with the GENCODE annotation to create a final consolidated set of transcripts. Final Cufflinks quantification is performed on the final merged annotation for each cell lines and downstream analysis of expression values and transcript characteristics is carried out. (C) Distinction between transcript expression estimation metrics used. In addition to the FPKM score corresponding to the most likely actual transcript abundance, for stringency purposes we use extensively the $\text{FPKM}_{conf-lo}$ lower limit of the 95% FPKM confidence interval provided by Cufflinks.

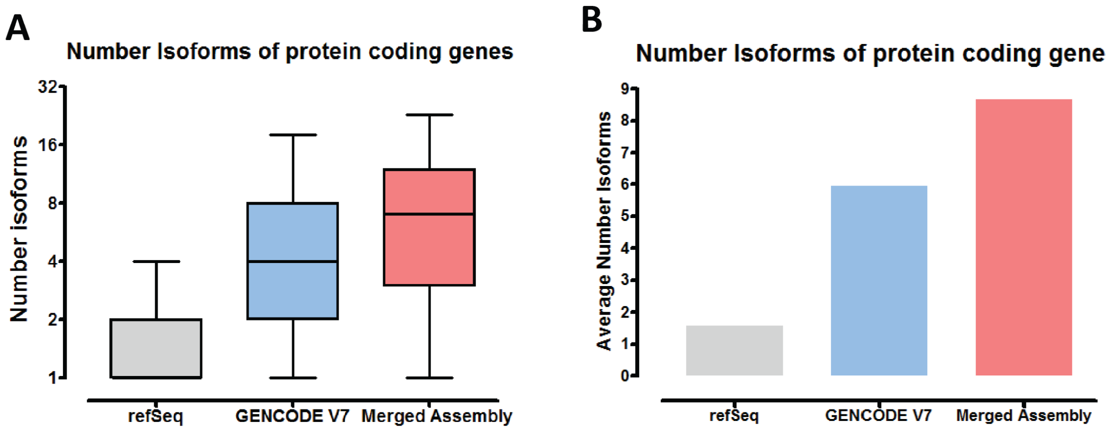


Figure 1.2: Number of isoforms per gene for protein coding genes in refSeq, GENCODE V7 and the final merged assembly based on ENCODE+HBM data. Number of isoforms per gene for protein coding genes in refSeq, GENCODE V7 and the final merged assembly based on ENCODE + HBM data. (A) Distribution of isoforms number (Y-axis is plotted on a $log2$ scale) (B) Average number of isoforms per gene.

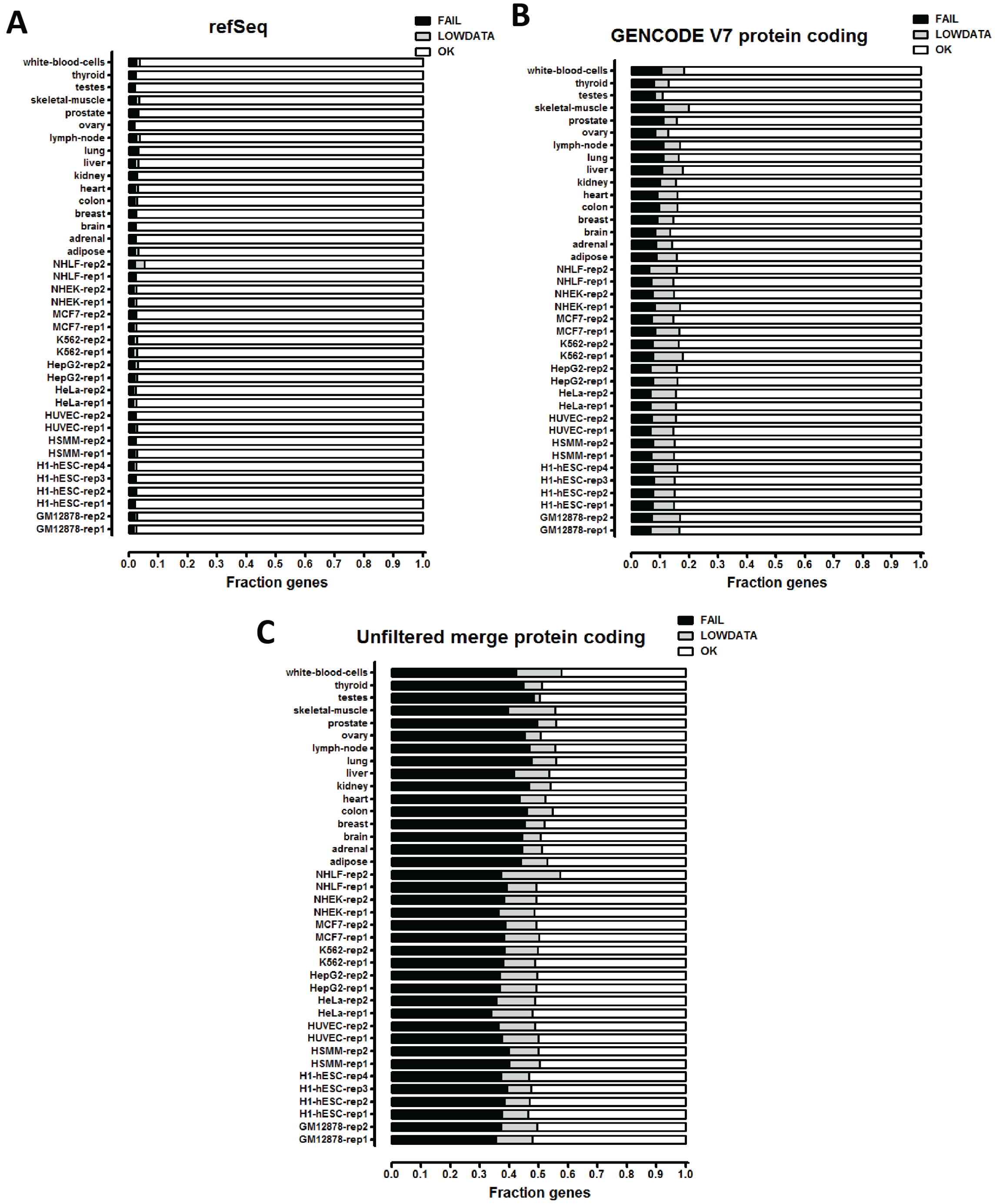
(Table 1.2).

1.2.1 Computational pipeline for uniform analysis of the transcriptome across multiple cell lines and tissues

To take advantage of the potential of RNA-seq to characterize both annotated and unannotated portions of the transcriptome, it is first necessary to define a full set of elements (exons, splice junctions and transcripts) that could then be compared and quantified between samples. A number of tools exist for *de novo* reconstruction of all transcript models from RNA-seq data (Trapnell et al. 2010; Guttman et al.; 2010). However, these strategies, as previously applied, produce results that are not directly comparable between individual samples. This problem is compounded by the fact that the resulting transcript models can be, and often are incomplete and imperfect, due to sequence read mapping errors, insufficient coverage of lowly expressed genes, and highly variable read coverage over some other genes. In order to address these issues, I devised a computational pipeline that combines *de novo*–generated transcript models from individual samples with existing annotated models while exerting a number of filters to reduce the number of artifactual and poorly supported transcripts. This single set of transcript models was then re-quantified across all samples.

I aimed for a relatively stringent set of novel isoform models of known genes plus transcripts of novel genes. This approach is expected to miss large numbers of “real” transcripts present in the data and to therefore underestimate transcriptome diversity. This is a necessary compromise between including all models for which there is some evidence and the ability of software and sequencing technology to reconstruct and resolve transcript abundance for complex loci. I note that as a result of Cufflinks’ abundance filters during *de novo* assembly and the additional stringency criteria imposed, final transcript level annotation does not incorporate all splice junctions for which there is sequence evidence; splice junctions are therefore examined separately from transcripts in later analysis.

Reads from individual samples were first aligned against the hg19 version of the human genome using TopHat (version 1.0.14; Trapnell et al. 2009) in *de novo* mode. The splice junctions identified this way were combined with the splice junctions in the GENCODE v4 annotation (Harrow et al. 2006) to create a final set of candidate junctions. This unified junctions set was then supplied to TopHat and all samples were remapped in order to include all reads mapping to annotated and candidate novel splices, that, due to low transcript abundance, low coverage or exons being too short, TopHat had not been


A
refSeq
FAIL
LOWDATA
OK
white-blood-cells
thyroid
testes
skeletal-muscle
prostate
ovary
lymph-node
lung
liver
kidney
heart
colon
breast
brain
adrenal
adipose
NHLF-rep2
NHLF-rep1
NHEK-rep2
NHEK-rep1
MCF7-rep2
MCF7-rep1
K562-rep2
K562-rep1
HepG2-rep2
HepG2-rep1
HeLa-rep2
HeLa-rep1
HUVEC-rep2
HUVEC-rep1
HSMM-rep2
HSMM-rep1
H1-hESC-rep4
H1-hESC-rep3
H1-hESC-rep2
H1-hESC-rep1
GM12878-rep2
GM12878-rep1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction genes
B
GENCODE V7 protein coding
FAIL
LOWDATA
OK
white-blood-cells
thyroid
testes
skeletal-muscle
prostate
ovary
lymph-node
lung
liver
kidney
heart
colon
breast
brain
adrenal
adipose
NHLF-rep2
NHLF-rep1
NHEK-rep2
NHEK-rep1
MCF7-rep2
MCF7-rep1
K562-rep2
K562-rep1
HepG2-rep2
HepG2-rep1
HeLa-rep2
HeLa-rep1
HUVEC-rep2
HUVEC-rep1
HSMM-rep2
HSMM-rep1
H1-hESC-rep4
H1-hESC-rep3
H1-hESC-rep2
H1-hESC-rep1
GM12878-rep2
GM12878-rep1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction genes
C
Unfiltered merge protein coding
FAIL
LOWDATA
OK
white-blood-cells
thyroid
testes
skeletal-muscle
prostate
ovary
lymph-node
lung
liver
kidney
heart
colon
breast
brain
adrenal
adipose
NHLF-rep2
NHLF-rep1
NHEK-rep2
NHEK-rep1
MCF7-rep2
MCF7-rep1
K562-rep2
K562-rep1
HepG2-rep2
HepG2-rep1
HeLa-rep2
HeLa-rep1
HUVEC-rep2
HUVEC-rep1
HSMM-rep2
HSMM-rep1
H1-hESC-rep4
H1-hESC-rep3
H1-hESC-rep2
H1-hESC-rep1
GM12878-rep2
GM12878-rep1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction genes

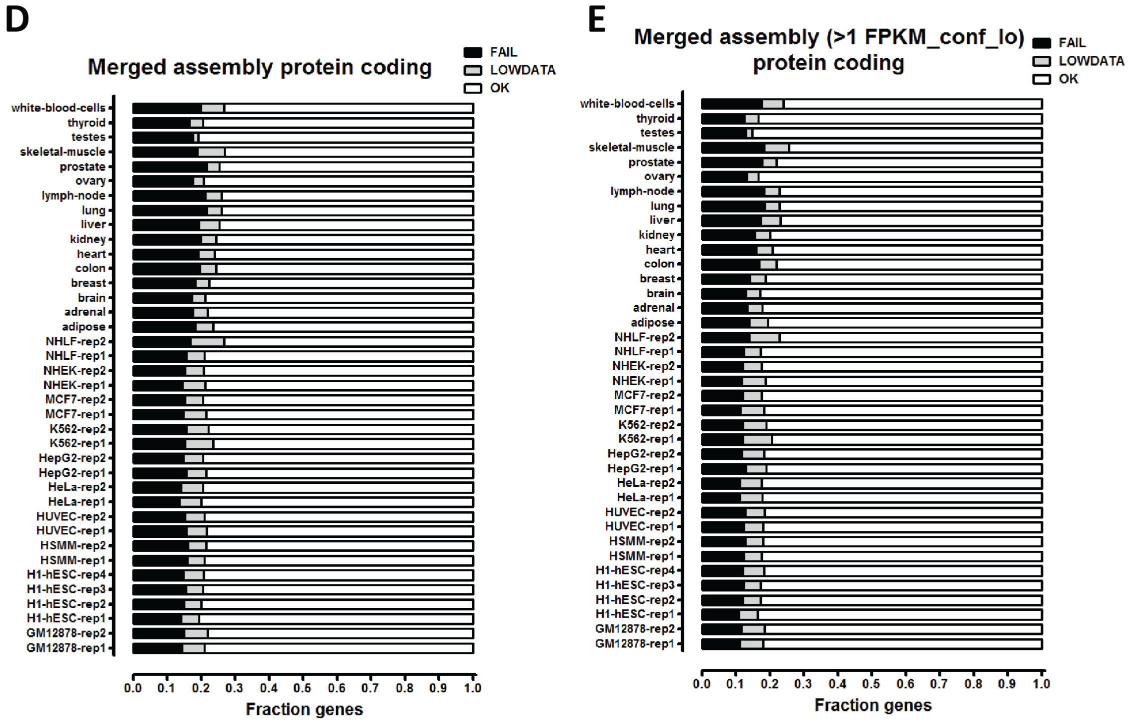


Figure 1.3: Number of genes for which isofor-level quantification is unidentifiable or faces other numerical issues. Cufflinks assigns a FAIL or LOWDATA status to genes where the algorithm can not confidently assign FPKMs to individual transcripts. (A) For the refSeq annotation, containing few isoforms, a very small percentage of genes are flagged in this manner (B) For GENCODE V7, 10-15% of protein coding genes are flagged. (C) For an unfiltered Cuffmerge assembly performed only on novel intergenic transcripts and novel isoforms with the GENCODE V7 annotation as a reference, more than half of protein coding genes are flagged. (D) A filtered assembly of all novel intergenic transcripts and novel isoforms still has ~5% more failed quantifications of protein coding genes than GENCODE V7 (E) A filtered assembly of all novel intergenic transcripts and novel isoforms with the added requirement that they should be present at >= 1 FPKM$_{conf\_lo}$ in the individual assemblies approaches the numbers observed for GENCODE v7 (the minimal annotation complexity we could work with). Total number of protein coding genes: ~20,500.

able to map in *de novo* mode.

Next, the resulting alignments were assembled into transcripts using Cufflinks (version 1.0.1; Trapnell et al. 2010) and the individual Cufflinks assemblies merged using the Cuffmerge program in the Cufflinks suite (Trapnell et al. 2012) with the GENCODE v7 annotation as a reference. The GENCODE annotation was chosen because it was adopted as the ENCODE analysis standard, selected as the most comprehensive set of curated transcript models for the human genome. *De novo* transcript assembly with Cufflinks can be done in a fully *de novo* mode or in a reference annotation based transcript (RABT) assembly mode (Roberts et al. 2011). The latter delivers more complete transcript models because incomplete assemblies typically arise in *de novo* mode due to stretches of low coverage or unmappable regions. In my experience, this class of artifacts is significant, even with very deeply sequenced datasets. However, the RABT mode produces a large number of artifactual transcript models when run on very complex annotations such as GENCODE v7, which contains 4 to 6 alternate isoforms on average for each gene (Figure 1.2). Ideally, these artifactual transcripts would be irrelevant to downstream analysis, because they would be assigned zero or very low expression values after requantification, but in practice reads are often dissipated across many models, due to uneven read coverage or the absence of reads allowing for un-

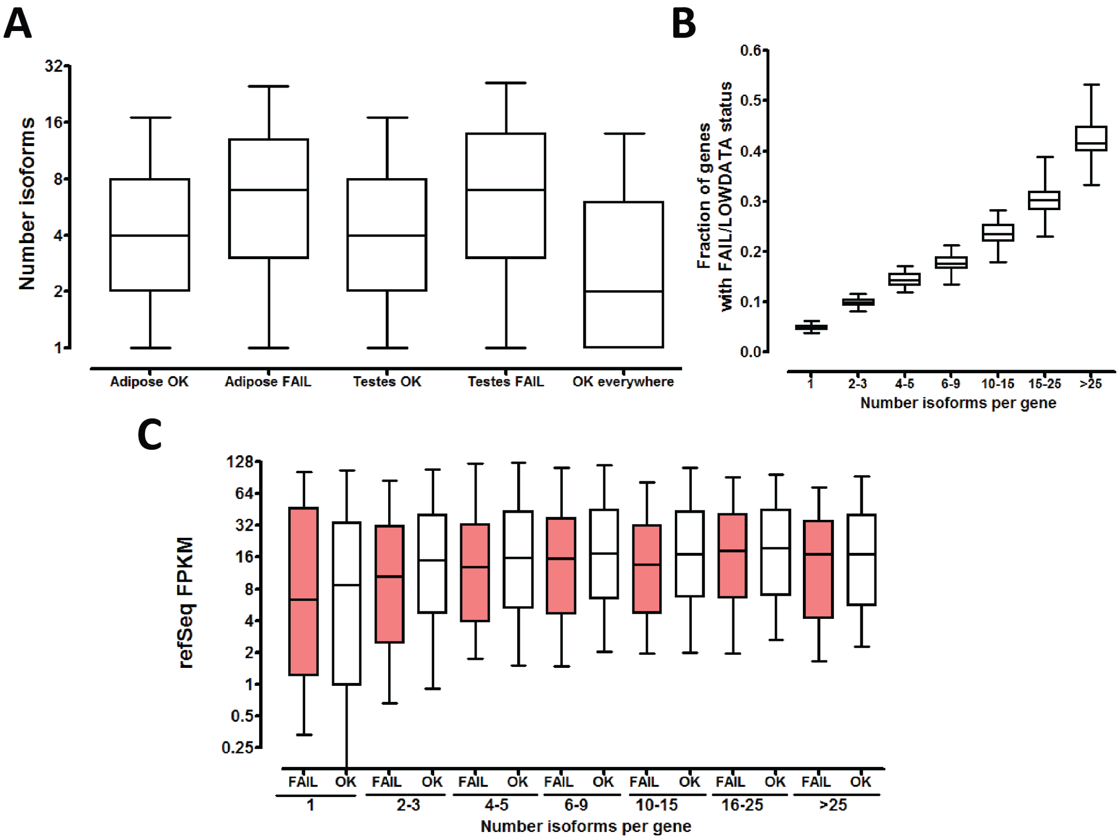


Figure 1.4: Relationship between "failure" of transcript-level quantification and locus complexity and expression levels. (A) Successfully quantified GENCODE v7 transcripts in adipose and testes tissue (two samples shown for brevity, results are similar for all cell lines) have a median of 4 isoforms per gene. Genes for which quantification fails in these samples have a median of 8 isoforms per gene. Finally, genes that are confidently quantified in all cell lines and tissues have a median of only 2 isoforms per gene. 5-95 percentile whiskers. (B) With increased locus complexity, an increasing number of genes become too complex to confidently quantify on the transcript level. Shown is the fraction of GENCODE v7 genes for which quantification fails as a function of the number of annotated isoforms for that gene. Box plots represent the distribution of that fraction across all samples used in this study. 5-95 percentile whiskers. (C) Weak correlation between expression levels and quantification failure. Plotted is the distribution of refSeq FPKMs for protein coding genes (here we used FPKMs calculated on the refSeq annotation to avoid the uncertainty arising from summing the FPKM estimates for individual transcripts in a genes in a complex annotation when transcript-level quantification is not reliable) as a function of their quantification status and isoform number in adipose tissue. 10-90 percentile whiskers.

ambiguously distinguishing between transcripts. Indeed, in the course of establishing the pipeline, it was found that a major challenge for downstream analysis arises from the rapid growth in the number of isoform models per gene, even after stringent filtering of anticipated artifacts. As more and more cell lines and tissues are analyzed, the number of isoforms becomes very large and the ability to confidently assign the still relatively short 75bp reads to individual isoforms is compromised (even using the GENCODE V7 annotation alone, it was not possible to confidently quantify the individual isoforms of about 2000 protein coding genes or about 10% of all; see Figure 1.3 and 1.4 for more detail, as well as the Discussion section for further treatment of the subject).

I therefore assembled transcripts for each

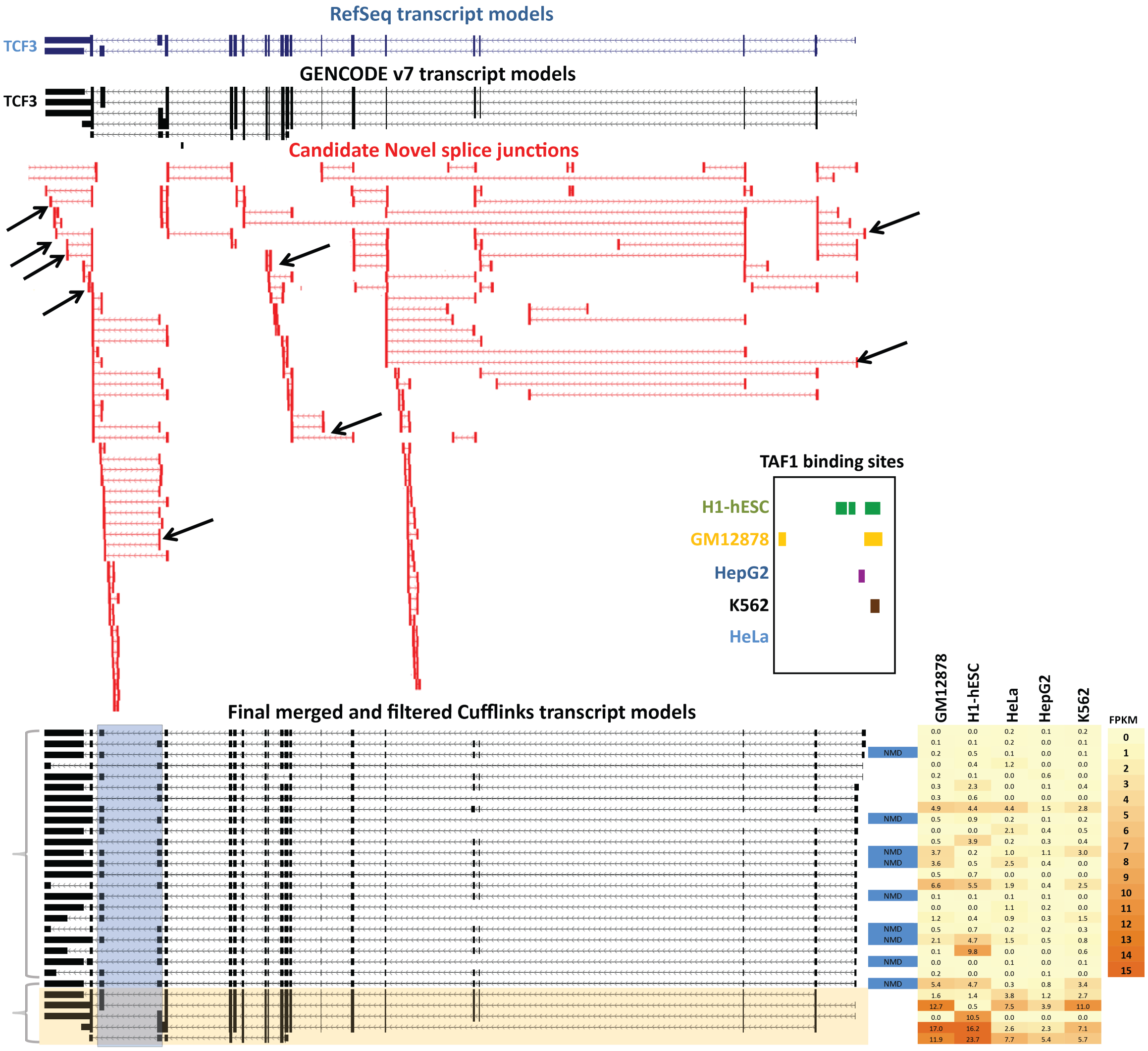


Figure 1.5: Isoform-level quantification, fragment support for known and novel junction, and TAF1 binding sites for the *TCF3* locus. The arrows point to the novel splice junctions incorporated in the novel isoforms annotated in the merged assembly.

sample individually in fully *de novo* mode, then applied a number of filters before and after the Cuffmerge step with the goal of deriving an as conservative a set of transcript models as possible. First, the individual assemblies were compared against the GENCODE annotation using Cuffcompare (Trapnell et al. 2010) in order to filter out intronic fragments and polymerase run-on fragments; only transcripts classified as intergenic or as novel isoforms of known genes were retained. I included all novel intergenic transcripts in the merge, but for novel isoforms of protein coding genes I required the lower 95% confidence **F**ragments **P**er **K**ilobase per **M**illion reads (FPKM) estimate (FPKM$_{conf\_lo}$, Figure 1.1) to be greater than 1. After merging transcripts with Cuffmerge, transcripts present in GENCODE V7 but missing from the resulting set of models were added back and major artifact classes such as retained introns and overtly long 3'UTRs were removed.

I illustrate the results of the pipeline in Figure 1.5 using the *TCF3* gene as an example. The *TCF3* gene encodes the E2A transcription factor, which plays important roles in myogenesis (Berkes & Tapscott 2005), lymphocyte development (Quong et al. 2002; Murre 2005), and in other systems. The *TCF3*/E2A locus is well known for producing two different proteins, E12 and E47, as a result of mutually exclusive al-

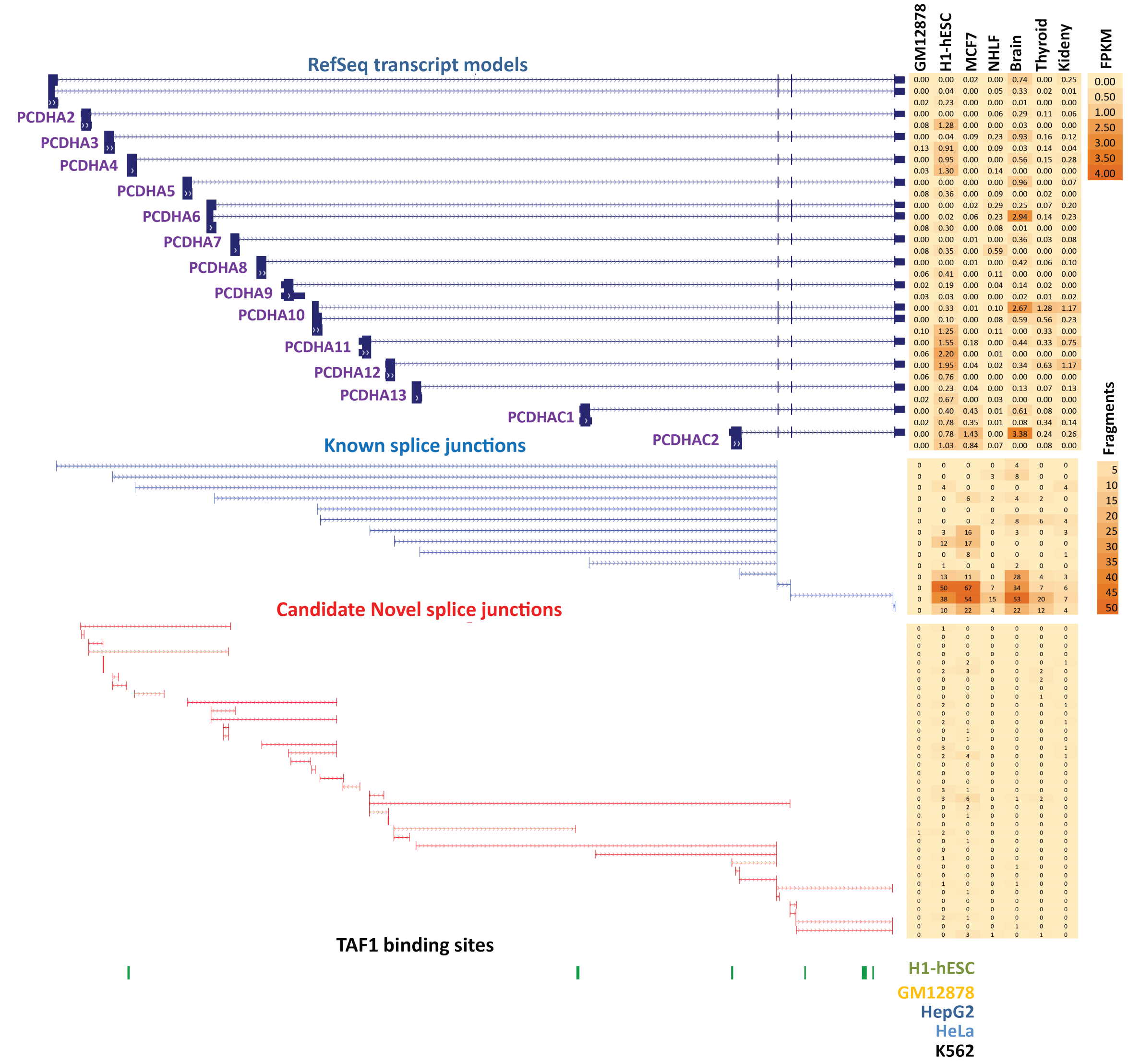


Figure 1.6: Isoform-level quantification, fragment support for known and novel junction, and TAF1 binding sites for the protocadherin-α cluster (*Pcdhα*).

ternative splicing of exons 17 and 18 (Murre et al. 1989a; Murre et al. 1989b; Figure 1.5). Two *TCF3* isoforms (one for E12 and one for E47) are annotated in the RefSeq set of transcript models, while 5 exist in GENCODE V7, with 2 and 3 alternative TSSs, respectively. A large number of unannotated splice junctions in the locus were detected, most of which turn out to be of low abundance when examined in detail. The final merged set of models contained additional 24 isoforms not present in GENCODE, with a new alternative TSS upstream of the 5'-most GENCODE TSS for the gene, thus greatly expanding the set of known *TCF3* isoforms. These newly assembled isoforms are of lower estimated abundance relative to the expression levels of the known ones. Finally, for two of the TSSs, one annotated and the one identified from RNA-seq data, we observed TAF1 binding overlapping the 5' exon.

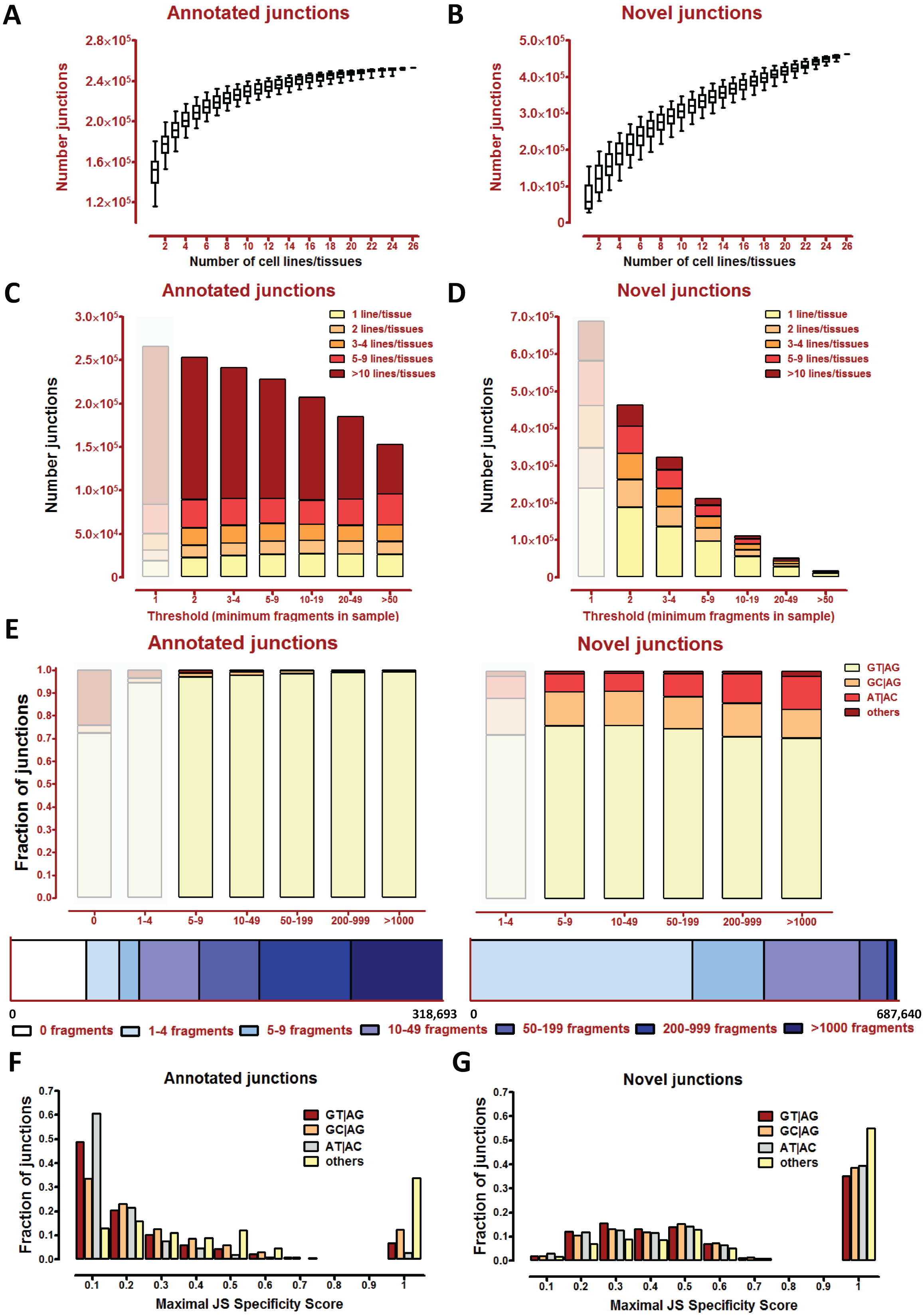
Another example of the utility of the integrated use of these datasets was the protocadherin-α (*Pcdhα*) cluster (Figure 1.6). Protocadherins are cell surface single-pass transmembrane proteins, particularly highly expressed in the nervous system and enriched in synaptic junctions, which have been proposed to play a major role in the precise specification of neuronal connectivity under the "chemoaffinity hypothesis" model of establishing neural circuits

(Zipursky & Janes, 2010). The *Pcdhα*, *Pcdhβ* and *Pcdhγ* genes exhibit a striking pattern of organization and clustering in the genome. All *Pcdhα* and all *Pcdhγ* protocadherins share three constant 3' exons which code for a portion of the intracellular domain of the protein, to which numerous unique alternative 5' exons, each with its own promoter, are alternatively spliced (Wu & Manitatis, 1999; Tasic et al. 2002; Wang et al. 2002); these 5' exons code for the extracellular, transmembrane, and parts of the intracelllar portions of the protein. The *Pcdhβ* cluster is similarly organized but there are no constant exons and each gene is transcribed individually. Protocadherins are transcribed monoallelically, i.e. only a single variable exon is used on each cluster allele, with which one exactly being determined stochastically, meaning that each cell produces one of a large number of combinations of protocadherins, potentially generating unique molecular identities for each neuron (Esumi et al. 2005). I examined *Pcdhα* expression in our datasets and observed the expected highest expression levels in brain tissue, with PCDHA6, PCDHA10 and PCDHAC2 being most highly expressed, and lower-level expression levels in several other tissues such as thyroid and kidney. Strikingly, I also found high (comparable to those in brain) expression levels of *Pcdhα* in human embryonic stem cells (which to the best of my knowledge has not been reported previously), and lower levels in a few other cell lines such as the breast cancer MCF7 cell line and the lung fibroblast NHLF cell line (Figure 1.6). TAF1 binding to the promoters of several of the more highly expressed *Pcdhα* genes was observed in H1-hESC. In addition, three TAF1 binding sites in the 3' intron of the *Pcdhα* cluster were detected, as well as a number of low-abundance novel splice junctions connecting the variable exons with each other (Figure 1.6); their significance is at present not clear and remains to be tested in future studies.

1.2.2 Catalog of splice junctions in the human genome

I compared the full set of splice junctions present in the TopHat mappings to the GENCODE V7 human genome annotation. Of the 318,693 splice junctions in the annotation, 266,311 were covered by at least one and 253,063 by at least two unique sequence fragments (to avoid counting PCR duplicates, a unique sequence fragment is defined as the number of non-identical read pairs crossing a junction and I refer to that number everywhere except where explicitly specified otherwise) (Figure 1.7C). This represents an approximate measurement of the breadth of coverage of the transcriptome in the data, with the junctions not detected consisting of a combination of junctions from rarely expressed genes not present in the cell lines and tissues examined, junctions from non-polyadenylated transcripts and possibly artifacts in the annotation. In addition to the annotated junctions, I also observed 687,638 candidate novel junctions supported by at least one, and 462,274 supported by at least two unique fragments. I note that the TopHat algorithm relies on first finding putative exons based on read coverage and then on identifying splice junctions nearby (Trapnell et al. 2009), i.e. it employs an "exon-first" approach to junction discovery. This junction set is therefore more conservative than those from some other *de novo* splice mapping algorithms relying on "seed-extend" strategies (Garber et al. 2011) to find splices (Dobin et al. 2013; De Bona et al. 2008; Wu et al. 2010), which are likely to find more junctions in the same dataset. I also note that I ran TopHat with default settings with respect to the genomic range over which new junctions can be discovered so the maximum distance between two splice sites is 500 kb. Only 81 annotated junction span genomic distances longer than 500kb so it is unlikely that many novel ones were missed due to this constraint. On average, around 150,000 annotated junctions were detected in each cell line or tissue (Figure 1.8A). Of the novel junctions, between 150,000 and 250,000 were found in each cell line, and 50-120,000 in each tissue (Figure 1.8B). The lower number in tissues likely reflects the fact that HBM data is a mixture of 2x50bp and 1x100bp reads, while the cell lines were sequenced as 2x75bp. This difference in read length is expected to make de novo junction discovery more difficult.

I next asked how exhaustively we had sampled the diversity of splicing events in the human transcriptome by looking at the saturation of junction detection as a function of the number of cell lines/tissues examined (Figure 1.7A and B). These cumulative plots show that annotated junctions exhibit a clear saturation trend, with more than 90% detected with less than half of the cell lines considered. In contrast, the trend for novel junction discovery indicates that fur-


A
Annotated junctions
Number junctions
2.8×10⁵
2.4×10⁵
2.0×10⁵
1.6×10⁵
1.2×10⁵
2 4 6 8 10 12 14 16 18 20 22 24 26
Number of cell lines/tissues
B
Novel junctions
Number junctions
5.0×10⁵
4.0×10⁵
3.0×10⁵
2.0×10⁵
1.0×10⁵
0
2 4 6 8 10 12 14 16 18 20 22 24 26
Number of cell lines/tissues
C
Annotated junctions
Number junctions
3.0×10⁵
2.5×10⁵
2.0×10⁵
1.5×10⁵
1.0×10⁵
5.0×10⁴
0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
1 2 3-4 5-9 10-19 20-49 >50
Threshold (minimum fragments in sample)
D
Novel junctions
Number junctions
7.0×10⁵
6.0×10⁵
5.0×10⁵
4.0×10⁵
3.0×10⁵
2.0×10⁵
1.0×10⁵
0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
1 2 3-4 5-9 10-19 20-49 >50
Threshold (minimum fragments in sample)
E
Annotated junctions
Fraction of junctions
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
0 1-4 5-9 10-49 50-199 200-999 >1000
Novel junctions
1-4 5-9 10-49 50-199 200-999 >1000
GT|AG
GC|AG
AT|AC
others
0
318,693
0
687,640
0 fragments
1-4 fragments
5-9 fragments
10-49 fragments
50-199 fragments
200-999 fragments
>1000 fragments
F
Annotated junctions
Fraction of junctions
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
GT|AG
GC|AG
AT|AC
others
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Maximal JS Specificity Score
G
Novel junctions
Fraction of junctions
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
GT|AG
GC|AG
AT|AC
others
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Maximal JS Specificity Score

ther sequencing of additional cell lines and tissues of different origin is likely to substantially increase the number of new candidate junctions.

An open question regarding alternative splicing events and unannotated transcripts in mammalian systems is to what extent they represent biologically functional events as opposed to well-tolerated transcriptional and splicing machinery noise (Wang et al. 2008; Pan et al. 2008; Melamud & Moult 2009; Sorek et al. 2004). I therefore sought to characterize the properties of novel junctions and compare them to those of annotated ones as a function of their expression levels. When the effect of different fragment support thresholds on junction discovery was examined (Figure 1.7C and D), a clear trend was observed: annotated junctions have high fragment count support (the splice-specific empirical surrogate for expression level) in multiple cell lines, while novel splices are mostly detected in one or a small number of cell lines. The majority of novel junctions were supported by only a few fragments, with their corresponding transcript isoforms being at levels of uncertain significance, assuming expression in most cells in the population. This is entirely consistent with a large fraction of them being noise. However, due to the very large total number of candidate novel junctions, significant numbers of highly supported novel junctions were still discovered: for example, 79,667 junctions were supported by more than 5 unique fragments in more than 3 cell lines/tissues, and 8,898 junctions supported by more than 20 fragments in more than 5 cell lines/tissues, thresholds that can be considered stringent and suggestive of biological functionality.

1.2.3 Splice junction motif preferences

Next, I asked how canonical (GT|AG) (Mount 1982) and non-canonical splice sites distribute in the junctions set (Figure 1.7E). A number of non-canonical splice site junctions are present in the GENCODE v7 annotation and I observed that they are most often found among those junctions that were not detected in any of our samples. The fraction of such junctions decreased with increased fragment support thresholds. These may represent artifacts in the annotation or transcripts which are depleted in polyA-selected RNA. Novel junctions were mostly of the canonical GT|AG type, but in addition, GC|AG and AT|AC, substrates of the minor U12 spliceosome (Burge et al. 1998; Patel & Steitz 2005; Will & Luhrmann 2005; Jackson 1991; Hall & Padgett 1994; Sharp & Burge 1997; Hall & Padgett 1996; Tarn & Steitz 1996a; Tarn & Steitz 1996b) were also very abundant irrespective of the level of fragment support. It is possible that this reflects a TopHat preference for such junctions rather than actual biological reality. About 10% of the novel canonical junctions, but a much smaller fraction of all non-canonical ones are supported by EST sequences (Figure 1.9). Finally, I explicitly examined the tissue specificity of junctions by calculating tis-

Figure 1.7 *(preceding page)*: Catalogue of splice junctions in the human genome. (A) and (B) Cumulative detection of annotated (A) and novel (B) splice junctions in ENCODE cell lines and HBM tissues. Unique fragment counts were summed where replicates were available, the order of the cell lines/tissues was permuted 10,000 times and the number of junctions detected with the addition of every cell lines/tissue was counted for each permutation. A threshold of 2 unique fragment counts was used. Note that the Y axis does not begin at 0. (C) and (D) Annotated splice junctions are much more abundant and widely used than novel ones. Plotted is the number of junctions detected at a given threshold with the color codes corresponding to the number of cell lines in which this threshold is passed. Most known junctions are detected at high fragment counts in multiple cell lines while the majority of novel junctions are supported by few reads and only in a small number of cell lines. Shaded area corresponds to support levels that we are least confident in. (A) Canonical and non-canonical splice-sites and total read support for annotated and known junctions. The sum of unique fragment counts across all samples for each junction is shown, and for each abundance category the fraction of canonical, major non-canonical (as reported by TopHat) and other splice sites was plotted. The total number of junctions in each category is shown in the blue bars below. (F) and (G) Tissue/cell type-specificity of splice junctions measured using the JS Specificity Score. High score indicates high tissue-specificity, low score indicates widespread abundance

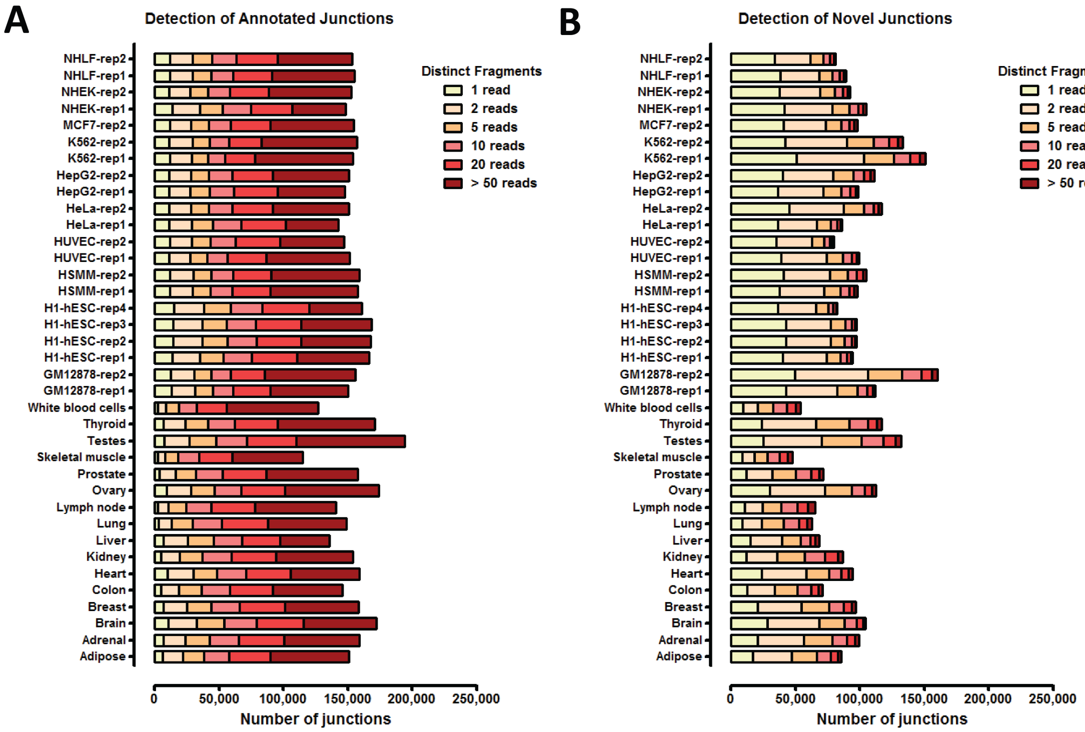


Figure 1.8: Number of splice junctions detected in each cell line and tissue. (A) Annotated (B) Novel

sue specificity score for each junction (JS score; see the Methods seciton for details). Annotated junctions mostly had low JS scores reflecting widespread abundance in multiple cell lines while novel junctions clustered in two groups - either with a JS score of 1 and perfect tissue specificity (due to detection in only a single cell line) or with a medium JS score and expression in a limited number of cell lines. In addition, canonical junctions had lower JS scores than non-canonical ones, suggesting detection of the latter in limited number of samples.

1.2.4 Classifying novel splice junctions relative to existing annotation

To better understand where novel junctions arise relative to existing gene structures, I classified all RNA-seq junctions into the classes depicted in Figure 1.10. I note that splice junctions connecting positions within a gene, for which no splice site is annotated (novel intragenic exons), need not originate from transcripts that belong to the gene in which they are embedded; they can instead result from nested, previously unannotated transcripts. Of all novel junctions, the most numerous category were junctions connecting an annotated exon to a novel exon within the same gene (class E, 264,121), followed by junctions connecting two novel intragenic exons (class C, 186,668) junctions connecting two annotated exons (class A, 75,147) and intergenic junctions outside of annotated genes (class H, 54,555) (Figure 1.10B).

Among all novel splice categories, the strongest in read support were the relatively small group of class B junctions that connect exons of two different annotated genes. Of these almost half arise from loci in which paralogs are adjacent and both are highly expressed in one or more of our samples (Figure 1.11A). One explanation is that they may represent computational artifacts, i.e. cases in which *de novo* junctions discovery incorrectly placed reads across two exons of different genes due to their high sequence similarity. A higher fraction of tandem paralog pairs had multiple such junctions connecting their exons (Figure 1.110B and C), and a high fraction of them had very similar donor or

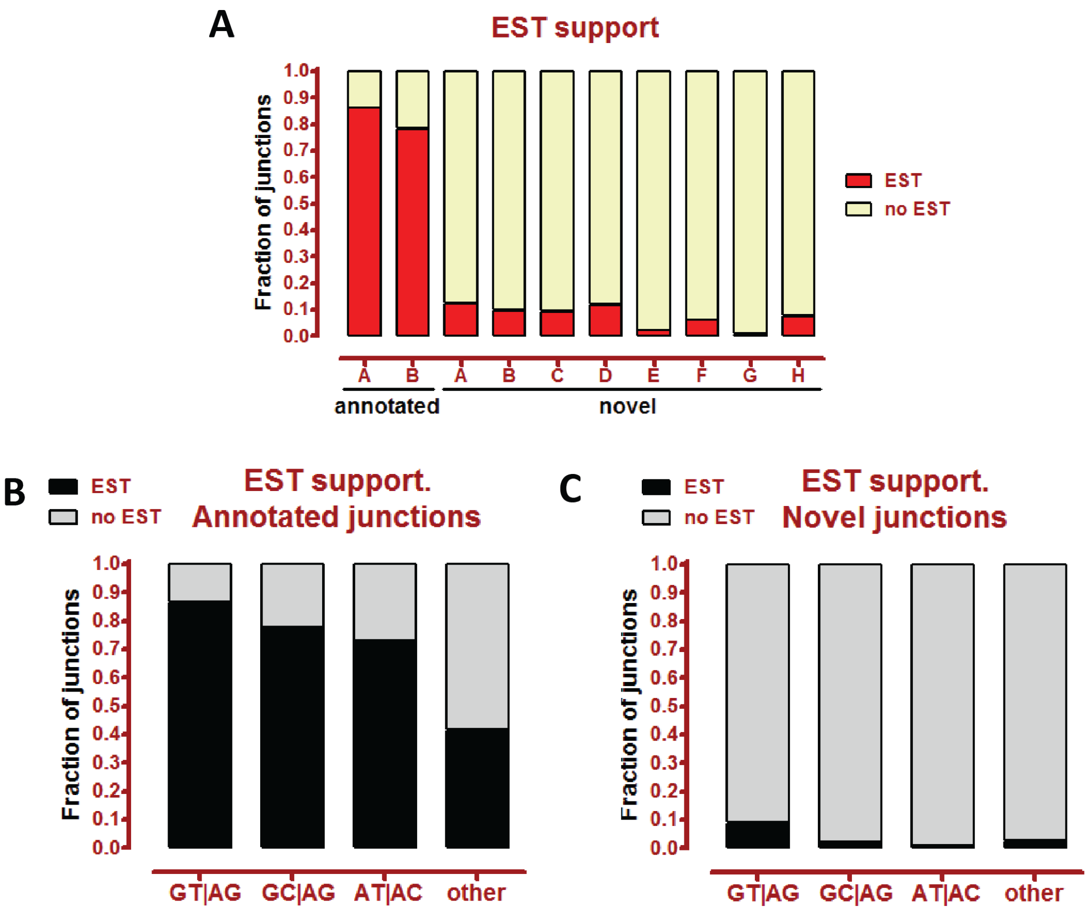


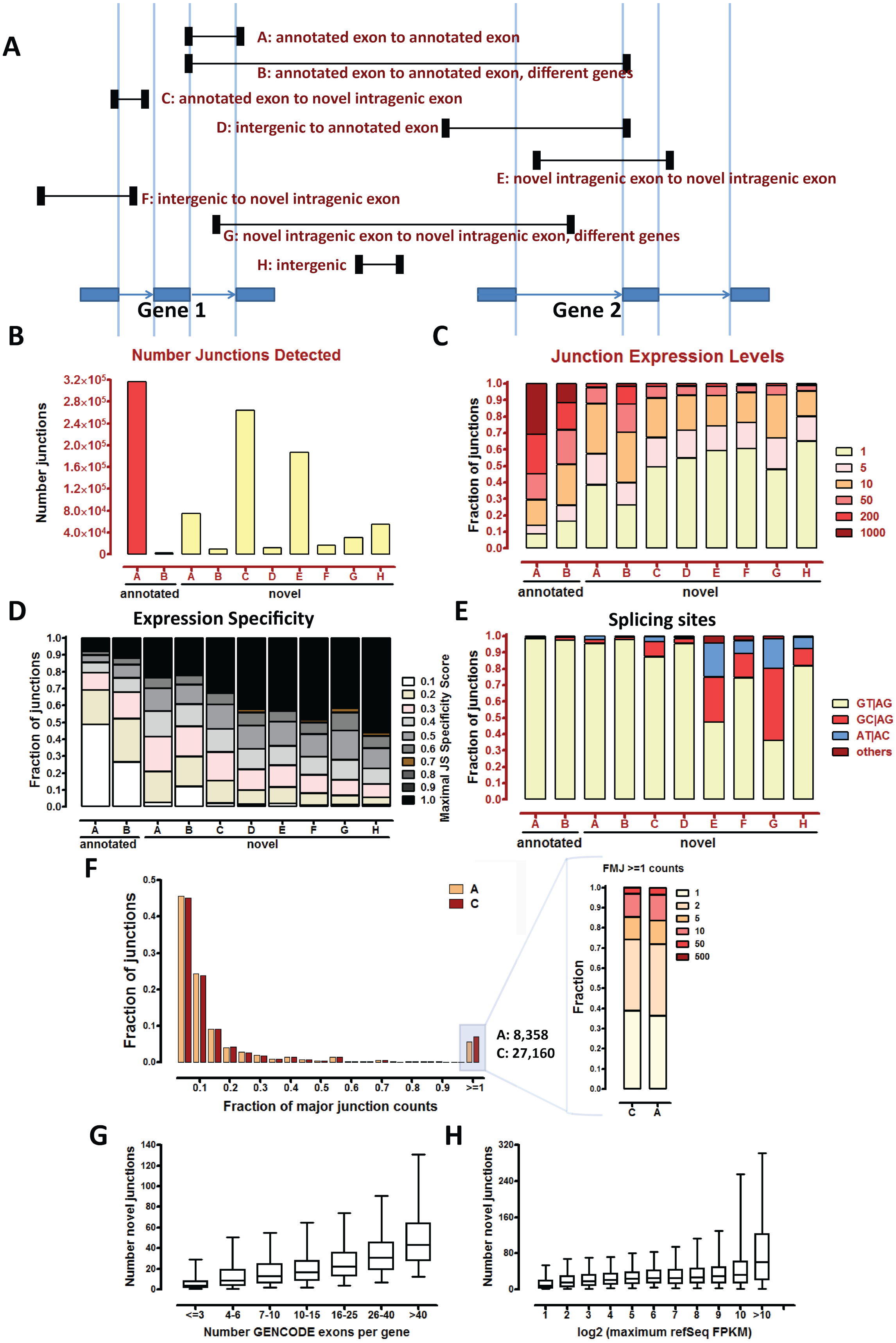
Figure 1.9: EST support for annotated and novel junctions. (A) EST support different junction connection categories (see Fig. 3) (B) EST support for annotated canonical and non-canonical junctions (C) EST support for novel canonical and non-canonical junctions.

acceptor sites in both genes compared to the rest of class B junctions (Figure 1.11E), consistent with a purely computational explanation. However, such junctions had higher fragment count support (Figure 1.11F) and the number of fragments in an individual sample correlated well with both genes being expressed in that sample (Figure 1.11G), which argues for their biochemical presence. Of the other class B junctions, about a third connect non-coding transcripts or protein coding transcripts to non-coding transcripts (Figure 1.11D) and on average, they originated from gene pairs with even higher expression than junctions connecting tandem paralogs (Figure 1.11G).

The next most abundant class of junctions were class A and class H junctions (Figure 1.10C), connecting known exons of a known gene and intragenic exons, respectively.

Because annotated splice sites are overwhelmingly canonical, we expect novel junctions connecting to an annotated exon to also be predominantly canonical, which is what is observed. Most non-canonical junctions belong to the E, F and G classes, which connect intragenic genomic positions. I note that completely intergenic, class H junctions exhibit a much higher proportion of canonical junctions than these three groups (Figure 1.10D). The most plausible interpretation of this observation is that a higher fraction of class H intergenic junctions represent functional transcripts while the other classes are mainly the result of biological and computational noise.

Previous studies have reported the existence of large numbers of alternative canonical splice


A
A: annotated exon to annotated exon
B: annotated exon to annotated exon, different genes
C: annotated exon to novel intragenic exon
D: intergenic to annotated exon
E: novel intragenic exon to novel intragenic exon
F: intergenic to novel intragenic exon
G: novel intragenic exon to novel intragenic exon, different genes
H: intergenic
Gene 1
Gene 2
B
Number Junctions Detected
Number junctions
3.2x10^5
2.8x10^5
2.4x10^5
2.0x10^5
1.6x10^5
1.2x10^5
8.0x10^4
4.0x10^4
0
A B A B C D E F G H
annotated
novel
C
Junction Expression Levels
Fraction of junctions
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
1
5
10
50
200
1000
A B A B C D E F G H
annotated
novel
D
Expression Specificity
Fraction of junctions
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Maximal JS Specificity Score
A B A B C D E F G H
annotated
novel
E
Splicing sites
Fraction of junctions
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
GT|AG
GC|AG
AT|AC
others
A B A B C D E F G H
annotated
novel
F
Fraction of junctions
0.5 0.4 0.3 0.2 0.1 0.0
A
C
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 >=1
Fraction of major junction counts
A: 8,358
C: 27,160
FMJ >=1 counts
Fraction
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
1
2
5
10
50
500
C A
G
Number novel junctions
140 120 100 80 60 40 20 0
<=3 4-6 7-10 10-15 16-25 26-40 >40
Number GENCODE exons per gene
H
Number novel junctions
320 240 160 80 0
1 2 3 4 5 6 7 8 9 10 >10
log2 (maximum refSeq FPKM)

acceptor sites separated by 3 bp from the main annotated acceptor site (”NAGNAG” splice acceptors) (Hiller et al. 2004; Akerman et al. 2006; Bradley et al. 2012). I found 1193 class C junctions of this kind, but this did not constitute the majority of such junctions – in addition to the classical NAGNAG events, I also observed large numbers of splice junctions representing other small shifts relative to the annotated splice donor sites and at both donor and acceptor ends. For a significant fraction of the junctions the shift was not divisible by 3 and therefore frame-preserving (Figure 1.12A and B) and there was not a large difference in the fraction of junctions that are canonical, in their fragment support or expression specificity (Figure 1.12C and D) between frame-preserving and non-frame preserving junctions.

The A and C classes of novel junctions connect known exons which have annotated junctions connecting to them. This allows us to ask what the abundance of these novel junctions relative to the associated annotated ones is, which I quantified as the fraction of major annotated junction counts (FMJ), where the major junction is the one with the highest fragment support in a given sample. For the majority of A and C novel junctions, this ratio was less than 0.1 (Figure 1.10F) arguing against their biological functionality. A small, (less than 10%) fraction had FMJ scores greater than 1 corresponding to preferential utilization of the novel junction over the annotated ones. However, around 80% of such cases have total read support of less than 5 fragments, i.e. these events mostly happen at junctions/genes that are lowly expressed, and biologically relevant preferential use of novel junctions is limited to the remaining few thousand junctions with high read coverage. Finally, I examined the cell type specificity of such events (Figure 1.13) and found that they mostly occur in a small number of cell lines/tissues, with testes, K562, H1-hESC and GM12878 exhibiting the highest number.

1.2.5 Correlation between presence of novel junctions and gene expression and loci complexity

Following the hypothesis that most novel junctions detected in RNA-seq data are the result of a combination of biological and experimental noise, I tested the correlation between detection of novel junctions for each gene and the expression levels and the number of exons for a given gene. The expectation is that highly expressed genes and genes with a large number of exons are likely to generate more novel junctions than genes with low expression levels and few exons. Our observations are indeed consistent with such an expectation as shown in Figure 1.10G and H.

1.2.6 Identification of novel intergenic transcripts

In recent years, long intergenic non-coding RNAs (lincRNA) have become a hot topic of research, with thousands of such transcripts identified using microarrays and RNA sequencing (Guttman et al. 2009; Khalil et al. 2009; Cabili et al. 2011). Individual lincRNAs have been implicated in a number of important biological processes (Guttman et al. 2011; Borsani et al, 1991; Brown et al. 1991; Lee et al. 1999; Azzalin et al. 2007; Huarte et al. 2010; Meller et al. 1997). To identify novel lincRNAs and characterize lincRNA expression patterns across cell types and tissues, I adapted previously pub-

Figure 1.10 ***(preceding page)*****: Relation of novel junctions to existing annotations.** (A) Different categories of junction connections relative to an annotation. (B) Number of junctions in each category (all annotated and novel ones included irrespective of read support). (C) Distribution of read support (across all samples) for each category in unique fragment counts. (D) JS specificity scores. (E) Canonical and non-canonical splice junctions. (F) Correlation between the number of novel junctions detected and the number of annotated exons for a given gene (only protein coding genes shown). (G) Correlation between the number of novel junctions detected and expression levels of genes (RefSeq FPKM values for protein coding genes shown). (H) Novel splice junctions at least one end of which is the same as that of an annotated splice junction are typically detected at a small fraction of the fragment counts of the major annotated junction (FMJ) sharing that splice site. For about 10% of them, the FMJ is greater than 1 but the majority are junctions with low fragment support

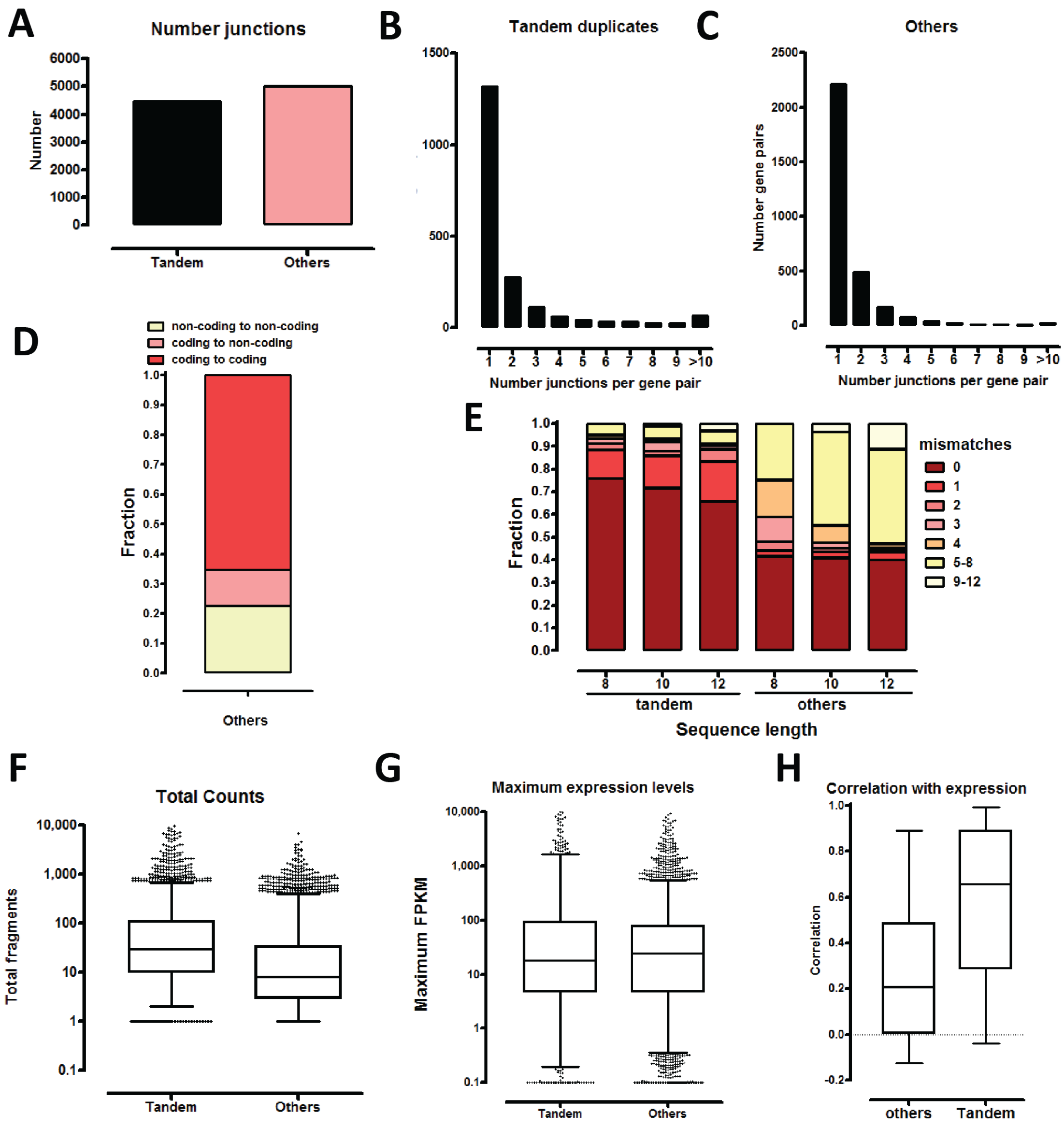


Figure 1.11: Splice junctions connecting known exons of different genes. (A) Number of junctions originating from pairs of tandem duplicate genes, and number of junctions originating from other genes. (B,C) Number junctions per gene pair. (D) Junctions connecting non-tandem duplicate genes according to whether they connect protein coding or non-coding genes (E) Minimal number mismatches between the donor or acceptor exon for gene A or gene B in a pair, respectively, and other downstream exons in gene A or upstream exons in gene B, respectively. TopHat requires at least 8bp on each side of a splice junction in order to map reads across it so lengths of 8, 10 and 12bp on each side of splice junctions were used. Note that 32 “tandem” junctions and 232 “others” junctions connected genes located on opposite genomic strands, and those are not included in the plot. (F) Total unique supporting fragment counts (G) Maximum expression level (in all cell lines and tissues) of the connected genes (H) Correlation between the minimum expression of genes in a pair and the distinct fragment counts mapping to the junctions in different samples.

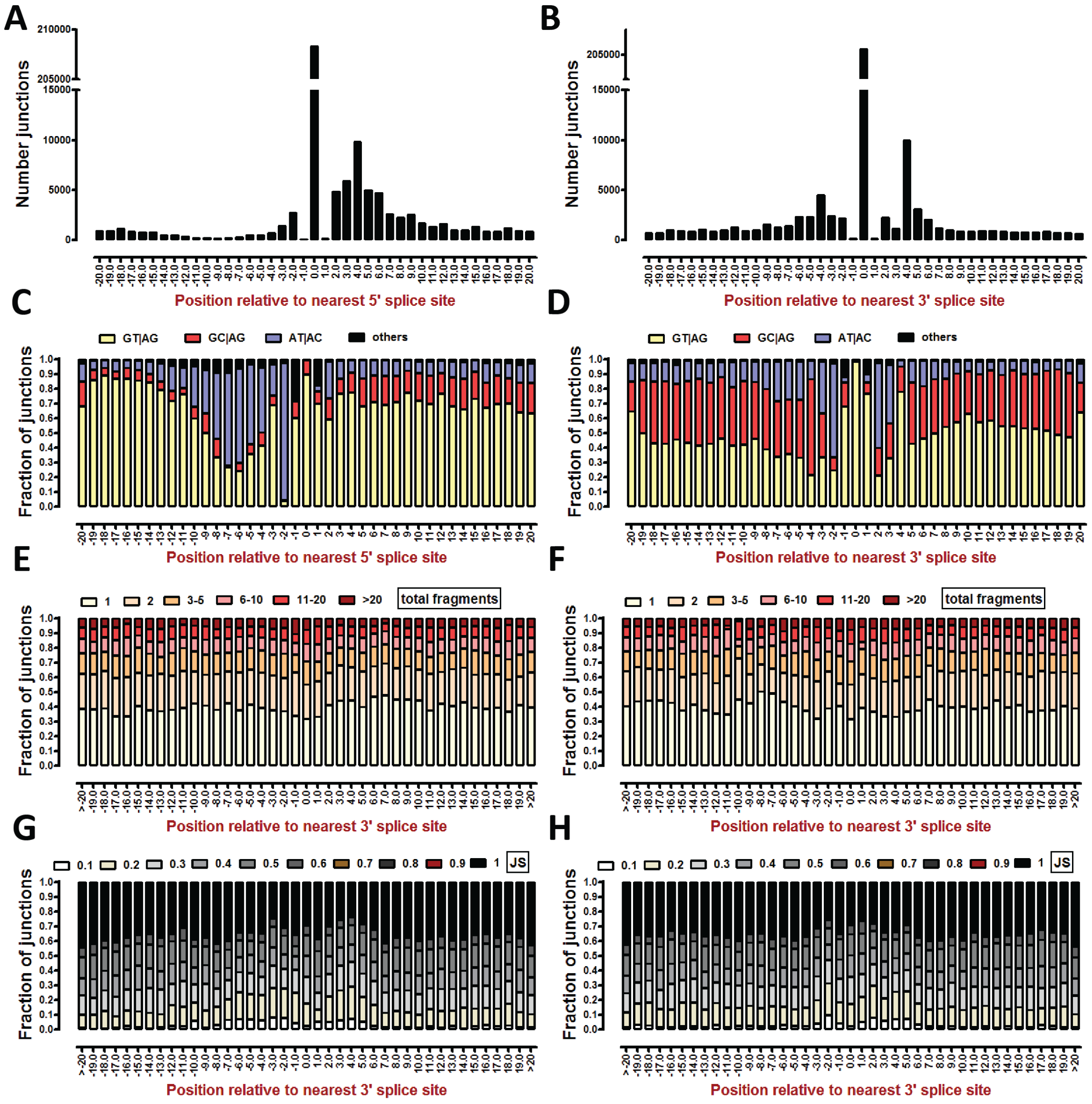


Figure 1.12: Distance of novel junction 5' and 3' ends to the nearest annotated splice. (A) 5' donor sites. (B) 3' acceptor sites (C) Distribution of canonical and non-canonical splice sites, 5' donor sites. (D) Distribution of canonical and non-canonical splice sites, 3' acceptor sites. (E) Total fragment support, 5' donor sites. (F) Total fragment support, 3' acceptor sites. (G) JS scores, 5' donor sites. (H) JS scores, 3' acceptor sites.

lished computational approaches for classifying intergenic transcripts (Guttman et al. 2010; Cabili et al. 2011). Briefly, for all intergenic multi-exonic transcript models in the final merged assembly, I first calculated the phylogenetic codon substitution frequency (PhyloCSF) score (Lin et al. 2011) and filtered out all transcripts with significantly constrained putative ORFs. I then scanned transcripts in all reading frames for the presence of protein domains annotated in the PFAM database (Punta et al. 2012) and removed all transcripts which contained such domains. The discarded transcripts were grouped together as transcripts of uncertain coding potential (TUCP) and analyzed separately. I identified 3591 candidate novel lincRNAs and 2592 TUCPs, numbers similar to those reported previously (Cabili et al. 2011; Guttman et al. 2009;

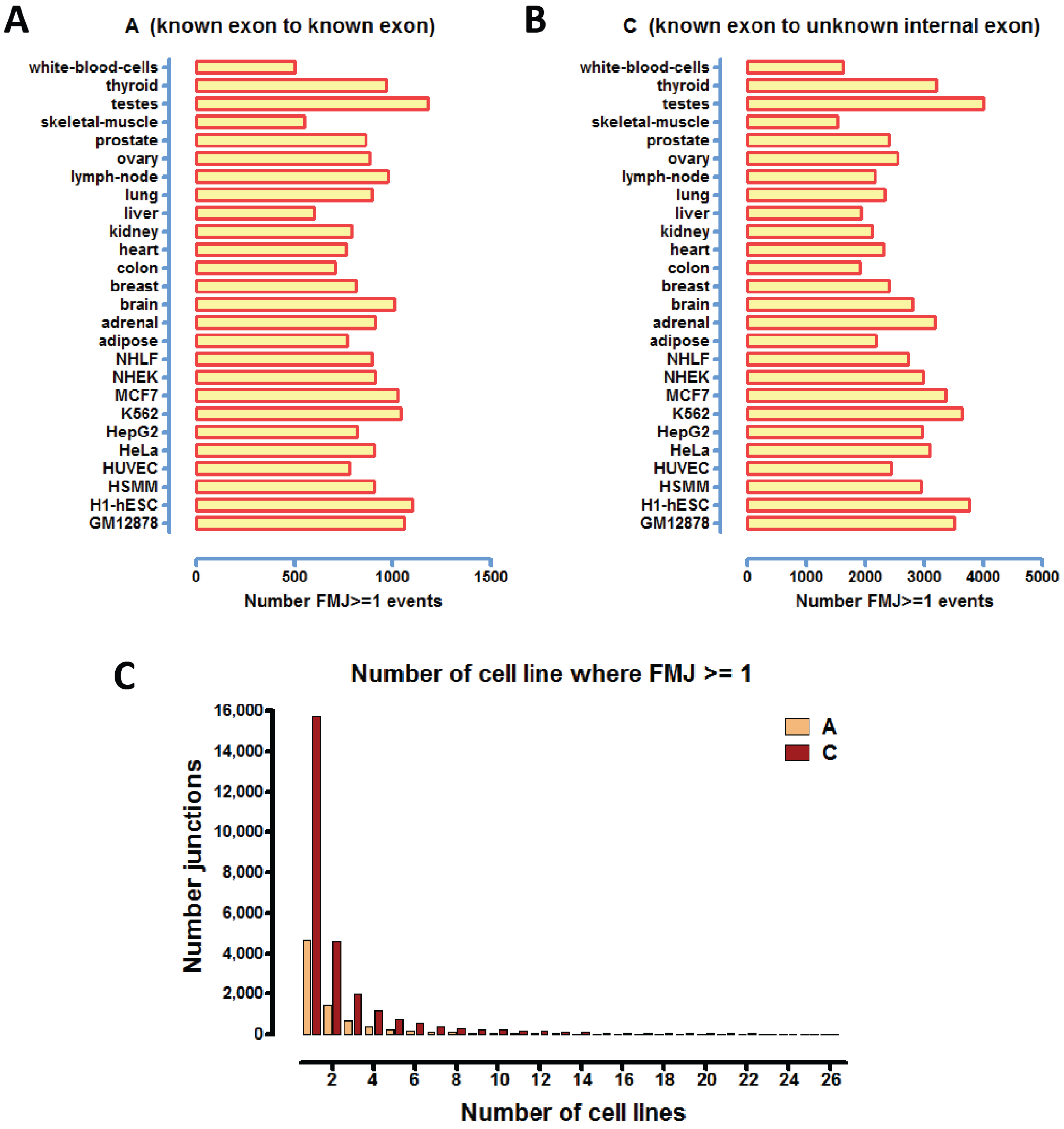
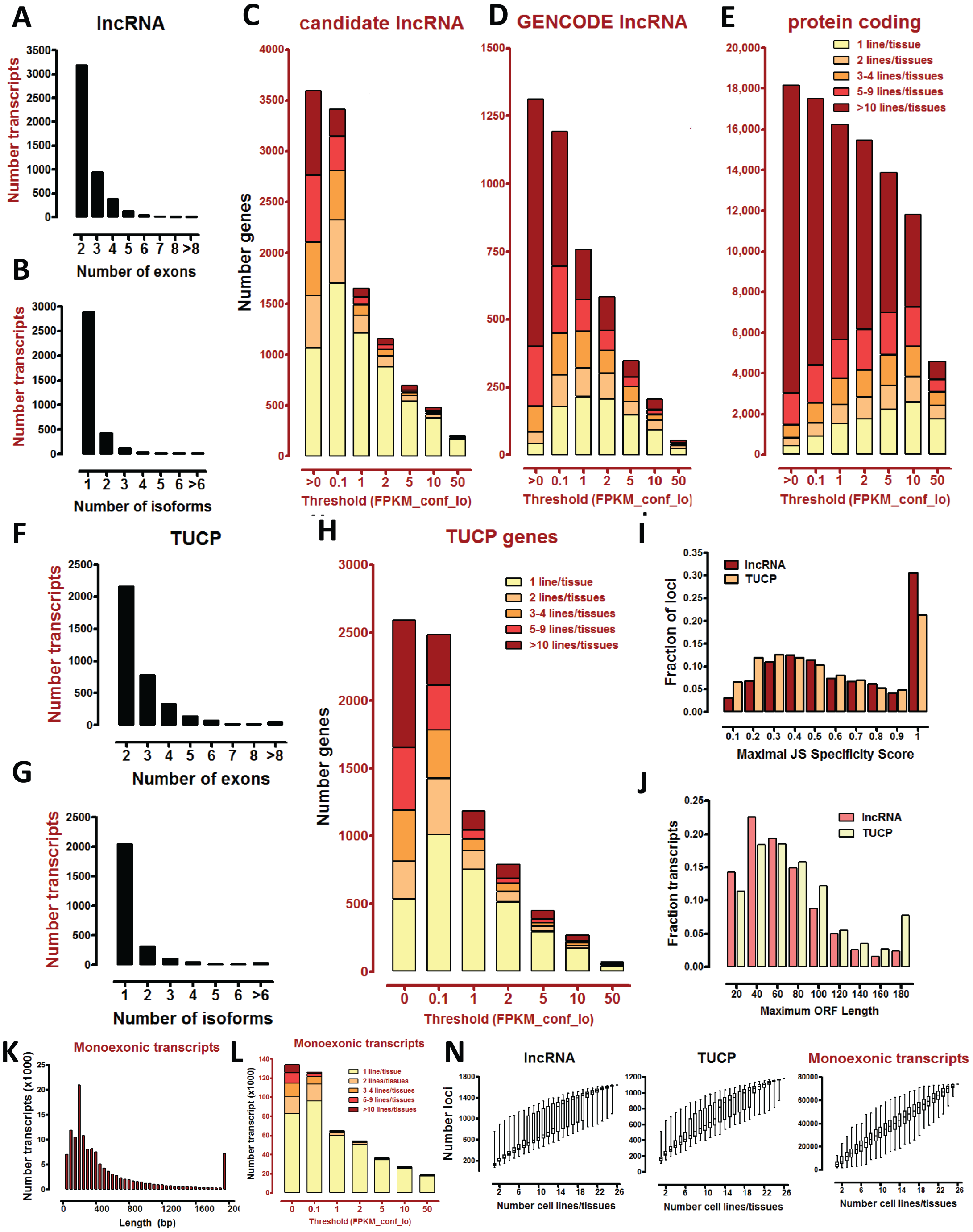


Figure 1.13: FMJ>=1 events. (A,B) FMJ>= 1 events per cell lines. (C) Number of cell lines in which each individual FMJ event is observed.

Khalil et al. 2010). In addition, the GENCODE v7 annotation contains 1368 annotated lincRNA genes which I analyzed in parallel. Most (67%) putative lincRNAs consisted of two exons and for 20% of them, more than one isoform was assembled (Figure 1.14A and B); for comparison, 68% of GENCODE v7 lincRNAs have 3 or more exons and 40% have multiple isoforms. I note that I also identified the longest ORF for each candidate lincRNA and TUCP and found ORFs of substantial length for significant fraction of both groups of transcripts (Figure 1.14J).

The majority of candidate lincRNAs were expressed at very low levels with only 695 (19%) expressed at FPKM$_{conf\_lo}$ greater than 5, and most only in one cell line/tissue (Figure 1.14C). The majority of protein coding genes pass that threshold (Figure 1.14E), and a higher proportion (26%) of GENCODE lincRNAs (Figure 1.14D). TUCP loci exhibited very similar characteristics in terms of number of exons and isoforms and expression patterns (Figure 1.14E,F


A
lncRNA
Number transcripts
3500
3000
2500
2000
1500
1000
500
0
2 3 4 5 6 7 8 >8
Number of exons
B
Number transcripts
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 >6
Number of isoforms
C
candidate lncRNA
Number genes
4000
3500
3000
2500
2000
1500
1000
500
0
>0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
D
GENCODE lncRNA
1500
1250
1000
750
500
250
0
>0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
E
protein coding
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
>0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
F
TUCP
Number transcripts
2500
2000
1500
1000
500
0
2 3 4 5 6 7 8 >8
Number of exons
G
Number transcripts
2500
2000
1500
1000
500
0
1 2 3 4 5 6 >6
Number of isoforms
H
TUCP genes
Number genes
3000
2500
2000
1500
1000
500
0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
I
Fraction of loci
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
lncRNA
TUCP
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Maximal JS Specificity Score
J
Fraction transcripts
0.25
0.20
0.15
0.10
0.05
0.00
lncRNA
TUCP
20 40 60 80 100 120 140 160 180
Maximum ORF Length
K
Monoexonic transcripts
Number transcripts (x1000)
25
20
15
10
5
0
0 400 800 1200 1600 200
Length (bp)
L
Monoexonic transcripts
Number transcript (x1000)
140
120
100
80
60
40
20
0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
N
lncRNA
Number loci
1800
1400
1000
600
200
2 6 10 14 18 22 26
Number cell lines/tissues
TUCP
1200
1000
800
600
400
200
0
2 6 10 14 18 22 26
Number cell lines/tissues
Monoexonic transcripts
80000
60000
40000
20000
0
2 6 10 14 18 22 26
Number cell lines/tissues

and G).

In addition to the set of spliced intergenic transcripts discussed above, the final merged assembly contained a very large number (≥130,000) of monoexonic transcripts, mostly shorter than 400 bp (Figure 1.14K). Due to the specifics of the merge procedure which fuses short overlapping fragments from multiple samples into a single larger one, and the short length of current RNA-seq reads, it is not possible to precisely define the start and end positions of these transcripts. A large number of them probably represent short spurious intergenic fragments yet there are still more than 20,000 expressed at a high-confidence threshold of more than 10 FPKM, strikingly almost always only in a single cell line (Figure 1.14M).

We are at present not certain how to interpret the nature of monoexonic loci as well as of candidate lincRNA and TUCP transcripts. There seems to be a large number of these transcripts expressed in highly cell type specific manner, therefore more are expected to be found if additional cell lines are sampled (Figure 1.14N). However they are mostly expressed at very low levels. Both candidate lincRNAs and TUCPs have high tissue specificity scores with lincRNAs being a little more tissue specific on average (Figure 1.14I). Each cell line and tissue expressed between 50 and 150 candidate lncRNAs at more than 1 FPKM$_{conf\_lo}$, with the notable exception of testes, where vastly more (more than 750) were detected (Figure 1.15, Figure 1.14N), and similar patterns were observed for TUCPs (Figure 1.16, Figure 1.14N). What the functional role and biological significance of all these transcripts is remains to be determined (See Discussion section for further discussion)

Combining all transcripts annotated in GENCODE v7 with novel isoforms of known genes, candidate lncRNAs, TUCPs, and monoexonic, I estimate that between 4 and 5 % of the human genome is expressed as exonic elements at ≥1 FPKM in at least one cell line or tissue in our dataset, and about 45% when introns are included (Figure 1.17).

1.2.7 Novel exons of annotated genes

After performing transcript-level quantification on the final merged assembly, I examined the nature and abundance of novel exons of known protein coding genes in the assembly. To this end, I assigned FPKM scores on exons derived from the sum of FPKMs of all individual transcripts containing them and classified exons according to their relation to the existing annotation (Figure 1.18A and B). The largest classes of novel exons were extensions of 5' and 3'UTRs. We expected this trend because of actual variation in the biology of transcription starts and processing (biology sources) and because of annotation imperfections at transcript ends (Hoskins et al. 2011; Carninci et al. 2006; Rach et al. 2011). The next most frequently observed novelties arise from extensions or shortenings of internal exons, consistent with our previous observation of a large number of novel splice sites located in introns and previously annotated exons.

Figure 1.14 *(preceding page)*: Identification of novel intergenic transcribed loci (lincRNAs and TUCPs). (A) Number of exons for candidate lncRNA genes. (B) Number of isoforms for candidate lncRNA genes (C), (D), (E) Expression of candidate lncRNA genes, annotated lncRNA genes and protein coding genes for comparison. While protein coding genes are widely expressed at high levels, annotated lncRNA are mostly expressed at low levels, and candidate novel lncRNAs are expressed at even lower levels and in few cell lines/tissues. FPKM$_{conf\_lo}$ thresholds were used for stringency purposes. (F), (G), (H) Transcripts of Uncertain Coding Potential (TUCP) are broadly similar in their characteristics and expression patterns to candidate lncRNAs. (I) Candidate lncRNAs are slightly more tissue-specific than TUCPs. (J) Substantial numbers of both lncRNAs and TUCPs contain ORFs of considerable length, with slightly more such ORFs observed in TUCPs (K) Large numbers of monoexonic intergenic transcripts are detected, mostly below 400bp of length (see text for detailed discussion). (L) Expression patterns of monoexonic intergenic transcripts. While mostly of low abundance and observed only in individual cell lines/tissues, there are still thousands of such transcripts expressed at significant levels, typically only in one cell lines (though again, usually in a single cell line or tissue). (N) Cumulative detection of novel intergenic transcripts. Threshold of FPKM$_{conf\_lo}$ ≥ 1 was used. Note the inflection of saturation caused by the testes sample in the lncRNA and TUCP plots.

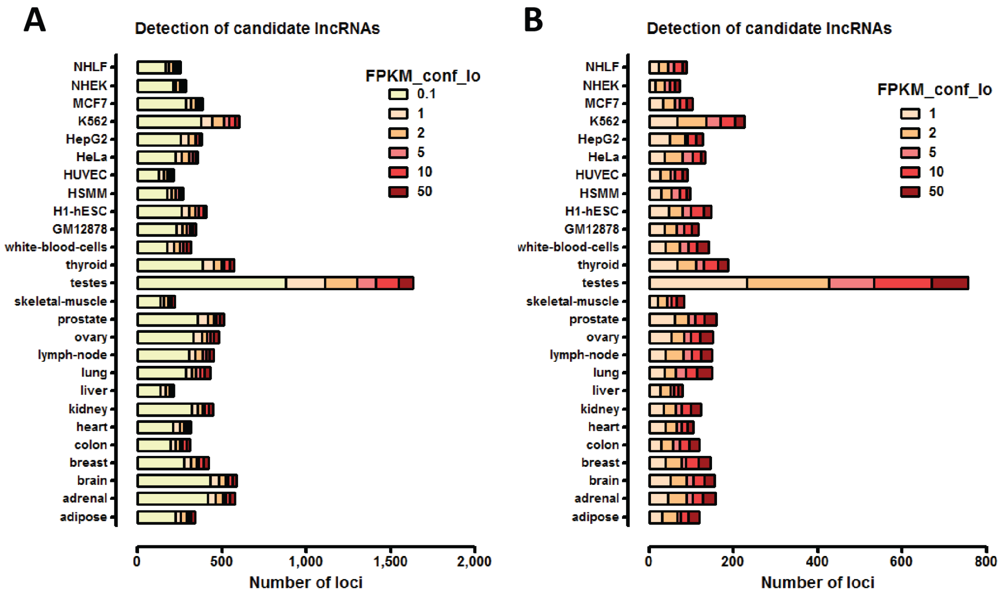


Figure 1.15: Expression of candidate lncRNA across cell lines. (A) At 0.1 FPKM$_{conf\_lo}$. (B) At 1 FPKM$_{conf\_lo}$ threshold.

Completely novel exons are rare, with evidence for 583 internal exons, 1279 novel 5 exons and 999 novel 3 exons at an FPKM cut-off of 5, for a total of 17,197 novel exons (Figure 1.18A).

1.2.8 Splicing isoform expression of protein coding genes

The final transcript set contained 42,775 novel isoforms of protein coding in addition to those already present in GENCODE. I examined the

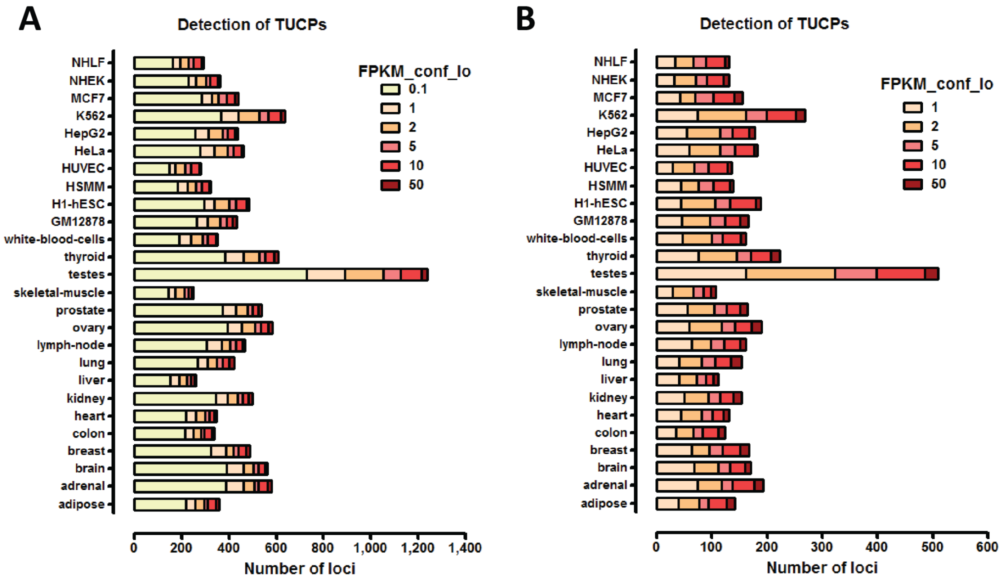


Figure 1.16: Expression of TUCP transcripts across cell lines. (A) At 0.1 FPKM$_{conf\_lo}$. (B) At 1 FPKM$_{conf\_lo}$ threshold.

expression patterns of annotated and novel isoforms and found that novel isoforms are on average expressed at lower levels than annotated ones (Figure 1.18C and D), yet they are similarly widely expressed (Figure 1.18I). Previous studies have suggested that almost all human genes undergo alternative splicing (Wang et al. 2008; Pan et al. 2008); however, alternative splicing is a noisy process and a large number of low-abundance isoforms might be generated without much biological relevance, so I aimed to understand isoform expression as a function of abundance estimates. At a conservative threshold of 5 FPKM$_{conf\_lo}$, 28,638 annotated isoforms and 3,374 novel ones were detected; this is an underestimate since where quantification was unreliable due to identifiability and other numerical issues, I assigned FPKM of 0 to all transcripts of a gene. Large numbers of isoforms were detected at lower thresholds and isoform detection did not clearly saturate at the level of 5 FPKM neither for annotated not for novel isoforms (Figure 1.185E and F). Using the same 5 FPKM$_{conf\_lo}$ threshold, I detect multiple annotated isoforms for 7,742 protein coding genes, and a novel isoform for 2,717 protein coding genes (Figure 1.18G and H), numbers that increase or decrease as thresholds are correspondingly relaxed or tightened.

Because transcription and splicing of very highly abundant genes can generate aberrant noise products that are still highly abundant when compared to rarely transcribed genes in the same cell lines/tissue, a more informative metric for evaluating alternative splicing isoform abundance is the ratio of a given isoform's abundance to that of the major isoform for the gene (fraction of major isoform, FMI). Across all cell lines and tissues, the median FMI value for the second most abundant isoform was stably between 0.4 and 0.5, between 0.1 and 0.2 for the third most abundant isoforms, and below 0.1 for lower-ranked isoforms (Figure 1.18J). FMI values of novel isoforms tend to be lower. For example, when ranked second, their FMI was below 0.2 rather than 0.4.

A different splicing isoform may be the major isoform in different cell lines, which is here referred to as major isoform switch. To determine how widespread this phenomenon is, I counted the different major isoforms for each gene in all cell lines and tissues at different detection cutoffs. Using the 5 FPKM$_{conf\_lo}$ threshold, I estimate that 7,541 genes express only a single major isoforms while 5,749 express multiple major isoforms, with 2308 expressing 3 or more (Figure 1.18K). For every pair of cell lines/tissues, between 600 and 2,800 genes switched their major isoform (Figure 1.1917).

The observations outlined above suggest a larger expression diversity on the level of individual transcripts than on the gene level. Indeed, when expression specificity was measured using the JS tissue specificity metric, it was usually higher for of individual transcripts than for the genes they belong to (Figure 1.18P).

1.2.9 Impact of splicing isoforms on protein sequence

The impact of alternative isoform expression on protein function depends on the difference in

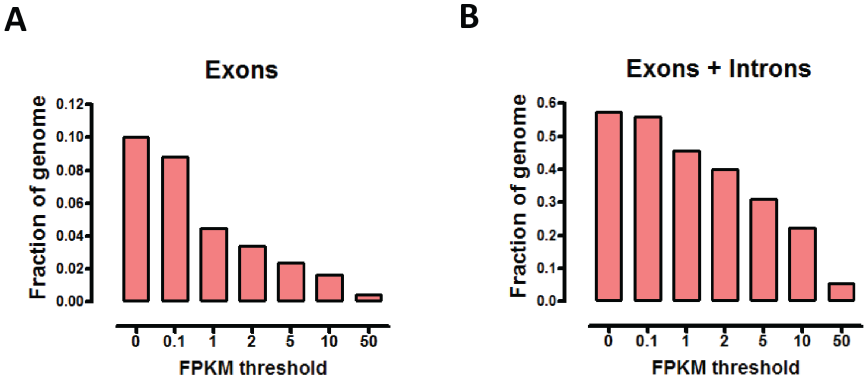
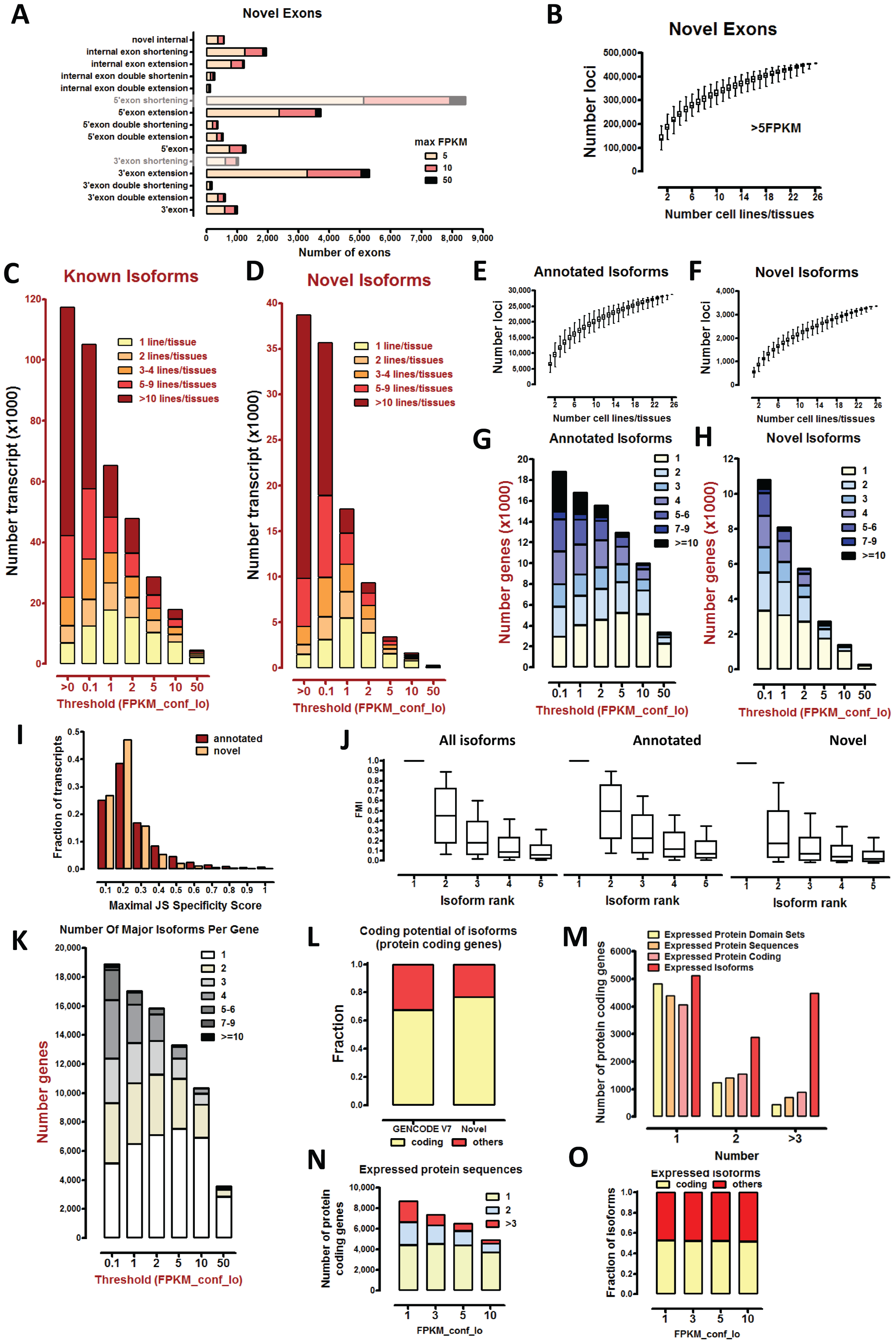
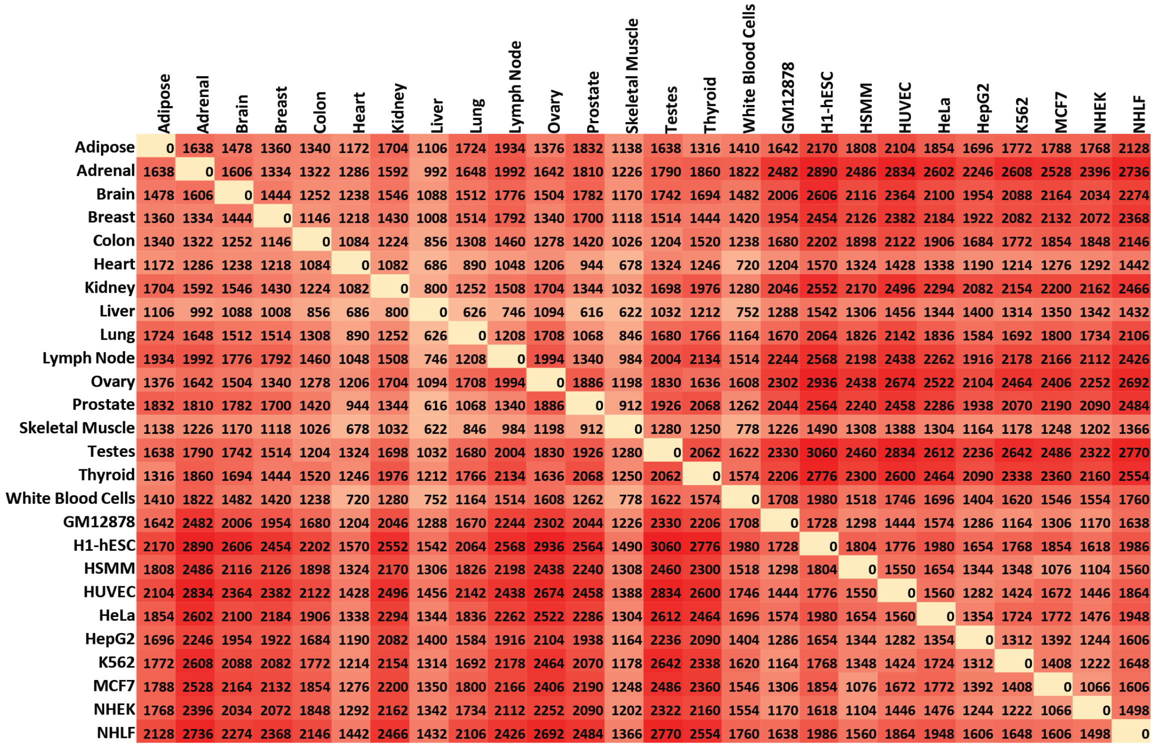


Figure 1.17: Fraction of genome expressed at a given FPKM threshold in at least one cell line or tissue, with (A) or without (B) the inclusion of introns.


A
Novel Exons
novel internal
internal exon shortening
internal exon extension
internal exon double shortenin
internal exon double extension
5'exon shortening
5'exon extension
5'exon double shortening
5'exon double extension
5'exon
3'exon shortening
3'exon extension
3'exon double shortening
3'exon double extension
3'exon
max FPKM
5
10
50
0 1,000 2,000 3,000 4,000 5,000 6,000 7,000 8,000 9,000
Number of exons
B
Novel Exons
Number loci
500,000 400,000 300,000 200,000 100,000 0
>5FPKM
2 6 10 14 18 22 26
Number cell lines/tissues
C
Known Isoforms
Number transcript (x1000)
120 100 80 60 40 20 0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
>0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
D
Novel Isoforms
Number transcript (x1000)
40 35 30 25 20 15 10 5 0
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
>0 0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
E
Annotated Isoforms
Number loci
30,000 25,000 20,000 15,000 10,000 5,000 0
2 6 10 14 18 22 26
Number cell lines/tissues
F
Novel Isoforms
Number loci
4,000 3,000 2,000 1,000 0
2 6 10 14 18 22 26
Number cell lines/tissues
G
Annotated Isoforms
Number genes (x1000)
20 18 16 14 12 10 8 6 4 2 0
1 2 3 4 5-6 7-9 >=10
0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
H
Novel Isoforms
Number genes (x1000)
12 10 8 6 4 2 0
1 2 3 4 5-6 7-9 >=10
0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
I
Fraction of transcripts
0.5 0.4 0.3 0.2 0.1 0.0
annotated
novel
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Maximal JS Specificity Score
J
All isoforms
Annotated
Novel
FMI
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
1 2 3 4 5
Isoform rank
1 2 3 4 5
Isoform rank
1 2 3 4 5
Isoform rank
K
Number Of Major Isoforms Per Gene
Number genes
20,000 18,000 16,000 14,000 12,000 10,000 8,000 6,000 4,000 2,000 0
1 2 3 4 5-6 7-9 >=10
0.1 1 2 5 10 50
Threshold (FPKM_conf_lo)
L
Coding potential of isoforms
(protein coding genes)
Fraction
1.0 0.8 0.6 0.4 0.2 0.0
GENCODE V7 Novel
coding others
M
Expressed Protein Domain Sets
Expressed Protein Sequences
Expressed Protein Coding
Expressed Isoforms
Number of protein coding genes
6000 5000 4000 3000 2000 1000 0
1 2 >3
Number
N
Expressed protein sequences
Number of protein coding genes
10,000 8,000 6,000 4,000 2,000 0
1 2 >3
1 3 5 10
FPKM_conf_lo
O
Expressed isoforms
coding others
Fraction of isoforms
1.0 0.8 0.6 0.4 0.2 0.0
1 3 5 10
FPKM_conf_lo

ORFs from alternative isoforms. Some isoforms with premature stop codons will likely be subject to nonsense-mediated decay (NMD) (Chang et al. 2007) and while regulatory roles for NMD alternative splicing events has been proposed (Cuccurese et al. 2005; Green et al. 2003; McGlincy & Smith 2008) many will likely have little biological impact. Similar expectations apply to transcripts with very large retained introns. More than a quarter of protein coding gene isoforms in the GENCODE V7 annotation are designated as non-coding for such reasons. I assigned novel isoforms into coding and non-coding following a similar requirement that protein coding isoforms contain an ORF and the ORF does not finish more than 50 bp downstream of the 3' exon splice junction. A similar but slightly higher (likely because stringent filters on retained intron transcripts were applied) fraction of novel isoforms was classified as coding in this manner (Figure 1.18L). Next, I examined the impact of expressed isoforms on the coding sequence of each gene (Figure 1.18M). I calculated four quantities for each gene at a given FPKM threshold: 1) the total number of isoforms expressed, 2) the number of protein coding isoforms expressed (excluding non-coding ones), 3) the number of different protein sequences expressed (if two isoforms only differ in such a way that there protein translation are the same, they were counted as one), and 4) the number of protein domain sets expressed (I scanned each transcript for the presence of domains annotated in the PFAM database; if two isoforms produced the number, type, order and sequence of PFAM domains, they were counted as one). At a conservative 5 FPKM$_{conf\_lo}$ threshold, 2,106 genes express multiple protein sequences, and PFAM

Figure 1.18 ***(preceding page)*****: Expression of annotated and novel isoforms of protein coding genes.** Genes and transcripts for which isoform-level quantification failed were excluded in all cases except for exons in (A) and (B). (A) New exons identified classified according to their relation to the existing annotation. Shortened 3' and 5'UTRs are shaded because the majority of these are likely to be the result of incomplete transcript assembly due to low read coverage. Exon FPKMs were defined as the sum of FPKMs for all individual transcripts containing the exon. The maximum such estimate for all samples is shown. (B) Cumulative detection of novel exons. C) and (D) Expression patterns of annotated and novel isoforms of protein coding genes. Annotated isoforms are on average more highly expressed than novel ones, however, novel ones are mostly as widely expressed as annotated ones. (E), (F) Cumulative detection of annotated and novel isoforms. (G), (H) Number of expressed annotated and novel isoforms per genes as a function of abundance levels. The plot shows the number of genes with a number of isoforms indicated by the color code expressed at level above the FPKM$_{conf\_lo}$ thresholds shown. (I). JS specificity scores for annotated and novel isoforms. (J) Isoform abundance as a fraction of the major isoform (FMI) for a gene. For each gene and each cell line/tissue, individual transcripts are ranked by their FPKM expression estimates. The isoform with the highest FPKM is the major one, the distribution of the ratio between the lower ranked isoforms and the major one for all genes and conditions is shown. (K) Number of major isoforms per gene. Genes may express different major isoforms in different cell lines; such events are more confidently identified when the expression level of the genes is high. Shown is the number of major isoforms per gene as indicated by the color code at the indicated FPKM$_{conf\_lo}$ thresholds for the major isoform. L) Coding potential of annotated and novel isoforms. The "other" category contains transcripts classified as NMD products, retained intron transcripts and other non-coding isoforms of coding genes. (M) Impact of isoforms on protein sequence. For each gene, the number of expressed isoforms, expressed protein coding isoforms (not all isoforms are protein coding), expressed protein sequences (some isoforms may only differ in their non-coding regions), and expressed domain sets was calculated. Domain sets were defined by scanning all transcripts for PFAM protein domains and counting as distinct only isoforms that differ in the identity and sequence of their protein domains. A threshold of 5 FPKM$_{conf\_lo}$ was used for this plot. (N) Number of expressed protein sequences as function of expression levels. The color code indicates the number of genes with 1, 2 or 3 and more protein sequences detected at the indicated FPKM$_{conf\_lo}$ threshold. (O) Fraction of expressed transcripts detected coding for proteins as a function of expression levels. (P) Expression specificity (JS score) of individual transcripts and the expression of the corresponding genes (protein coding genes only).



	Adipose	Adrenal	Brain	Breast	Colon	Heart	Kidney	Liver	Lung	Lymph Node	Ovary	Prostate	Skeletal Muscle	Testes	Thyroid	White Blood Cells	GM12878	H1-hESC	HSMM	HUVEC	HeLa	HepG2	K562	MCF7	NHEK	NHLF
Adipose	0	1638	1478	1360	1340	1172	1704	1106	1724	1934	1376	1832	1138	1638	1316	1410	1642	2170	1808	2104	1854	1696	1772	1788	1768	2128
Adrenal	1638	0	1606	1334	1322	1286	1592	992	1648	1992	1642	1810	1226	1790	1860	1822	2482	2890	2486	2834	2602	2246	2608	2528	2396	2736
Brain	1478	1606	0	1444	1252	1238	1546	1088	1512	1776	1504	1782	1170	1742	1694	1482	2006	2606	2116	2364	2100	1954	2088	2164	2034	2274
Breast	1360	1334	1444	0	1146	1218	1430	1008	1514	1792	1340	1700	1118	1514	1444	1420	1954	2454	2126	2382	2184	1922	2082	2132	2072	2368
Colon	1340	1322	1252	1146	0	1084	1224	856	1308	1460	1278	1420	1026	1204	1520	1238	1680	2202	1898	2122	1906	1684	1772	1854	1848	2146
Heart	1172	1286	1238	1218	1084	0	1082	686	890	1048	1206	944	678	1324	1246	720	1204	1570	1324	1428	1338	1190	1214	1276	1292	1442
Kidney	1704	1592	1546	1430	1224	1082	0	800	1252	1508	1704	1344	1032	1698	1976	1280	2046	2552	2170	2496	2294	2082	2154	2200	2162	2466
Liver	1106	992	1088	1008	856	686	800	0	626	746	1094	616	622	1032	1212	752	1288	1542	1306	1456	1344	1400	1314	1350	1342	1432
Lung	1724	1648	1512	1514	1308	890	1252	626	0	1208	1708	1068	846	1680	1766	1164	1670	2064	1826	2142	1836	1584	1692	1800	1734	2106
Lymph Node	1934	1992	1776	1792	1460	1048	1508	746	1208	0	1994	1340	984	2004	2134	1514	2244	2568	2198	2438	2262	1916	2178	2166	2112	2426
Ovary	1376	1642	1504	1340	1278	1206	1704	1094	1708	1994	0	1886	1198	1830	1636	1608	2302	2936	2438	2674	2522	2104	2464	2406	2252	2692
Prostate	1832	1810	1782	1700	1420	944	1344	616	1068	1340	1886	0	912	1926	2068	1262	2044	2564	2240	2458	2286	1938	2070	2190	2090	2484
Skeletal Muscle	1138	1226	1170	1118	1026	678	1032	622	846	984	1198	912	0	1280	1250	778	1226	1490	1308	1388	1304	1164	1178	1248	1202	1366
Testes	1638	1790	1742	1514	1204	1324	1698	1032	1680	2004	1830	1926	1280	0	2062	1622	2330	3060	2460	2834	2612	2236	2642	2486	2322	2770
Thyroid	1316	1860	1694	1444	1520	1246	1976	1212	1766	2134	1636	2068	1250	2062	0	1574	2206	2776	2300	2600	2464	2090	2338	2360	2160	2554
White Blood Cells	1410	1822	1482	1420	1238	720	1280	752	1164	1514	1608	1262	778	1622	1574	0	1708	1980	1518	1746	1696	1404	1620	1546	1554	1760
GM12878	1642	2482	2006	1954	1680	1204	2046	1288	1670	2244	2302	2044	1226	2330	2206	1708	0	1728	1298	1444	1574	1286	1164	1306	1170	1638
H1-hESC	2170	2890	2606	2454	2202	1570	2552	1542	2064	2568	2936	2564	1490	3060	2776	1980	1728	0	1804	1776	1980	1654	1768	1854	1618	1986
HSMM	1808	2486	2116	2126	1898	1324	2170	1306	1826	2198	2438	2240	1308	2460	2300	1518	1298	1804	0	1550	1654	1344	1348	1076	1104	1560
HUVEC	2104	2834	2364	2382	2122	1428	2496	1456	2142	2438	2674	2458	1388	2834	2600	1746	1444	1776	1550	0	1560	1282	1424	1672	1446	1864
HeLa	1854	2602	2100	2184	1906	1338	2294	1344	1836	2262	2522	2286	1304	2612	2464	1696	1574	1980	1654	1560	0	1354	1724	1772	1476	1948
HepG2	1696	2246	1954	1922	1684	1190	2082	1400	1584	1916	2104	1938	1164	2236	2090	1404	1286	1654	1344	1282	1354	0	1312	1392	1244	1606
K562	1772	2608	2088	2082	1772	1214	2154	1314	1692	2178	2464	2070	1178	2642	2338	1620	1164	1768	1348	1424	1724	1312	0	1408	1222	1648
MCF7	1788	2528	2164	2132	1854	1276	2200	1350	1800	2166	2406	2190	1248	2486	2360	1546	1306	1854	1076	1672	1772	1392	1408	0	1066	1606
NHEK	1768	2396	2034	2072	1848	1292	2162	1342	1734	2112	2252	2090	1202	2322	2160	1554	1170	1618	1104	1446	1476	1244	1222	1066	0	1498
NHLF	2128	2736	2274	2368	2146	1442	2466	1432	2106	2426	2692	2484	1366	2770	2554	1760	1638	1986	1560	1864	1948	1606	1648	1606	1498	0

Figure 1.19: Major isoform switch events. Major isoform switch events between cell lines at an FPKM$_{conf\_lo}$ threshold of 5. Shown is the number of genes for which the major isoform is different in each pair of cell line/tissues.

domains are affected by alternative isoform expression for 1,674 (Figure 1.18M). Relaxing the FPKM threshold results in higher estimates for the number of such genes (Figure 1.18N).

While performing this analysis, I noticed that approximately half of all expressed RNA isoforms, irrespective of detection threshold, are non-coding, a higher fraction than expected based on the fraction of such transcripts in the annotation (Figure 1.18O). This is a somewhat puzzling observation since the naive expectation would be that non-coding isoforms are mostly the result of transcriptional noise and that NMD isoforms are degraded relatively quickly, therefore they would be more frequently seen at low detection thresholds. Examples of such transcripts with regulatory function are known (Le Guiner et al. 2003; Sureau et al. 2001; Wollerton et al. 2004) so there may be biological functionality behind this observation. Further investigation will be needed to better understand this phenomenon.

1.2.10 Reconstruction of primary miRNA transcripts

We investigated whether any of the novel transcripts not in GENCODE V7 could correspond to miRNA primary transcripts. We compared the 2,104 miRNAs in miRBase V18 (Kozomara A & Griffiths-Jones 2011) to the GENCODE annotation and found that 57% were in the exons (9%) and introns (48%) of sense transcripts longer than 125 bp (Figure 1.20A). The inclusion of merged and filtered GENCODE+Cufflinks transcripts increases the percentage of overlapping known miRNAs to 59%, with an increase of microRNAs in exons to 15% (Figure 1.20A). However, it is likely that only a subset of miRBase microRNAs are expressed in our cell types and tissues. We therefore measured the expression of microRNAs in six ENCODE cell lines using Nanostring (Wyman et al. 2011) as described in the methods. We found 93 miRNAs expressed highly (≥200 counts) in one or more of the six cell lines. Whereas 57% of these miRNAs (9% exonic) overlapped a sense GENCODE transcript, we found that 62% (23% exonic) overlapped a merged and filtered GENCODE+Cufflinks sense transcript (Figure 1.20A). Given that a single Nanostring probe can map to more than one genomic location when only a subset may be transcribed even though we count all locations, our numbers are likely be an underestimate of the fraction of miRNAs that have evidence of primary transcripts in our RNA-seq data.

1.2.11 Complexity of TAF1 binding patterns in the human genome

Initiation of transcription at gene promoters is a primary point of regulation of transcriptional activity in eukaryotic cells, with many genes known to initiate transcription from multiple promoters (Landry et al. 2003; Wu et al. 1999; Tasic et al. 2002). For this reason, the characterization of the identity and activity of novel intergenic, novel alternative as well as annotated promoters is of great interest. To this end, we generated genome-wide ChIP-seq profiles for the TAF1 subunit of the TFIID general transcription factor, a component of the RNA Pol2 pre-initiation complex (PIC) (Buratowski et al. 1989; Näär et al. 2001), in GM12878, H1-hESC, HeLa, HepG2 and K562 cells. TAF1 binding is expected to mark all active promoters transcribed by RNA Pol2 and therefore be a good marker for discovery of new promoters.

I called TAF1 binding sites with ERANGE 4.0 (Johnson et al. 2007, `http://woldlab.caltech.edu/wiki/`) using relatively relaxed thresholds (see Methods) and calculated expression values in FPKM for all TSSs in GENCODE and the final merged set of transcripts models by summing the FPKM values for all transcripts sharing a given TSSs. I called between 9,000 and 20,000 TAF1 binding sites in individual replicates, with K562 and H1-hESC having the highest number (Figure 1.21A). The distribution of individual TAF1 binding site summits centered right on top of GENCODE TSS (Figure 1.21C) and TAF1 loading correlated positively with gene expression. However, I noticed that not all expressed TSSs are marked by TAF1 loading, with up to 25 % of TSSs expressed at more than 100 FPKM in H1-hESC not having a TAF1 binding sites, a proportion that grows with the decrease of expression levels (Figure 1.22A). In most cases this is not due to these TSSs containing repetitive sequences and sequencing reads failing to align as a result (Figure 1.22B). In the other cell lines we assayed, fewer TAF1 binding sites were called (Figure 1.21A) and an even higher number of highly expressed TSSs did not have a TAF1 binding site (Figure 1.22C). This could be due to technical variability in ChIP strength; however, the highest number of binding sites we identified in a single GM12878 TAF1 ChIP-seq replicate was less than 10,000, even though 12 different biological and technical replicates were generated, and similar results were obtained with two other lymphoblastoid cell lines, GM12891 and GM12892 (Figure 1.21B), which makes this explanation unlikely. It has been suggested that in certain cell lines and tissues, the composition of the PIC components varies (Deato & Tjian 2007; Goodrich & Tjian 2010; D'Alessio et al. 2011) which could explain the consistent differences between TAF1 binding observed in different cell lines, yet there was no negative correlation between TAF1 expression and the number of TAF1 binding sites. The other explanation is that there exists a class of promoters in the initiation of which TAF1 does not play a role. This is in agreement with previous tiling array-based studies profiling TAF1 distribution genome-wide (Kim et al. 2005).

In order to compare TAF1 binding across cell lines, I merged TAF1 binding sites summits that were close to each other from individual replicates across all cell lines (see Methods for details) and examined the binding patterns of the resulting set of 44,702 sites. 12,585 summits were within 100 bp of a GENCODE V7 TSS, additional 7,811 and 6,907 within 1 kb of a TSS, 8,538 were more than 1 kb upstream of the closest TSS and 7,864 downstream of it. Thus the majority of sites were associated with or close to known TSSs yet a sizeable fraction was located away from any known TSS. The strength of TAF1 binding as measured in RPM decreased with distance away from annotated TSSs with the majority of intergenic and intragenic sites being weaker than those close to TSSs (Figure 1.22D).

Using the merged set of TAF1 binding sites, I sought to determine whether the lack or presence of TAF1 binding was consistent between cell lines. To this end I compiled the set of all TSSs expressed at more than 1 FPKM$_{conf\_lo}$ in each of the five cell lines for which we have TAF1 binding data and compared the presence or absence of TAF1 binding by clustering the resulting data matrix. A large cluster of TSSs without TAF1 binding in all cell lines emerged from this analysis, and strikingly, it was also the group of TSSs without CpG island in their vicinity (Figure 1.22C)

We then asked how many of the intergenic or intragenic TAF1 sites we could explain with gene models derived from RNA-seq data. For this purpose we used a merged set of gene models generated without applying expression level filtering on the input data sets. About 20 % of TAF1 sites located more than 1 kb away from a TSS in each direction had a candidate novel TSS located within 1 kb of the peak summit, and close to 40 % of TAF1 sites between 100 bp and 1 kb upstream of known TSSs had candidate novel TSS within 100 bp (Figure 1.22D). Very few TAF1 sites downstream of TSS had a corresponding candidate TSS models, however, I note that the merge procedure is heavily biased against shortening of 5' exons and this might be the explanation. The other 80 % of intergenic and intragenic TAF1 sites may either be the result of RNA-seq assemblies bypassing the promoter region or falling short of it, or they may represent "shadows" of promoter looping to enhancer regions and not real promoters. The latter possibility is consistent with the lower strength of ChIP signal characteristic of these sites.

I grouped the sites into 9 groups depending

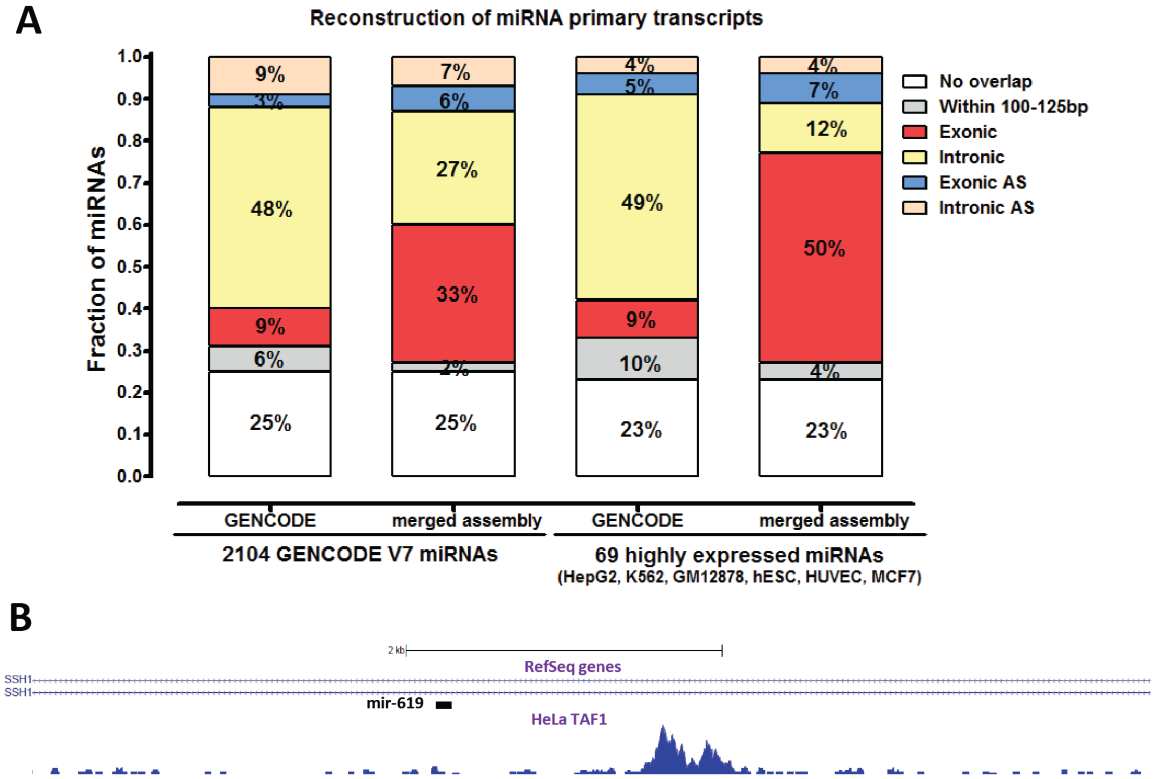


Figure 1.20: Reconstruction of primary miRNA transcripts. (A) Comparison of GENCODE and RNA-seq augmented annotations (merged assembly) to 1523 known miRNAs for evidence of primary miRNA transcripts (left) and to 69 highly expressed miRNAs (in at least one of GM12878, K562, human ES and HepG2, assayed with nanoString). Mature miRNAs were intersected with exonic and intronic regions of sense and antisense transcripts. The fraction of miRNAs for which a putative primary transcript was present increases in the merged assembly compared to GENCODE v7, which is even more pronounced when only the highly expressed miRNAs are considered. (B) Putative intronic promoter for mir-619, which is located within an intron of the SSH1 gene. A TAF1 site is situated upstream of the miRNA, suggesting the miRNA may be transcribed independently from the gene from its own promoter.

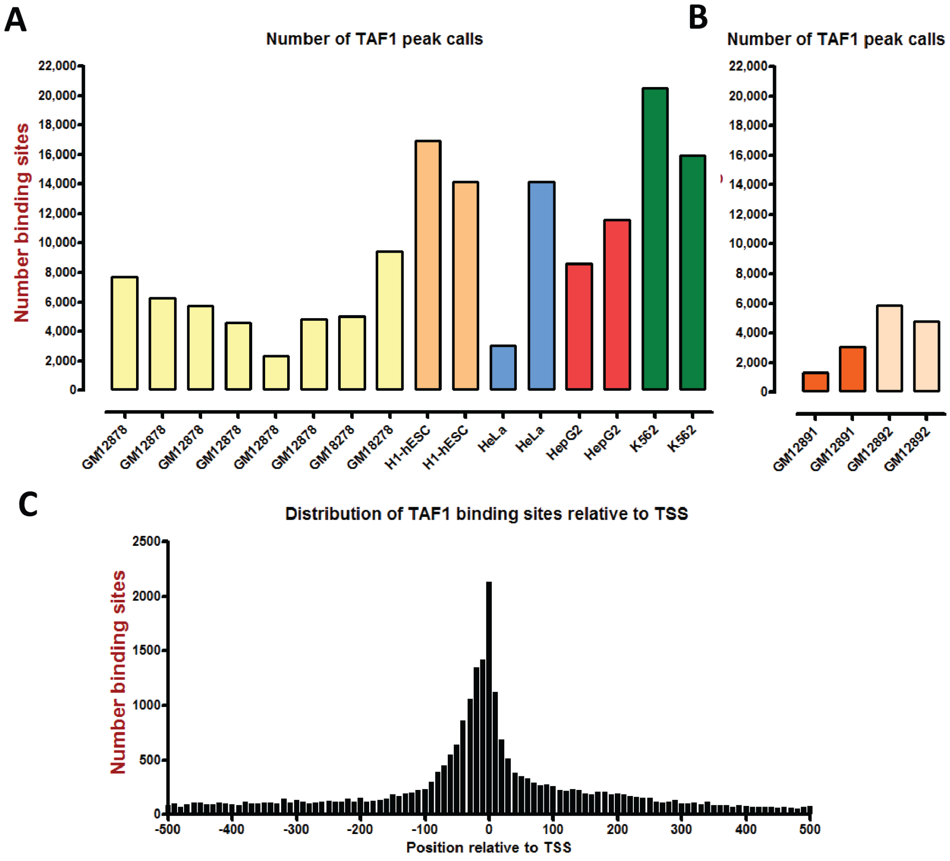
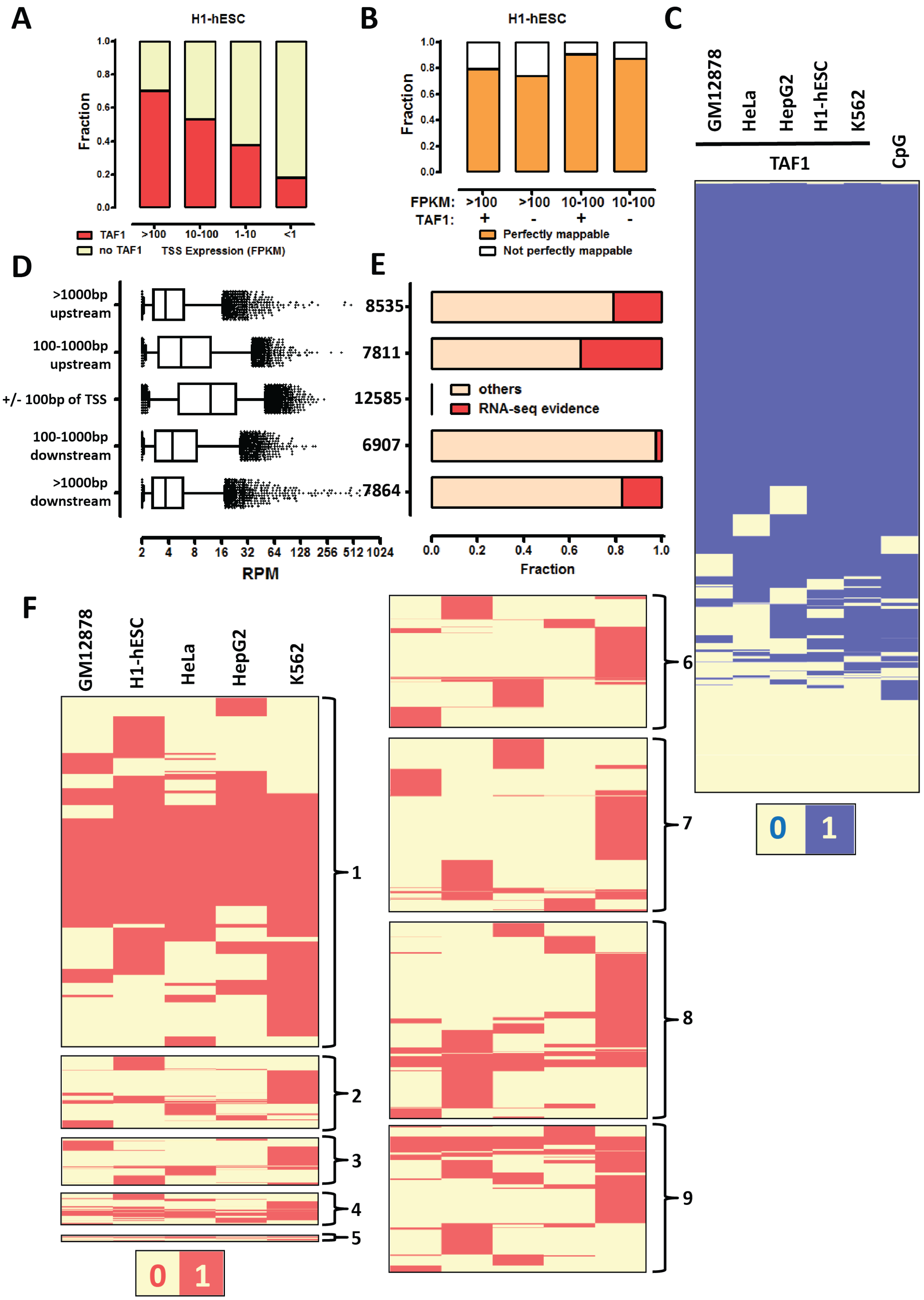


Figure 1.21: TAF1 binding sites. (A) Number of peak calls for individual replicates. (B) Number of peak calls for GM12891 and GM12892 cells, not used for subsequent analysis. (C) Distribution of TAF1 binding sites (combined set) relative to GENCODE V7 TSSs.

on their position relative to the GENCODE V7 reference and the set of RNA-seq-derived transcript models, and clustered them according to their presence or absence in each cell type. (Figure Fig.4.23E). Among the largest group (group 1), the GENCODE V7 TSS-associated sites, a large core of sites present and TSSs utilized in all cell lines is observed. In contrast, the sites located away from annotated TSS tend to be more highly cell type specific and present only in one cell line (groups 2-9).

1.2.12 Identification of novel 5' Transcription Start Sites

RNA-seq measurements have the potential to identify novel transcription start sites, however, there are several issue with the approach that need to be considered and that highlight the need for orthogonal information to increase confidence in predictions. There can be two different kinds of novel TSSs as illustrated in Figure 1.23 – novel 5' exons derived from alternative 5' end splicing events, and extensions of annotated 5' exons. As already, discussed, *de novo* transcript assemblies can, for a number of reasons, be incomplete and thus miss the actual TSS; in the same time, in cases in which internal exons serve as alternative promoters, a separate transcript may not be assembled due to the simultaneous expression of the longer isoform or the subsequent merge of the transcript into a longer model. Extensions of annotated 5' exons are particularly difficult to assess, as the phenomenon of imprecise transcriptional initiation occurring over a neighborhood of nucleotides is well established; while very long extensions are more likely to represent real new promoters, the interpretations of shorter ones is difficult. Promoters can in principle be both extended and shortened;


A
H1-hESC
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
TAF1
no TAF1
>100
10-100
1-10
<1
TSS Expression (FPKM)
B
H1-hESC
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
FPKM: >100 >100 10-100 10-100
TAF1: + - + -
Perfectly mappable
Not perfectly mappable
C
GM12878
HeLa
HepG2
H1-hESC
K562
TAF1
CpG
0
1
D
>1000bp upstream
100-1000bp upstream
+/- 100bp of TSS
100-1000bp downstream
>1000bp downstream
2
4
8
16
32
64
128
256
512
1024
RPM
8535
7811
12585
6907
7864
E
others
RNA-seq evidence
0.0
0.2
0.4
0.6
0.8
1.0
Fraction
F
GM12878
H1-hESC
HeLa
HepG2
K562
1
2
3
4
5
6
7
8
9
0
1

however, the latter is particularly challenging for assembly as RNA-seq library building is typically performed using random hexamer priming, which inherently results in lower coverage of the very end of transcripts.

For these reasons, I aimed at utilizing orthogonal evidence to assess the assembly of 5' transcript ends in our data. In addition to the TAF1 and RNA Polymerase II ChIP-seq data we generated, I also took advantage of genome-wide Capped Analysis of Gene Expression (CAGE) (Kodzius et al. 2006; Carninci et al. 2006) generated as part of the ENCODE consortium (ENCODE Project Consortium 2011) (See Methods for details on the use of CAGE data). I first examined the relation of TAF1 binding, RNA Polymerase II loading and the presence of CAGE clusters to the expression of the TSSs of GENCODE V7 protein coding genes (Figure 1.24A). As discussed above, not all highly expressed TSSs have associated TAF1 binding (Figure 1.24D), however, the sensitivity of CAGE clusters was much higher, with more than 90% of TSSs expressed at more than 10 FPKM being CAGE-positive (Figure 1.24E).

The set of merged transcript models contains 9,787 instances of novel 5' exons and 5,690 extensions of annotated 5' exons, in addition to the intergenic candidate lincRNAs and TUCP. Since 5' exon extensions are difficult to interpret we initially focused our attention on novel 5' exons. The expressions patterns of these 5' exons (where the expression of the exons is defined as the sum of the FPKMs of all transcripts containing it) (Figure 1.24K) were similar to those of novel isoforms of protein coding genes (Figure 1.18D). A lower fraction of these exons was supported by orthogonal TAF1 and CAGE evidence compared to annotated TSSs at similar expressions levels (Figure 1.24B,F and G). Strikingly, almost none of the intergenic spliced transcripts (lincRNA and TUCP) had TAF1 binding to its 5' end and a smaller fraction were positive CAGE clusters (Figure 1.24C,H and I). This indicates that *de novo* assembly of intergenic spliced transcripts may not be as complete as desired and/or some of them may utilize different mechanisms of their transcription initiation. The resolution of TAF1 ChIP-seq data is not high enough to be useful for assessing 5' exon extensions, but this can be done by asking for precise base pair matching of aligned CAGE read. A strikingly high proportion of 5' exon extensions, including the relatively few examples of 5' exon shortening, had orthogonal support in such manner. I use the *BHLHE40* transcription factor as a representative example of a gene with well supported novel TSSs in Figure 1.25.

Figure 1.22 *(preceding page)*: Complexity of genome-wide TAF1 binding patterns. (A) TAF1 binds to most but not all expressed transcription start sites (TSSs). (B) Absence of TAF1 is due in some but not the majority of cases to poor read mappability around the TSS. (C) TSSs without TAF1 binding sites tend to lack TAF1 binding in all cell lines and to also lack CpG islands in their vicinity. Shown are all TSSs expressed at more than 1 FPKM$_{conf\_lo}$ in all 5 cell lines examined; according to the presence or absence of TAF1 binding or CpG island, a score of 1 (blue) or 0 (light yellow) was assigned to it, and the resulting matrix was clustered hierarchically. (D,E) Distribution of TAF1 binding sites relative to the GENCODE V7 annotation. The total number of sites is indicated to the left of the plot in (E). (D) Binding sites found away from annotated TSSs tend to be weaker. The maximum RPM for a TAF1 binding sites across all datasets is plotted. (E) Orthogonal RNA-seq evidence from Cufflinks and Cuffmerge-derived transcript models for TAF1 binding sites not associated with annotated TSSs. For binding sites more than 1 kb away from a TSS, a transcript model TSS within 1 kb of the TAF1 binding site was required. For binding sites between 100 bp and 1 kb away from a TSS, a transcript model TSS within 100 bp of the TAF1 binding site was required. (F). TAF1 bindings sites not associated with GENCODE V7 TSS are mostly seen in one cell line. According to the presence or absence of a TAF1 binding site in a cell line, a score of 1 (red) or 0 (light yellow) was assigned to it, and the resulting matrix was clustered hierarchically for each of 9 groups of TAF1 binding sites. 1) TAF1 sites within 100 bp of a TSS, 2) TAF1 sites > 1 kb upstream of a TSS with RNA-seq evidence, 3) TAF1 sites > 1 kb downstream of a TSS with RNA-seq evidence, 4) TAF1 sites 100 bp to 1 kb upstream of a TSS with RNA-seq evidence, 5) TAF1 sites 100 bp to 1 kb downstream of a TSS with RNA-seq evidence, 6) other TAF1 sites > 1 kb upstream of a TSS, 7) other TAF1 sites > 1 kb downstream of a TSS, 8) other TAF1 sites 100 bp to 1 kb upstream, 9) other TAF1 sites 100 bp to 1 kb downstream of a TSS.

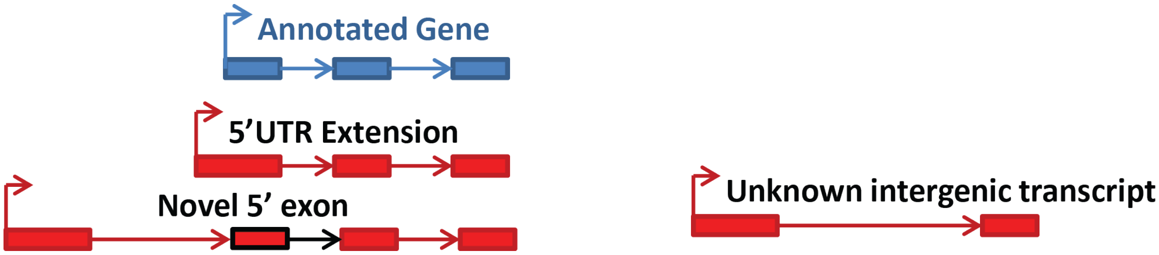


Figure 1.23: Different types of novel 5' transcirpt ends.

1.2.13 Alternative promoter usage

Initiation of transcription from alternative promoters is a well-established mechanism for generation of transcript diversity with a number of examples known (Landry et al. 2003; Wu et al. 1999; Tasic et al. 2002). To estimate how prevalent overall this phenomenon is in the human genome we examined the number of alternative TSSs utilized by each gene as a function of their expression levels. Of 9,939 genes with individual GENCODE v7 TSSs expressed at more than a conservative threshold of 5 FPKM$_{conf\_lo}$, 5,553 (~56%) expressed only a single TSS passing that threshold, 2,398 (24%) expressed two TSSs, and 1,988 (20%) expressed more than two TSSs. (Figure 1.24L). In addition, 1,494 genes had novel 5' exons expressed at more than 5 FPKM$_{conf\_lo}$ (Figure 1.18P), and for both annotated TSSs and novel 5' exons, relaxing this threshold results in the detection of a larger number of alternative promoter usage events (Figure 1.24M).

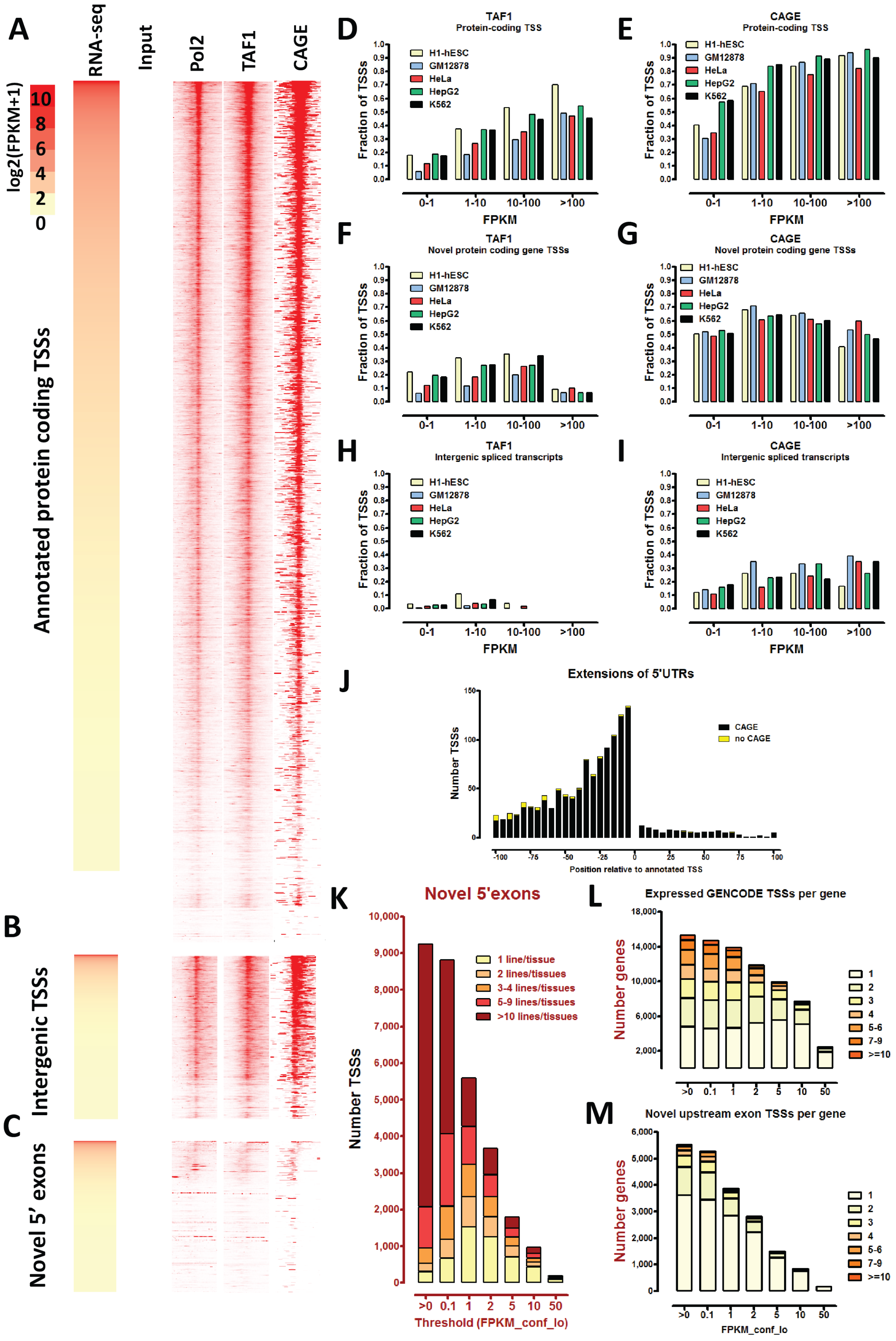
1.3 Discussion

A primary analysis of the human polyadenylated transcriptome was presented. The results reveal both the information richness of datasets generated with RNA-seq technology and the complexity of transcription in human cells. In the same time, they also highlight a number of challenges to data interpretation presented by the very same transcriptome complexity and the imperfections of current experimental and analytical tools. Below, I discuss the impact of this kind of RNA-seq measurements on the current status of our knowledge about the transcriptome, the major remaining areas of uncertainty and the expected further advances that will be needed to resolve them.

1.3.1 The growing complexity of the human transcriptome

As shown here and by others (Djebali et al. 2012), contemporary RNA-seq measurement have the potential to greatly increase both the number of isoforms of known genes and the number of transcripts belonging to various classes of intergenic, anti-sense and other more or less exotic types of transcription events (Gingeras 2009). However, this same sensitivity also presents a great challenge in distinguishing the products of transcriptional noise from functional transcripts. This is a problem to which in my opinion a satisfactory solution has not yet been found and I do not claim to have solved it here either[1]. Reasoning that erring towards a more conservative set of transcripts is more desirable for the purpose of generating interesting hypothesis with direct biological relevance for further investigation, a number of filters designed to remove as much of noise products and computational artifacts as possible were applied. Thus, the final set of transcripts expands on the GENCODE v7 annotation with less than 40,000 novel isoforms of protein coding genes, ~3,500 candidate lncRNAs and ~2,500 TUCPs. The increase in the number of splice junctions was proportionally significantly larger and even though the majority of them are poorly supported, large numbers of well-supported novel splice junctions were left out of the final set of transcript models at various steps in the computational analysis pipeline. For each set of novel or annotated elements of the transcriptome (splice junctions, exons, known isoforms of protein coding genes, novel isoforms of protein coding genes, intergenic non-coding RNAs) the same pattern

[1] This is just as true in 2014 as it was when these words were originally written in 2011



A
log2(FPKM+1)
10
8
6
4
2
0
RNA-seq
Input
Pol2
TAF1
CAGE
Annotated protein coding TSSs
B
Intergenic TSSs
C
Novel 5' exons
D
TAF1
Protein-coding TSS
E
CAGE
Protein-coding TSS
F
TAF1
Novel protein coding gene TSSs
G
CAGE
Novel protein coding gene TSSs
H
TAF1
Intergenic spliced transcripts
I
CAGE
Intergenic spliced transcripts
H1-hESC
GM12878
HeLa
HepG2
K562
Fraction of TSSs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0-1
1-10
10-100
>100
FPKM
J
Extensions of 5'UTRs
CAGE
no CAGE
Number TSSs
0
50
100
150
-100
-75
-50
-25
0
25
50
75
100
Position relative to annotated TSS
K
Novel 5'exons
1 line/tissue
2 lines/tissues
3-4 lines/tissues
5-9 lines/tissues
>10 lines/tissues
Number TSSs
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
>0
0.1
1
2
5
10
50
Threshold (FPKM_conf_lo)
L
Expressed GENCODE TSSs per gene
Number genes
2,000
6,000
10,000
14,000
18,000
>0
0.1
1
2
5
10
50
1
2
3
4
5-6
7-9
>=10
M
Novel upstream exon TSSs per gene
Number genes
0
1,000
2,000
3,000
4,000
5,000
6,000
>0
0.1
1
2
5
10
50
FPKM_conf_lo
1
2
3
4
5-6
7-9
>=10

is observed - very large numbers of poorly supported/abundant and a small number of highly abundant and well supported elements, with a continuum between them. Which elements are included and which are not is currently determined by setting thresholds that are somewhat biologically informed but still arbitrary. Finding the right balance in the necessary trade-off between sensitivity and specificity is an open challenge for the field; however, finding such a balance may be in principle impossible since functional transcripts can be expressed at relatively low levels while the noise products from highly expressed loci are expected to be also relatively highly abundant. For example, the important regulator of neuronal fate NRSF is usually observed to be expressed in single-digit FPKMs, and few lncRNAs are detected at high levels in each individual cell line even though large numbers of them were found to be of functional importance when knocked down in mouse embryonic stem cells by a recent study (Guttman et al. 2011) (although our data is for human cells, it is reasonable to expect that the general patterns of lncRNA expression are not drastically different between the two species).

The answers to several open questions in the field as well as the interpretation of observations for individual loci by researchers looking to more deeply investigate their gene of interest are highly dependent on the approach towards this problem. Both the extent of transcriptional activity in the human genome and the prevalence of functional alternative splicing events have been widely debated (Kapranov et al. 2002; Kapranov et al. 2007; Sorek et al. 2004; Wang et al. 2008; Dinger et al. 2009; van Bakel et al. 2010; Clark et al. 2011; Mercer et al. 2011); how abundance levels relate to distinguishing noise products from functional transcripts is at the heart of this debate.

Nevertheless, the number of annotated transcripts in the human genome is expected to grow considerably as more and more information derived from RNA-seq measurements is incorporated into annotations. This is a reasonable expectation given that we have surveyed a wide and diverse collection of cell lines and tissues and the discovery of most novel elements did not reach saturation (Figures 1.7B, 1.14N, 1.18B,E and F).

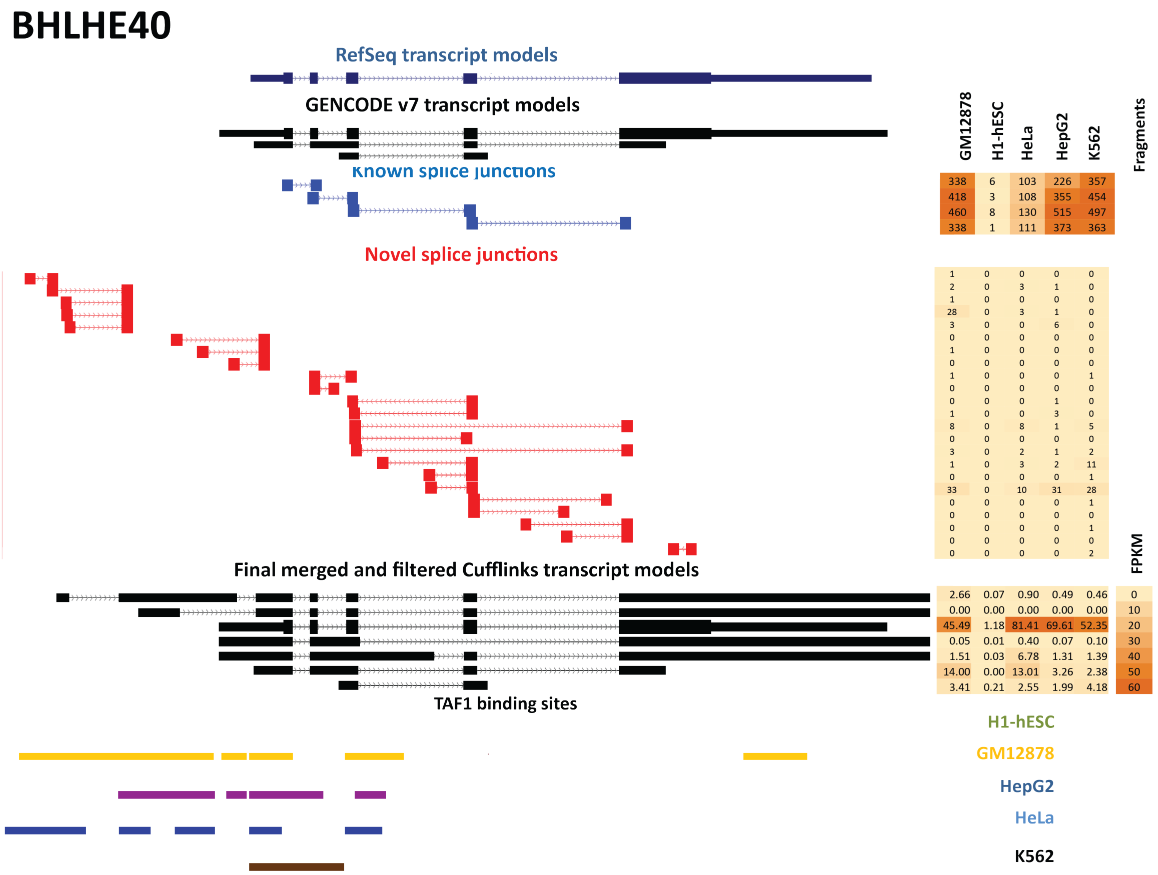
1.3.2 Reliability of transcript-level quantification

This growth in complexity, however, has the potential to even further complicate data analysis and results interpretation.

The accurate quantification of individual transcripts of a gene is of critical importance for the analysis of the prevalence and tissue-specificity of alternative splicing and alternative transcription initiation and termination events. However, accurately and confidently assigning the still short reads generated in RNA-seq experiment to transcripts in a complex locus is still not a trivial computational tas ; while current tools employ highly sophisticated algorithms for deconvolving the expression levels of individual isoforms, this becomes essentially impossible when locus complexity grows beyond a certain threshold as the statistical models employed often become unidentifiable. Yet, as more and more new transcripts are uncovered by the sequencing of wider panels of cell lines and tissues, the complexity of annotations is expected to grow further and further and make this an ever more intractable problem.

The current output of these program suggest the existence of a number of potentially interesting biological phenomena in the data, including the widespread occurrence of major isoform

Figure 1.24 ***(preceding page)*****: Identification and orthogonal support for novel 5' transcript ends.** (A-C) TAF1, RNA Polymerase II and CAGE cluster profiles around the TSS of GENCODE V7 protein coding genes (A), candidate novel 5' exon TSS of protein coding genes (B) and candidate lncRNAs and TUCPs (C). TSSs are sorted by decreasing expression level. (D), (F), (H) TAF1 coverage of expressed GENCODE V7 protein coding gene TSSs (C), candidate novel 5' exon TSSs (F) and candidate lncRNAs and TUCPs (H) in 5 ENCODE cell lines. (E), (G), (I) CAGE cluster coverage of expressed GENCODE V7 protein coding gene TSSs (E), candidate novel 5' exon TSSs (G) and candidate lncRNAs and TUCPs (I) in 5 ENCODE cell lines. (J). Support by CAGE reads for extended and shortened 5' exons. (K) Abundance levels and cell type specificity of novel TSSs. (L-M) Number of expressed annotated (L) and novel (M) TSS per gene as a function of expression levels.



GM12878	H1-hESC	HeLa	HepG2	K562
338	6	103	226	357
418	3	108	355	454
460	8	130	515	497
338	1	111	373	363

GM12878	H1-hESC	HeLa	HepG2	K562
1	0	0	0	0
2	0	3	1	0
1	0	0	0	0
28	0	3	1	0
3	0	0	6	0
0	0	0	0	0
1	0	0	0	0
0	0	0	0	0
1	0	0	0	1
0	0	0	0	0
0	0	0	1	0
1	0	0	3	0
8	0	8	1	5
0	0	0	0	0
3	0	2	1	2
1	0	3	2	11
0	0	0	0	1
33	0	10	31	28
0	0	0	0	1
0	0	0	0	0
0	0	0	0	1
0	0	0	0	0
0	0	0	0	2

GM12878	H1-hESC	HeLa	HepG2	K562	FPKM
2.66	0.07	0.90	0.49	0.46	0
0.00	0.00	0.00	0.00	0.00	10
45.49	1.18	81.41	69.61	52.35	20
0.05	0.01	0.40	0.07	0.10	30
1.51	0.03	6.78	1.31	1.39	40
14.00	0.00	13.01	3.26	2.38	50
3.41	0.21	2.55	1.99	4.18	60

Figure 1.25: Isoform-level quantification, fragment support for known and novel junction, and TAF1 binding sites for the BHLHE40 locus..

switch events with high tissue specificity (Figure 1.18K), the utilization of multiple alternative promoters, and the surprisingly high abundance of what appear to be NMD transcripts (Figure 1.18O), phenomena suggested to play significant role in the generation of proteome diversity and in gene regulation. However, their reality is to a large extent contingent on how accurately the underlying biological reality is reflected in this output. Thus, conclusive confirmation or refutation of these phenomena will have to await the arrival of data or computational tools that allow more confident deconvolution of transcript levels.

This is also relevant to downstream applications of RNA-seq quantification feeding into other areas of transcriptional biology. For example, complete understanding of the mechanisms of transcriptional regulation is not possible without complete understanding of the relationship between the interaction of sequence-specific transcription factors, general transcription factors, RNA polymerase and chromatin state at promoters, on one side, and transcript levels, on the other. Working with simpler annotations of the genome allows for mostly ignoring the issue; however, if alternative promoter use is indeed as ubiquitous as suggested by the data, the relative use of these TSSs will have to be very finely and accurately parsed and integrated with orthogonal ChIP-seq data for such understanding to be achieved.

Anecdotal evidence suggests that numerous suspicious quantification results can be found. For example, Figure 1.26 shows the case of the *FOSL2* gene, for which 5 isoforms are annotated in GENCODE, and additional 6 were presented in the merged assembly generated here. Requantification on the merged assembly suggested that two of the novel isoforms (originating from novel alternative promoters) are presented at FPKM levels comparable to those of the annotated isoforms; however, these isoforms are supported by just 1 spliced RNA-seq fragments spanning their unique splice junctions, while the corresponding unique splice junction of the annotated isoforms had coverage of 45 fragments, thus it is morel likely that the abundance of the novel isoforms is in fact significantly lower, even though these alternative promoters had orthogonal TAF1 occupancy support.

1.3.3 Transcript reconstruction and resolving transcript ends

Both alternative transcript initiation and alternative polyadenylation (Di Giammartino et al. 2011; Sandberg et al. 2008) have been suggested to play important role in gene expression regulation. Due to the nature of RNA-seq library-building protocols employing random hexamer priming, the extreme ends of transcripts are usually underrepresented in the final libraries, which, combined with the lower coverage naturally expected for lower-abundance transcripts, makes it difficult to precisely determine the exact beginning of a transcript or its polyadenylation site (reads containing portions of the polyA tail are also not expected to map to the genome). CAGE data provides information about capping events, and to the extent that capping events correspond to transcriptional initiation events (which is not always the case; Affymetrix ENCODE Transcriptome Project 2009), about promoters. In addition, several approaches have been devised to map polyadenylation sites (Oz-

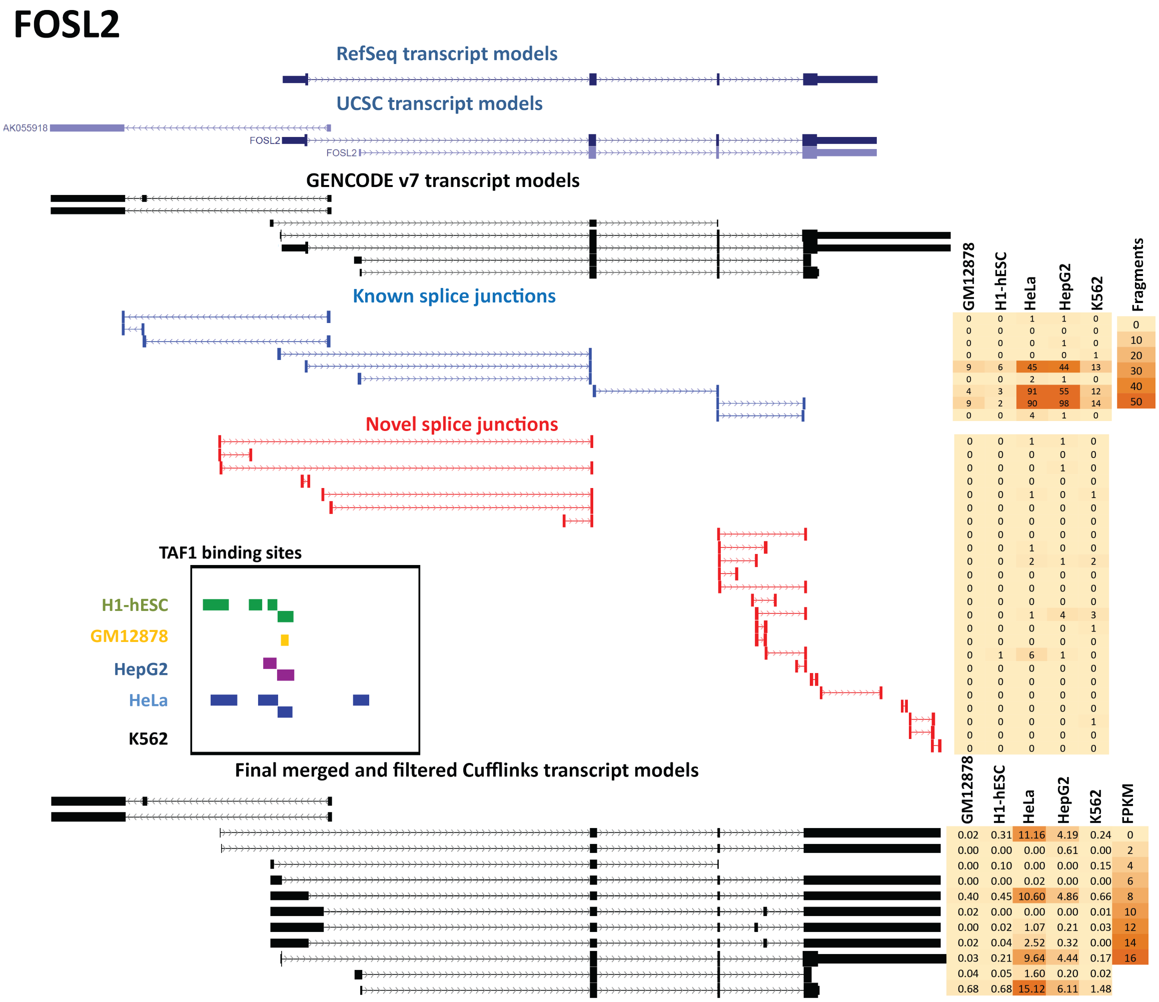


Figure 1.26: Isoform-level quantification, fragment support for known and novel junction, and TAF1 binding sites for the FOSL1 locus.

solak et al. 2010; Jan et al. 2011). However, building such libraries for large numbers of samples is a practical challenge, and their interpretation, as demonstrated by the discovery that CAGE tags do not always correspond to transcription initiation events and the fact that they still only provide information about the extremes of transcripts but not about the connectivity between, is not straightforward.

Here it has been possible to identify novel alternative 5' exons and to leverage CAGE data to confirm the extent of 5' extensions of known 5' exons. In addition, in many cases what seems to represent either 3'UTRs extending long past the annotated polyA site or unspliced transcripts originating in the 3'UTR vicinity was observed. Such cases are of great interest if shown to be continuous with the annotated transcript as they can change the set of miRNAs targeting it or play other, so far unappreciated, regulatory roles. However, we are at present unable to examine the nature of these transcriptional events as current short reads can be effectively used for transcript reconstruction when splice junctions are present but precisely defining transcripts for long stretches of continuously overlapping reads is challenging.

The same issue was confronted when analyzing intergenic transcripts. A very large number of monoexonic intergenic transcripts are observed (Figure 1.14K and L). A majority of these consist of single fragments mapping to intergenic space but large numbers of regions with high read coverage are also seen. Determining where these transcripts begin, and if they have biologically precisely defined ends, is of crucial importance for assessing their functional significance, and elucidating the mechanisms of regulation of their expression.

It was also observed that the 5' ends of intergenic spliced transcripts (candidate lncRNAs and TUCPs) as currently defined using reconstruction from RNA-seq are poorly supported by TAF1 binding and CAGE tags (Figure 1.24H, K, M and N). This suggests that due to the generally low expression levels of these transcripts, they have not been fully reconstructed and either large stretches of their first exons or whole first exons are missing. Alternatively, TAF1 loading and message capping may not play a role in the transcription initiation and biology of these transcripts. Either way, establishing that one of these options is the case by completing their transcript models is of great importance for understanding the biology of these RNAs.

1.3.4 Absolute numbers of transcripts per cell

FPKM values reflect the proportional abundance of transcripts in a sequencing library normalized for transcript length. Ideally, however, the actual numbers of copies of a transcript per individual cell should be obtained. This information is important both for evaluating the functional significance of transcriptional events (i.e. if a transcript is found at what amounts to one copy per ten cells, then it is more likely to be a product of transcriptional noise than if it is found at multiple copies per cell) and for deriving mechanistic insights into transcript functions. For example, in addition to other biological roles, both textitcis- and *trans*- action mechanisms have been suggested for how lncRNAs may participate in the regulation of transcription in the nucleus (Koziol & Rinn 2010). Naturally, this leads to the expectation that *cis*-acting transcripts that function at the genomic location which they are transcribed from, of which there are only two copies, should be present at very limited number of transcript copies per cell while *trans*-acting transcripts should be on average more abundant. For this issue to be resolved, measurements of the absolute transcript counts per cell are needed. At present, it is difficult to obtain that information from RNA-seq data, as RNA sequencing libraries are prepared from bulk RNA isolated from millions of cells. It is possible to calculate rough estimates of these numbers (Mortazavi et al. 2008); however, this requires precise tracking of cell numbers and the amount of RNA going into libraries. This is something that's not easily tractable for tissues, and even when it is available for cell lines, it is only a rough guess with major uncertainties associated with it.

1.3.5 Looking towards the future

I expect this issue and a number of the other challenges outlined so far to be resolved with the further advancement of sequencing technology. Very long read lengths, ideally covering the full length of transcripts, will be needed in order to enable the precise demarcation of transcript structure and transcript ends, particularly around polyadenylation sites and for intergenic non-coding transcripts. For truly accurate transcript-level quantification, an additional requirement for large numbers of such reads exists, in order to fully cover the dynamic range of transcript expression levels in bulk RNA preps. Single-cell transcriptomics (Islam et al. 2011; Tang et al. 2009; Tang et al. 2010; Tang et al. 2011) combined with single-molecule sequencing and single-molecule RNA FISH measurements should allow the determination of absolute transcript numbers on the level of individual cells, and resolve several of the outstanding questions in the field.

1.4 Methods

All data processing and analysis for which no software packages are referenced was performed using custom-written Python scripts.

1.4.1 Cell growth and RNA harvesting

Cells were grown according to established ENCODE protocols (`http://genome.ucsc.edu/ENCODE/protocols/cell/`) and RNA prepared following the protocol described in Mortazavi et al. 2008.

1.4.2 RNA-seq data generation

Total RNA was subjected to two rounds of polyA selection and libraries built following the protocol described in Mortazavi et al. 2008. Libraries were sequenced as 2x76bp reads on the Illumina Genome Analyzer. Human Body Map data was kindly provided by Dr. Gary Schroth and the Expression Applications group at Illumina.

1.4.3 Read mapping

The last base pair of each read was removed. The resulting 2x75bp reads were mapped using TopHat (Trapnell et al. 2009, version 1.0.14) in *de novo* mode against the hg19 verion of the human genome. The same procedure was applied to polyadenylated RNA-seq data from 16 tissues generated using Illumina HiSeq 2000 as part of the Human Body Map 2 project. The *de novo* discovered splice junctions from all cell lines and tissues were combined with the set of splice junctions in the GENCODE v4 annotation to derive an extended set of junctions. Reads were mapped again using TopHat (version 1.0.14) against the male or female version of the hg19 version of the human genome with the extended set of junctions supplied while keeping the de novo junction discovery option turned on. All subsequent analysis was done on the resulting alignments. Read mapping statistics are provided in Table 1.1.

1.4.4 Transcript models discovery, merging and quantification

Cufflinks (Trapnell et al. 2010; Trapnell et al. 2012; version 1.0.1) was used to assemble transcripts in *de novo* mode from the TopHat alignments. Each sample was processed individually. The assemblies from all the samples were merged together with Cuffmerge (version 1.1.0) into a large transcript super-set using GENCODE v7 as a reference annotation. Assembly was done in fully *de novo* mode rather than in reference annotation based transcript (RABT; Roberts et al. 2011) because RABT assemblies contain a large number of clearly artifactual transcripts (especially when a complex annotation with a large number of isoforms is used such GENCODE). Such false positives are often unique to each sample and when merged result in a very large number of isoforms per gene most of which do not correspond to real transcript molecules and which make accurate quantification impossible. I also found that merging transcripts using the unfiltered *de novo* Cufflinks assemblies also resulted in an unacceptably high number of likely artifactual transcript models (although significantly fewer than with RABT assemblies), particularly transcripts with extremely large retained introns. Therefore I aimed to minimize the number of artifacts in the final assemblies by applying multiple filters before and after the merge step.

As an initial step, I classified new transcripts according to their relation to the annotation using Cuffcompare. Only transcripts classified as unknown intergenic and novel isoforms of known genes (Cuffcompare class codes "j" and "u") were retained. In addition, I required that novel isoforms of known genes had FPKM$_{conf\_lo} \geq 1$. The resulting set of transcript models for each cell line was used to run Cuffmerge.

The Cuffmerge output was filtered as follows. First, all retained introns relative to the Cuffmerge output itself were filtered out, i.e. if an exon had the same start and end positions as the left exon and the right exon respectively in any pair of exons in the annotation, the transcript containing it was removed from annotation. Next, all GENCODE v7 transcripts that were not present in the merge were added to the assembly according to the following criteria: for multiexonic transcripts, if the exact chain of splice junctions of a GENCODE v7 transcript was not present in the merged assembly, the transcript was added to it; there is no good criteria to define presence of absence for monoexonic transcripts so those were considered present if there was a monoexonic transcript overlapping them. After that step retained introns were filtered out again, this time against the GENCODE v7 annotation. Finally, because multiple occasions of extremely long 3' UTRs being assembled (usually due to the presence of overlapping transcript models in multiple cell lines) were observed, which would artificially drive down FPKM estimates by increasing the length of transcripts, all 3'UTRs were trimmed down to a maximum length of 5kb.

1.4.5 Genome and transcript models, annotations, and classification

Two transcript and gene model annotation sets for the human genome were used - version 7 of the GENCODE annotation (Harrow et al. 2006; Harrow et al. 2012), downloaded from `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV7/beta/` and the refSeq annotation, downloaded from `http://genome.ucsc.edu/`. Transcripts and genes were classified into protein coding and various non-coding classes according to the biotype classification in GENCODE V7, and the same classification was used where necessary for refSeq genes. CpG island annotations were downloaded from `http://genome.ucsc.edu/` and TSSs were classified as CpG or non-CpG according to whether a CpG island was present within 1kb of the TSS. For novel transcript models, ORFs were annotated using the longest ORF found in the transcript; transcripts were classified as putative NMD substrates if the ORF ended more than 50bp before the position of the last splice junction.

1.4.6 Non-coding RNA annotation and classification

Novel non-coding RNA were classified following an approach similar to the computational pipeline for lncRNA annotation described in Cabili et al 2011. I only considered spliced intergenic unannotated transcripts as classified by Cuffcompare. For each transcript, the codon substitution frequency (CSF) score was calculated using PhyloCSF (Lin et al. 2011) and the 45 vertebrate multiple genome alignment for the hg19 version of human genome, downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/maf/`. PhyloCSF was also run on annotated lncRNA and protein coding transcript from GENCODE V7 to establish thresholds for determining whether a transcript is likely to be coding or not (Figure 1.27). In addition to that, each spliced intergenic transcript was translated in all reading frames in both orientations, and scanned for the presence of protein domain annotated in the PFAM database (Punta et al. 2012; `http://pfam.sanger.ac.uk/search`). Transcripts with positive CSF scores and transcripts containing PFAM domains were classified as TUCPs.

1.4.7 Tissue specificity score calculation

The JS tissue specificity score was calculated as follows (Cabili et al. 2011), with the modification that for splice junctions, due to the highly quantized nature of the fragment counts at the low end and the difficulty to properly normalize fragment counts for sequencing depth in such cases, a cap of 10 distinct fragment was applied to all numbers before calculating the JS score:

The Jensen-Shannon divergence of two discrete probability distributions D_1 and D_2 is defined as:

$$JSD(D_1, D_2) = H\left(\frac{D_1 + D_2}{2}\right) + \frac{H(D_1) + H(D_2)}{2} \quad (1.1)$$

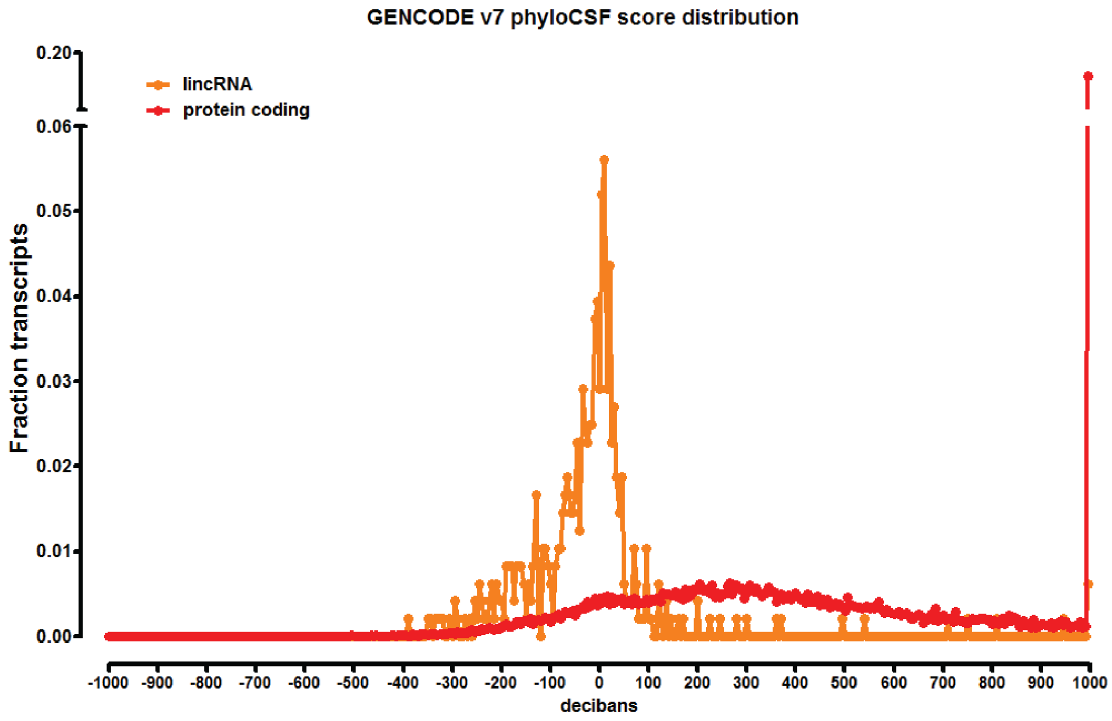


Figure 1.27: PhyloCSF score distribution for annotated in GENCODE V7 protein coding and lincRNA transcripts.

where $H(P)$ is the Shannon entropy for a discrete distribution P defined as:

$$H(P) = -\sum_{i=1}^{n} p_i * log(p_i) \tag{1.2}$$

The JS distance JS_{dist} is then defined as follows:

$$JS_{dist}(D_1, D_2) = \sqrt{JS(D_1, D_2)} \tag{1.3}$$

For a vector with expression values $E = \{e_1, e_2, ..., e_n\}$, a JS specificity score is then defined with respect to sample/tissue i as follows:

$$JS_{sp}(E|i) = 1 - JS_{dist}(E, E^i) \tag{1.4}$$

where E^i is the vector with maximum expression specificity, i.e. a positive FPKM value in sample/tissue i and FPKM $= 0$ everywhere else:

$$E^i := \{\delta_{i1}e_1, \delta_{i2}e_2, ..., \delta_{in}e_n\} \tag{1.5}$$

where δ_{ij} is the Kronecker delta function.

Finally, the JS specificity score JS_{sp} is the maximal specificity score across all samples/tissues, i.e.:

$$JS_{sp}(E) = \underset{i=\{1,..,|E|\}}{\text{argmax}} \ (JS_{sp}(E|i)) \tag{1.6}$$

1.4.8 Nanostring miRNA expression measurements

Measurements of miRNA expression using the miRNA Nanostring assay were performed on biological replicates following the manufacturer's instructions. Briefly, total RNA was extracted with the mirVana miRNA isolation kit and the remaining genomic DNA was removed by TURBO DNA-free kit (both kits are from Ambion, Life Technologies, NY). 100ng of total RNA, together with "spike-in" positive and negative control miRNAs, was annealed and ligated to the miRNAtags. After the unused miRNAtags were cleaned up, the chimeric miRNA:miRNAtag molecules were hybridized to the reporter codeset and capture probeset overnight. The hybridization mixture was purified on the nCounter Prep Station and the target molecules were immobilized and aligned on the nCounter cartridge. The nCounter cartridge was then scanned on the nCounter Digital Analyzer at maximum resolution. The collected data was further processed with nSolver analysis software to calculate the normalized miRNA expression level of each sample.

1.4.9 ChIP-seq data alignment and processing

ChIP-seq experiments were performed as described previously (Johnson et al. 2007), with the modification that a single round of PCR amplification was used instead of the majority of datasets (HeLa TAF1 being the only exception). The following antibodies were used: mouse monoclonal against TAF1 from Santa Cruz (sc-735), mouse monoclonal against RNA Polymerase II, clone 4H8 from Abcam (ab5408), mouse monoclonal against RNA Polymerase II, clone 8WG16 from (MMS-126R). Libraries were sequenced on the Illumina Genome Analyzer and reads of 36 bp size were generated. Each replicate contained at least 12 million uniquely aligned reads. Precise read mapping statistics are provided in Table 1.2.

Reads were aligned according to ENCODE standards against the male or female version of human genome (with random chromosomes and haplotypes excluded) depending on the sex of the cell line (male for H1-hESC and HepG2, female for HeLa, GM12878 and K562) using Bowtie (Langmead et al. 2009), version 0.12.7, with the following options: `-v 2 -t --best --strata`. TAF1 peak calling was done against appropriate input datasets using ERANGE 4.0 (Johnson et al. 2007), with the following settings: `‘‘--minimum 2 --ratio 3 --shift learn --revbackground --listPeak`. TAF1 peaks were merged according to the following procedure: if two peak summits were closer than 200 bp to each other, they were merged, with the new summit becoming the summit from the dataset whose reads per million (RPM) for the whole peak region were higher; this procedure was iterated across all datasets.

CAGE data processing

Tracks containing CAGE clusters and BAM files with individual read alignments were downloaded from the ENCODE portal at the UCSC Genome Browser (`http://genome.ucsc.edu/ENCODE/`. CAGE reads from all subcellular fractions were considered. In order for a TSS to be considered covered by a CAGE cluster, a CAGE cluster on the same strand as the direction of transcription was required. For the analysis of 5’ extensions, precise matching of 5’ ends of at least one CAGE read on the same strand as the direction of transcription was required.

Table 1.1: Read mapping statistics for the RNA-seq datasets

Cell Line	Read Length	Description	Insert Size	Rep	Unique	Unique Splices	Multi	Multi Splices
H1-hESC	2x75	ES cells	∼200	Rep4	45,317,222	8,618,025	1,933,958	92,172
H1-hESC	2x75	ES cells	∼200	Rep1	63,216,759	12,781,989	1,799,576	94,991
H1-hESC	2x75	ES cells	∼200	Rep2	64,849,492	12,937,223	1,996,326	106,373
H1-hESC	2x75	ES cells	∼200	Rep3	62,721,190	12,417,189	2,061,864	107,319
GM12878	2x75	lymphoblastoid	∼200	Rep2	152,774,148	23,930,558	6,259,715	420,288
GM12878	2x75	lymphoblastoid	∼200	Rep1	91,217,874	15,146,110	3,502,071	202,872
K562	2x75	myelogenous leukemia	∼200	Rep1	133,776,448	27,150,397	4,809,472	349,097
K562	2x75	myelogenous leukemia	∼200	Rep2	121,520,650	22,256,714	4,333,136	261,089
HSMM	2x75	myoblasts	∼200	Rep1	97,833,543	23,352,403	1,997,844	199,498
HSMM	2x75	myoblasts	∼200	Rep2	98,203,018	23,229,216	2,234,108	199,772
HUVEC	2x75	umbilical vein endothelial	∼200	Rep1	74,294,272	17,207,804	2,053,278	167,529
HUVEC	2x75	umbilical vein endothelial	∼200	Rep2	54,420,816	12,607,003	1,699,903	113,729
HeLa	2x75	HeLa	∼200	Rep1	49,453,158	9,301,487	2,060,411	125,644
HeLa	2x75	HeLa	∼200	Rep2	75,223,386	14,527,666	2,603,614	170,225
HepG2	2x75	liver carcinoma	∼200	Rep1	80,554,751	17,831,315	2,762,367	206,825
HepG2	2x75	liver carcinoma	∼200	Rep2	94,588,954	21,423,792	3,300,392	324,730
MCF7	2x75	breast cancer	∼200	Rep1	109,216,869	16,770,366	3,191,573	143,875
MCF7	2x75	breast cancer	∼200	Rep2	87,203,914	21,226,373	2,032,830	251,913
NHEK	2x75	keratinocytes	∼200	Rep1	79,396,401	12,110,678	2,642,317	292,371
NHEK	2x75	keratinocytes	∼200	Rep2	89,043,589	21,805,036	1,967,691	254,190
NHLF	2x75	lung fibroblasts	∼200	Rep1	87,308,499	20,557,003	1,786,840	149,586
NHLF	2x75	lung fibroblasts	∼200	Rep2	81,888,840	19,744,610	1,344,427	150,557
adipose	2x50+1x75		∼200	Rep1	184,034,305	18,186,787	9,474,379	317,902
adrenal	2x50+1x75		v200	Rep1	182,891,875	15,312,732	8,797,765	376,596
brain	2x50+1x75		∼200	Rep1	174,392,333	14,623,420	7,236,006	196,026
breast	2x50+1x75		∼200	Rep1	183,725,194	16,979,055	8,734,757	346,215
colon	2x50+1x75		∼200	Rep1	201,909,819	17,009,282	11,690,564	297,609
heart	2x50+1x75		∼200	Rep1	197,439,159	18,189,625	14,170,195	416,843
kidney	2x50+1x75		∼200	Rep1	192,378,197	15,596,359	11,063,331	281,585
liver	2x50+1x75		∼200	Rep1	187,757,362	25,697,250	10,039,834	1,193,880
lung	2x50+1x75		∼200	Rep1	194,249,068	19,991,574	9,938,722	702,441
lymph node	2x50+1x75		∼200	Rep1	193,396,478	18,473,375	12,780,186	1,271,624
ovary	2x50+1x75		∼200	Rep1	198,207,292	20,096,511	10,231,268	317,030
prostate	2x50+1x75		∼200	Rep1	205,065,901	21,109,090	10,372,722	302,181
muscle	2x50+1x75		∼200	Rep1	197,504,306	23,329,011	9,776,756	340,782
testes	2x50+1x75		∼200	Rep1	197,739,813	23,635,613	8,468,114	349,225
thyroid	2x50+1x75		∼200	Rep1	194,749,061	23,851,093	7,835,229	319,566
WBC	2x50+1x75		∼200	Rep1	199,299,458	24,007,769	10,147,780	366,024

Table 1.2: Read mapping statistics and library characteristics for ChIP-seq datasets

Cell Line	Factor/Antibody	Replicate	Uniquely aligned reads	Library Complexity
GM12878	Pol2-CTD-4H8	Rep1	28,110,098	0.78
GM12878	Pol2-CTD-4H8	Rep2	24,404,299	0.88
GM12878	Pol2-CTD-8WG16	Rep1	27,121,649	0.82
GM12878	Pol2-CTD-8WG16	Rep2	27,783,933	0.82
GM12878	TAF1	Rep1	18,374,847	0.55
GM12878	TAF1	Rep2	22,148,439	0.59
H1-hESC	Pol2-CTD-4H8	Rep1	17,359,575	0.89
H1-hESC	Pol2-CTD-4H8	Rep2	19,062,392	0.77
H1-hESC	Pol2-CTD-8WG16	Rep1	20,587,873	0.76
H1-hESC	Pol2-CTD-8WG16	Rep2	18,325,024	0.81
H1-hESC	TAF1	Rep1	14,023,010	0.87
H1-hESC	TAF1	Rep2	13,217,524	0.85
HeLa	Pol2-CTD-8WG16	Rep1	21,848,831	0.87
HeLa	Pol2-CTD-8WG16	Rep2	25,528,202	0.83
HeLa	TAF1	Rep1	28,472,126	0.53
HeLa	TAF1	Rep2	11,429,207	0.9
HepG2	Pol2-CTD-4H8	Rep1	18,242,505	0.91
HepG2	Pol2-CTD-4H8	Rep2	33,930,680	0.88
HepG2	Pol2-CTD-8WG16	Rep1	14,722,736	0.71
HepG2	Pol2-CTD-8WG16	Rep2	22,030,475	0.83
HepG2	TAF1	Rep1	18,580,720	0.84
HepG2	TAF1	Rep2	16,568,099	0.79
K562	Pol2-CTD-4H8	Rep1	9,798,768	0.85
K562	Pol2-CTD-4H8	Rep2	23,095,649	0.73
K562	Pol2-CTD-8WG16	Rep1	29,190,954	0.84
K562	Pol2-CTD-8WG16	Rep2	26,469,081	0.78
K562	TAF1	Rep1	17,018,556	0.89
K562	TAF1	Rep2	19,987,210	0.78

2

Simulation-based characterization of transcript assembly and quantification from short-read RNA-seq data

This chapter contains the results of a simulation aimed at understanding the performance of software performing transcript-level quantification and/or assembly. It was carried out after the work presented in the previous chapter was completed and as a result did not inform it; however, it does shed light on the interpretation of the results from it, which I discuss here.

Abstract

The reliability of the analysis of transcriptome diversity using short-read RNA-seq data is inherently limited by the performance of the software used to carry it out. Anecdotal evidence has presented numerous examples of computational artifacts significantly affecting biological conclusions. To clarify some of these issues, a simulation study of some of the most often used RNA-seq quantification and transcript reconstruction tools was carried out. Its results place minimum bounds on the fraction of false positives and false negatives in the real-data analysis presented in the previous chapter. I also examine the effect of several characteristics of RNA-seq datasets that are suspected to influence quantification and/or assembly but simulations published in the past have so far not modeled.

2.1 Introduction

The currently existing high-throughput sequencing technologies that are capable of delivering the needed for RNA-seq sequencing depth all produce short reads, much shorter than the length of mRNA molecules. Read lengths have increased significantly with the development of the technology, from 25bp around 2007 to up to 2x250bp and even longer now. However, the longer reads are not necessarily optimal for RNA-seq applications (unless they cover full-length mRNAs, which at present they do not), for reasons outlined in the Methods section of this chapter, thus the analysis of RNA-seq data faces the following common challenges:

1. Aligning of short reads to the genome, in a splice-aware manner that allows the discovery of previously unannotated splice junctions that are present in the data

2. The quantification of gene expression levels, at the gene and at the transcript level. The latter is important on its own as it would ideally provide reliable information on any differential regulation of splicing, transcriptional initiation or polyadenylation between samples, but it is also vital for the accurate quantification on the gene level (again, see discussion below in the Methods section).

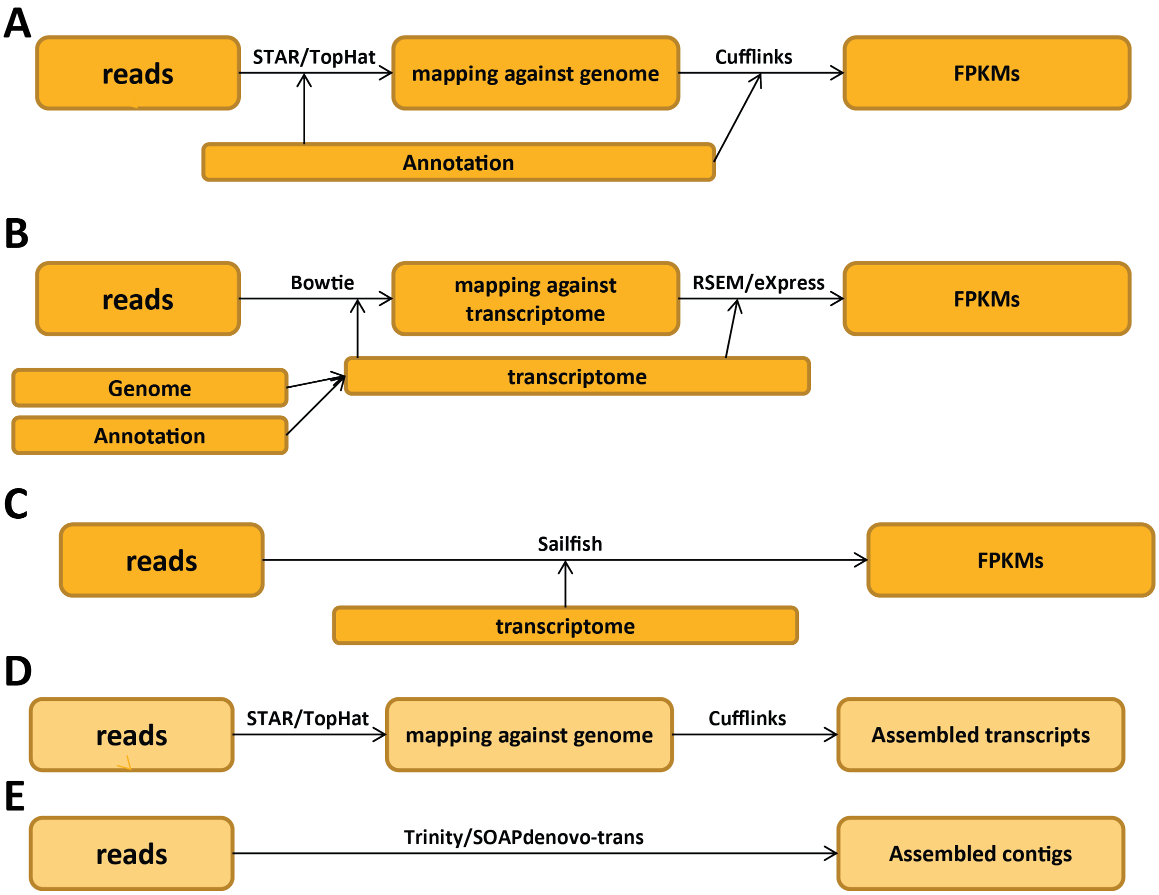


Figure 2.1: Strategies for carrying out isoform-level quantification and assembly for RNA-seq data. There are three approaches adopted in the literature for carrying out transcript-level quantification of RNA-seq data: alignment and quantification in genomic space (A), alignment and quantification in transcriptome space (B), and the alignment-free k-mer-based quantification approach adopted by Sailfish (C). See text for more details. Here, genomic alignment and quantification were carried out using TopHat or STAR and Cufflinks, transcriptome alignment and quantification using Bowtie and RSEM or eXpress. There are two main approaches for *de novo* transcript reconstruction: alignment-based reconstruction (D), and alignment-free *de novo* assembly from reads (E). Here, STAR and TopHat mappings plus Cufflinks assembly were used for the former, while Trinity and SOPAdenovo-trans were used for the latter

3. The *de novo* reconstruction of expressed transcripts from short reads. This is needed for the discovery of novel transcripts in sequenced and annotated genomes, for the annotation of newly sequenced genomes and often for the sequencing and analysis of the transcriptomes of species for which a genome assembly does not exist.

A wide variety of computational tools have been developed to carry out these tasks. Dozens of RNA-seq mappers, which carry out read mapping and *de novo* splice junction detection, have been published. These include TopHat (Trapnell et al. 2009; Trapnell et al. 2012), STAR (Dobin et al. 2013), RUM (Grant et al. 2013), SplitSeek (Ameur et al. 2011), SpliceMap (Au et al. 2010), Map-Next (Bao et al. 2009), Supersplat (Bryant et al. 2010), QPALMA (De Bona et al. 2008), HMMSplicer (Dimon et al. 2010), OSA (Hu et al. 2012), SOAPsplice (Huang et al. 2011), PALMapper (Jean et al. 2010), SeqMap (Jiang & Wong 2008), MapAl (Labaj et al. 2012), TrueSight (Li et al. 2013), Subread (Liao et al. 2013), GEM (Marco-Sola et al. 2012), PASTA (Tang & Riva 2013), MapSplice (Wang et al. 2010), X-MATE (Wood et al. 2011) , GSNAP (Wu & Nacu 2010), OLego (Wu et al. 2013), and others. The ENCODE Project used both TopHat and STAR. TopHat was used for most of the analyses presented in this thesis.

A similarly diverse set of transcript-level quantification algorithms is available, including Cufflinks (Trapnell et al. 2010; Trapnell et al. 2012; Trapnell et al. 2013; Roberts et al. 2011a; Roberts 2011b), eXpress (Roberts & Pachter 2013), RSEM (Li et al. 2010; Li et al. 2011), Sailfish (Patro et al. 2014), CEM/IsoLasso (Li et al. 2011; Li & Jiang 2012), Flux-Capacitor, IQSeq (Du et al. 2012), iReckon (Mezlini et al. 2013), IsoEM (Nicolae et al. 2011), MMSeq (Turro et al. 2011), PennSeq (Hu et al. 2014), RNAExpress (Forster et al. 2013), SLIDE (Li et al. 2011), and Traph (Jo et al. 2014), Oqtans (Sreedharan et al. 2014), rQuant (Bohnert & Rätsch 2010), RNASEQR (Chen et al. 2012), RDiff (Drewe et al. 2013), Montebello (Hiller & Wong 2013), IsoformEx (Kim et al. 2011), NEUMA (Lee et al. 2011), EBSeq (Leng et al. 2013), SASeq (Nguyen et al. 2013), NSMAP (Xia et al. 2011), MITIE (Behr et al. 2013), iQuant (iQuant et al. 2011), and others (Jiang & Wong 2009; Bohnert et al. 2009, Feng et al. 2010; Feng et al. 2011).

In addition to transcript-level quantification software, a number of packages focusing on quantifying splicing inclusion at the level of individual alternative splicing events (rather than the more complicated problem of analyzing full transcripts) have been developed, including MISO (Katz et al. 2010), KISSPLICE (Sacomoto et al. 2012), MATS (Shen et al. 2012), DiffSplice (Hu et al. 2013), MMES (Wang et al. 2010), SpliceTrap (Wu et al. 2011), DEXSeq (Anders et al. 2012), SplicingCompass (Aschoff et al. 2013), PSGInfer (LeGault & Dewey 2013), and others.

Finally, the assembly problem has been addressed by multiple approaches too. Those based on aligning reads to a reference genome include Cufflinks, mGene (Behr et al. 2010), RNASEQR (Chen et al. 2012), G-Mo.R-Se (Denoeud et al. 2008), Montebello (Hiller & Wong 2013), Rnnotator (Martin et al. 2010), DRUT (Mangul et al. 2012), GRIST (Boley et al. 2014), CRAC (Philippe et al. 2013), MITIE (Behr et al. 2013), and others (Jackson et al. 2009; Bao et al. 2013; Seok et al. 2012). Alignment-free *de novo* reconstruction programs include Oases (Schulz et al. 2012), Trinity (Grabherr et al. 2011; Haas et al. 2013), SOAPdenovo-Trans

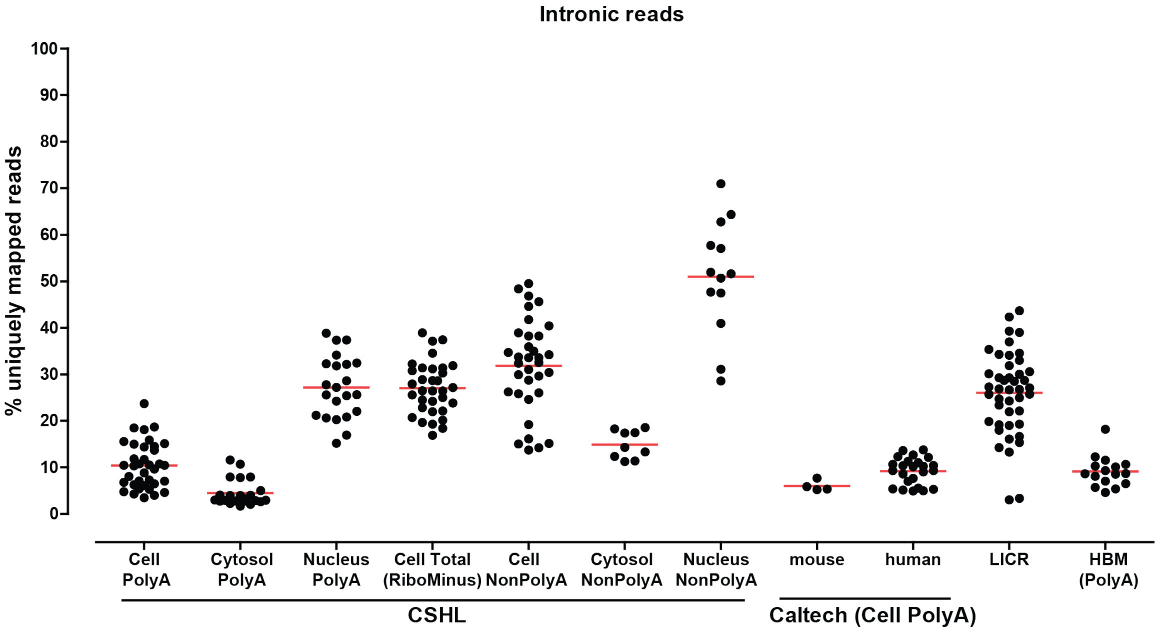


Figure 2.2: Distribution of the fraction of intronic reads in ENCODE datasets. Shown is the fraction of intronic reads in different ENCODE datasets (downloaded from the USCS Genome Browser) as well as the Human Body Map dataset (HBM).

(Xie et al. 2014), Trans-ABySS (Robertson et al. 2010) and EBARDenovo (Chu et al. 2013).

Most transcript-level quantification programs adopt a variation of a common likelihood-based approach to the problem (first discussed in Xing et al. 2006):

$$\mathcal{L}(\Theta) = P(\mathcal{O}|\Theta) \tag{2.1}$$

Where Θ refers to the unknown parameters of the model (for example, the relative abundances of individual isoforms) and $\mathcal{O}$ is the set of observations (for example, the set of alignments to the genome or the transcriptome).

Perhaps the most general version of this likelihood function, which incorporates the majority of complexities that are modeled by various quantification algorithms, is the following (Pachter 2011):

$$\mathcal{L}(\Theta) = \prod_{(t_G)\in(G,T)} \prod_{f\in F_{(G,T)}} \sum_{(t,i)\in(t_G)} \frac{1}{\widetilde{l}_{t_G}} \Theta_{t_G} \frac{D_{FL}(l_{t_G}(f))}{\sum_{k=1}^{i-1} D_{FL}(i-k)} w^{3'}_{(t,i)} w^{5'}_{(t,i-l_t(f)+1)} w^{pos}_{\frac{i}{l_{t_G}}} e_{t_G,f} \tag{2.2}$$

where:

- t_G refers to a transcript t belonging to gene G.
- (G,T) refers to the set of genes G and their transcripts T between which reads are to be allocated.
- Θ refers to the isoform abundance assignments. For a given gene G, $\sum_{t\in G} \Theta_{t_G} = 1$.
- f is a sequencing fragment; both ends of a fragment are sequenced in paired-end format.
- $F_{(G,T)}$ refers to the set of fragments aligning to transcripts T in a gene G, or a set of genes $\{G_1, ..., G_n\}$ such that a subset of fragments $F_s \subseteq F$ align ambiguously to transcripts of more than one gene.
- (t,i) refers to position i in transcript t.
- D_{FL} is the fragment length distribution.
- $w^{3'}_{(t,i)}$ is a term accounting for coverage bias at the 3' end of fragments (Li et al. 2010).
- $w^{5'}_{(t,j)}$ is a term accounting for coverage bias at the 5' end of fragments.
- $w^{pos}_{\frac{i}{l_t}}$ is a positional bias term, accounting for systematic coverage biases along the length of the transcript.
- $e_{t_G,f}$ is the probability that the alignment is correct; it accounts for mapping errors.
- $\widetilde{l}_{t_G}$ is the effective length of each transcript, calculated as follows:

$$\widetilde{l}_{t_G} = \sum_{i\in t_G} \left(\sum_{j=1}^{i-1} \frac{D_{FL}(i-j)}{\sum_{k=1}^{i-1} D_{FL}(i-k)} w^{3'}_{(t,i)} w^{5'}_{(t,i-l_t(f)+1)} w^{pos}_{\frac{i}{l_{t_G}}} \right) \tag{2.3}$$

In most cases the parameters are inferred using some variation of the expectation-maximization (EM) algorithm (Dempster et al. 1977). This is done in three general ways (Figure 2.1A-C): from splice-aware alignments to the genome (for example, by Cufflinks), from alignments to the transcriptome (examples include eXpress and RSEM), and without any alignments (the k-mer counting approach adopted by Sailfish). Unfortunately, the likelihood model is not always identifiable (see discussion in Hiller et al. 2009 and the supplement of Trapnell et al. 2010). Identifiability becomes increasingly difficult to achieve with the increase of isoform complexity (as shown empirically in the previous chapter), which in plain terms is the result of the fact that the more isoforms there are in the annotation, the more likely it is that no fragments that can unambiguously distinguish all of them are present in the data.

The approaches to the *de novo* assembly are somewhat more varied. For example, the most popular alignment-based approach (Cufflinks; Trapnell et al. 2010) aims to return the minimal set of transcripts that can explain the observed data (subject to some constraints on absolute abundance), while the approach adopted by GRIT (Boley et al. 2014) is to identify all possible expressed isoforms and then rank them by their estimated abundance. Alignment-free assembly algorithms usually employ de Bruijn graphs (de Bruijn 1946) to tackle the problem, which have been extensively used for assembling genomes from short reads (Pevzner & Tang 2001; Pevzner et al. 2011; Zebrino & Birney 2008; Butler et al. 2008; Gnerre et al. 2011; Luo et al. 2012; Bankevich et al. 2012; Simpson et al. 2009; Zimin et al. 2013).

All the results presented in the previous chapter depend critically on the ability of the software used to faithfully carry out the tasks of read mapping and transcript quantification and reconstruction, thus which programs return the most reliable output and to what extent it can be trusted is of utmost importance for their interpretation. However, an interesting phenomenon is observed in the literature: each publication of a new package concludes that it outperforms all other existing tools, usually by carrying out simulations that demonstrate this is the case against known ground truth. This is problematic, first, because of its clear logical impossibility, and second, because the simulations are usually not very realistic as they do not model some data properties that working with data has lead me to suspect are actually having a significant negative effect on results – for example, it is usually the case that isoforms from the refSeq annotation (which does not contain many alternative splicing products) are simulated, with no reads coming from the intronic or intergenic space, making the problem much easier to solve than the challenge presented by real data.

There are multiple known or suspected variables that affect both how difficult the problems of isoform abundance estimation and transcript reconstruction are and how well they can be solved. These include:

1. **Annotation complexity**. As already mentioned, more complex annotations present a greater challenge to quantification software even if only a single isoform is actually expressed as reads have to be properly allocated between more and more transcripts. Annotations range from simple (i.e. refSeq, mostly one or two isoforms per gene) to intermediately complex (i.e. UCSC) to very complex (i.e. GENCODE, with up to 10 isoforms per gene on average).

2. **Isoform expression complexity**. The more isoforms are expressed in the sample, the more different splice junctions there are to be parsed between them, which would be expected to be more difficult to do than if only a single isoform is expressed. This affects both quantification and assembly.

3. **Data quality**. PolyA-selected RNA-seq can suffer from several kinds of data deficiencies. First, suboptimal PolyA selection can result in larger amounts of intronic reads (although this can also be a purely biological phenomenon). At high sequencing depths, this could pose problems for both transcript assembly and quantification as shorter introns can get completely filled-in with reads leading to incorrect inference of retained intron isoforms. Wide variation of the fraction of intronic reads is observed between different protocols, production centers and biological sources (especially subcellular fractions), as shown in Figure 2.2. Second, RNA degradation can result in coverage being skewed towards the 3' end, which makes parsing alternative splicing events around the 5' end more difficult (even if algorithms try to normalize for such biases; i.e. through the $w^{pos}_{\frac{i}{l_t}}$ term above)

4. **Library construction protocol**. Both stranded and unstranded protocols are in wide use for RNA-seq. Stranded libraries are expected to provide more power for accurate transcript reconstruction and

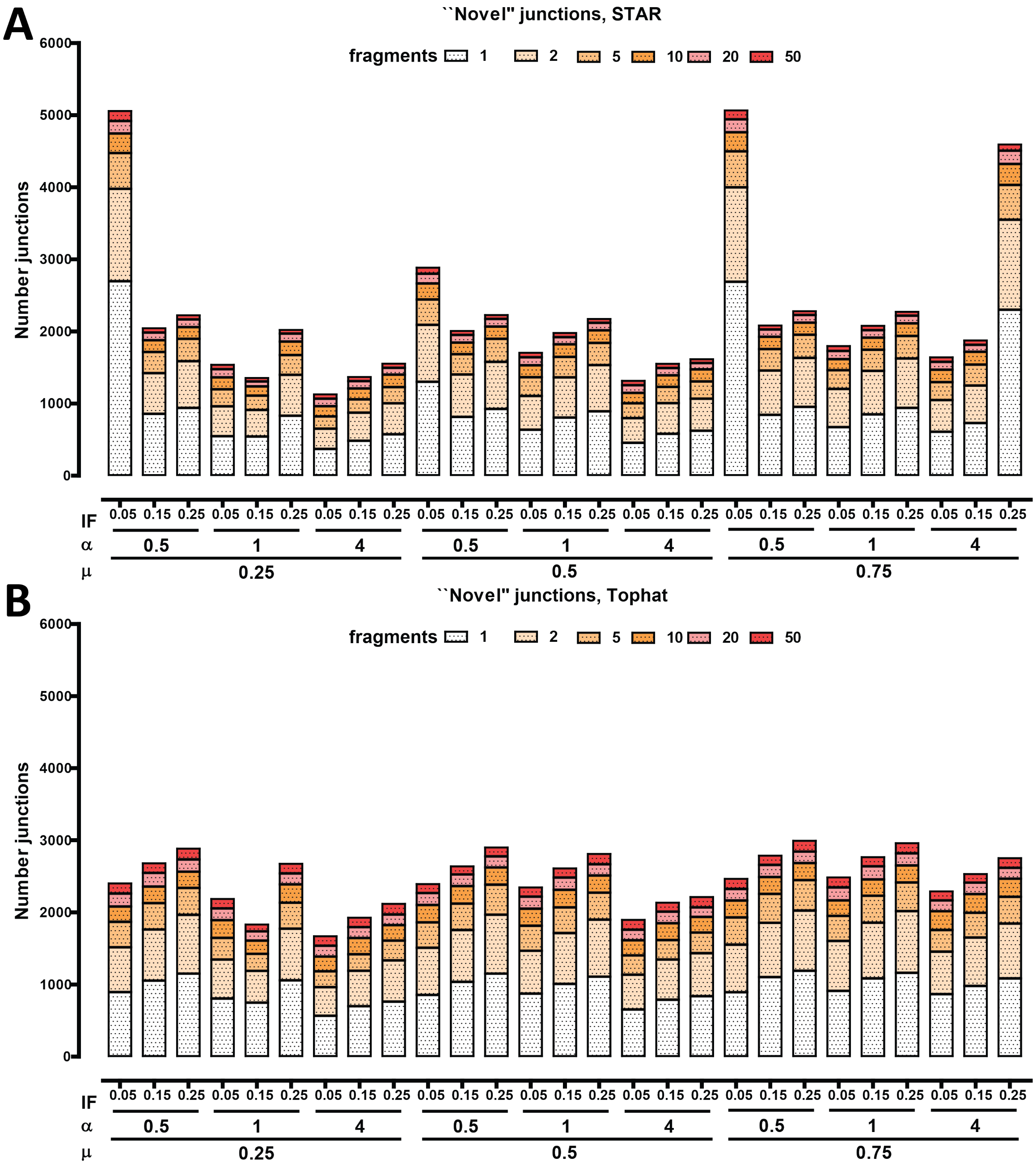


Figure 2.3: Number of "novel" splice junctions detected by STAR and TopHat. Shown is the number of junctions not annotated in GENCODE V16 detected at different levels of coverage (measured in collapsed, unique fragments) by the two mappers. Note that only annotated transcripts were used in the simulation, i.e. no novel junctions are expected to be detected, and the ones that are represent false positives.

quantification as they allow the resolution of overlapping sense and anti-sense transcripts. However, it has to be noted that some stranded protocols (dUTP in particular) are not absolutely strand-specific.

5. **Fragment length distribution**. During library construction, RNA is fragmented, usually to pieces of 200 to 300 nucleotides length. The exact fragment length can have a significant effect on transcript as-

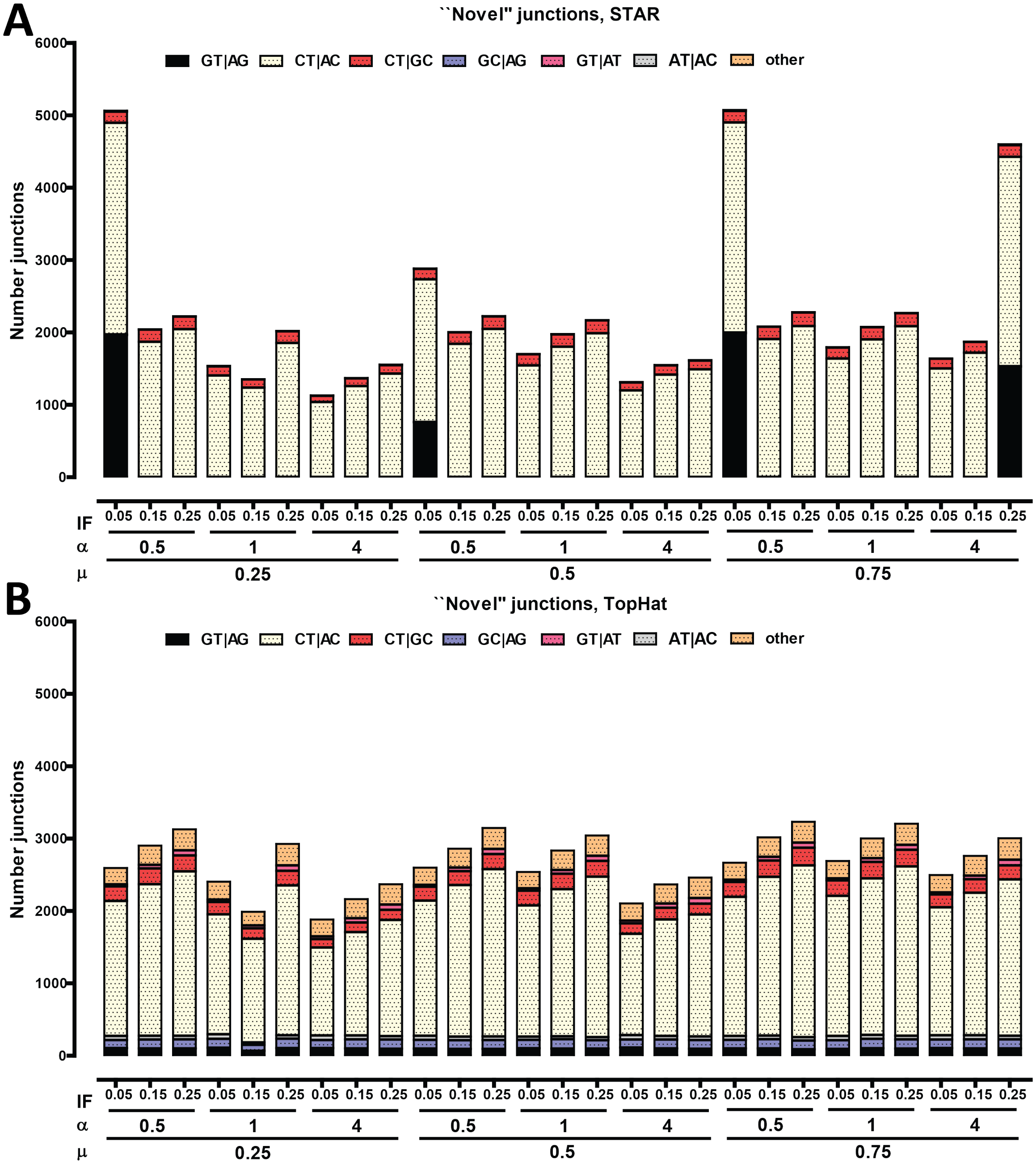


Figure 2.4: Sequence type of "novel" splice junctions detected by STAR and TopHat. Shown is the number of junctions not annotated in GENCODE V16 detected at different levels of coverage by the two mappers split by the sequence of their splicing motifs. Canonical junctions recognized by the major spliceosome are of the GT|AG type, the two major classes of non-canonical junctions are GC|AG and AT—AC.

sembly and quantification. Longer fragment lengths provide greater connectivity of distant sequences, but they lead to stronger coverage and representation biases.

6. **Read length**. For obvious reasons, it is intuitive to think that longer reads will always result in better assembly and quan-

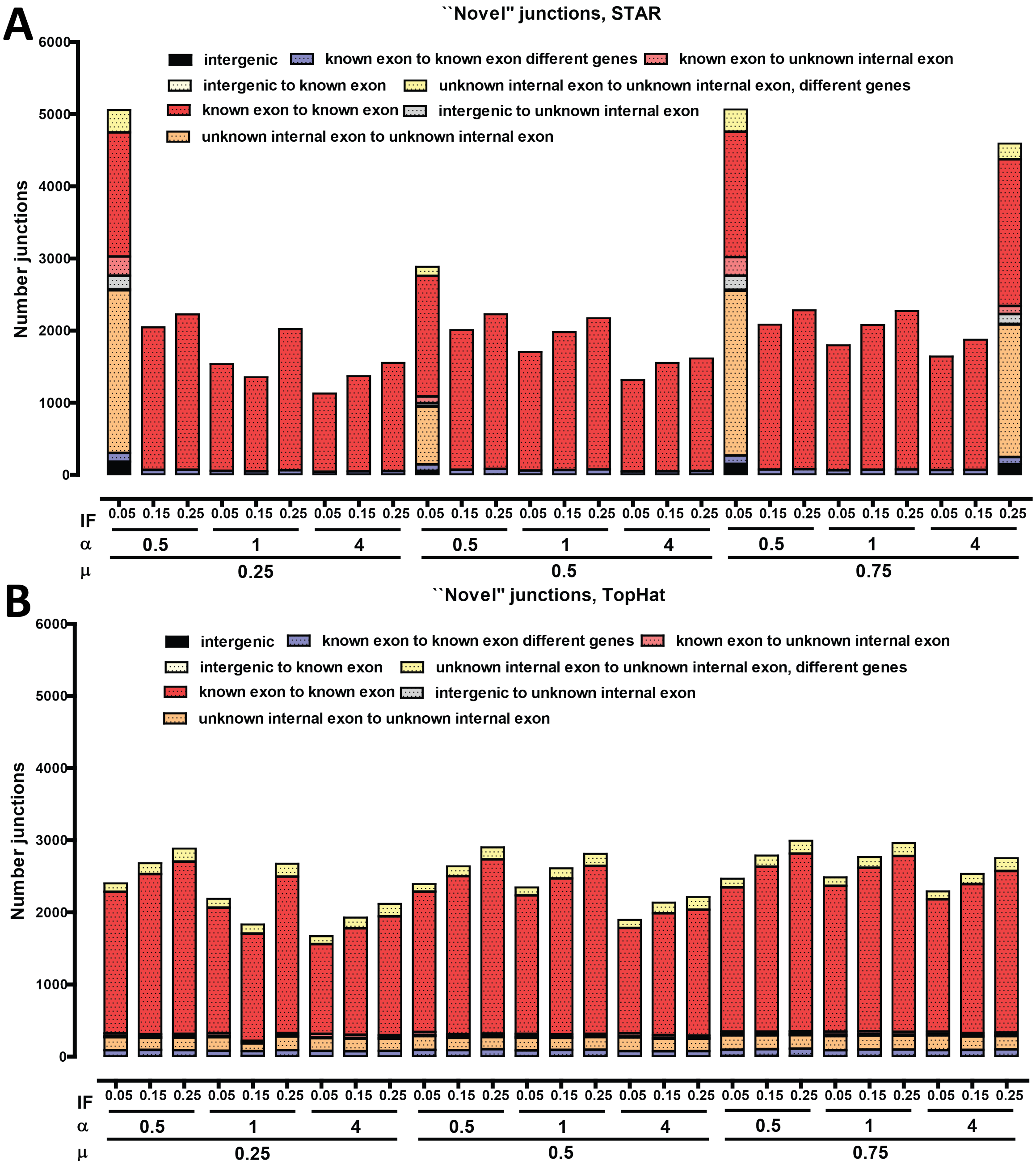


Figure 2.5: Classification of "novel" splice junctions detected by STAR and TopHat relative to the annotation. Shown is the number of junctions not annotated in GENCODE V16 divided according to how they relate to the annotation (GENCODE V16). The categories are introduced and detailed in the previous chapter.

tification. However, long reads only make sense if the fragment size distribution is correspondingly long, and as mentioned above, longer fragment distribution leads to poorer quantification results.

7. **Sequencing depth**. Again, for obvious reasons, deeper sequencing provides more quantification and assembly power.

It is not practically possible to examine all these variables due to the high dimensionality of the parameter space, much less against the very large number of software tools (new versions of which, as well as new algorithms, are continuously being published). I chose to focus on a limited set of the most popular analysis packages that still represents the range of existing approaches to the problems and on the data characteristics that are in my opinion most relevant to ENCODE results and least studied, while picking optimal parameter values for the others. These parameters were the isoform expression complexity and the impact of data quality, in particular the prevalence of intronic reads (due to the presence of numerous retained introns in Cufflinks assemblies discussed in the previous chapter).

2.2 Methods

2.2.1 Simulation parameters

For the purposes of this comparison, it is irrelevant what gene-level expression values are used for the simulation although matching real-life data is in no way a negative. Therefore, Cufflinks-derived gene-level quantification estimates for actual samples were used as a starting point from which isoform expression levels were assigned to individual transcripts. These estimates are in FPKM (**F**ragments **P**er **K**ilobase per **M**illion fragments), where we define FPKM for a transcript as follows:

$$FPKM_T = \frac{\text{Number fragments mapping to a transcript}}{\frac{\text{Total number of mapped fragments}}{1{,}000{,}000} * \frac{\text{Length of transcript}}{1{,}000}} \tag{2.4}$$

Here a fragment is defined as a pair of reads when both ends of a paired-end read are mapped or as the read itself when it is a singleton or the sequencing data is single-end.

For a gene G which contains N individual transcripts $T_{G_{0,\ldots,N}}$, there are two ways to define FPKMs on the gene level:

$$FPKM_G = \frac{|\text{fragments}_{b \in B_G}|}{\frac{\text{Total number of mapped fragments}}{1{,}000{,}000} * \frac{|B_G|}{1{,}000}} \tag{2.5}$$

Here we count all fragments $fragments_b$ mapping to a base pair b belonging to all base pairs B_G annotated as part of the gene G and normalize against the total number of base pairs $|B_G|$.

The alternative option is to calculate FPKM as follows:

$$FPKM_G = \sum_{T \in G} FPKM_{T_G} \tag{2.6}$$

Which is the sum of the FPKMs estimated for each individual transcript.

The latter is the more biologically correct way of calculating FPKMs as it normalizes better for cases in which an isoform that is very large or very short relative to the total gene length is expressed (Trapnell et al. 2010; Pachter 2011), and is therefore the one adopted here.

For each transcript of a gene, we can define the FMI (**F**raction of **M**ajor **I**soform) quantity as follows:

$$FMI_{T_G} = \frac{FPKM_{T_G}}{\max\limits_{T \in G}(FPKM_{T_G})} \tag{2.7}$$

The FMI values can be used to determine isoform expression complexity. Examination of the distribution of FMI values on real data (with the caveat that real-life isoform-level quantification is unreliable, although this is not relevant for simulation purposes) using Cufflinks showed that the median FMI of the second major isoform is around 0.5, of the third major isoform between 0.25 and 0.3, of the fourth major isoform between 0.10 and 0.15, etc. (see previous chapter). However, the distribution of the FMI for the second major isoform is not normal but actually roughly uniform with some bias towards 0. Uniform distribution is not well suited for the goals of this simulation exercise because a way to vary the isoform complexity is needed and this would be better manipulated through shifting the means (see below). Therefore the FMI distribution was modeled with a Gaussian, with mean μ and variance σ^2, which is dispersed and truncated by requiring that $\mu = \sigma$, i.e. for any FMI μ that is picked, the left $1 - \sigma$ position in the distribution is 0. For each gene G with N individual isoform, $T_{G_{0,\dots,N}}$ ranked by expression such that $FMI_{(T_G)_i} > FMI_{(T_G)_{i+1}}$, the FMI for each isoform is chosen as follows:

$$FMI_{(T_G)_i} = \begin{cases} 1 & \text{if } i = 0; \\ \max(0, FMI_{(T_G)_{i+1}} \sim \dfrac{1}{\displaystyle\int_{-\infty}^{FMI_{i-1}} \mathcal{N}_{FMI_i}} \{\mathcal{N}_{FMI_i} : FMI_i < FMI_{i-1}\}) & \text{if } i > 0 \end{cases} \tag{2.8}$$

Where:

$$\mathcal{N}_{FMI_i} = \mathcal{N}(\mu_1^{i\alpha}, \sigma_i^2 = {\mu_1^{i\alpha}}^2) \tag{2.9}$$

$\mathcal{N}$ refers to a Gaussian, and they key parameters are the mean FMI for the second ranked isoform (μ_1) and α, which are used to scale the global isoform complexity (higher α will lead to much quicker decay of the mean FMI). Note that the Gaussian is rescaled to take into account the fact that only the parts of it between $-\infty$ and the FMI of the next more highly expressed isoform of the gene are considered, so that if the randomly chosen FMI value was less than zero, it was set to 0, at which point all subsequent isoforms were set to zero too. The isoform ranking was also picked at random for each gene.

2.2.2 Read simulation

A reasonable very deeply sequenced RNA-seq dataset contains ~200×10^6 reads, or about one lane of HiSeq worth of reads. It is also what ENCODE produced for most of its samples (Djebali et al. 2012). For this reason, the total sequencing depth was fixed at $R = 200 \times 10^6$ read pairs (or double the ENCODE number, i.e. a very deeply sequenced sample). It is known that long fragment sizes actually degrade the performance of RNA-seq. This is because:

1. Short transcripts are underrepresented by reducing their effective length (there are only 200 positions in which a 400bp-long fragment can originate from a 600bp-long trancsript, but 1600 such positions for a 2kb-long transcript). Quantification programs perform an effective length normalization, which takes some of these biases into account. However, another issue still remains unresolved, and it is experimental in nature:

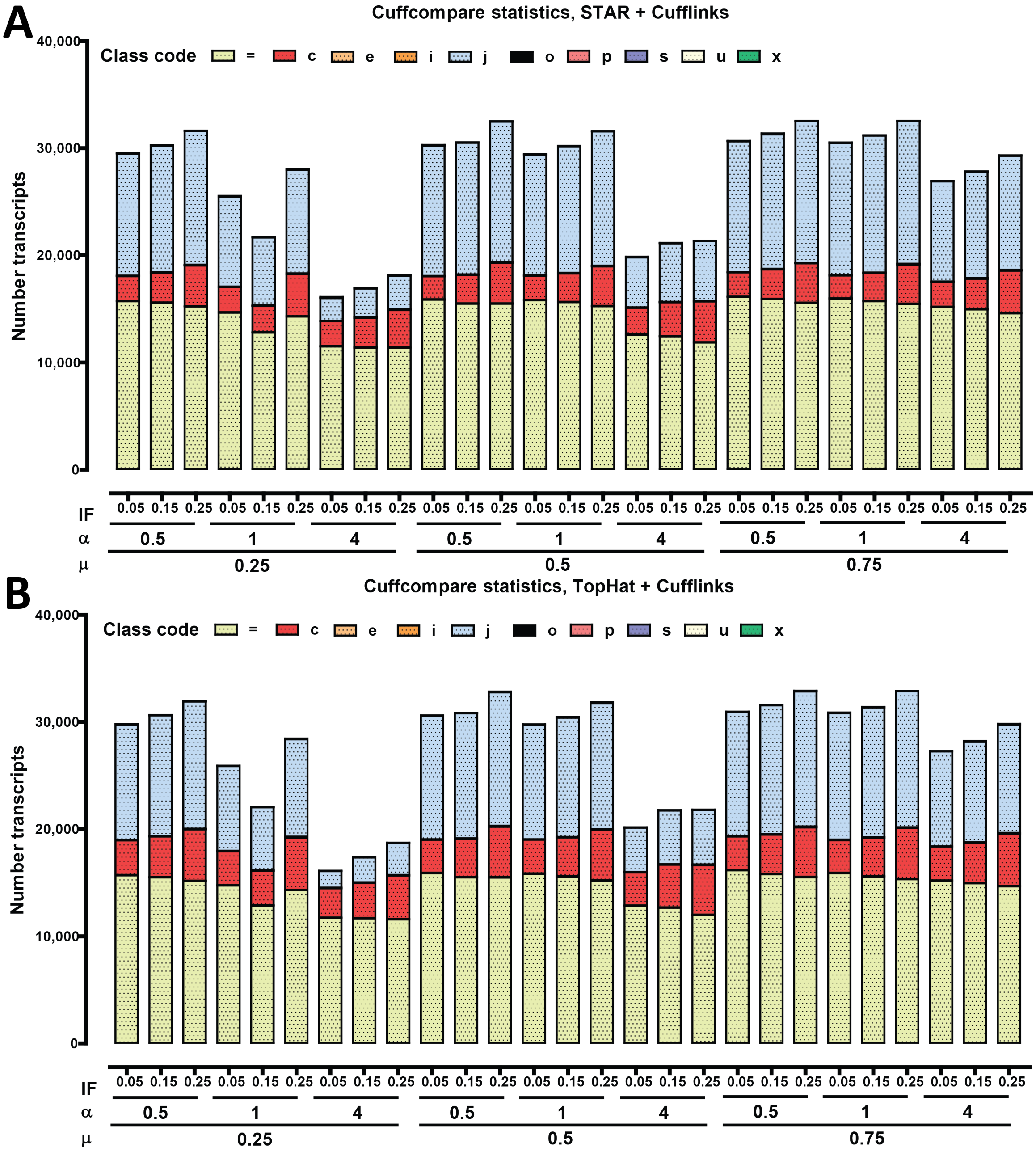
2. Short transcripts are underrepresented in the sequencing libraries. Suppose fragments were size-selected so that they are distributed as a Gaussian, i.e. $D_{FL} \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu = 500$ and $\sigma = 100$, and consider the same case of the 600bp-long transcript and the 2kb-long transcript described above. Fragments are generated by random fragmentation of either RNA molecules or cDNA, depending on the protocol used. Assuming this fragmentations is random, on average only 1/3 of fragments will be within 100bp of the mean of the size-selection range for the 600bp-long transcripts, i.e. each transcript or full-length cDNA molecule will be represented in the library 1/3 of the time, while the 2kb-long one will usually contribute 3 fragments to it.

3. A significant contributor to uneven sequencing coverage in RNA-seq seem to be RNA secondary structures and more and more complex such structures are formed in longer RNA molecules. Depending on the protocol used, this may have a more or less severe negative effect on transcript representation and coverage in the final libraries.

For these reasons, reads were sampled from a fragment size distribution centered around 250bp with standard deviation of 50 (i.e. $D_{FL} \sim \mathcal{N}(250, 50^2)$), and the length of the reads was limited to 2x100bp. The `mason` read simulator (Holtgrewe 2010) was used for simulating the reads. The simulation was carried out as follows:

1. Separate "chromosomes" were generated for each transcript using the GENCODE V16 annotation and the human genome reference sequence.

2. Separate "chromosomes" were also generated for the unspliced, pre-mRNA form of each GENCODE V16 transcript. This is not entirely realistic as in reality splicing is predominantly cotranscriptional (Dujardin et al. 2013), and pre-mRNAs exist in a partially spliced state but rarely in a completely unspliced one. But this process is generally poorly understood so for simplic-



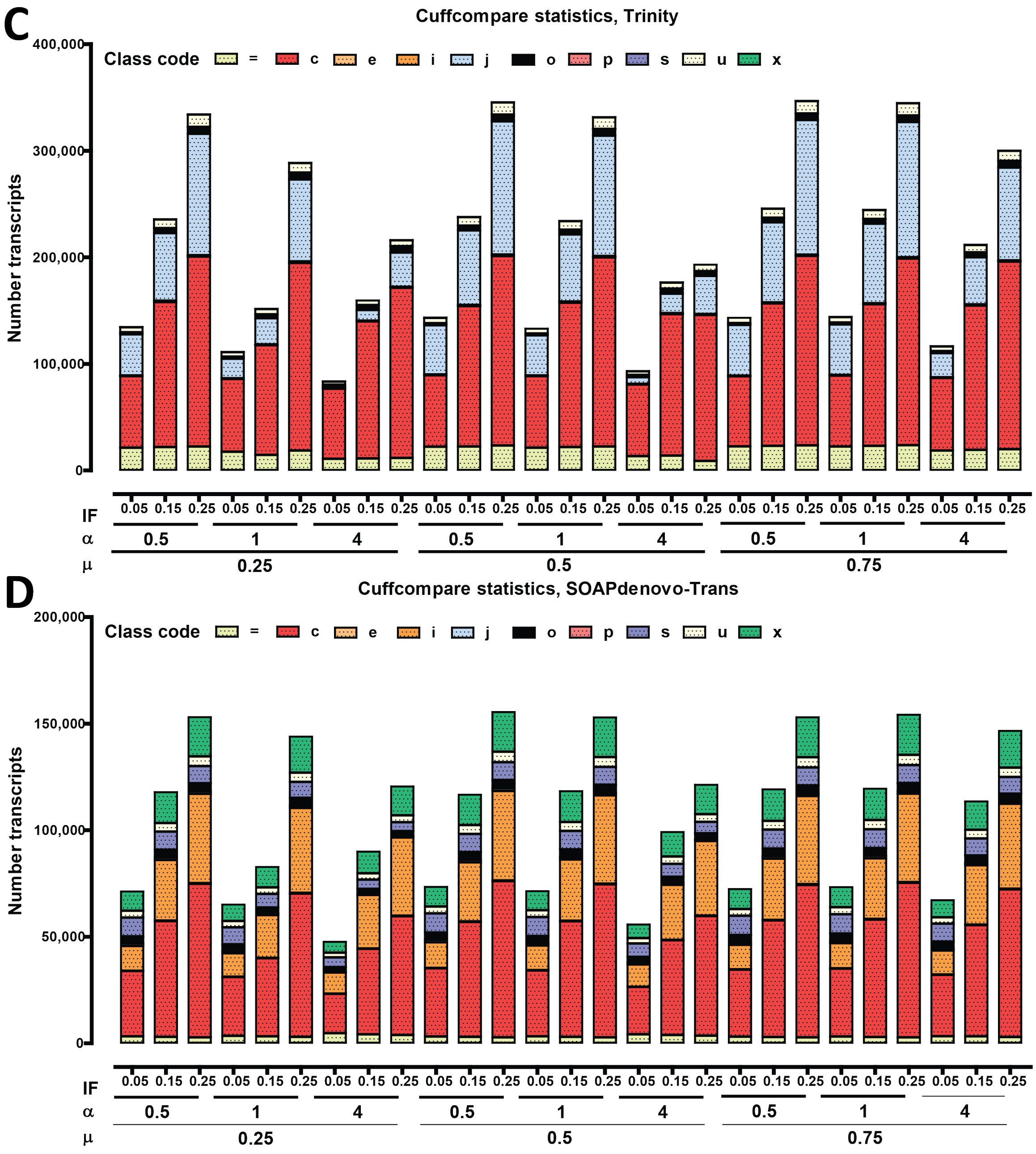


Figure 2.6: Number of transcripts in each assembly according to the Cuffcompare classification. The number of transcripts in each Cuffcompare category is shown for Cufflinks assemblies on STAR (A) and TopHat (B) mapping and for de novo Trinity (C) and SOAPdenovo-trans (D) assemblies. The Cuffcompare codes are defined (and prioritized during classification in the same order) as follows (Trapnell et al. 2010; Trapnell et al. 2012): *"**=**": Complete match of intron chain; "**c**": Contained; "**j**": Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript; "**e**": Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment; "**i**": A transfrag falling entirely within a reference intron; "**o**": Generic exonic overlap with a reference transcript; "**p**": Possible polymerase run-on fragment (within 2Kbases of a reference transcript); "**r**": Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case; "**u**": Unknown, intergenic transcript; "**x**": Exonic overlap with reference on the opposite strand; "**s**": An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)*

ity here it is assumed that transcription and splicing are completely uncoupled.

3. An ENCODE K562 sample RNA-seq sampled was used to obtain real-life gene-level FPKM estimates using Cufflinks (version 2.0.2). Note that only protein coding genes were included, which was done deliberately, with the goal of examining the performance of quantification software with respect to pseudogenes and lincRNAs, for which mapping artifacts might confound output (due to close sequence homology with protein coding genes in the case of pseudogenes, and due to the presence of repetitive elements in many lincRNAs).

4. Isoform-level FPKMs were simulated from the gene-level FPKMs as described above, using all 9 combinations of $\mu = 0.25, 0.5$ or 0.75 and $\alpha = 0.5, 1$, or 4. A value of $\alpha = 4$ means almost no alternative isoform expression), while when $\alpha = 0.5$, the 10th highest isoform will still have on average $\Theta = 0.03$ (see below for definition of Θ).

5. For each such combination, 3 datasets with a different intronic fraction of reads were simulated (IF = 0.05, 0.15 or 0.25). IF = 0.05% corresponds to some of the best polyA-selection cases we have observed in practice, IF = 0.15 can be considered intermediate level of intronic reads, and IF = 0.25 is what is often observed in some nuclear subcellular fractions in ENCODE data (though much higher values have also been seen; Figure 2.2).

6. Using the IF and transcript-level FPKM values, the number of reads that should be simulated for each transcript containing introns and its corresponding pre-mRNA was calculated. The intronic fraction was constant for all transcripts.

7. Stranded RNA-seq reads were generated for each mRNA and pre-mRNA using `mason`, with the following settings: `illumina --read-length 100 --library-length-mean 250 -le 50 --include-read-information --forward-only --simulate-qualities --mate-pairs --prob-insert 0 --prob-delete 0 --haplotype-snp-rate 0 --haplotype-indel-rate 0`. Reads were subsequently renamed to records their origin and proper mapping.

8. FPKMs were rescaled according to the IF value so that intronic reads are excluded from the denominator in the calculation of the true FPKM value. The true FPKM values were recorded and saved.

The resulting simulated set of reads represents a somewhat easier to solve problem than real-life data does, as it does not model transcript coverage non-uniformity (the sources of which are not entirely understood). However, it does provide a measure of the relative performance of programs, as well as minimum bounds on the fraction of incorrectly quantified and assembled transcripts, which is still informative with respect to the interpretation of the results in the previous chapter.

2.2.3 Read Mapping

Reads were mapped to the hg19 assembly of the human genome using both the STAR (version 2.3.0e; Dobin et al. 2013) and TopHat (Version 2.0.8; Trapnell et al. 2009; Trapnell et al. 2012b) aligners, using the GENCODE V16 as a source of annotated transcripts and junctions to aid mapping. The following settings were used for STAR; default settings were used for TopHat. `--outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000`.

For RSEM and eXpress quantifications, reads were mapped against a GENCODE V16 transcriptome index, using Bowtie (version 0.12.7; Langmead et al. 2009), with the following settings: `-e 200 -a --offrate 1 -t -X 1000`.

2.2.4 Transcript assembly and reconstruction

Cufflinks (version 2.0.2; Trapnell et al. 2010; Trapnell et al. 2012a) was used for assembly on STAR and TopHat alignemtns, with default settings except for specifying that the libraries are stranded. Scripture (Guttman et al. 2010) was also tested; however, its computational requirements were too large and made running it on all simulated datasets practically impossible.

For alignment-free assembly, Trinity (Grabherr et al. 2011; Haas et al. 2013) was used with the following settings: `--SS_lib_type FR --min_kmer_cov 2`, and SOAPdenovo-Trans (Xie et al. 2014) was run, with the following settings: `SOAPdenovo-Trans-31mer max_rd_len=100 avg_ins=250 reverse_seq=0`. BLAT (Kent 2002) was used to map the resulting contigs back to the genome, with only contigs longer than 200bp considered. Custom-written python scripts were used to convert the resulting PSL-format output to GTF format, while retaining only the best alignment(s) for each contig.

2.2.5 Isoform-level quantification

Cufflinks (version 2.0.2; Trapnell et al. 2010; Trapnell et al. 2012a) was run on both STAR and TopHat alignemtns, with default settings except for specifying that the libraries are stranded.

RSEM (version 1.2.7; Li et al. 2010; Li et al. 2011): was run as follows: `--calc-ci --forward-prob 1`. eXpress (version 1.5.0; Roberts & Pachter 2013) was run with default settings. Both were run on Bowtie alignments.

Sailfish (version 0.5.0; Patro et al. 2014) was run with default settings and $k = 20$.

A number of other packages were also tested: CEM/IsoLasso (Li et al. 2011), Flux-Capacitor, IQSeq (Du et al. 2012), iReckon (Mezlini et al. 2013), IsoEM (Nicolae et al. 2011), MMSeq (Turro et al. 2011), PennSeq (Hu et al. 2014), RNAExpress (Forster et al. 2013), SLIDE (Li et al. 2011), and Traph (Jo et al. 2014), However, all of them turned out to be practically impossible to run due to dependency issues with software no longer being maintained and/or computational requirements (for example, Penn-Seq took more than a week running on 8CPUs and 40GB of memory without showing any signs of convergence).

2.2.6 Metrics for evaluation of quantification performance

The following metrics were used to evaluate quantification performance:

1. The Pearson correlation r between the true FPKMs and the estimated FPKMs on the gene level
2. The Pearson correlation r between the true FPKMs and the estimated FPKMs on the transcript level
3. The mean total Θ difference between the true relative isoform abundances in each gene and the estimated isoform abundances:

$$MT\Theta_{diff} = \frac{\sum_{G}\sum_{T\in G}|\Theta_{E(T_G)} - \Theta_{T(T_G)}|}{N_G} \tag{2.10}$$

Where N_G is the total number of annotated genes considered, Θ_E is the estimated Θ and Θ_T is the true Θ for each isoform of a gene, and Θ is defined as:

$$\Theta_{T_G} = \frac{FPKM_{T_G}}{\sum_{T\in G}(FPKM_{T_G})} \tag{2.11}$$

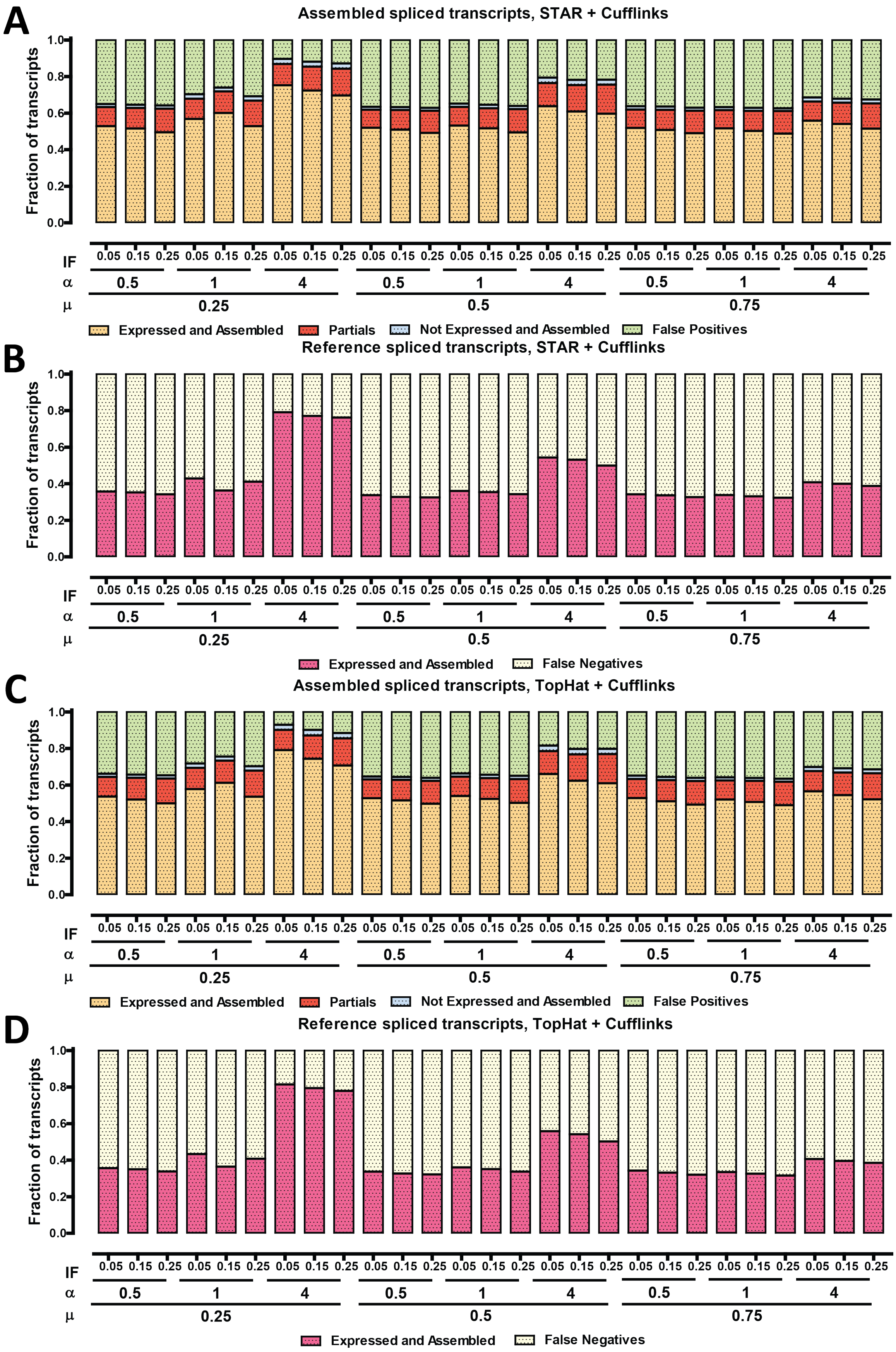
Note that the possible values of $MT\Theta_{diff}$ are limited to $MT\Theta_{diff} \in [0, 2]$, with $MT\Theta_{diff} = 0$ corresponding to perfectly accurate parsing of reads between isoforms and $MT\Theta_{diff} = 2$ to complete misallocation (for example, if only one isoform is expressed but it received 0 FPKM and the reads were instead allocated to other isoforms).

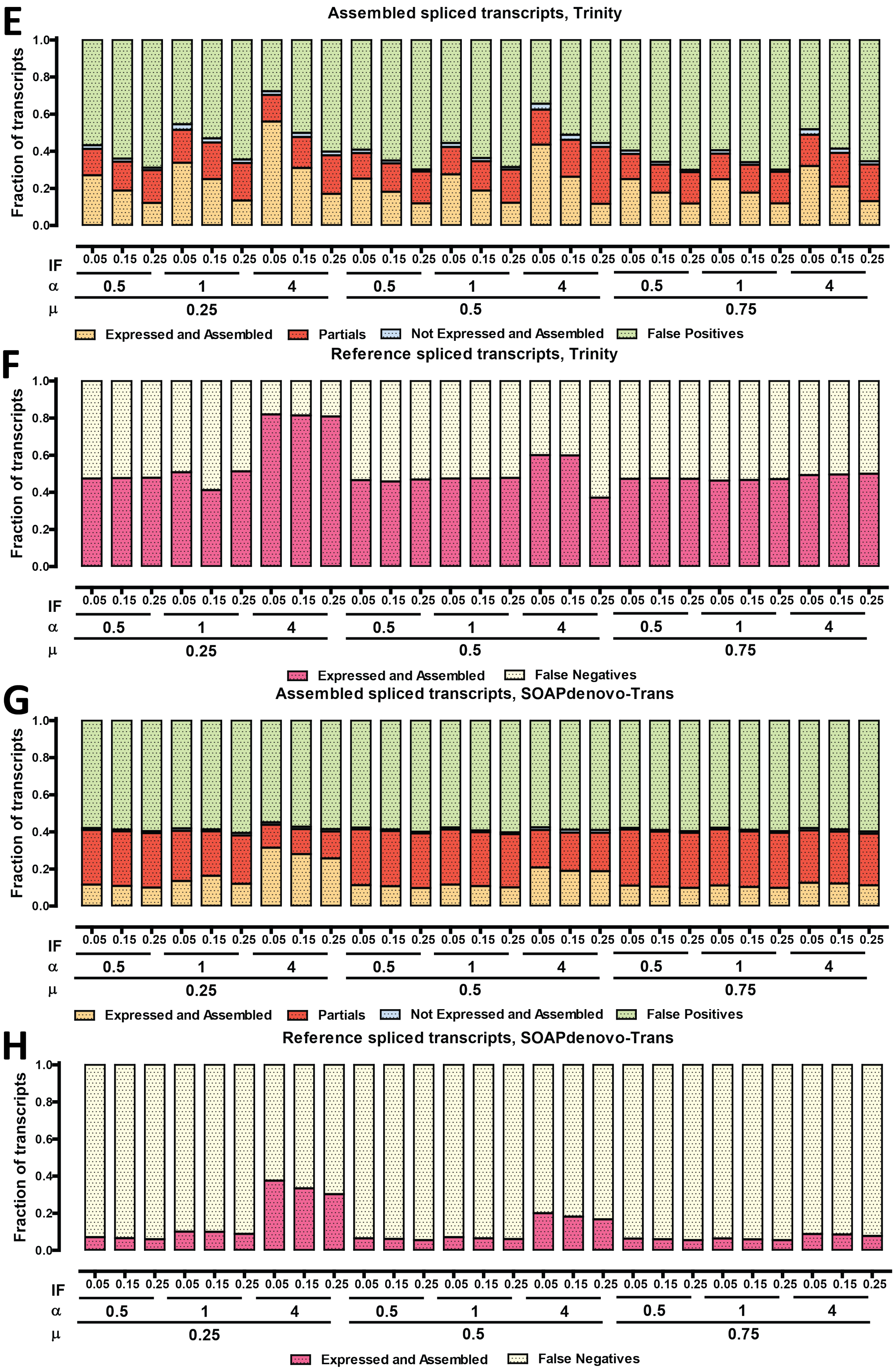
4. The fraction of genes with an incorrectly assigned major isoform, i.e.:

$$\underset{T_G}{\operatorname{argmax}}(\max_{T\in G}(\Theta_{E(T_G)})) \neq \underset{T_G}{\operatorname{argmax}}(\max_{T\in G}(\Theta_{T(T_G)}))$$

5. The fraction of genes with false positive isoforms
6. The fraction of false positive isoforms
7. The fraction of genes with false negative isoforms
8. The fraction of false negative isoforms

Here, a false positive isoform was defined as one with $\Theta_{T(T_G)} = 0$ and $\Theta_{E(T_G)} \geq 0.05$, and a false negative isoform as one with $\Theta_{T(T_G)} \geq 0.05$ and $\Theta_{E(T_G)} \leq 0.001$.


A
Assembled spliced transcripts, STAR + Cufflinks
Fraction of transcripts
1.0
0.8
0.6
0.4
0.2
0.0
IF
0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α
0.5 1 4 0.5 1 4 0.5 1 4
μ
0.25 0.5 0.75
Expressed and Assembled
Partials
Not Expressed and Assembled
False Positives
B
Reference spliced transcripts, STAR + Cufflinks
Fraction of transcripts
Expressed and Assembled
False Negatives
C
Assembled spliced transcripts, TopHat + Cufflinks
Fraction of transcripts
Expressed and Assembled
Partials
Not Expressed and Assembled
False Positives
D
Reference spliced transcripts, TopHat + Cufflinks
Fraction of transcripts
Expressed and Assembled
False Negatives


E
Assembled spliced transcripts, Trinity
Fraction of transcripts
0.0
0.2
0.4
0.6
0.8
1.0
IF
0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α
0.5 1 4 0.5 1 4 0.5 1 4
μ
0.25 0.5 0.75
Expressed and Assembled
Partials
Not Expressed and Assembled
False Positives
F
Reference spliced transcripts, Trinity
Fraction of transcripts
Expressed and Assembled
False Negatives
G
Assembled spliced transcripts, SOAPdenovo-Trans
Fraction of transcripts
Expressed and Assembled
Partials
Not Expressed and Assembled
False Positives
H
Reference spliced transcripts, SOAPdenovo-Trans
Fraction of transcripts
Expressed and Assembled
False Negatives

2.2.7 Metrics for evaluation of assembly performance

The following metrics were used to evaluate assemblies relative to the GENCODE V16 annotation that was used to generate the data.

1. The number of transcripts in the various Cuffcompare classes (Cuffcompare is a program in the Cufflinks suite used to compare annotations). See the legend of Figure 2.6 for detailed explanation.

2. The number of perfectly matching intron chains for expressed spliced transcripts, where an intron chain is defined as follows. Every transcript T_G in gene G is defined according to its exonic coordinates as the ordered set of exon left and right positions: $T_G := \{(l_1, r_1), ..., (l_n, r_n)\}$. An intron chain IC is defined as the ordered set of left and right intronic positions, i.e.: $IC_{T_G} := \{(r_1, l_2), ..., (r_{n-1}, l_n)\}$. Comparing the intron chains allows the 5' and 3' ends, which are very difficult to assemble precisely (and are often not precisely defined biologically to begin with) to differ. An annotated transcript T_G was defined as expressed if $FPKM_{T_G} > 0$.

3. The number of assembled but not expressed genes, i.e. transcripts with $FPKM_{T_G} = 0$, which were nevertheless expressed. This may sound counterintuitive, but is not impossible, and does in fact happen occasionally.

4. The number of partially assembled spliced transcripts, i.e. transcripts, the intron chain IC_A of which is a strict subset of the intron chain of some annotated transcript T_G, i.e. $IC_A \subset IC_{T_G}$.

5. The number of false positive spliced transcripts, i.e. transcripts with an intron chain that is inconsistent with the intron chains found in the annotation.

6. The number of false negative spliced transcripts, transcripts that were expressed but not assembled. A threshold of 1 FPKM was set to define a transcript as assembled.

2.3 Results

2.3.1 Splice junction discovery

The main goals of this simulation were to assess transcript quantification and reconstruction. For this purpose, reads were simulated from the protein coding portion of the GENCODE V16 transcriptome, and then it was again GENCODE V16 that was used when mapping the reads, i.e. there are no novel junctions to discover and the mapping process is maximally aided by the annotation, which in this case completely matches the source of the reads. Nevertheless the simulation is useful with respect to the minimum number of false positive junctions observed in real-life data, and their nature.

Reads were mapped with both TopHat and STAR, and the junctions detected extracted. The strand of the junctions was annotated based on the directionality of the reads. Figure 2.3 shows the number of "novel" splice junctions detected by each algorithm in each of the 27

Figure 2.7 *(preceding page)*: Assembly statistics for spliced transcripts. (A,C,E,G) The distribution of true positives ("Expressed and Assembled"), partial true positives ("Partials"), partial false positives ("Not Expressed and Assembled") and false positives ("False Positives") among *de novo* assembled transcripts is shown. The categories are defined as follows: "Expressed and Assembled" refers to transcripts that were expressed at > 0 FPKM in simulation and we assembled completely, i.e. have a complete intron chain match in the annotation; "Partials" refers to assembled transcripts the intron chain of which is a subset of the intron chain of an annotated transcript; "Not Expressed and Assembled" refers to transcripts with FPKM= 0 in the simulation, which were nevertheless assembled with a complete intron chain (this is not impossible in complex loci even if rare); the "false positives" are transcripts with intron chains that are not found in the annotation, neither as complete chains nor as subsets of annotated intron chains. (B,D,F,H) The distribution of true positives ("Expressed and Assembled") and false negatives ("False Negatives") among annotated transcripts expressed at > 1 FPKM in the simulation. A false negative is a transcript the complete intron chain of which was not found among the *de novo* assembled transcripts. (A,B) Cufflinks on STAR alignments; (C,D) Cufflinks on TopHat alignments; (E,F) Trinity; (G,H) SOAPdenovo-trans

Pearson correlation	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
	Cufflinks STAR	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	Cufflinks TopHat	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.95	0.95	0.97	0.96	0.97	0.97	0.97	0.97	0.96	0.96	0.93	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	eXpress	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	RSEM	0.97	0.97	0.97	0.97	0.97	0.97	0.95	0.95	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	Sailfish	0.95	0.93	0.92	0.95	0.94	0.92	0.94	0.92	0.91	0.95	0.93	0.92	0.95	0.93	0.92	0.95	0.93	0.89	0.95	0.93	0.92	0.95	0.93	0.91	0.95	0.93	0.91

0.90
0.92
0.94
0.96
0.98
1.00

Figure 2.8: Correlation between true and estimated FPKMs on the gene level. The Pearson correlation between the true gene-level FPKMs and the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters.

Pearson correlation	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
	Cufflinks STAR	0.87	0.86	0.86	0.86	0.85	0.84	0.82	0.80	0.80	0.87	0.85	0.86	0.87	0.86	0.86	0.84	0.83	0.78	0.87	0.87	0.86	0.88	0.87	0.86	0.86	0.85	0.85
	Cufflinks TopHat	0.85	0.84	0.84	0.84	0.84	0.82	0.80	0.78	0.78	0.86	0.83	0.84	0.85	0.84	0.84	0.83	0.81	0.76	0.86	0.85	0.84	0.86	0.85	0.84	0.85	0.84	0.83
	eXpress	0.90	0.88	0.86	0.90	0.87	0.84	0.87	0.82	0.79	0.90	0.86	0.86	0.90	0.88	0.86	0.88	0.84	0.78	0.91	0.88	0.87	0.91	0.88	0.87	0.90	0.87	0.85
	RSEM	0.92	0.89	0.88	0.91	0.88	0.86	0.88	0.82	0.80	0.92	0.87	0.88	0.92	0.89	0.87	0.89	0.85	0.79	0.92	0.90	0.88	0.92	0.90	0.88	0.91	0.88	0.86
	Sailfish	0.77	0.72	0.68	0.74	0.70	0.65	0.68	0.61	0.58	0.77	0.71	0.69	0.76	0.72	0.68	0.71	0.64	0.59	0.77	0.72	0.69	0.77	0.73	0.69	0.75	0.69	0.66

0.70
0.75
0.80
0.85
0.90
0.95

Figure 2.9: Correlation between true and estimated FPKMs on the transcript level. The Pearson correlation between the true transcript-level FPKMs and the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters.

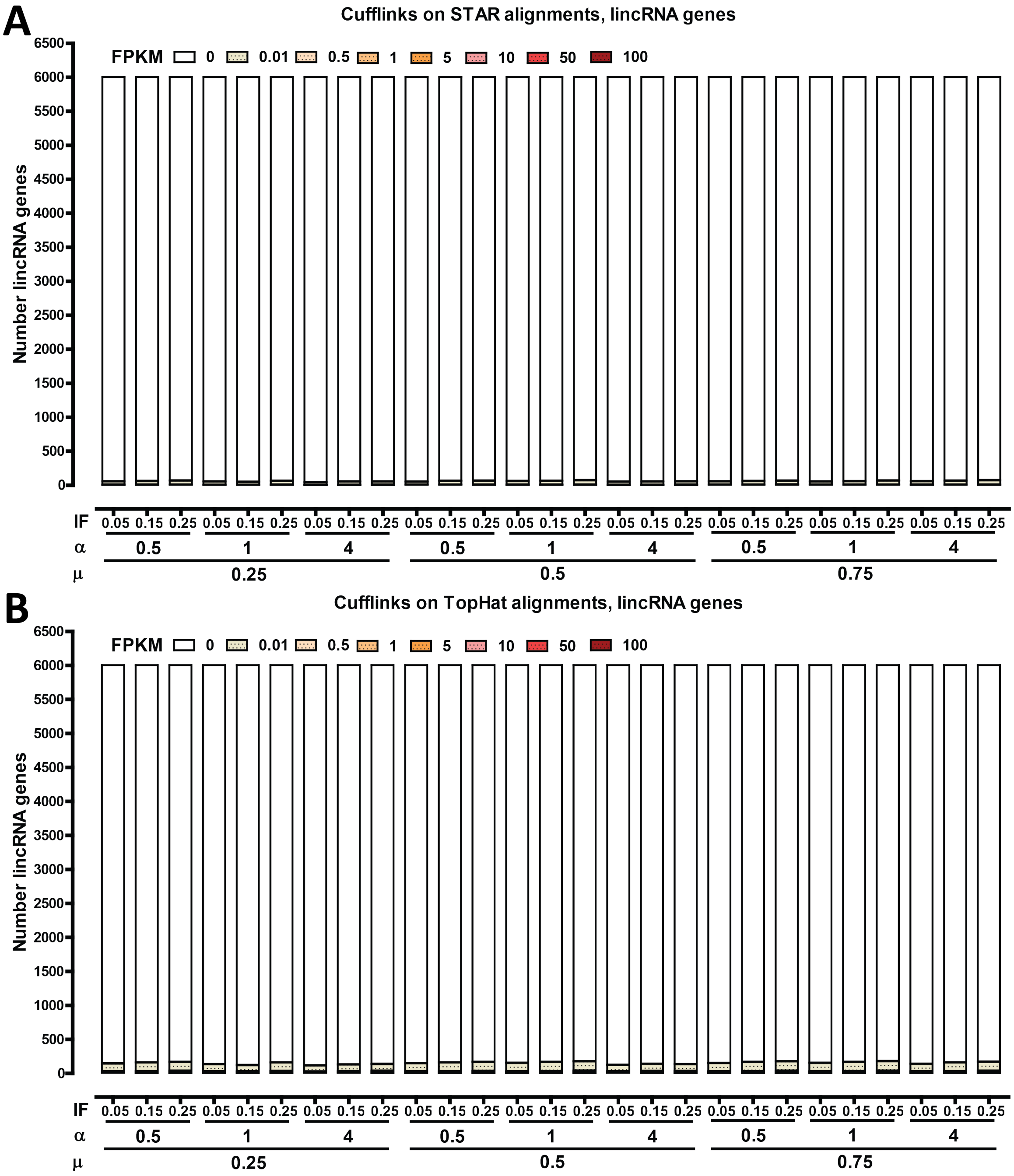
simulated libraries and their fragment support (I use “novel” in quotation marks to indicate that they are false positives). TopHat detected between 2000 and 3000 false positive junctions, while STAR found on average slightly fewer ones, but in a few cases it produced substantially more of them for unknown at present reasons. In both cases there was a positive correlation between the IF parameter and the number of false positive junctions. Figure 2.4 shows the intronic motifs of these junctions, and Figure 2.5 shows how they relate to the annotation, following the convention adopted in the previous chapter. Remarkably, most of the “novel” junctions turned out to be anti-sense to known transcripts and connecting known exons, as indicated by the fact that the dominant intronic motif was CT|AC (which is the antisense to GT|AG, the canonical splice motif). This was not the case only in the anomalous STAR mappings where a substantial number of CT|AC junctions were still present. Both STAR and TopHat found junctions with a CT|GC motif but only TopHat returned GC|AG, GT|AT, AT|AC and splices with other sequence motifs. A large number of CT|AC junctions was not observed in TopHat alignments of real RNA-seq data suggesting that the majority of “novel” junctions seen in the simulation were the result of strand assignment issues in this particular set of alignments, possibly due to the version of the software used. However, antisense junctions can be easily spotted and filtered, thus bringing down the real number of false positives to just a few hundreds, meaning that the majority of splicing complexity observed in real RNA-seq data is not due to computational artifacts.

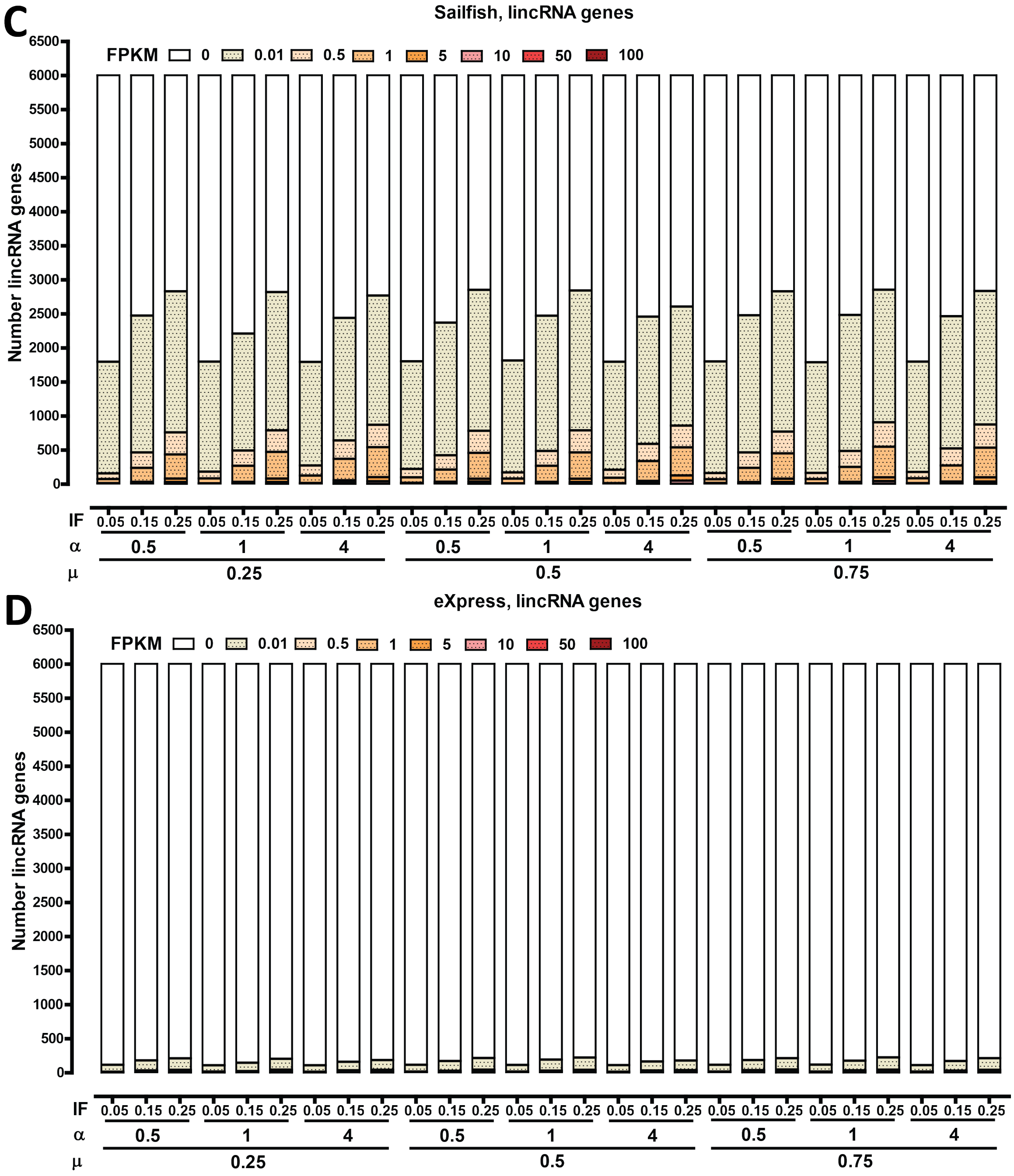
2.3.2 Accuracy of *de novo* transcript assembly

Transcript reconstruction of STAR and TopHat alignments was carried out using Cufflinks. In parallel alignment-free assemblies were generated using Trinity and SOAPdenovo-Trans, then the resulting contigs were mapped back to the genome using BLAT, and converted to GTF file format. As a first assembly evaluation step, all four sets of GTF files were run through Cuffcompare, the GTF comparison module in the Cufflinks suite of programs, against the GENCODE V16 reference, and the fraction of transcripts classified under the different Cuffcompare classes counted. The results are shown in Figure 2.6. Cufflinks produced very similar results on STAR and TopHat alignments, generating between 11,000 and 15,000 fully matched transcripts (Cuffcompare class “=”) depending on the expressed isoform complexity (Figure 2.6A and B). A few notable trends emerged when the fraction of partial assemblies (Cuffcompare class “*c*”) and “new isoforms” (Cuffcompare class “*j*”) assembled were considered – higher values of the α parameter ($\alpha = 4$), i.e. lower isoform complexity, resulted in a relatively small fraction of “new isoforms” (as only annotated transcripts were simulated and the same GENCODE V16 annotation was used as a reference, no “new isoforms” were expected; all such transcripts are therefore false positives), but in the simulations with $\alpha = 0.5$ and $\alpha = 1$ more than 10,000 such isoforms were assembled. Increasing the intronic fraction also had a negative effect on assembly though not as pronounced as the effect of isoform complexity, with the fraction of true positives decreasing slightly and the fraction of partial assemblies and false positives increasing.

Trinity and SOAPdenovo-Trans results were striking in comparison (Figure 2.6 C and D). Trinity actually generated a few hundred more true positive transcripts than Cufflinks, although this is not clearly visible in the figure, which in turn is because of the extremely large number of partial assemblies and false positives it produced – in the hundreds of thousands. The number of such contigs was strongly correlated with the intronic fraction of reads. These results are a combination of assembling each true transcripts into multiple short fragmentary assemblies and of the assembly of many isoforms with retained introns. In contrast, SOAPdenovo-Trans did not assemble almost any “new isoforms”, instead generating a large number of transcripts classified as “intronic”, suggesting it might be dealing better with intronic reads. However, it also assembled very few true transcripts (only ~3,000 on average) and it also generated many partial assemblies, the number of which also correlated strongly with the intronic fraction of reads.

To better understand the assemblies, I carried out a more direct comparison using only the assembled spliced transcripts/contigs using the additional true/false positive and false negative metrics listed in the Methods section (Figure 2.7). This was done in two ways: from the perspective of the assemblies (Figure 2.7A,C,E,G), and from the point of view of the set of expressed transcripts (Figure 2.7B,D,F,H). In the


A
Cufflinks on STAR alignments, lincRNA genes
FPKM
0
0.01
0.5
1
5
10
50
100
Number lincRNA genes
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
IF
0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α
0.5
1
4
0.5
1
4
0.5
1
4
μ
0.25
0.5
0.75
B
Cufflinks on TopHat alignments, lincRNA genes
FPKM
0
0.01
0.5
1
5
10
50
100
Number lincRNA genes
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
IF
0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α
0.5
1
4
0.5
1
4
0.5
1
4
μ
0.25
0.5
0.75


C
Sailfish, lincRNA genes
FPKM ☐ 0 ▨ 0.01 ▨ 0.5 ▨ 1 ▨ 5 ▨ 10 ▨ 50 ▨ 100
Number lincRNA genes
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
IF 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α 0.5 1 4 0.5 1 4 0.5 1 4
μ 0.25 0.5 0.75
D
eXpress, lincRNA genes
FPKM ☐ 0 ▨ 0.01 ▨ 0.5 ▨ 1 ▨ 5 ▨ 10 ▨ 50 ▨ 100
Number lincRNA genes
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
IF 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α 0.5 1 4 0.5 1 4 0.5 1 4
μ 0.25 0.5 0.75

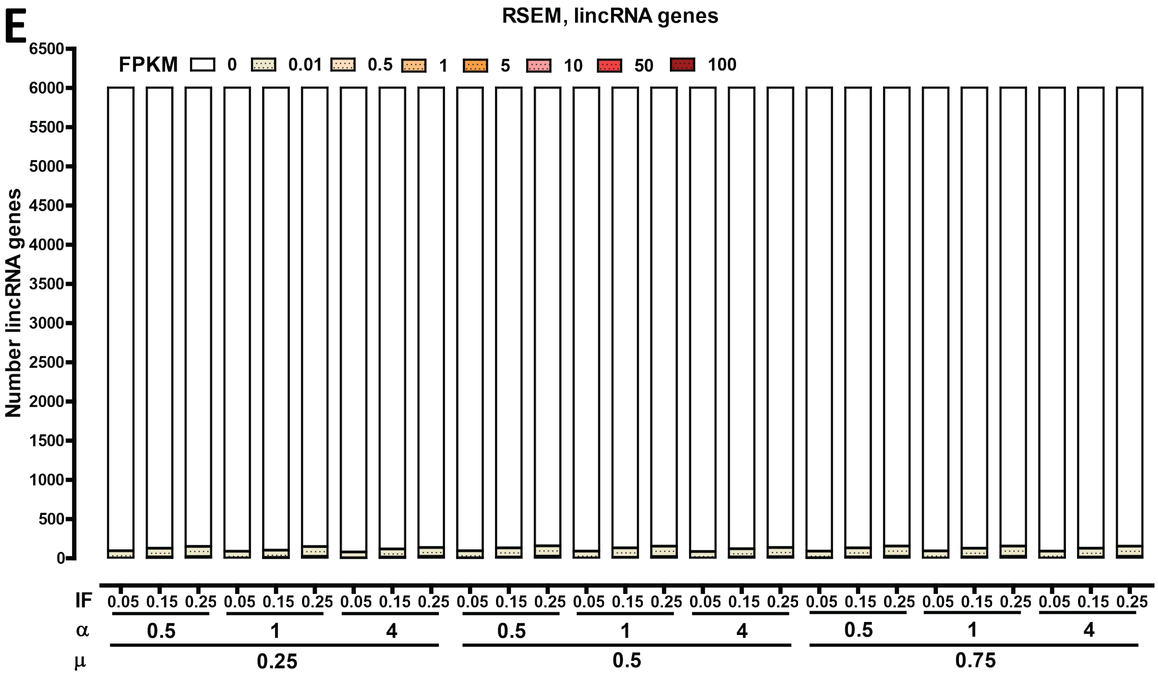


Figure 2.10: Distribution of estimated FPKMs for lincRNA genes. The number of lincRNAs "detected" at different FPKM cutoffs in the output of RNA-seq quantification programs is shown. Note that lincRNAs were not included in the original simulation therefore the true expression values should be zero for all of them.

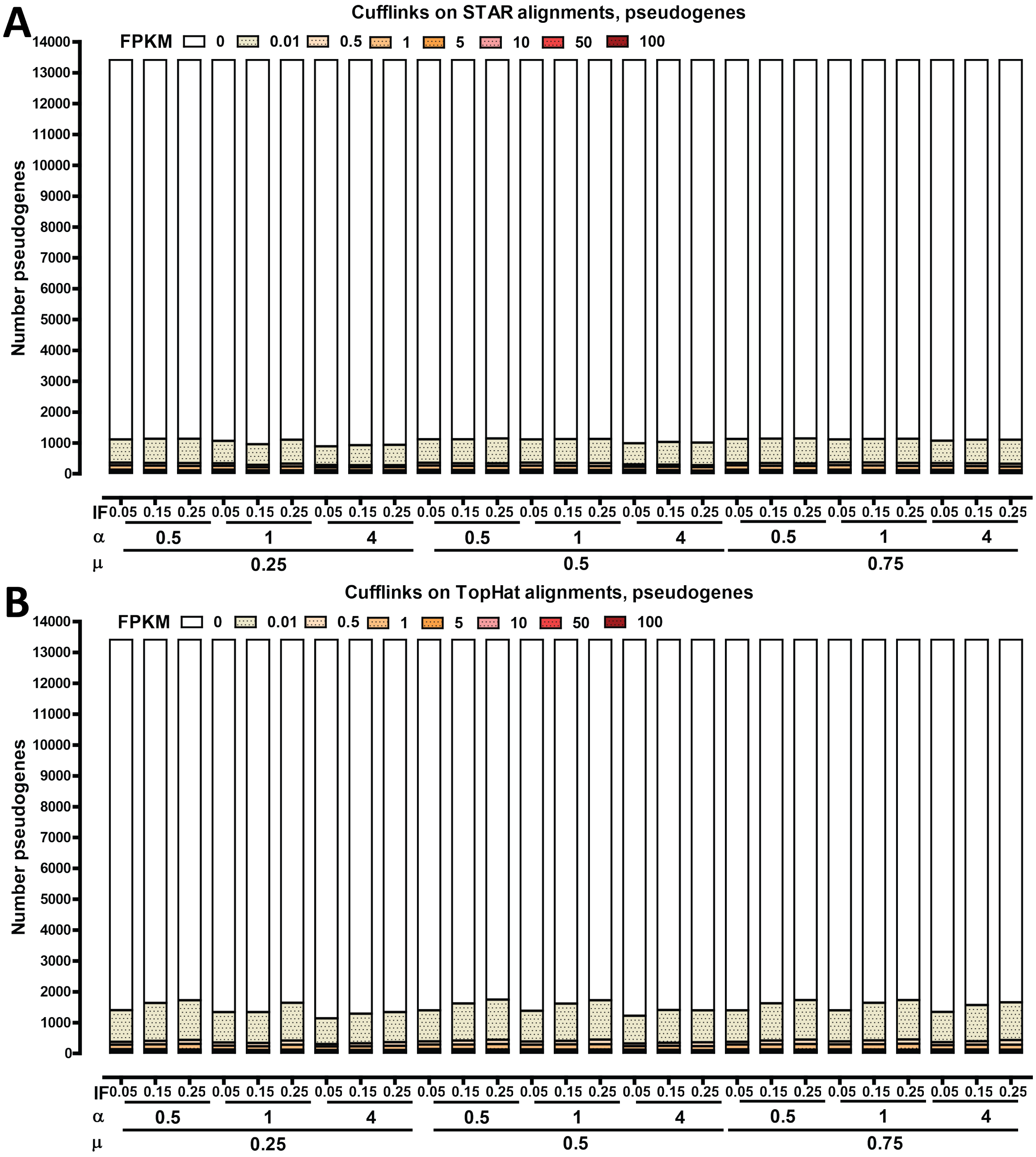
former case, we define a true positive as a transcript that is both expressed in the simulation and assembled (at the level of its intron chain), a partial true positive is a partially assembled expressed transcripts, and a false positive is a transcript, the intron chain of which is incompatible with the annotation. In the annotation-centered comparison, true positives (expressed at ≥1 FPKM and assembled) and false negatives (expressed at ≥1 FPKM but not assembled) transcripts are counted.

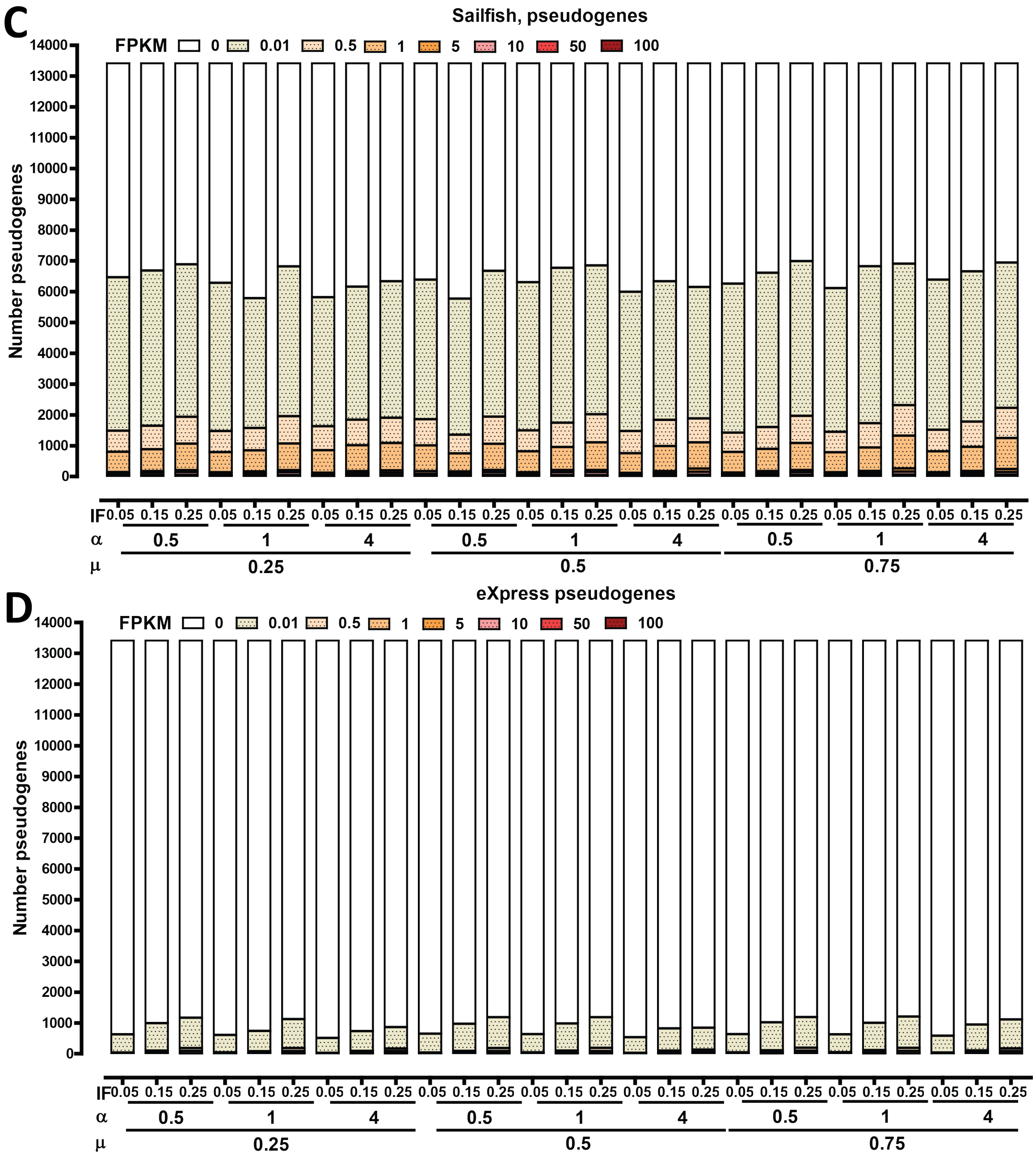
STAR+Cufflinks and TopHat+Cufflinks results were again comparable, with a slight advantage to the TopHat+Cufflinks combination. Once again, the negative effect on the accuracy of the results of isoform complexity was highlighted. At $IF = 0.05$, $\mu = 0.25$, and $\alpha = 4$, i.e high-purity polyA-selection on samples in which almost always only one isoform is expressed, nearly 80% of assembled transcripts were true positives, with ≤10% being false positives (Figure 2.7C), and >80% of expressed transcripts were assembled, with <20% being false negatives (Figure 2.7D). However, when $\mu = 0.75$, and $\alpha = 0.5$, only ∼50% of assembled transcripts were true positives, nearly 40% were false positives, and only ∼35% of expressed transcripts were successfully assembled (with ∼65% remaining as false negatives).

Trinity and SOAPdenovo-Trans results followed the same trend across the parameter space, but were worse in terms of absolute performance. Trinity successfully assembled a higher fraction of the expressed transcripts than Cufflinks did (Figure 2.7F); however, this was at the cost of a much larger fraction of false positives (Figure 2.7E). Notably, this fraction was highly sensitive to the value of the IF parameter. SOAPdenovo-trans was again less sensitive to intronic reads but its performance was very poor in absolute terms (Figure 2.7G and H)).

2.3.3 Accuracy of isoform-level quantification

The accuracy of gene and transcript expression quantification was assessed using the multiple metrics listed in the Methods section. Figure 2.8 shows the Pearson correlation between the estimated FPKMs on the gene level and the true


A
Cufflinks on STAR alignments, pseudogenes
FPKM
0
0.01
0.5
1
5
10
50
100
Number pseudogenes
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
IF 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α 0.5 1 4 0.5 1 4 0.5 1 4
μ 0.25 0.5 0.75
B
Cufflinks on TopHat alignments, pseudogenes
FPKM
0
0.01
0.5
1
5
10
50
100
Number pseudogenes
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
IF 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α 0.5 1 4 0.5 1 4 0.5 1 4
μ 0.25 0.5 0.75


C
Sailfish, pseudogenes
FPKM
0
0.01
0.5
1
5
10
50
100
Number pseudogenes
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
IF 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α 0.5 1 4 0.5 1 4 0.5 1 4
μ 0.25 0.5 0.75
D
eXpress pseudogenes
FPKM
0
0.01
0.5
1
5
10
50
100
Number pseudogenes
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
IF 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25 0.05 0.15 0.25
α 0.5 1 4 0.5 1 4 0.5 1 4
μ 0.25 0.5 0.75

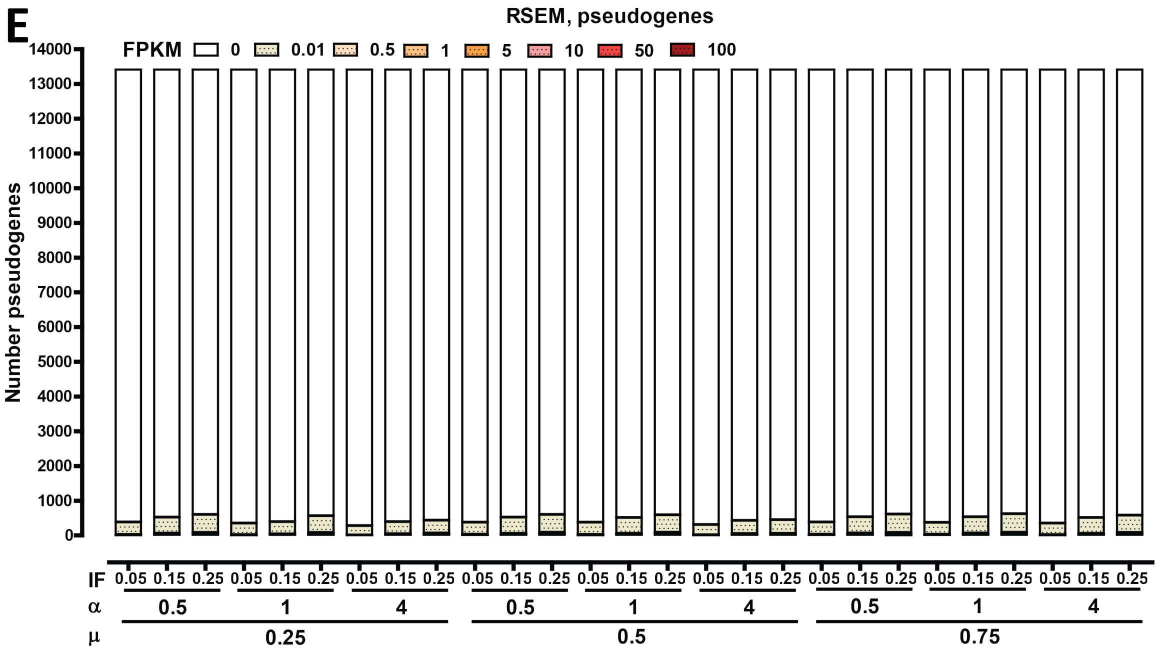


Figure 2.11: Distribution of estimated FPKMs for pseudogenes. The number of pseudogenes "detected" at different FPKM cutoffs in the output of RNA-seq quantification programs is shown. Note that pseudogenes were not included in the original simulation therefore the true expression values should be zero for all of them.

A

	μ	0.25								
	α	0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
MTΘdiff all	Cufflinks STAR	0.52	0.53	0.53	0.46	0.46	0.47	0.18	0.20	0.20
	Cufflinks TopHat	0.55	0.55	0.55	0.48	0.48	0.49	0.21	0.22	0.22
	eXpress	0.48	0.48	0.49	0.40	0.41	0.42	0.14	0.16	0.17
	RSEM	0.48	0.49	0.49	0.40	0.41	0.42	0.13	0.15	0.16
	Sailfish	0.54	0.58	0.61	0.50	0.54	0.58	0.33	0.41	0.47
MTΘdiff 0-1 FPKM	Cufflinks STAR	0.46	0.47	0.47	0.43	0.43	0.44	0.29	0.30	0.30
	Cufflinks TopHat	0.48	0.49	0.49	0.45	0.45	0.46	0.31	0.32	0.32
	eXpress	0.44	0.45	0.45	0.41	0.42	0.43	0.28	0.30	0.30
	RSEM	0.45	0.45	0.46	0.41	0.42	0.42	0.27	0.29	0.29
	Sailfish	0.58	0.62	0.64	0.56	0.59	0.63	0.49	0.56	0.59
MTΘdiff 1-5 FPKM	Cufflinks STAR	0.48	0.49	0.50	0.39	0.39	0.40	0.07	0.07	0.07
	Cufflinks TopHat	0.52	0.53	0.54	0.44	0.44	0.44	0.11	0.11	0.11
	eXpress	0.45	0.46	0.47	0.37	0.38	0.39	0.05	0.07	0.09
	RSEM	0.44	0.45	0.45	0.36	0.37	0.38	0.03	0.05	0.06
	Sailfish	0.49	0.59	0.66	0.45	0.54	0.65	0.30	0.48	0.58
MTΘdiff 5-10 FPKM	Cufflinks STAR	0.53	0.52	0.53	0.42	0.43	0.43	0.07	0.08	0.08
	Cufflinks TopHat	0.56	0.56	0.56	0.46	0.46	0.47	0.10	0.10	0.11
	eXpress	0.52	0.52	0.52	0.42	0.42	0.43	0.05	0.07	0.09
	RSEM	0.50	0.49	0.50	0.39	0.40	0.41	0.03	0.05	0.07
	Sailfish	0.50	0.54	0.59	0.45	0.51	0.56	0.22	0.32	0.41

	μ	0.5								
	α	0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
MTΘdiff all	Cufflinks STAR	0.54	0.56	0.54	0.52	0.53	0.53	0.32	0.32	0.39
	Cufflinks TopHat	0.56	0.58	0.56	0.54	0.55	0.55	0.34	0.34	0.41
	eXpress	0.49	0.52	0.50	0.48	0.48	0.49	0.25	0.26	0.35
	RSEM	0.50	0.52	0.50	0.48	0.49	0.49	0.24	0.25	0.34
	Sailfish	0.54	0.59	0.61	0.54	0.58	0.61	0.40	0.46	0.56
MTΘdiff 0-1 FPKM	Cufflinks STAR	0.47	0.49	0.48	0.47	0.47	0.48	0.35	0.35	0.44
	Cufflinks TopHat	0.49	0.51	0.50	0.48	0.49	0.50	0.37	0.37	0.46
	eXpress	0.45	0.47	0.46	0.44	0.45	0.46	0.33	0.34	0.43
	RSEM	0.46	0.48	0.47	0.45	0.46	0.46	0.32	0.33	0.43
	Sailfish	0.58	0.63	0.63	0.58	0.62	0.63	0.52	0.57	0.64
MTΘdiff 1-5 FPKM	Cufflinks STAR	0.49	0.53	0.50	0.48	0.49	0.50	0.20	0.20	0.40
	Cufflinks TopHat	0.54	0.57	0.54	0.52	0.53	0.54	0.25	0.25	0.44
	eXpress	0.47	0.51	0.48	0.46	0.46	0.46	0.18	0.20	0.40
	RSEM	0.46	0.50	0.47	0.44	0.45	0.46	0.16	0.17	0.38
	Sailfish	0.50	0.61	0.67	0.50	0.60	0.69	0.35	0.51	0.76
MTΘdiff 5-10 FPKM	Cufflinks STAR	0.53	0.57	0.54	0.52	0.51	0.52	0.22	0.22	0.30
	Cufflinks TopHat	0.56	0.60	0.57	0.55	0.55	0.55	0.25	0.25	0.32
	eXpress	0.52	0.56	0.53	0.51	0.52	0.52	0.19	0.20	0.28
	RSEM	0.50	0.55	0.51	0.49	0.49	0.50	0.17	0.18	0.27
	Sailfish	0.50	0.57	0.59	0.50	0.54	0.58	0.31	0.38	0.58

	μ	0.75								
	α	0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
MTΘdiff all	Cufflinks STAR	0.54	0.54	0.54	0.54	0.54	0.54	0.49	0.49	0.49
	Cufflinks TopHat	0.56	0.56	0.57	0.56	0.56	0.57	0.51	0.51	0.51
	eXpress	0.49	0.50	0.50	0.49	0.50	0.50	0.43	0.44	0.45
	RSEM	0.49	0.50	0.50	0.50	0.50	0.51	0.43	0.44	0.45
	Sailfish	0.55	0.58	0.61	0.54	0.58	0.61	0.51	0.56	0.59
MTΘdiff 0-1 FPKM	Cufflinks STAR	0.47	0.48	0.48	0.47	0.48	0.48	0.44	0.45	0.45
	Cufflinks TopHat	0.49	0.50	0.50	0.49	0.49	0.50	0.46	0.46	0.47
	eXpress	0.45	0.46	0.46	0.45	0.46	0.46	0.42	0.43	0.44
	RSEM	0.46	0.46	0.47	0.46	0.46	0.46	0.43	0.43	0.44
	Sailfish	0.58	0.62	0.64	0.58	0.62	0.63	0.57	0.61	0.63
MTΘdiff 1-5 FPKM	Cufflinks STAR	0.49	0.50	0.51	0.50	0.50	0.51	0.43	0.43	0.44
	Cufflinks TopHat	0.54	0.54	0.55	0.54	0.54	0.55	0.47	0.47	0.48
	eXpress	0.47	0.48	0.48	0.47	0.48	0.49	0.40	0.41	0.42
	RSEM	0.46	0.46	0.47	0.46	0.47	0.48	0.39	0.40	0.41
	Sailfish	0.50	0.59	0.67	0.51	0.60	0.69	0.47	0.59	0.68
MTΘdiff 5-10 FPKM	Cufflinks STAR	0.53	0.54	0.54	0.54	0.54	0.54	0.45	0.46	0.46
	Cufflinks TopHat	0.55	0.56	0.58	0.57	0.58	0.58	0.49	0.49	0.49
	eXpress	0.52	0.52	0.53	0.54	0.54	0.54	0.45	0.46	0.46
	RSEM	0.51	0.50	0.52	0.51	0.51	0.51	0.43	0.44	0.44
	Sailfish	0.52	0.54	0.59	0.51	0.55	0.61	0.47	0.51	0.58

0.20
0.40
0.60
0.80
1.00

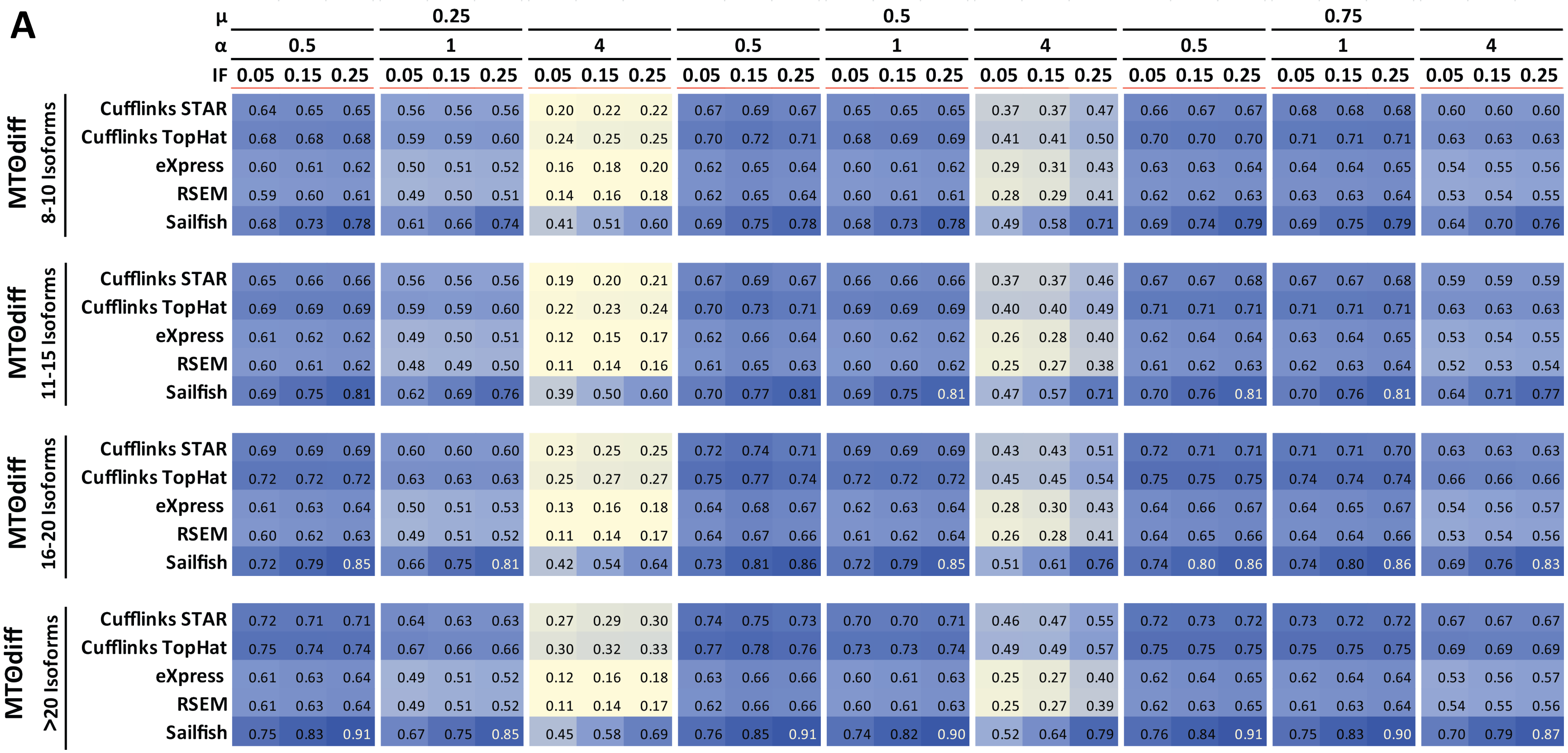
B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
MTΘdiff 10-50 FPKM	Cufflinks STAR	0.55	0.55	0.56	0.47	0.47	0.48	0.09	0.10	0.11	0.57	0.60	0.58	0.55	0.55	0.56	0.30	0.30	0.32	0.57	0.58	0.58	0.58	0.58	0.59	0.50	0.50	0.51
	Cufflinks TopHat	0.58	0.58	0.59	0.50	0.50	0.50	0.11	0.12	0.13	0.61	0.63	0.61	0.58	0.58	0.59	0.32	0.32	0.34	0.60	0.61	0.61	0.61	0.61	0.62	0.53	0.53	0.54
	eXpress	0.53	0.54	0.54	0.42	0.43	0.44	0.04	0.05	0.07	0.56	0.59	0.57	0.53	0.53	0.54	0.20	0.21	0.24	0.56	0.56	0.57	0.57	0.57	0.57	0.47	0.47	0.48
	RSEM	0.53	0.53	0.54	0.42	0.42	0.43	0.02	0.04	0.06	0.55	0.58	0.56	0.53	0.53	0.54	0.19	0.20	0.23	0.55	0.56	0.56	0.56	0.56	0.57	0.46	0.47	0.47
	Sailfish	0.50	0.53	0.57	0.44	0.47	0.52	0.18	0.25	0.33	0.52	0.56	0.57	0.50	0.53	0.56	0.29	0.35	0.42	0.52	0.54	0.58	0.52	0.54	0.58	0.47	0.50	0.54
MTΘdiff 50-100 FPKM	Cufflinks STAR	0.64	0.65	0.64	0.62	0.62	0.62	0.17	0.17	0.19	0.65	0.66	0.65	0.64	0.63	0.63	0.40	0.40	0.40	0.65	0.65	0.65	0.65	0.65	0.65	0.63	0.63	0.62
	Cufflinks TopHat	0.65	0.66	0.66	0.62	0.62	0.62	0.18	0.20	0.20	0.67	0.68	0.68	0.64	0.64	0.64	0.41	0.42	0.41	0.66	0.66	0.67	0.66	0.66	0.66	0.64	0.64	0.63
	eXpress	0.51	0.51	0.52	0.40	0.40	0.41	0.03	0.05	0.07	0.53	0.55	0.54	0.50	0.51	0.52	0.17	0.18	0.21	0.53	0.54	0.54	0.53	0.53	0.54	0.44	0.45	0.46
	RSEM	0.52	0.52	0.53	0.40	0.40	0.41	0.02	0.04	0.06	0.54	0.55	0.55	0.51	0.51	0.52	0.17	0.18	0.20	0.54	0.54	0.54	0.53	0.54	0.54	0.45	0.45	0.46
	Sailfish	0.51	0.53	0.57	0.44	0.48	0.52	0.19	0.25	0.33	0.52	0.55	0.58	0.51	0.54	0.57	0.28	0.34	0.39	0.52	0.54	0.58	0.51	0.54	0.58	0.47	0.50	0.54
MTΘdiff 100-500 FPKM	Cufflinks STAR	0.74	0.75	0.77	0.65	0.66	0.67	0.20	0.22	0.22	0.76	0.77	0.77	0.75	0.77	0.78	0.41	0.43	0.45	0.75	0.77	0.78	0.75	0.75	0.76	0.69	0.70	0.72
	Cufflinks TopHat	0.73	0.74	0.76	0.65	0.66	0.66	0.20	0.21	0.23	0.75	0.76	0.76	0.75	0.76	0.77	0.40	0.42	0.44	0.75	0.77	0.77	0.75	0.76	0.76	0.67	0.68	0.70
	eXpress	0.49	0.49	0.50	0.39	0.40	0.40	0.02	0.04	0.07	0.51	0.52	0.52	0.49	0.49	0.50	0.17	0.19	0.22	0.51	0.51	0.51	0.51	0.51	0.52	0.43	0.44	0.44
	RSEM	0.51	0.51	0.52	0.40	0.40	0.41	0.02	0.04	0.06	0.53	0.54	0.53	0.51	0.51	0.52	0.18	0.19	0.23	0.52	0.52	0.53	0.52	0.53	0.53	0.44	0.45	0.45
	Sailfish	0.53	0.56	0.59	0.47	0.51	0.55	0.23	0.29	0.38	0.54	0.56	0.60	0.54	0.56	0.61	0.33	0.38	0.41	0.55	0.57	0.61	0.54	0.56	0.58	0.50	0.53	0.56
MTΘdiff >500 FPKM	Cufflinks STAR	0.82	0.85	0.89	0.69	0.72	0.76	0.18	0.22	0.28	0.84	0.88	0.89	0.79	0.82	0.84	0.38	0.40	0.44	0.88	0.89	0.90	0.86	0.88	0.90	0.72	0.73	0.75
	Cufflinks TopHat	0.81	0.83	0.88	0.64	0.68	0.73	0.15	0.19	0.25	0.81	0.85	0.88	0.76	0.78	0.82	0.36	0.38	0.43	0.85	0.86	0.89	0.83	0.85	0.88	0.68	0.70	0.72
	eXpress	0.44	0.45	0.45	0.36	0.37	0.38	0.02	0.05	0.08	0.44	0.47	0.46	0.43	0.44	0.45	0.15	0.17	0.20	0.45	0.46	0.47	0.44	0.45	0.46	0.37	0.38	0.40
	RSEM	0.53	0.54	0.54	0.44	0.44	0.45	0.02	0.04	0.07	0.54	0.58	0.56	0.52	0.52	0.53	0.18	0.20	0.23	0.56	0.56	0.57	0.55	0.55	0.55	0.44	0.45	0.46
	Sailfish	0.54	0.56	0.60	0.47	0.53	0.57	0.27	0.34	0.45	0.52	0.56	0.58	0.51	0.54	0.58	0.34	0.42	0.44	0.53	0.56	0.60	0.53	0.56	0.58	0.49	0.52	0.55

Figure 2.12: Mean Total Θ difference as function of gene expression level. The Mean Total Θ difference between true transcript-level FPKMs and the output of different quantification programs was calculated for each gene and then averaged, at varying values of the μ, α and IF parameters and at varying gene-level expression cutoffs. The averaging was performed only relative to the number of genes considered in each set.

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
MTΘdiff all	Cufflinks STAR	0.52	0.53	0.53	0.46	0.46	0.47	0.18	0.20	0.20	0.54	0.56	0.54	0.52	0.53	0.53	0.32	0.32	0.39	0.54	0.54	0.54	0.54	0.54	0.54	0.49	0.49	0.49
	Cufflinks TopHat	0.55	0.55	0.55	0.48	0.48	0.49	0.21	0.22	0.22	0.56	0.58	0.56	0.54	0.55	0.55	0.34	0.34	0.41	0.56	0.56	0.57	0.56	0.56	0.57	0.51	0.51	0.51
	eXpress	0.48	0.48	0.49	0.40	0.41	0.42	0.14	0.16	0.17	0.49	0.52	0.50	0.48	0.48	0.49	0.25	0.26	0.35	0.49	0.50	0.50	0.49	0.50	0.50	0.43	0.44	0.45
	RSEM	0.48	0.49	0.49	0.40	0.41	0.42	0.13	0.15	0.16	0.50	0.52	0.50	0.48	0.49	0.49	0.24	0.25	0.34	0.49	0.50	0.50	0.50	0.50	0.51	0.43	0.44	0.45
	Sailfish	0.54	0.58	0.61	0.50	0.54	0.58	0.33	0.41	0.47	0.54	0.59	0.61	0.54	0.58	0.61	0.40	0.46	0.56	0.55	0.58	0.61	0.54	0.58	0.61	0.51	0.56	0.59
MTΘdiff 2 Isoforms	Cufflinks STAR	0.41	0.41	0.42	0.38	0.38	0.39	0.17	0.18	0.18	0.42	0.44	0.43	0.41	0.41	0.42	0.27	0.27	0.34	0.41	0.42	0.42	0.42	0.42	0.42	0.39	0.39	0.40
	Cufflinks TopHat	0.42	0.43	0.43	0.39	0.39	0.40	0.18	0.20	0.19	0.43	0.46	0.44	0.42	0.42	0.43	0.28	0.29	0.35	0.42	0.43	0.43	0.43	0.43	0.43	0.40	0.40	0.41
	eXpress	0.39	0.39	0.39	0.36	0.36	0.36	0.18	0.19	0.19	0.39	0.42	0.40	0.39	0.39	0.39	0.26	0.27	0.33	0.39	0.39	0.39	0.40	0.39	0.39	0.37	0.37	0.38
	RSEM	0.41	0.40	0.40	0.37	0.37	0.37	0.16	0.17	0.17	0.41	0.44	0.41	0.40	0.40	0.40	0.25	0.26	0.32	0.40	0.41	0.40	0.41	0.41	0.41	0.38	0.38	0.39
	Sailfish	0.43	0.44	0.46	0.41	0.43	0.44	0.28	0.34	0.37	0.43	0.47	0.46	0.43	0.44	0.45	0.34	0.37	0.44	0.43	0.44	0.45	0.43	0.44	0.46	0.42	0.43	0.46
MTΘdiff 3-4 Isoforms	Cufflinks STAR	0.55	0.55	0.56	0.49	0.49	0.49	0.19	0.20	0.21	0.56	0.59	0.57	0.55	0.55	0.56	0.32	0.33	0.41	0.57	0.57	0.57	0.56	0.57	0.57	0.51	0.52	0.52
	Cufflinks TopHat	0.57	0.58	0.58	0.51	0.51	0.52	0.22	0.23	0.23	0.58	0.61	0.60	0.57	0.57	0.58	0.34	0.35	0.43	0.59	0.59	0.60	0.59	0.59	0.60	0.54	0.54	0.55
	eXpress	0.53	0.53	0.53	0.46	0.47	0.48	0.18	0.20	0.22	0.54	0.56	0.54	0.52	0.53	0.53	0.29	0.31	0.39	0.54	0.54	0.54	0.54	0.55	0.54	0.49	0.49	0.50
	RSEM	0.54	0.54	0.55	0.47	0.47	0.48	0.17	0.19	0.20	0.55	0.57	0.55	0.54	0.54	0.54	0.29	0.30	0.39	0.55	0.55	0.56	0.55	0.55	0.56	0.50	0.50	0.51
	Sailfish	0.55	0.58	0.60	0.52	0.55	0.58	0.35	0.42	0.48	0.55	0.60	0.61	0.56	0.58	0.60	0.42	0.48	0.57	0.56	0.58	0.61	0.56	0.59	0.61	0.53	0.56	0.59
MTΘdiff 5-7 Isoforms	Cufflinks STAR	0.63	0.63	0.64	0.55	0.55	0.55	0.18	0.18	0.19	0.65	0.67	0.65	0.63	0.63	0.63	0.35	0.35	0.43	0.65	0.65	0.66	0.65	0.65	0.65	0.58	0.58	0.58
	Cufflinks TopHat	0.66	0.66	0.67	0.57	0.57	0.58	0.20	0.21	0.21	0.67	0.70	0.68	0.66	0.66	0.66	0.37	0.38	0.45	0.67	0.68	0.68	0.67	0.68	0.68	0.60	0.60	0.61
	eXpress	0.59	0.59	0.60	0.51	0.51	0.52	0.16	0.18	0.19	0.61	0.64	0.62	0.59	0.60	0.60	0.30	0.31	0.41	0.61	0.62	0.63	0.61	0.62	0.62	0.54	0.55	0.55
	RSEM	0.60	0.60	0.60	0.50	0.51	0.52	0.14	0.16	0.17	0.62	0.64	0.62	0.60	0.60	0.61	0.29	0.30	0.39	0.62	0.62	0.63	0.62	0.62	0.62	0.54	0.54	0.55
	Sailfish	0.64	0.68	0.71	0.59	0.64	0.68	0.37	0.46	0.52	0.64	0.69	0.71	0.64	0.68	0.71	0.46	0.53	0.64	0.64	0.68	0.71	0.64	0.68	0.71	0.61	0.65	0.69

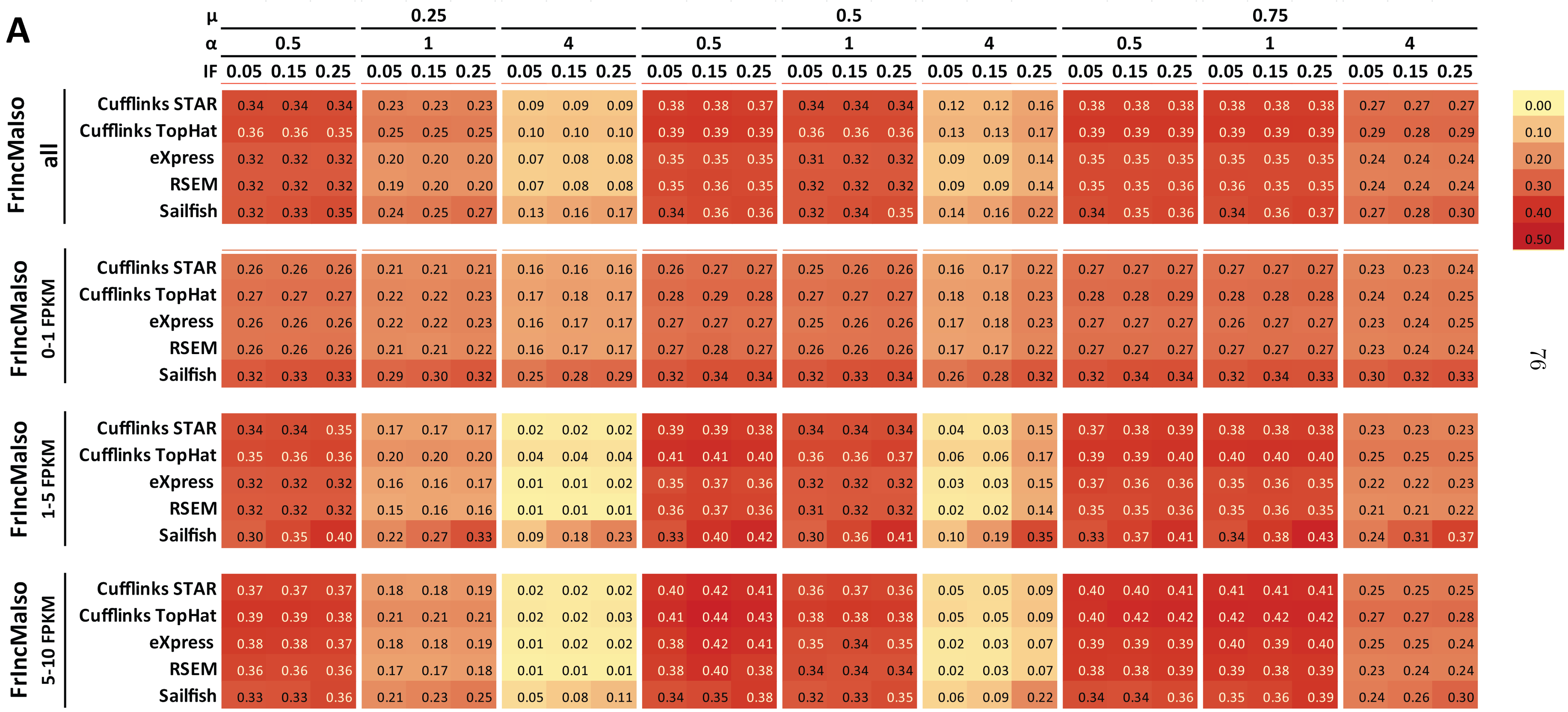
0.20
0.40
0.60
0.80
1.00



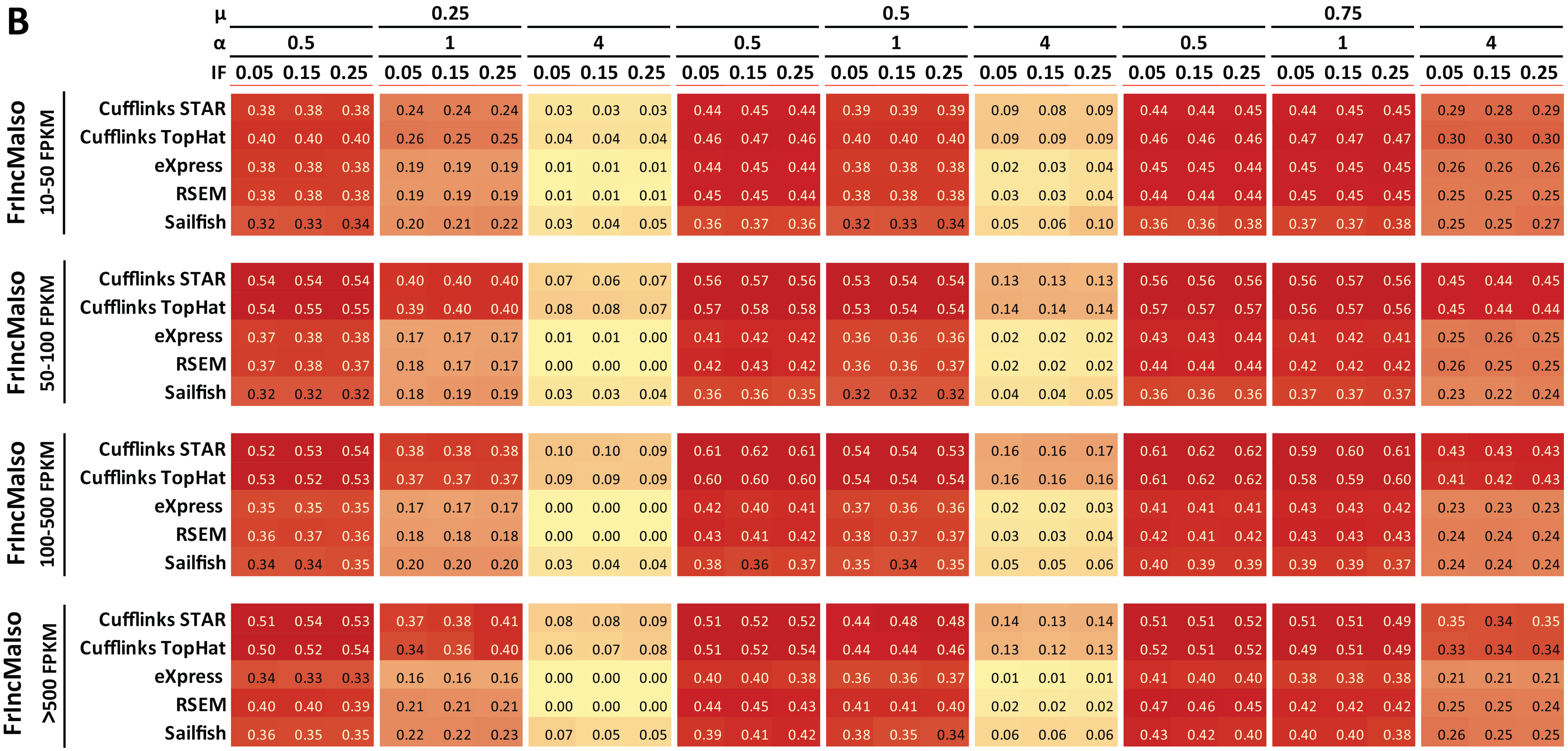
A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
MTΘdiff 8-10 Isoforms	Cufflinks STAR	0.64	0.65	0.65	0.56	0.56	0.56	0.20	0.22	0.22	0.67	0.69	0.67	0.65	0.65	0.65	0.37	0.37	0.47	0.66	0.67	0.67	0.68	0.68	0.68	0.60	0.60	0.60
	Cufflinks TopHat	0.68	0.68	0.68	0.59	0.59	0.60	0.24	0.25	0.25	0.70	0.72	0.71	0.68	0.69	0.69	0.41	0.41	0.50	0.70	0.70	0.70	0.71	0.71	0.71	0.63	0.63	0.63
	eXpress	0.60	0.61	0.62	0.50	0.51	0.52	0.16	0.18	0.20	0.62	0.65	0.64	0.60	0.61	0.62	0.29	0.31	0.43	0.63	0.63	0.64	0.64	0.64	0.65	0.54	0.55	0.56
	RSEM	0.59	0.60	0.61	0.49	0.50	0.51	0.14	0.16	0.18	0.62	0.65	0.64	0.60	0.61	0.61	0.28	0.29	0.41	0.62	0.62	0.63	0.63	0.63	0.64	0.53	0.54	0.55
	Sailfish	0.68	0.73	0.78	0.61	0.66	0.74	0.41	0.51	0.60	0.69	0.75	0.78	0.68	0.73	0.78	0.49	0.58	0.71	0.69	0.74	0.79	0.69	0.75	0.79	0.64	0.70	0.76
MTΘdiff 11-15 Isoforms	Cufflinks STAR	0.65	0.66	0.66	0.56	0.56	0.56	0.19	0.20	0.21	0.67	0.69	0.67	0.66	0.66	0.66	0.37	0.37	0.46	0.67	0.67	0.68	0.67	0.67	0.68	0.59	0.59	0.59
	Cufflinks TopHat	0.69	0.69	0.69	0.59	0.59	0.60	0.22	0.23	0.24	0.70	0.73	0.71	0.69	0.69	0.69	0.40	0.40	0.49	0.71	0.71	0.71	0.71	0.71	0.71	0.63	0.63	0.63
	eXpress	0.61	0.62	0.62	0.49	0.50	0.51	0.12	0.15	0.17	0.62	0.66	0.64	0.60	0.62	0.62	0.26	0.28	0.40	0.62	0.64	0.64	0.63	0.64	0.65	0.53	0.54	0.55
	RSEM	0.60	0.61	0.62	0.48	0.49	0.50	0.11	0.14	0.16	0.61	0.65	0.63	0.60	0.60	0.62	0.25	0.27	0.38	0.61	0.62	0.63	0.62	0.63	0.64	0.52	0.53	0.54
	Sailfish	0.69	0.75	0.81	0.62	0.69	0.76	0.39	0.50	0.60	0.70	0.77	0.81	0.69	0.75	0.81	0.47	0.57	0.71	0.70	0.76	0.81	0.70	0.76	0.81	0.64	0.71	0.77
MTΘdiff 16-20 Isoforms	Cufflinks STAR	0.69	0.69	0.69	0.60	0.60	0.60	0.23	0.25	0.25	0.72	0.74	0.71	0.69	0.69	0.69	0.43	0.43	0.51	0.72	0.71	0.71	0.71	0.71	0.70	0.63	0.63	0.63
	Cufflinks TopHat	0.72	0.72	0.72	0.63	0.63	0.63	0.25	0.27	0.27	0.75	0.77	0.74	0.72	0.72	0.72	0.45	0.45	0.54	0.75	0.75	0.75	0.74	0.74	0.74	0.66	0.66	0.66
	eXpress	0.61	0.63	0.64	0.50	0.51	0.53	0.13	0.16	0.18	0.64	0.68	0.67	0.62	0.63	0.64	0.28	0.30	0.43	0.64	0.66	0.67	0.64	0.65	0.67	0.54	0.56	0.57
	RSEM	0.60	0.62	0.63	0.49	0.51	0.52	0.11	0.14	0.17	0.64	0.67	0.66	0.61	0.62	0.64	0.26	0.28	0.41	0.64	0.65	0.66	0.64	0.64	0.66	0.53	0.54	0.56
	Sailfish	0.72	0.79	0.85	0.66	0.75	0.81	0.42	0.54	0.64	0.73	0.81	0.86	0.72	0.79	0.85	0.51	0.61	0.76	0.74	0.80	0.86	0.74	0.80	0.86	0.69	0.76	0.83
MTΘdiff >20 Isoforms	Cufflinks STAR	0.72	0.71	0.71	0.64	0.63	0.63	0.27	0.29	0.30	0.74	0.75	0.73	0.70	0.70	0.71	0.46	0.47	0.55	0.72	0.73	0.72	0.73	0.72	0.72	0.67	0.67	0.67
	Cufflinks TopHat	0.75	0.74	0.74	0.67	0.66	0.66	0.30	0.32	0.33	0.77	0.78	0.76	0.73	0.73	0.74	0.49	0.49	0.57	0.75	0.75	0.75	0.75	0.75	0.75	0.69	0.69	0.69
	eXpress	0.61	0.63	0.64	0.49	0.51	0.52	0.12	0.16	0.18	0.63	0.66	0.66	0.60	0.61	0.63	0.25	0.27	0.40	0.62	0.64	0.65	0.62	0.64	0.64	0.53	0.56	0.57
	RSEM	0.61	0.63	0.64	0.49	0.51	0.52	0.11	0.14	0.17	0.62	0.66	0.66	0.60	0.61	0.63	0.25	0.27	0.39	0.62	0.63	0.65	0.61	0.63	0.64	0.54	0.55	0.56
	Sailfish	0.75	0.83	0.91	0.67	0.75	0.85	0.45	0.58	0.69	0.76	0.85	0.91	0.74	0.82	0.90	0.52	0.64	0.79	0.76	0.84	0.91	0.75	0.83	0.90	0.70	0.79	0.87

Figure 2.13: Mean Total Θ difference as function of annotated isoform complexity. The Mean Total Θ difference between true transcript-level FPKMs and the output of different quantification programs was calculated for each gene and then averaged, at varying values of the μ, α and IF parameters and for genes with different number of isoforms annotated in GENCODE V16. The averaging was performed only relative to the number of genes considered in each set.



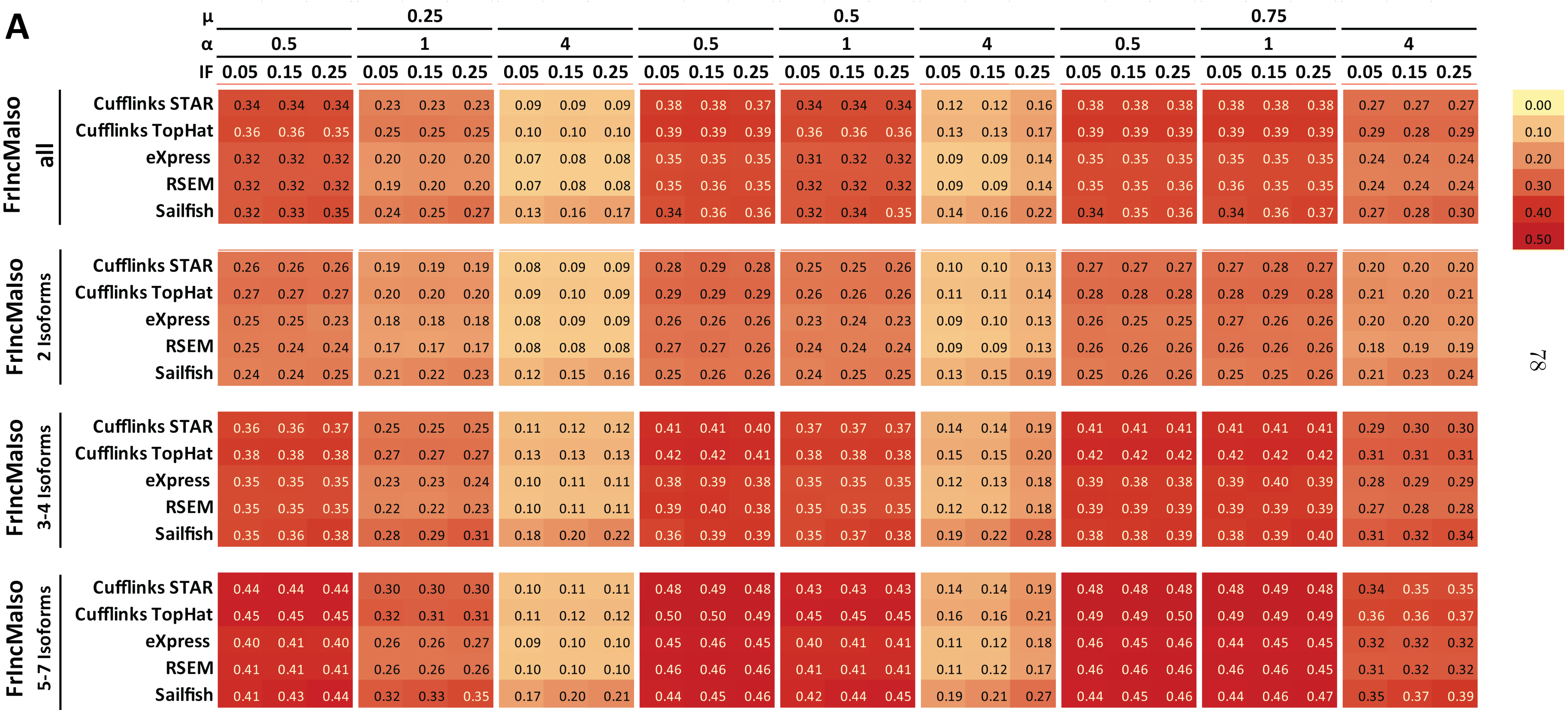
	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FrIncMalso all	Cufflinks STAR	0.34	0.34	0.34	0.23	0.23	0.23	0.09	0.09	0.09	0.38	0.38	0.37	0.34	0.34	0.34	0.12	0.12	0.16	0.38	0.38	0.38	0.38	0.38	0.38	0.27	0.27	0.27
	Cufflinks TopHat	0.36	0.36	0.35	0.25	0.25	0.25	0.10	0.10	0.10	0.39	0.39	0.39	0.36	0.36	0.36	0.13	0.13	0.17	0.39	0.39	0.39	0.39	0.39	0.39	0.29	0.28	0.29
	eXpress	0.32	0.32	0.32	0.20	0.20	0.20	0.07	0.08	0.08	0.35	0.35	0.35	0.31	0.32	0.32	0.09	0.09	0.14	0.35	0.35	0.35	0.35	0.35	0.35	0.24	0.24	0.24
	RSEM	0.32	0.32	0.32	0.19	0.20	0.20	0.07	0.08	0.08	0.35	0.36	0.35	0.32	0.32	0.32	0.09	0.09	0.14	0.35	0.35	0.36	0.36	0.35	0.35	0.24	0.24	0.24
	Sailfish	0.32	0.33	0.35	0.24	0.25	0.27	0.13	0.16	0.17	0.34	0.36	0.36	0.32	0.34	0.35	0.14	0.16	0.22	0.34	0.35	0.36	0.34	0.36	0.37	0.27	0.28	0.30
FrIncMalso 0-1 FPKM	Cufflinks STAR	0.26	0.26	0.26	0.21	0.21	0.21	0.16	0.16	0.16	0.26	0.27	0.27	0.25	0.26	0.26	0.16	0.17	0.22	0.27	0.27	0.27	0.27	0.27	0.27	0.23	0.23	0.24
	Cufflinks TopHat	0.27	0.27	0.27	0.22	0.22	0.23	0.17	0.18	0.17	0.28	0.29	0.28	0.27	0.27	0.27	0.18	0.18	0.23	0.28	0.28	0.29	0.28	0.28	0.28	0.24	0.24	0.25
	eXpress	0.26	0.26	0.26	0.22	0.22	0.23	0.16	0.17	0.17	0.27	0.27	0.27	0.25	0.26	0.26	0.17	0.18	0.23	0.27	0.27	0.27	0.26	0.27	0.27	0.23	0.24	0.25
	RSEM	0.26	0.26	0.26	0.21	0.21	0.22	0.16	0.17	0.17	0.27	0.28	0.27	0.26	0.26	0.26	0.17	0.17	0.22	0.27	0.27	0.27	0.27	0.27	0.27	0.23	0.24	0.24
	Sailfish	0.32	0.33	0.33	0.29	0.30	0.32	0.25	0.28	0.29	0.32	0.34	0.34	0.32	0.33	0.34	0.26	0.28	0.32	0.32	0.34	0.34	0.32	0.34	0.33	0.30	0.32	0.33
FrIncMalso 1-5 FPKM	Cufflinks STAR	0.34	0.34	0.35	0.17	0.17	0.17	0.02	0.02	0.02	0.39	0.39	0.38	0.34	0.34	0.34	0.04	0.03	0.15	0.37	0.38	0.39	0.38	0.38	0.38	0.23	0.23	0.23
	Cufflinks TopHat	0.35	0.36	0.36	0.20	0.20	0.20	0.04	0.04	0.04	0.41	0.41	0.40	0.36	0.36	0.37	0.06	0.06	0.17	0.39	0.39	0.40	0.40	0.40	0.40	0.25	0.25	0.25
	eXpress	0.32	0.32	0.32	0.16	0.16	0.17	0.01	0.01	0.02	0.35	0.37	0.36	0.32	0.32	0.32	0.03	0.03	0.15	0.37	0.36	0.36	0.35	0.36	0.35	0.22	0.22	0.23
	RSEM	0.32	0.32	0.32	0.15	0.16	0.16	0.01	0.01	0.01	0.36	0.37	0.36	0.31	0.32	0.32	0.02	0.02	0.14	0.35	0.35	0.36	0.35	0.36	0.35	0.21	0.21	0.22
	Sailfish	0.30	0.35	0.40	0.22	0.27	0.33	0.09	0.18	0.23	0.33	0.40	0.42	0.30	0.36	0.41	0.10	0.19	0.35	0.33	0.37	0.41	0.34	0.38	0.43	0.24	0.31	0.37
FrIncMalso 5-10 FPKM	Cufflinks STAR	0.37	0.37	0.37	0.18	0.18	0.19	0.02	0.02	0.02	0.40	0.42	0.41	0.36	0.37	0.36	0.05	0.05	0.09	0.40	0.40	0.41	0.41	0.41	0.41	0.25	0.25	0.25
	Cufflinks TopHat	0.39	0.39	0.38	0.21	0.21	0.21	0.02	0.02	0.03	0.41	0.44	0.43	0.38	0.38	0.38	0.05	0.05	0.09	0.40	0.42	0.42	0.42	0.42	0.42	0.27	0.27	0.28
	eXpress	0.38	0.38	0.37	0.18	0.18	0.19	0.01	0.02	0.02	0.38	0.42	0.41	0.35	0.34	0.35	0.02	0.03	0.07	0.39	0.39	0.39	0.40	0.39	0.40	0.25	0.25	0.24
	RSEM	0.36	0.36	0.36	0.17	0.17	0.18	0.01	0.01	0.01	0.38	0.40	0.38	0.34	0.34	0.34	0.02	0.03	0.07	0.38	0.38	0.39	0.39	0.38	0.39	0.23	0.24	0.24
	Sailfish	0.33	0.33	0.36	0.21	0.23	0.25	0.05	0.08	0.11	0.34	0.35	0.38	0.32	0.33	0.35	0.06	0.09	0.22	0.34	0.34	0.36	0.35	0.36	0.39	0.24	0.26	0.30



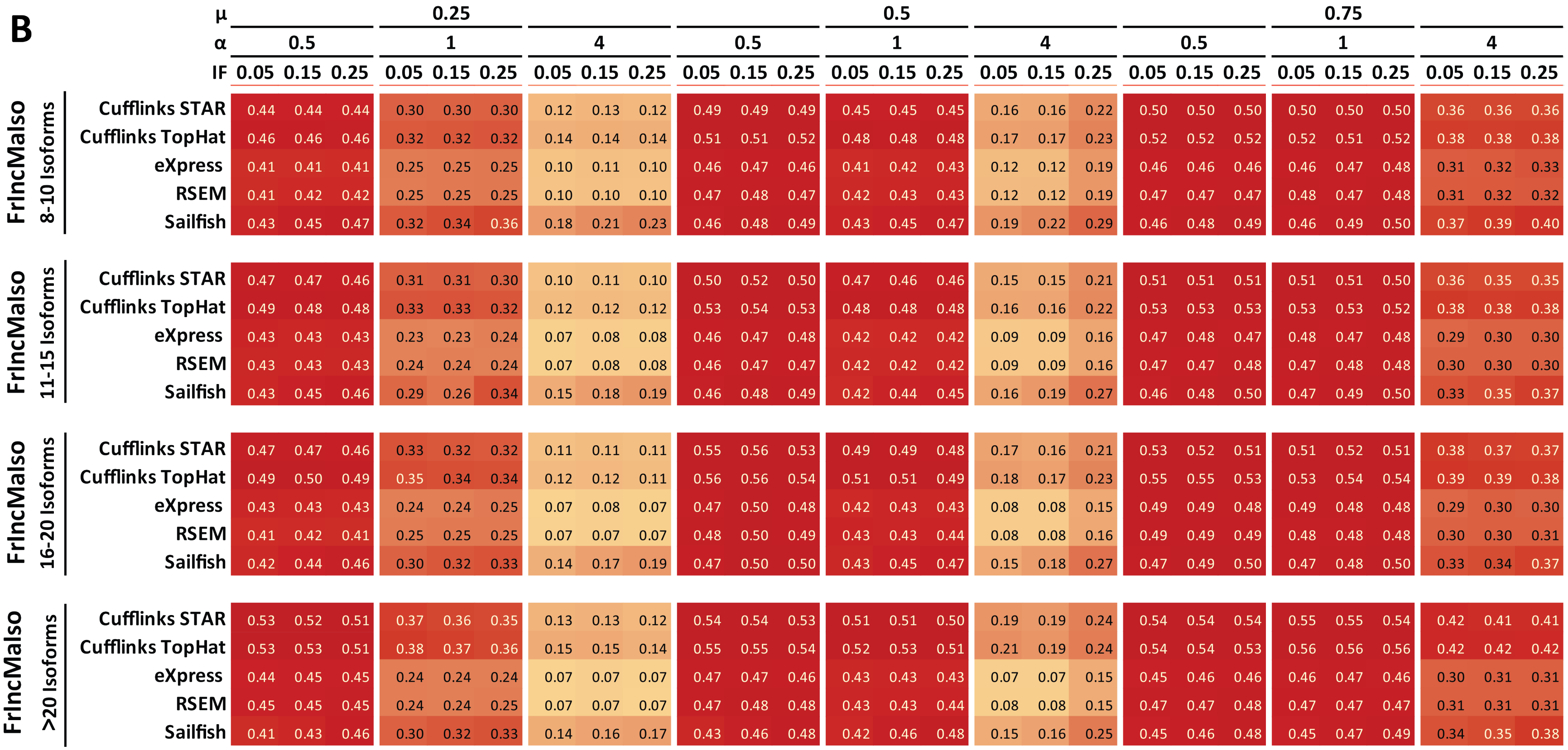
	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FrIncMaIso 10-50 FPKM	Cufflinks STAR	0.38	0.38	0.38	0.24	0.24	0.24	0.03	0.03	0.03	0.44	0.45	0.44	0.39	0.39	0.39	0.09	0.08	0.09	0.44	0.44	0.45	0.44	0.45	0.45	0.29	0.28	0.29
	Cufflinks TopHat	0.40	0.40	0.40	0.26	0.25	0.25	0.04	0.04	0.04	0.46	0.47	0.46	0.40	0.40	0.40	0.09	0.09	0.09	0.46	0.46	0.46	0.47	0.47	0.47	0.30	0.30	0.30
	eXpress	0.38	0.38	0.38	0.19	0.19	0.19	0.01	0.01	0.01	0.44	0.45	0.44	0.38	0.38	0.38	0.02	0.03	0.04	0.45	0.45	0.44	0.45	0.45	0.45	0.26	0.26	0.26
	RSEM	0.38	0.38	0.38	0.19	0.19	0.19	0.01	0.01	0.01	0.45	0.45	0.44	0.38	0.38	0.38	0.03	0.03	0.04	0.44	0.44	0.44	0.45	0.45	0.45	0.25	0.25	0.25
	Sailfish	0.32	0.33	0.34	0.20	0.21	0.22	0.03	0.04	0.05	0.36	0.37	0.36	0.32	0.33	0.34	0.05	0.06	0.10	0.36	0.36	0.38	0.37	0.37	0.38	0.25	0.25	0.27
FrIncMaIso 50-100 FPKM	Cufflinks STAR	0.54	0.54	0.54	0.40	0.40	0.40	0.07	0.06	0.07	0.56	0.57	0.56	0.53	0.54	0.54	0.13	0.13	0.13	0.56	0.56	0.56	0.56	0.57	0.56	0.45	0.44	0.45
	Cufflinks TopHat	0.54	0.55	0.55	0.39	0.40	0.40	0.08	0.08	0.07	0.57	0.58	0.58	0.53	0.54	0.54	0.14	0.14	0.14	0.57	0.57	0.57	0.56	0.57	0.56	0.45	0.44	0.44
	eXpress	0.37	0.38	0.38	0.17	0.17	0.17	0.01	0.01	0.00	0.41	0.42	0.42	0.36	0.36	0.36	0.02	0.02	0.02	0.43	0.43	0.44	0.41	0.42	0.41	0.25	0.26	0.25
	RSEM	0.37	0.38	0.37	0.18	0.17	0.17	0.00	0.00	0.00	0.42	0.43	0.42	0.36	0.36	0.37	0.02	0.02	0.02	0.44	0.44	0.44	0.42	0.42	0.42	0.26	0.25	0.25
	Sailfish	0.32	0.32	0.32	0.18	0.19	0.19	0.03	0.03	0.04	0.36	0.36	0.35	0.32	0.32	0.32	0.04	0.04	0.05	0.36	0.36	0.36	0.37	0.37	0.37	0.23	0.22	0.24
FrIncMaIso 100-500 FPKM	Cufflinks STAR	0.52	0.53	0.54	0.38	0.38	0.38	0.10	0.10	0.09	0.61	0.62	0.61	0.54	0.54	0.53	0.16	0.16	0.17	0.61	0.62	0.62	0.59	0.60	0.61	0.43	0.43	0.43
	Cufflinks TopHat	0.53	0.52	0.53	0.37	0.37	0.37	0.09	0.09	0.09	0.60	0.60	0.60	0.54	0.54	0.54	0.16	0.16	0.16	0.61	0.62	0.62	0.58	0.59	0.60	0.41	0.42	0.43
	eXpress	0.35	0.35	0.35	0.17	0.17	0.17	0.00	0.00	0.00	0.42	0.40	0.41	0.37	0.36	0.36	0.02	0.02	0.03	0.41	0.41	0.41	0.43	0.43	0.42	0.23	0.23	0.23
	RSEM	0.36	0.37	0.36	0.18	0.18	0.18	0.00	0.00	0.00	0.43	0.41	0.42	0.38	0.37	0.37	0.03	0.03	0.04	0.42	0.41	0.42	0.43	0.43	0.43	0.24	0.24	0.24
	Sailfish	0.34	0.34	0.35	0.20	0.20	0.20	0.03	0.04	0.04	0.38	0.36	0.37	0.35	0.34	0.35	0.05	0.05	0.06	0.40	0.39	0.39	0.39	0.39	0.37	0.24	0.24	0.24
FrIncMaIso >500 FPKM	Cufflinks STAR	0.51	0.54	0.53	0.37	0.38	0.41	0.08	0.08	0.09	0.51	0.52	0.52	0.44	0.48	0.48	0.14	0.13	0.14	0.51	0.51	0.52	0.51	0.51	0.49	0.35	0.34	0.35
	Cufflinks TopHat	0.50	0.52	0.54	0.34	0.36	0.40	0.06	0.07	0.08	0.51	0.52	0.54	0.44	0.44	0.46	0.13	0.12	0.13	0.52	0.51	0.52	0.49	0.51	0.49	0.33	0.34	0.34
	eXpress	0.34	0.33	0.33	0.16	0.16	0.16	0.00	0.00	0.00	0.40	0.40	0.38	0.36	0.36	0.37	0.01	0.01	0.01	0.41	0.40	0.40	0.38	0.38	0.38	0.21	0.21	0.21
	RSEM	0.40	0.40	0.39	0.21	0.21	0.21	0.00	0.00	0.00	0.44	0.45	0.43	0.41	0.41	0.40	0.02	0.02	0.02	0.47	0.46	0.45	0.42	0.42	0.42	0.25	0.25	0.24
	Sailfish	0.36	0.35	0.35	0.22	0.22	0.23	0.07	0.05	0.05	0.39	0.41	0.42	0.38	0.35	0.34	0.06	0.06	0.06	0.43	0.42	0.40	0.40	0.40	0.38	0.26	0.25	0.25

Figure 2.14: Fraction of genes with an incorrectly assigned major isoform as a function of gene expression levels. The fraction of genes with an incorrectly assigned major isoform ("FrIncMaIso") in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and at varying gene-level expression cutoffs.

A



	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FrlncMaIso all	Cufflinks STAR	0.34	0.34	0.34	0.23	0.23	0.23	0.09	0.09	0.09	0.38	0.38	0.37	0.34	0.34	0.34	0.12	0.12	0.16	0.38	0.38	0.38	0.38	0.38	0.38	0.27	0.27	0.27
	Cufflinks TopHat	0.36	0.36	0.35	0.25	0.25	0.25	0.10	0.10	0.10	0.39	0.39	0.39	0.36	0.36	0.36	0.13	0.13	0.17	0.39	0.39	0.39	0.39	0.39	0.39	0.29	0.28	0.29
	eXpress	0.32	0.32	0.32	0.20	0.20	0.20	0.07	0.08	0.08	0.35	0.35	0.35	0.31	0.32	0.32	0.09	0.09	0.14	0.35	0.35	0.35	0.35	0.35	0.35	0.24	0.24	0.24
	RSEM	0.32	0.32	0.32	0.19	0.20	0.20	0.07	0.08	0.08	0.35	0.36	0.35	0.32	0.32	0.32	0.09	0.09	0.14	0.35	0.35	0.36	0.36	0.35	0.35	0.24	0.24	0.24
	Sailfish	0.32	0.33	0.35	0.24	0.25	0.27	0.13	0.16	0.17	0.34	0.36	0.36	0.32	0.34	0.35	0.14	0.16	0.22	0.34	0.35	0.36	0.34	0.36	0.37	0.27	0.28	0.30
FrlncMaIso 2 Isoforms	Cufflinks STAR	0.26	0.26	0.26	0.19	0.19	0.19	0.08	0.09	0.09	0.28	0.29	0.28	0.25	0.25	0.26	0.10	0.10	0.13	0.27	0.27	0.27	0.27	0.28	0.27	0.20	0.20	0.20
	Cufflinks TopHat	0.27	0.27	0.27	0.20	0.20	0.20	0.09	0.10	0.09	0.29	0.29	0.29	0.26	0.26	0.26	0.11	0.11	0.14	0.28	0.28	0.28	0.28	0.29	0.28	0.21	0.20	0.21
	eXpress	0.25	0.25	0.23	0.18	0.18	0.18	0.08	0.09	0.09	0.26	0.26	0.26	0.23	0.24	0.23	0.09	0.10	0.13	0.26	0.25	0.25	0.27	0.26	0.26	0.20	0.20	0.20
	RSEM	0.25	0.24	0.24	0.17	0.17	0.17	0.08	0.08	0.08	0.27	0.27	0.26	0.24	0.24	0.24	0.09	0.09	0.13	0.26	0.26	0.26	0.26	0.26	0.26	0.18	0.19	0.19
	Sailfish	0.24	0.24	0.25	0.21	0.22	0.23	0.12	0.15	0.16	0.25	0.26	0.26	0.24	0.25	0.25	0.13	0.15	0.19	0.25	0.26	0.26	0.25	0.25	0.26	0.21	0.23	0.24
FrlncMaIso 3-4 Isoforms	Cufflinks STAR	0.36	0.36	0.37	0.25	0.25	0.25	0.11	0.12	0.12	0.41	0.41	0.40	0.37	0.37	0.37	0.14	0.14	0.19	0.41	0.41	0.41	0.41	0.41	0.41	0.29	0.30	0.30
	Cufflinks TopHat	0.38	0.38	0.38	0.27	0.27	0.27	0.13	0.13	0.13	0.42	0.42	0.41	0.38	0.38	0.38	0.15	0.15	0.20	0.42	0.42	0.42	0.42	0.42	0.42	0.31	0.31	0.31
	eXpress	0.35	0.35	0.35	0.23	0.23	0.24	0.10	0.11	0.11	0.38	0.39	0.38	0.35	0.35	0.35	0.12	0.13	0.18	0.39	0.38	0.38	0.39	0.40	0.39	0.28	0.29	0.29
	RSEM	0.35	0.35	0.35	0.22	0.22	0.23	0.10	0.11	0.11	0.39	0.40	0.38	0.35	0.35	0.35	0.12	0.12	0.18	0.39	0.39	0.39	0.39	0.39	0.39	0.27	0.28	0.28
	Sailfish	0.35	0.36	0.38	0.28	0.29	0.31	0.18	0.20	0.22	0.36	0.39	0.39	0.35	0.37	0.38	0.19	0.22	0.28	0.38	0.38	0.39	0.38	0.39	0.40	0.31	0.32	0.34
FrlncMaIso 5-7 Isoforms	Cufflinks STAR	0.44	0.44	0.44	0.30	0.30	0.30	0.10	0.11	0.11	0.48	0.49	0.48	0.43	0.43	0.43	0.14	0.14	0.19	0.48	0.48	0.48	0.48	0.49	0.48	0.34	0.35	0.35
	Cufflinks TopHat	0.45	0.45	0.45	0.32	0.31	0.31	0.11	0.12	0.12	0.50	0.50	0.49	0.45	0.45	0.45	0.16	0.16	0.21	0.49	0.49	0.50	0.49	0.49	0.49	0.36	0.36	0.37
	eXpress	0.40	0.41	0.40	0.26	0.26	0.27	0.09	0.10	0.10	0.45	0.46	0.45	0.40	0.41	0.41	0.11	0.12	0.18	0.46	0.45	0.46	0.44	0.45	0.45	0.32	0.32	0.32
	RSEM	0.41	0.41	0.41	0.26	0.26	0.26	0.10	0.10	0.10	0.46	0.46	0.46	0.41	0.41	0.41	0.11	0.12	0.17	0.46	0.46	0.46	0.46	0.46	0.45	0.31	0.32	0.32
	Sailfish	0.41	0.43	0.44	0.32	0.33	0.35	0.17	0.20	0.21	0.44	0.45	0.46	0.42	0.44	0.45	0.19	0.21	0.27	0.44	0.45	0.46	0.44	0.46	0.47	0.35	0.37	0.39



B

		μ	0.25									0.5									0.75								
		α	0.5			1			4			0.5			1			4			0.5			1			4		
		IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FrIncMaIso	8-10 Isoforms	Cufflinks STAR	0.44	0.44	0.44	0.30	0.30	0.30	0.12	0.13	0.12	0.49	0.49	0.49	0.45	0.45	0.45	0.16	0.16	0.22	0.50	0.50	0.50	0.50	0.50	0.50	0.36	0.36	0.36
		Cufflinks TopHat	0.46	0.46	0.46	0.32	0.32	0.32	0.14	0.14	0.14	0.51	0.51	0.52	0.48	0.48	0.48	0.17	0.17	0.23	0.52	0.52	0.52	0.52	0.51	0.52	0.38	0.38	0.38
		eXpress	0.41	0.41	0.41	0.25	0.25	0.25	0.10	0.11	0.10	0.46	0.47	0.46	0.41	0.42	0.43	0.12	0.12	0.19	0.46	0.46	0.46	0.46	0.47	0.48	0.31	0.32	0.33
		RSEM	0.41	0.42	0.42	0.25	0.25	0.25	0.10	0.10	0.10	0.47	0.48	0.47	0.42	0.43	0.43	0.12	0.12	0.19	0.47	0.47	0.47	0.48	0.47	0.48	0.31	0.32	0.32
		Sailfish	0.43	0.45	0.47	0.32	0.34	0.36	0.18	0.21	0.23	0.46	0.48	0.49	0.43	0.45	0.47	0.19	0.22	0.29	0.46	0.48	0.49	0.46	0.49	0.50	0.37	0.39	0.40
FrIncMaIso	11-15 Isoforms	Cufflinks STAR	0.47	0.47	0.46	0.31	0.31	0.30	0.10	0.11	0.10	0.50	0.52	0.50	0.47	0.46	0.46	0.15	0.15	0.21	0.51	0.51	0.51	0.51	0.51	0.50	0.36	0.35	0.35
		Cufflinks TopHat	0.49	0.48	0.48	0.33	0.33	0.32	0.12	0.12	0.12	0.53	0.54	0.53	0.48	0.48	0.48	0.16	0.16	0.22	0.53	0.53	0.53	0.53	0.53	0.52	0.38	0.38	0.38
		eXpress	0.43	0.43	0.43	0.23	0.23	0.24	0.07	0.08	0.08	0.46	0.47	0.48	0.42	0.42	0.42	0.09	0.09	0.16	0.47	0.48	0.47	0.48	0.47	0.48	0.29	0.30	0.30
		RSEM	0.43	0.43	0.43	0.24	0.24	0.24	0.07	0.08	0.08	0.46	0.47	0.47	0.42	0.42	0.42	0.09	0.09	0.16	0.47	0.47	0.48	0.47	0.48	0.48	0.30	0.30	0.30
		Sailfish	0.43	0.45	0.46	0.29	0.26	0.34	0.15	0.18	0.19	0.46	0.48	0.49	0.42	0.44	0.45	0.16	0.19	0.27	0.46	0.48	0.50	0.47	0.49	0.50	0.33	0.35	0.37
FrIncMaIso	16-20 Isoforms	Cufflinks STAR	0.47	0.47	0.46	0.33	0.32	0.32	0.11	0.11	0.11	0.55	0.56	0.53	0.49	0.49	0.48	0.17	0.16	0.21	0.53	0.52	0.51	0.51	0.52	0.51	0.38	0.37	0.37
		Cufflinks TopHat	0.49	0.50	0.49	0.35	0.34	0.34	0.12	0.12	0.11	0.56	0.56	0.54	0.51	0.51	0.49	0.18	0.17	0.23	0.55	0.55	0.53	0.53	0.54	0.54	0.39	0.39	0.38
		eXpress	0.43	0.43	0.43	0.24	0.24	0.25	0.07	0.08	0.07	0.47	0.50	0.48	0.42	0.43	0.43	0.08	0.08	0.15	0.49	0.49	0.48	0.49	0.48	0.48	0.29	0.30	0.30
		RSEM	0.41	0.42	0.41	0.25	0.25	0.25	0.07	0.07	0.07	0.48	0.50	0.49	0.43	0.43	0.44	0.08	0.08	0.16	0.49	0.49	0.49	0.48	0.48	0.48	0.30	0.30	0.31
		Sailfish	0.42	0.44	0.46	0.30	0.32	0.33	0.14	0.17	0.19	0.47	0.50	0.50	0.43	0.45	0.47	0.15	0.18	0.27	0.47	0.49	0.50	0.47	0.48	0.50	0.33	0.34	0.37
FrIncMaIso	>20 Isoforms	Cufflinks STAR	0.53	0.52	0.51	0.37	0.36	0.35	0.13	0.13	0.12	0.54	0.54	0.53	0.51	0.51	0.50	0.19	0.19	0.24	0.54	0.54	0.54	0.55	0.55	0.54	0.42	0.41	0.41
		Cufflinks TopHat	0.53	0.53	0.51	0.38	0.37	0.36	0.15	0.15	0.14	0.55	0.55	0.54	0.52	0.53	0.51	0.21	0.19	0.24	0.54	0.54	0.53	0.56	0.56	0.56	0.42	0.42	0.42
		eXpress	0.44	0.45	0.45	0.24	0.24	0.24	0.07	0.07	0.07	0.47	0.47	0.46	0.43	0.43	0.43	0.07	0.07	0.15	0.45	0.46	0.46	0.46	0.47	0.46	0.30	0.31	0.31
		RSEM	0.45	0.45	0.45	0.24	0.24	0.25	0.07	0.07	0.07	0.47	0.48	0.48	0.43	0.43	0.44	0.08	0.08	0.15	0.47	0.47	0.48	0.47	0.47	0.47	0.31	0.31	0.31
		Sailfish	0.41	0.43	0.46	0.30	0.32	0.33	0.14	0.16	0.17	0.43	0.46	0.48	0.42	0.46	0.48	0.15	0.16	0.25	0.45	0.46	0.48	0.45	0.47	0.49	0.34	0.35	0.38

Figure 2.15: Fraction of genes with an incorrectly assigned major isoform as a function of annotated isoform complexity. The fraction of genes with an incorrectly assigned major isoform (“FrIncMaIso”) in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and for genes with different number of isoforms annotated in GENCODE V16.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Genes all	Cufflinks STAR	0.06	0.06	0.07	0.08	0.08	0.10	0.12	0.13	0.15	0.06	0.06	0.07	0.06	0.06	0.07	0.10	0.12	0.13	0.05	0.06	0.06	0.06	0.06	0.06	0.07	0.08	0.09
	Cufflinks TopHat	0.08	0.08	0.09	0.10	0.10	0.12	0.14	0.16	0.18	0.07	0.07	0.08	0.08	0.08	0.09	0.13	0.14	0.15	0.07	0.07	0.08	0.07	0.08	0.08	0.09	0.10	0.11
	eXpress	0.06	0.07	0.09	0.08	0.09	0.12	0.09	0.12	0.17	0.06	0.07	0.08	0.06	0.08	0.09	0.08	0.11	0.15	0.06	0.07	0.08	0.06	0.07	0.08	0.07	0.08	0.11
	RSEM	0.03	0.04	0.06	0.04	0.05	0.08	0.04	0.07	0.11	0.03	0.04	0.05	0.03	0.04	0.06	0.04	0.06	0.10	0.03	0.04	0.05	0.03	0.04	0.05	0.04	0.05	0.07
	Sailfish	0.24	0.28	0.33	0.29	0.35	0.41	0.35	0.46	0.55	0.23	0.26	0.31	0.24	0.28	0.33	0.33	0.43	0.48	0.23	0.27	0.31	0.22	0.26	0.31	0.27	0.34	0.39
FN Genes 0-1 FPKM	Cufflinks STAR	0.10	0.11	0.11	0.13	0.13	0.15	0.17	0.19	0.20	0.10	0.10	0.10	0.11	0.11	0.11	0.16	0.17	0.17	0.09	0.10	0.10	0.10	0.10	0.10	0.12	0.13	0.13
	Cufflinks TopHat	0.12	0.12	0.12	0.15	0.15	0.16	0.20	0.21	0.22	0.11	0.11	0.12	0.12	0.12	0.13	0.18	0.19	0.19	0.11	0.11	0.12	0.11	0.11	0.11	0.13	0.14	0.15
	eXpress	0.15	0.16	0.17	0.18	0.18	0.21	0.22	0.25	0.28	0.14	0.14	0.16	0.15	0.16	0.17	0.20	0.23	0.23	0.13	0.15	0.16	0.14	0.15	0.16	0.17	0.18	0.20
	RSEM	0.08	0.09	0.10	0.10	0.11	0.13	0.11	0.15	0.17	0.08	0.08	0.10	0.09	0.10	0.10	0.11	0.14	0.15	0.08	0.09	0.09	0.08	0.09	0.10	0.09	0.11	0.12
	Sailfish	0.28	0.32	0.34	0.35	0.39	0.42	0.49	0.56	0.60	0.27	0.30	0.33	0.29	0.32	0.34	0.43	0.49	0.51	0.26	0.31	0.33	0.27	0.31	0.33	0.33	0.37	0.39
FN Genes 1-5 FPKM	Cufflinks STAR	0.04	0.04	0.04	0.05	0.05	0.05	0.08	0.08	0.09	0.03	0.03	0.04	0.03	0.04	0.04	0.06	0.06	0.07	0.04	0.04	0.04	0.04	0.04	0.03	0.05	0.05	0.06
	Cufflinks TopHat	0.06	0.06	0.07	0.09	0.07	0.09	0.12	0.12	0.13	0.05	0.05	0.06	0.06	0.06	0.06	0.10	0.10	0.11	0.05	0.06	0.06	0.06	0.06	0.06	0.07	0.08	0.09
	eXpress	0.04	0.05	0.07	0.06	0.08	0.10	0.07	0.10	0.15	0.04	0.05	0.07	0.04	0.06	0.07	0.06	0.09	0.13	0.04	0.05	0.06	0.04	0.05	0.07	0.05	0.07	0.09
	RSEM	0.03	0.04	0.05	0.03	0.05	0.07	0.03	0.06	0.10	0.02	0.03	0.05	0.03	0.04	0.05	0.03	0.06	0.09	0.03	0.03	0.05	0.02	0.03	0.05	0.04	0.04	0.06
	Sailfish	0.21	0.27	0.31	0.28	0.38	0.41	0.38	0.52	0.60	0.21	0.24	0.28	0.21	0.27	0.32	0.33	0.46	0.52	0.19	0.24	0.28	0.19	0.24	0.28	0.25	0.34	0.38
FN Genes 5-10 FPKM	Cufflinks STAR	0.03	0.03	0.04	0.04	0.04	0.04	0.07	0.08	0.09	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.05	0.06	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04
	Cufflinks TopHat	0.04	0.05	0.05	0.06	0.07	0.07	0.10	0.11	0.12	0.04	0.05	0.05	0.05	0.05	0.05	0.08	0.09	0.09	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06
	eXpress	0.04	0.05	0.06	0.06	0.06	0.09	0.07	0.10	0.14	0.04	0.04	0.06	0.04	0.05	0.07	0.05	0.07	0.11	0.04	0.05	0.06	0.04	0.04	0.06	0.05	0.06	0.08
	RSEM	0.03	0.03	0.05	0.03	0.03	0.06	0.03	0.06	0.09	0.02	0.02	0.04	0.02	0.03	0.04	0.02	0.04	0.07	0.02	0.02	0.04	0.02	0.03	0.04	0.02	0.03	0.05
	Sailfish	0.20	0.24	0.28	0.25	0.33	0.38	0.31	0.44	0.52	0.17	0.22	0.27	0.19	0.23	0.28	0.30	0.40	0.47	0.18	0.23	0.27	0.17	0.22	0.28	0.23	0.30	0.34

0.00
0.10
0.20
0.30
0.40

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Genes 10-50 FPKM	Cufflinks STAR	0.04	0.04	0.04	0.05	0.05	0.06	0.09	0.11	0.13	0.03	0.04	0.04	0.04	0.04	0.04	0.08	0.10	0.11	0.03	0.04	0.04	0.03	0.03	0.04	0.04	0.05	0.05
	Cufflinks TopHat	0.06	0.06	0.06	0.07	0.07	0.09	0.12	0.13	0.15	0.05	0.06	0.06	0.06	0.06	0.06	0.10	0.12	0.13	0.05	0.06	0.06	0.05	0.05	0.06	0.06	0.07	0.08
	eXpress	0.03	0.04	0.06	0.04	0.05	0.07	0.04	0.07	0.11	0.03	0.03	0.05	0.04	0.04	0.06	0.04	0.06	0.09	0.03	0.04	0.06	0.03	0.03	0.05	0.04	0.05	0.07
	RSEM	0.02	0.02	0.04	0.02	0.02	0.05	0.02	0.04	0.09	0.02	0.02	0.03	0.02	0.02	0.04	0.02	0.04	0.07	0.01	0.02	0.04	0.02	0.02	0.03	0.02	0.02	0.05
	Sailfish	0.21	0.26	0.31	0.25	0.32	0.39	0.27	0.39	0.50	0.19	0.24	0.29	0.21	0.25	0.32	0.26	0.38	0.43	0.21	0.26	0.30	0.19	0.24	0.27	0.24	0.31	0.37
FN Genes 50-100 FPKM	Cufflinks STAR	0.05	0.06	0.07	0.11	0.10	0.13	0.16	0.19	0.24	0.06	0.06	0.08	0.06	0.06	0.07	0.16	0.17	0.21	0.05	0.05	0.06	0.05	0.05	0.06	0.09	0.09	0.10
	Cufflinks TopHat	0.07	0.07	0.08	0.12	0.11	0.14	0.17	0.20	0.23	0.08	0.08	0.09	0.07	0.07	0.07	0.16	0.19	0.22	0.06	0.06	0.07	0.07	0.07	0.07	0.11	0.12	0.12
	eXpress	0.02	0.03	0.05	0.02	0.04	0.07	0.02	0.05	0.11	0.02	0.02	0.04	0.02	0.03	0.05	0.02	0.04	0.09	0.02	0.03	0.06	0.02	0.03	0.05	0.02	0.04	0.07
	RSEM	0.01	0.01	0.03	0.01	0.02	0.06	0.01	0.03	0.09	0.01	0.01	0.03	0.01	0.01	0.04	0.01	0.03	0.08	0.01	0.02	0.04	0.01	0.02	0.03	0.01	0.02	0.05
	Sailfish	0.25	0.30	0.37	0.28	0.32	0.44	0.27	0.38	0.50	0.24	0.28	0.35	0.25	0.29	0.36	0.27	0.39	0.47	0.23	0.29	0.35	0.24	0.28	0.34	0.27	0.35	0.41
FN Genes 100-500 FPKM	Cufflinks STAR	0.09	0.10	0.13	0.13	0.15	0.17	0.16	0.21	0.26	0.07	0.08	0.09	0.10	0.12	0.16	0.14	0.19	0.23	0.08	0.09	0.12	0.08	0.08	0.10	0.11	0.13	0.17
	Cufflinks TopHat	0.10	0.11	0.12	0.14	0.15	0.18	0.17	0.22	0.27	0.08	0.09	0.11	0.11	0.13	0.16	0.17	0.20	0.24	0.10	0.10	0.11	0.09	0.10	0.11	0.13	0.15	0.17
	eXpress	0.02	0.02	0.05	0.02	0.03	0.07	0.02	0.04	0.10	0.02	0.02	0.05	0.02	0.02	0.05	0.02	0.03	0.10	0.02	0.02	0.04	0.01	0.02	0.04	0.02	0.03	0.06
	RSEM	0.01	0.01	0.04	0.01	0.02	0.06	0.01	0.04	0.10	0.01	0.02	0.04	0.01	0.02	0.04	0.01	0.03	0.09	0.01	0.01	0.04	0.01	0.01	0.04	0.01	0.02	0.05
	Sailfish	0.27	0.31	0.39	0.33	0.36	0.46	0.31	0.40	0.53	0.27	0.28	0.36	0.29	0.30	0.40	0.32	0.41	0.46	0.28	0.31	0.38	0.25	0.28	0.33	0.31	0.36	0.43
FN Genes >500 FPKM	Cufflinks STAR	0.11	0.17	0.16	0.13	0.17	0.25	0.17	0.28	0.32	0.09	0.16	0.19	0.08	0.14	0.18	0.15	0.25	0.30	0.06	0.13	0.15	0.08	0.14	0.15	0.14	0.22	0.24
	Cufflinks TopHat	0.09	0.16	0.19	0.11	0.16	0.24	0.14	0.25	0.28	0.09	0.15	0.20	0.08	0.16	0.18	0.13	0.24	0.29	0.05	0.12	0.17	0.10	0.14	0.18	0.14	0.22	0.24
	eXpress	0.01	0.03	0.07	0.01	0.03	0.10	0.01	0.06	0.17	0.01	0.02	0.08	0.00	0.02	0.08	0.02	0.04	0.17	0.01	0.02	0.08	0.01	0.02	0.05	0.01	0.04	0.10
	RSEM	0.01	0.02	0.06	0.01	0.01	0.09	0.01	0.05	0.16	0.01	0.02	0.07	0.00	0.02	0.08	0.01	0.04	0.14	0.00	0.02	0.08	0.01	0.02	0.06	0.01	0.03	0.10
	Sailfish	0.25	0.30	0.38	0.28	0.34	0.52	0.29	0.48	0.67	0.24	0.30	0.39	0.27	0.31	0.43	0.31	0.49	0.57	0.23	0.29	0.39	0.25	0.33	0.40	0.30	0.38	0.50

Figure 2.16: Fraction of genes with false negative isoforms as a function of gene expression levels. The fraction of genes with false negative isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and at varying gene-level expression cutoffs. A false negative isoform is defined as a transcript with true $\Theta > 0.05$ and estimated $\Theta < 0.001$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Genes all	Cufflinks STAR	0.06	0.06	0.07	0.08	0.08	0.10	0.12	0.13	0.15	0.06	0.06	0.07	0.06	0.06	0.07	0.10	0.12	0.13	0.05	0.06	0.06	0.06	0.06	0.06	0.07	0.08	0.09
	Cufflinks TopHat	0.08	0.08	0.09	0.10	0.10	0.12	0.14	0.16	0.18	0.07	0.07	0.08	0.08	0.08	0.09	0.13	0.14	0.15	0.07	0.07	0.08	0.07	0.08	0.08	0.09	0.10	0.11
	eXpress	0.06	0.07	0.09	0.08	0.09	0.12	0.09	0.12	0.17	0.06	0.07	0.08	0.06	0.08	0.09	0.08	0.11	0.15	0.06	0.07	0.08	0.06	0.07	0.08	0.07	0.08	0.11
	RSEM	0.03	0.04	0.06	0.04	0.05	0.08	0.04	0.07	0.11	0.03	0.04	0.05	0.03	0.04	0.06	0.04	0.06	0.10	0.03	0.04	0.05	0.03	0.04	0.05	0.04	0.05	0.07
	Sailfish	0.24	0.28	0.33	0.29	0.35	0.41	0.35	0.46	0.55	0.23	0.26	0.31	0.24	0.28	0.33	0.33	0.43	0.48	0.23	0.27	0.31	0.22	0.26	0.31	0.27	0.34	0.39
FN Genes 2 Isoforms	Cufflinks STAR	0.01	0.01	0.01	0.01	0.01	0.01	0.09	0.10	0.10	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Cufflinks TopHat	0.01	0.01	0.01	0.01	0.01	0.01	0.10	0.11	0.11	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	eXpress	0.03	0.03	0.03	0.03	0.03	0.03	0.12	0.15	0.17	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.03	0.03	0.02	0.03	0.03
	RSEM	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sailfish	0.06	0.08	0.10	0.06	0.07	0.10	0.25	0.35	0.40	0.06	0.08	0.10	0.06	0.08	0.10	0.07	0.09	0.10	0.06	0.08	0.10	0.06	0.08	0.10	0.06	0.08	0.10
FN Genes 3-4 Isoforms	Cufflinks STAR	0.01	0.01	0.01	0.02	0.02	0.02	0.11	0.12	0.13	0.01	0.01	0.01	0.01	0.01	0.01	0.08	0.09	0.09	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Cufflinks TopHat	0.01	0.02	0.02	0.03	0.03	0.03	0.14	0.15	0.16	0.01	0.02	0.02	0.01	0.01	0.01	0.10	0.11	0.11	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
	eXpress	0.02	0.03	0.03	0.03	0.04	0.05	0.12	0.15	0.20	0.02	0.03	0.03	0.02	0.02	0.02	0.09	0.11	0.14	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03
	RSEM	0.01	0.01	0.01	0.02	0.02	0.02	0.05	0.09	0.13	0.01	0.01	0.02	0.01	0.01	0.01	0.04	0.06	0.09	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
	Sailfish	0.06	0.08	0.09	0.08	0.10	0.13	0.33	0.45	0.55	0.06	0.08	0.10	0.05	0.07	0.09	0.26	0.36	0.39	0.05	0.07	0.09	0.06	0.08	0.09	0.06	0.09	0.11
FN Genes 5-7 Isoforms	Cufflinks STAR	0.03	0.04	0.04	0.08	0.07	0.09	0.11	0.12	0.14	0.03	0.04	0.04	0.04	0.04	0.04	0.11	0.12	0.13	0.03	0.03	0.04	0.03	0.03	0.03	0.06	0.07	0.07
	Cufflinks TopHat	0.05	0.05	0.05	0.10	0.09	0.12	0.14	0.15	0.17	0.05	0.05	0.05	0.05	0.05	0.05	0.14	0.15	0.15	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.09	0.09
	eXpress	0.04	0.05	0.06	0.09	0.10	0.14	0.10	0.13	0.17	0.04	0.04	0.06	0.05	0.06	0.07	0.11	0.13	0.17	0.04	0.04	0.06	0.04	0.05	0.06	0.07	0.08	0.10
	RSEM	0.02	0.03	0.04	0.04	0.05	0.08	0.05	0.08	0.13	0.02	0.03	0.04	0.02	0.03	0.04	0.06	0.08	0.12	0.02	0.03	0.04	0.02	0.03	0.04	0.04	0.05	0.07
	Sailfish	0.15	0.18	0.22	0.29	0.35	0.44	0.38	0.50	0.60	0.14	0.16	0.20	0.15	0.18	0.22	0.36	0.48	0.54	0.14	0.17	0.21	0.13	0.16	0.19	0.24	0.30	0.36

0.00
0.10
0.20
0.30
0.40

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Genes 8-10 Isoforms	Cufflinks STAR	0.08	0.09	0.09	0.12	0.12	0.13	0.14	0.15	0.17	0.08	0.07	0.08	0.09	0.09	0.10	0.14	0.15	0.16	0.07	0.07	0.08	0.07	0.08	0.08	0.11	0.11	0.13
	Cufflinks TopHat	0.11	0.12	0.12	0.15	0.14	0.17	0.17	0.19	0.21	0.10	0.09	0.10	0.11	0.12	0.12	0.17	0.18	0.20	0.09	0.10	0.11	0.09	0.10	0.10	0.13	0.14	0.15
	eXpress	0.09	0.11	0.13	0.12	0.14	0.17	0.10	0.13	0.19	0.09	0.09	0.12	0.09	0.10	0.13	0.11	0.13	0.19	0.08	0.10	0.12	0.08	0.09	0.12	0.12	0.13	0.17
	RSEM	0.05	0.06	0.09	0.06	0.08	0.12	0.05	0.08	0.14	0.05	0.05	0.08	0.05	0.07	0.09	0.05	0.08	0.13	0.04	0.06	0.08	0.05	0.06	0.07	0.06	0.07	0.11
	Sailfish	0.32	0.39	0.46	0.41	0.48	0.58	0.41	0.53	0.64	0.27	0.32	0.39	0.32	0.38	0.46	0.42	0.54	0.60	0.28	0.35	0.41	0.27	0.33	0.39	0.39	0.48	0.56
FN Genes 11-15 Isoforms	Cufflinks STAR	0.16	0.17	0.18	0.18	0.18	0.22	0.22	0.26	0.30	0.16	0.16	0.19	0.16	0.16	0.19	0.21	0.24	0.26	0.16	0.16	0.19	0.16	0.17	0.18	0.17	0.18	0.21
	Cufflinks TopHat	0.19	0.21	0.22	0.21	0.21	0.25	0.23	0.28	0.32	0.19	0.20	0.22	0.20	0.20	0.22	0.23	0.26	0.29	0.20	0.21	0.23	0.19	0.20	0.22	0.22	0.22	0.25
	eXpress	0.10	0.12	0.16	0.09	0.11	0.15	0.06	0.11	0.15	0.09	0.11	0.17	0.10	0.12	0.16	0.07	0.10	0.17	0.09	0.10	0.15	0.10	0.12	0.15	0.08	0.11	0.15
	RSEM	0.06	0.08	0.12	0.07	0.09	0.13	0.05	0.08	0.14	0.07	0.08	0.12	0.06	0.08	0.13	0.05	0.09	0.13	0.07	0.08	0.12	0.07	0.08	0.12	0.07	0.09	0.12
	Sailfish	0.59	0.67	0.75	0.58	0.67	0.75	0.49	0.64	0.73	0.61	0.66	0.74	0.61	0.67	0.74	0.51	0.66	0.73	0.60	0.67	0.74	0.61	0.67	0.74	0.57	0.67	0.74
FN Genes 16-20 Isoforms	Cufflinks STAR	0.14	0.15	0.15	0.17	0.16	0.18	0.19	0.22	0.24	0.14	0.15	0.17	0.14	0.15	0.16	0.17	0.20	0.22	0.14	0.15	0.16	0.14	0.14	0.15	0.15	0.16	0.17
	Cufflinks TopHat	0.17	0.18	0.18	0.19	0.19	0.21	0.20	0.23	0.26	0.17	0.18	0.20	0.17	0.17	0.18	0.20	0.23	0.25	0.18	0.18	0.18	0.17	0.18	0.18	0.18	0.20	0.20
	eXpress	0.11	0.13	0.16	0.10	0.13	0.17	0.08	0.12	0.18	0.10	0.12	0.16	0.11	0.14	0.18	0.09	0.12	0.19	0.12	0.13	0.17	0.10	0.11	0.16	0.11	0.13	0.17
	RSEM	0.06	0.08	0.11	0.07	0.08	0.13	0.05	0.07	0.14	0.06	0.08	0.11	0.07	0.09	0.11	0.05	0.08	0.12	0.07	0.08	0.11	0.06	0.08	0.11	0.07	0.08	0.12
	Sailfish	0.50	0.57	0.67	0.51	0.60	0.69	0.45	0.57	0.68	0.50	0.55	0.65	0.51	0.58	0.68	0.46	0.59	0.67	0.50	0.58	0.67	0.49	0.57	0.67	0.52	0.63	0.69
FN Genes >20 Isoforms	Cufflinks STAR	0.11	0.12	0.12	0.12	0.13	0.15	0.13	0.16	0.19	0.10	0.10	0.11	0.11	0.12	0.13	0.13	0.15	0.17	0.10	0.11	0.12	0.10	0.11	0.12	0.12	0.13	0.14
	Cufflinks TopHat	0.14	0.15	0.16	0.16	0.17	0.19	0.17	0.19	0.22	0.13	0.13	0.15	0.15	0.16	0.17	0.17	0.18	0.21	0.13	0.14	0.15	0.14	0.14	0.16	0.16	0.17	0.18
	eXpress	0.10	0.12	0.16	0.10	0.13	0.16	0.08	0.11	0.17	0.09	0.10	0.13	0.09	0.12	0.15	0.08	0.11	0.17	0.09	0.11	0.14	0.10	0.10	0.14	0.10	0.12	0.16
	RSEM	0.06	0.07	0.10	0.05	0.07	0.11	0.04	0.07	0.12	0.05	0.05	0.08	0.05	0.07	0.10	0.05	0.07	0.12	0.05	0.06	0.10	0.05	0.07	0.10	0.06	0.07	0.11
	Sailfish	0.42	0.50	0.58	0.45	0.54	0.64	0.42	0.54	0.65	0.40	0.47	0.56	0.42	0.49	0.60	0.43	0.55	0.63	0.40	0.49	0.56	0.38	0.46	0.54	0.45	0.55	0.63

Figure 2.17: Fraction of genes with false negative isoforms as a function of annotated isoform complexity. The fraction of genes with false negative isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and for genes with different number of isoforms annotated in GENCODE V16. A false negative isoform is defined as a transcript with true $\Theta > 0.05$ and estimated $\Theta < 0.001$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Genes all	Cufflinks STAR	0.05	0.05	0.05	0.04	0.04	0.04	0.01	0.01	0.01	0.05	0.06	0.05	0.05	0.05	0.05	0.03	0.03	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04
	Cufflinks TopHat	0.05	0.06	0.06	0.04	0.04	0.05	0.01	0.01	0.01	0.06	0.07	0.06	0.05	0.05	0.06	0.03	0.03	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05
	eXpress	0.04	0.04	0.04	0.03	0.03	0.03	0.00	0.01	0.01	0.03	0.04	0.04	0.04	0.04	0.04	0.02	0.02	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.03	0.03	0.03
	RSEM	0.09	0.09	0.09	0.08	0.08	0.09	0.01	0.01	0.01	0.09	0.15	0.10	0.09	0.09	0.10	0.06	0.06	0.14	0.09	0.10	0.10	0.09	0.09	0.10	0.08	0.09	0.09
	Sailfish	0.14	0.16	0.17	0.13	0.14	0.16	0.04	0.06	0.06	0.15	0.20	0.18	0.15	0.17	0.18	0.11	0.12	0.18	0.15	0.17	0.18	0.15	0.17	0.18	0.13	0.15	0.16
FP Genes 0-1 FPKM	Cufflinks STAR	0.14	0.14	0.14	0.11	0.11	0.11	0.02	0.02	0.01	0.14	0.15	0.14	0.14	0.13	0.14	0.08	0.08	0.11	0.14	0.14	0.14	0.14	0.14	0.15	0.13	0.13	0.13
	Cufflinks TopHat	0.15	0.15	0.15	0.12	0.12	0.13	0.02	0.02	0.02	0.15	0.17	0.16	0.15	0.15	0.15	0.08	0.08	0.12	0.15	0.16	0.16	0.15	0.15	0.16	0.14	0.14	0.14
	eXpress	0.13	0.13	0.14	0.10	0.10	0.11	0.01	0.02	0.02	0.13	0.14	0.14	0.13	0.13	0.14	0.07	0.07	0.10	0.13	0.13	0.14	0.13	0.13	0.14	0.11	0.12	0.12
	RSEM	0.34	0.35	0.36	0.32	0.32	0.33	0.04	0.04	0.05	0.36	0.40	0.38	0.35	0.36	0.37	0.24	0.25	0.32	0.36	0.37	0.38	0.36	0.37	0.37	0.33	0.33	0.34
	Sailfish	0.42	0.44	0.44	0.39	0.40	0.41	0.15	0.18	0.19	0.46	0.48	0.46	0.43	0.44	0.45	0.32	0.34	0.40	0.43	0.45	0.46	0.44	0.45	0.46	0.41	0.42	0.44
FP Genes 1-5 FPKM	Cufflinks STAR	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.02	0.02	0.01	0.01	0.01	0.05	0.01	0.02	0.02	0.01	0.01	0.02	0.02	0.02	0.01
	Cufflinks TopHat	0.03	0.03	0.03	0.02	0.02	0.02	0.00	0.01	0.00	0.03	0.04	0.03	0.03	0.03	0.03	0.01	0.01	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	eXpress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
	RSEM	0.03	0.04	0.04	0.03	0.03	0.04	0.00	0.00	0.00	0.03	0.10	0.04	0.03	0.03	0.04	0.01	0.01	0.14	0.03	0.03	0.04	0.03	0.04	0.04	0.03	0.03	0.03
	Sailfish	0.08	0.15	0.20	0.07	0.12	0.17	0.02	0.05	0.05	0.12	0.20	0.22	0.09	0.15	0.22	0.06	0.11	0.24	0.10	0.15	0.21	0.09	0.15	0.24	0.07	0.13	0.19
FP Genes 5-10 FPKM	Cufflinks STAR	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01
	Cufflinks TopHat	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.02	0.03	0.02	0.03	0.03	0.02	0.01	0.01	0.03	0.03	0.02	0.02	0.03	0.03	0.02	0.02	0.02	0.02
	eXpress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RSEM	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.09	0.01	0.01	0.01	0.01	0.00	0.00	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Sailfish	0.04	0.06	0.09	0.03	0.05	0.07	0.01	0.01	0.02	0.05	0.12	0.10	0.04	0.07	0.09	0.02	0.03	0.13	0.05	0.06	0.09	0.05	0.07	0.10	0.04	0.05	0.09

0.00
0.10
0.20
0.30
0.40

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Genes 10-50 FPKM	Cufflinks STAR	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01
	Cufflinks TopHat	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01
	eXpress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RSEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sailfish	0.02	0.03	0.03	0.02	0.03	0.03	0.00	0.00	0.00	0.02	0.07	0.04	0.02	0.03	0.04	0.01	0.02	0.06	0.03	0.03	0.04	0.02	0.03	0.05	0.01	0.02	0.03
FP Genes 50-100 FPKM	Cufflinks STAR	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01
	Cufflinks TopHat	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01
	eXpress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RSEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sailfish	0.01	0.02	0.02	0.02	0.02	0.03	0.00	0.00	0.00	0.02	0.05	0.03	0.02	0.03	0.03	0.01	0.02	0.04	0.02	0.02	0.03	0.02	0.02	0.04	0.01	0.01	0.02
FP Genes 100-500 FPKM	Cufflinks STAR	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01
	Cufflinks TopHat	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01
	eXpress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RSEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sailfish	0.01	0.02	0.02	0.01	0.01	0.02	0.00	0.00	0.00	0.02	0.05	0.03	0.02	0.03	0.03	0.01	0.01	0.04	0.03	0.03	0.04	0.02	0.03	0.03	0.01	0.02	0.02
FP Genes >500 FPKM	Cufflinks STAR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00
	Cufflinks TopHat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.00
	eXpress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RSEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sailfish	0.03	0.03	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.06	0.02	0.01	0.02	0.02	0.00	0.00	0.04	0.01	0.02	0.02	0.03	0.04	0.04	0.02	0.02	0.02

Figure 2.18: Fraction of genes with false positive isoforms as a function of gene expression levels. The fraction of genes with false positive isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and at varying gene-level expression cutoffs. A false positive isoform is defined as a transcript with true $\Theta = 0$ and estimated $\Theta > 0.05$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Genes all	Cufflinks STAR	0.05	0.05	0.05	0.04	0.04	0.04	0.01	0.01	0.01	0.05	0.06	0.05	0.05	0.05	0.05	0.03	0.03	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04
	Cufflinks TopHat	0.05	0.06	0.06	0.04	0.04	0.05	0.01	0.01	0.01	0.06	0.07	0.06	0.05	0.05	0.06	0.03	0.03	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05
	eXpress	0.04	0.04	0.04	0.03	0.03	0.03	0.00	0.01	0.01	0.03	0.04	0.04	0.04	0.04	0.04	0.02	0.02	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.03	0.03	0.03
	RSEM	0.09	0.09	0.09	0.08	0.08	0.09	0.01	0.01	0.01	0.09	0.15	0.10	0.09	0.09	0.10	0.06	0.06	0.14	0.09	0.10	0.10	0.09	0.09	0.10	0.08	0.09	0.09
	Sailfish	0.14	0.16	0.17	0.13	0.14	0.16	0.04	0.06	0.06	0.15	0.20	0.18	0.15	0.17	0.18	0.11	0.12	0.18	0.15	0.17	0.18	0.15	0.17	0.18	0.13	0.15	0.16
FP Genes 2 Isoforms	Cufflinks STAR	0.06	0.06	0.06	0.06	0.06	0.06	0.01	0.01	0.01	0.07	0.08	0.07	0.07	0.07	0.07	0.05	0.05	0.06	0.07	0.07	0.07	0.06	0.06	0.06	0.07	0.07	0.07
	Cufflinks TopHat	0.06	0.06	0.07	0.06	0.06	0.07	0.02	0.02	0.02	0.07	0.08	0.08	0.07	0.08	0.08	0.05	0.05	0.06	0.08	0.08	0.08	0.06	0.06	0.07	0.07	0.08	0.08
	eXpress	0.04	0.04	0.04	0.04	0.04	0.04	0.01	0.01	0.01	0.05	0.06	0.05	0.05	0.05	0.05	0.03	0.03	0.04	0.05	0.05	0.05	0.04	0.04	0.05	0.05	0.05	0.05
	RSEM	0.11	0.11	0.12	0.13	0.13	0.14	0.01	0.02	0.02	0.12	0.16	0.12	0.12	0.12	0.12	0.13	0.13	0.20	0.12	0.12	0.12	0.11	0.11	0.11	0.12	0.12	0.13
	Sailfish	0.13	0.14	0.15	0.15	0.15	0.16	0.03	0.04	0.05	0.14	0.19	0.16	0.14	0.15	0.16	0.14	0.14	0.19	0.13	0.15	0.16	0.14	0.15	0.15	0.14	0.16	0.16
FP Genes 3-4 Isoforms	Cufflinks STAR	0.07	0.07	0.07	0.05	0.05	0.05	0.01	0.01	0.01	0.06	0.06	0.06	0.06	0.07	0.07	0.03	0.03	0.05	0.07	0.07	0.06	0.07	0.06	0.06	0.06	0.06	0.06
	Cufflinks TopHat	0.08	0.08	0.08	0.06	0.06	0.07	0.01	0.01	0.01	0.07	0.08	0.07	0.07	0.07	0.07	0.04	0.04	0.05	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.08
	eXpress	0.06	0.06	0.06	0.04	0.04	0.04	0.01	0.00	0.01	0.05	0.06	0.05	0.06	0.06	0.06	0.02	0.02	0.04	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.06
	RSEM	0.14	0.14	0.15	0.14	0.14	0.14	0.01	0.01	0.02	0.13	0.19	0.14	0.14	0.14	0.15	0.10	0.10	0.18	0.13	0.14	0.15	0.14	0.14	0.15	0.14	0.14	0.14
	Sailfish	0.18	0.20	0.21	0.18	0.19	0.20	0.05	0.06	0.07	0.18	0.24	0.22	0.19	0.21	0.22	0.14	0.15	0.19	0.19	0.21	0.22	0.19	0.21	0.22	0.18	0.19	0.21
FP Genes 5-7 Isoforms	Cufflinks STAR	0.05	0.05	0.06	0.04	0.04	0.04	0.01	0.01	0.01	0.06	0.06	0.06	0.05	0.05	0.05	0.03	0.03	0.05	0.05	0.06	0.06	0.05	0.06	0.06	0.05	0.05	0.05
	Cufflinks TopHat	0.06	0.06	0.07	0.05	0.05	0.05	0.01	0.01	0.01	0.07	0.07	0.07	0.06	0.06	0.06	0.03	0.03	0.05	0.07	0.07	0.07	0.06	0.06	0.07	0.05	0.06	0.06
	eXpress	0.05	0.05	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.05	0.05	0.05	0.04	0.04	0.05	0.02	0.02	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.04	0.04
	RSEM	0.12	0.12	0.13	0.10	0.10	0.11	0.01	0.01	0.01	0.12	0.19	0.13	0.12	0.12	0.13	0.07	0.07	0.15	0.12	0.12	0.13	0.12	0.12	0.12	0.10	0.11	0.11
	Sailfish	0.18	0.20	0.22	0.15	0.16	0.18	0.05	0.07	0.07	0.19	0.24	0.22	0.19	0.20	0.21	0.12	0.14	0.20	0.19	0.21	0.22	0.18	0.20	0.21	0.16	0.18	0.19

0.00
0.10
0.20
0.30
0.40

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Genes 8-10 Isoforms	Cufflinks STAR	0.05	0.05	0.05	0.04	0.04	0.04	0.01	0.01	0.01	0.05	0.06	0.05	0.05	0.05	0.05	0.03	0.03	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.04	0.05
	Cufflinks TopHat	0.06	0.06	0.06	0.05	0.05	0.05	0.01	0.01	0.01	0.06	0.07	0.06	0.05	0.06	0.06	0.03	0.03	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05
	eXpress	0.04	0.04	0.04	0.03	0.04	0.04	0.01	0.01	0.01	0.04	0.05	0.05	0.04	0.04	0.04	0.02	0.02	0.05	0.04	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04
	RSEM	0.10	0.10	0.10	0.08	0.08	0.09	0.01	0.01	0.01	0.10	0.18	0.11	0.10	0.10	0.10	0.06	0.06	0.16	0.10	0.11	0.11	0.10	0.10	0.11	0.08	0.09	0.09
	Sailfish	0.16	0.19	0.21	0.14	0.17	0.19	0.06	0.08	0.08	0.18	0.25	0.23	0.17	0.19	0.21	0.13	0.14	0.22	0.18	0.20	0.23	0.17	0.20	0.22	0.15	0.17	0.19
FP Genes 11-15 Isoforms	Cufflinks STAR	0.04	0.04	0.04	0.03	0.03	0.03	0.01	0.01	0.01	0.04	0.06	0.04	0.04	0.04	0.05	0.02	0.02	0.05	0.04	0.04	0.04	0.04	0.05	0.05	0.03	0.03	0.03
	Cufflinks TopHat	0.04	0.05	0.05	0.03	0.03	0.04	0.01	0.01	0.01	0.05	0.07	0.05	0.05	0.05	0.05	0.02	0.02	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.04	0.04	0.04
	eXpress	0.02	0.02	0.03	0.02	0.02	0.02	0.00	0.01	0.00	0.02	0.03	0.03	0.02	0.03	0.03	0.02	0.02	0.05	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02
	RSEM	0.06	0.07	0.07	0.06	0.06	0.06	0.01	0.01	0.01	0.07	0.15	0.08	0.08	0.08	0.08	0.04	0.04	0.13	0.07	0.07	0.07	0.08	0.08	0.09	0.06	0.06	0.07
	Sailfish	0.14	0.16	0.18	0.13	0.14	0.16	0.05	0.06	0.07	0.15	0.22	0.20	0.14	0.17	0.19	0.10	0.12	0.20	0.15	0.18	0.19	0.16	0.18	0.20	0.13	0.15	0.17
FP Genes 16-20 Isoforms	Cufflinks STAR	0.04	0.04	0.05	0.04	0.04	0.04	0.01	0.01	0.01	0.06	0.08	0.06	0.05	0.05	0.05	0.03	0.03	0.07	0.05	0.05	0.05	0.06	0.06	0.06	0.04	0.04	0.04
	Cufflinks TopHat	0.05	0.05	0.06	0.04	0.04	0.05	0.01	0.01	0.01	0.07	0.09	0.07	0.06	0.06	0.06	0.03	0.03	0.08	0.06	0.06	0.06	0.06	0.06	0.06	0.04	0.05	0.05
	eXpress	0.02	0.02	0.03	0.02	0.02	0.02	0.00	0.01	0.00	0.03	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.07	0.02	0.02	0.02	0.03	0.03	0.04	0.02	0.02	0.02
	RSEM	0.06	0.07	0.07	0.06	0.06	0.06	0.01	0.01	0.01	0.07	0.15	0.07	0.07	0.07	0.07	0.05	0.05	0.15	0.06	0.07	0.07	0.08	0.08	0.08	0.06	0.07	0.07
	Sailfish	0.13	0.15	0.17	0.11	0.13	0.14	0.05	0.06	0.06	0.14	0.22	0.19	0.14	0.16	0.17	0.10	0.11	0.20	0.15	0.18	0.19	0.14	0.16	0.18	0.12	0.14	0.16
FP Genes >20 Isoforms	Cufflinks STAR	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.03	0.05	0.04	0.03	0.03	0.03	0.02	0.02	0.08	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03
	Cufflinks TopHat	0.04	0.04	0.04	0.03	0.03	0.03	0.01	0.01	0.01	0.04	0.06	0.05	0.04	0.04	0.04	0.03	0.02	0.08	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.03
	eXpress	0.01	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.07	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02
	RSEM	0.04	0.04	0.05	0.04	0.05	0.05	0.01	0.00	0.01	0.05	0.13	0.06	0.04	0.04	0.05	0.03	0.04	0.14	0.04	0.05	0.05	0.04	0.04	0.05	0.04	0.04	0.04
	Sailfish	0.12	0.16	0.19	0.11	0.13	0.16	0.05	0.06	0.06	0.14	0.22	0.21	0.13	0.16	0.19	0.09	0.11	0.20	0.15	0.17	0.19	0.13	0.17	0.18	0.10	0.13	0.15

Figure 2.19: Fraction of genes with false positive isoforms as a function of annotated isoform complexity. The fraction of genes with false positive isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and for genes with different number of isoforms annotated in GENCODE V16. A false positive isoform is defined as a transcript with true $\Theta = 0$ and estimated $\Theta > 0.05$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Isoforms all	Cufflinks STAR	0.009	0.009	0.010	0.012	0.012	0.014	0.017	0.021	0.023	0.008	0.009	0.010	0.009	0.009	0.011	0.015	0.017	0.020	0.008	0.009	0.010	0.008	0.009	0.009	0.011	0.012	0.013
	Cufflinks TopHat	0.011	0.012	0.013	0.016	0.015	0.018	0.022	0.025	0.028	0.011	0.011	0.012	0.012	0.012	0.013	0.020	0.022	0.024	0.010	0.011	0.012	0.011	0.011	0.012	0.014	0.015	0.016
	eXpress	0.011	0.013	0.015	0.014	0.016	0.020	0.017	0.021	0.029	0.010	0.011	0.014	0.011	0.013	0.016	0.015	0.018	0.025	0.010	0.011	0.014	0.010	0.012	0.014	0.013	0.014	0.018
	RSEM	0.005	0.006	0.008	0.006	0.007	0.011	0.006	0.010	0.017	0.004	0.005	0.008	0.005	0.006	0.009	0.006	0.009	0.015	0.004	0.005	0.008	0.005	0.005	0.008	0.005	0.007	0.010
	Sailfish	0.039	0.050	0.062	0.048	0.061	0.079	0.059	0.083	0.108	0.037	0.045	0.058	0.040	0.049	0.062	0.055	0.077	0.090	0.037	0.048	0.059	0.036	0.047	0.057	0.045	0.060	0.073
FN Isoforms 0-1 FPKM	Cufflinks STAR	0.023	0.024	0.024	0.030	0.026	0.032	0.036	0.040	0.043	0.022	0.022	0.023	0.024	0.025	0.026	0.034	0.036	0.035	0.020	0.021	0.023	0.021	0.022	0.023	0.027	0.028	0.029
	Cufflinks TopHat	0.026	0.027	0.027	0.033	0.030	0.036	0.043	0.046	0.048	0.023	0.024	0.026	0.027	0.027	0.028	0.039	0.040	0.040	0.023	0.024	0.025	0.024	0.024	0.025	0.030	0.031	0.033
	eXpress	0.042	0.045	0.046	0.053	0.047	0.060	0.067	0.075	0.082	0.040	0.040	0.045	0.041	0.045	0.049	0.061	0.066	0.063	0.038	0.040	0.044	0.039	0.042	0.044	0.048	0.050	0.053
	RSEM	0.018	0.020	0.023	0.022	0.020	0.029	0.022	0.030	0.036	0.017	0.017	0.021	0.018	0.021	0.023	0.023	0.029	0.030	0.016	0.018	0.021	0.017	0.019	0.021	0.020	0.023	0.026
	Sailfish	0.066	0.078	0.085	0.082	0.086	0.106	0.115	0.137	0.154	0.061	0.068	0.081	0.067	0.077	0.084	0.103	0.120	0.123	0.061	0.072	0.081	0.062	0.074	0.083	0.077	0.090	0.098
FN Isoforms 1-5 FPKM	Cufflinks STAR	0.007	0.007	0.007	0.009	0.008	0.009	0.014	0.015	0.016	0.007	0.007	0.007	0.007	0.007	0.007	0.010	0.010	0.012	0.007	0.006	0.007	0.007	0.006	0.006	0.009	0.009	0.010
	Cufflinks TopHat	0.011	0.012	0.012	0.015	0.012	0.016	0.021	0.022	0.024	0.010	0.010	0.011	0.011	0.010	0.012	0.017	0.018	0.020	0.010	0.010	0.010	0.011	0.010	0.010	0.013	0.014	0.015
	eXpress	0.007	0.010	0.014	0.011	0.013	0.019	0.012	0.018	0.028	0.008	0.009	0.012	0.007	0.011	0.014	0.010	0.015	0.022	0.007	0.010	0.012	0.007	0.010	0.013	0.009	0.012	0.017
	RSEM	0.006	0.007	0.009	0.006	0.007	0.013	0.005	0.010	0.019	0.004	0.005	0.008	0.005	0.007	0.010	0.005	0.009	0.015	0.005	0.006	0.009	0.005	0.006	0.010	0.006	0.008	0.012
	Sailfish	0.045	0.059	0.071	0.056	0.077	0.093	0.082	0.117	0.139	0.046	0.052	0.066	0.045	0.060	0.072	0.067	0.103	0.116	0.039	0.054	0.065	0.041	0.055	0.065	0.052	0.074	0.087
FN Isoforms 5-10 FPKM	Cufflinks STAR	0.005	0.005	0.006	0.006	0.006	0.007	0.011	0.013	0.014	0.005	0.006	0.005	0.004	0.004	0.005	0.007	0.008	0.009	0.004	0.004	0.005	0.004	0.005	0.005	0.005	0.005	0.006
	Cufflinks TopHat	0.007	0.007	0.008	0.011	0.010	0.011	0.016	0.018	0.019	0.008	0.008	0.008	0.008	0.008	0.008	0.012	0.014	0.014	0.007	0.007	0.008	0.007	0.007	0.007	0.009	0.009	0.009
	eXpress	0.007	0.008	0.010	0.009	0.009	0.014	0.010	0.015	0.021	0.006	0.006	0.010	0.006	0.009	0.011	0.008	0.011	0.017	0.007	0.008	0.011	0.006	0.007	0.010	0.008	0.010	0.012
	RSEM	0.004	0.005	0.007	0.004	0.005	0.009	0.004	0.008	0.014	0.003	0.004	0.006	0.003	0.005	0.007	0.004	0.007	0.011	0.003	0.004	0.006	0.004	0.004	0.007	0.004	0.005	0.008
	Sailfish	0.034	0.046	0.057	0.043	0.058	0.078	0.054	0.090	0.111	0.031	0.043	0.056	0.034	0.046	0.058	0.051	0.079	0.094	0.034	0.046	0.056	0.032	0.045	0.056	0.041	0.059	0.071

0.000	0.010	0.020	0.030	0.040	0.050

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Isoforms 10-50 FPKM	Cufflinks STAR	0.005	0.005	0.005	0.006	0.007	0.008	0.012	0.015	0.017	0.004	0.005	0.005	0.005	0.005	0.005	0.010	0.012	0.014	0.004	0.005	0.005	0.004	0.004	0.005	0.005	0.006	0.007
	Cufflinks TopHat	0.008	0.008	0.008	0.010	0.009	0.011	0.016	0.018	0.021	0.008	0.008	0.008	0.008	0.008	0.009	0.014	0.016	0.017	0.007	0.007	0.008	0.007	0.007	0.008	0.008	0.009	0.010
	eXpress	0.004	0.005	0.008	0.005	0.005	0.009	0.005	0.008	0.015	0.004	0.004	0.007	0.005	0.005	0.008	0.005	0.007	0.012	0.004	0.005	0.007	0.004	0.004	0.007	0.005	0.006	0.009
	RSEM	0.002	0.003	0.005	0.002	0.003	0.007	0.002	0.005	0.012	0.002	0.003	0.005	0.002	0.003	0.005	0.002	0.005	0.010	0.002	0.003	0.005	0.002	0.003	0.005	0.002	0.003	0.006
	Sailfish	0.029	0.040	0.053	0.035	0.048	0.067	0.039	0.062	0.087	0.029	0.036	0.048	0.030	0.039	0.054	0.038	0.059	0.076	0.030	0.039	0.051	0.027	0.037	0.049	0.035	0.048	0.064
FN Isoforms 50-100 FPKM	Cufflinks STAR	0.005	0.006	0.007	0.012	0.012	0.016	0.019	0.024	0.030	0.006	0.006	0.008	0.006	0.006	0.007	0.018	0.020	0.027	0.006	0.006	0.007	0.005	0.006	0.007	0.010	0.011	0.011
	Cufflinks TopHat	0.007	0.007	0.008	0.014	0.013	0.017	0.021	0.027	0.031	0.008	0.008	0.009	0.007	0.007	0.008	0.019	0.024	0.028	0.007	0.007	0.008	0.007	0.007	0.008	0.012	0.013	0.013
	eXpress	0.003	0.003	0.006	0.003	0.004	0.008	0.002	0.005	0.013	0.002	0.003	0.005	0.002	0.003	0.006	0.002	0.004	0.010	0.002	0.003	0.006	0.002	0.003	0.005	0.002	0.004	0.008
	RSEM	0.001	0.001	0.003	0.001	0.002	0.006	0.001	0.003	0.011	0.001	0.001	0.003	0.001	0.001	0.004	0.001	0.003	0.008	0.001	0.002	0.004	0.001	0.002	0.003	0.001	0.002	0.006
	Sailfish	0.030	0.037	0.052	0.034	0.041	0.063	0.033	0.052	0.078	0.028	0.034	0.048	0.031	0.036	0.050	0.034	0.052	0.068	0.029	0.037	0.048	0.028	0.034	0.048	0.033	0.043	0.060
FN Isoforms 100-500 FPKM	Cufflinks STAR	0.010	0.012	0.014	0.014	0.019	0.021	0.020	0.028	0.035	0.008	0.009	0.011	0.010	0.013	0.019	0.017	0.024	0.030	0.009	0.010	0.013	0.008	0.008	0.010	0.013	0.016	0.021
	Cufflinks TopHat	0.011	0.013	0.015	0.016	0.019	0.023	0.022	0.030	0.037	0.009	0.010	0.012	0.012	0.015	0.019	0.021	0.028	0.034	0.011	0.012	0.014	0.009	0.010	0.012	0.015	0.018	0.021
	eXpress	0.002	0.002	0.005	0.002	0.003	0.008	0.002	0.005	0.012	0.002	0.002	0.005	0.002	0.002	0.005	0.002	0.003	0.011	0.002	0.002	0.005	0.001	0.002	0.005	0.002	0.003	0.006
	RSEM	0.001	0.001	0.004	0.001	0.002	0.006	0.001	0.004	0.011	0.001	0.001	0.004	0.001	0.002	0.005	0.001	0.003	0.010	0.001	0.001	0.004	0.001	0.001	0.004	0.001	0.002	0.005
	Sailfish	0.033	0.040	0.055	0.039	0.047	0.071	0.040	0.059	0.089	0.032	0.036	0.051	0.033	0.038	0.056	0.040	0.058	0.070	0.032	0.039	0.053	0.030	0.036	0.047	0.036	0.047	0.063
FN Isoforms >500 FPKM	Cufflinks STAR	0.012	0.018	0.020	0.012	0.018	0.031	0.018	0.034	0.047	0.009	0.017	0.024	0.009	0.015	0.020	0.016	0.029	0.042	0.006	0.015	0.018	0.007	0.014	0.018	0.015	0.025	0.029
	Cufflinks TopHat	0.011	0.017	0.024	0.011	0.019	0.029	0.014	0.031	0.045	0.009	0.016	0.026	0.009	0.017	0.021	0.014	0.030	0.039	0.005	0.013	0.020	0.010	0.015	0.022	0.017	0.027	0.030
	eXpress	0.001	0.003	0.007	0.001	0.003	0.010	0.001	0.008	0.021	0.001	0.003	0.008	0.000	0.002	0.008	0.002	0.005	0.022	0.000	0.002	0.009	0.001	0.002	0.006	0.001	0.003	0.011
	RSEM	0.001	0.002	0.007	0.001	0.002	0.009	0.001	0.006	0.019	0.001	0.003	0.008	0.000	0.002	0.008	0.001	0.005	0.019	0.000	0.002	0.008	0.001	0.002	0.007	0.001	0.003	0.010
	Sailfish	0.040	0.051	0.073	0.044	0.057	0.096	0.048	0.083	0.136	0.036	0.046	0.070	0.039	0.047	0.072	0.050	0.084	0.095	0.033	0.046	0.070	0.035	0.046	0.064	0.044	0.061	0.087

Figure 2.20: Fraction of genes with false negative isoforms as a function of gene expression levels. The fraction of false negative isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and at varying gene-level expression cutoffs. A false negative isoform is defined as a transcript with true $\Theta > 0.05$ and estimated $\Theta < 0.001$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Isoforms all	Cufflinks STAR	0.009	0.009	0.010	0.012	0.012	0.014	0.017	0.021	0.023	0.008	0.009	0.010	0.009	0.009	0.011	0.015	0.017	0.020	0.008	0.009	0.010	0.008	0.009	0.009	0.011	0.012	0.013
	Cufflinks TopHat	0.011	0.012	0.013	0.016	0.015	0.018	0.022	0.025	0.028	0.011	0.011	0.012	0.012	0.012	0.013	0.020	0.022	0.024	0.010	0.011	0.012	0.011	0.011	0.012	0.014	0.015	0.016
	eXpress	0.011	0.013	0.015	0.014	0.016	0.020	0.017	0.021	0.029	0.010	0.011	0.014	0.011	0.013	0.016	0.015	0.018	0.025	0.010	0.011	0.014	0.010	0.012	0.014	0.013	0.014	0.018
	RSEM	0.005	0.006	0.008	0.006	0.007	0.011	0.006	0.010	0.017	0.004	0.005	0.008	0.005	0.006	0.009	0.006	0.009	0.015	0.004	0.005	0.008	0.005	0.005	0.008	0.005	0.007	0.010
	Sailfish	0.039	0.050	0.062	0.048	0.061	0.079	0.059	0.083	0.108	0.037	0.045	0.058	0.040	0.049	0.062	0.055	0.077	0.090	0.037	0.048	0.059	0.036	0.047	0.057	0.045	0.060	0.073
FN Isoforms 2 isoforms	Cufflinks STAR	0.005	0.005	0.005	0.005	0.005	0.005	0.044	0.049	0.051	0.004	0.004	0.004	0.005	0.004	0.004	0.006	0.006	0.005	0.004	0.004	0.003	0.004	0.004	0.004	0.005	0.005	0.004
	Cufflinks TopHat	0.007	0.007	0.007	0.006	0.005	0.006	0.052	0.057	0.057	0.006	0.006	0.006	0.006	0.006	0.006	0.007	0.007	0.005	0.004	0.006	0.005	0.005	0.005	0.006	0.006	0.006	0.005
	eXpress	0.019	0.019	0.019	0.018	0.017	0.018	0.067	0.080	0.091	0.019	0.018	0.020	0.017	0.019	0.019	0.018	0.020	0.021	0.017	0.017	0.018	0.018	0.019	0.019	0.018	0.019	0.018
	RSEM	0.003	0.003	0.003	0.002	0.002	0.003	0.014	0.027	0.037	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	Sailfish	0.034	0.046	0.056	0.033	0.038	0.054	0.128	0.177	0.206	0.034	0.042	0.054	0.035	0.045	0.054	0.036	0.049	0.056	0.033	0.045	0.054	0.033	0.044	0.056	0.034	0.043	0.053
FN Isoforms 3-4 isoforms	Cufflinks STAR	0.003	0.004	0.004	0.006	0.005	0.007	0.034	0.038	0.041	0.003	0.003	0.005	0.002	0.003	0.003	0.023	0.026	0.028	0.002	0.003	0.003	0.003	0.003	0.003	0.003	0.004	0.004
	Cufflinks TopHat	0.005	0.005	0.005	0.008	0.007	0.009	0.044	0.048	0.052	0.004	0.005	0.006	0.003	0.004	0.004	0.030	0.033	0.034	0.003	0.003	0.004	0.004	0.004	0.004	0.004	0.005	0.005
	eXpress	0.010	0.011	0.012	0.013	0.013	0.017	0.041	0.052	0.068	0.009	0.011	0.012	0.009	0.010	0.010	0.030	0.037	0.044	0.009	0.009	0.010	0.009	0.010	0.011	0.011	0.011	0.013
	RSEM	0.003	0.004	0.005	0.005	0.005	0.008	0.014	0.027	0.041	0.004	0.004	0.007	0.003	0.004	0.005	0.011	0.019	0.027	0.003	0.003	0.004	0.004	0.004	0.005	0.004	0.005	0.006
	Sailfish	0.019	0.027	0.032	0.027	0.034	0.043	0.109	0.156	0.193	0.021	0.026	0.034	0.019	0.025	0.030	0.083	0.114	0.127	0.019	0.025	0.030	0.020	0.027	0.032	0.022	0.030	0.036
FN Isoforms 5-7 isoforms	Cufflinks STAR	0.006	0.006	0.007	0.015	0.014	0.018	0.022	0.024	0.027	0.006	0.007	0.007	0.007	0.007	0.008	0.020	0.023	0.025	0.005	0.006	0.007	0.005	0.006	0.006	0.012	0.013	0.014
	Cufflinks TopHat	0.008	0.008	0.009	0.019	0.017	0.022	0.028	0.031	0.034	0.008	0.009	0.009	0.008	0.009	0.010	0.026	0.029	0.031	0.007	0.007	0.008	0.007	0.008	0.008	0.015	0.016	0.017
	eXpress	0.010	0.010	0.013	0.020	0.021	0.028	0.025	0.031	0.041	0.010	0.010	0.013	0.011	0.013	0.015	0.025	0.029	0.038	0.009	0.010	0.012	0.009	0.011	0.013	0.016	0.017	0.020
	RSEM	0.005	0.005	0.007	0.008	0.009	0.016	0.009	0.016	0.026	0.004	0.005	0.008	0.005	0.006	0.008	0.010	0.016	0.024	0.004	0.005	0.008	0.004	0.005	0.007	0.007	0.009	0.012
	Sailfish	0.029	0.036	0.045	0.056	0.069	0.092	0.082	0.115	0.149	0.027	0.033	0.043	0.029	0.037	0.046	0.076	0.108	0.124	0.027	0.036	0.044	0.025	0.034	0.040	0.046	0.061	0.074

0.000
0.010
0.020
0.030
0.040
0.050

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FN Isoforms 8-10 isoforms	Cufflinks STAR	0.011	0.011	0.012	0.016	0.015	0.018	0.019	0.022	0.024	0.010	0.010	0.011	0.012	0.012	0.014	0.019	0.020	0.023	0.009	0.010	0.011	0.010	0.010	0.011	0.015	0.015	0.017
	Cufflinks TopHat	0.014	0.015	0.016	0.020	0.019	0.023	0.024	0.027	0.029	0.013	0.013	0.014	0.015	0.015	0.016	0.024	0.025	0.027	0.012	0.013	0.014	0.012	0.012	0.013	0.018	0.019	0.020
	eXpress	0.015	0.016	0.019	0.020	0.022	0.028	0.020	0.025	0.034	0.014	0.014	0.017	0.014	0.016	0.019	0.020	0.023	0.033	0.013	0.014	0.017	0.013	0.015	0.017	0.019	0.020	0.025
	RSEM	0.007	0.008	0.012	0.008	0.011	0.016	0.006	0.010	0.019	0.006	0.007	0.010	0.007	0.008	0.011	0.007	0.010	0.018	0.005	0.007	0.010	0.006	0.007	0.010	0.007	0.010	0.015
	Sailfish	0.043	0.057	0.070	0.058	0.073	0.096	0.064	0.090	0.120	0.037	0.044	0.059	0.044	0.055	0.070	0.062	0.088	0.105	0.039	0.051	0.063	0.037	0.049	0.059	0.056	0.073	0.090
FN Isoforms 11-15 isoforms	Cufflinks STAR	0.010	0.011	0.012	0.012	0.013	0.015	0.013	0.017	0.020	0.009	0.009	0.011	0.010	0.011	0.012	0.013	0.015	0.018	0.009	0.010	0.011	0.010	0.010	0.011	0.011	0.013	0.014
	Cufflinks TopHat	0.014	0.015	0.016	0.016	0.017	0.019	0.017	0.021	0.025	0.012	0.013	0.014	0.014	0.015	0.016	0.017	0.020	0.023	0.012	0.013	0.015	0.013	0.014	0.015	0.015	0.017	0.018
	eXpress	0.011	0.013	0.016	0.011	0.015	0.018	0.009	0.012	0.018	0.010	0.010	0.013	0.011	0.013	0.016	0.009	0.012	0.018	0.009	0.012	0.015	0.010	0.011	0.015	0.011	0.014	0.017
	RSEM	0.006	0.007	0.010	0.005	0.007	0.011	0.004	0.006	0.012	0.004	0.005	0.008	0.005	0.006	0.010	0.004	0.006	0.012	0.005	0.005	0.009	0.005	0.006	0.009	0.005	0.007	0.011
	Sailfish	0.046	0.058	0.072	0.049	0.065	0.084	0.049	0.069	0.092	0.041	0.051	0.067	0.045	0.055	0.073	0.049	0.068	0.084	0.042	0.054	0.068	0.040	0.052	0.065	0.048	0.064	0.081
FN Isoforms 16-20 isoforms	Cufflinks STAR	0.010	0.011	0.011	0.012	0.012	0.014	0.014	0.017	0.019	0.011	0.011	0.012	0.010	0.011	0.012	0.013	0.015	0.016	0.010	0.011	0.012	0.010	0.010	0.011	0.012	0.012	0.013
	Cufflinks TopHat	0.013	0.013	0.013	0.015	0.015	0.016	0.016	0.019	0.021	0.013	0.013	0.014	0.013	0.013	0.014	0.016	0.018	0.019	0.013	0.013	0.014	0.012	0.013	0.013	0.014	0.015	0.015
	eXpress	0.009	0.012	0.013	0.009	0.012	0.014	0.008	0.010	0.017	0.008	0.010	0.013	0.009	0.010	0.014	0.008	0.010	0.016	0.010	0.011	0.014	0.008	0.010	0.013	0.009	0.011	0.014
	RSEM	0.005	0.006	0.009	0.005	0.006	0.010	0.003	0.005	0.011	0.004	0.006	0.008	0.005	0.006	0.009	0.004	0.005	0.010	0.005	0.006	0.009	0.004	0.005	0.008	0.004	0.005	0.009
	Sailfish	0.043	0.053	0.069	0.044	0.059	0.076	0.041	0.058	0.076	0.042	0.050	0.064	0.044	0.053	0.069	0.042	0.060	0.072	0.044	0.054	0.068	0.041	0.052	0.066	0.046	0.061	0.073
FN Isoforms >20 isoforms	Cufflinks STAR	0.008	0.009	0.010	0.010	0.010	0.012	0.012	0.015	0.018	0.008	0.008	0.010	0.008	0.009	0.010	0.011	0.013	0.016	0.008	0.009	0.011	0.008	0.009	0.010	0.010	0.010	0.012
	Cufflinks TopHat	0.010	0.011	0.013	0.012	0.012	0.015	0.013	0.017	0.020	0.010	0.011	0.012	0.011	0.012	0.013	0.013	0.015	0.017	0.011	0.011	0.013	0.010	0.011	0.012	0.012	0.013	0.014
	eXpress	0.006	0.007	0.009	0.007	0.007	0.010	0.005	0.008	0.011	0.007	0.008	0.011	0.006	0.008	0.010	0.005	0.007	0.011	0.005	0.006	0.009	0.006	0.007	0.008	0.005	0.007	0.009
	RSEM	0.003	0.004	0.006	0.003	0.004	0.006	0.002	0.004	0.007	0.003	0.004	0.006	0.003	0.003	0.006	0.003	0.004	0.006	0.003	0.003	0.006	0.003	0.004	0.006	0.003	0.004	0.006
	Sailfish	0.037	0.046	0.055	0.036	0.044	0.054	0.031	0.043	0.056	0.038	0.043	0.054	0.038	0.044	0.054	0.033	0.044	0.053	0.037	0.044	0.053	0.038	0.045	0.053	0.037	0.045	0.054

Figure 2.21: Fraction of false negative isoforms a function of annotated isoform complexity. The fraction of false negative isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and for genes with different number of isoforms annotated in GENCODE V16. A false negative isoform is defined as a transcript with true $\Theta > 0.05$ and estimated $\Theta < 0.001$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Isoforms all	Cufflinks STAR	0.007	0.007	0.008	0.005	0.005	0.006	0.001	0.001	0.001	0.007	0.009	0.008	0.007	0.007	0.007	0.003	0.003	0.006	0.007	0.008	0.008	0.008	0.008	0.008	0.006	0.006	0.006
	Cufflinks TopHat	0.008	0.009	0.009	0.006	0.006	0.006	0.001	0.001	0.001	0.009	0.010	0.009	0.008	0.008	0.009	0.004	0.004	0.007	0.009	0.009	0.009	0.009	0.009	0.010	0.007	0.007	0.007
	eXpress	0.006	0.007	0.007	0.004	0.006	0.005	0.001	0.001	0.001	0.007	0.008	0.008	0.006	0.006	0.007	0.002	0.003	0.006	0.007	0.007	0.007	0.007	0.007	0.008	0.005	0.005	0.006
	RSEM	0.016	0.017	0.018	0.013	0.013	0.014	0.001	0.001	0.001	0.017	0.026	0.019	0.017	0.017	0.018	0.008	0.008	0.018	0.018	0.019	0.019	0.018	0.019	0.019	0.014	0.014	0.015
	Sailfish	0.030	0.035	0.038	0.023	0.026	0.029	0.005	0.007	0.008	0.033	0.043	0.041	0.031	0.035	0.038	0.015	0.017	0.026	0.033	0.038	0.041	0.033	0.037	0.040	0.025	0.029	0.031
FP Isoforms 0-1 FPKM	Cufflinks STAR	0.034	0.034	0.034	0.024	0.024	0.024	0.003	0.003	0.003	0.035	0.037	0.036	0.032	0.031	0.033	0.014	0.014	0.019	0.034	0.035	0.035	0.036	0.036	0.038	0.028	0.028	0.028
	Cufflinks TopHat	0.036	0.037	0.038	0.026	0.026	0.027	0.003	0.003	0.003	0.039	0.042	0.040	0.035	0.035	0.036	0.015	0.015	0.020	0.038	0.039	0.039	0.040	0.041	0.042	0.030	0.030	0.031
	eXpress	0.035	0.035	0.038	0.023	0.024	0.026	0.003	0.003	0.003	0.036	0.039	0.041	0.034	0.034	0.037	0.013	0.014	0.019	0.037	0.038	0.039	0.037	0.039	0.042	0.027	0.028	0.030
	RSEM	0.097	0.099	0.102	0.076	0.130	0.079	0.007	0.008	0.009	0.106	0.112	0.111	0.099	0.100	0.104	0.046	0.047	0.058	0.106	0.110	0.113	0.106	0.109	0.111	0.082	0.083	0.086
	Sailfish	0.139	0.148	0.157	0.106	0.143	0.119	0.026	0.033	0.035	0.160	0.163	0.170	0.143	0.149	0.158	0.067	0.075	0.086	0.147	0.161	0.170	0.150	0.160	0.174	0.117	0.128	0.138
FP Isoforms 1-5 FPKM	Cufflinks STAR	0.003	0.003	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.002	0.004	0.002	0.002	0.002	0.002	0.001	0.001	0.006	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	Cufflinks TopHat	0.005	0.005	0.005	0.003	0.003	0.004	0.001	0.001	0.001	0.005	0.006	0.004	0.005	0.004	0.004	0.002	0.002	0.006	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004
	eXpress	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.000	0.005	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.000	0.001
	RSEM	0.006	0.006	0.007	0.004	0.005	0.006	0.000	0.000	0.000	0.005	0.016	0.007	0.004	0.005	0.007	0.002	0.002	0.019	0.005	0.005	0.007	0.005	0.006	0.007	0.004	0.004	0.005
	Sailfish	0.016	0.029	0.043	0.012	0.019	0.031	0.002	0.007	0.008	0.025	0.040	0.049	0.016	0.031	0.046	0.009	0.018	0.036	0.018	0.031	0.048	0.017	0.032	0.057	0.012	0.025	0.039
FP Isoforms 5-10 FPKM	Cufflinks STAR	0.003	0.003	0.003	0.002	0.002	0.002	0.001	0.001	0.001	0.003	0.004	0.002	0.003	0.002	0.003	0.001	0.001	0.004	0.003	0.003	0.002	0.003	0.003	0.002	0.002	0.001	0.002
	Cufflinks TopHat	0.003	0.003	0.003	0.003	0.003	0.002	0.001	0.001	0.001	0.003	0.004	0.003	0.004	0.004	0.003	0.001	0.001	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.002	0.002
	eXpress	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RSEM	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.001	0.013	0.002	0.002	0.001	0.002	0.000	0.000	0.010	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.001
	Sailfish	0.006	0.010	0.015	0.004	0.007	0.011	0.001	0.001	0.002	0.009	0.017	0.018	0.006	0.011	0.015	0.003	0.005	0.017	0.008	0.009	0.015	0.007	0.010	0.019	0.006	0.008	0.015

0.000
0.010
0.020
0.030
0.040
0.050

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Isoforms 10-50 FPKM	Cufflinks STAR	0.002	0.002	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.002	0.003	0.002	0.002	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
	Cufflinks TopHat	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.003	0.002	0.002	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	eXpress	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RSEM	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Sailfish	0.002	0.003	0.004	0.002	0.002	0.003	0.000	0.000	0.000	0.003	0.008	0.005	0.002	0.004	0.005	0.001	0.002	0.007	0.003	0.004	0.005	0.003	0.004	0.006	0.002	0.002	0.004
FP Isoforms 50-100 FPKM	Cufflinks STAR	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001
	Cufflinks TopHat	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.002	0.002	0.002	0.001	0.001	0.001	0.000	0.000	0.001	0.002	0.002	0.001	0.002	0.002	0.001	0.001	0.001	0.001
	eXpress	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RSEM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Sailfish	0.001	0.002	0.002	0.001	0.001	0.002	0.000	0.000	0.000	0.002	0.005	0.003	0.002	0.003	0.003	0.001	0.002	0.004	0.002	0.002	0.002	0.001	0.002	0.004	0.001	0.001	0.002
FP Isoforms 100-500 FPKM	Cufflinks STAR	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001
	Cufflinks TopHat	0.001	0.001	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001
	eXpress	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RSEM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Sailfish	0.001	0.002	0.002	0.001	0.001	0.002	0.000	0.000	0.000	0.001	0.004	0.003	0.002	0.003	0.003	0.001	0.001	0.004	0.003	0.003	0.003	0.002	0.002	0.003	0.001	0.001	0.002
FP Isoforms >500 FPKM	Cufflinks STAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.001	0.001	0.000	0.000	0.000
	Cufflinks TopHat	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000	0.002	0.000	0.000	0.000	0.002	0.001	0.001	0.000	0.000	0.000
	eXpress	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	RSEM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Sailfish	0.003	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.002	0.006	0.002	0.001	0.002	0.002	0.000	0.000	0.004	0.001	0.002	0.002	0.003	0.004	0.004	0.002	0.002	0.002

Figure 2.22: Fraction of false positive isoforms a function of gene expression levels. The fraction of false positive isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and at varying gene-level expression cutoffs. A false positive isoform is defined as a transcript with true $\Theta = 0$ and estimated $\Theta > 0.05$.

A

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Isoforms all	Cufflinks STAR	0.007	0.007	0.008	0.005	0.005	0.006	0.001	0.001	0.001	0.007	0.009	0.008	0.007	0.007	0.007	0.003	0.003	0.006	0.007	0.008	0.008	0.008	0.008	0.008	0.006	0.006	0.006
	Cufflinks TopHat	0.008	0.009	0.009	0.006	0.006	0.006	0.001	0.001	0.001	0.009	0.010	0.009	0.008	0.008	0.009	0.004	0.004	0.007	0.009	0.009	0.009	0.009	0.009	0.010	0.007	0.007	0.007
	eXpress	0.006	0.007	0.007	0.004	0.006	0.005	0.001	0.001	0.001	0.007	0.008	0.008	0.006	0.006	0.007	0.002	0.003	0.006	0.007	0.007	0.007	0.007	0.007	0.008	0.005	0.005	0.006
	RSEM	0.016	0.017	0.018	0.013	0.013	0.014	0.001	0.001	0.001	0.017	0.026	0.019	0.017	0.017	0.018	0.008	0.008	0.018	0.018	0.019	0.019	0.018	0.019	0.019	0.014	0.014	0.015
	Sailfish	0.030	0.035	0.038	0.023	0.026	0.029	0.005	0.007	0.008	0.033	0.043	0.041	0.031	0.035	0.038	0.015	0.017	0.026	0.033	0.038	0.041	0.033	0.037	0.040	0.025	0.029	0.031
FP Isoforms 2 Isoforms	Cufflinks STAR	0.029	0.029	0.031	0.030	0.031	0.032	0.006	0.006	0.006	0.033	0.039	0.036	0.034	0.035	0.037	0.024	0.024	0.029	0.036	0.035	0.035	0.028	0.029	0.032	0.037	0.037	0.037
	Cufflinks TopHat	0.031	0.032	0.033	0.032	0.033	0.035	0.008	0.008	0.008	0.036	0.041	0.038	0.037	0.038	0.038	0.024	0.025	0.030	0.038	0.038	0.038	0.031	0.032	0.033	0.037	0.038	0.038
	eXpress	0.021	0.020	0.021	0.020	0.020	0.021	0.004	0.004	0.004	0.024	0.029	0.027	0.025	0.026	0.027	0.015	0.015	0.020	0.027	0.026	0.027	0.019	0.021	0.023	0.027	0.027	0.027
	RSEM	0.056	0.054	0.058	0.065	0.067	0.068	0.007	0.008	0.008	0.058	0.080	0.061	0.059	0.060	0.061	0.066	0.066	0.098	0.062	0.061	0.062	0.054	0.057	0.055	0.062	0.062	0.066
	Sailfish	0.067	0.070	0.077	0.077	0.080	0.082	0.015	0.022	0.027	0.071	0.094	0.080	0.070	0.075	0.080	0.070	0.072	0.094	0.067	0.076	0.082	0.071	0.076	0.077	0.072	0.079	0.082
FP Isoforms 3-4 Isoforms	Cufflinks STAR	0.025	0.025	0.026	0.017	0.017	0.018	0.002	0.002	0.002	0.022	0.023	0.022	0.023	0.023	0.024	0.009	0.009	0.013	0.024	0.024	0.022	0.024	0.024	0.025	0.019	0.020	0.020
	Cufflinks TopHat	0.028	0.030	0.030	0.020	0.020	0.021	0.002	0.002	0.002	0.027	0.028	0.027	0.026	0.027	0.027	0.011	0.010	0.015	0.028	0.028	0.028	0.029	0.028	0.029	0.024	0.024	0.025
	eXpress	0.021	0.023	0.023	0.015	0.015	0.015	0.002	0.001	0.002	0.020	0.021	0.021	0.022	0.022	0.023	0.008	0.007	0.013	0.023	0.022	0.022	0.022	0.023	0.024	0.017	0.017	0.019
	RSEM	0.057	0.058	0.062	0.052	0.051	0.050	0.004	0.004	0.005	0.054	0.073	0.056	0.056	0.058	0.060	0.029	0.030	0.053	0.058	0.061	0.062	0.057	0.058	0.060	0.053	0.053	0.054
	Sailfish	0.079	0.088	0.094	0.067	0.071	0.076	0.014	0.017	0.020	0.079	0.099	0.095	0.083	0.090	0.096	0.044	0.048	0.062	0.082	0.091	0.098	0.081	0.090	0.097	0.073	0.079	0.085
FP Isoforms 5-7 Isoforms	Cufflinks STAR	0.012	0.012	0.012	0.008	0.008	0.009	0.001	0.001	0.001	0.013	0.015	0.014	0.011	0.011	0.012	0.005	0.005	0.009	0.012	0.013	0.014	0.013	0.013	0.014	0.009	0.010	0.010
	Cufflinks TopHat	0.013	0.014	0.014	0.009	0.009	0.010	0.001	0.001	0.001	0.015	0.016	0.016	0.013	0.014	0.014	0.005	0.005	0.009	0.015	0.015	0.015	0.015	0.015	0.017	0.011	0.011	0.011
	eXpress	0.011	0.011	0.013	0.007	0.008	0.009	0.001	0.001	0.001	0.013	0.014	0.014	0.011	0.011	0.011	0.003	0.004	0.008	0.012	0.013	0.013	0.012	0.013	0.014	0.008	0.009	0.010
	RSEM	0.031	0.032	0.033	0.022	0.023	0.025	0.002	0.002	0.002	0.033	0.046	0.036	0.031	0.033	0.034	0.012	0.013	0.027	0.034	0.035	0.037	0.033	0.034	0.034	0.024	0.025	0.026
	Sailfish	0.054	0.060	0.063	0.039	0.043	0.046	0.008	0.011	0.012	0.059	0.071	0.068	0.055	0.060	0.063	0.024	0.027	0.040	0.058	0.065	0.069	0.056	0.063	0.066	0.044	0.048	0.051

0.000
0.010
0.020
0.030
0.040
0.050

B

	μ	0.25									0.5									0.75								
	α	0.5			1			4			0.5			1			4			0.5			1			4		
	IF	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25	0.05	0.15	0.25
FP Isoforms 8-10 Isoforms	Cufflinks STAR	0.008	0.008	0.008	0.006	0.006	0.006	0.001	0.001	0.001	0.007	0.009	0.008	0.007	0.007	0.007	0.003	0.003	0.007	0.008	0.008	0.008	0.008	0.009	0.009	0.006	0.006	0.007
	Cufflinks TopHat	0.010	0.010	0.010	0.006	0.006	0.007	0.001	0.001	0.001	0.009	0.011	0.010	0.008	0.008	0.008	0.004	0.004	0.007	0.009	0.009	0.010	0.009	0.010	0.010	0.007	0.007	0.008
	eXpress	0.008	0.008	0.009	0.005	0.005	0.006	0.001	0.001	0.001	0.008	0.010	0.010	0.007	0.007	0.008	0.003	0.003	0.006	0.008	0.008	0.009	0.008	0.009	0.009	0.006	0.006	0.006
	RSEM	0.018	0.019	0.019	0.013	0.013	0.014	0.001	0.002	0.002	0.020	0.030	0.022	0.017	0.018	0.019	0.007	0.008	0.019	0.019	0.020	0.020	0.020	0.021	0.022	0.014	0.015	0.016
	Sailfish	0.035	0.040	0.045	0.025	0.028	0.033	0.006	0.009	0.009	0.039	0.051	0.049	0.034	0.039	0.044	0.017	0.019	0.028	0.038	0.044	0.049	0.038	0.044	0.049	0.027	0.032	0.036
FP Isoforms 11-15 Isoforms	Cufflinks STAR	0.004	0.004	0.005	0.002	0.002	0.003	0.001	0.001	0.001	0.004	0.006	0.004	0.004	0.004	0.005	0.002	0.002	0.004	0.004	0.004	0.005	0.005	0.005	0.005	0.003	0.003	0.003
	Cufflinks TopHat	0.004	0.005	0.005	0.003	0.003	0.003	0.001	0.001	0.001	0.005	0.006	0.005	0.005	0.005	0.005	0.002	0.002	0.004	0.005	0.005	0.006	0.006	0.006	0.006	0.004	0.004	0.004
	eXpress	0.003	0.003	0.004	0.002	0.002	0.003	0.000	0.000	0.000	0.003	0.004	0.004	0.003	0.003	0.004	0.001	0.001	0.004	0.003	0.004	0.004	0.003	0.004	0.004	0.003	0.003	0.003
	RSEM	0.008	0.008	0.009	0.006	0.006	0.007	0.001	0.001	0.001	0.010	0.017	0.011	0.010	0.010	0.010	0.003	0.004	0.010	0.010	0.010	0.010	0.011	0.011	0.012	0.007	0.007	0.007
	Sailfish	0.021	0.025	0.026	0.016	0.018	0.020	0.004	0.005	0.005	0.023	0.032	0.030	0.022	0.025	0.028	0.009	0.011	0.018	0.023	0.028	0.030	0.024	0.028	0.031	0.017	0.020	0.022
FP Isoforms 16-20 Isoforms	Cufflinks STAR	0.003	0.003	0.003	0.002	0.002	0.003	0.000	0.001	0.000	0.004	0.006	0.005	0.003	0.003	0.003	0.002	0.001	0.004	0.004	0.004	0.004	0.005	0.005	0.005	0.002	0.003	0.003
	Cufflinks TopHat	0.004	0.004	0.004	0.003	0.003	0.003	0.001	0.001	0.001	0.005	0.006	0.005	0.004	0.004	0.004	0.002	0.002	0.005	0.004	0.004	0.004	0.005	0.005	0.005	0.003	0.003	0.003
	eXpress	0.002	0.002	0.002	0.001	0.002	0.002	0.000	0.000	0.000	0.002	0.003	0.003	0.002	0.002	0.003	0.001	0.001	0.004	0.002	0.003	0.003	0.003	0.004	0.004	0.001	0.002	0.002
	RSEM	0.006	0.006	0.007	0.004	0.004	0.005	0.001	0.001	0.001	0.007	0.012	0.008	0.006	0.006	0.007	0.003	0.003	0.009	0.007	0.008	0.008	0.008	0.008	0.008	0.005	0.005	0.005
	Sailfish	0.015	0.017	0.018	0.010	0.012	0.013	0.003	0.003	0.004	0.017	0.023	0.021	0.015	0.017	0.018	0.007	0.008	0.013	0.017	0.020	0.022	0.017	0.018	0.020	0.011	0.013	0.014
FP Isoforms >20 Isoforms	Cufflinks STAR	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.003	0.001	0.001	0.002	0.001	0.002	0.002	0.001	0.001	0.001
	Cufflinks TopHat	0.002	0.002	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.002	0.003	0.002	0.002	0.002	0.002	0.001	0.001	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
	eXpress	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.003	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001
	RSEM	0.002	0.003	0.003	0.002	0.002	0.002	0.000	0.000	0.000	0.003	0.006	0.004	0.002	0.003	0.003	0.001	0.001	0.005	0.003	0.003	0.003	0.002	0.003	0.003	0.002	0.002	0.002
	Sailfish	0.009	0.011	0.013	0.007	0.008	0.009	0.002	0.002	0.002	0.010	0.015	0.015	0.010	0.012	0.014	0.004	0.005	0.009	0.011	0.013	0.014	0.010	0.012	0.013	0.007	0.009	0.010

Figure 2.23: Fraction of genes positive isoforms a function of annotated isoform complexity. The fraction of false positive isoforms in the output of different quantification programs was calculated, at varying values of the μ, α and IF parameters and for genes with different number of isoforms annotated in GENCODE V16. A false positive isoform is defined as a transcript with true $\Theta = 0$ and estimated $\Theta > 0.05$.

simulated locus-level FPKMs. The correlation was mostly very high – 0.97 for most settings of the isoform complexity and IF parameters and for most programs – except for Sailfish FPKMs, which exhibited a consistently poorer correlation with the true values. Sailfish also displayed the highest sensitivity to the increased presence of intronic reads, with correlation dropping by between 0.03 and 0.06 from IF = 0.05 to IF == 0.25. For the other programs, however, this effect was much more modest, usually less than 0.01 decrease in correlation.

Figure 2.9 shows the Pearson correlation between the estimated and true FPKMs for individual transcripts. Sailfish was once again the worst performing program, and it was once again most sensitive to the intronic fraction parameter; this was also true for all other comparison metrics, thus for the rest of this exposition I will focus on the other four options, without mentioning Sailfish specifically. Overall, the correlation between the estimated and true values on the transcript level was significantly worse than that for locus-level quantification, usually being between 0.8 and 0.9. It was also significantly more sensitive to the intronic fraction of reads, dropping by as much as 0.10 in some cases when going from IF = 0.05 to IF = 0.25. Some difference between the different programs were apparent. Cufflinks on STAR alignments performed consistently better than Cufflinks on TopHat alignments, in curious contrast to the slight advantage the latter had in assembly; however, these are not mutually exclusive possibilities. But in both cases, Cufflinks was outperformed by both eXpress and RSEM, with RSEM producing slightly better results than eXpress. For example, at $\alpha = 0.5$, $\mu = 0.25$ and IF = 0.05, the correlation was 0.92 for RSEM, 0.9 for eXpress, 0.85 for TopHat+Cufflinks, and 0.87 for STAR+Cufflinks. A counterintuitive observation was that the correlation decreased when the α parameter was increased, i.e. when the isoform complexity decreased. This is most likely explained by false positive FPKM values being generated for transcripts that are in fact either not expressed or expressed at relatively low levels.

I next examined how many lincRNAs and pseudogenes received positive FPKM values (Figures 2.10 and 2.11, respectively). As previously mentioned, only protein coding genes were included in the simulation; therefore, all lincRNAs and pseudogenes should have received 0 FPKMs, and the extent to which this is not the case provides useful insight into the performance of the different programs (it is also of interest with respect to the reliability of results concerning lincRNAs and pseudogenes that are based on the output of these programs). Sailfish was a clear outlier in this comparison, quantifying many lincRNAs and even more pseudogenes (up to half) as "expressed", sometimes at quite high levels. Smaller in magnitude, but still substantial in some cases differences were observed between the other programs too. Cufflinks quantified fewer lincRNAs and pseudogenes as expressed when run on STAR alignments than it did when TopHat alignments was used as input. STAR+Cufflinks was the best performing combination with respect to lincRNAs, while RSEM and eXpress had the fewest false positives when pseudogenes were considered. The significance of these differences is elaborated on in the Discussion section.

The distribution of the mean total Θ difference was examined next, for all transcripts, and across the range of FPKM values (Figure 2.12) and annotation complexity (Figure 2.13). Once again, overall, RSEM and eXpress outperformed Cufflinks, RSEM produced slightly better results than eXpress, as did Cufflinks on STAR alignments compared to Cufflinks on TopHat alignments. However, some interesting patterns were observed. Quantification was quite reliable when the isoform expression complexity was low – for example, at $\alpha = 4$, $\mu = 0.25$ and IF = 0.05, the $MT\Theta_{diff}$ values for RSEM and eXpress were 0.13 and 0.14, respectively – but with increased isoform expression complexity, performance deteriorated significantly – the RSEM and eXpress $MT\Theta_{diff}$ values at $\alpha = 0.5$, $\mu = 0.75$ and IF = 0.05, they were 0.50 and 0.49. A striking pattern was observed when the relationship between gene expression levels and the ability to accurately parse reads between isoforms was examined: the higher the gene-level FPKMs, the worse Cufflinks's performance was, while in contrast $MT\Theta_{diff}$ values for RSEM and eXpress either remained constant or decreased (Figure 2.12). This is not entirely surprising given the way Cufflinks's likelihood optimization proceeds (Trapnell et al. 2010; and C. Trapnell, personal communication), but was nevertheless an intriguing observation. Cufflinks's performance also deteriorated consistently with the increase in isoform complexity in the annotation (Figure 2.12), but the pattern observed for RSEM and eXpress was more complicated. In the case of $\alpha = 4$, $\mu = 0.25$ and IF = 0.05, i.e. the lowest isoform expression complexity, the $MT\Theta_{diff}$ actually went down for both, from 0.18 for eXpress and 0.16 for RSEM for genes with 2 annotated isoforms, to 0.12 for eXpress and 0.11 for RSEM for genes with >20 annotated isoforms. However, in the simulated libraries with high isoform expression complexity, the $MT\Theta_{diff}$ values increased for genes with more annotated isoforms.

The fraction of genes with an incorrectly assigned major isoform is shown in Figures 2.14 (as a function of gene expression levels) and 2.15 (as a function of annotation complexity). It was highly dependent on the isoform expression complexity of the simulated libraries: for example, RSEM assigned the wrong isoform 7% of the time when $\alpha = 4$, $\mu = 0.25$ and IF = 0.05, but it did so for 35% of genes when $\alpha = 0.5$, $\mu = 0.75$ and IF = 0.05. The relative performance of the programs according to this metric was very similar to that revealed by the $MT\Theta_{diff}$. RSEM and eXpress produced better results than Cufflinks, and STAR+Cufflinks was more reliable than TopHat+Cufflinks. Cufflinks was once again performing worse on more highly expressed genes (though in this case, this was the group of genes expressed in the $(100, 500]$ FPKM range and not the highest expressed genes, the $\geq$500 FPKM ones, where the worst values were observed). Annotation complexity had a significant negative effect in the samples expressing a complex mixture of isoforms (but, notably, not so much in the ones where $\alpha = 4$): 24–26% of genes with 2 annotated isoforms had an incorrectly assigned major isoform by RSEM in $\alpha = 4$ libraries, but this number rose to 44-48% for genes with $\geq$20 annotated isoforms.

Figures 2.16 and 2.17 show the fraction of genes with false negative isoforms in the various quantification sets as a function of gene expression levels and annotation complexity, Figures 2.18 and 2.19 show the fraction of false negative isoforms among all transcripts, while Figures 2.20, 2.21, 2.22, and 2.23 show the corresponding values for false positive isoforms (how false positive and false negative isoforms are defined is described in the Methods section). False positive isoforms were generally rare, except for genes expressed at very low levels, while false negatives were a considerably more common occurrence. The relative performance of the programs followed the patterns established by the previous metrics, with one notable difference. RSEM returned consistently fewer false negatives than eXpress did, but it also generated more false positives than observed in eXpress quantifications.

2.4 Discussion

The results of this simulation highlight the deficiencies of current transcriptomic measurement and analysis practices and also inform the interpretation of the results presented in the previous chapter. On a most general level, the conclusion is that splice junction detection and discovery work relatively well, as is the case for gene-level quantification. However, isoform assembly and isoform-level quantification not only remain unresolved problems, but are in fact likely unsolvable computationally as long as the nature of the underlying data remains the same. Of note, similar in their nature conclusions were reached by the RGASP (RNA-seq Genome Annotation Assessment Project; Steijger et al. 2013; Engström et al. 2013) initiative, which also used simulations in addition to real RNA-seq datasets to evaluate the performance of RNA-seq mapping and transcript assembly and genome annotation software.

While the simulation was simplistic and did not present much of a challenge with respect to splice junction discovery, the fact that so few false positive junctions were returned is encouraging. However, the problem of assembling transcripts *de novo* was not solved in satisfactory way by any of the programs tested. Except for the simples cases, in which only one isoform was expressed for most genes, up to 40% of assembled transcripts were false positives and up to 60% of the expressed transcripts were false negatives. Given that the simulation did not model real-life complicating factors such as the nonuniformity of sequence coverage, it is likely that results on real RNA-seq data are in fact even worse. Another important insight gained from the simulation was that genome-free assembly approaches are extremely sensitive to the fraction of intronic reads present in the sample, producing many more false positive and/or partially assembled transcripts when the IF parameter was increased from 0.05 to 0.25, something of significant importance for the practice of transcriptomic analysis in the absence of a corresponding genomic assembly – to the best knowledge of the author, little attention has been paid to the issue so far in such cases, but even if that was not the case, the intronic fraction is almost impossible to measure without an assembled and annotated genome, presenting a difficult to resolve conundrum. In any case, assembling the genome and then carrying out transcript reconstruction is clearly the better option (if, of course, such a choice is available) by a significant margin.

Isoform-level quantification is also not quite up to the desired level. As with assembly, it is highly likely that performance on real-life datasets is worse than what was observed on simulated data. But even in the simplified simulation, in all of the complex samples ($\mu = 0.5$ or $\mu = 0.75$, and $\alpha = 0.5$ or $\alpha = 1$), 1/3 of genes had an incorrectly assigned major isoform, and this rose to ∼50% for genes with a large number of annotated isoforms. Some clear difference between the programs emerged. Sailfish, which uses k-mer frequencies instead of alignments provided the most unreliable sets of FPKMs; this is not surprising as the naive expectation is that the problem of parsing k-mers between genes and transcripts would be considerably more difficult to solve than the problem of doing the same with alignments. In this context, it is also not surprising that Sailfish generated so many false positives in pseudogenes and lincRNAs. It is possible that different values of k than the default value used here will generate better results, but it is unlikely that they will ever reach the performance of the alignment-based approaches. A bit more surprising was the fact that the transcriptome-space programs performed better than Cufflinks, but this in fact makes sense given the nature of the alignments each such programs is presented with and how they affect their output. Cufflinks works with alignments in genomic coordinates and does not really "see" all alignments a read might have to other genes, even though such reads are recognized as multireads and treated accordingly. In contrast, programs like eXpress stream reads and directly weigh alignments between all places in the transcriptome they map to; this leads to better parsing of reads between paralogous genes and fewer false positives due to mapping issues, though their performance with respect only to the isoforms of protein coding genes also seems to be superior. Based on the simulation results, working in transcriptome space should be the preferred approach if reliability of output is the top priority of the analysis. However, it has to remembered that even eXpress and RSEM do not completely solve the problem and only provide a marginal in comparison to the deviation from the truth improvement over Cufflinks.

These results are not surprising given the background of a wide variety of accumulated anecdotal examples of questionable quantification output, but they do present an explicit illustration of the magnitude of the problem. They also provide some context for the interpretation of real-life results if we are willing to engage in some Bayesian reasoning: simulated samples with low complexity of expressed isoforms consistently returned results closest to the ground truth, while samples with high isoform complexity fared the worst. This means that if a major isoform is observed with minor isoforms with very low FMI values, then it is significantly more likely that the quantification output represents the underlying biochemical reality well. Conversely, if multiple isoforms are scored as expressed at high level, it is much more difficult to tell whether their ranking is correct, which one is the major isoform, and by how much. These considerations are of major importance to the question of how much regulated alternative isoform switching happens between different cell types; unfortunately, a major fraction of putative such events belong to the second category, making any definitive conclusions about the phenomenon difficult to defend.

The simulation also once again confirmed the theoretical expectation that the complexity of the annotation being quantified has a significant negative effect on the output of the quantification: the more isoforms there are in the annotation, the more likely it is that the maximum likelihood model becomes unidentifiable. The unfolding of a quite unsettling scenario is thus entirely possible in the near future: as RNA-seq probes ever deeper into the complexity of the transcriptome, and genome annotations become updated to reflect that newly acquired knowledge, an ever higher fraction of genes in these annotations will become impossible to confidently quantify as they will contain too many isoforms that cannot be unambiguously distinguished from one another based on short reads. In the same time, while the length of reads keeps increasing, the length of the fragments they originate from cannot increase without introducing deeply problematic on their own biases in libraries (see discussion in the Methods section), meaning that the short-read sequencing format of RNA-seq cannot really go beyond

2x150bp. The only meaningful solution to these issues will be the advent of sequencing technologies that can produce full transcript-length reads at a sufficient sequencing depth (meaning tens of millions of reads). Such a technology will have to also achieve that without the limitations imposed by size selection that exist for current long-read platform such as PacBio (Sharon et al. 2013; Au et al. 2013) so that both short and very long transcripts are sequenced equally efficiently. It would be also highly desirable for it to perform direct RNA-sequencing, i.e. without the need to convert RNA into cDNA, as reverse transcription might be a significant source of biases (for example, due to the presence of secondary structures in RNA molecules, internal polyA priming sites if oilgo-dT priming is used, etc.). A technology that has the potential to deliver such a radical paradigm shift in the field is nanopore sequencing (Branton et al. 2008), even if the development of functional direct RNA sequencing based on it is still at least a few years into the future. Until then the analysis of alternative splicing and processing using RNA-seq at the level of whole transcripts and the whole transcriptome (as opposed to the targeted sequencing of individual genes and the analysis of localized splicing events) will remain a complicated and fraught with epistemological difficulties enterprise.

3

Single-cell RNA-seq in human lymphoblastoid cells: stochasticity in gene expression and RNA splicing

The majority of the material in this chapter was published as:

<u>Marinov GK*</u>, Williams BA*, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24**:496–510. doi: 10.1101/gr.161034

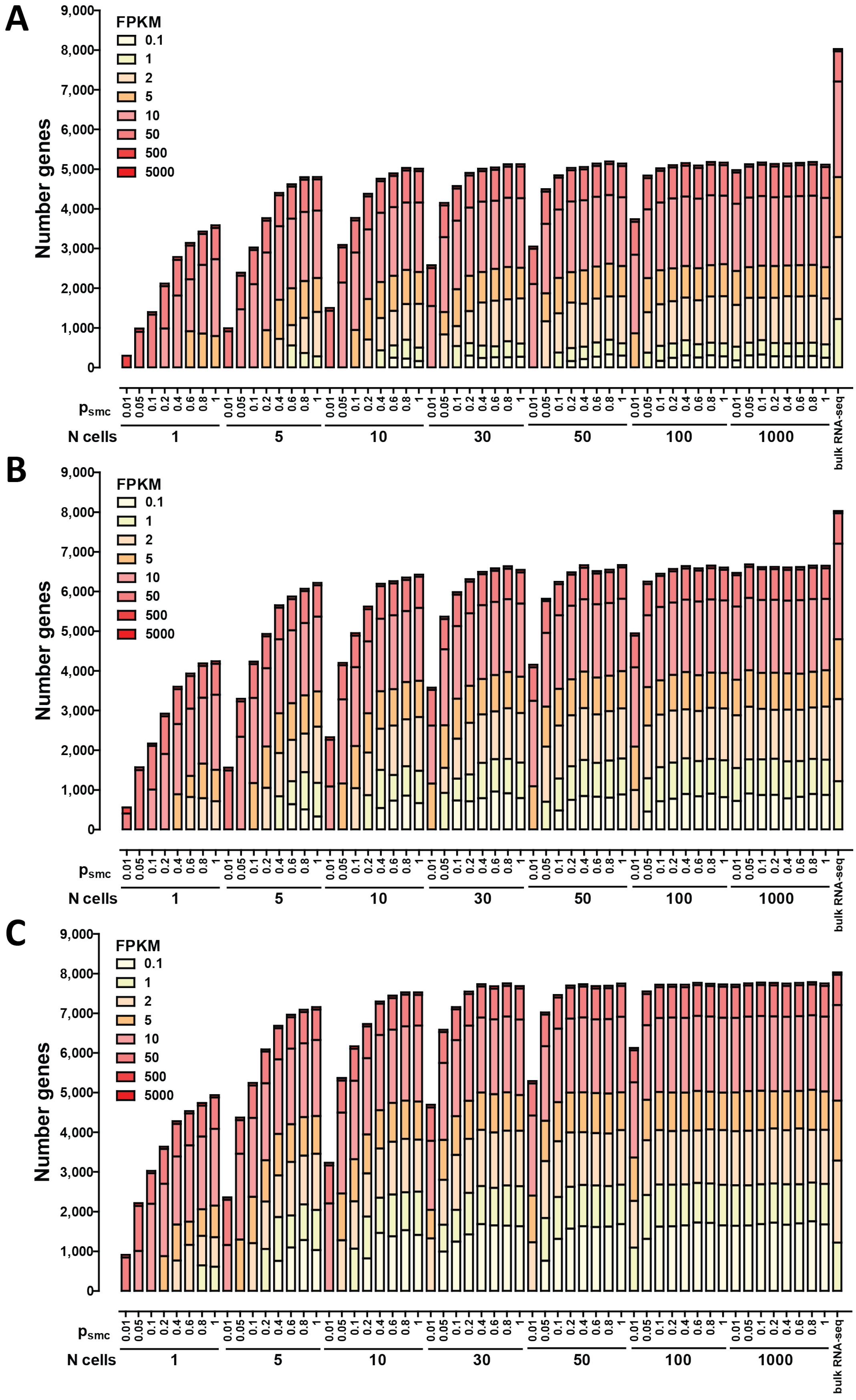
The paper is reprinted in Appendix J. The single-cell RNA-seq data on which it is based was generated by Brian Williams in the Wold lab. My contribution is the computational analytical framework for analysis and the analysis itself as well as some input into the experimental design.

Abstract

In this work, we applied the SMART-seq low-input RNA-seq protocol to study cell-to-cell variation in gene expression, alternative splicing and allelic bias in the reference lymphoblastoid cell line GM12878. We also identified and addressed the technical noise issues intrinsic to single-cell RNA-seq, by devising experimental and computational approaches to distinguish between biological and technical variation in measurements. By using spike-in quantification standards we estimated the absolute number of RNA molecules per cell for each gene and found significant variation in total mRNA content, between 50,000 to 300,000 transcripts per cell. We directly measured technical stochasticity by a pool/split design, and found that there are significant differences in expression between individual cells, over and above technical variation. We identified specific gene coexpression modules that were preferentially expressed in subsets of individual cells, including one enriched for mRNA processing and splicing factors. We assessed cell-to-cell variation in alternative splicing and allelic bias, and found evidence for significant differences in splice site usage between individual cells that exceed the observed variation in the pool/split comparison. We also found similar cell-to-cell differences in allelic bias suggesting widespread random monallelic expression, however, such differences were also observed (although at lower levels) in pool/splits and have to be considered a provisional result until further improvements in experimental protocols. Finally, we showed that transcriptomes from small pools of 30-100 cells approach the information content and reproducibility of RNA-seq from large amounts of input material.

3.1 Introduction

Gene expression levels can differ widely between superficially similar cells. One source of variation is stochastic transcriptional “bursting”


A
Number genes
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
FPKM
0.1
1
2
5
10
50
500
5000
p_{smc}
0.01 0.05 0.1 0.2 0.4 0.6 0.8 1
N cells
1
5
10
30
50
100
1000
bulk RNA-seq
B
Number genes
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
FPKM
0.1
1
2
5
10
50
500
5000
p_{smc}
0.01 0.05 0.1 0.2 0.4 0.6 0.8 1
N cells
1
5
10
30
50
100
1000
bulk RNA-seq
C
Number genes
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
FPKM
0.1
1
2
5
10
50
500
5000
p_{smc}
0.01 0.05 0.1 0.2 0.4 0.6 0.8 1
N cells
1
5
10
30
50
100
1000
bulk RNA-seq

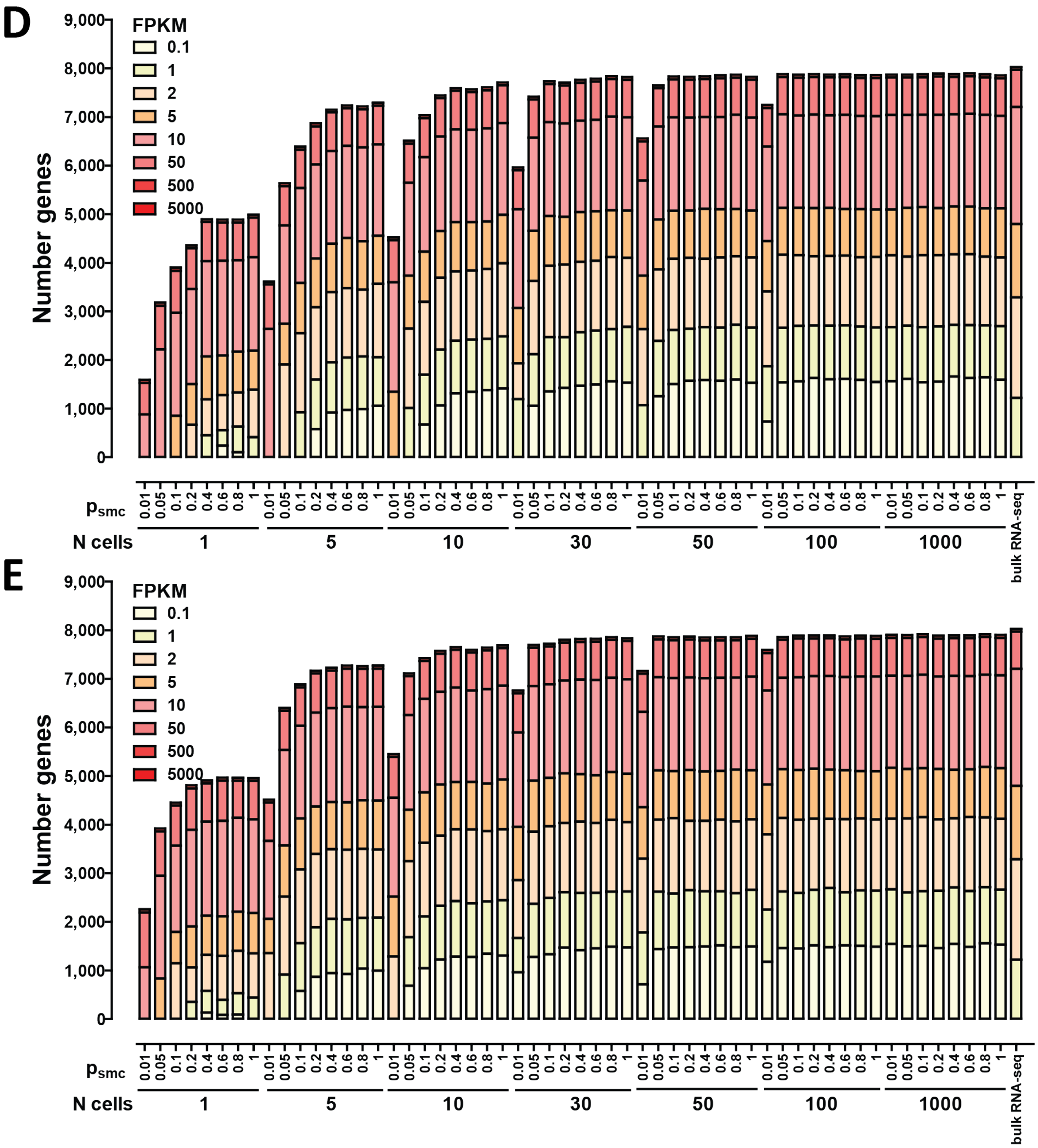


Figure 3.1: Detection of expressed genes in simulated datasets as a function of the single molecule capture efficiency, the number of cells and the average number of transcripts per cell. (A) Average of 50,000 mRNAs per cell. (B) Average of 100,000 mRNAs per cell. (C) Average of 200,000 mRNAs per cell. (D) Average of 500,000 mRNAs per cell. (E) Average of 1,000,000 mRNAs per cell. See the Methods section for full details on how the simulation was carried out.

(Elowitz et al. 2002; Ozbudak et al. 2002; Blake et al. 2003; Raser & O'Shea 2005; Kaufmann & van Oudenaarden 2007). Those studies mainly used fluorescent protein fusion genes to monitor the expression of one or a few genes. They revealed dynamic fluctuations through time that are seen as "salt-and-pepper" variation across a cell population at any given time. In addition to this bursting behavior, individual cells are expected to display controlled and coordinated differences in the expression of genes engaged in dynamic physiologic processes, such as cell cycle phase progression, paracrine or autocrine signaling response, or stress response.

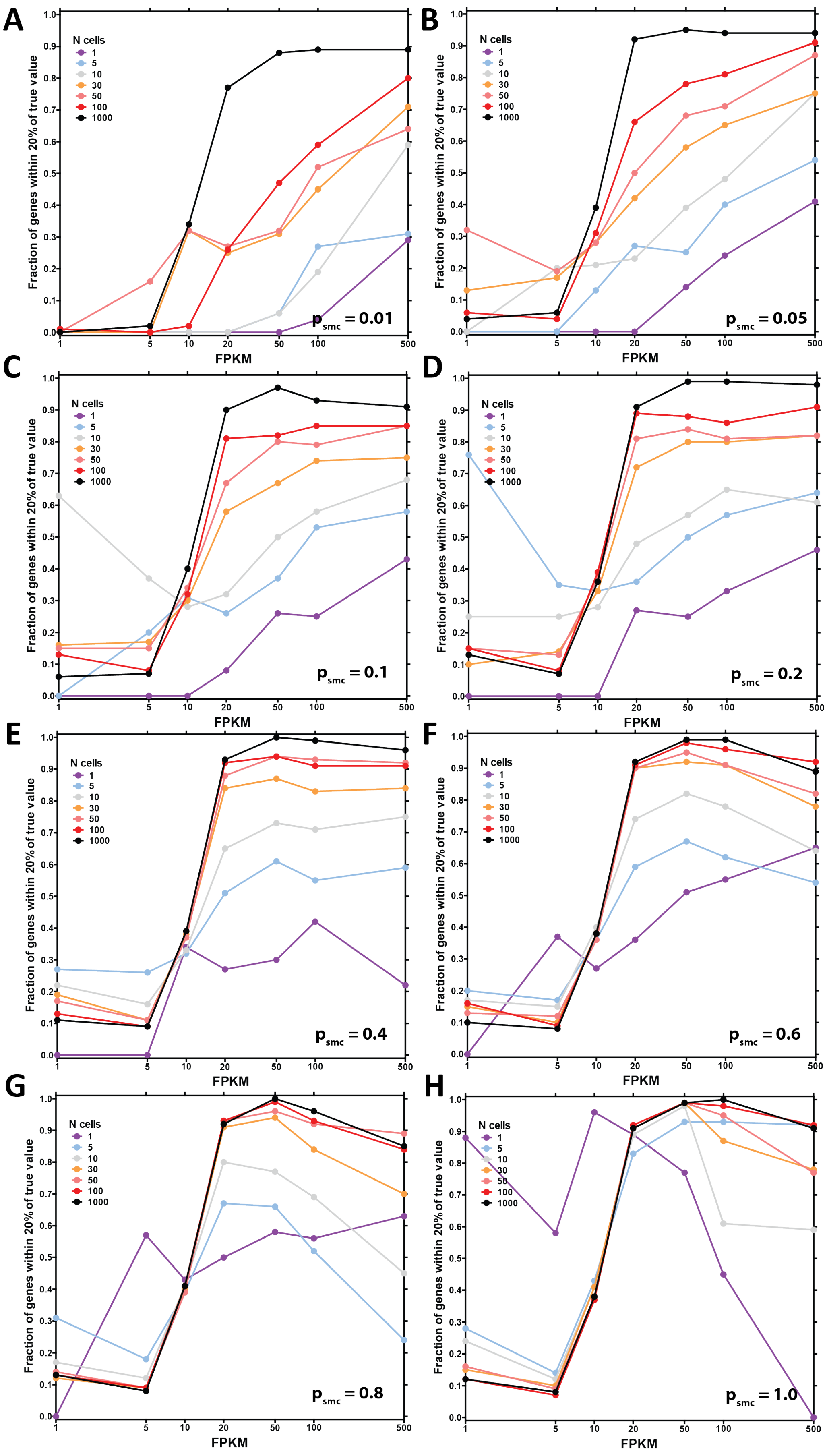
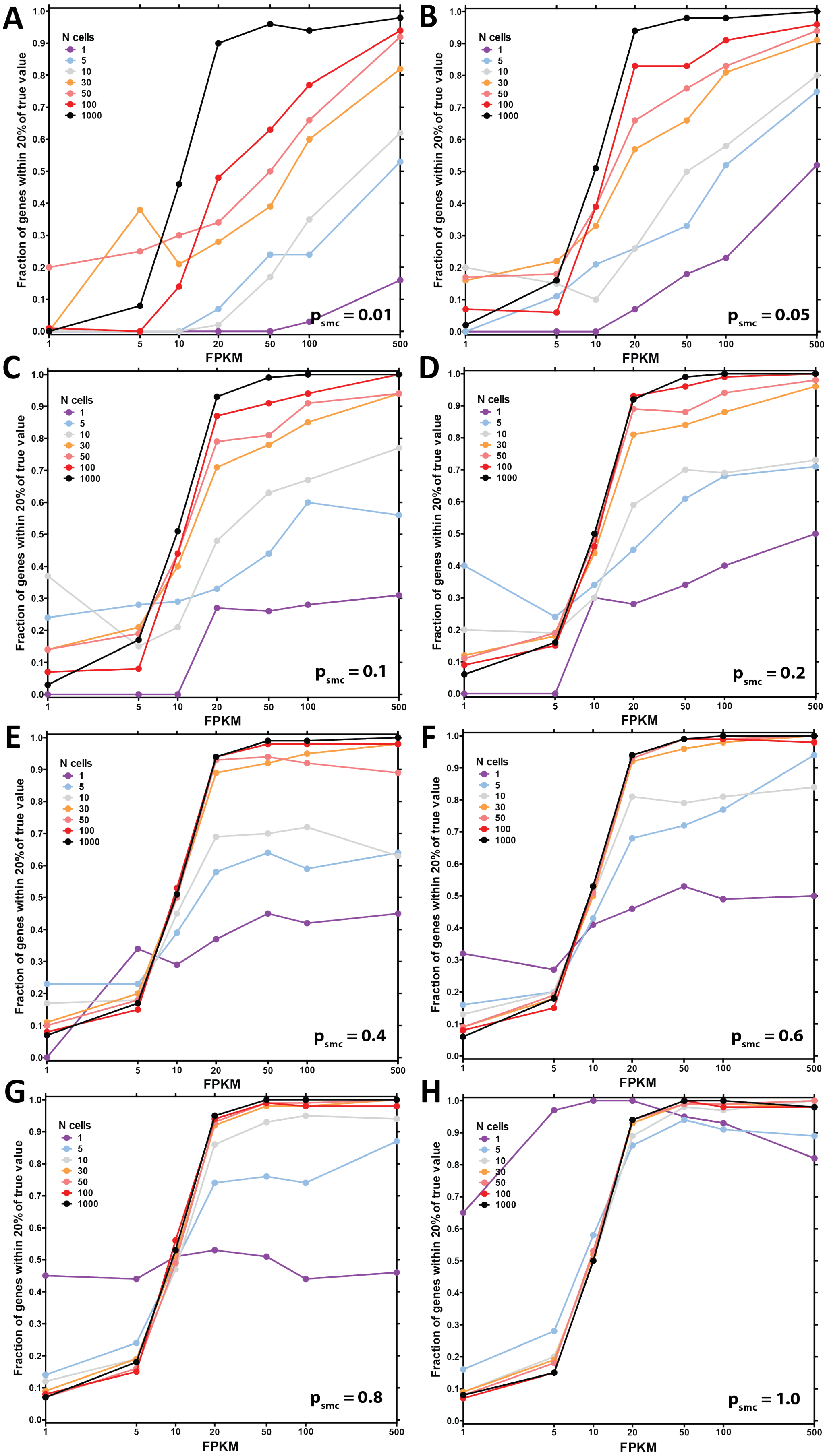

A
B
C
D
E
F
G
H
N cells
1
5
10
30
50
100
1000
Fraction of genes within 20% of true value
FPKM
1
5
10
20
50
100
500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

Figure 3.2 ***(preceding page)*****: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability**. Average of 50,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.

Figure 3.3: (following page) Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 100,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.


A
B
C
D
E
F
G
H
N cells
1
5
10
30
50
100
1000
Fraction of genes within 20% of true value
FPKM
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
5
10
20
50
100
500
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

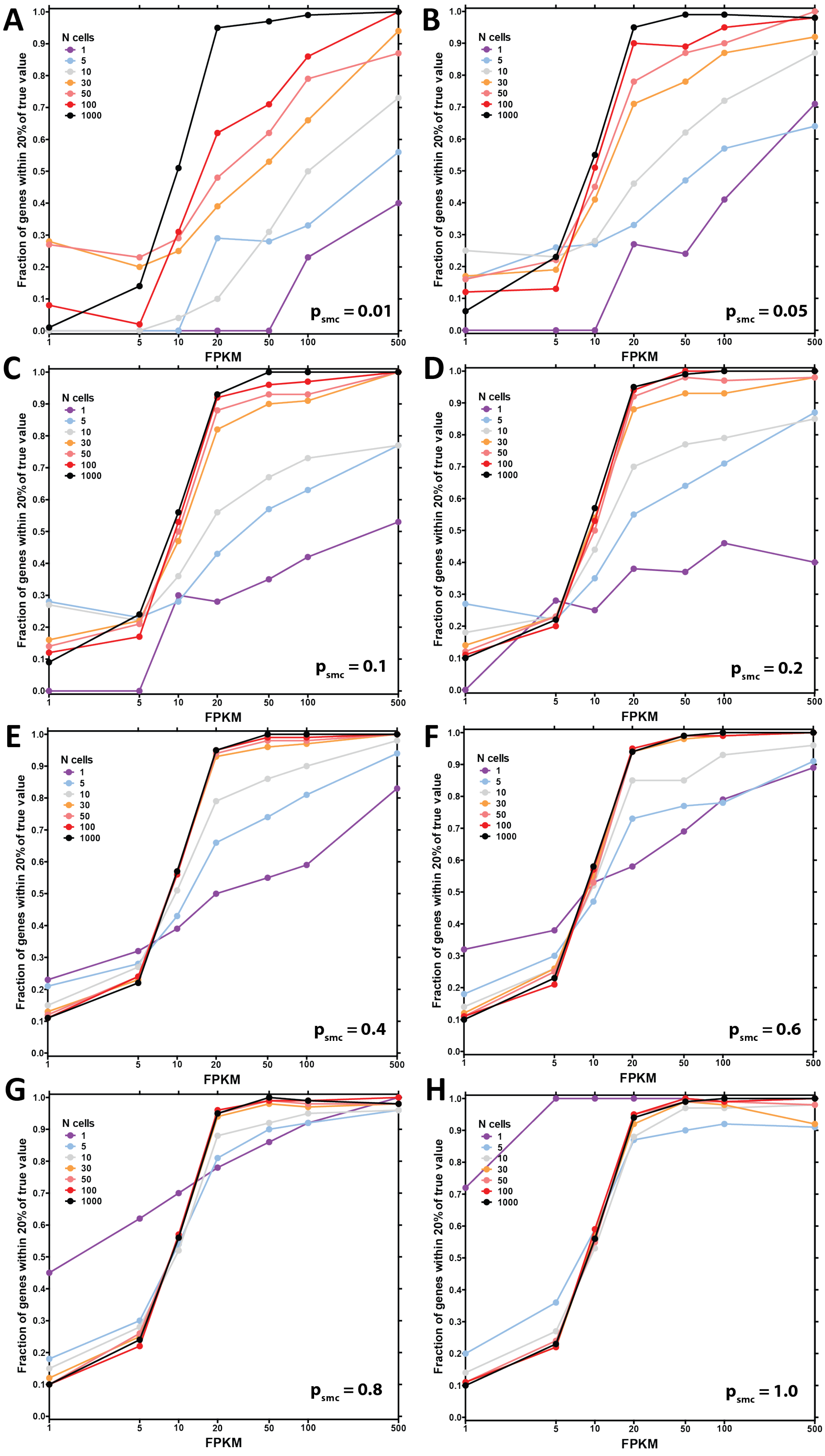
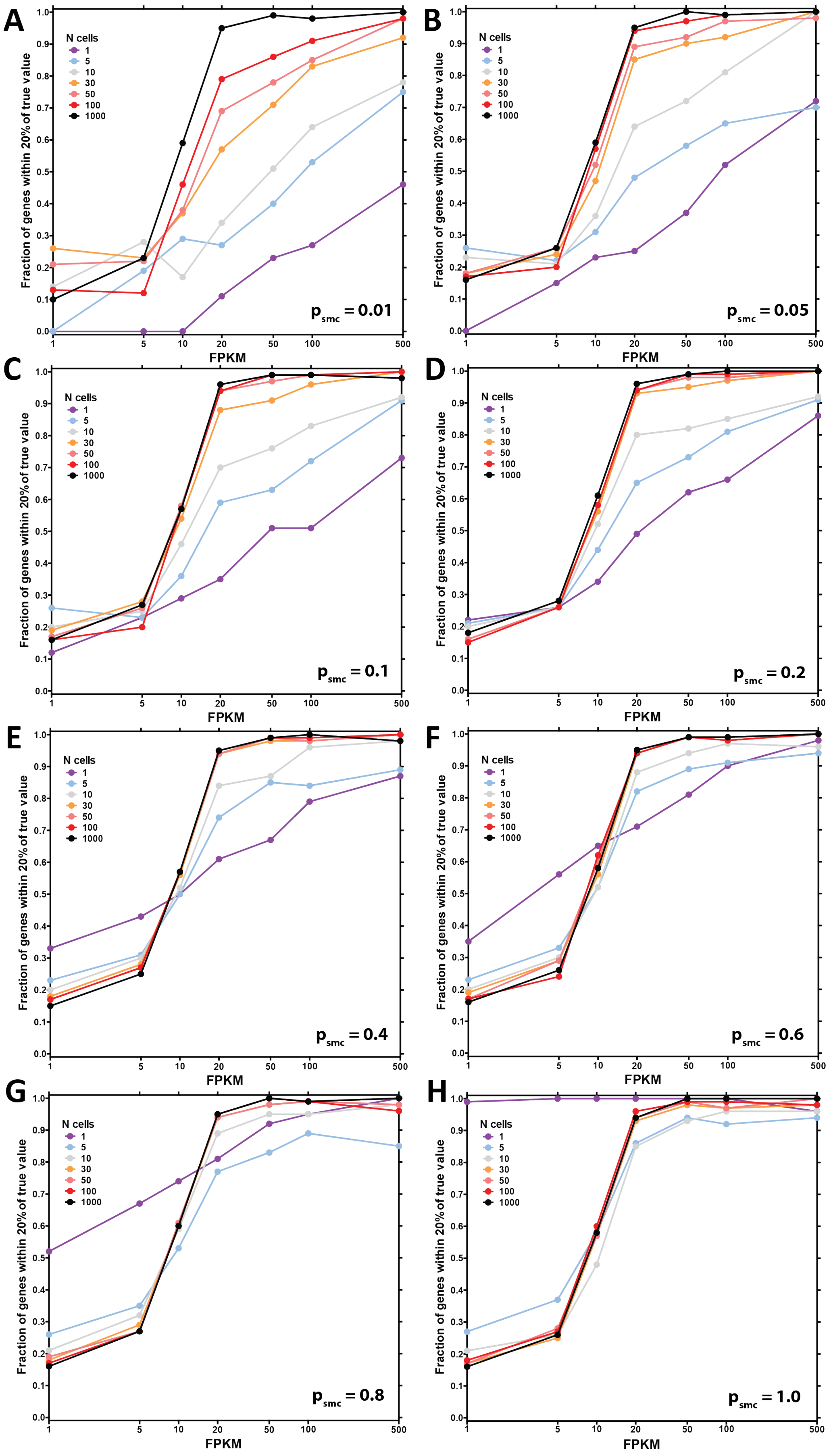

A
B
C
D
E
F
G
H
Fraction of genes within 20% of true value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
FPKM
1
5
10
20
50
100
500
N cells
1
5
10
30
50
100
1000
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

Figure 3.4 ***(preceding page)*****: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability**. Average of 200,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.

Figure 3.5: (following page) Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 500,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.


A
B
C
D
E
F
G
H
Fraction of genes within 20% of true value
FPKM
1
5
10
20
50
100
500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
N cells
1
5
10
30
50
100
1000
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

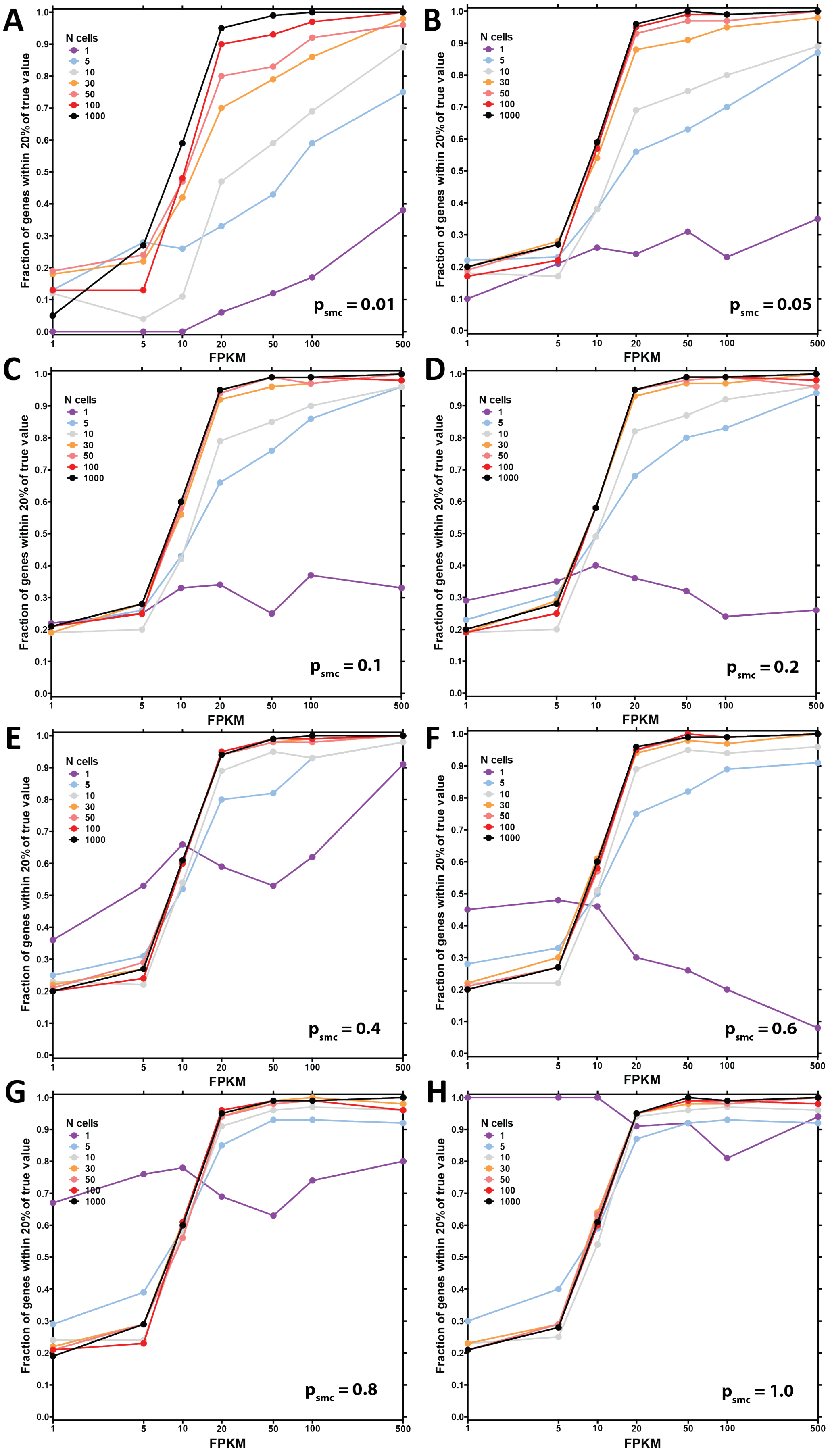
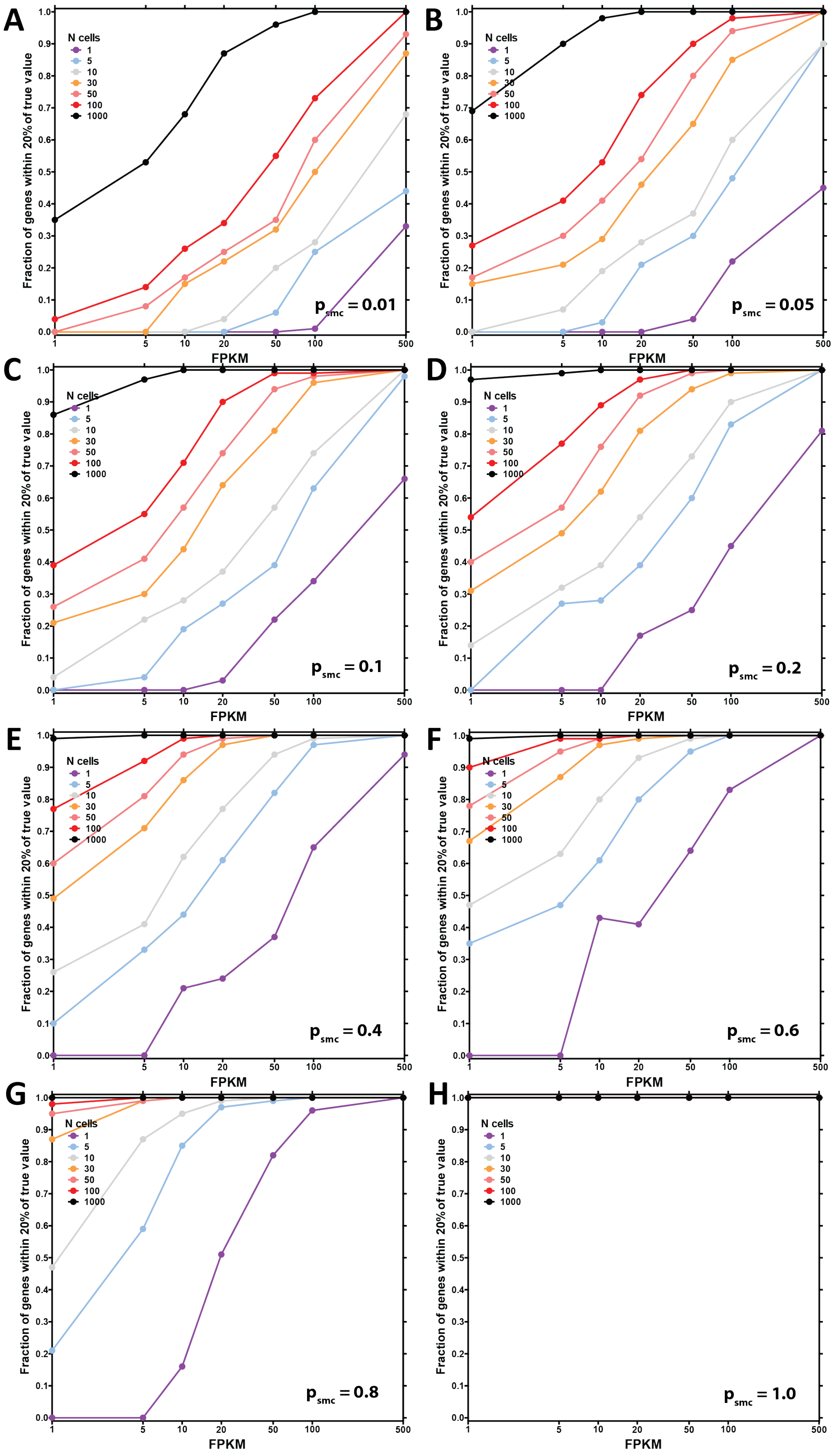

A
B
C
D
E
F
G
H
N cells
1
5
10
30
50
100
1000
Fraction of genes within 20% of true value
FPKM
1
5
10
20
50
100
500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

Figure 3.6 ***(preceding page)*****: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability**. Average of 1,000,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.

Figure 3.7: (following page) Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 50,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.


A
B
C
D
E
F
G
H
Fraction of genes within 20% of true value
FPKM
N cells
1
5
10
30
50
100
1000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
5
10
20
50
100
500
p_{smc} = 0.01
p_{smc} = 0.05
p_{smc} = 0.1
p_{smc} = 0.2
p_{smc} = 0.4
p_{smc} = 0.6
p_{smc} = 0.8
p_{smc} = 1.0

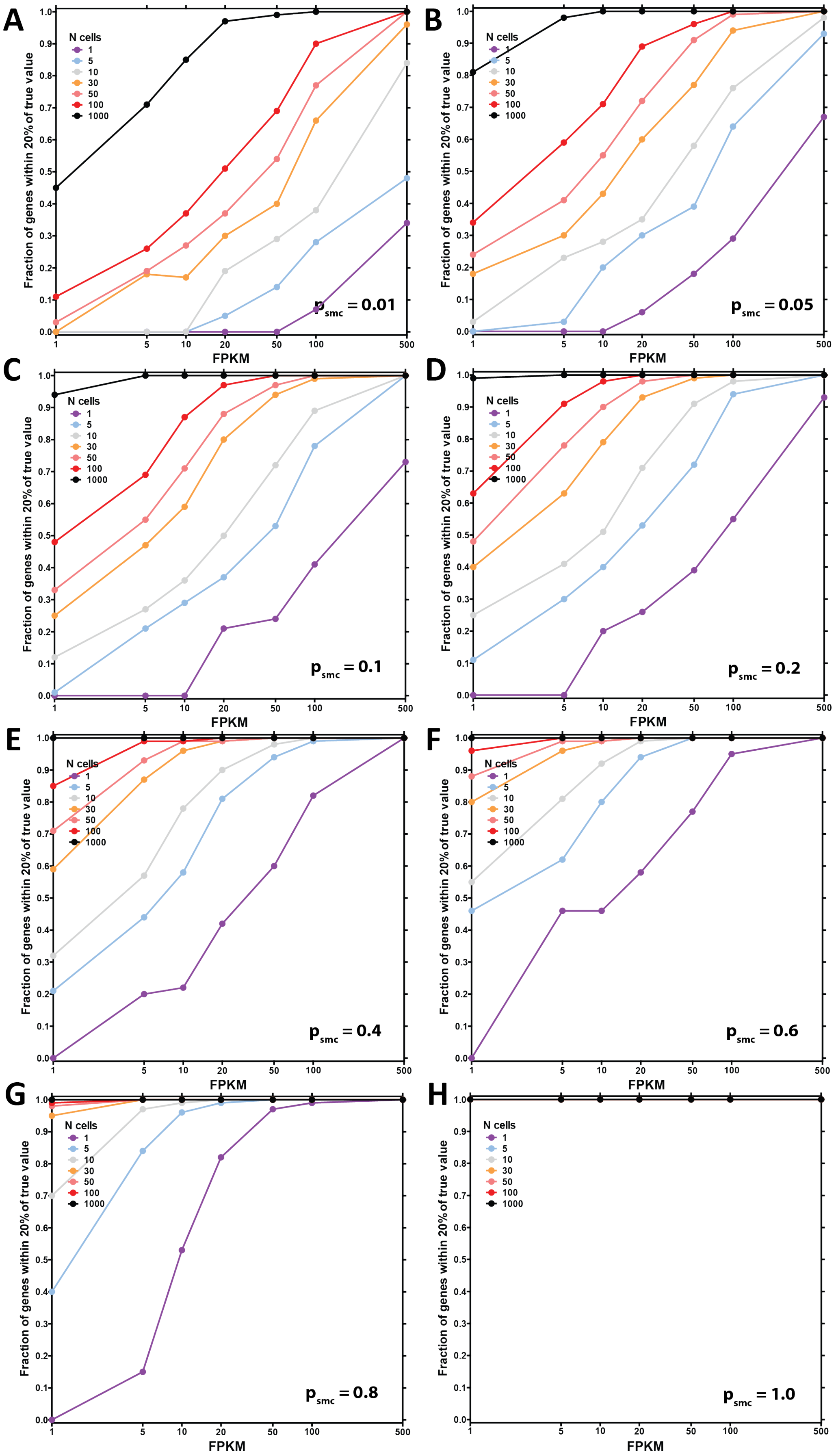
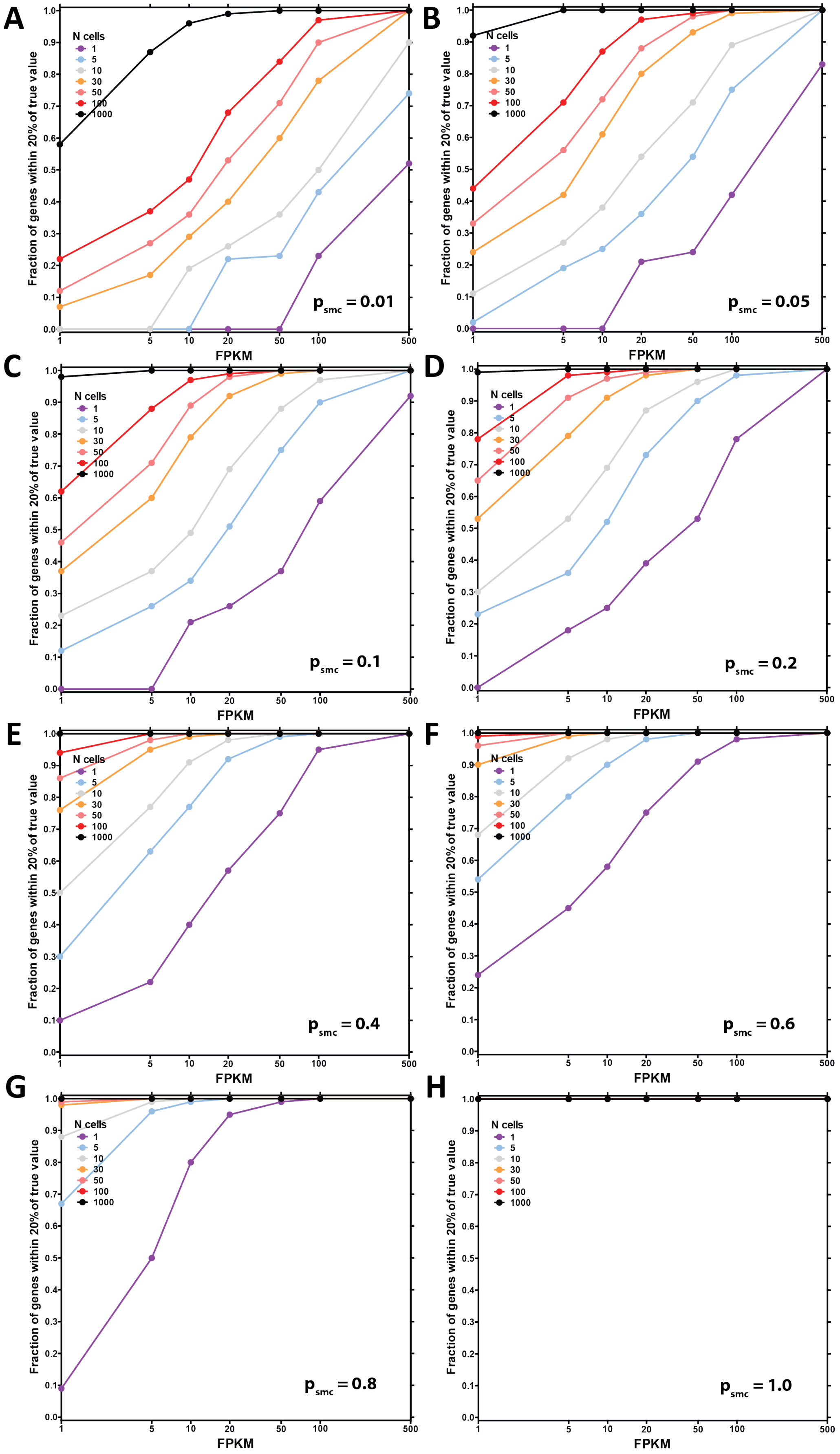

A
B
C
D
E
F
G
H
N cells
1
5
10
30
50
100
1000
Fraction of genes within 20% of true value
FPKM
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
5
10
20
50
100
500
p_smc = 0.01
p_smc = 0.05
p_smc = 0.1
p_smc = 0.2
p_smc = 0.4
p_smc = 0.6
p_smc = 0.8
p_smc = 1.0

Figure 3.8 ***(preceding page)*****: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability**. Average of 100,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.

Figure 3.9: (following page) Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 200,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.


A
B
C
D
E
F
G
H
N cells
1
5
10
30
50
100
1000
Fraction of genes within 20% of true value
FPKM
1
5
10
20
50
100
500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

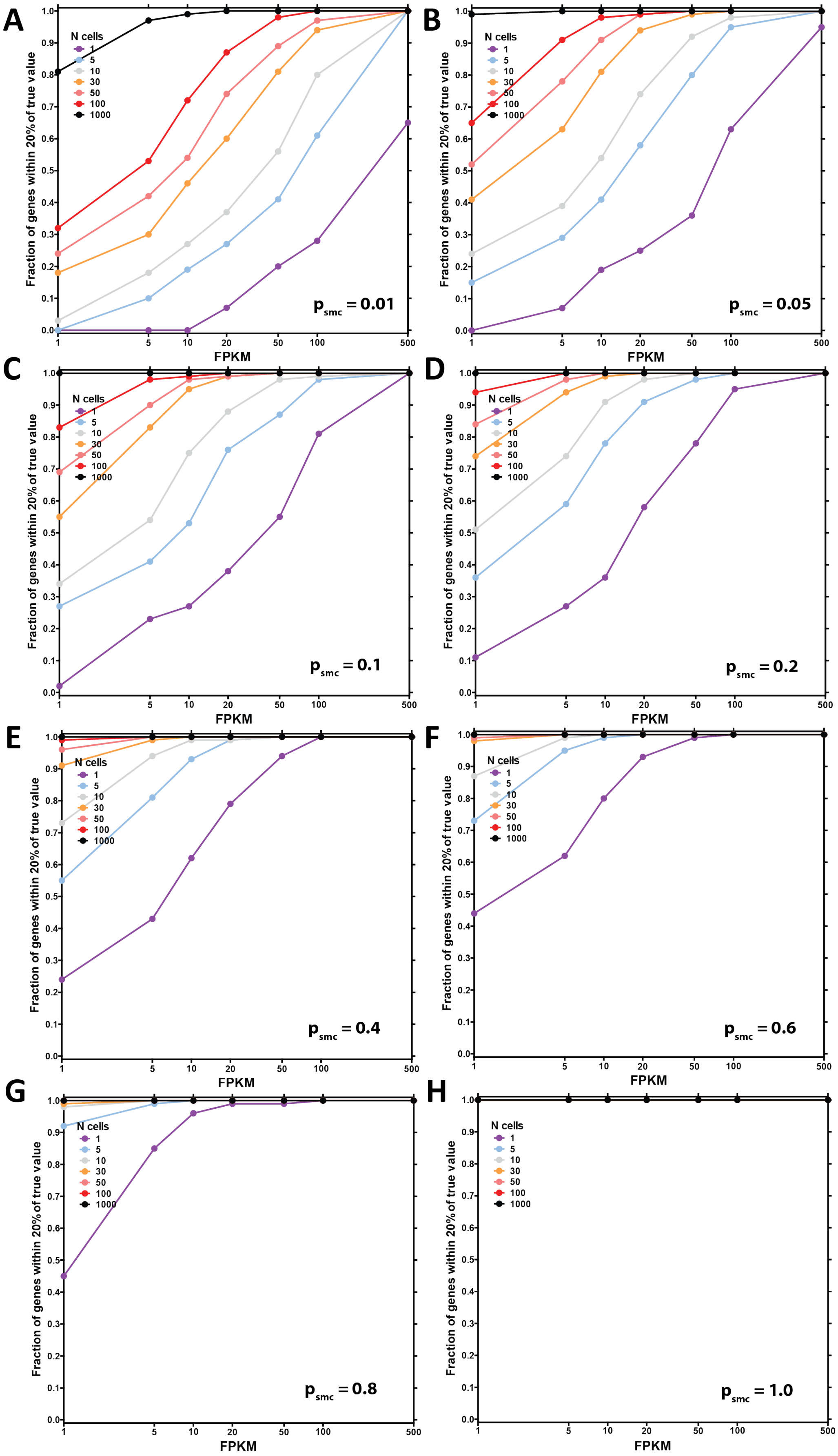
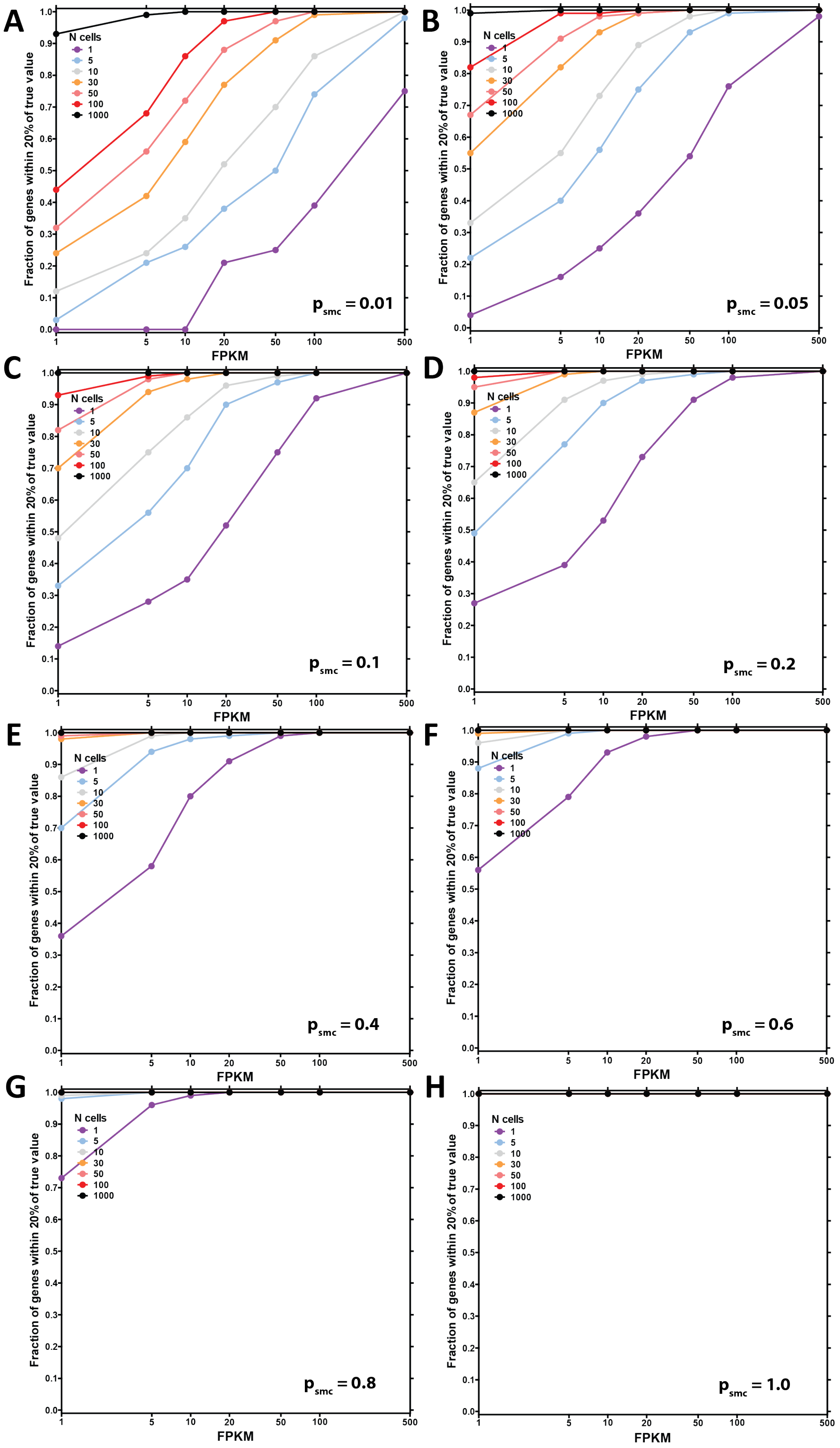

A
B
C
D
E
F
G
H
Fraction of genes within 20% of true value
FPKM
N cells
1
5
10
30
50
100
1000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
5
10
20
50
100
500
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

Figure 3.10 ***(preceding page)*****: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability**. Average of 500,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.

Figure 3.11: (following page) Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 1,000,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.


A
B
C
D
E
F
G
H
Fraction of genes within 20% of true value
FPKM
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
5
10
20
50
100
500
N cells
1
5
10
30
50
100
1000
$p_{smc} = 0.01$
$p_{smc} = 0.05$
$p_{smc} = 0.1$
$p_{smc} = 0.2$
$p_{smc} = 0.4$
$p_{smc} = 0.6$
$p_{smc} = 0.8$
$p_{smc} = 1.0$

A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
16		0.00	0.00	0.00	0.01	0.04	0.03	0.00	0.00	0.00	0.00	0.00
32		0.00	0.00	0.00	0.00	0.03	0.10	0.07	0.03	0.01	0.00	0.00
64		0.00	0.00	0.00	0.00	0.00	0.07	0.14	0.14	0.08	0.03	0.01
128		0.00	0.00	0.00	0.00	0.00	0.03	0.14	0.33	0.31	0.21	0.21
256		0.00	0.00	0.00	0.00	0.00	0.01	0.08	0.31	0.50	0.43	0.44
512		0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.21	0.43	0.52	0.54
1024		0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.21	0.44	0.54	0.73

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.01	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.03	0.07	0.07	0.03	0.01	0.00	0.00	0.00	0.00
16		0.00	0.00	0.00	0.07	0.19	0.17	0.10	0.10	0.07	0.05	0.04
32		0.00	0.00	0.00	0.03	0.17	0.34	0.33	0.33	0.31	0.31	0.30
64		0.00	0.00	0.00	0.01	0.10	0.33	0.52	0.54	0.51	0.50	0.46
128		0.00	0.00	0.00	0.00	0.10	0.33	0.54	0.69	0.67	0.63	0.57
256		0.00	0.00	0.00	0.00	0.07	0.31	0.51	0.67	0.76	0.77	0.69
512		0.00	0.00	0.00	0.00	0.05	0.31	0.50	0.63	0.77	0.84	0.78
1024		0.00	0.00	0.00	0.00	0.04	0.30	0.46	0.57	0.69	0.78	0.75

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.02	0.05	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.04	0.15	0.14	0.10	0.08	0.07	0.05	0.06	0.04
16		0.00	0.00	0.01	0.14	0.31	0.32	0.30	0.30	0.30	0.34	0.32
32		0.00	0.00	0.01	0.10	0.32	0.50	0.50	0.47	0.47	0.47	0.46
64		0.00	0.00	0.00	0.08	0.30	0.50	0.67	0.64	0.61	0.62	0.59
128		0.00	0.00	0.00	0.07	0.30	0.47	0.64	0.73	0.72	0.74	0.70
256		0.00	0.00	0.00	0.05	0.30	0.47	0.61	0.72	0.78	0.81	0.79
512		0.00	0.00	0.00	0.06	0.34	0.47	0.62	0.74	0.81	0.84	0.84
1024		0.00	0.00	0.00	0.04	0.32	0.46	0.59	0.70	0.79	0.84	0.83

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.04	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.04	0.11	0.11	0.07	0.06	0.05	0.04	0.04	0.03	0.03
8		0.00	0.01	0.11	0.24	0.26	0.26	0.25	0.26	0.26	0.27	0.27
16		0.00	0.00	0.07	0.26	0.46	0.49	0.46	0.47	0.44	0.48	0.46
32		0.00	0.00	0.06	0.26	0.49	0.67	0.65	0.65	0.59	0.65	0.61
64		0.00	0.00	0.05	0.25	0.46	0.65	0.77	0.79	0.71	0.78	0.74
128		0.00	0.00	0.04	0.26	0.47	0.65	0.79	0.88	0.80	0.89	0.84
256		0.00	0.00	0.04	0.26	0.44	0.59	0.71	0.80	0.75	0.83	0.79
512		0.00	0.00	0.03	0.27	0.48	0.65	0.78	0.89	0.83	0.94	0.89
1024		0.00	0.00	0.03	0.27	0.46	0.61	0.74	0.84	0.79	0.89	0.83

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.10	0.09	0.07	0.05	0.05	0.04	0.04	0.03	0.02	0.04
4		0.00	0.09	0.19	0.21	0.22	0.23	0.23	0.24	0.24	0.23	0.24
8		0.00	0.07	0.21	0.35	0.41	0.42	0.42	0.40	0.40	0.39	0.40
16		0.00	0.05	0.22	0.41	0.62	0.67	0.67	0.64	0.64	0.62	0.64
32		0.00	0.05	0.23	0.42	0.67	0.81	0.83	0.78	0.79	0.76	0.78
64		0.00	0.04	0.23	0.42	0.67	0.83	0.88	0.84	0.86	0.83	0.85
128		0.00	0.04	0.24	0.40	0.64	0.78	0.84	0.81	0.83	0.80	0.82
256		0.00	0.03	0.24	0.40	0.64	0.79	0.86	0.83	0.84	0.82	0.84
512		0.00	0.02	0.23	0.39	0.62	0.76	0.83	0.80	0.82	0.79	0.81
1024		0.00	0.04	0.24	0.40	0.64	0.78	0.85	0.82	0.84	0.81	0.83

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.14	0.15	0.15	0.17	0.18	0.19	0.20	0.20	0.20	0.20
4		0.00	0.15	0.26	0.30	0.34	0.35	0.36	0.36	0.36	0.36	0.35
8		0.00	0.15	0.30	0.41	0.49	0.50	0.51	0.51	0.51	0.50	0.50
16		0.00	0.17	0.34	0.49	0.68	0.72	0.74	0.75	0.74	0.73	0.72
32		0.00	0.18	0.35	0.50	0.72	0.80	0.82	0.84	0.83	0.82	0.81
64		0.00	0.19	0.36	0.51	0.74	0.82	0.85	0.87	0.86	0.85	0.84
128		0.00	0.20	0.36	0.51	0.75	0.84	0.87	0.89	0.88	0.87	0.86
256		0.00	0.20	0.36	0.51	0.74	0.83	0.86	0.88	0.87	0.85	0.85
512		0.00	0.20	0.36	0.50	0.73	0.82	0.85	0.87	0.85	0.84	0.84
1024		0.00	0.20	0.35	0.50	0.72	0.81	0.84	0.86	0.85	0.84	0.83

G

$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.20	0.22	0.26	0.30	0.32	0.32	0.32	0.32	0.31	0.32
4		0.00	0.22	0.32	0.37	0.42	0.43	0.42	0.42	0.43	0.40	0.42
8		0.00	0.26	0.37	0.47	0.55	0.58	0.59	0.58	0.60	0.56	0.59
16		0.00	0.30	0.42	0.55	0.69	0.74	0.75	0.74	0.77	0.72	0.75
32		0.00	0.32	0.43	0.58	0.74	0.80	0.81	0.80	0.84	0.78	0.82
64		0.00	0.32	0.42	0.59	0.75	0.81	0.82	0.81	0.85	0.78	0.82
128		0.00	0.32	0.42	0.58	0.74	0.80	0.81	0.80	0.84	0.77	0.82
256		0.00	0.32	0.43	0.60	0.77	0.84	0.85	0.84	0.89	0.81	0.86
512		0.00	0.31	0.40	0.56	0.72	0.78	0.78	0.77	0.81	0.74	0.79
1024		0.00	0.32	0.42	0.59	0.75	0.82	0.82	0.82	0.86	0.79	0.83

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.25	0.29	0.33	0.37	0.39	0.39	0.39	0.40	0.39	0.40
4		0.00	0.29	0.32	0.38	0.43	0.46	0.47	0.46	0.47	0.46	0.48
8		0.00	0.33	0.38	0.46	0.55	0.59	0.60	0.59	0.61	0.59	0.62
16		0.00	0.37	0.43	0.55	0.66	0.73	0.74	0.73	0.76	0.72	0.77
32		0.00	0.39	0.46	0.59	0.73	0.82	0.82	0.82	0.85	0.81	0.86
64		0.00	0.39	0.47	0.60	0.74	0.82	0.83	0.83	0.86	0.81	0.87
128		0.00	0.39	0.46	0.59	0.73	0.82	0.83	0.82	0.85	0.81	0.86
256		0.00	0.40	0.47	0.61	0.76	0.85	0.86	0.85	0.89	0.84	0.90
512		0.00	0.39	0.46	0.59	0.72	0.81	0.81	0.81	0.84	0.79	0.85
1024		0.00	0.40	0.48	0.62	0.77	0.86	0.87	0.86	0.90	0.85	0.91

Figure 3.12 ***(preceding page)*****: Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes**. A single cell, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\dfrac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bulk}}}$$

and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.13: (following page) Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 5 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\dfrac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bulk}}}$$

and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.01	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.02	0.07	0.06	0.02	0.01	0.00	0.00	0.00	0.00
16		0.00	0.00	0.00	0.06	0.18	0.16	0.09	0.07	0.06	0.05	0.03
32		0.00	0.00	0.00	0.02	0.16	0.35	0.33	0.28	0.30	0.31	0.30
64		0.00	0.00	0.00	0.01	0.09	0.33	0.55	0.52	0.52	0.53	0.51
128		0.00	0.00	0.00	0.00	0.07	0.28	0.52	0.68	0.66	0.64	0.58
256		0.00	0.00	0.00	0.00	0.06	0.30	0.52	0.66	0.81	0.83	0.76
512		0.00	0.00	0.00	0.00	0.05	0.31	0.53	0.64	0.83	0.98	0.92
1024		0.00	0.00	0.00	0.00	0.03	0.30	0.51	0.58	0.76	0.92	0.93

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.05	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.04	0.12	0.11	0.09	0.08	0.07	0.07	0.08	0.07	0.08
8		0.00	0.01	0.11	0.29	0.33	0.32	0.33	0.35	0.37	0.36	0.36
16		0.00	0.01	0.09	0.33	0.58	0.59	0.56	0.55	0.55	0.53	0.52
32		0.00	0.00	0.08	0.32	0.59	0.77	0.75	0.73	0.73	0.70	0.68
64		0.00	0.00	0.07	0.33	0.56	0.75	0.90	0.89	0.88	0.85	0.83
128		0.00	0.00	0.07	0.35	0.55	0.73	0.89	0.97	0.97	0.95	0.92
256		0.00	0.00	0.08	0.37	0.55	0.73	0.88	0.97	1.00	1.00	0.96
512		0.00	0.00	0.07	0.36	0.53	0.70	0.85	0.95	1.00	1.00	0.99
1024		0.00	0.00	0.08	0.36	0.52	0.68	0.83	0.92	0.96	0.99	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.10	0.09	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07
4		0.00	0.09	0.23	0.25	0.26	0.30	0.31	0.32	0.32	0.32	0.32
8		0.00	0.06	0.25	0.48	0.52	0.51	0.48	0.47	0.46	0.45	0.45
16		0.00	0.06	0.26	0.52	0.74	0.75	0.71	0.70	0.69	0.67	0.65
32		0.00	0.07	0.30	0.51	0.75	0.92	0.90	0.88	0.85	0.83	0.81
64		0.00	0.07	0.31	0.48	0.71	0.90	0.96	0.96	0.93	0.90	0.87
128		0.00	0.07	0.32	0.47	0.70	0.88	0.96	1.00	0.99	0.95	0.92
256		0.00	0.07	0.32	0.46	0.69	0.85	0.93	0.99	1.00	0.97	0.94
512		0.00	0.07	0.32	0.45	0.67	0.83	0.90	0.95	0.97	0.99	0.96
1024		0.00	0.07	0.32	0.45	0.65	0.81	0.87	0.92	0.94	0.96	0.94

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.17	0.18	0.18	0.23	0.25	0.26	0.26	0.27	0.26	0.27
4		0.00	0.18	0.34	0.38	0.39	0.37	0.35	0.34	0.34	0.34	0.33
8		0.00	0.18	0.38	0.59	0.64	0.61	0.59	0.57	0.56	0.55	0.55
16		0.00	0.23	0.39	0.64	0.88	0.88	0.84	0.82	0.82	0.78	0.81
32		0.00	0.25	0.37	0.61	0.88	0.97	0.96	0.94	0.93	0.87	0.92
64		0.00	0.26	0.35	0.59	0.84	0.96	0.98	0.97	0.96	0.89	0.94
128		0.00	0.26	0.34	0.57	0.82	0.94	0.97	0.99	0.99	0.92	0.96
256		0.00	0.27	0.34	0.56	0.82	0.93	0.96	0.99	1.00	0.97	0.99
512		0.00	0.26	0.34	0.55	0.78	0.87	0.89	0.92	0.97	1.00	0.99
1024		0.00	0.27	0.33	0.55	0.81	0.92	0.94	0.96	0.99	0.99	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.31	0.32	0.35	0.34	0.31	0.29	0.28	0.28	0.28	0.28
4		0.00	0.32	0.48	0.51	0.48	0.44	0.41	0.40	0.39	0.39	0.40
8		0.00	0.35	0.51	0.73	0.76	0.71	0.67	0.64	0.64	0.62	0.63
16		0.00	0.34	0.48	0.76	0.93	0.93	0.89	0.87	0.86	0.84	0.83
32		0.00	0.31	0.44	0.71	0.93	0.99	0.98	0.96	0.95	0.91	0.89
64		0.00	0.29	0.41	0.67	0.89	0.98	0.98	0.98	0.96	0.91	0.89
128		0.00	0.28	0.40	0.64	0.87	0.96	0.98	0.99	0.99	0.94	0.90
256		0.00	0.28	0.39	0.64	0.86	0.95	0.96	0.99	1.00	0.98	0.92
512		0.00	0.28	0.39	0.62	0.84	0.91	0.91	0.94	0.98	1.00	0.96
1024		0.00	0.28	0.40	0.63	0.83	0.89	0.89	0.90	0.92	0.96	0.97

F $p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.34	0.36	0.35	0.32	0.29	0.27	0.27	0.26	0.26	0.26
4		0.00	0.36	0.52	0.54	0.51	0.46	0.43	0.43	0.42	0.40	0.40
8		0.00	0.35	0.54	0.74	0.77	0.72	0.68	0.66	0.65	0.64	0.63
16		0.00	0.32	0.51	0.77	0.97	0.96	0.94	0.92	0.92	0.93	0.89
32		0.00	0.29	0.46	0.72	0.96	0.99	0.98	0.97	0.96	0.99	0.94
64		0.00	0.27	0.43	0.68	0.94	0.98	0.99	0.99	0.98	0.99	0.95
128		0.00	0.27	0.43	0.66	0.92	0.97	0.99	1.00	1.00	1.00	0.95
256		0.00	0.26	0.42	0.65	0.92	0.96	0.98	1.00	1.00	1.00	0.96
512		0.00	0.26	0.40	0.64	0.93	0.99	0.99	1.00	1.00	1.00	0.95
1024		0.00	0.26	0.40	0.63	0.89	0.94	0.95	0.95	0.96	0.95	0.94

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.42	0.42	0.40	0.34	0.29	0.27	0.26	0.25	0.25	0.24
4		0.00	0.42	0.55	0.58	0.54	0.48	0.46	0.45	0.43	0.43	0.42
8		0.00	0.40	0.58	0.77	0.80	0.76	0.73	0.71	0.70	0.68	0.69
16		0.00	0.34	0.54	0.80	0.97	0.96	0.94	0.93	0.92	0.89	0.92
32		0.00	0.29	0.48	0.76	0.96	0.99	0.98	0.98	0.95	0.93	0.96
64		0.00	0.27	0.46	0.73	0.94	0.98	0.99	0.99	0.97	0.93	0.96
128		0.00	0.26	0.45	0.71	0.93	0.98	0.99	1.00	0.99	0.95	0.97
256		0.00	0.25	0.43	0.70	0.92	0.95	0.97	0.99	1.00	0.94	0.96
512		0.00	0.25	0.43	0.68	0.89	0.93	0.93	0.95	0.94	0.98	0.96
1024		0.00	0.24	0.42	0.69	0.92	0.96	0.96	0.97	0.96	0.96	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.48	0.47	0.45	0.39	0.34	0.30	0.30	0.29	0.28	0.32
4		0.00	0.47	0.59	0.60	0.57	0.51	0.47	0.47	0.46	0.45	0.48
8		0.00	0.45	0.60	0.78	0.80	0.75	0.71	0.70	0.69	0.69	0.70
16		0.00	0.39	0.57	0.80	0.98	0.96	0.93	0.94	0.93	0.94	0.86
32		0.00	0.34	0.51	0.75	0.96	1.00	0.97	0.98	0.96	0.97	0.86
64		0.00	0.30	0.47	0.71	0.93	0.97	0.98	0.98	0.97	0.95	0.81
128		0.00	0.30	0.47	0.70	0.94	0.98	0.98	1.00	1.00	0.99	0.83
256		0.00	0.29	0.46	0.69	0.93	0.96	0.97	1.00	1.00	1.00	0.89
512		0.00	0.28	0.45	0.69	0.94	0.97	0.95	0.99	1.00	1.00	1.00
1024		0.00	0.32	0.48	0.70	0.86	0.86	0.81	0.83	0.89	1.00	1.00

A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.00	0.02	0.05	0.04	0.02	0.01	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.04	0.15	0.16	0.10	0.06	0.07	0.04	0.03	0.02
16		0.00	0.00	0.02	0.16	0.34	0.34	0.29	0.31	0.29	0.29	0.31
32		0.00	0.00	0.01	0.10	0.34	0.56	0.55	0.54	0.52	0.51	0.53
64		0.00	0.00	0.00	0.06	0.29	0.55	0.75	0.74	0.71	0.67	0.67
128		0.00	0.00	0.00	0.07	0.31	0.54	0.74	0.88	0.87	0.80	0.80
256		0.00	0.00	0.00	0.04	0.29	0.52	0.71	0.87	0.99	0.95	0.95
512		0.00	0.00	0.00	0.03	0.29	0.51	0.67	0.80	0.95	0.99	0.98
1024		0.00	0.00	0.00	0.02	0.31	0.53	0.67	0.80	0.95	0.98	1.00

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.09	0.09	0.05	0.06	0.07	0.07	0.06	0.06	0.07	0.05
4		0.00	0.09	0.23	0.24	0.26	0.29	0.30	0.30	0.31	0.32	0.31
8		0.00	0.05	0.24	0.48	0.52	0.50	0.48	0.47	0.46	0.45	0.45
16		0.00	0.06	0.26	0.52	0.76	0.77	0.73	0.70	0.68	0.70	0.67
32		0.00	0.07	0.29	0.50	0.77	0.93	0.91	0.87	0.85	0.87	0.82
64		0.00	0.07	0.30	0.48	0.73	0.91	0.97	0.96	0.93	0.95	0.89
128		0.00	0.06	0.30	0.47	0.70	0.87	0.96	0.99	0.99	0.99	0.94
256		0.00	0.06	0.31	0.46	0.68	0.85	0.93	0.99	1.00	1.00	0.96
512		0.00	0.07	0.32	0.45	0.70	0.87	0.95	0.99	1.00	1.00	0.98
1024		0.00	0.05	0.31	0.45	0.67	0.82	0.89	0.94	0.96	0.98	1.00

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.16	0.17	0.17	0.21	0.24	0.25	0.25	0.25	0.26	0.25
4		0.00	0.17	0.33	0.38	0.39	0.37	0.35	0.34	0.33	0.32	0.33
8		0.00	0.17	0.38	0.61	0.65	0.61	0.58	0.57	0.56	0.55	0.55
16		0.00	0.21	0.39	0.65	0.89	0.89	0.86	0.84	0.83	0.82	0.81
32		0.00	0.24	0.37	0.61	0.89	0.98	0.97	0.96	0.95	0.95	0.92
64		0.00	0.25	0.35	0.58	0.86	0.97	1.00	0.99	0.99	0.98	0.95
128		0.00	0.25	0.34	0.57	0.84	0.96	0.99	1.00	1.00	1.00	0.96
256		0.00	0.25	0.33	0.56	0.83	0.95	0.99	1.00	1.00	1.00	0.98
512		0.00	0.26	0.32	0.55	0.82	0.95	0.98	1.00	1.00	1.00	1.00
1024		0.00	0.25	0.33	0.55	0.81	0.92	0.95	0.96	0.98	1.00	1.00

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.27	0.30	0.31	0.31	0.27	0.25	0.25	0.24	0.23	0.23
4		0.00	0.30	0.48	0.50	0.48	0.43	0.40	0.39	0.39	0.37	0.37
8		0.00	0.31	0.50	0.71	0.75	0.69	0.64	0.62	0.61	0.60	0.59
16		0.00	0.31	0.48	0.75	0.97	0.96	0.94	0.91	0.90	0.91	0.90
32		0.00	0.27	0.43	0.69	0.96	1.00	0.99	0.98	0.97	0.98	0.97
64		0.00	0.25	0.40	0.64	0.94	0.99	1.00	0.99	0.98	0.99	0.97
128		0.00	0.25	0.39	0.62	0.91	0.98	0.99	1.00	1.00	0.99	0.97
256		0.00	0.24	0.39	0.61	0.90	0.97	0.98	1.00	1.00	1.00	0.99
512		0.00	0.23	0.37	0.60	0.91	0.98	0.99	0.99	1.00	1.00	1.00
1024		0.00	0.23	0.37	0.59	0.90	0.97	0.97	0.97	0.99	1.00	1.00

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.43	0.43	0.41	0.37	0.32	0.30	0.29	0.28	0.28	0.30
4		0.00	0.43	0.56	0.58	0.54	0.47	0.44	0.43	0.42	0.41	0.44
8		0.00	0.41	0.58	0.79	0.82	0.76	0.72	0.70	0.69	0.68	0.69
16		0.00	0.37	0.54	0.82	0.99	0.98	0.96	0.95	0.95	0.93	0.91
32		0.00	0.32	0.47	0.76	0.98	1.00	1.00	0.98	1.00	0.97	0.93
64		0.00	0.30	0.44	0.72	0.96	1.00	1.00	0.99	1.00	0.97	0.92
128		0.00	0.29	0.43	0.70	0.95	0.98	0.99	1.00	1.00	0.97	0.92
256		0.00	0.28	0.42	0.69	0.95	1.00	1.00	1.00	1.00	0.99	0.93
512		0.00	0.28	0.41	0.68	0.93	0.97	0.97	0.97	0.99	1.00	0.96
1024		0.00	0.30	0.44	0.69	0.91	0.93	0.92	0.92	0.93	0.96	0.97

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.41	0.43	0.41	0.35	0.30	0.28	0.27	0.28	0.26	0.28
4		0.00	0.43	0.59	0.61	0.55	0.48	0.45	0.43	0.43	0.41	0.44
8		0.00	0.41	0.61	0.82	0.84	0.78	0.74	0.71	0.71	0.69	0.71
16		0.00	0.35	0.55	0.84	1.00	0.99	0.98	0.98	0.97	0.97	0.97
32		0.00	0.30	0.48	0.78	0.99	1.00	1.00	1.00	0.99	0.99	0.98
64		0.00	0.28	0.45	0.74	0.98	1.00	1.00	1.00	0.99	0.99	0.98
128		0.00	0.27	0.43	0.71	0.98	1.00	1.00	1.00	0.99	1.00	0.98
256		0.00	0.28	0.43	0.71	0.97	0.99	0.99	0.99	1.00	1.00	0.98
512		0.00	0.26	0.41	0.69	0.97	0.99	0.99	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.44	0.71	0.97	0.98	0.98	0.98	0.98	1.00	1.00

G

$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.45	0.45	0.44	0.36	0.30	0.28	0.27	0.28	0.27	0.26
4		0.00	0.45	0.60	0.63	0.57	0.49	0.46	0.45	0.45	0.44	0.43
8		0.00	0.44	0.63	0.83	0.84	0.77	0.73	0.71	0.71	0.69	0.69
16		0.00	0.36	0.57	0.84	0.99	0.99	0.98	0.97	0.97	0.95	0.96
32		0.00	0.30	0.49	0.77	0.99	1.00	1.00	1.00	0.99	0.97	0.98
64		0.00	0.28	0.46	0.73	0.98	1.00	1.00	1.00	1.00	0.97	0.98
128		0.00	0.27	0.45	0.71	0.97	1.00	1.00	1.00	1.00	0.99	0.98
256		0.00	0.28	0.45	0.71	0.97	0.99	1.00	1.00	1.00	1.00	0.99
512		0.00	0.27	0.44	0.69	0.95	0.97	0.97	0.99	1.00	1.00	1.00
1024		0.00	0.26	0.43	0.69	0.96	0.98	0.98	0.98	0.99	1.00	1.00

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.46	0.47	0.45	0.38	0.32	0.30	0.29	0.28	0.28	0.27
4		0.00	0.47	0.63	0.65	0.59	0.51	0.47	0.45	0.44	0.44	0.43
8		0.00	0.45	0.65	0.84	0.85	0.79	0.75	0.72	0.70	0.71	0.70
16		0.00	0.38	0.59	0.85	0.99	0.99	0.98	0.96	0.96	0.97	0.97
32		0.00	0.32	0.51	0.79	0.99	1.00	1.00	0.99	0.99	1.00	1.00
64		0.00	0.30	0.47	0.75	0.98	1.00	1.00	0.99	0.99	1.00	1.00
128		0.00	0.29	0.45	0.72	0.96	0.99	0.99	0.99	1.00	1.00	0.99
256		0.00	0.28	0.44	0.70	0.96	0.99	0.99	1.00	1.00	1.00	1.00
512		0.00	0.28	0.44	0.71	0.97	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.43	0.70	0.97	1.00	1.00	0.99	1.00	1.00	1.00

Figure 3.14 ***(preceding page)*****: Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes**. A pool of 10 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\dfrac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.15: (following page) Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 30 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\dfrac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.05	0.04	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
4		0.00	0.04	0.14	0.15	0.12	0.13	0.13	0.13	0.13	0.14	0.14
8		0.00	0.02	0.15	0.37	0.40	0.41	0.42	0.43	0.43	0.43	0.44
16		0.00	0.01	0.12	0.40	0.63	0.65	0.61	0.59	0.56	0.56	0.55
32		0.00	0.01	0.13	0.41	0.65	0.84	0.83	0.80	0.77	0.76	0.75
64		0.00	0.00	0.13	0.42	0.61	0.83	0.95	0.94	0.90	0.89	0.89
128		0.00	0.00	0.13	0.43	0.59	0.80	0.94	1.00	0.99	0.98	0.97
256		0.00	0.00	0.13	0.43	0.56	0.77	0.90	0.99	1.00	1.00	1.00
512		0.00	0.00	0.14	0.43	0.56	0.76	0.89	0.98	1.00	1.00	1.00
1024		0.00	0.00	0.14	0.44	0.55	0.75	0.89	0.97	1.00	1.00	1.00

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.22	0.24	0.26	0.29	0.29	0.28	0.28	0.27	0.27	0.28
4		0.00	0.24	0.42	0.46	0.45	0.41	0.39	0.38	0.37	0.37	0.38
8		0.00	0.26	0.46	0.69	0.72	0.68	0.64	0.62	0.61	0.61	0.61
16		0.00	0.29	0.45	0.72	0.94	0.94	0.91	0.90	0.88	0.88	0.88
32		0.00	0.29	0.41	0.68	0.94	0.99	0.99	0.99	0.98	0.98	0.97
64		0.00	0.28	0.39	0.64	0.91	0.99	1.00	1.00	1.00	1.00	1.00
128		0.00	0.28	0.38	0.62	0.90	0.99	1.00	1.00	1.00	1.00	1.00
256		0.00	0.27	0.37	0.61	0.88	0.98	1.00	1.00	1.00	1.00	1.00
512		0.00	0.27	0.37	0.61	0.88	0.98	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.38	0.61	0.88	0.97	1.00	1.00	1.00	1.00	1.00

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.34	0.35	0.36	0.32	0.28	0.27	0.26	0.26	0.26	0.26
4		0.00	0.35	0.52	0.55	0.50	0.45	0.42	0.41	0.40	0.40	0.39
8		0.00	0.36	0.55	0.80	0.82	0.75	0.71	0.69	0.67	0.68	0.66
16		0.00	0.32	0.50	0.82	0.99	0.99	0.97	0.96	0.96	0.96	0.95
32		0.00	0.28	0.45	0.75	0.99	1.00	1.00	0.99	1.00	1.00	0.99
64		0.00	0.27	0.42	0.71	0.97	1.00	1.00	1.00	1.00	1.00	0.99
128		0.00	0.26	0.41	0.69	0.96	0.99	1.00	1.00	1.00	1.00	0.99
256		0.00	0.26	0.40	0.67	0.96	1.00	1.00	1.00	1.00	1.00	0.99
512		0.00	0.26	0.40	0.68	0.96	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.26	0.39	0.66	0.95	0.99	0.99	0.99	0.99	1.00	1.00

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.41	0.43	0.43	0.36	0.31	0.29	0.29	0.28	0.28	0.28
4		0.00	0.43	0.60	0.63	0.56	0.47	0.44	0.43	0.41	0.41	0.41
8		0.00	0.43	0.63	0.86	0.86	0.78	0.73	0.71	0.69	0.69	0.69
16		0.00	0.36	0.56	0.86	1.00	1.00	0.99	0.98	0.98	0.98	0.98
32		0.00	0.31	0.47	0.78	1.00	1.00	1.00	1.00	0.99	1.00	1.00
64		0.00	0.29	0.44	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.29	0.43	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.41	0.69	0.98	0.99	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.41	0.69	0.98	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.41	0.69	0.98	1.00	1.00	1.00	1.00	1.00	1.00

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.45	0.46	0.44	0.37	0.30	0.28	0.28	0.28	0.27	0.27
4		0.00	0.46	0.62	0.64	0.56	0.48	0.45	0.43	0.44	0.42	0.42
8		0.00	0.44	0.64	0.86	0.86	0.79	0.75	0.73	0.72	0.71	0.70
16		0.00	0.37	0.56	0.86	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.30	0.48	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.45	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.28	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.44	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.27	0.42	0.71	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.42	0.70	0.99	1.00	1.00	1.00	1.00	1.00	1.00

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.45	0.47	0.44	0.35	0.28	0.25	0.24	0.23	0.23	0.23
4		0.00	0.47	0.66	0.68	0.59	0.50	0.47	0.45	0.44	0.44	0.43
8		0.00	0.44	0.68	0.88	0.88	0.82	0.78	0.76	0.74	0.73	0.73
16		0.00	0.35	0.59	0.88	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.28	0.50	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.25	0.47	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.24	0.45	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.23	0.44	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.23	0.44	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.23	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00

G

$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.50	0.51	0.47	0.36	0.28	0.26	0.26	0.25	0.23	0.26
4		0.00	0.51	0.68	0.68	0.58	0.49	0.45	0.43	0.43	0.40	0.43
8		0.00	0.47	0.68	0.88	0.88	0.81	0.76	0.74	0.74	0.71	0.73
16		0.00	0.36	0.58	0.88	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.28	0.49	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.26	0.45	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.43	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.43	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.23	0.40	0.71	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.26	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.49	0.51	0.48	0.38	0.31	0.28	0.28	0.27	0.28	0.27
4		0.00	0.51	0.66	0.67	0.57	0.48	0.44	0.43	0.41	0.42	0.41
8		0.00	0.48	0.67	0.89	0.89	0.82	0.79	0.77	0.76	0.76	0.75
16		0.00	0.38	0.57	0.89	1.00	1.00	1.00	1.00	1.00	0.99	1.00
32		0.00	0.31	0.48	0.82	1.00	1.00	1.00	1.00	1.00	0.99	1.00
64		0.00	0.28	0.44	0.79	1.00	1.00	1.00	1.00	1.00	0.99	1.00
128		0.00	0.28	0.43	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.27	0.41	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.42	0.76	0.99	0.99	0.99	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.41	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00

A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.08	0.08	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05
4		0.00	0.08	0.21	0.23	0.24	0.27	0.29	0.29	0.30	0.30	0.29
8		0.00	0.05	0.23	0.47	0.51	0.51	0.48	0.48	0.46	0.46	0.46
16		0.00	0.05	0.24	0.51	0.75	0.77	0.74	0.71	0.70	0.70	0.70
32		0.00	0.06	0.27	0.51	0.77	0.93	0.92	0.89	0.88	0.88	0.87
64		0.00	0.05	0.29	0.48	0.74	0.92	0.99	0.98	0.97	0.98	0.97
128		0.00	0.05	0.29	0.48	0.71	0.89	0.98	1.00	1.00	1.00	1.00
256		0.00	0.05	0.30	0.46	0.70	0.88	0.97	1.00	1.00	1.00	1.00
512		0.00	0.05	0.30	0.46	0.70	0.88	0.98	1.00	1.00	1.00	1.00
1024		0.00	0.05	0.29	0.46	0.70	0.87	0.97	1.00	1.00	1.00	1.00

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.30	0.31	0.34	0.33	0.30	0.28	0.28	0.28	0.28	0.28
4		0.00	0.31	0.49	0.53	0.47	0.42	0.38	0.37	0.37	0.37	0.37
8		0.00	0.34	0.53	0.80	0.81	0.75	0.70	0.68	0.68	0.67	0.67
16		0.00	0.33	0.47	0.81	0.99	0.98	0.96	0.95	0.94	0.94	0.94
32		0.00	0.30	0.42	0.75	0.98	1.00	1.00	1.00	0.99	1.00	1.00
64		0.00	0.28	0.38	0.70	0.96	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.28	0.37	0.68	0.95	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.37	0.68	0.94	0.99	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.37	0.67	0.94	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.37	0.67	0.94	1.00	1.00	1.00	1.00	1.00	1.00

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.39	0.40	0.39	0.35	0.30	0.29	0.28	0.28	0.28	0.27
4		0.00	0.40	0.58	0.59	0.53	0.46	0.43	0.42	0.41	0.42	0.40
8		0.00	0.39	0.59	0.81	0.83	0.77	0.73	0.71	0.71	0.71	0.69
16		0.00	0.35	0.53	0.83	0.99	0.99	0.99	0.98	0.98	0.98	0.97
32		0.00	0.30	0.46	0.77	0.99	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.29	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.28	0.42	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.41	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.42	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.40	0.69	0.97	1.00	1.00	1.00	1.00	1.00	1.00

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.44	0.45	0.43	0.35	0.29	0.27	0.27	0.27	0.26	0.27
4		0.00	0.45	0.62	0.64	0.56	0.48	0.44	0.43	0.43	0.42	0.43
8		0.00	0.43	0.64	0.85	0.86	0.79	0.75	0.73	0.73	0.72	0.73
16		0.00	0.35	0.56	0.86	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.29	0.48	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.27	0.44	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.27	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.26	0.42	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.52	0.52	0.49	0.38	0.31	0.28	0.27	0.27	0.27	0.28
4		0.00	0.52	0.67	0.68	0.59	0.49	0.46	0.44	0.43	0.44	0.44
8		0.00	0.49	0.68	0.88	0.87	0.79	0.76	0.74	0.72	0.72	0.73
16		0.00	0.38	0.59	0.87	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.31	0.49	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.46	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.44	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.27	0.43	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.27	0.44	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.44	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.47	0.49	0.44	0.35	0.29	0.26	0.26	0.25	0.24	0.25
4		0.00	0.49	0.67	0.68	0.59	0.50	0.46	0.44	0.43	0.42	0.43
8		0.00	0.44	0.68	0.87	0.88	0.82	0.78	0.76	0.75	0.73	0.73
16		0.00	0.35	0.59	0.88	1.00	1.00	1.00	1.00	0.99	0.99	0.99
32		0.00	0.29	0.50	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.26	0.46	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.43	0.75	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.24	0.42	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.25	0.43	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00

G

$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.49	0.51	0.47	0.35	0.28	0.26	0.25	0.24	0.24	0.23
4		0.00	0.51	0.69	0.69	0.59	0.50	0.46	0.44	0.43	0.43	0.41
8		0.00	0.47	0.69	0.88	0.88	0.81	0.78	0.76	0.74	0.74	0.72
16		0.00	0.35	0.59	0.88	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.28	0.50	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.26	0.46	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.25	0.44	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.24	0.43	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.24	0.43	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.23	0.41	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.52	0.53	0.49	0.39	0.32	0.29	0.29	0.29	0.28	0.29
4		0.00	0.53	0.69	0.69	0.58	0.49	0.46	0.45	0.45	0.44	0.45
8		0.00	0.49	0.69	0.90	0.89	0.83	0.79	0.77	0.77	0.76	0.77
16		0.00	0.39	0.58	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.32	0.49	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.29	0.46	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.29	0.45	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.29	0.45	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.29	0.45	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 3.16 *(preceding page)*: Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 50 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.17: (following page) Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 100 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.16	0.16	0.16	0.21	0.24	0.24	0.24	0.25	0.24	0.24
4		0.00	0.16	0.33	0.38	0.40	0.37	0.35	0.34	0.33	0.34	0.34
8		0.00	0.16	0.38	0.65	0.68	0.64	0.61	0.59	0.58	0.59	0.58
16		0.00	0.21	0.40	0.68	0.92	0.92	0.89	0.87	0.86	0.86	0.85
32		0.00	0.24	0.37	0.64	0.92	0.99	0.98	0.98	0.97	0.97	0.95
64		0.00	0.24	0.35	0.61	0.89	0.98	1.00	1.00	0.99	0.99	0.96
128		0.00	0.24	0.34	0.59	0.87	0.98	1.00	1.00	1.00	1.00	0.98
256		0.00	0.25	0.33	0.58	0.86	0.97	0.99	1.00	1.00	1.00	0.98
512		0.00	0.24	0.34	0.59	0.86	0.97	0.99	1.00	1.00	1.00	1.00
1024		0.00	0.24	0.34	0.58	0.85	0.95	0.96	0.98	0.98	1.00	1.00

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.39	0.41	0.40	0.34	0.30	0.28	0.28	0.28	0.27	0.28
4		0.00	0.41	0.59	0.61	0.52	0.44	0.41	0.40	0.39	0.39	0.40
8		0.00	0.40	0.61	0.84	0.84	0.77	0.72	0.70	0.68	0.68	0.69
16		0.00	0.34	0.52	0.84	1.00	0.99	0.99	0.99	0.98	0.98	0.98
32		0.00	0.30	0.44	0.77	0.99	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.41	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.28	0.40	0.70	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.39	0.68	0.98	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.27	0.39	0.68	0.98	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.40	0.69	0.98	1.00	1.00	1.00	1.00	1.00	1.00

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.44	0.46	0.44	0.35	0.29	0.27	0.26	0.26	0.25	0.26
4		0.00	0.46	0.62	0.64	0.55	0.47	0.43	0.42	0.42	0.40	0.41
8		0.00	0.44	0.64	0.86	0.86	0.79	0.75	0.73	0.72	0.70	0.71
16		0.00	0.35	0.55	0.86	1.00	1.00	1.00	1.00	1.00	0.99	0.99
32		0.00	0.29	0.47	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.27	0.43	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.42	0.73	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.26	0.42	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.40	0.70	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.26	0.41	0.71	0.99	1.00	1.00	1.00	1.00	1.00	1.00

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.45	0.48	0.47	0.38	0.31	0.29	0.29	0.29	0.28	0.27
4		0.00	0.48	0.67	0.69	0.58	0.49	0.45	0.44	0.44	0.42	0.41
8		0.00	0.47	0.69	0.89	0.88	0.82	0.77	0.75	0.75	0.73	0.72
16		0.00	0.38	0.58	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.31	0.49	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.29	0.45	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.29	0.44	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.29	0.44	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.42	0.73	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.41	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.53	0.53	0.48	0.36	0.29	0.27	0.26	0.25	0.25	0.28
4		0.00	0.53	0.69	0.69	0.59	0.49	0.45	0.44	0.42	0.42	0.45
8		0.00	0.48	0.69	0.90	0.90	0.84	0.80	0.78	0.77	0.76	0.79
16		0.00	0.36	0.59	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.29	0.49	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.27	0.45	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.44	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.42	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.42	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.28	0.45	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.50	0.52	0.48	0.37	0.30	0.27	0.26	0.25	0.25	0.26
4		0.00	0.52	0.69	0.69	0.57	0.46	0.42	0.41	0.40	0.40	0.41
8		0.00	0.48	0.69	0.92	0.90	0.84	0.80	0.78	0.77	0.76	0.77
16		0.00	0.37	0.57	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.30	0.46	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.27	0.42	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.41	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.40	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.40	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.26	0.41	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G

$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.54	0.53	0.48	0.36	0.29	0.27	0.26	0.26	0.25	0.27
4		0.00	0.53	0.69	0.69	0.58	0.49	0.46	0.44	0.44	0.42	0.44
8		0.00	0.48	0.69	0.88	0.88	0.82	0.79	0.76	0.76	0.74	0.76
16		0.00	0.36	0.58	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.29	0.49	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.27	0.46	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.26	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.42	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.51	0.52	0.49	0.38	0.30	0.28	0.27	0.28	0.26	0.25
4		0.00	0.52	0.69	0.69	0.58	0.48	0.45	0.43	0.43	0.41	0.39
8		0.00	0.49	0.69	0.90	0.89	0.82	0.79	0.76	0.76	0.74	0.72
16		0.00	0.38	0.58	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.30	0.48	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.45	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.43	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.43	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.26	0.41	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.25	0.39	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00

A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.42	0.45	0.42	0.35	0.29	0.28	0.27	0.26	0.25	0.26
4		0.00	0.45	0.64	0.66	0.56	0.47	0.44	0.42	0.42	0.39	0.41
8		0.00	0.42	0.66	0.88	0.88	0.81	0.77	0.75	0.74	0.71	0.73
16		0.00	0.35	0.56	0.88	1.00	1.00	1.00	0.99	0.99	0.99	0.99
32		0.00	0.29	0.47	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.44	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.42	0.75	0.99	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.26	0.42	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.39	0.71	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.26	0.41	0.73	0.99	1.00	1.00	1.00	1.00	1.00	1.00

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.53	0.53	0.49	0.38	0.30	0.28	0.27	0.27	0.27	0.27
4		0.00	0.53	0.69	0.70	0.59	0.49	0.45	0.43	0.43	0.42	0.43
8		0.00	0.49	0.70	0.90	0.89	0.83	0.78	0.76	0.75	0.75	0.75
16		0.00	0.38	0.59	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.30	0.49	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.45	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.43	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.27	0.43	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.27	0.42	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.43	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.49	0.51	0.47	0.36	0.29	0.26	0.25	0.25	0.25	0.24
4		0.00	0.51	0.70	0.70	0.60	0.50	0.47	0.46	0.44	0.45	0.43
8		0.00	0.47	0.70	0.90	0.89	0.82	0.78	0.76	0.74	0.74	0.73
16		0.00	0.36	0.60	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.29	0.50	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.26	0.47	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.25	0.46	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.44	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.45	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.24	0.43	0.73	1.00	1.00	1.00	1.00	1.00	1.00	1.00

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.51	0.52	0.48	0.36	0.29	0.26	0.26	0.25	0.25	0.24
4		0.00	0.52	0.69	0.70	0.59	0.49	0.45	0.44	0.43	0.42	0.41
8		0.00	0.48	0.70	0.91	0.90	0.85	0.80	0.78	0.77	0.76	0.75
16		0.00	0.36	0.59	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.29	0.49	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.26	0.45	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.26	0.44	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.43	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.42	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.24	0.41	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.52	0.54	0.50	0.40	0.32	0.29	0.28	0.28	0.28	0.27
4		0.00	0.54	0.72	0.72	0.60	0.50	0.46	0.44	0.44	0.44	0.42
8		0.00	0.50	0.72	0.92	0.90	0.84	0.80	0.78	0.77	0.76	0.75
16		0.00	0.40	0.60	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.32	0.50	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.29	0.46	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.28	0.44	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.28	0.44	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.28	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.42	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F $p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.51	0.53	0.49	0.38	0.31	0.28	0.27	0.27	0.27	0.27
4		0.00	0.53	0.70	0.71	0.61	0.52	0.48	0.46	0.46	0.45	0.45
8		0.00	0.49	0.71	0.90	0.90	0.84	0.80	0.77	0.77	0.76	0.76
16		0.00	0.38	0.61	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.31	0.52	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.48	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.46	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.27	0.46	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.27	0.45	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.27	0.45	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.54	0.54	0.48	0.36	0.28	0.26	0.25	0.25	0.25	0.26
4		0.00	0.54	0.71	0.71	0.60	0.49	0.45	0.43	0.43	0.42	0.44
8		0.00	0.48	0.71	0.90	0.89	0.83	0.79	0.76	0.76	0.75	0.75
16		0.00	0.36	0.60	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.28	0.49	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.26	0.45	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.25	0.43	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.25	0.43	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.25	0.42	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.26	0.44	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.00	0.51	0.53	0.50	0.38	0.31	0.28	0.27	0.26	0.26	0.25
4		0.00	0.53	0.70	0.71	0.60	0.50	0.46	0.44	0.43	0.43	0.41
8		0.00	0.50	0.71	0.90	0.89	0.82	0.78	0.76	0.75	0.74	0.73
16		0.00	0.38	0.60	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32		0.00	0.31	0.50	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64		0.00	0.28	0.46	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128		0.00	0.27	0.44	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256		0.00	0.26	0.43	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512		0.00	0.26	0.43	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024		0.00	0.25	0.41	0.73	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 3.18 *(preceding page)*: Accuracy of estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 1000 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\dfrac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.19: (following page) Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A single cell, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\dfrac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4			0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8			0.00	0.01	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
16			0.00	0.00	0.02	0.05	0.04	0.00	0.00	0.00	0.00	0.00
32			0.00	0.00	0.00	0.04	0.10	0.07	0.02	0.00	0.00	0.00
64			0.00	0.00	0.00	0.00	0.07	0.14	0.13	0.06	0.02	0.00
128			0.00	0.00	0.00	0.00	0.02	0.13	0.31	0.29	0.18	0.18
256			0.00	0.00	0.00	0.00	0.00	0.06	0.29	0.52	0.45	0.48
512			0.00	0.00	0.00	0.00	0.00	0.02	0.18	0.45	0.58	0.63
1024			0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.48	0.63	0.89

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4			0.00	0.07	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00
8			0.00	0.05	0.12	0.10	0.04	0.01	0.00	0.00	0.00	0.00
16			0.00	0.01	0.10	0.22	0.20	0.11	0.08	0.06	0.04	0.02
32			0.00	0.00	0.04	0.20	0.35	0.32	0.29	0.29	0.29	0.27
64			0.00	0.00	0.01	0.11	0.32	0.55	0.55	0.56	0.57	0.53
128			0.00	0.00	0.00	0.08	0.29	0.55	0.70	0.71	0.68	0.61
256			0.00	0.00	0.00	0.06	0.29	0.56	0.71	0.89	0.91	0.84
512			0.00	0.00	0.00	0.04	0.29	0.57	0.68	0.91	1.00	0.98
1024			0.00	0.00	0.00	0.02	0.27	0.53	0.61	0.84	0.98	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4			0.01	0.11	0.07	0.02	0.01	0.00	0.00	0.00	0.00	0.00
8			0.00	0.07	0.25	0.22	0.15	0.10	0.07	0.06	0.04	0.02
16			0.00	0.02	0.22	0.39	0.37	0.34	0.35	0.37	0.37	0.39
32			0.00	0.01	0.15	0.37	0.53	0.53	0.52	0.52	0.50	0.51
64			0.00	0.00	0.10	0.34	0.53	0.72	0.73	0.70	0.67	0.66
128			0.00	0.00	0.07	0.35	0.52	0.73	0.90	0.89	0.86	0.85
256			0.00	0.00	0.06	0.37	0.52	0.70	0.89	0.97	0.96	0.94
512			0.00	0.00	0.04	0.37	0.50	0.67	0.86	0.96	1.00	1.00
1024			0.00	0.00	0.02	0.39	0.51	0.66	0.85	0.94	1.00	1.00

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.11	0.09	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4			0.09	0.25	0.23	0.15	0.12	0.07	0.06	0.06	0.03	0.01
8			0.01	0.23	0.45	0.43	0.40	0.38	0.38	0.40	0.39	0.39
16			0.00	0.15	0.43	0.60	0.60	0.58	0.58	0.58	0.58	0.59
32			0.00	0.12	0.40	0.60	0.75	0.75	0.70	0.69	0.67	0.67
64			0.00	0.07	0.38	0.58	0.75	0.91	0.90	0.88	0.86	0.85
128			0.00	0.06	0.38	0.58	0.70	0.90	0.99	0.98	0.97	0.97
256			0.00	0.06	0.40	0.58	0.69	0.88	0.98	1.00	1.00	1.00
512			0.00	0.03	0.39	0.58	0.67	0.86	0.97	1.00	1.00	1.00
1024			0.00	0.01	0.39	0.59	0.67	0.85	0.97	1.00	1.00	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.25	0.25	0.11	0.04	0.01	0.00	0.00	0.00	0.00	0.00
4			0.25	0.49	0.48	0.46	0.46	0.46	0.47	0.47	0.45	0.46
8			0.11	0.48	0.69	0.69	0.67	0.63	0.62	0.61	0.60	0.60
16			0.04	0.46	0.69	0.84	0.85	0.82	0.80	0.79	0.77	0.79
32			0.01	0.46	0.67	0.85	0.97	0.96	0.94	0.93	0.92	0.92
64			0.00	0.46	0.63	0.82	0.96	1.00	0.99	0.99	0.99	0.99
128			0.00	0.47	0.62	0.80	0.94	0.99	1.00	1.00	1.00	1.00
256			0.00	0.47	0.61	0.79	0.93	0.99	1.00	1.00	1.00	1.00
512			0.00	0.45	0.60	0.77	0.92	0.99	1.00	1.00	1.00	1.00
1024			0.00	0.46	0.60	0.79	0.92	0.99	1.00	1.00	1.00	1.00

F $p_{smc} = 0.6$

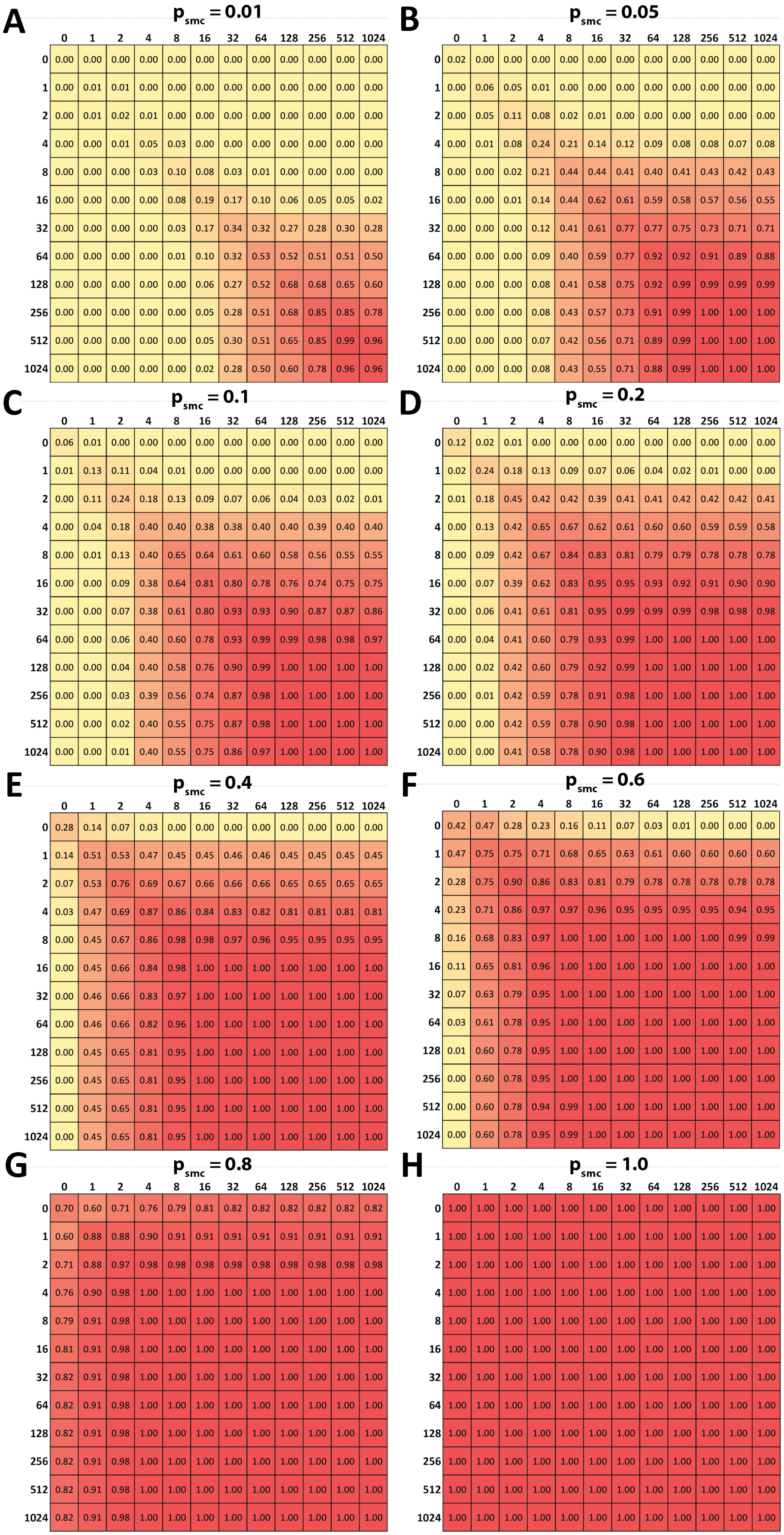
	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.42	0.32	0.25	0.23	0.19	0.16	0.11	0.04	0.00	0.00
4			0.32	0.77	0.70	0.72	0.70	0.68	0.65	0.62	0.59	0.59
8			0.25	0.70	0.89	0.88	0.84	0.83	0.81	0.80	0.80	0.80
16			0.23	0.72	0.88	0.96	0.96	0.95	0.95	0.94	0.94	0.94
32			0.19	0.70	0.84	0.96	1.00	1.00	0.99	0.99	0.99	0.99
64			0.16	0.68	0.83	0.95	1.00	1.00	1.00	1.00	1.00	1.00
128			0.11	0.65	0.81	0.95	0.99	1.00	1.00	1.00	1.00	1.00
256			0.04	0.62	0.80	0.94	0.99	1.00	1.00	1.00	1.00	1.00
512			0.00	0.59	0.80	0.94	0.99	1.00	1.00	1.00	1.00	1.00
1024			0.00	0.59	0.80	0.94	0.99	1.00	1.00	1.00	1.00	1.00

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			0.68	0.59	0.72	0.75	0.77	0.80	0.81	0.81	0.81	0.81
4			0.59	0.91	0.91	0.92	0.93	0.93	0.94	0.94	0.94	0.94
8			0.72	0.91	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
16			0.75	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32			0.77	0.93	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64			0.80	0.93	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128			0.81	0.94	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256			0.81	0.94	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512			0.81	0.94	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024			0.81	0.94	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0												
1												
2			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00



A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.01	0.05	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.03	0.10	0.08	0.03	0.01	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.08	0.19	0.17	0.10	0.06	0.05	0.05	0.02
32	0.00	0.00	0.00	0.00	0.03	0.17	0.34	0.32	0.27	0.28	0.30	0.28
64	0.00	0.00	0.00	0.00	0.01	0.10	0.32	0.53	0.52	0.51	0.51	0.50
128	0.00	0.00	0.00	0.00	0.00	0.06	0.27	0.52	0.68	0.68	0.65	0.60
256	0.00	0.00	0.00	0.00	0.00	0.05	0.28	0.51	0.68	0.85	0.85	0.78
512	0.00	0.00	0.00	0.00	0.00	0.05	0.30	0.51	0.65	0.85	0.99	0.96
1024	0.00	0.00	0.00	0.00	0.00	0.02	0.28	0.50	0.60	0.78	0.96	0.96

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.06	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.05	0.11	0.08	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.01	0.08	0.24	0.21	0.14	0.12	0.09	0.08	0.08	0.07	0.08
8	0.00	0.00	0.02	0.21	0.44	0.44	0.41	0.40	0.41	0.43	0.42	0.43
16	0.00	0.00	0.01	0.14	0.44	0.62	0.61	0.59	0.58	0.57	0.56	0.55
32	0.00	0.00	0.00	0.12	0.41	0.61	0.77	0.77	0.75	0.73	0.71	0.71
64	0.00	0.00	0.00	0.09	0.40	0.59	0.77	0.92	0.92	0.91	0.89	0.88
128	0.00	0.00	0.00	0.08	0.41	0.58	0.75	0.92	0.99	0.99	0.99	0.99
256	0.00	0.00	0.00	0.08	0.43	0.57	0.73	0.91	0.99	1.00	1.00	1.00
512	0.00	0.00	0.00	0.07	0.42	0.56	0.71	0.89	0.99	1.00	1.00	1.00
1024	0.00	0.00	0.00	0.08	0.43	0.55	0.71	0.88	0.99	1.00	1.00	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.01	0.13	0.11	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.11	0.24	0.18	0.13	0.09	0.07	0.06	0.04	0.03	0.02	0.01
4	0.00	0.04	0.18	0.40	0.40	0.38	0.38	0.40	0.40	0.39	0.40	0.40
8	0.00	0.01	0.13	0.40	0.65	0.64	0.61	0.60	0.58	0.56	0.55	0.55
16	0.00	0.00	0.09	0.38	0.64	0.81	0.80	0.78	0.76	0.74	0.75	0.75
32	0.00	0.00	0.07	0.38	0.61	0.80	0.93	0.93	0.90	0.87	0.87	0.86
64	0.00	0.00	0.06	0.40	0.60	0.78	0.93	0.99	0.99	0.98	0.98	0.97
128	0.00	0.00	0.04	0.40	0.58	0.76	0.90	0.99	1.00	1.00	1.00	1.00
256	0.00	0.00	0.03	0.39	0.56	0.74	0.87	0.98	1.00	1.00	1.00	1.00
512	0.00	0.00	0.02	0.40	0.55	0.75	0.87	0.98	1.00	1.00	1.00	1.00
1024	0.00	0.00	0.01	0.40	0.55	0.75	0.86	0.97	1.00	1.00	1.00	1.00

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.12	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.02	0.24	0.18	0.13	0.09	0.07	0.06	0.04	0.02	0.01	0.00	0.00
2	0.01	0.18	0.45	0.42	0.42	0.39	0.41	0.41	0.42	0.42	0.42	0.41
4	0.00	0.13	0.42	0.65	0.67	0.62	0.61	0.60	0.60	0.59	0.59	0.58
8	0.00	0.09	0.42	0.67	0.84	0.83	0.81	0.79	0.79	0.78	0.78	0.78
16	0.00	0.07	0.39	0.62	0.83	0.95	0.95	0.93	0.92	0.91	0.90	0.90
32	0.00	0.06	0.41	0.61	0.81	0.95	0.99	0.99	0.99	0.98	0.98	0.98
64	0.00	0.04	0.41	0.60	0.79	0.93	0.99	1.00	1.00	1.00	1.00	1.00
128	0.00	0.02	0.42	0.60	0.79	0.92	0.99	1.00	1.00	1.00	1.00	1.00
256	0.00	0.01	0.42	0.59	0.78	0.91	0.98	1.00	1.00	1.00	1.00	1.00
512	0.00	0.00	0.42	0.59	0.78	0.90	0.98	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.00	0.41	0.58	0.78	0.90	0.98	1.00	1.00	1.00	1.00	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.28	0.14	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.14	0.51	0.53	0.47	0.45	0.45	0.46	0.46	0.45	0.45	0.45	0.45
2	0.07	0.53	0.76	0.69	0.67	0.66	0.66	0.66	0.65	0.65	0.65	0.65
4	0.03	0.47	0.69	0.87	0.86	0.84	0.83	0.82	0.81	0.81	0.81	0.81
8	0.00	0.45	0.67	0.86	0.98	0.98	0.97	0.96	0.95	0.95	0.95	0.95
16	0.00	0.45	0.66	0.84	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.00	0.46	0.66	0.83	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.00	0.46	0.66	0.82	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.00	0.45	0.65	0.81	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.00	0.45	0.65	0.81	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.00	0.45	0.65	0.81	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.45	0.65	0.81	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F $p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.42	0.47	0.28	0.23	0.16	0.11	0.07	0.03	0.01	0.00	0.00	0.00
1	0.47	0.75	0.75	0.71	0.68	0.65	0.63	0.61	0.60	0.60	0.60	0.60
2	0.28	0.75	0.90	0.86	0.83	0.81	0.79	0.78	0.78	0.78	0.78	0.78
4	0.23	0.71	0.86	0.97	0.97	0.96	0.95	0.95	0.95	0.95	0.94	0.95
8	0.16	0.68	0.83	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
16	0.11	0.65	0.81	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.07	0.63	0.79	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.03	0.61	0.78	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.01	0.60	0.78	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.00	0.60	0.78	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.00	0.60	0.78	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.60	0.78	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.70	0.60	0.71	0.76	0.79	0.81	0.82	0.82	0.82	0.82	0.82	0.82
1	0.60	0.88	0.88	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
2	0.71	0.88	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
4	0.76	0.90	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.79	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.81	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 3.20 ***(preceding page)*****: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability**. A pool of 5 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\dfrac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bulk}}}$$

and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.21: (following page) Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 10 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\dfrac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bulk}}}$$

and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.01	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.02	0.06	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.03	0.11	0.09	0.03	0.01	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.09	0.22	0.19	0.12	0.07	0.05	0.04	0.01	0.00
16	0.00	0.00	0.00	0.03	0.19	0.36	0.35	0.29	0.29	0.29	0.27	0.29
32	0.00	0.00	0.00	0.01	0.12	0.35	0.55	0.56	0.53	0.52	0.51	0.53
64	0.00	0.00	0.00	0.00	0.07	0.29	0.56	0.76	0.74	0.71	0.69	0.70
128	0.00	0.00	0.00	0.00	0.05	0.29	0.53	0.74	0.88	0.87	0.83	0.83
256	0.00	0.00	0.00	0.00	0.04	0.29	0.52	0.71	0.87	0.98	0.96	0.96
512	0.00	0.00	0.00	0.00	0.01	0.27	0.51	0.69	0.83	0.96	1.00	1.00
1024	0.00	0.00	0.00	0.00	0.00	0.29	0.53	0.70	0.83	0.96	1.00	1.00

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.03	0.12	0.09	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.01	0.09	0.23	0.19	0.13	0.10	0.09	0.07	0.06	0.06	0.06	0.05
4	0.00	0.03	0.19	0.44	0.42	0.40	0.41	0.42	0.42	0.43	0.44	0.43
8	0.00	0.01	0.13	0.42	0.64	0.64	0.61	0.59	0.57	0.56	0.55	0.55
16	0.00	0.00	0.10	0.40	0.64	0.80	0.80	0.76	0.73	0.73	0.73	0.71
32	0.00	0.00	0.09	0.41	0.61	0.80	0.93	0.92	0.89	0.88	0.87	0.85
64	0.00	0.00	0.07	0.42	0.59	0.76	0.92	0.99	0.99	0.98	0.98	0.97
128	0.00	0.00	0.06	0.42	0.57	0.73	0.89	0.99	1.00	1.00	1.00	0.99
256	0.00	0.00	0.06	0.43	0.56	0.73	0.88	0.98	1.00	1.00	1.00	1.00
512	0.00	0.00	0.06	0.44	0.55	0.73	0.87	0.98	1.00	1.00	1.00	1.00
1024	0.00	0.00	0.05	0.43	0.55	0.71	0.85	0.97	0.99	1.00	1.00	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.06	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.05	0.27	0.23	0.16	0.14	0.10	0.09	0.08	0.07	0.06	0.05	0.06
2	0.02	0.23	0.39	0.39	0.38	0.36	0.37	0.39	0.39	0.39	0.39	0.39
4	0.00	0.16	0.39	0.66	0.65	0.61	0.59	0.58	0.58	0.57	0.57	0.57
8	0.00	0.14	0.38	0.65	0.82	0.80	0.77	0.75	0.74	0.73	0.72	0.72
16	0.00	0.10	0.36	0.61	0.80	0.93	0.93	0.91	0.90	0.89	0.88	0.87
32	0.00	0.09	0.37	0.59	0.77	0.93	0.99	0.98	0.98	0.97	0.97	0.96
64	0.00	0.08	0.39	0.58	0.75	0.91	0.98	1.00	1.00	1.00	1.00	1.00
128	0.00	0.07	0.39	0.58	0.74	0.90	0.98	1.00	1.00	1.00	1.00	1.00
256	0.00	0.06	0.39	0.57	0.73	0.89	0.97	1.00	1.00	1.00	1.00	1.00
512	0.00	0.05	0.39	0.57	0.72	0.88	0.97	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.06	0.39	0.57	0.72	0.87	0.96	1.00	1.00	1.00	1.00	1.00

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.17	0.15	0.11	0.06	0.05	0.04	0.03	0.02	0.01	0.00	0.00	0.00
1	0.15	0.48	0.44	0.42	0.42	0.43	0.44	0.44	0.45	0.45	0.45	0.45
2	0.11	0.44	0.66	0.64	0.61	0.60	0.59	0.60	0.60	0.59	0.60	0.59
4	0.06	0.42	0.64	0.84	0.83	0.81	0.79	0.79	0.79	0.79	0.78	0.78
8	0.05	0.42	0.61	0.83	0.95	0.94	0.93	0.92	0.91	0.91	0.90	0.91
16	0.04	0.43	0.60	0.81	0.94	0.99	0.99	0.99	0.98	0.98	0.97	0.97
32	0.03	0.44	0.59	0.79	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00
64	0.02	0.44	0.60	0.79	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00
128	0.01	0.45	0.60	0.79	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00
256	0.00	0.45	0.59	0.79	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00
512	0.00	0.45	0.60	0.78	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.45	0.59	0.78	0.91	0.97	1.00	1.00	1.00	1.00	1.00	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.33	0.36	0.33	0.32	0.32	0.32	0.32	0.32	0.31	0.31	0.31	0.31
1	0.36	0.68	0.69	0.67	0.65	0.63	0.63	0.62	0.62	0.62	0.62	0.62
2	0.33	0.69	0.86	0.87	0.85	0.84	0.83	0.83	0.83	0.82	0.82	0.82
4	0.32	0.67	0.87	0.97	0.97	0.95	0.94	0.93	0.93	0.93	0.93	0.92
8	0.32	0.65	0.85	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.32	0.63	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.32	0.63	0.83	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.32	0.62	0.83	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.31	0.62	0.83	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.31	0.62	0.82	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.31	0.62	0.82	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.31	0.62	0.82	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F $p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.59	0.59	0.56	0.54	0.49	0.47	0.44	0.42	0.42	0.42	0.42	0.42
1	0.59	0.87	0.89	0.87	0.85	0.85	0.84	0.84	0.84	0.84	0.84	0.84
2	0.56	0.89	0.97	0.97	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
4	0.54	0.87	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.49	0.85	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.47	0.85	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.44	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.42	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.42	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.42	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.42	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.42	0.84	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.81	0.84	0.84	0.87	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.89
1	0.84	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2	0.84	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.87	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.88	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.02	0.09	0.07	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.07	0.14	0.12	0.06	0.03	0.01	0.00	0.00	0.00	0.00	0.00
4	0.00	0.02	0.12	0.28	0.27	0.21	0.18	0.17	0.17	0.16	0.17	0.17
8	0.00	0.00	0.06	0.27	0.48	0.47	0.45	0.45	0.46	0.45	0.46	0.46
16	0.00	0.00	0.03	0.21	0.47	0.65	0.66	0.61	0.59	0.57	0.56	0.56
32	0.00	0.00	0.01	0.18	0.45	0.66	0.84	0.83	0.80	0.77	0.76	0.76
64	0.00	0.00	0.00	0.17	0.45	0.61	0.83	0.95	0.95	0.92	0.91	0.91
128	0.00	0.00	0.00	0.17	0.46	0.59	0.80	0.95	1.00	0.99	0.99	0.99
256	0.00	0.00	0.00	0.16	0.45	0.57	0.77	0.92	0.99	1.00	1.00	1.00
512	0.00	0.00	0.00	0.17	0.46	0.56	0.76	0.91	0.99	1.00	1.00	1.00
1024	0.00	0.00	0.00	0.17	0.46	0.56	0.76	0.91	0.99	1.00	1.00	1.00

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.10	0.09	0.05	0.03	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
1	0.09	0.38	0.36	0.33	0.31	0.31	0.31	0.32	0.32	0.32	0.32	0.32
2	0.05	0.36	0.54	0.53	0.52	0.51	0.51	0.51	0.50	0.50	0.50	0.49
4	0.03	0.33	0.53	0.71	0.72	0.67	0.65	0.64	0.63	0.63	0.62	0.62
8	0.02	0.31	0.52	0.72	0.90	0.89	0.86	0.85	0.85	0.84	0.84	0.84
16	0.01	0.31	0.51	0.67	0.89	0.97	0.97	0.95	0.95	0.94	0.93	0.93
32	0.01	0.31	0.51	0.65	0.86	0.97	1.00	1.00	0.99	0.99	0.99	0.99
64	0.00	0.32	0.51	0.64	0.85	0.95	1.00	1.00	1.00	1.00	1.00	1.00
128	0.00	0.32	0.50	0.63	0.85	0.95	0.99	1.00	1.00	1.00	1.00	1.00
256	0.00	0.32	0.50	0.63	0.84	0.94	0.99	1.00	1.00	1.00	1.00	1.00
512	0.00	0.32	0.50	0.62	0.84	0.93	0.99	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.32	0.49	0.62	0.84	0.93	0.99	1.00	1.00	1.00	1.00	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.18	0.20	0.18	0.17	0.16	0.16	0.16	0.16	0.16	0.16	0.17	0.17
1	0.20	0.57	0.58	0.55	0.55	0.55	0.54	0.54	0.54	0.54	0.54	0.54
2	0.18	0.58	0.75	0.74	0.70	0.68	0.66	0.65	0.65	0.65	0.64	0.64
4	0.17	0.55	0.74	0.89	0.88	0.86	0.83	0.82	0.82	0.82	0.82	0.81
8	0.16	0.55	0.70	0.88	0.98	0.98	0.97	0.96	0.95	0.95	0.95	0.95
16	0.16	0.55	0.68	0.86	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.16	0.54	0.66	0.83	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.16	0.54	0.65	0.82	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.16	0.54	0.65	0.82	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.16	0.54	0.65	0.82	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.17	0.54	0.64	0.82	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.17	0.54	0.64	0.81	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.30	0.36	0.34	0.33	0.33	0.33	0.33	0.32	0.32	0.32	0.32	0.32
1	0.36	0.78	0.77	0.73	0.71	0.70	0.69	0.69	0.69	0.69	0.69	0.69
2	0.34	0.77	0.91	0.90	0.88	0.86	0.86	0.85	0.85	0.85	0.85	0.85
4	0.33	0.73	0.90	0.98	0.98	0.97	0.96	0.96	0.95	0.95	0.95	0.95
8	0.33	0.71	0.88	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.33	0.70	0.86	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.33	0.69	0.86	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.32	0.69	0.85	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.32	0.69	0.85	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.32	0.69	0.85	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.32	0.69	0.85	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.32	0.69	0.85	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.51	0.62	0.60	0.60	0.59	0.59	0.58	0.58	0.58	0.58	0.58	0.58
1	0.62	0.95	0.95	0.94	0.92	0.92	0.91	0.91	0.91	0.91	0.91	0.91
2	0.60	0.95	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
4	0.60	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.59	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.59	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.58	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.58	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.58	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.58	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.58	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.58	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F $p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.71	0.78	0.76	0.74	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
1	0.78	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2	0.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.74	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.90	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
1	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 3.22 *(preceding page)*: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 30 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

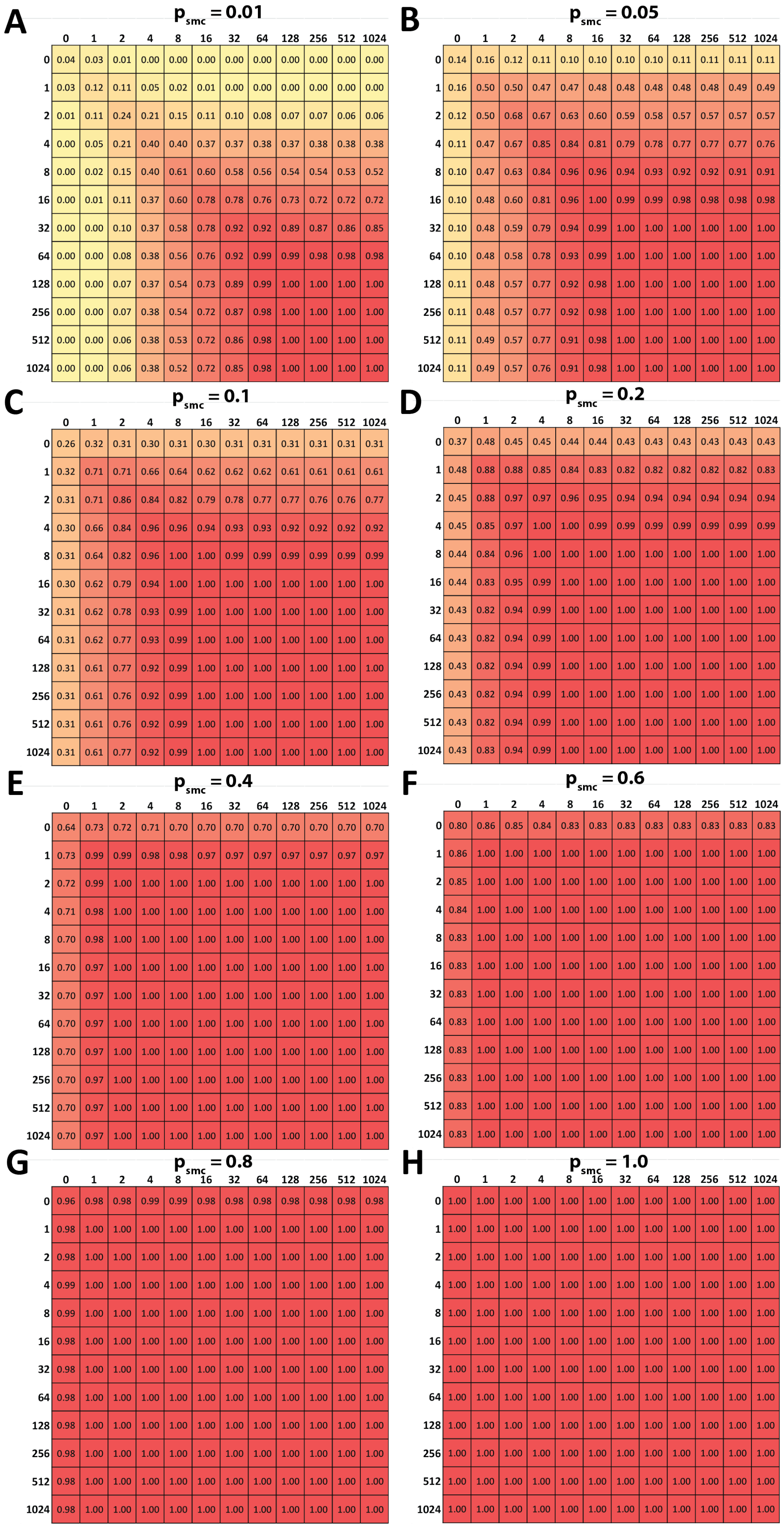
$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.23: (following page) Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 50 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.03	0.12	0.11	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
2	0.01	0.11	0.24	0.21	0.15	0.11	0.10	0.08	0.07	0.07	0.06	0.06
4	0.00	0.05	0.21	0.40	0.40	0.37	0.37	0.38	0.37	0.38	0.38	0.38
8	0.00	0.02	0.15	0.40	0.61	0.60	0.58	0.56	0.54	0.54	0.53	0.52
16	0.00	0.01	0.11	0.37	0.60	0.78	0.78	0.76	0.73	0.72	0.72	0.72
32	0.00	0.00	0.10	0.37	0.58	0.78	0.92	0.92	0.89	0.87	0.86	0.85
64	0.00	0.00	0.08	0.38	0.56	0.76	0.92	0.99	0.99	0.98	0.98	0.98
128	0.00	0.00	0.07	0.37	0.54	0.73	0.89	0.99	1.00	1.00	1.00	1.00
256	0.00	0.00	0.07	0.38	0.54	0.72	0.87	0.98	1.00	1.00	1.00	1.00
512	0.00	0.00	0.06	0.38	0.53	0.72	0.86	0.98	1.00	1.00	1.00	1.00
1024	0.00	0.00	0.06	0.38	0.52	0.72	0.85	0.98	1.00	1.00	1.00	1.00

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.14	0.16	0.12	0.11	0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.11
1	0.16	0.50	0.50	0.47	0.47	0.48	0.48	0.48	0.48	0.48	0.49	0.49
2	0.12	0.50	0.68	0.67	0.63	0.60	0.59	0.58	0.57	0.57	0.57	0.57
4	0.11	0.47	0.67	0.85	0.84	0.81	0.79	0.78	0.77	0.77	0.77	0.76
8	0.10	0.47	0.63	0.84	0.96	0.96	0.94	0.93	0.92	0.92	0.91	0.91
16	0.10	0.48	0.60	0.81	0.96	1.00	0.99	0.99	0.98	0.98	0.98	0.98
32	0.10	0.48	0.59	0.79	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00
64	0.10	0.48	0.58	0.78	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00
128	0.11	0.48	0.57	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00
256	0.11	0.48	0.57	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00
512	0.11	0.49	0.57	0.77	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.11	0.49	0.57	0.76	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.26	0.32	0.31	0.30	0.31	0.30	0.31	0.31	0.31	0.31	0.31	0.31
1	0.32	0.71	0.71	0.66	0.64	0.62	0.62	0.62	0.61	0.61	0.61	0.61
2	0.31	0.71	0.86	0.84	0.82	0.79	0.78	0.77	0.77	0.76	0.76	0.77
4	0.30	0.66	0.84	0.96	0.96	0.94	0.93	0.93	0.92	0.92	0.92	0.92
8	0.31	0.64	0.82	0.96	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
16	0.30	0.62	0.79	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.31	0.62	0.78	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.31	0.62	0.77	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.31	0.61	0.77	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.31	0.61	0.76	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.31	0.61	0.76	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.31	0.61	0.77	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.37	0.48	0.45	0.45	0.44	0.44	0.43	0.43	0.43	0.43	0.43	0.43
1	0.48	0.88	0.88	0.85	0.84	0.83	0.82	0.82	0.82	0.82	0.82	0.83
2	0.45	0.88	0.97	0.97	0.96	0.95	0.94	0.94	0.94	0.94	0.94	0.94
4	0.45	0.85	0.97	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99
8	0.44	0.84	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.44	0.83	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.43	0.82	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.43	0.82	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.43	0.82	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.43	0.82	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.43	0.82	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.43	0.83	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.64	0.73	0.72	0.71	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
1	0.73	0.99	0.99	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97
2	0.72	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.70	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.80	0.86	0.85	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
1	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G

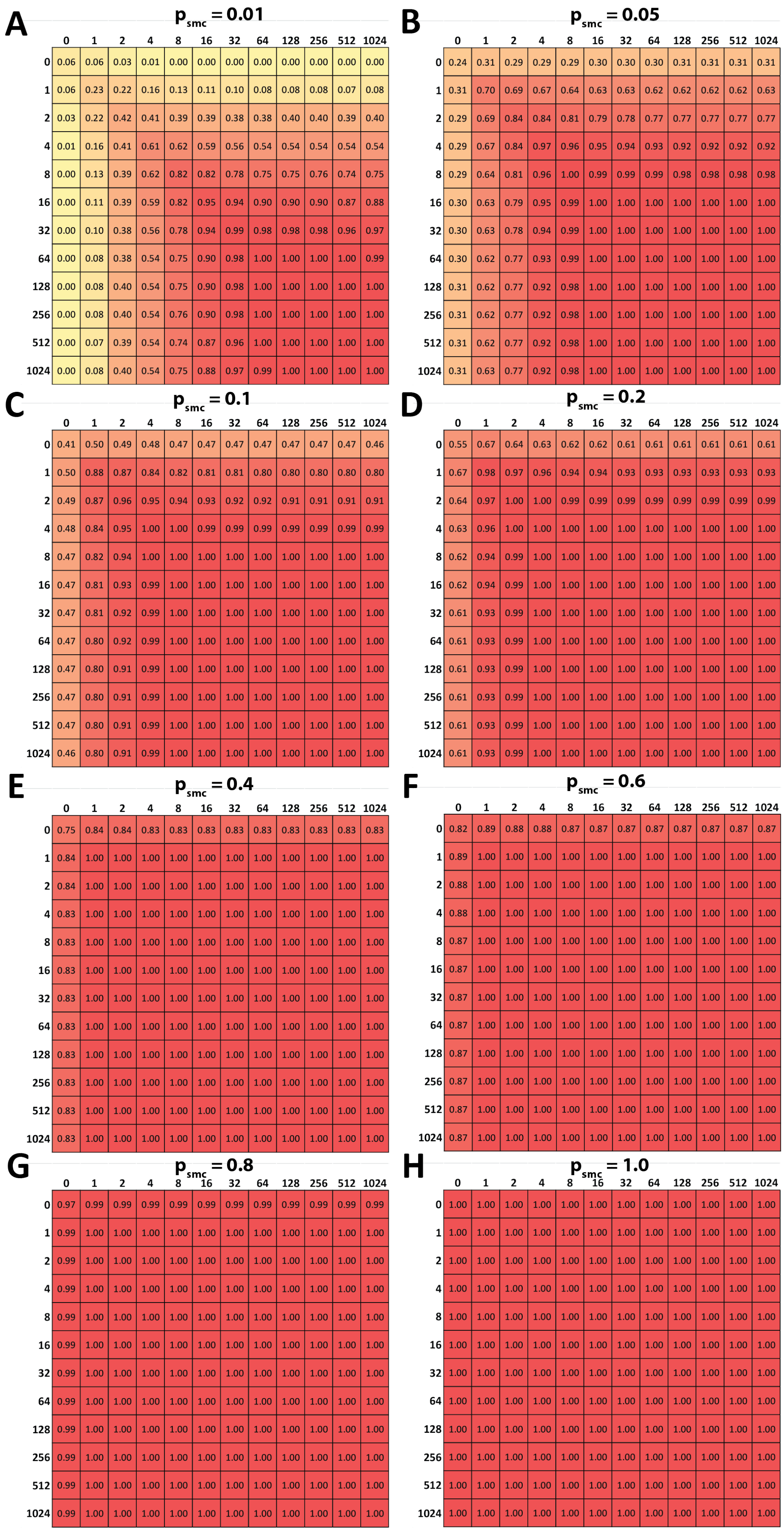
$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.96	0.98	0.98	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98
1	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00



A

$p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.06	0.06	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.06	0.23	0.22	0.16	0.13	0.11	0.10	0.08	0.08	0.08	0.07	0.08
2	0.03	0.22	0.42	0.41	0.39	0.39	0.38	0.38	0.40	0.40	0.39	0.40
4	0.01	0.16	0.41	0.61	0.62	0.59	0.56	0.54	0.54	0.54	0.54	0.54
8	0.00	0.13	0.39	0.62	0.82	0.82	0.78	0.75	0.75	0.76	0.74	0.75
16	0.00	0.11	0.39	0.59	0.82	0.95	0.94	0.90	0.90	0.90	0.87	0.88
32	0.00	0.10	0.38	0.56	0.78	0.94	0.99	0.98	0.98	0.98	0.96	0.97
64	0.00	0.08	0.38	0.54	0.75	0.90	0.98	1.00	1.00	1.00	1.00	0.99
128	0.00	0.08	0.40	0.54	0.75	0.90	0.98	1.00	1.00	1.00	1.00	1.00
256	0.00	0.08	0.40	0.54	0.76	0.90	0.98	1.00	1.00	1.00	1.00	1.00
512	0.00	0.07	0.39	0.54	0.74	0.87	0.96	1.00	1.00	1.00	1.00	1.00
1024	0.00	0.08	0.40	0.54	0.75	0.88	0.97	0.99	1.00	1.00	1.00	1.00

B

$p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.24	0.31	0.29	0.29	0.29	0.30	0.30	0.30	0.31	0.31	0.31	0.31
1	0.31	0.70	0.69	0.67	0.64	0.63	0.63	0.62	0.62	0.62	0.62	0.63
2	0.29	0.69	0.84	0.84	0.81	0.79	0.78	0.77	0.77	0.77	0.77	0.77
4	0.29	0.67	0.84	0.97	0.96	0.95	0.94	0.93	0.92	0.92	0.92	0.92
8	0.29	0.64	0.81	0.96	1.00	0.99	0.99	0.99	0.98	0.98	0.98	0.98
16	0.30	0.63	0.79	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.30	0.63	0.78	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.30	0.62	0.77	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.31	0.62	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.31	0.62	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.31	0.62	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.31	0.63	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

C

$p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.41	0.50	0.49	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.46
1	0.50	0.88	0.87	0.84	0.82	0.81	0.81	0.80	0.80	0.80	0.80	0.80
2	0.49	0.87	0.96	0.95	0.94	0.93	0.92	0.92	0.91	0.91	0.91	0.91
4	0.48	0.84	0.95	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99
8	0.47	0.82	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.47	0.81	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.47	0.81	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.47	0.80	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.47	0.80	0.91	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.47	0.80	0.91	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.47	0.80	0.91	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.46	0.80	0.91	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

D

$p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.55	0.67	0.64	0.63	0.62	0.62	0.61	0.61	0.61	0.61	0.61	0.61
1	0.67	0.98	0.97	0.96	0.94	0.94	0.93	0.93	0.93	0.93	0.93	0.93
2	0.64	0.97	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
4	0.63	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.62	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.62	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.61	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.61	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.61	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.61	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.61	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.61	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

E

$p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.75	0.84	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
1	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F

$p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.82	0.89	0.88	0.88	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
1	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G

$p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
1	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H

$p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 3.24 ***(preceding page)*****: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability**. A pool of 100 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

Figure 3.25: (following page) Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 1000 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bulk}}}$$

and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.

A $p_{smc} = 0.01$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.34	0.45	0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.42	0.43
1	0.45	0.84	0.84	0.80	0.78	0.77	0.76	0.76	0.75	0.75	0.75	0.75
2	0.44	0.84	0.96	0.95	0.93	0.92	0.91	0.90	0.90	0.90	0.90	0.90
4	0.43	0.80	0.95	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.97	0.97
8	0.43	0.78	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.43	0.77	0.92	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.43	0.76	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.43	0.76	0.90	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.43	0.75	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.43	0.75	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.42	0.75	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.43	0.75	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

B $p_{smc} = 0.05$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.68	0.79	0.78	0.77	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
1	0.79	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.76	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

C $p_{smc} = 0.1$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.75	0.85	0.85	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
1	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

D $p_{smc} = 0.2$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.87	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
1	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

E $p_{smc} = 0.4$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.88	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
1	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

F $p_{smc} = 0.6$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.91	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
1	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

G $p_{smc} = 0.8$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
1	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

H $p_{smc} = 1.0$

	0	1	2	4	8	16	32	64	128	256	512	1024
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
128	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
256	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
512	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1024	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

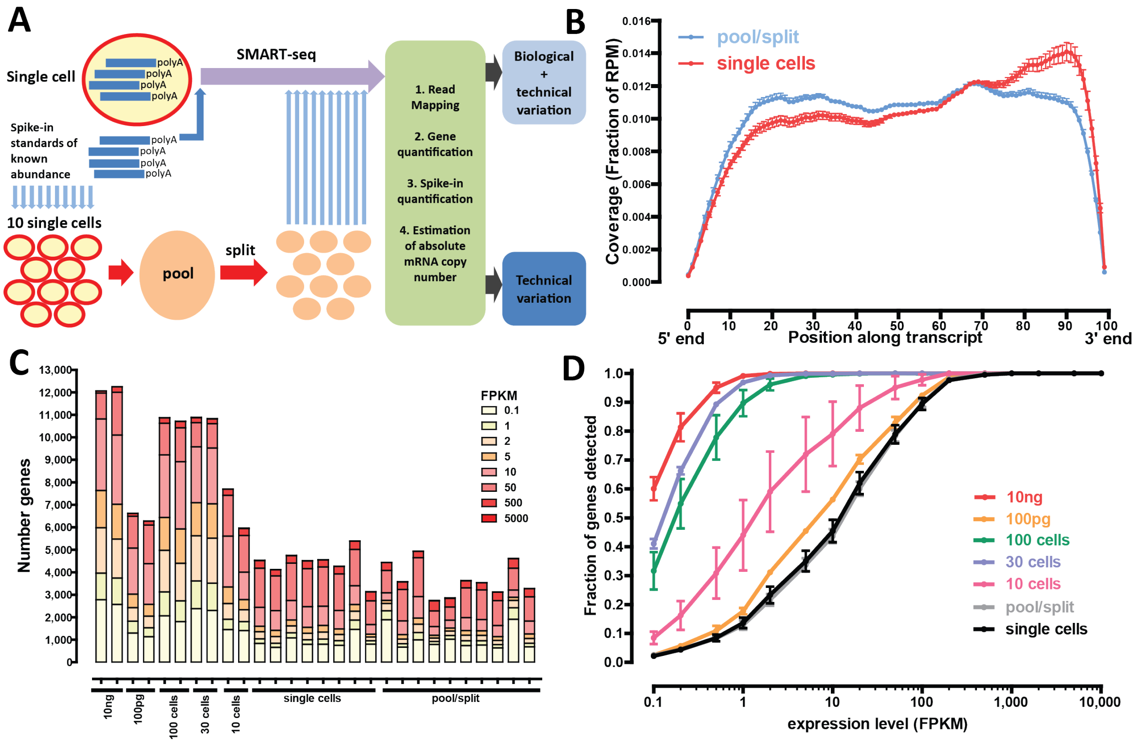


Figure 3.26: Simulated and measured transcriptome profiles from individual cells and small cell pools. (A) Number of detected genes in simulated datasets as a function of the number of cells pooled and the single molecule capture efficiency (p_{smc}) (assuming 100,000 mRNA molecules per cell). See Supplementary Figure 1 for full details. (B and C) Accuracy of gene expression estimation as a function of the number of cells pooled and the single molecule capture efficiency; $p_{smc} = 0.1$ in (B) and $p_{smc} = 0.8$ in (C), 100,000 mRNA molecules per cell assumed. Shown is the fraction of genes at the indicated expression levels in FPKM, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value, after modeling the stochasticity due to the single-molecule capture efficiency of the library-building protocol. See the Methods section and Supplementary Figures 2-11 for full details. Note that the simulation is intended to illuminate the relative effects of the various parameters studied, and the absolute numbers of genes should not be directly compared to the real-life data shown in (G). (D) Experimental design. Single cells are combined with spike-in quantification standards and SMART-seq libraries are generated. In parallel, multiple single cells are pooled together and combined with spikes, then lysed and split into the same number of reactions and converted into SMART-seq libraries. Libraries are then sequenced, data processed computationally and estimates for the absolute number of copies per cell are derived based on the spikes. Variation in pool/split experiments is due to technical stochasticity, while variation in single-cell libraries is a combination of biological variation and technical noise. (E) Uniformity of transcript coverage. Shown is the average coverage along the length of an mRNA for single cells and pool/split experiments. Only mRNAs longer than 1kb from genes with a single annotated isoform in the refSeq annotation set were included. See Supplementary Figure 29 for more details. (F) Number of detected protein coding genes for libraries built from 10ng and 100pg of polyA RNA, pools of 100, 30 and 10 cells, representative pool/split experiments (individually and summed across all libraries) and representative single cells (individually and summed across all libraries). (G) Fraction of genes from 100 ng bulk polyA+ RNA libraries that were detected in pools of 100, 30 or 10 cells, 100pg of polyA+ RNA, pools/split experiments and single-cells. FPKM is shown on the X-axis.

Table 3.1: Initial amounts of spiked-in sequences in absolute number of RNA copies. Note that two more spikes, "Lambda 9786 clone F" (9786bp) and "Lambda 11300 clone G" (11290bp) were included in libraries, however, they exhibit highly non-uniform read coverage leading to unreliable quantification estimates and were thus excluded.

Spike-in	Libraries 12515-12543	Libraries 12818-13303
AGP23	100	5
AP2	5	50
EPR1	20	10
OBF5	10	500
PDF1	40	20
VATG3	5000	5000

Beyond such already appreciated heterogeneity lie currently unknown cell-to-cell differences with biological implications for defining cell states, metabolic function, and in complex tissues, cell identity.

Measuring RNA transcripts in single cells is now done in multiple ways, and similar conclusions about variability are emerging from the higher sensitivity methods. For individual genes, Single Molecule RNA Fluorescence In Situ Hybridization (SM-RNA FISH) is highly informative (Femino et al. 1998; Raj et al. 2008), and multiplexed versions now enable multiple genes to be measured in parallel (Lubeck & Cai 2012). A major advantage of SM-RNA FISH is the ability to accurately count the absolute number of transcripts in a cell. A second and older approach is multiplexed single-cell RT-qPCR (Cornelison & Wold 1997), which has now been advanced to increasingly high throughput formats (White et al. 2011; Sanchez-Freire et al. 2012, Livak et al. 2013). It produces semi-quantitative relative comparisons between individual cells. However, neither SM-RNA-FISH nor the current forms of multiplex RT-qPCR cover the entire transcriptome or have the single-nucleotide resolution needed to study fine-structure features of gene expression such as allele specificity, RNA editing and alternative splicing.

To address these and other limitations, elegant methods have recently been developed for performing RNA-seq with very small amounts of RNA, down to the level of individual cells. These are broadly referred to as "single-cell RNA-seq" (Tang et al. 2009; Tang et al. 2010; Tang et al. 2011; Ozsolak et al. 2010; Islam et al. 2011; Hashimshony et al. 2012; Qiu et al. 2012; Ramsköld & Luo et al. 2012; Brouilette et al. 2012; Pan et al. 2012, Cann et al. 2012). Despite these significant advances, there are substantial shortcomings in these methods, and a robust method for comprehensive and accurate measurement of the transcriptome of a single cell is not yet available.

A particular challenge for single-cell methods is the efficiency and uniformity with which each mRNA in copied into cDNA, and ultimately represented in the library. This challenge intersects in crucial ways with transcriptome structure. Specifically, thousands of genes are expressed in the range of 1 to 30 mRNA copies per cell, including many essential mRNAs (for example, key transcription factors, Zenklusen et al. 2008). Even lower transcript levels, averaging < 1 mRNA per cell on the population level, are now being reliably detected by RNA-seq. This raises questions whether very rare RNAs represent background biological noise, or alternatively, are functional in only a small fraction of cells. Single-cell RNA-seq has the potential to address these issues, but their resolution depends on how faithfully and efficiently RNAs are captured and represented in sequencing libraries (referred to throughout as the "single-molecule capture efficiency", p_{smc}). In addition, the uniformity of transcript coverage in early single-cell RNA-seq protocols has typically been heavily biased towards the 3' end, which affects both gene expression estimates and the ability to analyze alternative splicing, RNA editing and allelic bias.

A second major use for single-cell RNA-seq is the transcriptomic characterization of rare cells. The human body consists of hundreds of dis-

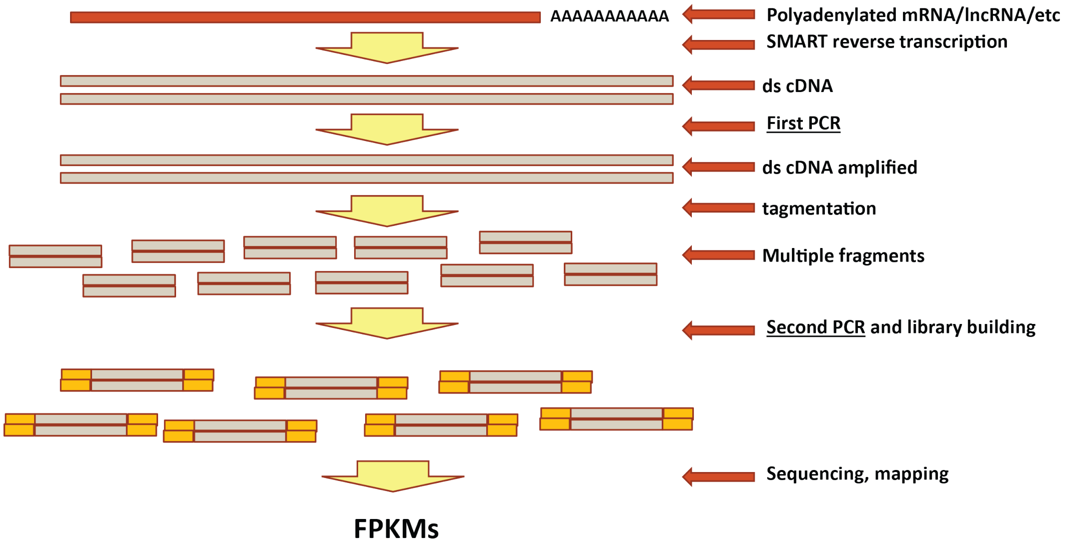


Figure 3.27: Outline of the single-cell SMART-seq RNA-seq library generation workflow.

tinct cell types, plus large numbers of neuronal and transient developmental cell types. Many of these are numerically minor components of complex tissues, making them inaccessible to standard methods relying on large RNA inputs. Isolation of single cells based on the cell surface markers or using microdissection coupled with single-cell RNA-seq could fill this gap in our understanding of gene expression patterns in complex multicellular organisms. However, the feasibility of this approach also depends on the experimental robustness of single-cell RNA-seq protocols. Alternatively, single-cell resolution may not be absolutely required for this purpose and small pools of cells may be sufficient to characterize rare cell type transcriptomes; an open unresolved question is how small such pools can be to adequately meet that goal.

We addressed the issues highlighted above using the SMART-seq protocol (Ramsköld & Luo et al. 2012) to measure the transcriptome of single cells and small cell pools from the GM12878 lymphoblastoid cell line. This line is derived from the NA12878 individual, for which a fully sequenced genome with completely phased heterozygous single nucleotide polymorphisms (SNPs) and indels is available (1000 Genomes Project Consortium 2012). GM12878 cells have also been the subject of an extensive functional genomic characterization by the ENCODE Consortium (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012) and have been used in prior studies aiming at characterizing allele-biased gene expression and transcription factor occupancy (Rozowsky et al. 2011; Reddy et al, 2012).

Using spike-in quantification standards of known abundance (Mortazavi & Williams et al. 2008) we derived estimates for the absolute number of transcript copies for each gene in each cell, and directly measured the average value of p_{smc}. "Pool/split" experiments (consisting of pooling RNA from multiple single cells, splitting the pool into the same number of separate reactions and building libraries from them) allowed us to measure the extent and impact of and control for technical variation. We found that the p_{smc} value is quite low (∼0.1). An analysis framework accounting for technical stochasticity was developed, which we used to assess variability in gene expression, allelic bias, and alternative splicing at the single cell level. Distinct from prior studies, our approach allowed us to parse findings into those that are just as likely to be of technical origins and those that are more likely to be of biological interest.

We found evidence of significant variability in the total number of mRNA molecules per cell, which underscores the importance of working with absolute copies-per-cell estimates when analyzing single cells as opposed to the widely used R/FPKM (**R**eads/**F**ragments **P**er **K**ilobase per

Million mapped reads/fragments) metric that only measures the relative abundance of genes in a library (FPKMs are still the better metric to use for larger cell pools). We identified biologically coherent modules of coexpressed genes specifically expressed in individual cells or groups of cells. These include expected variation associated with cell cycle phases, and an unexpected module enriched mRNA processing and splicing genes. We observed evidence of higher levels of autosomal allelic exclusion on the single-cell level, potentially associated with transcription bursts, however it is at present difficult to confidently distinguish from technical variability. In contrast, we found much stronger evidence for widespread major splice site usage switches between individual cells. Finally, our analysis of similarly constructed small cell pools (30 to 100 cells) revealed a high robustness and reproducibility, approaching that of bulk RNA measurements. This presents a reliable path forward towards the future comprehensive

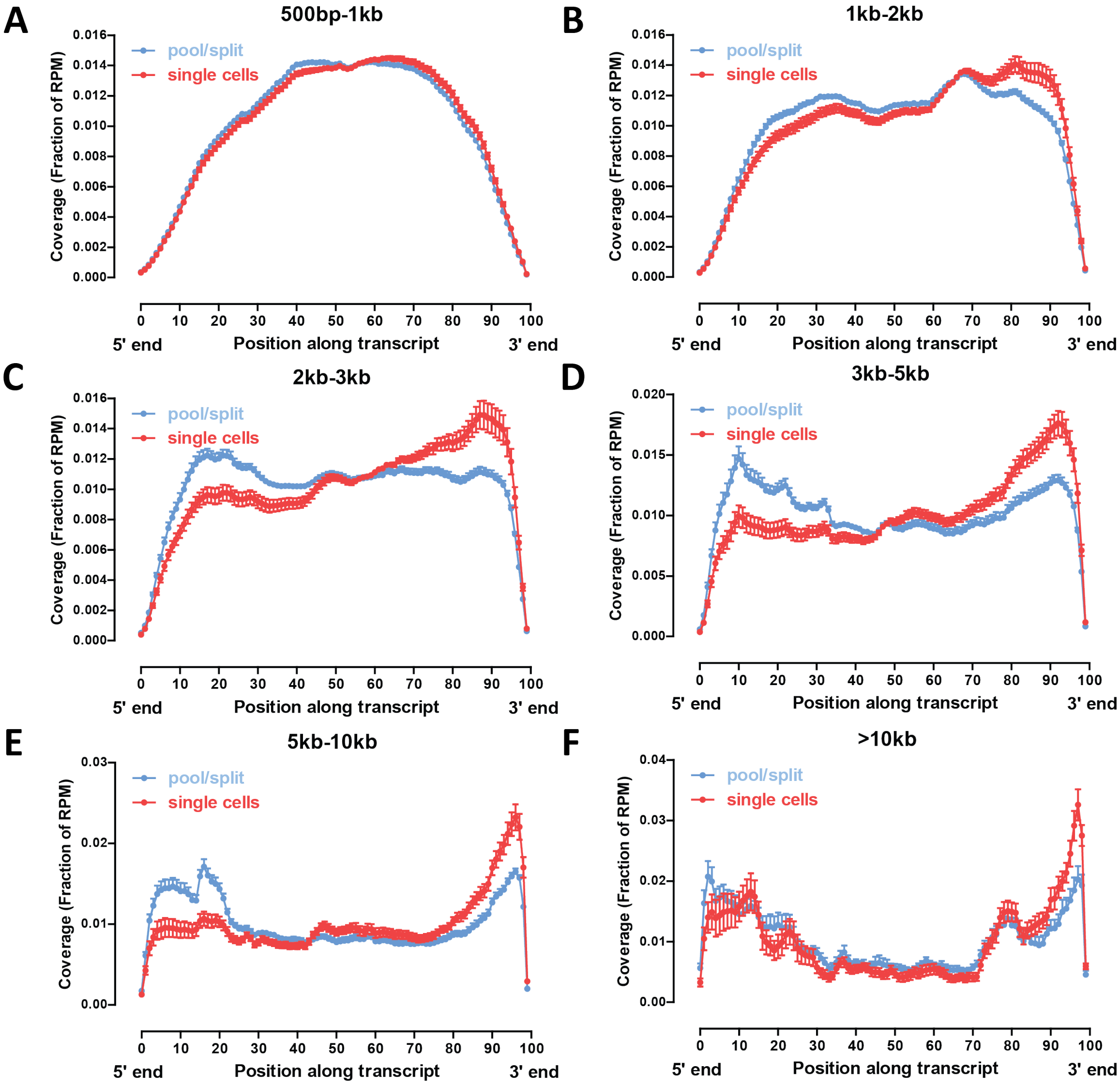


Figure 3.28: Uniformity of transcript coverage as a function of transcript length. Shown is the average coverage along the length of an mRNA for single cells and pool/split experiments. Only mRNAs with a single annotated isoform in the refSeq annotation set and within the indicated length limits were included.

transcriptomic characterization of rare cell types.

3.2 Results

3.2.1 *In silico* examination of major variables affecting informativeness of single-cell and small cell-pool RNA-seq

We began this study with two goals: first, to study gene expression heterogeneity in GM12878 cells on the single-cell level, and second, to determine the minimal optimal size of cell pools that is informative of the characteristics of the larger cell population, with the goal of applying that approach to rare cell types in future studies. How well these goals are achieved depends on several parameters affecting biological and technical stochasticity and detection sensitivity, the values of which were unknown. To understand their influence, we carried out a simulation of single-cell and cell-pool transcriptomes (see the Methods section for details) by varying the following parameters:

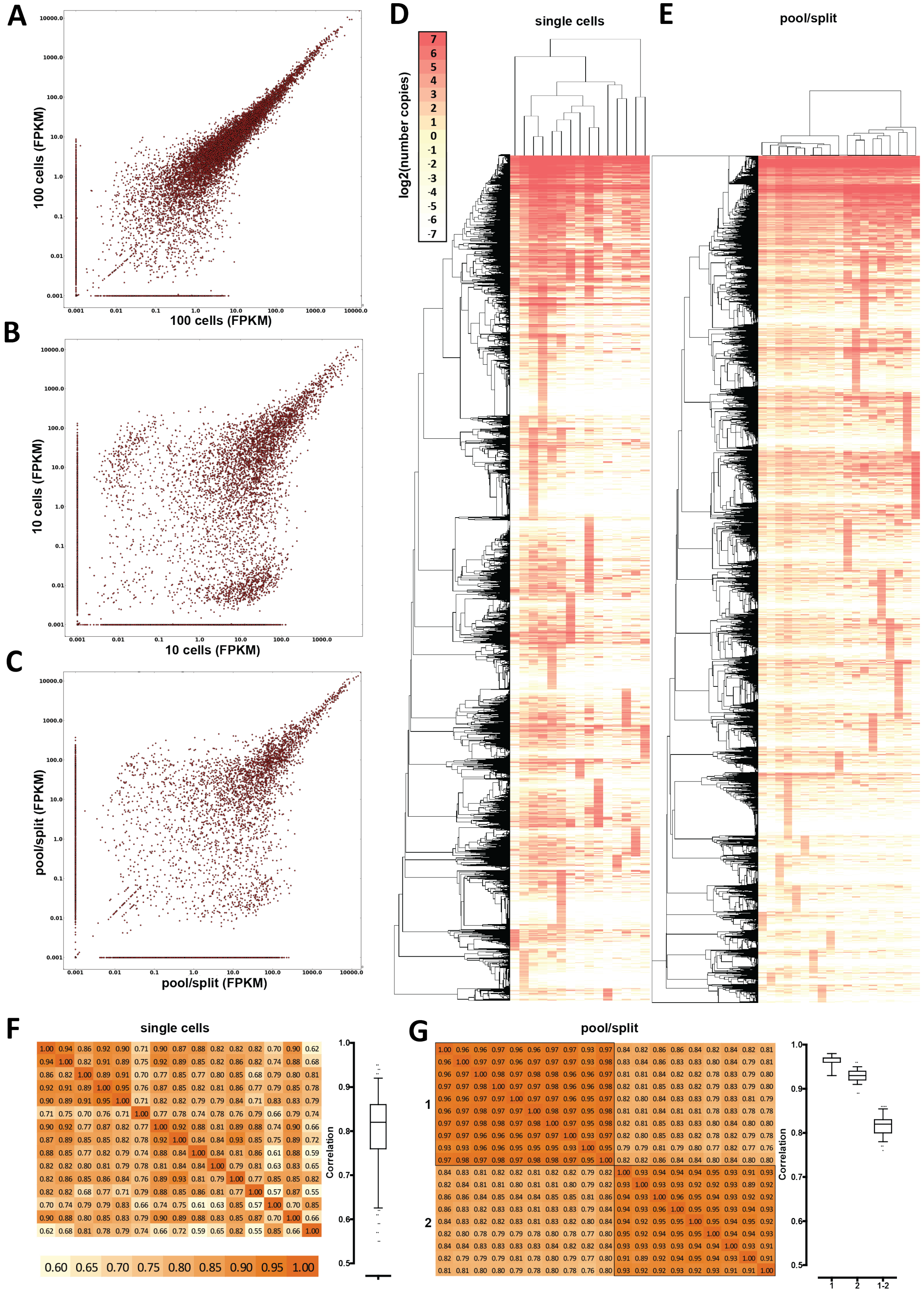
1. **Single-molecule capture efficiency** (p_{smc}). In contrast to bulk RNA-seq libraries, an individual cell contains a very limited total number of mRNA molecules. Individual genes can be present in single-digit transcript numbers. If only a fraction of mRNAs are successfully represented in a library, a technical stochasticity component is introduced. Depending on its magnitude, data interpretability can be significantly affected due to false negatives and a distortion of relative gene abundance estimates. The p_{smc} parameter is the probability that any given original RNA molecule is captured in the final library. We examined the effect on expression quantification of p_{smc} ranging from 0.01 to 1.

2. **Total number of mRNA molecules per cell**. The impact of low p_{smc} on expression measurements will be more severe if fewer mRNA molecules are present in a cell. The average total number of mRNA molecules in a single cell is not known for most cell types, but it is expected to vary with cell size, metabolic status, and even cell cycle phase. This means that single-cell expression measurements in some cell types are likely to be more robust to technical noise than in others. We varied the total number of mRNAs from 50,000 to 1,000,000 (while keeping the number of genes expressed constant).

3. **Frequency of expression of individual genes in single cells.** From prior studies we expect that some genes will be expressed in all or most cells, while others will be expressed in only a subset of cells. Genes detected at lower levels in bulk RNA-seq are the most obvious candidates to be expressed in a subset of cells in a population, although we do not know what fraction of low-abundance RNAs behave in such a way. This is particularly relevant to cell pools: a gene expressed at 50 copies per cell but only in 10% of cells would still be stochastically represented in a pool of 10 cells even if p_{smc} is high. In the absence of reliable data on this, we modeled the probability of expression in a given single cell with a distribution centered around very high values for genes highly expressed in bulk RNA-seq measurements, and progressively lower values with decreasing expression levels (details in the Methods section).

The simulation results are summarized in Figure 3.1 and Figures 3.2-3.25. As expected, low p_{smc} has a profoundly negative impact on gene expression quantification accuracy and reliability, leading to frequent false negatives Figure 3.1, and to poor estimates of expression levels. For example, in a single cell with 100,000 mRNAs, $p_{smc} = 0.1$ results in only 40% of genes expressed at 100 FPKM receiving FPKMs within 20% of the true value (Supplementary Figure 1C), but this fraction rises to nearly 100% if $p_{smc} = 0.8$ Figure 3.1G. The quantification of relative expression levels is similarly affected, with only the most highly expressed genes being consistently well quantified relative to each other at low p_{smc} (Figure 3.12-3.25.

In contrast, our simulation results indicate that cell pools are much more robust to technical noise, with 90% of genes expressed at 10 FPKM receiving FPKM estimates within 20% of their true value Figure 3.1C at $p_{smc} = 0.1$ in a pool of 100 cells. They also represent the expression profiles of the general population reasonably


A
10000.0
1000.0
100.0
10.0
1.0
0.1
0.01
0.001
100 cells (FPKM)
0.001
0.01
0.1
1.0
10.0
100.0
1000.0
10000.0
100 cells (FPKM)
B
10000.0
1000.0
100.0
10.0
1.0
0.1
0.01
0.001
10 cells (FPKM)
0.001
0.01
0.1
1.0
10.0
100.0
1000.0
10 cells (FPKM)
C
10000.0
1000.0
100.0
10.0
1.0
0.1
0.01
0.001
pool/split (FPKM)
0.001
0.01
0.1
1.0
10.0
100.0
1000.0
10000.0
pool/split (FPKM)
D
single cells
log2(number copies)
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
E
pool/split
F
single cells
1.00 0.94 0.86 0.92 0.90 0.71 0.90 0.87 0.88 0.82 0.82 0.82 0.70 0.90 0.62
0.94 1.00 0.82 0.91 0.89 0.75 0.92 0.89 0.85 0.82 0.86 0.82 0.74 0.88 0.68
0.86 0.82 1.00 0.89 0.91 0.70 0.77 0.85 0.77 0.80 0.85 0.68 0.79 0.80 0.81
0.92 0.91 0.89 1.00 0.95 0.76 0.87 0.85 0.82 0.81 0.86 0.77 0.79 0.85 0.78
0.90 0.89 0.91 0.95 1.00 0.71 0.82 0.82 0.79 0.79 0.84 0.71 0.83 0.83 0.79
0.71 0.75 0.70 0.76 0.71 1.00 0.77 0.78 0.74 0.78 0.76 0.79 0.66 0.79 0.74
0.90 0.92 0.77 0.87 0.82 0.77 1.00 0.92 0.88 0.81 0.89 0.88 0.74 0.90 0.66
0.87 0.89 0.85 0.85 0.82 0.78 0.92 1.00 0.84 0.84 0.93 0.85 0.75 0.89 0.72
0.88 0.85 0.77 0.82 0.79 0.74 0.88 0.84 1.00 0.84 0.81 0.86 0.61 0.88 0.59
0.82 0.82 0.80 0.81 0.79 0.78 0.81 0.84 0.84 1.00 0.79 0.81 0.63 0.83 0.65
0.82 0.86 0.85 0.86 0.84 0.76 0.89 0.93 0.81 0.79 1.00 0.77 0.85 0.85 0.82
0.82 0.82 0.68 0.77 0.71 0.79 0.88 0.85 0.86 0.81 0.77 1.00 0.57 0.87 0.55
0.70 0.74 0.79 0.79 0.83 0.66 0.74 0.75 0.61 0.63 0.85 0.57 1.00 0.70 0.85
0.90 0.88 0.80 0.85 0.83 0.79 0.90 0.89 0.88 0.83 0.85 0.87 0.70 1.00 0.66
0.62 0.68 0.81 0.78 0.79 0.74 0.66 0.72 0.59 0.65 0.82 0.55 0.85 0.66 1.00
0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
1.0
0.9
0.8
0.7
0.6
0.5
Correlation
G
pool/split
1
2
1.00 0.96 0.96 0.97 0.96 0.96 0.97 0.97 0.93 0.97 0.84 0.82 0.86 0.86 0.84 0.82 0.84 0.82 0.81
0.96 1.00 0.97 0.97 0.96 0.97 0.97 0.97 0.93 0.98 0.83 0.84 0.86 0.83 0.83 0.80 0.84 0.79 0.81
0.96 0.97 1.00 0.98 0.97 0.98 0.98 0.96 0.96 0.97 0.81 0.81 0.84 0.82 0.82 0.78 0.83 0.79 0.80
0.97 0.97 0.98 1.00 0.97 0.97 0.98 0.96 0.96 0.98 0.82 0.81 0.85 0.83 0.82 0.79 0.83 0.79 0.80
0.96 0.96 0.97 0.97 1.00 0.97 0.97 0.96 0.95 0.97 0.82 0.80 0.84 0.83 0.83 0.79 0.83 0.81 0.79
0.96 0.97 0.98 0.97 0.97 1.00 0.98 0.97 0.95 0.98 0.81 0.81 0.84 0.81 0.81 0.78 0.83 0.78 0.78
0.97 0.97 0.98 0.98 0.97 0.98 1.00 0.97 0.95 0.98 0.82 0.81 0.85 0.83 0.83 0.80 0.84 0.80 0.80
0.97 0.97 0.96 0.96 0.96 0.97 0.97 1.00 0.93 0.97 0.82 0.80 0.85 0.83 0.82 0.78 0.82 0.79 0.78
0.93 0.93 0.96 0.96 0.95 0.95 0.95 0.93 1.00 0.95 0.79 0.79 0.81 0.79 0.80 0.77 0.82 0.77 0.76
0.97 0.98 0.97 0.98 0.97 0.98 0.98 0.97 0.95 1.00 0.82 0.82 0.86 0.84 0.84 0.80 0.84 0.80 0.80
0.84 0.83 0.81 0.82 0.82 0.81 0.82 0.82 0.79 0.82 1.00 0.93 0.94 0.94 0.94 0.95 0.93 0.91 0.93
0.82 0.84 0.81 0.81 0.80 0.81 0.81 0.80 0.79 0.82 0.93 1.00 0.93 0.93 0.92 0.92 0.93 0.89 0.92
0.86 0.86 0.84 0.85 0.84 0.84 0.85 0.85 0.81 0.86 0.94 0.93 1.00 0.96 0.95 0.94 0.93 0.92 0.92
0.86 0.83 0.82 0.83 0.83 0.81 0.83 0.83 0.79 0.84 0.94 0.93 0.96 1.00 0.95 0.95 0.93 0.94 0.93
0.84 0.83 0.82 0.82 0.83 0.81 0.83 0.82 0.80 0.84 0.94 0.92 0.95 0.95 1.00 0.95 0.94 0.95 0.92
0.82 0.80 0.78 0.79 0.79 0.78 0.80 0.78 0.77 0.80 0.95 0.92 0.94 0.95 0.95 1.00 0.94 0.94 0.93
0.84 0.84 0.83 0.83 0.83 0.83 0.84 0.82 0.82 0.84 0.93 0.93 0.93 0.93 0.94 0.94 1.00 0.93 0.91
0.82 0.79 0.79 0.79 0.81 0.78 0.80 0.79 0.77 0.80 0.91 0.89 0.92 0.94 0.95 0.94 0.93 1.00 0.91
0.81 0.81 0.80 0.80 0.79 0.78 0.80 0.78 0.76 0.80 0.93 0.92 0.92 0.93 0.92 0.93 0.91 0.91 1.00
1.0
0.9
0.8
0.7
0.6
0.5
Correlation
1
2
1-2

well Figure 3.1, even at low p_{smc}, starting from a size of ∼30 cells (10-cell pools seem not to be sufficient to achieve this). Finally, as expected, the larger the number of total mRNA molecules per cell, the greater is the buffer against technical noise, resulting in more robust quantification Figures 3.2-3.11.

3.2.2 Transcriptome measurements of individual single cells and companion pool/splits

The simulation results informed our experimental design, which aimed to gain a firm grasp on technical stochasticity in two ways Figure 3.26A. First, we generated single-cell RNA-seq libraries and in parallel carried out "pool/split" experiments. In a pool/split, multiple cells are pooled and lysed together, then split into the same number of reactions, from which libraries are built. Variation between these libraries should be purely technical (with stochastic splitting possibly playing a role at the low end). Variation observed at similar levels in both single cells and pool/splits cannot be confidently considered real, even if this leads to some true biological variation being obscured. However, variation above the pool-split level can be identified and ascribed to biological sources.

We generated single-cell RNA-seq libraries from 15 single GM12878 cells and from two pairs of 10-cell pool/split experiments. We also sequenced replicates of pools of multiple cells (10, 30 and 100 cells) as well as 100pg and 10ng samples of bulk RNA (corresponding to ∼10 and ∼1000 cells), to assess the stability of measurements as a function of the amount of starting material.

We used the SMART-seq protocol (Ramsköld & Luo et al. 2012) (Figure 3.27) to generate our libraries. A detailed description of the protocol, as we implemented it, is presented in the Methods section. We obtained nearly uniform full-length transcript coverage (Figure 3.26B, Figure 3.28). Uniformity of coverage, which depends on the intactness of RNAs and the successful copying of full-length molecules, is highly desirable for several reasons. First, RNA-seq data quantification using the RPKM/FPKM metric (Mortazavi & Williams et al. 2008; Trapnell et al. 2010), makes an implicit assumption of full coverage. Second, it enables the analysis of alternative splicing and allelic bias as read coverage of 5'-proximal splice sites and heterozygous positions is ensured.

We added spike-in quantification standards of known abundance (in absolute number of RNA copies, Table 3) at the very beginning of cDNA synthesis. This allows us to, first, estimate p_{smc}, and second, derive gene expression estimates in absolute numbers of copies per cell. The latter is important because while FPKM is useful for comparing expression levels within a library, it can only be used to compare directly across different libraries when the total amount of RNA in each starting sample is roughly the same (Anders & Huber 2010). This assumption is usually only mildly violated when working with bulk samples, but when single cells are compared, it becomes significantly more problematic as the variation in the total amount of RNA in each cell is expected to be much larger.

The extent of variation in the total amount of RNA between single cells is not known a priori, but it will often be larger than that between the averages of large populations of cells. For this reason, the ideal single-cell RNA-seq protocol would directly measure the absolute num-

Figure 3.29 *(preceding page)*: Technical and biological variation in single-cell RNA-seq measurements of gene expression. (A) Correlation between expression levels (in FPKM) between two pools of 100 cells. (B) Correlation between expression levels (in FPKM) between two pools of 10 cells. (C) Correlation between expression levels (in FPKM) between two representative pool/split libraries. A pseudocount of 0.001 was added to each data point in the scatter plots for visualization purposes. (D and E) Hierarchical clustering of estimated copies-per-cell values for protein coding genes in single-cell (D) and pool/split (E) libraries. Pearson correlation was used as a distance metric and only genes expressed at a level of at least one estimated copy in at least one library were included. (F and G) Correlation between estimated copies-per-cell values for protein coding genes in single-cell libraries (F) and pool/split libraries (G). Two sets of pool/split experiments (1 and 2) are shown and "1-2" in the box-plot refers to correlations between the two sets while "1" and "2" refer to correlation within each experiment. Similar plots but using Spearman correlation are shown in Supplementary Figure 32.

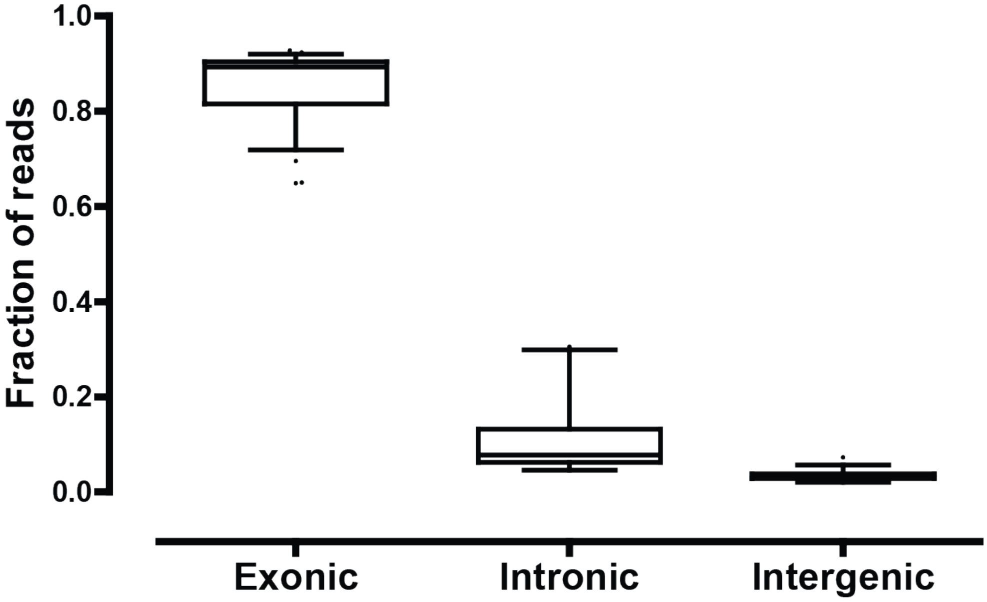


Figure 3.30: Efficiency of enrichment for polydenylated messages. Shown is the fraction of reads mapping to exons, introns or integenic space (GENCODE V13 annotation).

ber of transcripts per cell. This is not possible with the protocols existing at the time of this work, including SMART-seq. Each original cDNA molecule is amplified to a large number of copies which are then subjected to tagmentation and a second round of PCR; this erases any relation between original molecules and the fragments in the final sequencing library as each founder molecule results in multiple overlapping smaller fragments in the final library.

Figures 3.26 and 3.29 summarize the technical characterization of the SMART-seq protocol applied to GM12878 cells. In addition to the mostly complete coverage along transcript length, sequencing libraries were also highly enriched for exonic sequences (Figure 3.30), indicating a high efficiency of enrichment for polyadenylated molecules.

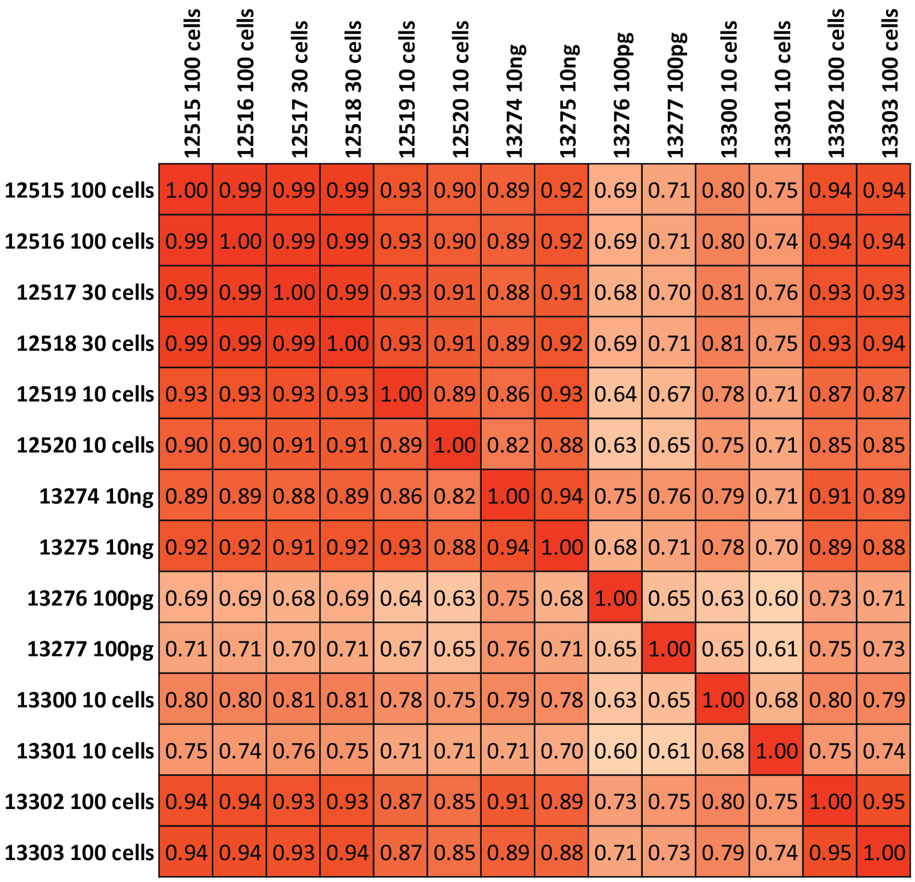
3.2.3 Gene detection in single cells versus pools of varied sizes

We compared single cell and pool/split libraries, as well as cell pools, with bulk RNA samples from GM12878 cells (Figure 3.26C). In bulk RNA libraries, we detect about 12,000 genes expressed at more than 0.1FPKM. A similar number of genes, between 4,000 and 5,000, is detected in both single cell and pool/split libraries. These differences between single cells and bulk libraries are due mostly to genes expressed at low levels. Genes expressed at more than 100 FPKM in 10ng bulk RNA samples are detected in almost all libraries, while only ∼30% of genes expressed at ∼10 FPKM and 10% of genes expressed at ∼1FPKM were detected in any given single cell (Figure 3.26D). Notably, the number of genes detected in both 100-cell and 30-cell pools was similar to that detected in the 10ng libraries (∼11,000). In contrast, in the 10-cell pools and 100pg libraries lower numbers of genes were detected, between 6,000 and 7,000. This is consistent with simulation results suggesting that 30 cells is the lower limit of cell number at which the transcriptome library complexity begins to approach that of the larger cell population. This is corroborated by the correlation between the expression levels of replicate measurements (Figure 3.29A, Figure 3.31). In contrast, a sizable population of genes present at high levels in one replicate and at very low levels or completely absent in the other appears in 10-cell pools (Figure 3.29B) and especially, in pool/split libraries (Figure 3.29C). Finally, union sets of genes detected in all individual cell libraries and in all pool-split libraries was ∼10,000, which was in the range seen for 30-100-cell pools.

3.2.4 Pool/splits measure technical variation and reveal biological variation among single cells

The observed variations in gene expression levels and detection can be explained as a combination of some genes not being expressed in each and every cell and low p_{smc} resulting in large numbers of false negatives. We calculated the average p_{smc} across all libraries based on the detection of spike-ins (details in Methods). This number is in our estimates ~0.1. We also estimated that for GM12878 single cells one transcript copy corresponds to on average to ~10 FPKM 3.32. This agrees well with the observation that detection of genes becomes unstable below ~100FPKM (Figure 3.29B and 3.29C), which in turn is consistent with previous observations (Ramsköld & Luo et al. 2012).

We compared expression measurements in single-cell and pool/split libraries. Hierarchical clustering of each group is shown in Figures 3.29D and 3.29E (with two independent biological replicate pool/spit experiments shown in Figure 3.29E). The distances between the expression profiles within the same pool/split experiment were significantly smaller than those for individual single cells (branch lengths in Figures 3.29D and 3.29E) and average correlations between single cells were accordingly lower than those between libraries from the same pool/split (Figures 3.29F and 3.29G). A notable feature of the data are small clusters of genes present at high levels in only one library. These are more prominent in single cells than in pool/splits, yet they are clearly present in all samples. In single cells, this is due to a mixture of stochastic capture ef-



	12515 100 cells	12516 100 cells	12517 30 cells	12518 30 cells	12519 10 cells	12520 10 cells	13274 10ng	13275 10ng	13276 100pg	13277 100pg	13300 10 cells	13301 10 cells	13302 100 cells	13303 100 cells
12515 100 cells	1.00	0.99	0.99	0.99	0.93	0.90	0.89	0.92	0.69	0.71	0.80	0.75	0.94	0.94
12516 100 cells	0.99	1.00	0.99	0.99	0.93	0.90	0.89	0.92	0.69	0.71	0.80	0.74	0.94	0.94
12517 30 cells	0.99	0.99	1.00	0.99	0.93	0.91	0.88	0.91	0.68	0.70	0.81	0.76	0.93	0.93
12518 30 cells	0.99	0.99	0.99	1.00	0.93	0.91	0.89	0.92	0.69	0.71	0.81	0.75	0.93	0.94
12519 10 cells	0.93	0.93	0.93	0.93	1.00	0.89	0.86	0.93	0.64	0.67	0.78	0.71	0.87	0.87
12520 10 cells	0.90	0.90	0.91	0.91	0.89	1.00	0.82	0.88	0.63	0.65	0.75	0.71	0.85	0.85
13274 10ng	0.89	0.89	0.88	0.89	0.86	0.82	1.00	0.94	0.75	0.76	0.79	0.71	0.91	0.89
13275 10ng	0.92	0.92	0.91	0.92	0.93	0.88	0.94	1.00	0.68	0.71	0.78	0.70	0.89	0.88
13276 100pg	0.69	0.69	0.68	0.69	0.64	0.63	0.75	0.68	1.00	0.65	0.63	0.60	0.73	0.71
13277 100pg	0.71	0.71	0.70	0.71	0.67	0.65	0.76	0.71	0.65	1.00	0.65	0.61	0.75	0.73
13300 10 cells	0.80	0.80	0.81	0.81	0.78	0.75	0.79	0.78	0.63	0.65	1.00	0.68	0.80	0.79
13301 10 cells	0.75	0.74	0.76	0.75	0.71	0.71	0.71	0.70	0.60	0.61	0.68	1.00	0.75	0.74
13302 100 cells	0.94	0.94	0.93	0.93	0.87	0.85	0.91	0.89	0.73	0.75	0.80	0.75	1.00	0.95
13303 100 cells	0.94	0.94	0.93	0.94	0.87	0.85	0.89	0.88	0.71	0.73	0.79	0.74	0.95	1.00

Figure 3.31: Correlation between expression estimates based on different cell pools sizes and different amounts of input bulk RNA. Correlation coefficients were calculated on the $log_2(FPKM+1)$ transform of the FPKM estimates for the refSeq annotation, with only protein coding genes present at ≥ 1 FPKM in at least one library included.

fects and real biological variation. In pool/splits, stochastic capture is the predominant source. It is important to note that given the low p_{smc}, it is difficult to determine the cause of variation for any given gene. Nevertheless, the major conclusion at the transcriptome level is that there are biological differences between single cells because the technical stochasticity in pool/splits is significantly less than variation across single cells.

3.2.5 Estimating absolute transcript levels in single cells

Absolute transcript counts are the biologically relevant values ideally obtained from a single-cell gene expression profiling experiment because, as discussed above, FPKM is a poor metric for comparing gene expression levels in individual cells if the total amount of RNA varies a lot. We derived transcript number estimates for each gene based on the FPKM values of spike-ins. We observed good agreement between the input number of spike-in RNA copies and the corresponding FPKM values in the final libraries (Figures 3.33 and 3.34).

We used the transcripts-per-cell estimates for all subsequent analyses. Previous studies have reported that genes can be separated into two distinct groups based on their expression levels - one group expressed at high (> 1FPKM) levels and one at very low (<< 1 FPKM) (Hebenstreit et al. 2011). We examined the distribution of estimated copies per cell in single cells in pool/splits (Figure 3.35A). We found that in individual cells, most protein coding genes are expressed at levels between 1 and ~50 copies per cell. The distribution suggests a roughly equal number of genes at each level except for a larger group of transcripts with

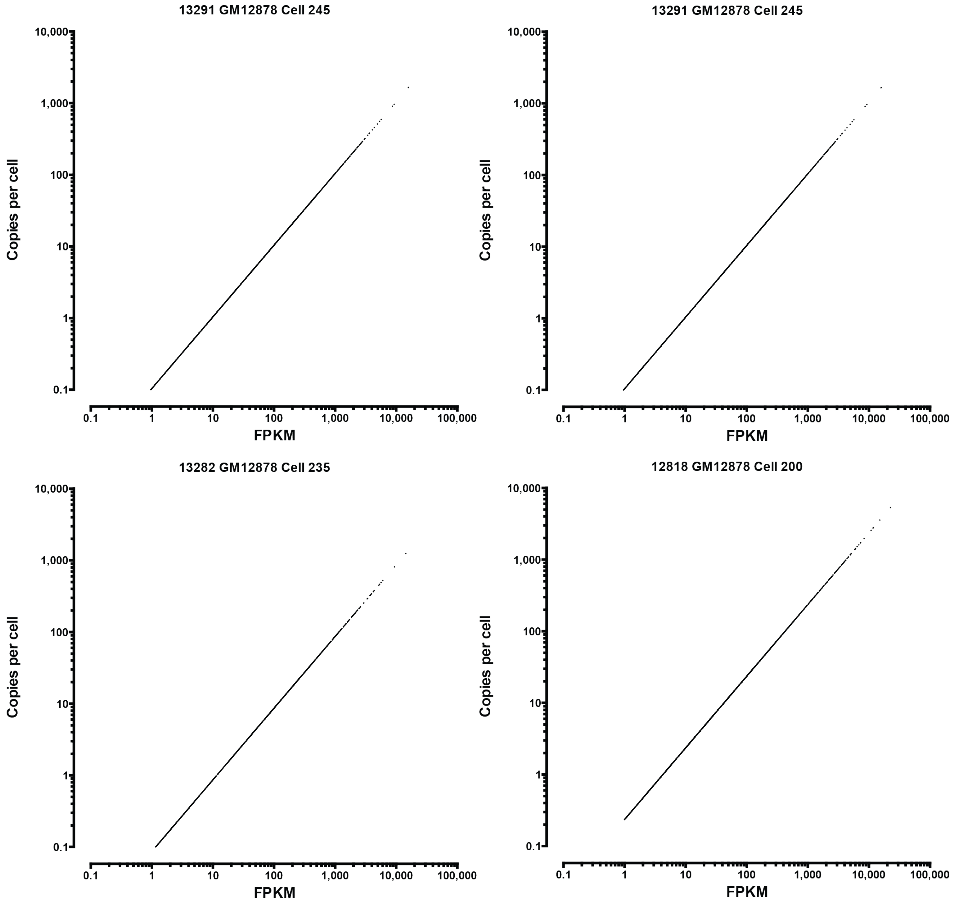


Figure 3.32: Relation between FPKMs and copies-per-cell estimates in representative single-cell libraries.

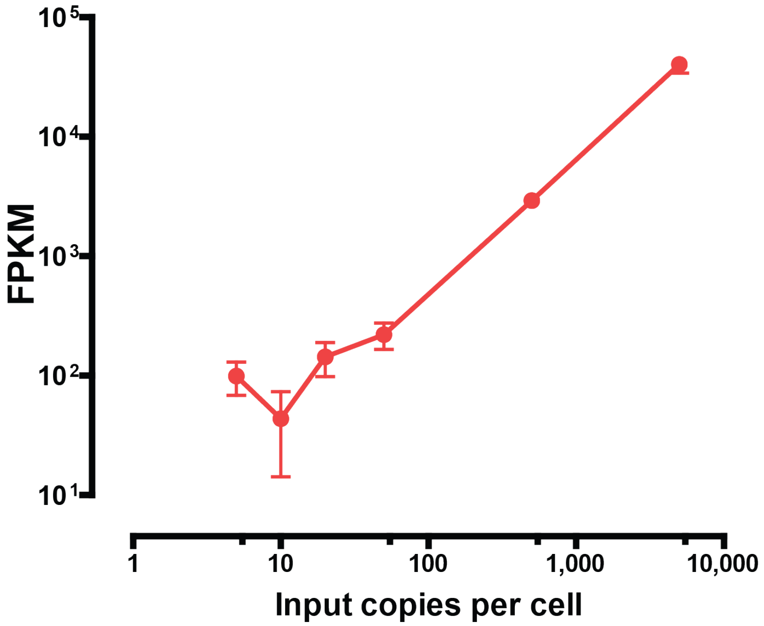


Figure 3.33: Correspondence between initial spike-in amounts and spike abundance in sequenced libraries as measured in FPKMs. Error bars represent the standard error of the mean.

fractional transcript-per-cell values. Obviously, single-cell determinations are constrained in a way that population level measurements cannot be: one transcript per cell is the minimum

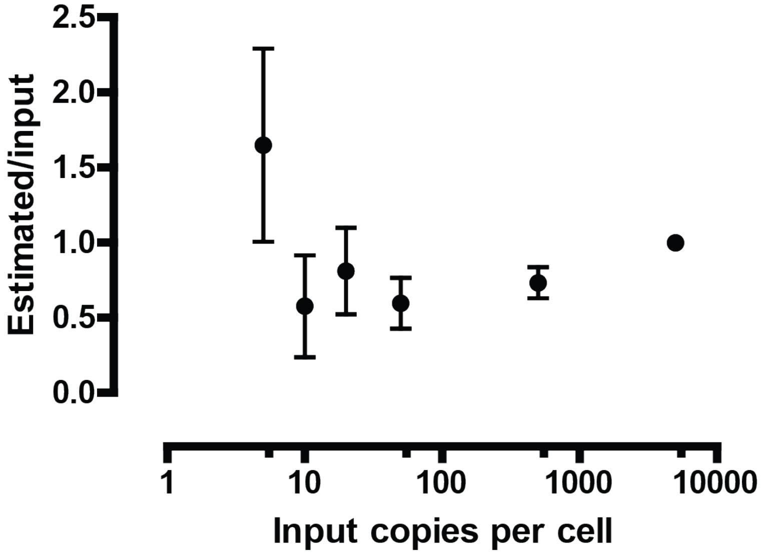
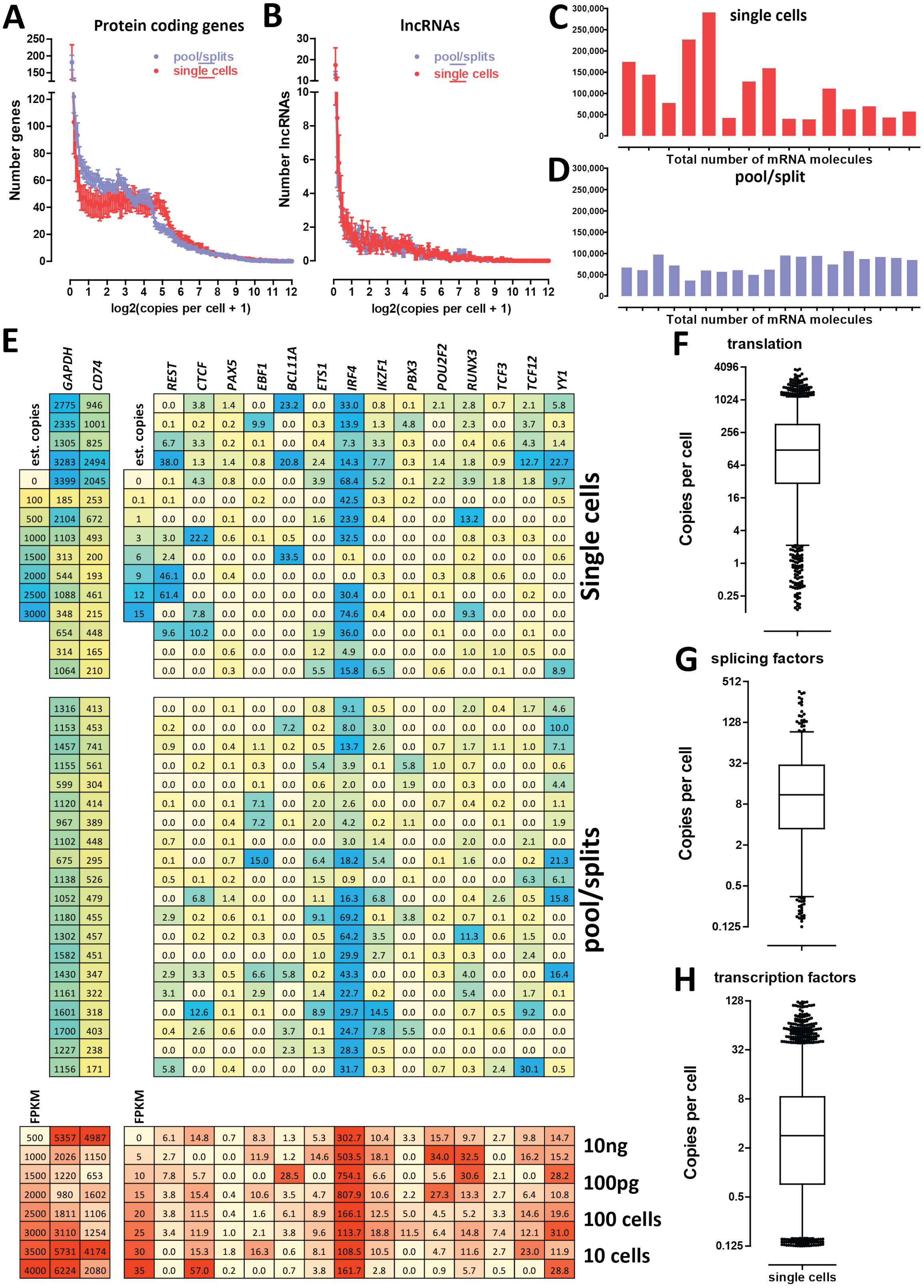


Figure 3.34: Stability of copies per cell estimation. Spike-in sequences of known abundance (Supplementary Table 2) were added to each reaction prior to library building. A linear regression calibration was derived based on RPKM/FPKM values calculated for each. Shown is the average ratio of estimated copies per cell and the actual spiked in copies per cell for these spike sequences. Error bars represent the standard error of the mean.



est. copies
0
100
500
1000
1500
2000
2500
3000

est. copies
0
0.1
1
3
6
9
12
15

Single cells

GAPDH	*CD74*	*REST*	*CTCF*	*PAX5*	*EBF1*	*BCL11A*	*ETS1*	*IRF4*	*IKZF1*	*PBX3*	*POU2F2*	*RUNX3*	*TCF3*	*TCF12*	*YY1*
2775	946	0.0	3.8	1.4	0.0	23.2	0.0	33.0	0.8	0.1	2.1	2.8	0.7	2.1	5.8
2335	1001	0.1	0.2	0.2	9.9	0.0	0.3	13.9	1.3	4.8	0.0	2.3	0.0	3.7	0.3
1305	825	6.7	3.3	0.2	0.1	0.0	0.4	7.3	3.3	0.3	0.0	0.4	0.6	4.3	1.4
3283	2494	38.0	1.3	1.4	0.8	20.8	2.4	14.3	7.7	0.3	1.4	1.8	0.9	12.7	22.7
3399	2045	0.1	4.3	0.8	0.0	0.0	3.9	68.4	5.2	0.1	2.2	3.9	1.8	1.8	9.7
185	253	0.1	0.0	0.0	0.2	0.0	0.0	42.5	0.3	0.2	0.0	0.0	0.0	0.0	0.5
2104	672	0.0	0.0	0.1	0.0	0.0	1.6	23.9	0.4	0.0	0.0	13.2	0.0	0.0	0.0
1103	493	3.0	22.2	0.6	0.1	0.5	0.0	32.5	0.0	0.0	0.0	0.8	0.3	0.3	0.0
313	200	2.4	0.0	0.0	0.0	33.5	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.6
544	193	46.1	0.0	0.4	0.0	0.0	0.0	0.0	0.3	0.0	0.3	0.8	0.6	0.0	0.0
1088	461	61.4	0.0	0.0	0.0	0.0	0.0	30.4	0.0	0.1	0.1	0.0	0.0	0.2	0.0
348	215	0.0	7.8	0.0	0.0	0.0	0.0	74.6	0.4	0.0	0.0	9.3	0.0	0.0	0.0
654	448	9.6	10.2	0.0	0.0	0.0	1.9	36.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0
314	165	0.0	0.0	0.6	0.0	0.0	1.2	4.9	0.0	0.0	0.0	1.0	1.0	0.5	0.0
1064	210	0.0	0.0	0.3	0.0	0.0	5.5	15.8	6.5	0.0	0.6	0.0	0.1	0.0	8.9

pool/splits

GAPDH	*CD74*	*REST*	*CTCF*	*PAX5*	*EBF1*	*BCL11A*	*ETS1*	*IRF4*	*IKZF1*	*PBX3*	*POU2F2*	*RUNX3*	*TCF3*	*TCF12*	*YY1*
1316	413	0.0	0.0	0.1	0.0	0.0	0.8	9.1	0.5	0.0	0.0	2.0	0.4	1.7	4.6
1153	453	0.2	0.0	0.0	0.0	7.2	0.2	8.0	3.0	0.0	0.0	0.0	0.0	0.0	10.0
1457	741	0.9	0.0	0.4	1.1	0.2	0.5	13.7	2.6	0.0	0.7	1.7	1.1	1.0	7.1
1155	561	0.0	0.2	0.1	0.3	0.0	5.4	3.9	0.1	5.8	1.0	0.7	0.0	0.0	0.6
599	304	0.0	0.0	0.0	0.1	0.0	0.6	2.0	0.0	1.9	0.0	0.3	0.0	0.0	4.4
1120	414	0.1	0.0	0.1	7.1	0.0	2.0	2.6	0.0	0.0	0.7	0.4	0.2	0.0	1.1
967	389	0.0	0.0	0.4	7.2	0.1	2.0	4.2	0.2	1.1	0.0	0.0	0.1	0.0	1.9
1102	448	0.7	0.0	0.1	0.0	0.0	0.0	3.0	1.4	0.0	0.0	2.0	0.0	2.1	0.0
675	295	0.1	0.0	0.7	15.0	0.0	6.4	18.2	5.4	0.0	0.1	1.6	0.0	0.2	21.3
1138	526	0.5	0.1	0.2	0.0	0.0	1.5	0.9	0.0	0.1	0.0	0.0	0.0	6.3	6.1
1052	479	0.0	6.8	1.4	0.0	0.0	1.1	16.3	6.8	0.0	0.0	0.4	2.6	0.5	15.8
1180	455	2.9	0.2	0.6	0.1	0.0	9.1	69.2	0.1	3.8	0.2	0.7	0.1	0.2	0.0
1302	457	0.0	0.2	0.2	0.3	0.0	0.5	64.2	3.5	0.0	0.0	11.3	0.6	1.5	0.0
1582	451	0.0	0.0	0.0	0.0	0.0	1.0	29.9	2.7	0.1	0.3	0.3	0.0	2.4	0.0
1430	347	2.9	3.3	0.5	6.6	5.8	0.2	43.3	0.0	0.0	0.3	4.0	0.0	0.0	16.4
1161	322	3.1	0.0	0.1	2.9	0.0	1.4	22.7	0.2	0.0	0.0	5.4	0.0	1.7	0.1
1601	318	0.0	12.6	0.1	0.1	0.0	8.9	29.7	14.5	0.0	0.0	0.7	0.5	9.2	0.0
1700	403	0.4	2.6	0.6	0.0	3.7	0.1	24.7	7.8	5.5	0.0	0.1	0.6	0.0	0.0
1227	238	0.0	0.0	0.0	0.0	2.3	1.3	28.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0
1156	171	5.8	0.0	0.4	0.0	0.0	0.0	31.7	0.3	0.0	0.7	0.3	2.4	30.1	0.5

FPKM			FPKM														
500	5357	4987	0	6.1	14.8	0.7	8.3	1.3	5.3	302.7	10.4	3.3	15.7	9.7	2.7	9.8	14.7
1000	2026	1150	5	2.7	0.0	0.0	11.9	1.2	14.6	503.5	18.1	0.0	34.0	32.5	0.0	16.2	15.2
1500	1220	653	10	7.8	5.7	0.0	0.0	28.5	0.0	754.1	6.6	0.0	5.6	30.6	2.1	0.0	28.2
2000	980	1602	15	3.8	15.4	0.4	10.6	3.5	4.7	807.9	10.6	2.2	27.3	13.3	2.7	6.4	10.8
2500	1811	1106	20	3.8	11.5	0.4	1.6	6.1	8.9	166.1	12.5	5.0	4.5	5.2	3.3	14.6	19.6
3000	3110	1254	25	3.4	11.9	1.0	2.1	3.8	9.6	113.7	18.8	11.5	6.4	14.8	7.4	12.1	31.0
3500	5731	4174	30	0.0	15.3	1.8	16.3	0.6	8.1	108.5	10.5	0.0	4.7	11.6	2.7	23.0	11.9
4000	6224	2080	35	0.0	57.0	0.2	0.0	0.7	3.8	161.7	2.8	0.0	0.9	5.7	0.5	0.0	28.8

Figure 3.35 *(preceding page)*: Absolute expression levels at the single cell level. FPKM values converted to estimated copies per cell using the spike-in quantification standards are shown. (A) Distribution of expression levels of refSeq protein coding genes in estimated copies per cell in single-cells and pool/split experiments. (B) Distribution of expression levels of GENCODE V13 lncRNA protein coding genes in estimated copies per cell in single-cells (red) and pool/split experiments (blue). (C) Total number of mRNA copies per cell in single cells. (D) Total number of mRNA copies in pool/split experiments; (E) Expression levels of house-keeping and highly expressed genes (*GAPDH*, *CD74*, left panel) and general (*CTCF*, *REST*, *YY1*) and B-cell regulatory (*PAX5*, *EBF1*, *BCL11A*, *ETS1*, *IRF4*, *IKZF1*, *PBX3*, *POU2F2*, *RUNX3*, *TCF3*, *TCF12*) transcription factors (right panel). Upper and middle panels show the estimated copies-per-cell numbers for single-cells and pool/splits respectively. The lower panel shows FPKM values for cell pools and bulk RNA libraries. (F, G and H) Distribution of absolute expression levels in copies per cell in single cells for translation initiation, elongation and termination proteins (F), splicing regulators (G) and transcription factors (H). The list of translation proteins was retrieved from the corresponding GO category annotations downloaded from FuncAssociate 2.0 (Berriz et al. 2009). The list of splicing regulators was obtained from the SpliceAid-F database of human splicing factors (Giulietti et al. 2013). The list of transcription factors used was the one from Vaquerizas et al. 2009. Note that only values ≥0.1 estimated copies per cell were included in these plots, i.e. libraries in which the genes was not detected were excluded.

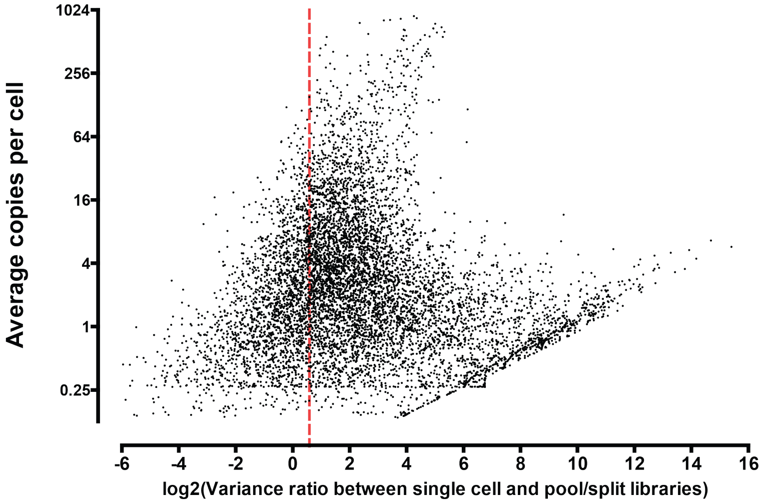


Figure 3.36: Ratio of the variance of single cell and pool/split libraries vs. average estimated number of mRNA molecules. The vertical line corresponds to a variance ratio of 1.5. Genes with a variance ratio higher than 1.5 were retained for network construction. Most genes with a lower ratio (and correspondingly high variance in pool/split libraries) have a relatively low average estimated number of mRNA molecules per cell.

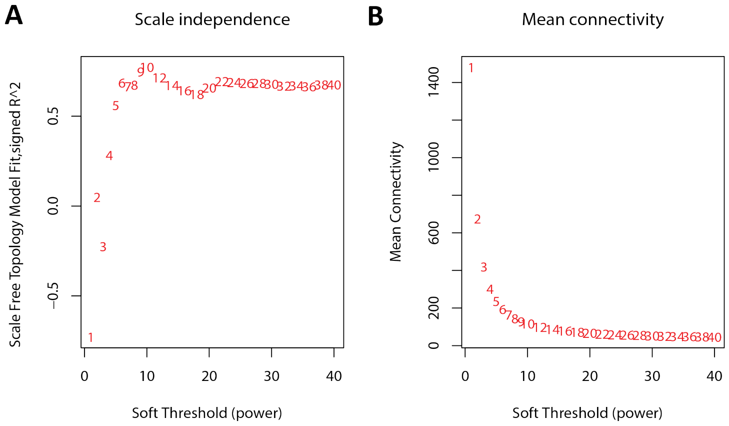


Figure 3.37: Optimization of the soft threshold parameter for constructing weighted correlation gene expression network. (A) Scale independence (B) Mean connectivity. A value of $\beta = 6$ was used for network construction.

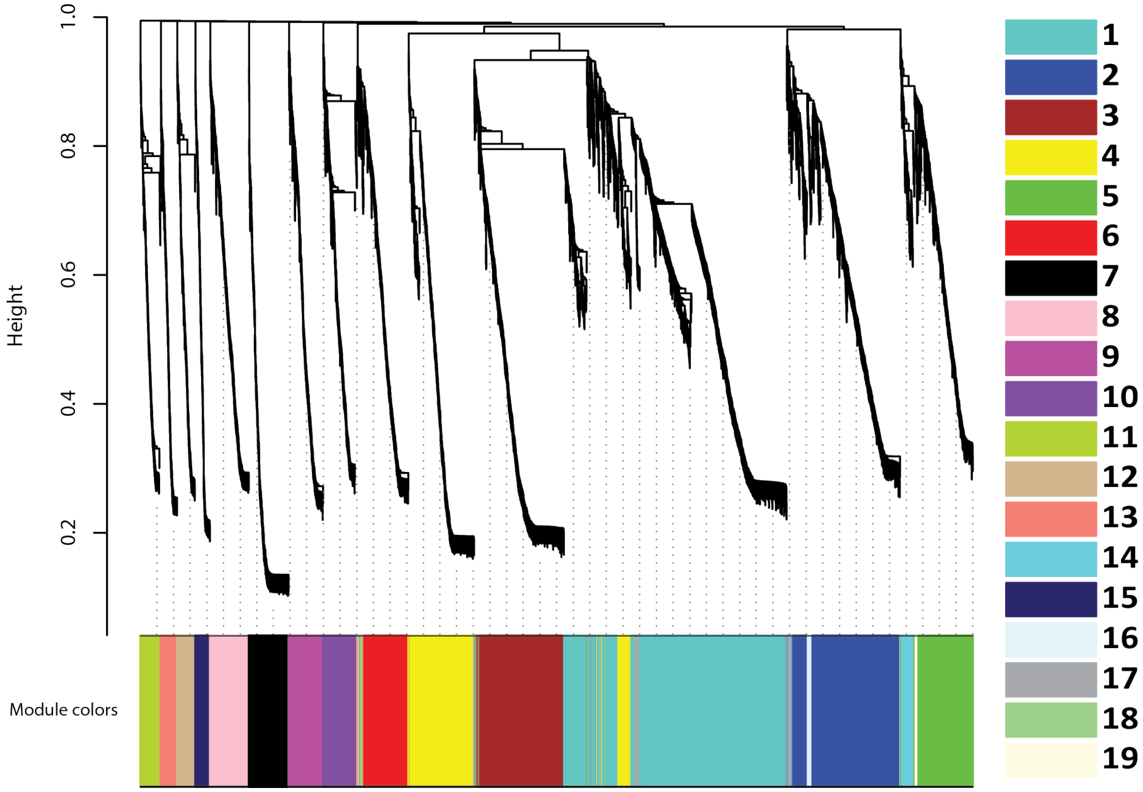


Figure 3.38: Cluster dendrogram of gene coexpression modules derived from single GM12878 cells..

non-zero value possible. The lower values likely represent a combination of mapping artifacts (due to high sequence homology of paralogs) and RNAs that were both present at low levels and poorly represented (due, for example, to the fragmentation of a single original RNA molecule resulting in artificially low FPKMs as a result to coverage only at the 3' end). The distribution of estimated copies in pool/split libraries exhibited a more linear decrease in the number of more highly expressed genes, consistent with averaging of variation between cells.

We also examined the distribution of the expression levels of long non-coding RNAs (lncRNAs, Guttman et al. 2009). Consistent with previous observations (Ramsköld et al. 2009; Guttman et al. 2010; Djebali & Davis et al. 2012), lncRNAs have generally much lower expression levels compared to protein coding genes (Figure 3.35B). We note that accurate quantification of the absolute number of copies is of great relevance to understanding lncRNA biology as both *cis* and *trans* models for the function of lncRNAs have been proposed (Koziol & Rinn, 2010; Rinn & Chang, 2012) and lncRNAs functioning in *cis* are expected to be expressed at lower levels (possibly only one or two copies per cell) compared to lncRNAs acting in *trans.* At present, copies-per-cell estimates are not sufficiently reliable for this issue to be conclusively resolved (in addition, the SMART-seq protocol is specific for polyA+ RNAs while it cannot be assumed that lncRNAs, especially the *cis*-acting ones, are polyadenylated); nevertheless, we expect future improvements in single-cell RNA-seq methodology to be highly informative in understanding lncRNA biology.

We were also able to directly assess the total number of mRNAs present in each cell (Figures 3.35C and 3.35D). Based on the average mass of RNA in each cell (derived from bulk RNA samples from know number of cells) and the average length of mRNAs in the human genome, we estimated that each GM12878 cell contains on average 80,000 mRNAs. However, we observed striking cell-to-cell differences in the total transcript number of single cells, with some cells expressing <50,000 mRNAs and others almost 300,000. In contrast, pool/split experiments exhibited remarkable uniformity (between 50,000 and 100,000 transcripts), and agree well with prior expectations. It is therefore unlikely that the observed cell-to-cell variability is due to technical noise.

Because transcriptional regulators play a crucial role in defining the gene expression state of cells, we examined the expression of several well-known general transcription factors as well as major regulators of B-cell differentiation (Figure 3.35E). Remarkably, except for *IRF4*, which was usually expressed at several dozen copies, most factors were detected at <10 copies per cell, and were often not detected at all. We stress that this does not mean that they are not expressed. Given the 10% p_{smc} of the protocol, these observations are consistent with simple technical failure to detect them. It is also possible that there are no mRNA copies in some cells at the moment of harvest, especially if they are infrequently transcribed. Extending these observations to other functional groups, we assessed proteins involved in translation (as a major group of genes with housekeeping functions, Figure 3.35F), splicing regulators (Figure 3.35G) and all transcription factors (Figure 3.35H). The median number of copies per cell was ~100 for translation proteins, ~10 for splicing regulators, and strikingly, only ~3 for transcription factors. This highlights the differences that exist between certain functional categories of genes in the robustness of their quantification in single-cell RNA-seq analysis, depending on their expression levels.

3.2.6 Identification of modules of coexpressed genes

Cell-to-cell gene expression variability may occur on the level of individual genes, but it can also occur in a coordinated fashion. A well-studied example is cell cycle phase-specific gene expression. In an asynchronous culture of cell, groups of genes expressed at specific times during the cell cycle will be present in a fraction of cells proportional to the time cells spend in each such phase.

To test whether we are able to identify the expected cell cycle-associated variation, and to search for any novel functional modules, we carried out Weighted Gene Coexpression Network Analysis (WCGNA, Zhang & Horvath 2005) using the copies per cell estimates for single cells and removing genes that were highly variant in pool/split libraries in order to minimize technical noise (see Methods and Figures 3.36 and 3.37). We identified 19 coexpression modules containing ≥10 genes each (Figure 3.38). The expression patterns of these modules were

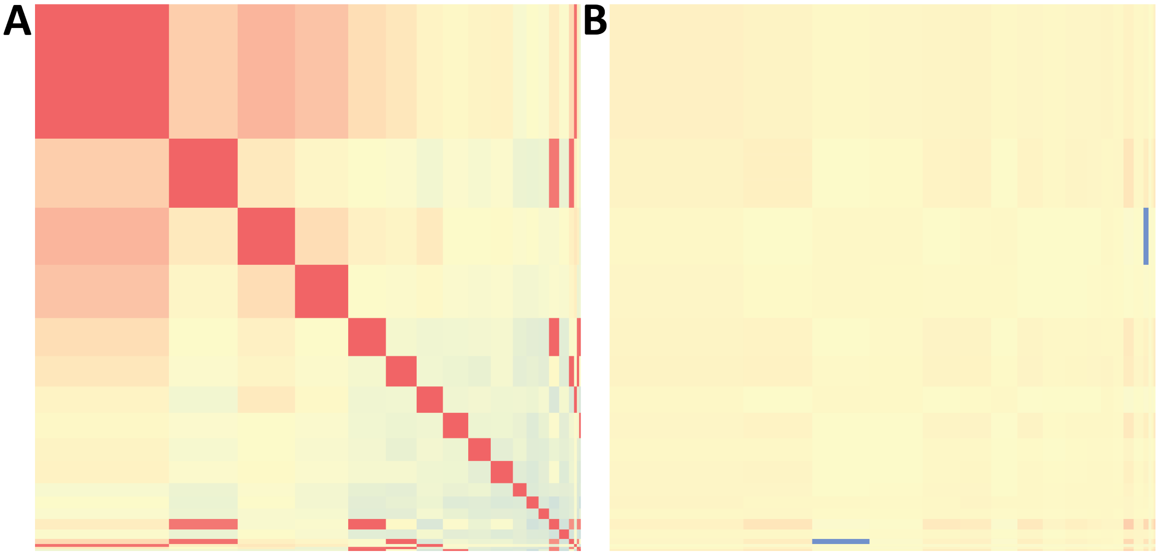


Figure 3.39: Average correlation within and between coexpression modules in single cells and pool/splits. Modules are sorted by decreasing size. (A) Single cells. (B) Pool/splits
.

mostly well differentiated among single cells and were absent from pool/split libraries (Figure 3.40A and Figure 3.39).

We then determined the Gene Ontology (GO) category enrichment of each module. The largest module (module 1) was highly enriched for GO categories relating to housekeeping gene functions (Table 3.2 and 3.3) and also for the G_1 and S phases of the cell cycle, and contained most genes that are generally highly expressed (Figure 3.40A). Module 6 was enriched for genes involved in the M phase of the cell cycle, likely corresponding to a single cell which was in that phase. We tested the plausibility of this explanation by measuring the fraction of unsynchronized GM12878 cells in the G_0+G_1, S, and M phases of the cell cycle using flow cytometry. About 14% of cells were in M phase, and the probability of capturing exactly one such cell out of 15 is 0.25; that is, these observations are consistent with this cell alone being in the M phase of the cell cycle (Figure 3.40B).

A more surprising observation was that the second largest module (module 2) was enriched for genes involved in splicing and mRNA processing. It is driven by an individual cell and two additional cells with a somewhat similar expression profile. This cell, however, was not an outlier when splice site usage patterns were compared between individual cells (data not shown). One interpretation of these observations is that there is a general upregulation of splicing and mRNA processing factors in this cell that does not necessarily result in a distinctive alternative splicing program.

Module 3 was enriched for metabolic cofactor and iron-sulfur cluster binding proteins, including proteins involved in mitochondrial respiratory chains. This is an intriguing observation as module 3 was mostly driven by the two cells exhibiting the highest total number of mRNA molecules per cell (Figure 3.35C, 4th and 5th columns in clustergram in Figure 3.40A), consistent with a generally elevated metabolic state.

We also carried out a mirrored analysis WCGNA where pool/splits were treated as single cells and vice versa. We did not observe significant GO enrichment beyond trivial terms in the largest modules (Figure 3.41 and Table 3.4).

In addition to the coexpression analysis, we also examined the relationship between the expression variability of genes and various genomic data about their promoters, including long-range chromatin interactions, DNA methylation status, histone marks, transcription start site sequence elements, and CpG islands. No robust explanatory correlation was evident (Figures 3.42-3.46), and we expect that data with less technical stochasticity will be needed to illuminate relationships of this kind.

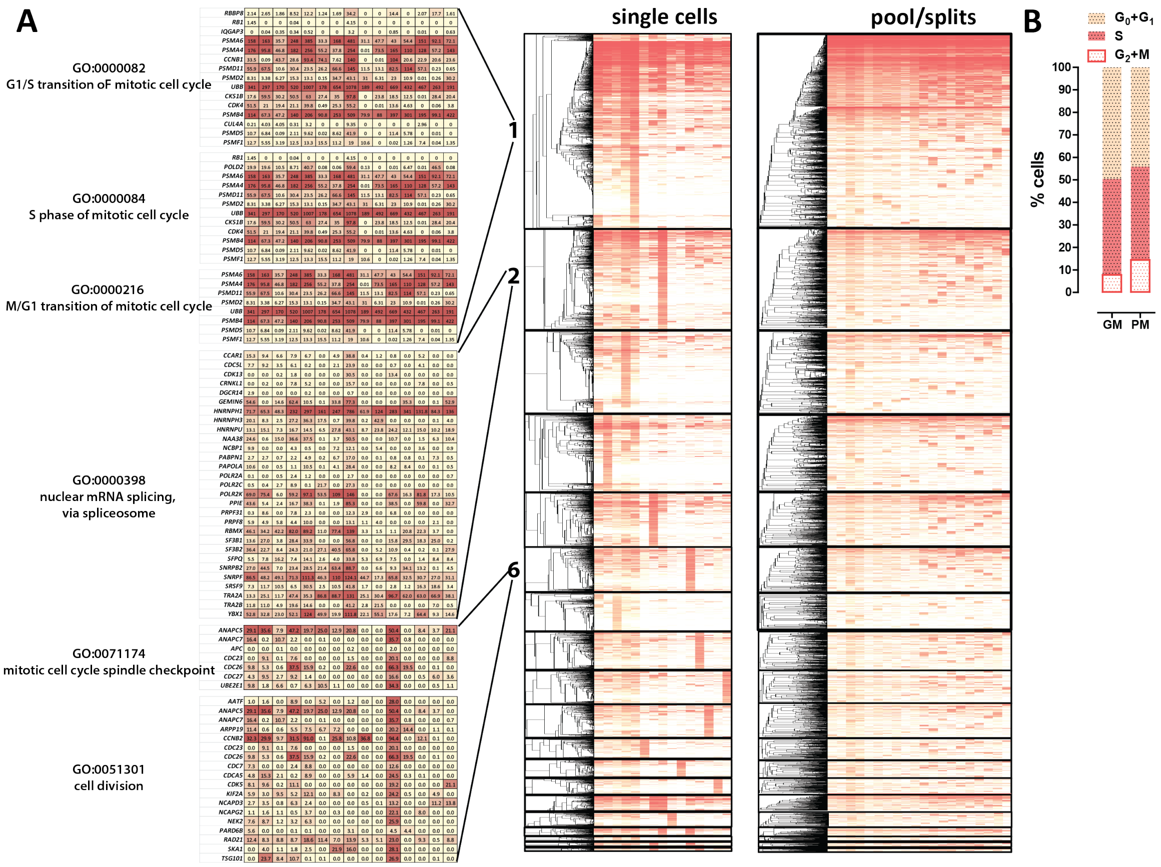


Figure 3.40: Gene coexpression modules derived from single GM12878 cells. Weighted gene correlation networks were constructed using the WCGNA R package (Langfelder & Horvath 2008). (A). Expression levels and hierarchical clustering of genes within modules (modules are sorted by number, which corresponds to their size) in single cells and pool/split experiments. Only genes are clustered (dendrograms on the left) and the identity of the cells and pool/split experiments is the same in each column (two right panels). The absolute expression values of genes belonging to representative GO categories associated with cell cycle phases (modules 1 and 6) and mRNA processing and splicing (module 2) are also shown. (B) Distribution of cell cycle states in a representative GM12878 cell population, in growth media (GM) and picking media (PM). The fraction of cells in M phase is consistent with 1 such cell being picked in a sample of 15.

3.2.7 Allele-biased expression at the single-cell level

Allele-specific gene expression has been previously reported to be widespread (Gimelbrant et al. 2007; Pickrell et al. 2010; Rozowsky et al. 2011; Reddy et al. 2012; Zhang & Borevitz 2009; McManus et al. 2010). An intriguing phenomenon observed for hundreds of genes in clonal lymphoblastoid cell lines (Gimelbrant et al. 2007; Chess 2012) is the random monoallelic expression of autosomal genes. However, those studies were conducted on large pools of cells, producing a snapshot of average allelic bias in the population, and leaving open the possibility that monoallelic expression is even more widespread on the single-cell level.

GM12878 cells are a good system for addressing this issue, as its fully phased heterozygous genome sequence is available (1000 Genomes Project Consortium 2012). We aligned RNA-seq reads in an allele-specific manner to the heterozygous GM12878 transcriptome and calculated allelic bias for each gene as the fraction of reads mapping to the maternal allele. We applied very stringent criteria for determining statistically significant allele-biased expression events based on the absolute transcript number

Table 3.2: Representative Gene Ontology categories enriched in coexpressed gene modules. Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al. 2009). The full list of enriched categories is available in Table 3.3.

Adjusted p-value	GO attrib ID	attrib name
		Module 1
<0.001	GO:0006415	translational termination
<0.001	GO:0006414	translational elongation
<0.001	GO:0070469	respiratory chain
<0.001	GO:0071845	cellular component disassembly at cellular level
<0.001	GO:0004129	cytochrome-c oxidase activity
<0.001	GO:0022904	respiratory electron transport chain
<0.001	GO:0030964	NADH dehydrogenase complex
<0.001	GO:0072413	signal transduction involved in mitotic cell cycle checkpoint
0.019	GO:0006626	protein targeting to mitochondrion
<0.001	GO:0048002	antigen processing and presentation of peptide antigen
<0.001	GO:0010467	gene expression
<0.001	GO:0006839	mitochondrial transport
0.007	GO:0006458	de novo' protein folding
<0.001	GO:0016071	mRNA metabolic process
<0.001	GO:0000216	M/G1 transition of mitotic cell cycle
0.014	GO:0000502	proteasome complex
0.005	GO:0060333	interferon-gamma-mediated signaling pathway
<0.001	GO:0000084	S phase of mitotic cell cycle
<0.001	GO:0000082	G1/S transition of mitotic cell cycle
0.005	GO:0000209	protein polyubiquitination
<0.001	GO:0008380	RNA splicing
		Module 2
<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
0.017	GO:0005681	spliceosomal complex
<0.001	GO:0006397	mRNA processing
		Module 3
<0.001	GO:0051186	cofactor metabolic process
0.002	GO:0051539	4 iron, 4 sulfur cluster binding
0.021	GO:0051536	iron-sulfur cluster binding
		Module 6
0.027	GO:0005680	anaphase-promoting complex
0.001	GO:0007094	mitotic cell cycle spindle assembly checkpoint

estimates and taking into account the challenges presented by the nature of single-cell RNA-seq data (see Methods for details). Previous studies have evaluated allele-biased expression examining the ratio of reads mapping to each allele; this approach, however, is not directly applicable to single-cell data generated with the SMART-seq protocol because of the large number of heterozygous reads that may be sequenced from a very small number of original founder molecules in the cell. For this reason, we used the estimated absolute number of mRNA molecules in the cell to derive estimates for the absolute number of mRNA molecules from each allele, and required that the allelic ratio for both reads and estimated mRNA copies passes a binomial test for significance. Finally, we also tested for the possibility that apparent allelic biases are in fact due to differential stochastic capture of the two alleles (the details are described in the Methods

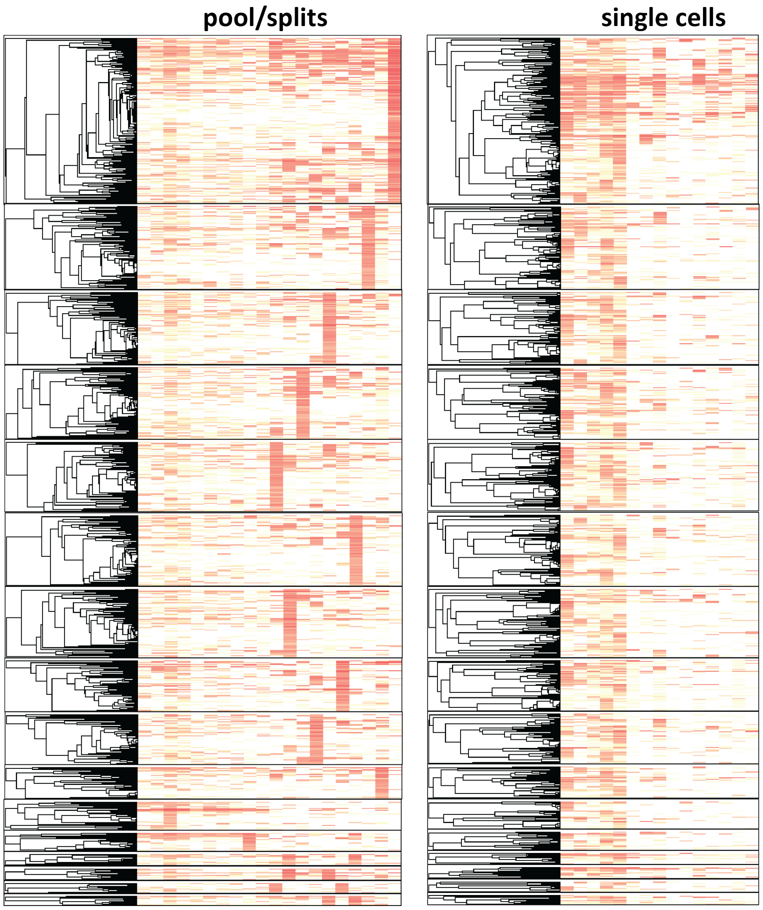


Figure 3.41: Mirrored coexpression analysis of pool/split and single cell datasets. The same analysis presented in Figure 4 was carried out treating pool/splits as single cells and vice versa.

section). This analysis was carried out on both single-cell and pool/split libraries, and also on 10ng bulk RNA libraries (for which only allelic bias on the level of reads was evaluated).

GM12878 are derived from a female donor and it is well-documented that in mammalian females, the X chromosome undergoes random inactivation early in embryonic development (Lyon, 1961). As a validation of our pipeline, we first examined the allelic bias of genes located on the X chromosome, and found that in all single cells, expression was exclusively from the maternal X chromosome. We performed all subsequent analysis excluding X-chromosome genes.

We observed good reproducibility of allelic bias profiles in 10ng bulk RNA libraries (Figure 3.47A), with most genes being expressed from both alleles (Figure 3.47D). Allelic bias was also highly reproducible in 30-cell and 100-cell pools (Figure 3.48). In contrast, allelic bias profiles of single cells correlated poorly with each other and a large fraction of genes were apparently monoallelically expressed from different alleles in different cells (Figure 3.47B). The majority of highly expressed genes (≥100 copies per cell) exhibited biallelic expression while most genes at low expression levels were measured as monoallelically expressed (Figure 3.47F). We then compared allelic bias variability for individual genes across individual single cells, focusing only on cells in which statistically significant allelic bias was observed, and observed frequent switching between the two alleles (Figure 3.47G, Figure 3.49A)

These observations can be explained as a combination of three factors. First, it has been previously reported that allelic bias is more common among genes expressed at low levels (Gimelbrant et al. 2007, Reddy et al. 2012). A second explanation is the phenomenon of "transcriptional bursting" (Raj & van Oudenaarden 2008;

Dar et al. 2012). A single transcription burst produces several mRNA molecules from a single allele. If all mRNAs from a gene in a given cell at a given moment are the product of one or a small number of such bursts, all copies would originate from only that allele. Finally, stochastic effects due to the low single-molecule capture efficiency of the protocol undoubtedly play a role. The fewer founder molecules are captured, the more likely it is that they belong to only one allele. We therefore performed the same analyses in pool/split libraries and observed a broadly similar (although always lower) fraction of genes passing all significance tests for allelic bias (Figures 3.47C, 3.47E and 3.49). Thus, it is at present difficult to draw confident conclusions about the prevalence of random monoallelic expression at the single cell level. Lowering the level of technical stochasticity will be necessary for this issue to be resolved.

3.2.8 Alternative splicing at the single-cell level

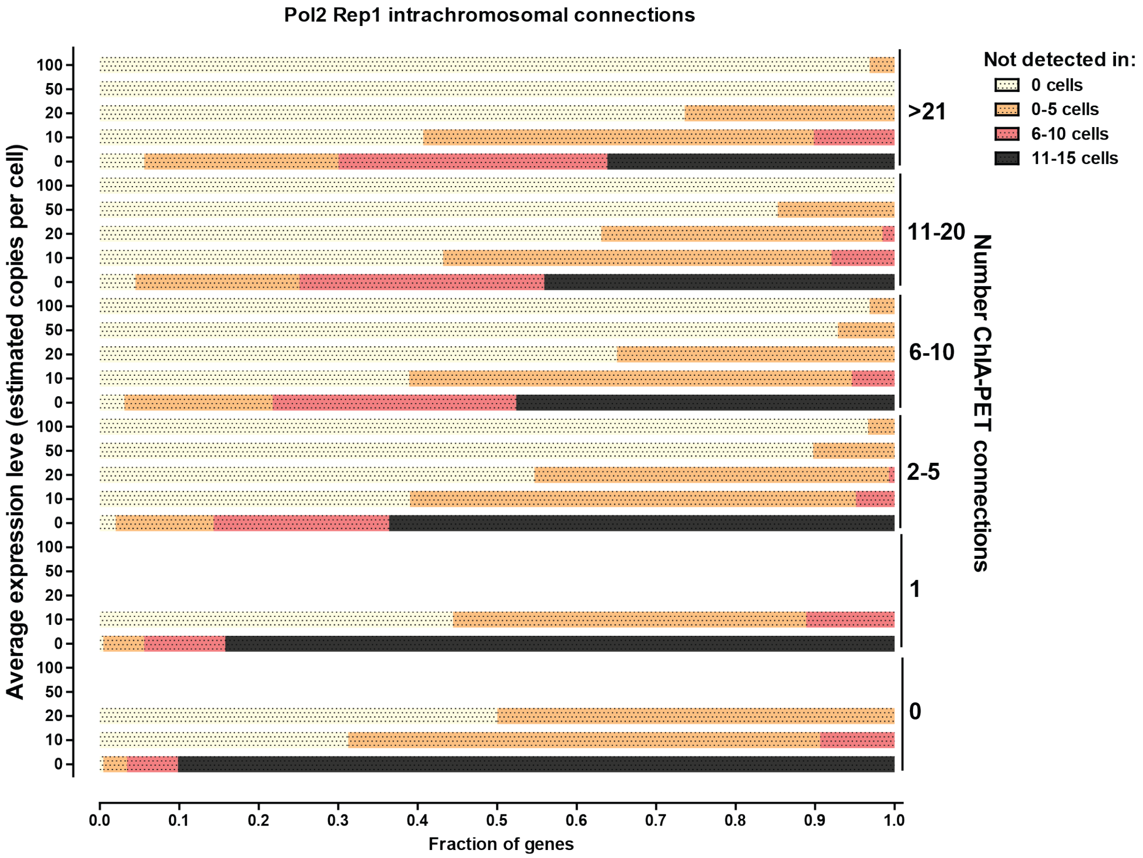
Previous studies have suggested that most genes in mammalian genomes undergo some alternative splicing (Mortazavi & Williams et al. 2008; Wang et al. 2008; Djebali & Davis et al. 2012). At present, however, the biological relevance of the majority of these alternative isoforms is still uncertain and stochastic noise in the splicing machinery is one explanation (Sorek et al. 2004; Melamud & Moult 2009). Characterizing alternative splicing at the single-cell level is highly relevant to elucidating the importance of alternative splicing events, as it in principle provides detailed information about their frequency both within single cells and populations of cells that is otherwise masked in bulk RNA-seq measurements. A second important question is how consistent the alternative splicing patterns observed on the population level are when examined at the level of individual cells.

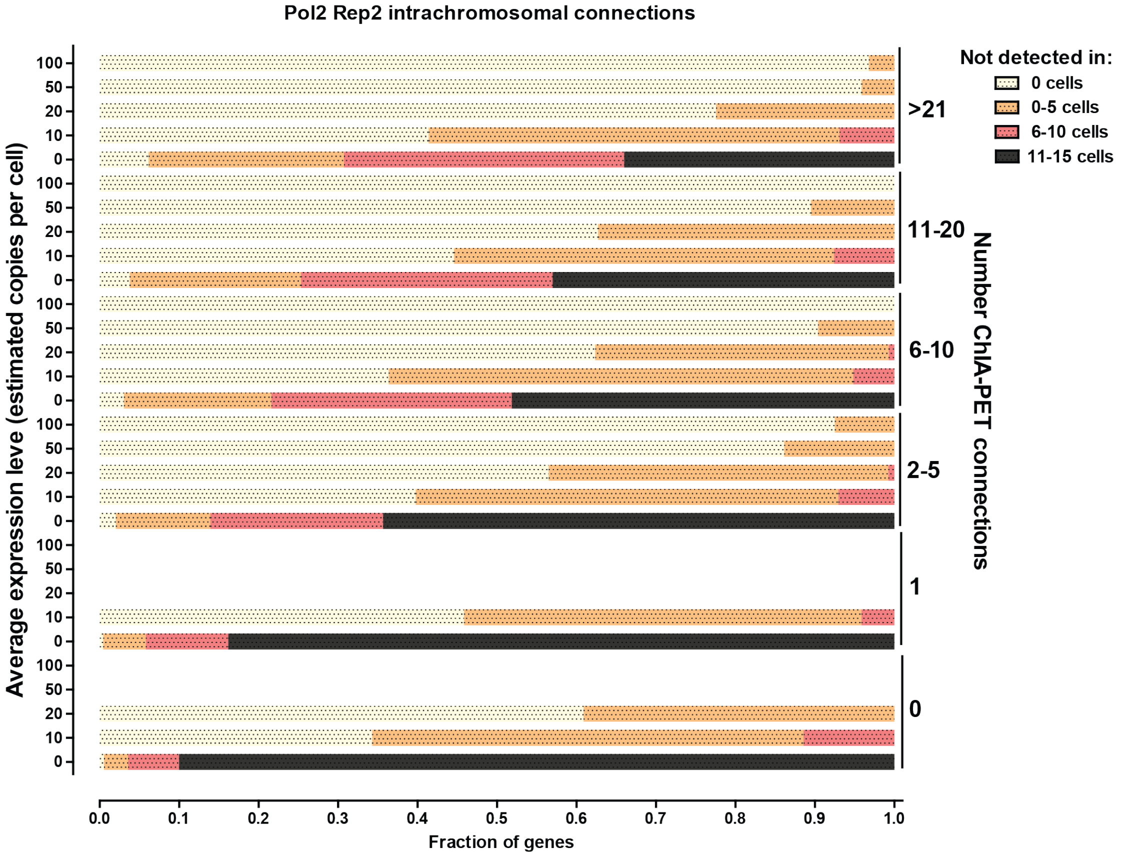
We quantified alternative splicing using the intron-centric splice inclusion ψ score approach (Pervouchine et al. 2013). Details of our mapping and analysis pipeline are described in the Methods section. For reasons given there, we adopted a conservative approach and only analyzed novel splice junctions for which at least one of the donor or acceptor sites has already been annotated in GENCODE V13 (Harrow et al. 2012), thus avoiding library-building artifacts.

We detected between 200 and 2000 novel splice junctions satisfying these criteria in each individual cell (Figure 3.50). This number is certainly an underestimate given the low p_{smc}. About 35% of novel junctions connected two annotated exons (Figure 3.51A, Figure 3.52A); most of these represent novel exon skipping events. In another 60% the unannotated donor or acceptor site was internal to the gene. These were concentrated close to already annotated splice sites (Figure 3.52B and C). In particular, novel acceptor sites peaked at the +3 and −3 position downstream of annotated sites representing mostly instances of NAGNAG tandem acceptor sites (Hiller et al. 2004; Bradley et al. 2012). Novel 5' donor sites were fewer in number and peaked at +4 and −4 positions relative to annotated donor sites thus shifting the coding frame of the transcript. This is a phenomenon we previously also observed in bulk RNA-seq data (See Chapter 1), the significance of which is at present not clear. The proportions observed were independent of the read coverage and estimated number of copies per cell thresholds applied (Figure 3.54A).

We also examined the distribution of novel splices across individual single cells and found that the majority of them were found in only a single cell, with <10% found in two cells, and very few in three or more cells (Figure 3.51B, 3.53B). While this result could be greatly affected by p_{smc} issues, it was independent of the read and estimated transcript copies threshold used (Figure 3.54), suggesting that most novel splices are indeed only present in a small fraction of cells.

We asked how often multiple splice sites are used at the single-cell level. In bulk RNA-seq at a threshold of 15 distinct read fragments, a numeric minority of ψ scores were equal to 1 (i.e. exclusive use of only one donor-acceptor pair). The presence of alternative splice sites is thus widespread. Nevertheless, in most cases, ψ was close to 1. The vast majority of novel splices received very low inclusion scores (Figure 3.51C) and would generally be considered to be the result of biological noise in the splicing system). In contrast, in single cells, one dominant splice site was the norm except for very highly expressed genes ($\geq$ 100 copies per cell), for which a wide diversity of splice site usage was seen (Figure 3.51D, Figure 3.55). As this observation was true even for genes expressed at $\geq$ 50 copies per cell, we believe it is not a p_{smc} arti-


Pol2 Rep1 intrachromosomal connections
Not detected in:
0 cells
0-5 cells
6-10 cells
11-15 cells
Average expression leve (estimated copies per cell)
100
50
20
10
0
>21
11-20
6-10
2-5
1
0
Number ChIA-PET connections
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction of genes


Pol2 Rep2 intrachromosomal connections
Not detected in:
0 cells
0-5 cells
6-10 cells
11-15 cells
Average expression leve (estimated copies per cell)
100
50
20
10
0
>21
11-20
6-10
2-5
1
0
Number ChIA-PET connections
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction of genes

fact. It is an interesting and open question why very highly expressed genes (enriched for genes with housekeeping function) exhibit very high levels of alternative splicing in single cells. These results differ significantly from the same analysis carried out on novel splice junctions (Figure 3.51E, Figure 3.56). Somewhat surprisingly, we found that a significant proportion of novel splices had ψ scores of 1 in single cells; this was true, however, only for genes expressed at lower levels ($\leq$ 50 copies) and it is therefore possible that it is mostly a p_{smc} artifact. In contrast, in highly expressed genes, no novel junctions received a dominant (≥ 0.5) ψ score. However, the scores were still consistently higher than what is observed for novel splices in bulk RNA samples.

Finally, we evaluated the consistency of splice site usage between individual cells. We applied a statistical framework similar to the one used to analyze allelic bias and derived a list of dominant splice junctions in each cell, taking into account the estimated absolute number of copies and the stochastic capture effects. We asked how often the dominant splice site changes between different cells. We found 282 such genes in single cells, suggesting the phenomenon may be widespread. The genes involved were enriched for ribosomal and translation proteins, and also, intriguingly, for proteins involved in RNA splicing and processing (Table 3.6). We tested this single-cell variation against pool/split experiments, in which we found very few genes with different dominant splice sites across libraries. (Figure 3.51F and 3.51G, Figure 3.57). This argues that much of the observed alternative splicing variation is in fact due to biological differences between cells, and is in agreement with the bimodality of splicing in individual mouse immune cells observed previously (Shalek et al. 2013).

3.3 Discussion

The two major goals for single-cell RNA-seq are to obtain high-resolution transcriptomes for rare cell types or states and to measure the differences in RNA expression and processing between individual cells. We showed that the first goal can be achieved by studying 30-100 cell pool samples even in the absence of perfect capture of each and every individual RNA molecule. Our conclusion is consistent with independent 80-cell measurements (Ramsköld & Luo et al. 2012). The pools reproduce the expression profiles (Figure 3.31) and allelic-bias patterns (Figure 3.48) of the larger population, and similar numbers of genes and splice junctions are detected in them (Figure 3.58, Figure 3.31). The approach is applicable to cells collected by laser-capture, micromanipulation, or cell sorting based on molecular markers or reporter-gene expression. This defines a path forward for the transcriptomic characterization of many previously inaccessible rare cell types and states, including transient cell types in embryonic development, diverse neuronal types in the brain, and cells in tumors.

To understand single-cell variation in the GM18278 reference cell line, we generated and analyzed high-quality, state-of-the-art single-cell RNA-seq data individual GM12878 cells. Nevertheless, like prior studies, our data display significant stochasticity. We present experimental and analytical approaches for measuring and accounting for technical stochasticity. We introduced and measured single-molecule capture efficiency, the key parameter influencing technical stochasticity and find that its value is around 0.1 with the current SMART-seq protocol. We controlled for technical stochasticity experimentally by carrying out pool/split experiments, which allowed us to identify significant biological variation over and above technical variation.

In line with previous observations, we find great cell-to-cell variability in gene expression

Figure 3.42 ***(preceding page)*****: Relation between the long-range chromosomal element connectivity of TSSs and gene expression stochasticity**. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the number of ENCODE ChIA-PET connections to TSSs in K562 cells (replicates 1 and 2). K562 was used as the closest cell line to GM12878 for which such data is currently available; ChIA-PET connections were downloaded from the UCSC Genome Browser. Within each group of genes defined by the number of ChIA-PET connections, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.

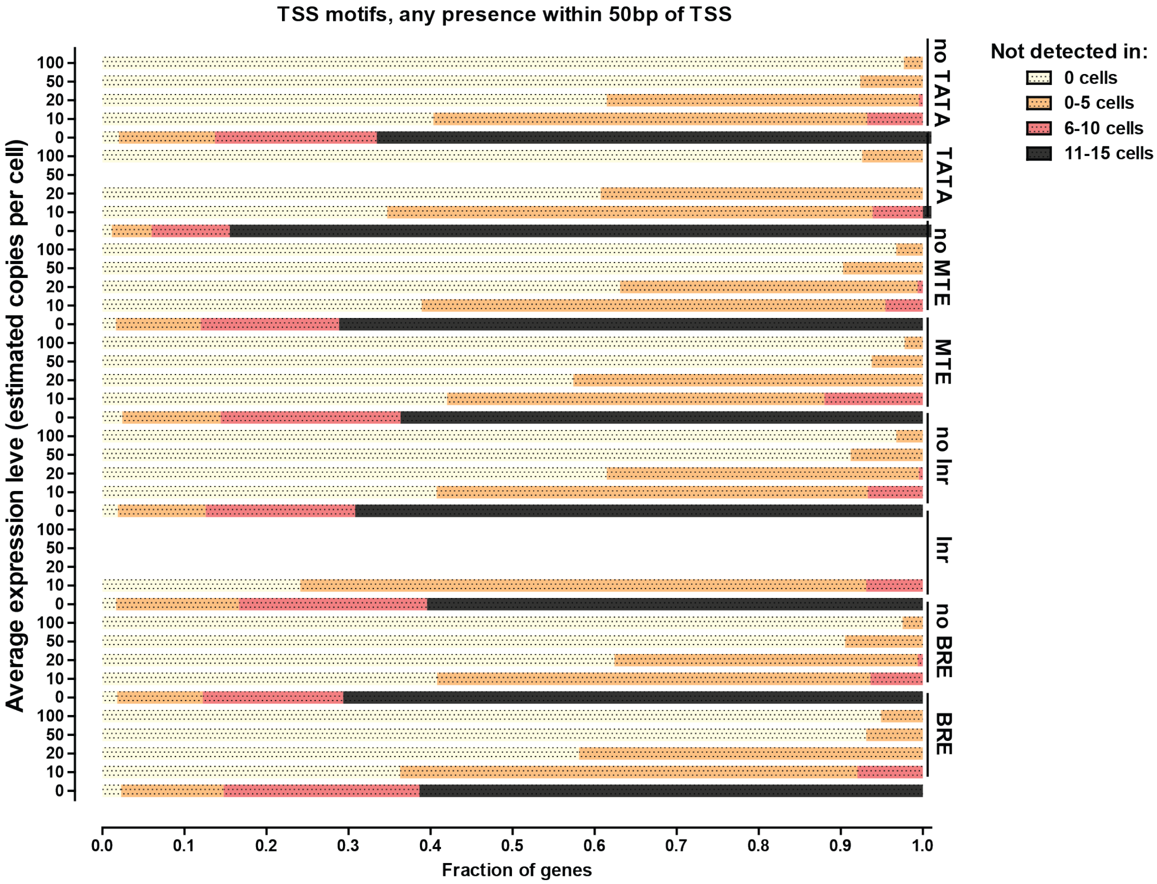


Figure 3.43: Relation between the presence of TSS-associated sequence elements and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the presence or absence of sequence motifs at TSSs (defined by FIMO using position weight matrices obtained from Jin et al., 2006). Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.

levels. We demonstrate that at least some of this variation is due to coordinated differences in the expression of biologically coherent sets of genes, for example, genes associated with different phases of the cell cycle, as well as the surprising observation of a coexpression module enriched for genes involved in mRNA processing and splicing.

We also observed unexpected levels of cell-to-cell variation in autosomal allelic expression bias and alternative splicing. The observation of allele switching between single cells could be explained as a technical artifact given that a similar, although always lower, level of switching was observed in pool/split libraries. We therefore consider this a provisional result in need of further investigation with improved experimental protocols. The observed frequency of major splice switching in single cells is a stronger effect, and based on comparison with pool/split experiments, it is unlikely to be the sole result of technical stochasticity.

Transcriptional bursting is a main candidate for a biological explanation for these observations. If a gene is expressed in a series of infrequent relative to the half life of its mRNAs such bursts, at any given time the population of mRNAs in the cell is likely to originate from only one allele. This can also explain the observed variation in alternative splicing. It is possible that the same set of factors influencing splice site choice maintain physical association with the gene dur-

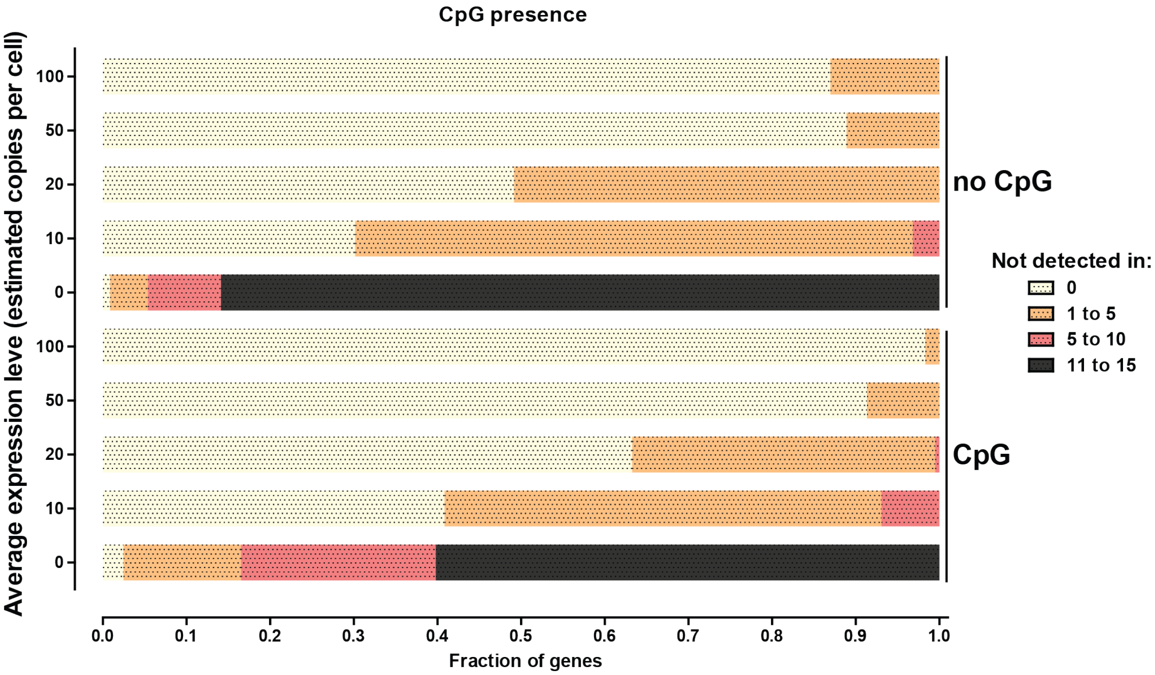


Figure 3.44: Relation between the presence of CpG islands near TSSs and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the presence or absence of CpG islands within 1kb of the TSS. Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.

ing a transcriptional burst leading to a particular splicing pattern being highly favored locally. Future studies should shed light on these intriguing questions. In-depth investigation of individual cases by other methods will naturally be needed to validate the initial global observations. The limitations of the current single-cell RNA-seq assay make it possible to capture the general pattern, and the data are a source of candidates for detailed validation and study, but no single candidate event is assured of reproducing.

Much biology involves genes whose transcript levels are in the range highly affected by technical variation. Considerable improvement in the single-molecule capture efficiency is therefore needed. Based on our simulations and results from pool/split experiments, we estimate that an increase in p_{smc} from 0.1 to 0.5 would be a major leap forward, while $p_{smc} \geq 0.8$ would provide sufficient measurement reliability for virtually any biological use. The experimental framework provided here would be highly useful for evaluating future improvements in protocols.

We also found that the amount of mRNA per cell is highly variable between individual cells. This is both biologically interesting and important for analysis pipelines as RPKM-type metrics are not reliable given such large difference in total RNA per cell (Lovén et al. 2012; Lin et al. 2012). At present, the direct relationship between the absolute number of mRNA copies per cell and the number of sequencing reads in a library is lost due to the fragmentation of amplified cDNA molecules that is a common feature of available protocols resulting in multiple distinct but overlapping sequencing fragments for each founder RNA molecule. mRNA copy number therefore has to be estimated back from FPKMs with the help of spike-in sequences. This is far from a flawless method for doing so, as first, it depends on the accuracy of quantification of the spike-ins, and second, it assumes the absence of systemic differences between spike-in RNAs and endogenous RNAs. The ideal single-cell RNA-seq assay would combine a very high single-molecule capture efficiency with an amplification-free, and preferably, also reverse transcription-free, direct RNA sequencing that

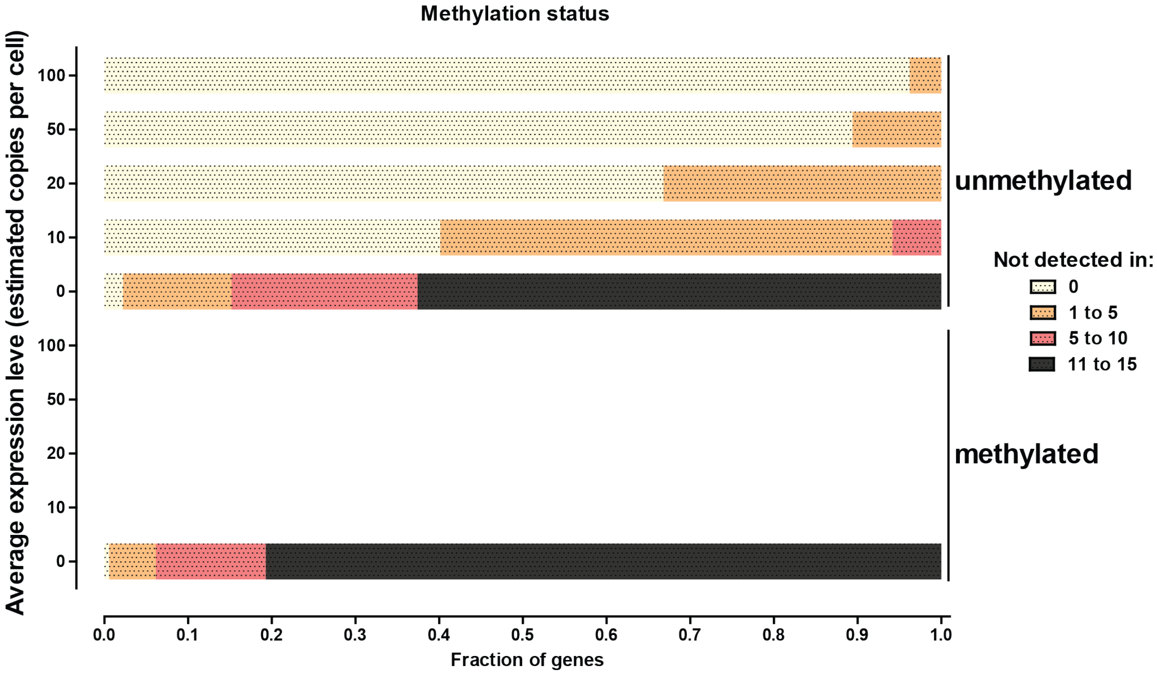


Figure 3.45: Relation between the methylation status of promoters and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the methylation status of their promoters as defined using ENCODE reduced-representation bisulfite sequencing data (RRBS) for the GM12878 cell line from Varley et al., 2013, downloaded from the UCSC Genome Browser. Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.

allows the direct counting of transcript copies. Emerging sequencing technologies (Branton et al. 2008; Schadt et al. 2010) already hold promise for such radical improvements.

3.4 Addendum: More Recent Developments in the Field

Between the completion of this work and the writing of this chapter, a number of studies appeared, which addressed some of the issues adressed in it.

The observation that certain genes exhibit dramatic variation in splice site usage between individual cells in a population was confirmed independently (Shalek et al. 2013), including an orthogonal validation by SM-FISH.

Two groups reported widespread random monoallelic expression between individual cells (Xue et al. 2013; Deng et al. 2014), however they paid significantly less attention to the problem of technical noise than we did (this was especially true in the case of Xue et al. 2013), thus the question whether there indeed is widespread such variation cannot be considered fully resolved yet.

In this work, we generated our libraries generated manually, however over the course of 2013, automated microfluidics-based methods for carrying out RNA-seq became very popular, in particular the Fluidigm C1 system. This has allowed very large numbers of individual cells to be profiled, and there are reasons to think that the p_{smc} is higher for libraries generated on the Fluidigm (Wu et al. 2014; Islam et al. 2014) though still not nearly as high as desired.

A new version of SMART-seq protocol, SMART-seq2, was described (Picelli et al. 2013; Picelli et al. 2014), which supposedly also improves the p_{smc}, maybe up to 0.4-0.5, although this was not measured in a way comparable

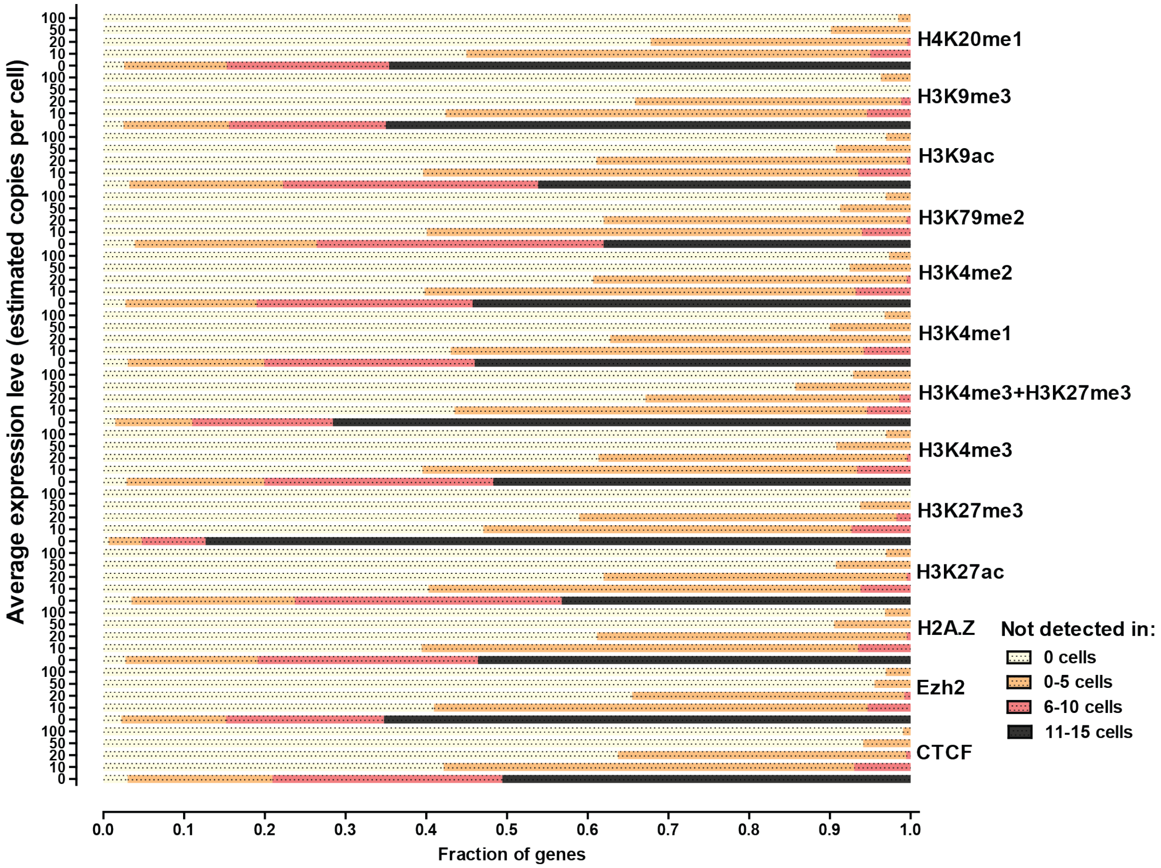


Figure 3.46: Relation between the histone modification status of promoters and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the presence or absence of the various histone marks, the bivalent H3K4me3 + H3K27me3 combination of marks, CTCF and Ezh2 as defined from ENCODE data for the GM12878 cell line using the peak calls available from the UCSC Genome Browser. Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.

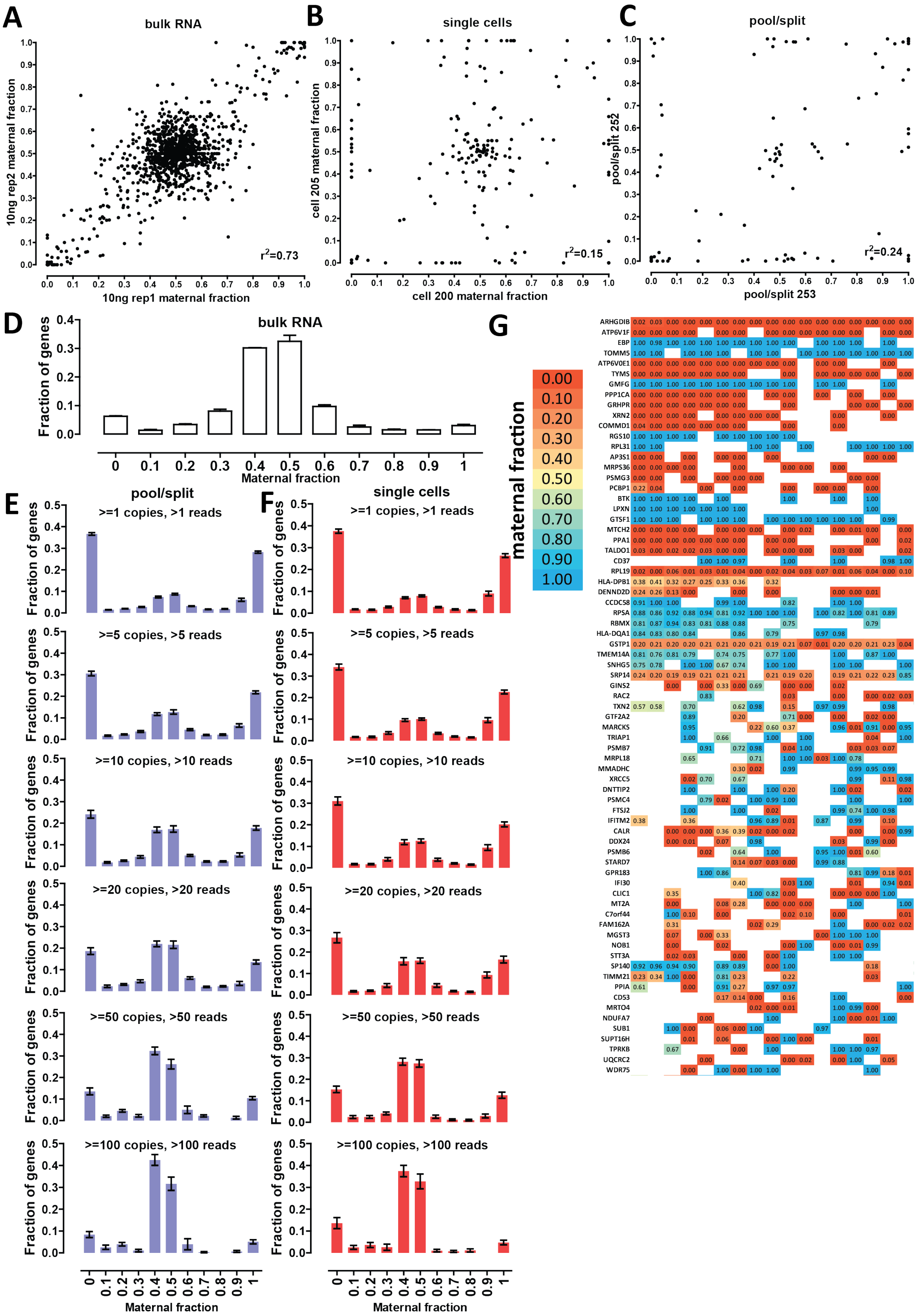
to the way we did it, giving hopes that the combination of SMART-seq2 and Fluidigm may achieve even higher efficiencies. As I write these words, this has not yet been tested.

None of these improvements, however, address the issue of directly counting individual transcripts. A new approach towards accomplishing this was described recently (Islam et al. 2014), however it suffers from the problem of being a 5'-tagging confounding analyses requiring capture of the whole transcript. Full-length single-molecule RNA sequencing remains the goal for the future.

3.5 Methods

3.5.1 Single cell collection

Single cell harvesting from live cultures requires a micropipet with a polished glass tip with an approximate diameter of 40μm. Borosilicate glass microfiber pipettes (FHC omega dot fiber 30-30-0) were pulled on a Sutter Instruments P80/PC microcapillary puller with the following parameters: 750 heat, 150 pull, 100 velocity, 5 time. After pulling, the microcapillary tips were mounted on a glass microscope slide using modeling clay, and broken by closing a pair of #5 Dumont forceps around the glass. We used a scaled eyepiece reticle to judge the width of the break


A
bulk RNA
10ng rep2 maternal fraction
10ng rep1 maternal fraction
r²=0.73
B
single cells
cell 205 maternal fraction
cell 200 maternal fraction
r²=0.15
C
pool/split
pool/split 252
pool/split 253
r²=0.24
D
bulk RNA
Fraction of genes
Maternal fraction
E
pool/split
F
single cells
>=1 copies, >1 reads
>=5 copies, >5 reads
>=10 copies, >10 reads
>=20 copies, >20 reads
>=50 copies, >50 reads
>=100 copies, >100 reads
Fraction of genes
Maternal fraction
G
maternal fraction
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
ARHGDIB
ATP6V1F
EBP
TOMM5
ATP6V0E1
TYMS
GMFG
PPP1CA
GRHPR
XRN2
COMMD1
RGS10
RPL31
AP3S1
MRPS36
PSMG3
PCBP1
BTK
LPXN
GTSF1
MTCH2
PPA1
TALDO1
CD37
RPL19
HLA-DPB1
DENND2D
CCDC58
RPSA
RBMX
HLA-DQA1
GSTP1
TMEM14A
SNHG5
SRP14
GINS2
RAC2
TXN2
GTF2A2
MARCKS
TRIAP1
PSMB7
MRPL18
MMADHC
XRCC5
DNTTIP2
PSMC4
FTSJ2
IFITM2
CALR
DDX24
PSMB6
STARD7
GPR183
IFI30
CLIC1
MT2A
C7orf44
FAM162A
MGST3
NOB1
STT3A
SP140
TIMM21
PPIA
CD53
MRTO4
NDUFA7
SUB1
SUPT16H
TPRKB
UQCRC2
WDR75

at about 40μm. After breaking, the tips were smoothed using a microforge. To prevent sticking of cells to the interior of the capillary, we treated the pipettes with Sigmacote by attaching Tygon tubing and a syringe to the blunt end of the microcapillary, and drawing the Sigmacote solution into the tip. This also provided assurance that the forged tips had not closed. The capillaries were then rinsed with distilled water twice using the same technique, and allowed to dry at room temperature overnight.

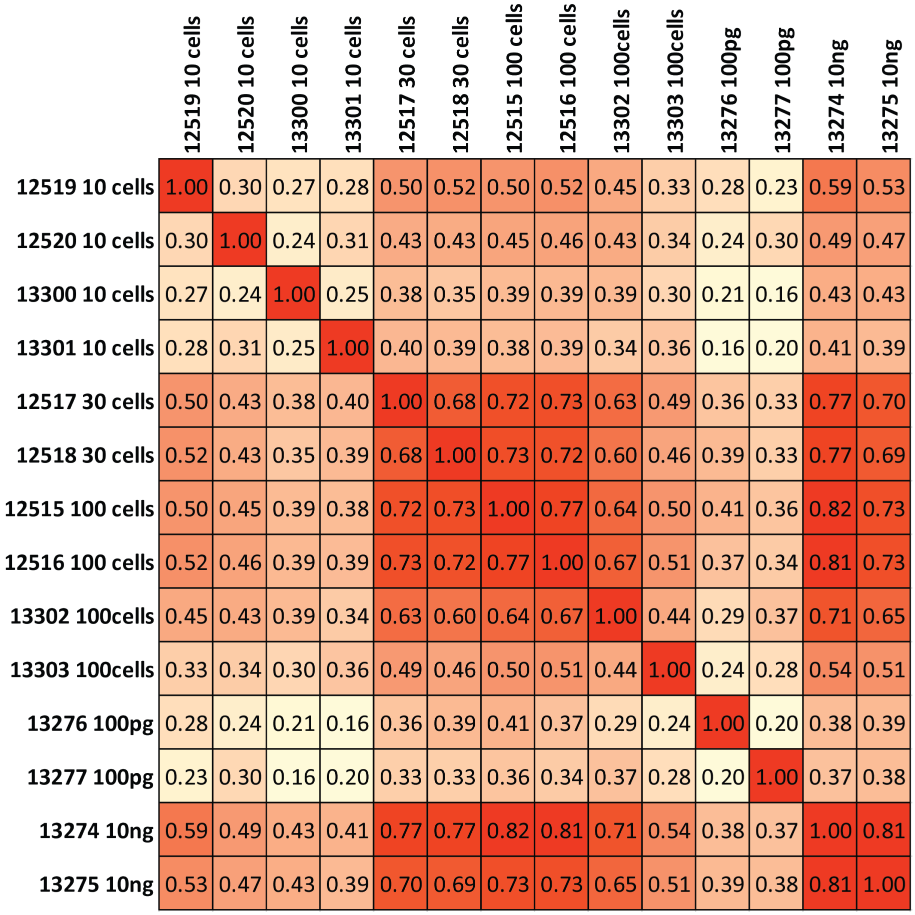
An aliquot (5×10^6 cells) of GM12878 cells were thawed rapidly and cultured in 10mL of medium (RPMI 1640, 15% FBS, 2mM L-glutamine, 1% penicillin-streptomycin). The cells were grown at density of $2 \times 10^5 - 2 \times 10^6$ cells/mL of medium for 11 days until harvest. On the day prior to harvest, the culture volume was increased to 100mL by the addition of fresh medium, bringing the density to 2×10^5 cells/mL. At harvest time (23 hours later), cells were triturated using a 10mL pipette, and a small aliquot (~100μLs) of the culture was removed. A few μLs of the cell suspension was added to a 250μL volume of "cell picking medium" (RPMI1640 with 15% Superblock (Pierce catalog #37515) and 2mM glutamine). This diluted cell suspension was then placed in a 3cm culture dish and returned to the 37 °C incubator for 10 minutes prior to single cell harvesting.

The microcapillary pipet was mounted on a micromanipulator and attached to a 100μL glass syringe via Tygon tubing. A dish of picking medium was brought to the illuminated stage on the phase contrast scope, and the tip was submerged using the micromanipulator. Picking medium was drawn up into the microcapillary to a height of about 75mm. The tip was removed from the picking medium, re-submerged into a dish from the incubator containing the dilute cell suspension, and lowered gently to the floor of the dish. Individual cells were aspirated into the pipet by gentle vacuum applied via the glass syringe. When a single cell had been aspirated, the tip was rapidly lifted out of the picking medium, and the picking dish was removed from the illuminated area of the stage. A small sliver of silanated cover glass (Molecular Dimensions, catalog #MD406) was then placed on a glass slide on the stage, and a 4.5μL drop of cell lysis solution was placed on the sliver with a Rainin P10 micropipette. The lysis solution contains 2.5μL of reaction buffer (Clontech SMARTer Ultra Low RNA kit), 1μL of 3 SMART CDS Primer IIA (Clontech) and 1μL of spike-in quantification standards. The drop of lysis solution was visualized on the illuminated area of the stage, and the pipette tip containing the picked cell was lowered into it. Gentle pressure was applied to the syringe to expel the cell from the pipette, and the tip was then lifted from the lysis solution. Visual confirmation was made at high power, while the cell dissolved in the lysis solution. The glass sliver was lifted from the stage using forceps, and placed in the bottom of a 200μL PCR tube. The tube was spun for 15 seconds at 10,000g, the sliver was removed, and the lysed cell was immediately frozen on dry ice. Twenty individual cells were collected in this way. We also collected two samples of ten cell pools into the same volume of lysis buffer, using the pipette picking method.

For ~100 cell pools, cells were first diluted in picking buffer to a concentration of 10 cells/μL. 10μL of the dilute cell suspension were added to 90μL of picking buffer in a 200μL PCR tube, and spun at 2500g for 90 seconds to pellet the cells. The tube was then mounted sideways in modeling clay on a glass slide, and the pellet was visualized under the phase contrast scope. A drawn glass pipette tip attached to the micromanipulator was advanced into the picking medium and the excess picking medium was withdrawn using the syringe. A 4.5μL aliquot of lysis buffer was then added to the cell pellet, and the lysate was spun and frozen as for the above samples.

After picking the individual and pooled cell samples, the remainder of the culture ($\sim 2 \times 10^7$

Figure 3.47 *(preceding page)*: Allele-biased expression at the single-cell level. (A,B and C) Correlation between allele bias between 10ng bulk RNA replicates (A), between two individual single cells (B) and between two pool/split libraries (C). Shown is the maternal fraction of reads for genes with at least 15 reads covering heterozygous positions for 10ng libraries and for genes with at least 10 reads covering heterozygous positions and expressed at more than 10 copies per cell for single cells and pool/splits. (D). Distribution of allele bias in bulk RNA samples ($\geq$15 reads covering positions). (E and F). Distribution of allele bias as a function of the read and copies threshold in single cell (E) and pool/split (F) libraries.



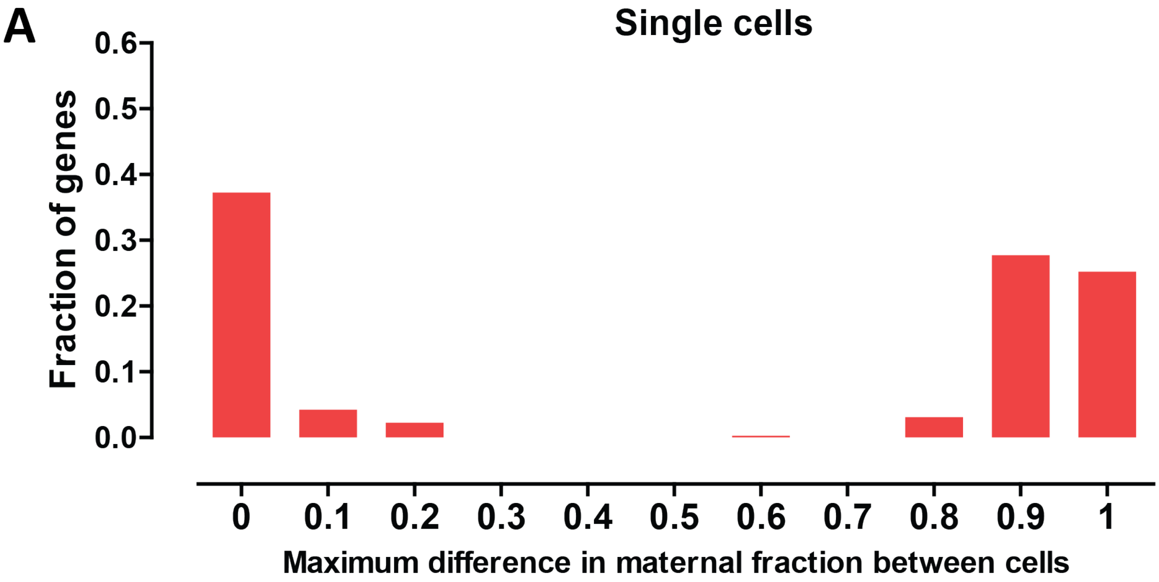
	12519 10 cells	12520 10 cells	13300 10 cells	13301 10 cells	12517 30 cells	12518 30 cells	12515 100 cells	12516 100 cells	13302 100cells	13303 100cells	13276 100pg	13277 100pg	13274 10ng	13275 10ng
12519 10 cells	1.00	0.30	0.27	0.28	0.50	0.52	0.50	0.52	0.45	0.33	0.28	0.23	0.59	0.53
12520 10 cells	0.30	1.00	0.24	0.31	0.43	0.43	0.45	0.46	0.43	0.34	0.24	0.30	0.49	0.47
13300 10 cells	0.27	0.24	1.00	0.25	0.38	0.35	0.39	0.39	0.39	0.30	0.21	0.16	0.43	0.43
13301 10 cells	0.28	0.31	0.25	1.00	0.40	0.39	0.38	0.39	0.34	0.36	0.16	0.20	0.41	0.39
12517 30 cells	0.50	0.43	0.38	0.40	1.00	0.68	0.72	0.73	0.63	0.49	0.36	0.33	0.77	0.70
12518 30 cells	0.52	0.43	0.35	0.39	0.68	1.00	0.73	0.72	0.60	0.46	0.39	0.33	0.77	0.69
12515 100 cells	0.50	0.45	0.39	0.38	0.72	0.73	1.00	0.77	0.64	0.50	0.41	0.36	0.82	0.73
12516 100 cells	0.52	0.46	0.39	0.39	0.73	0.72	0.77	1.00	0.67	0.51	0.37	0.34	0.81	0.73
13302 100cells	0.45	0.43	0.39	0.34	0.63	0.60	0.64	0.67	1.00	0.44	0.29	0.37	0.71	0.65
13303 100cells	0.33	0.34	0.30	0.36	0.49	0.46	0.50	0.51	0.44	1.00	0.24	0.28	0.54	0.51
13276 100pg	0.28	0.24	0.21	0.16	0.36	0.39	0.41	0.37	0.29	0.24	1.00	0.20	0.38	0.39
13277 100pg	0.23	0.30	0.16	0.20	0.33	0.33	0.36	0.34	0.37	0.28	0.20	1.00	0.37	0.38
13274 10ng	0.59	0.49	0.43	0.41	0.77	0.77	0.82	0.81	0.71	0.54	0.38	0.37	1.00	0.81
13275 10ng	0.53	0.47	0.43	0.39	0.70	0.69	0.73	0.73	0.65	0.51	0.39	0.38	0.81	1.00

Figure 3.48: Correlation between allelic bias in cell pools of different sizes.

cells) was spun down in two aliquots for 5 minutes at 1000g at 4 °C. The culture medium was removed, the pellet was rinsed with PBS, and re-spun as above. After the removal of PBS, both pellets were lysed with 1.2mL lysis buffer from the Ambion mirVana kit (catalog #AM1560). The lysates were then processed according to the manufacturer's protocol. After eluting total RNA from the columns, we performed a DNA digestion step to remove residual contaminating genomic DNA, using the DNA-free kit from Ambion (catalog #AM1907). After quality control with Qubit and the Agilent BioAnalyzer, the bulk prep total RNA was diluted to both 10ng/μL and 100pg/μL concentrations. We then added single microliter aliquots to the lysis buffer described above, and froze the samples for processing using the single cell protocol.

3.5.2 First strand cDNA synthesis and amplification

The frozen samples were brought to the lab bench on dry ice. Lysis and denaturation were accomplished by heating the samples for 3 minutes at 72 °C. The samples were spun down and placed in a cooling rack at 4 °C. 5.5μL of first strand reaction buffer (Clontech) was then added (2μL of buffer, 1μl of RNAse inhibitor, 1μl of dNTPs, 0.25μl of DTT, 1μL of SMARTer IIa oligos, and 1μl of SMARTScribe reverse transcriptase). The samples were reverse transcribed at 42 °C for 90 minutes and denatured at 70 °C for 10 minutes. After denaturation, the samples were spun down and 25μL of Ampure XP SPRI beads (Beckman Coulter genomics) were added. The samples were incubated for 8 minutes at room temperature, then the beads were separated on a magnet for 5 minutes. The supernatant solution was removed with a pipette,



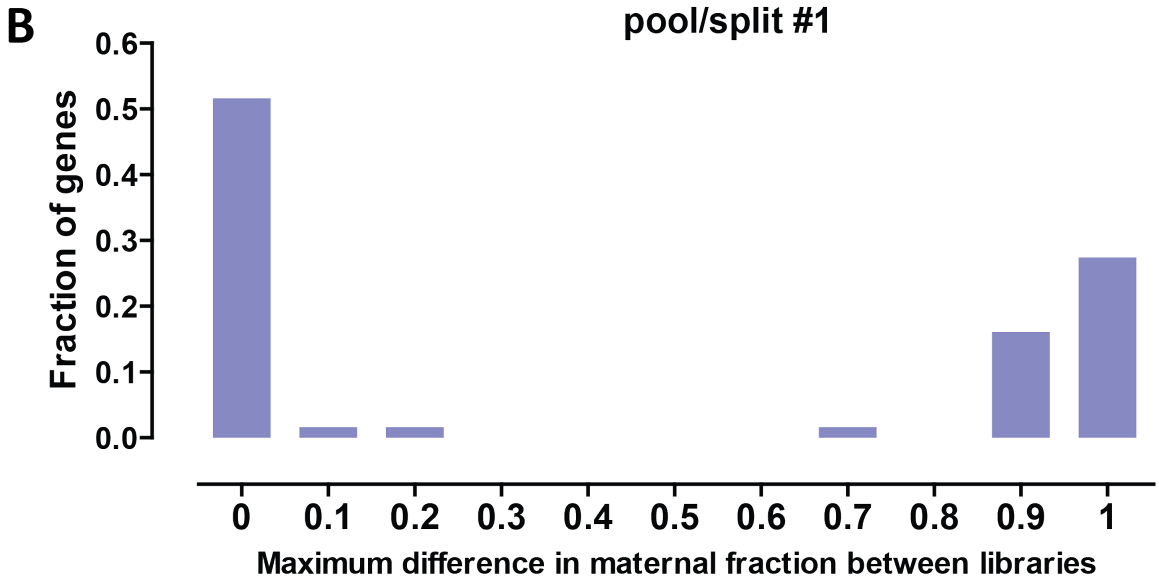


Figure 3.49: Changes in allele expression bias between individual cells and between individual libraries in pool/split experiment 1. Shown is the maximum difference between the maternal fraction of reads in single-cells (A) and the pool/split (B). Only gene/library pairs for which the ψ score passed all three tests for statistical significance of bias towards one splice (described in Methods) were included

and the beads were spun at 1000g for 1 minute to pellet. The sample was then placed back on the magnet, and excess supernatant was removed with a 10μL Rainin pipet tip. 50μL of amplification solution were then used to resuspend the beads (5μL of PCR buffer, 2μL of dNTPs, 2μL of amplification primers and 2μL of Advantage2 polymerase mix), and the samples were amplified under the following conditions: 1 minute at 95 °C, followed by cycles of 15 seconds at 95 °C, 30 seconds at 65 °C, 6 minutes at 68 °C, and final elongation for 10 minutes at 72 °C. Single cell and pool/split samples were amplified for 26 cycles, the 10 cell pools were amplified for 22 cycles, the 100 cell pools were amplified for 18 cycles, and the bulk prep RNA samples were

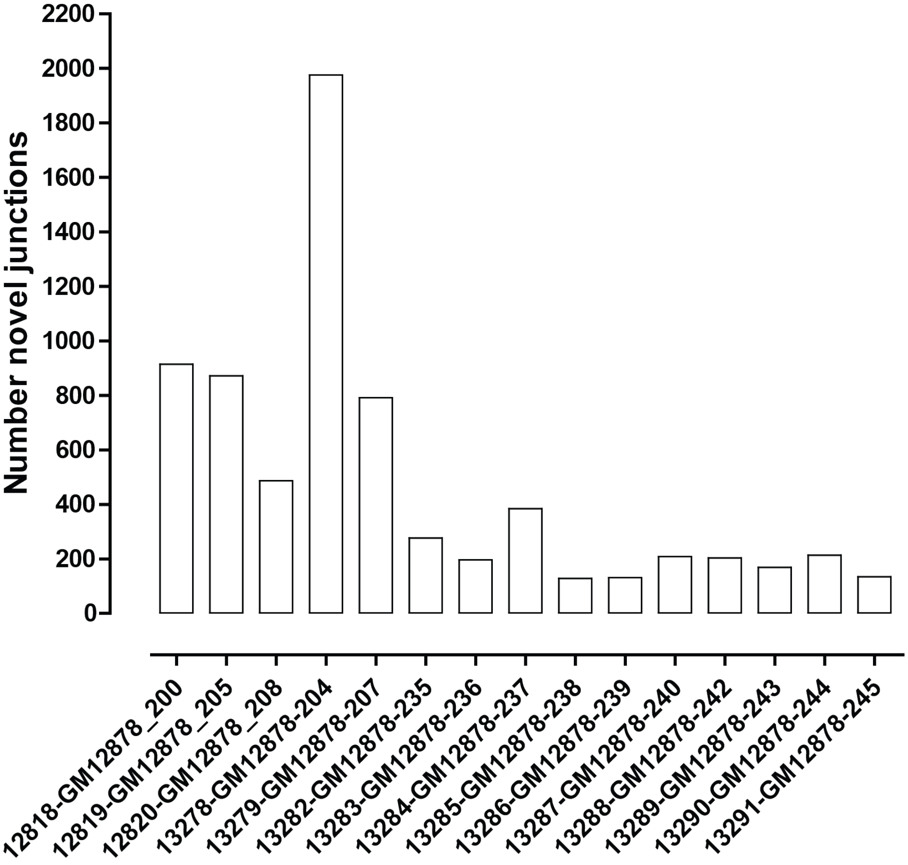


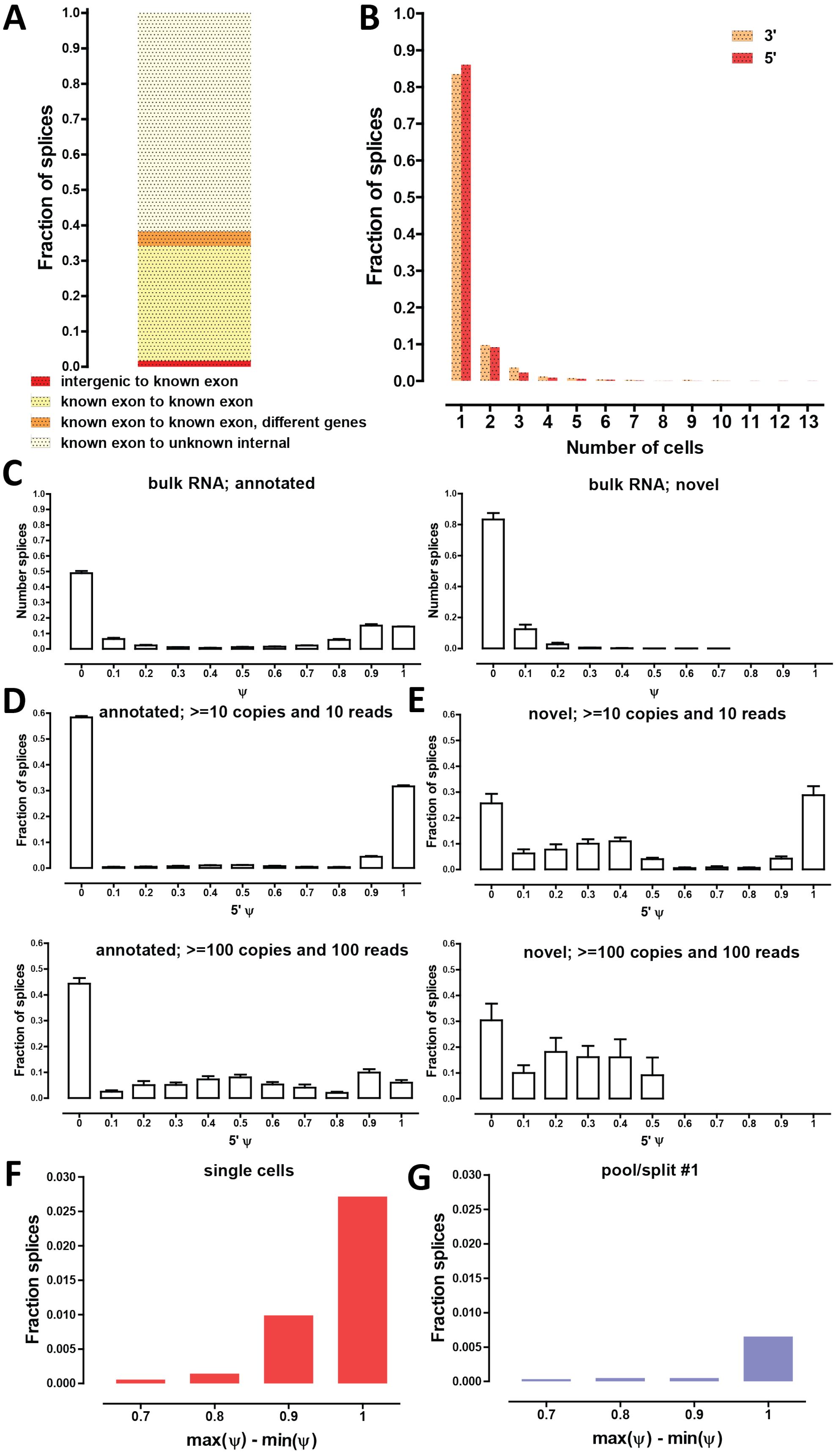
Figure 3.50: Number of novel splice junctions (connecting to annotated donor and/or acceptor sites) detected in individual cells.

amplified for 15 cycles. The amplified cDNA was spun down, and 90μL of Ampure XP beads were added. The beads were incubated with the amplified product for 8 minutes, then separated on a magnet for 5 minutes. The reaction solution was removed and the beads were washed twice with 200μL of freshly prepared 80% ethanol for 30 seconds. After the second ethanol wash, the beads were pelleted at 1000g for 1 minute, the residual ethanol was removed with a P10 Rainin pipette tip, and the beads were allowed to dry until the pellet showed signs of cracking. The beads were then resuspended in 20μL of 10mM Tris-HCl pH 8.5 for 10 minutes, and then separated on the magnet for 5 minutes. The supernatant containing the amplified cDNA was then withdrawn and 1μL was used for quantification with Qubit HS DNA reagents (Lifetech). An additional 1μL aliquot of the amplified sample was diluted to 3ng/μL, and then used for fragment length estimation on the Agilent BioAnalyzer using the HS cDNA kit.

Ten of the single cell samples were reverse transcribed and amplified as single cell aliquots. Ten were lysed and denatured, then pooled together and re-split to homogenize the mRNA populations in each (pool/split samples). The 10 and 100 cell pools were processed as the single cell aliquots, except they were amplified for 22 and 18 cycles each.

3.5.3 Tagmentation

Tagmentation (Illumina/Nextera) uses a transposase mixture to simultaneously fragment and tag the ends of fragmented cDNA with amplification primers. 50ng aliquots of the SMART amplified cDNA were combined with tagmentation reagents according to the manufacturers protocol. After tagmentation, the reaction was cleaned up using 1.5 volumes of QG buffer (Qiagen) and 1.8 volumes of Ampure XP SPRI beads, according to the protocol of Gertz et al. 2012.


A
Fraction of splices
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
intergenic to known exon
known exon to known exon
known exon to known exon, different genes
known exon to unknown internal
B
Fraction of splices
3'
5'
Number of cells
1
2
3
4
5
6
7
8
9
10
11
12
13
C
bulk RNA; annotated
Number splices
ψ
bulk RNA; novel
Number splices
ψ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
D
annotated; >=10 copies and 10 reads
Fraction of splices
5' ψ
E
novel; >=10 copies and 10 reads
Fraction of splices
5' ψ
annotated; >=100 copies and 100 reads
Fraction of splices
5' ψ
novel; >=100 copies and 100 reads
Fraction of splices
5' ψ
F
single cells
Fraction splices
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.7
0.8
0.9
1
max(ψ) - min(ψ)
G
pool/split #1
Fraction splices
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.7
0.8
0.9
1
max(ψ) - min(ψ)

The tagmented cDNA was eluted from the beads in 20μL of Tris-HCL pH 8.5, and subjected to an additional 5 rounds of amplification, according to the manufacturers protocol. The amplified and tagmented cDNA was cleaned up using 0.8 volumes of SPRI beads, washed twice with 200μL of 80% ethanol, dried and eluted with 30μL of Tris-HCl pH 8.5.

The tagmented libraries were quantified with Qubit HS DNA reagents, and 3ng from each sample were assayed on the Agilent BioAnalyzer using the HS cDNA kit. Libraries were judged to be acceptable if they showed a peak in the 300-400bp range. Library sequencing was performed on the HiSeq 2000 Illumina instrument, using the single read, 100 bp format.

3.5.4 Preparation of quantitation standards

The quantification spike-in standards are designed to test a range of copy number concentrations over 3 factors of 10. We chose two size ranges (∼ 300nt and ∼ 1400nt) to test the effect of transcript length on counting accuracy. The following transcripts were amplified from *Arabidopsis* total RNA for use as quantitation standards: VATG (376nt), OBF5 (1444nt), Apetala2 (1405nt), PDF (348nt), EPR (1451nt), AGP (323nt). These amplified cDNAs were cloned into a modified cloning vector containing the pBluescript II promoters and multiple cloning site, flanking an elongated polyA sequence. The resulting clones were linearized downstream of the polyA sequence, so that in vitro transcription would result in the automatic inclusion of a polyA tail, without the need for polyA polymerase. In vitro transcription was performed using the EpiCentre Ampliscribe T3 in vitro transcription kit (catalog #AS3103). The reactions were cleaned up using a Qiagen RNA cleanup column (Qiagen catalog #74124). The transcribed products were quantified using Qubit RNA reagents (3 repeated measures) and then size verified on the Agilent BioAnalyzer using RNA Nano reagents. The transcripts were then diluted in diluent containing yeast tRNA as a carrier (Ambion Catalog #AM7119) and RNAse inhibitor (Clontech catalog #2313A), and then combined into a cocktail for use as 1μL aliquots. The final concentrations for tRNA was 100pg/μL. The final concentrations of the spike-in standards are listed in Table 3.

3.5.5 *In silico* simulation of single-cell and cell pool transcriptomes

We aimed primarily to examine the effects of the levels of technical stochasticity and the amount of input, but also tried to approximate what a real population of cells might look like, with all the variation of gene expression on the single-cell level that exists in it. To this end, we used the following model.

Let $|S|$ be the number of cells pooled, and p_{E_g} be the probability that a gene g belonging to the set of polyadenylated genes G is expressed in any given cell $S_i \in S$. There likely exist a

Figure 3.51 ***(preceding page)*****: Alternative splicing at the single-cell level**. (A) Classification of new junctions connecting known splice sites. (B) Frequency of detection of novel splice junctions. Novel junctions for which neither the donor nor acceptor site has been annotated were excluded for reasons described in the main text in both (A) and (B). A threshold of 10 estimated copies and a coverage of 10 reads was applied, but results are essentially the same independent of the thresholds used (Supplementary Figure 40A). (C). Distribution of ψ scores in bulk RNA samples for annotated and novel splice junctions. A threshold of 15 reads combined for all splice junctions in which a donor or acceptor site participates was applied. Note that for each ψ_1 score there is at least one matching $\psi_2 \leq 1 - \psi_1$ score corresponding to the other alternative junction; in some cases, more than two alternative donor or acceptor sites exist, thus the relative height of the $0 \leq \psi \leq 0.1$ bar. (D - upper and lower) Distribution of 5' ψ scores for annotated splice junctions at two different detection thresholds in single-cell libraries (see Supplementary Figure 41 for more detail). (E - upper and lower) Distribution of 5' ψ scores for novel splice junctions at two different detection thresholds in single-cell libraries (see Supplementary Figure 42 for more detail). (F) and (G) Frequency of major splice site usage switches between individual cells (F) and individual libraries in a pool/split experiment (G). Note the strong support for major splice site use switching across the collection of single cells.

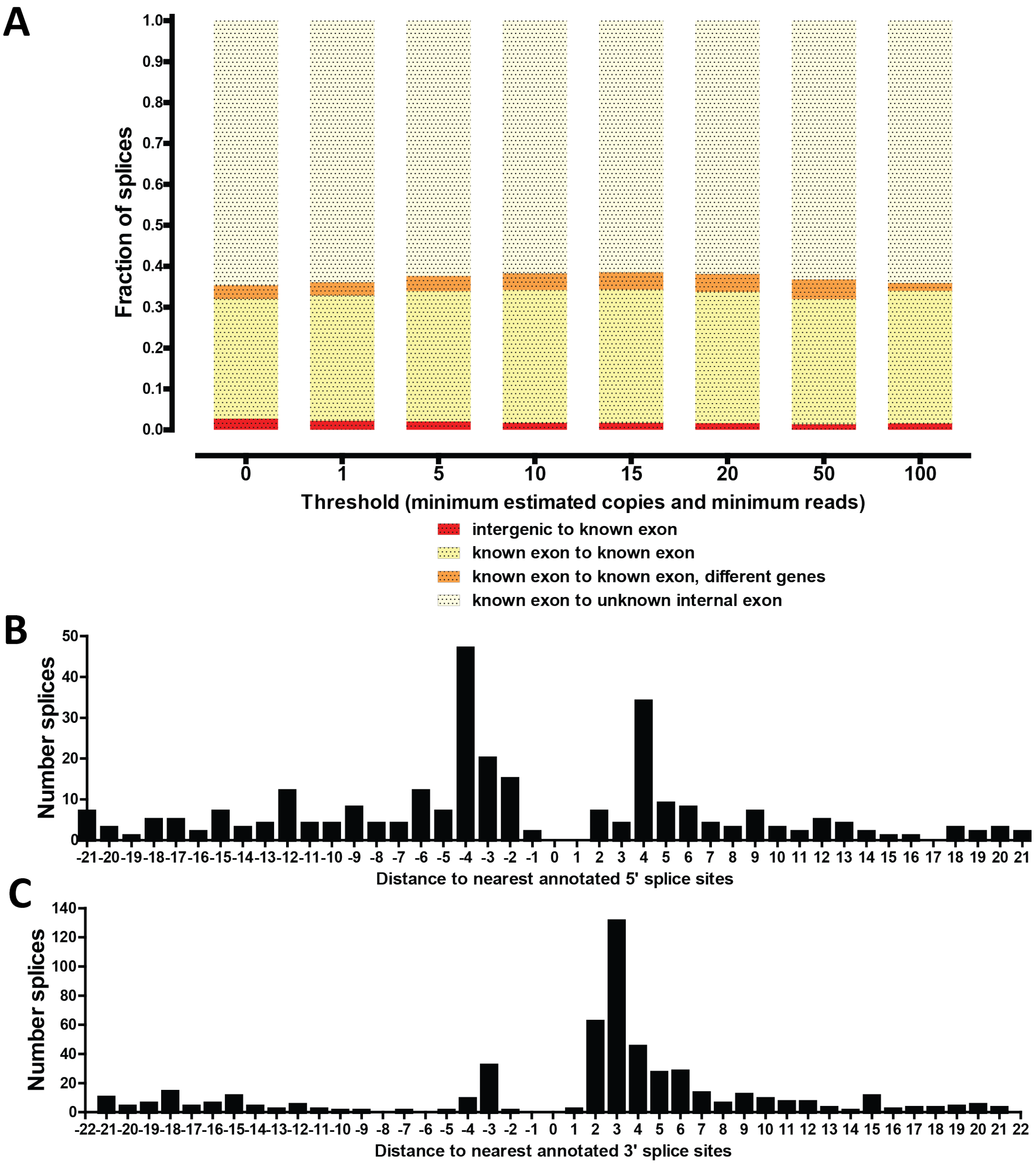


Figure 3.52: Relationship of novel splice junctions to annotation. (A) Relation to annotated exons. The detection threshold (in both estimated number of copies and reads mapping to heterozygous positions) was varied as shown and the fraction of junctions belonging to each class was calculated. (B) Distance of the donor site to the nearest annotated 5' splice site. (C) Distance of the acceptor site to the nearest annotated 3' splice sites. All detected junctions were included in (B) and (C).

group of housekeeping genes for which $p_{E_g} \approx 1$, and then there is a continuum of genes for which $p_{E_g} < 1$. Finally, there likely exist genes that are present only in a small fraction of cells for which $p_{E_g} \ll 1$. We denote with T the total number of mRNA molecules expressed in each cell S_i, with C_{C_g} the true number of transcript copies per cell for each gene $g \in G$, where G is the set

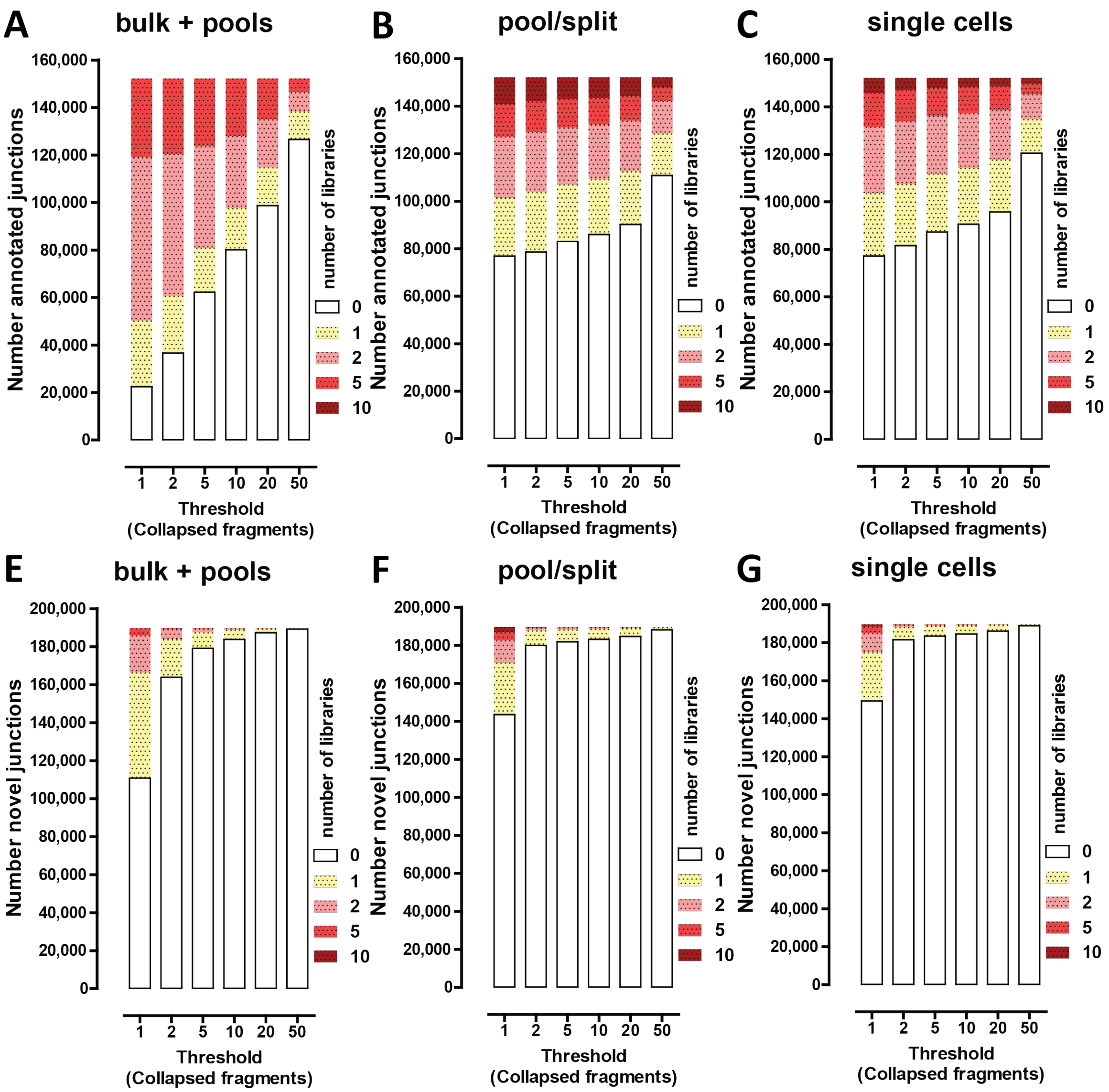


Figure 3.53: Splice junctions detection. The total number of annotated or novel junctions in all libraries is included in each plot and junctions that are not detected in each group of experiments are represented by a white bar. (A, B and C) Annotated junctions in bulk and pool libraries (A), pool/split experiments (B) and single cells (C). (D, E and F) Novel junctions in bulk and pool libraries (D), pool/split experiments (E) and single cells (F). Shown are all junctions detected in pools, pool/splits or single cells; when a junction is detected in 0 libraries, only the libraries in the indicated group are referred to.

of all genes. By definition, $T = \sum_{g \in G} C_{C_g}$. For simplicity, we assume it is constant for each cell.

We derive FPKM estimates FPKM_g for each gene based on bulk RNA-seq measurements. For simplicity, and since this does not in any way affect the conclusions of the simulations, we assume that the ratios of FPKM values between genes are equal to the ratios between the their absolute number of transcript molecules in the very large cell pool from which the library was built. We then derive an estimate for the average value of C_{C_g} when a gene is expressed in a given cell as follows:

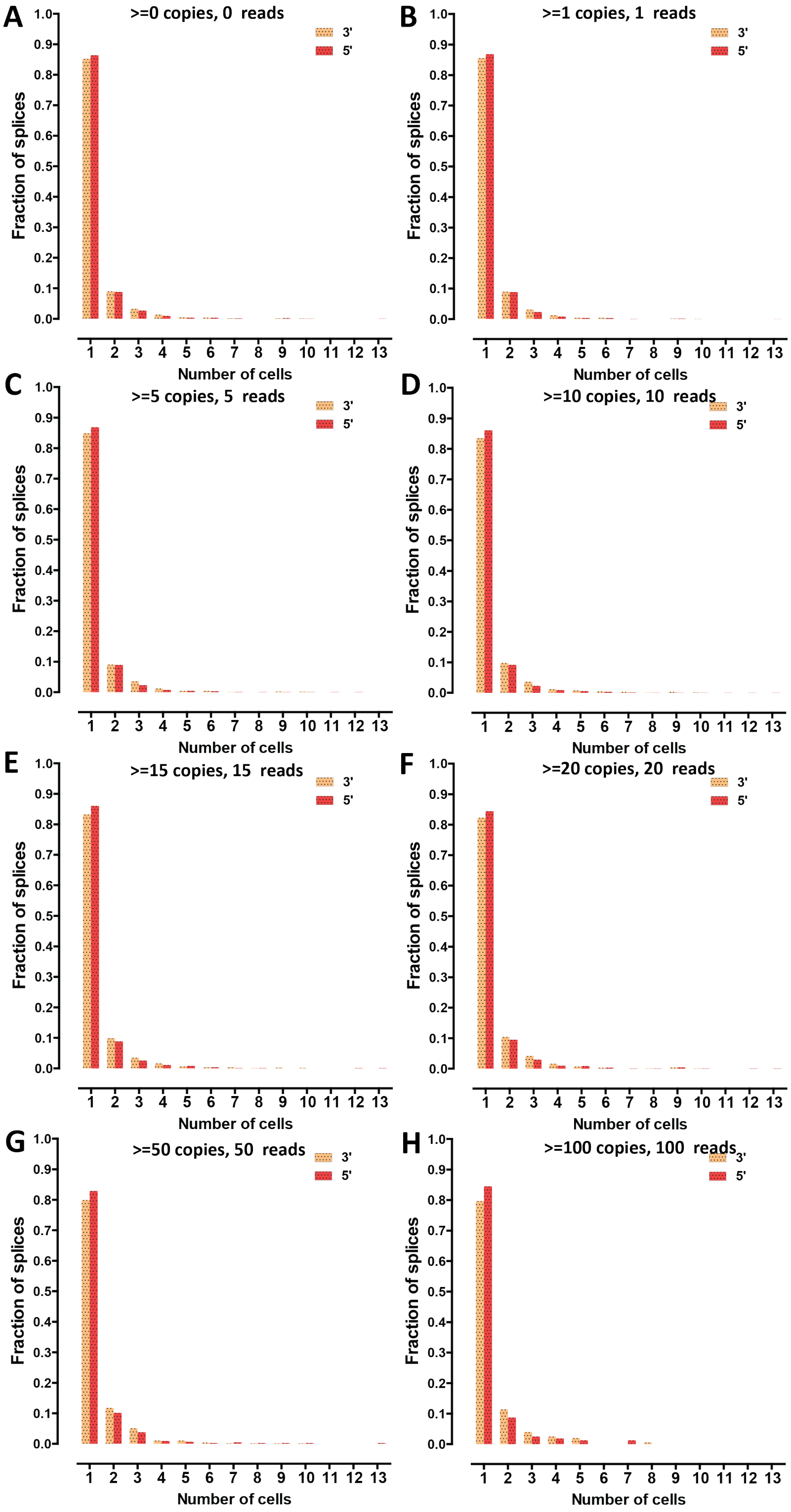


Figure 3.54: Number of cells in which a novel junctions is detected. The detection threshold (in both estimated number of copies and reads mapping to heterozygous positions) was varied as shown and the fraction of splices detected in a give number of cells plotted.

$$C_{C_g} = \frac{E_g * FPKM_g}{\sum\limits_{g\in G} E_g * FPKM_g} * T \qquad (3.1)$$

where we account for the fact that only a portion of cells express the gene by setting $E_g = 1$ when a gene is expressed in a given cell, and

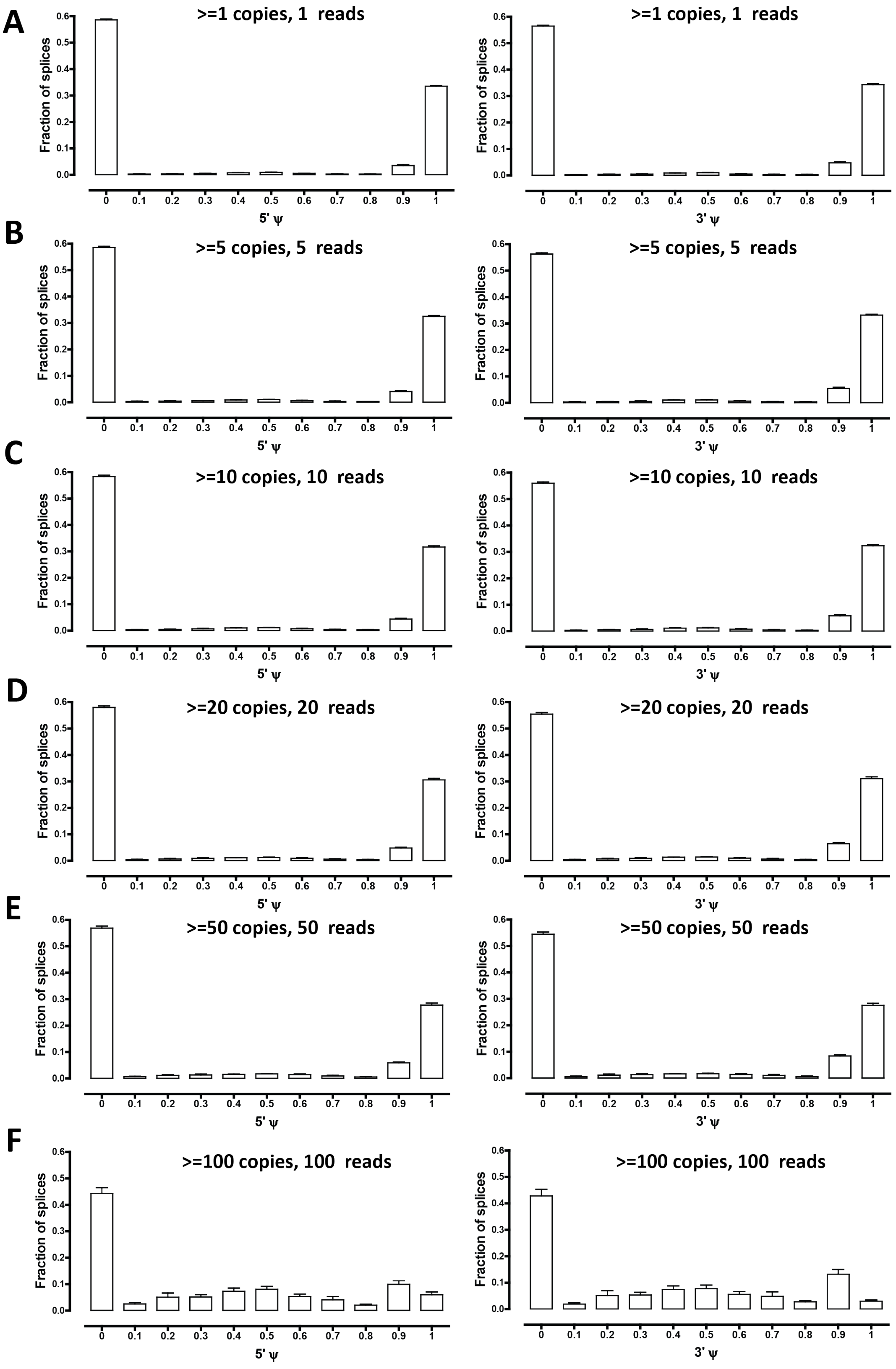


Figure 3.55: Distribution of 5' and 3' ψ scores as a function of the expression and splice junction spanning reads threshold.

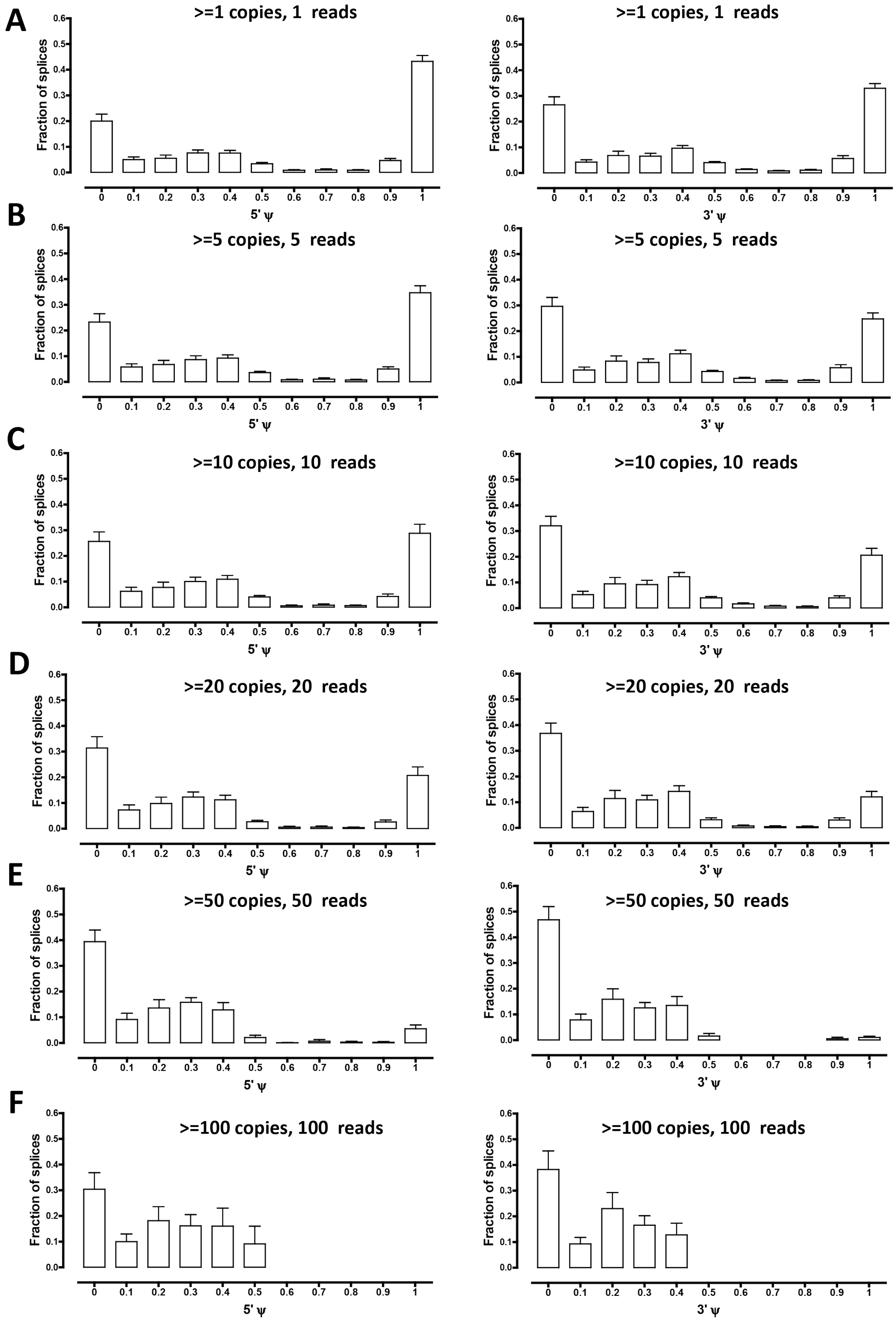


Figure 3.56: Distribution of 5' and 3' ψ scores as a function of the expression and splice junction spanning reads threshold for novel splice junctions. Only novel splice junctions connecting at least one of the donor or acceptor site for which is annotated are included.

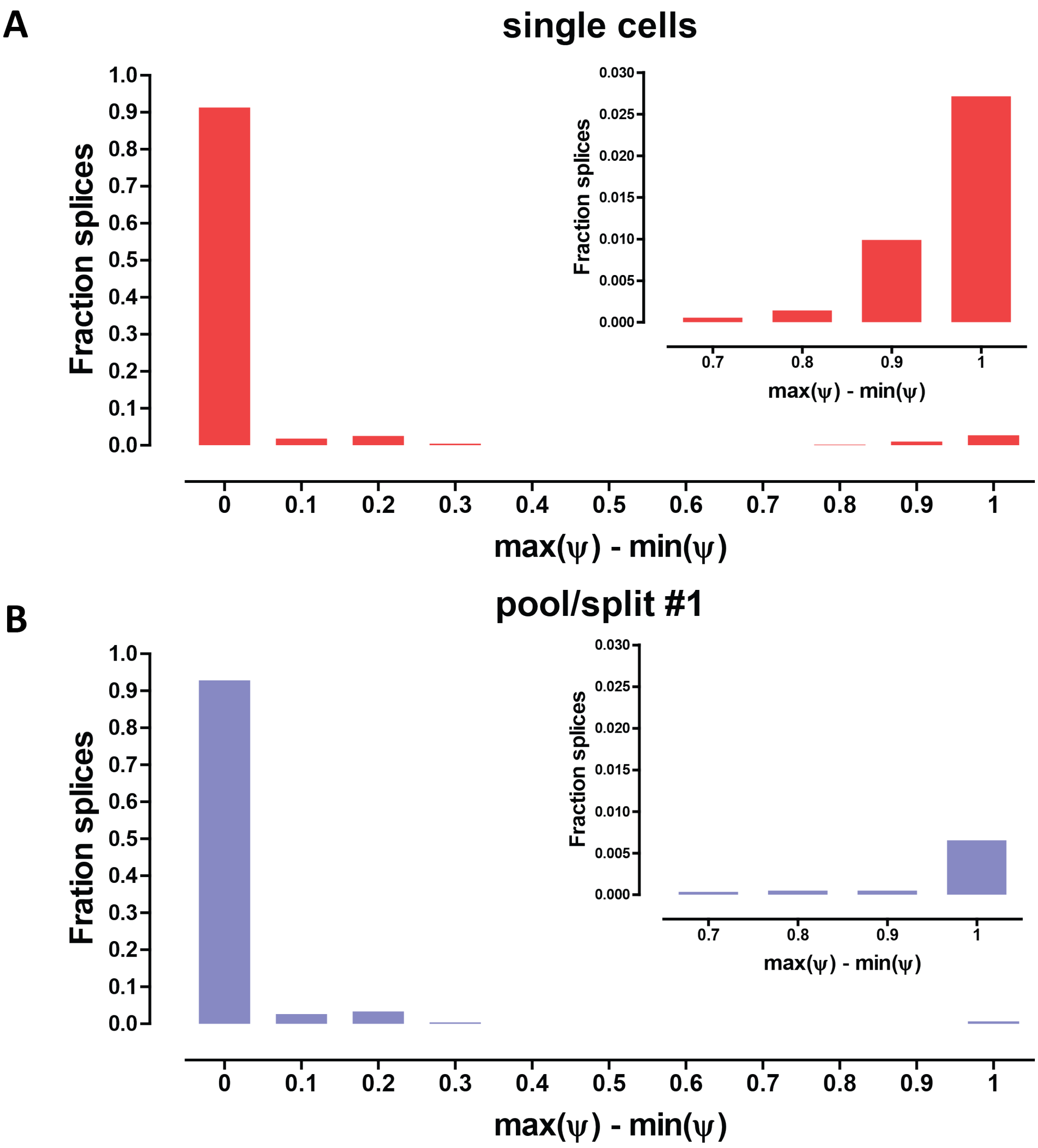


Figure 3.57: Major splice site switches between individual cells. Shown is the maximum difference between ψ scores in single-cells (A) and individual libraries in pool/split experiment 1 (B). Only gene/library pairs for which the ψ score passed all three tests for statistical significance of bias towards one splice (described in Methods) were included

$E_g = 0$ when it is not (E_g is set based on the probability p_{E_g}, as described further below).

Finally, we define the single-molecule capture efficiency p_{smc} as the probability that any given RNA molecule in a cell will be converted into cDNA, amplified and eventually present in the sequencing library.

We use the following algorithm for generating *in silico* cell pool transcriptomes and then the FPKM values in the corresponding libraries. We denote the number of original transcript copies present in the final library (after the effects of

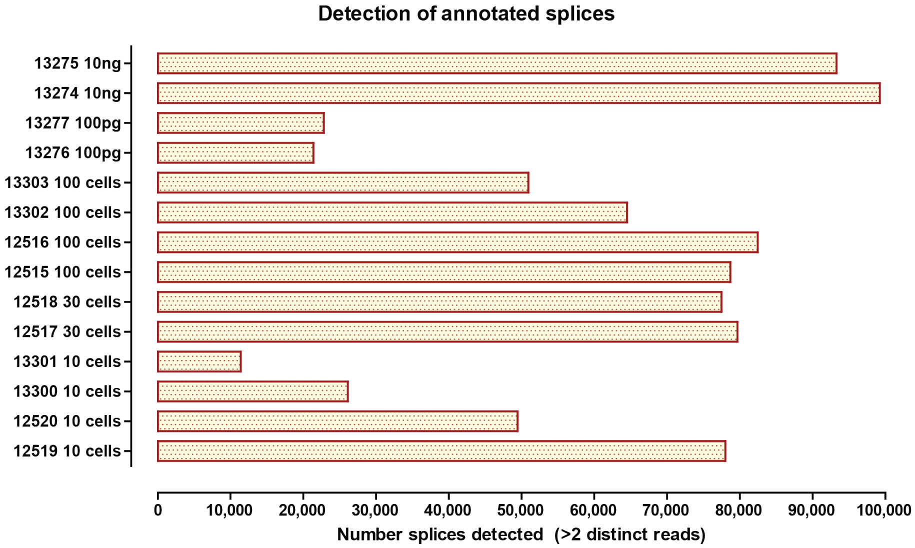


Figure 3.58: Detection of annotated splice junctions in cell pools of different sizes.

technical stochasticity have been modeled) with C_{L_g}

Algorithm 1 Cell pool RNA-seq simulation

```
for $g \in G$ do
    $C_{L_g} \leftarrow 0$
end for
for $i = 1 \rightarrow |S|$ do
    for $g \in G$ do
        $p \leftarrow$ random number $\in [0, 1]$
        if $p \leq p_{E_g}$ then
            $E_g \leftarrow 1$
        else
            $E_g \leftarrow 0$
        end if
    end for
    for $g \in G$ do
        $C_{C_g} \leftarrow \frac{E_g * FPKM_g}{\sum_{g \in G} E_g * FPKM_g} * T$
        for $i = 1 \rightarrow C_{C_g}$ do
            $p \leftarrow$ random number $\in [0, 1]$
            if $p \leq p_{smc}$ then
                $C_{L_g} = C_{L_g} + 1$
            end if
        end for
    end for
end for
for $g \in G$ do
    1. $FPKM_{L_g} \leftarrow \frac{C_{L_g}}{\sum_{g \in G} C_{L_g}} \sum_{g \in G} FPKM_g$
    2. $FPKM_{C_g} \leftarrow \frac{C_{C_g}}{\sum_{g \in G} C_{C_g}} \sum_{g \in G} FPKM_g$
    3. compare $FPKM_{L_g}$ with $FPKM_{C_g}$
end for
```

In practice, we have no reliable estimates of what the distribution of p_{E_g} might be across the whole transcriptome (this in itself is a major open research question). We assigned p_{E_g} values to genes by first splitting all genes expressed at FPKM ≥ 1 in 10 percentile groups in order of increased expression: $PG \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. For each expression percentile group PG we modeled the distribution of the p_{E_g} values with the normalized Gaussian density $\mathcal{N}(\mu, \sigma)$ over the interval $[F, 1]$ where $\mu = PG$ and $\sigma = |0.9 - PG|$, and F is the minimal fraction of cells a gene can be expressed in (which we set to 0.01).

3.5.6 Sequence alignment and gene expression quantification

Reads were aligned against a combined Bowtie (Langmead et al. 2009) index of the NCBI GRCh37 (hg19) version of the human genome (downloaded from UCSC) excluding the Y chromosome (as GM12878 cells are of female origin) and random chromosomes and the spike-in sequences using TopHat version 1.4.1 (Trapnell et al. 2009; Trapnell et al. 2012) and the GENCODE V13 annotation (Harrow et al. 2012) with the "--GTF" option. Read mapping statistics are available in Table 3.7. Gene expression was quantified using Cufflinks version 2.0.2 (Trapnell et al. 2010; Trapnell et al. 2012) for both the GENCODE V13 and refSeq annotations. FPKMs were converted to estimates for copies-per-cell numbers using spike-in sequences of known abundance (Supplementary Table 2); FPKMs were calculated for the spike-ins and used to create a calibration curve for each library (forcing the regression through 0 to avoid the assignment of positive copies per cell to genes with 0 FPKMs) on the basis of which and the Cufflinks FPKMs copies-per-cell estimates were derived for each gene.

3.5.7 Single-molecule capture efficiency estimation

We estimated the average single-molecule capture efficiency based on the number of libraries with 0 FPKM for each spike and the number of input molecules at which that spike was present in the reaction. The actual single-molecule capture efficiency need not be exactly the same for all libraries. It is a binomial process, but it is not possible to estimate it precisely by directly modeling the outcome with a binomial distribution as only the number of complete failures (libraries with 0 FPKM for a given spike, in which all C_s trials where C_s is the number of input copies for spike s) is known. The number of successes (and the corresponding exact number of failures) is not known because multiple copies of each spike are used as input, and as a result, in a library with a non-zero FPKM it is only known that some copies were successfully captured but not how many exactly. We derived an approximate estimate for the single-molecule capture efficiency by treating individual libraries as single trials in a binomial process, then dividing the estimated single-molecule capture efficiency by the number of input copies:

$$p_{smc} = \frac{1}{C_s} \arg\max_{p} \mathcal{L}(p|L_0 + L_1, L_1) \tag{3.2}$$

Where L_0 is the number of libraries with 0 FPKM and L_1 is the number of libraries with non-zero FPKM for the spike. This is a relatively crude way to estimate p_{smc} and it works well only when its value is small but in practice the p_{smc} value is indeed small.

For the AGP23 spike (spiked-in at 5 copies), the estimated single-molecule capture efficiency was 0.138 (95% confidence interval 0.106 to 0.164); for the EPR1 spike (10 copies), the estimated single-molecule capture efficiency was 0.053 (95% confidence interval 0.037 to 0.068), and for the PDF1 spike (20 copies), it was 0.045 (95% confidence interval 0.038 to 0.048). As these are approximate estimates, for simplicity we used an average single-molecule capture efficiency $p_{smc} = 0.10$ in subsequent calculations.

3.5.8 Analysis of allele-biased expression

The diploid (May 2011 release) NA12878 genome containing phased SNPs and indels based on the NCBI build 36 (hg18) version of the human genome was downloaded from http://sv.gersteinlab.org/NA12878_diploid/. Coordinates for the refSeq annotation for hg18 (downloaded from the UCSC genome browser) were converted into paternal and maternal coordinates. Heterozygous transcriptomes containing two copies of each transcript were built and reads were aligned using Bowtie (Langmead et al. 2009) (version 0.12.7) with the following settings: "-v 0 -a --best --strata", i.e. with no mismatches allowed. Reads were assigned to an allele if they mapped only to one of the alleles of a gene. All identical reads were collapsed into a single count in order to eliminate PCR amplification artifacts. Allele-biased expression was assessed as follows. First, for each gene using the total number of allele-specific reads for each allele (over all heterozygous positions), a binomial test with a uniform read distribution expectation, a 0.05 p-value cutoff, and a Bonferroni multiple-hypothesis testing correction where the correction factor is the number of genes with sufficiently many allele-specific reads for the binomial test to pass the specified p-value in the case of complete dominance of one of the alleles. Second, the number of copies for each gene was used to derive an estimate for the absolute number of copies per cell for each allele, i.e., for alleles A and a and a per-cell copies estimate for the gene C_E:

$$C_{E_A} = \frac{N_{reads}(A)}{N_{reads}(A) + N_{reads}(a)} C_E \tag{3.3}$$

Another binomial test similar to the one described above was then run using the C_{E_A} and

C_{E_a} estimates. As it is possible that only a small number of reads map differentially to the two alleles of a gene (due, for example, to heterozygous positions being located in a region of poor sequencing coverage) while the gene itself is expressed highly, thus resulting in a significant binomial test using the copies-per-cell estimates that is, however, poorly supported on the read level, both tests were required to pass statistical significance for an allele bias call to be made.

Finally, due to the imperfect single-molecule capture efficiency of the single-cell RNA-seq library building process, it is possible that apparent allele biases are the result of purely stochastic differences between the capture efficiency for the two alleles in a given library. For this reason, we applied a third filter for allele-biased expression calls, which required that the probability of obtaining apparently statistically significant differences in the estimated copies per cell for the two alleles C_{E_A} and C_{E_a} by chance from two independent binomial process with the estimated single-molecule capture efficiency p_{smc} is low ($p \leq 0.05$ after applying Bonferroni multiple hypothesis testing correction):

$$p = \sum_{C_E}^{C_C} \frac{NB(C_C - C_E, p_{smc})}{\sum_{C_E}^{C_C} NB(C_C - C_E, p_{smc})} \sum_{0}^{C_{E_a}} B(C_{C_a}, p_{smc}) \sum_{C_{E_A}}^{C_A} B(C_{C_A}, p_{smc}) \qquad (3.4)$$

Where $C_C = 2{*}C_{C_a} = 2{*}C_{C_A}$ refer to the actual number of copies per cell (as opposed to the estimated number of copies $C_E = C_{E_a} + C_{E_A}$), $NB(C_C - C_E, p_{smc})$ refers to the negative binomial probability that the actual number of copies is C_C given the estimated number of copies C_E:

$$NB(C_C - C_E, p_{smc}) = \binom{C_E + (C_C - C_E) - 1}{C_E - 1} p_c^{C_E} (1 - p_{smc})^{C_C - C_E}$$

and the binomial probabilities $B(C_{C_A}, p_{smc})$ and $B(C_{C_a}, p_c)$ are defined as:

$$B(C_{C_A}, p_{smc}) = \binom{C_{C_A}}{C_{E_A}} {p_{smc}}^{C_{E_A}} (1{-}p_{smc})^{C_{C_A} - C_{E_A}}$$

and

$$B(C_{C_a}, p_{smc}) = \binom{C_{C_a}}{C_{E_a}} {p_{smc}}^{C_{E_a}} (1{-}p_{smc})^{C_{C_a} - C_{E_a}}$$

The probability was evaluated for possible values of the actual number of copies per cell up to $C_C = min(5000, 100 * C_E)$.

Genes on the X chromosome were excluded from all analysis as the GM18278 cell line is female. The inactivation of the X chromosome leading to a corresponding allelic exclusion was observed as expected (data not shown).

3.5.9 Alternative splicing analysis

We mapped reads using TopHat with *de novo* junction discovery turned on; such alignments are in principle suited for the discovery and analysis of novel splice junctions, a large number of which has been recently reported by the ENCODE consortium (Djebali & Davis et al. 2012). An important step in such analysis is distinguishing between true novel splice junctions on one hand and mapping and library-building artifacts on the other. Such artifacts certainly exist as we observe "novel junctions" in our spike-in quantification standards, which are not spliced (Table). Confidence in the reality of newly discovered splice junctions in traditional RNA-seq is boosted by the number of distinct sequencing fragments supporting them, and by replication in other libraries. However, the former line of evidence is not applicable to single-cell RNA-seq due to the one-to-many relationship between original founder RNA molecules and sequencing fragments in the final library, while the latter is difficult to apply in all cases given the uniqueness of each individual single cell. For these reasons, we restricted alternative splicing analysis to known splice junctions and novel junctions, at least one end of which was annotated as splice site in GENCODE V13.

We calculated 5' and 3' splicing inclusion ψ scores as follows (Pervouchine et al. 2013):

$$\psi_5(D,A) = \frac{N_{reads}(D,A)}{\sum\limits_{A_i \in A} N_{reads}(D,A_i)} \tag{3.5}$$

$$\psi_3(D,A) = \frac{N_{reads}(D,A)}{\sum\limits_{D_i \in D} N_{reads}(D_i,A)} \tag{3.6}$$

Where D and A refer to the donor and acceptor splice sites, respectively, and the number of reads N_{reads} refers to the number of spliced reads crossing a splice or donor sites after apparent PCR duplicates have been collapsed into a single count. We note that any given donor or acceptor splice site need not be included in all transcripts expressed from the gene it belongs to. Since isoform-level quantification is not a completely solved problem and it is even less clear what its relative stability is for single-cell RNA-seq compared to the bulk RNA datasets for which algorithms have been designed, we only included alternative splice sites for which the donor or acceptor site was found in all annotated transcripts for the gene (GENCODE V13 annotation) as well as novel junctions (compared to the GENCODE V13 annotation) derived from the TopHat mappings involving such splice sites. This allows us to use gene-level FPKM estimates, which are in general more reliable than isoform-level ones, and the mRNA copies-per-cell estimates based on those to derive the approximate absolute number of transcripts containing a given splice junction in each. The statistical significance of bias towards one of the sites was assessed analogously to the approach described for allele-biased expression above, with one significant modification: in cases of more than two possible A_i acceptor sites, for a donor site D or D_i sites for an acceptor site A, the major pair (the one with the most reads) was compared to the sum of reads for all other pairs as if those pairs constituted as single pair. This approach was adopted so that a maximum number of alternative splicing events are included in the analysis and with a focus on identifying cases of robust and statistically significant splice site use switches between individual single cells. When the major (D,A) pair did not have more than half of all reads, the site was excluded from further analysis.

3.5.10 Gene expression clustering and weighted correlation network analysis

Weighted correlation networks (Zhang & Horvath 2005) were constructed from the single-cell vectors of estimated mRNA copies using the WGCNA R package (Langfelder & Horvath 2008) using the `blockwiseModules` function with $\beta = 6$ (Supplementary Figure 34) and a minimum module size of 10 genes. Input genes were filtered as follows: first, we required that genes be expressed at more than one estimated copy per cell C_E in at least one cell; second, we required that the ratio between the C_E variance in single cells and the C_E variance in pool/split libraries be more than 1.5. The latter requirement was imposed in order to minimize the identification of apparently coexpressed gene modules due to purely stochastic differences in transcript capture (see Supplementary Figure 33 for more detail).

Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al. 2009). Gene expression clustering was carried out using Cluster 3.0 (de Hoon et al. 2004) and visualized using TreeView (Saldanha 2004).

Table 3.3: Full list of Gene Ontology categories enriched in coexpressed gene modules. Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al., 2009).

N	X	LOD	P	P adj	attrib ID	attrib name
Module 1						
80	85	2.22750847233495	2.42885421570 57e-80	<0.001	GO:0019083	viral transcription
35	38	2.0543096800611	9.94569501487657e-35	<0.001	GO:0022625	cytosolic large ribosomal subunit
79	88	1.98429369667081	3.81815501340532e-75	<0.001	GO:0006415	translational termination
82	92	1.95783499865395	2.25183855262354e-77	<0.001	GO:0019058	viral infectious cycle
33	37	1.91934326185776	1.0606448232178e-31	<0.001	GO:0022627	cytosolic small ribosomal subunit
84	99	1.79959130536831	2.83136121010728e-75	<0.001	GO:0006414	translational elongation
79	98	1.67170885189082	2.54715717317477e-67	<0.001	GO:0043624	cellular protein complex disassembly
79	99	1.64996220882795	1.16143543924047e-66	<0.001	GO:0043241	protein complex disassembly
5	6	1.60339903304627	2.31518111579124e-05	0.042	GO:0042719	mitochondrial intermembrane space protein transporter complex
12	16	1.48484395467934	1.55655372018477e-10	<0.001	GO:0005753	mitochondrial proton-transporting ATP synthase complex
12	16	1.48484395467934	1.55655372018477e-10	<0.001	GO:0045259	proton-transporting ATP synthase complex
8	11	1.42532291837978	3.17730434970578e-07	<0.001	GO:0042274	ribosomal small subunit biogenesis
42	60	1.41110133078985	7.7723344218279e-32	<0.001	GO:0015935	small ribosomal subunit
10	14	1.40852193295129	1.23601718399618e-08	<0.001	GO:0042776	mitochondrial ATP synthesis coupled proton transport
82	119	1.40425322218248	2.90687191098935e-60	<0.001	GO:0034623	cellular macromolecular complex disassembly
43	62	1.39861873942686	2.70656133560673e-32	<0.001	GO:0015934	large ribosomal subunit
82	120	1.39279633589538	8.45636479922126e-60	<0.001	GO:0032984	macromolecular complex disassembly
14	20	1.39014850585389	1.95202402281072e-11	<0.001	GO:0015985	energy coupled proton transport, down electrochemical gradient
14	20	1.39014850585389	1.95202402281072e-11	<0.001	GO:0015986	ATP synthesis coupled proton transport
9	13	1.36475589719287	1.06059816949165e-07	<0.001	GO:0045263	proton-transporting ATP synthase complex, coupling factor F(o)
102	154	1.35835395745315	3.83509033779718e-72	<0.001	GO:0003735	structural constituent of ribosome
80	121	1.34883668562943	1.30325679955416e-56	<0.001	GO:0031018	endocrine pancreas development
6	9	1.3082177020168	2.3318724760727e-05	0.049	GO:0042273	ribosomal large subunit biogenesis
92	156	1.22084501073779	1.79769896957827e-58	<0.001	GO:0022415	viral reproductive process
9	15	1.20500031764798	6.35922215735707e-07	<0.001	GO:0042613	MHC class II protein complex
90	159	1.17815283094918	4.63835969773452e-55	<0.001	GO:0005840	ribosome
11	20	1.12368307791708	1.16201671228436e-07	<0.001	GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
126	243	1.1058240710311	8.64889403866865e-71	<0.001	GO:0006412	translation
33	66	1.04671661076047	1.04054778845837e-18	<0.001	GO:0070469	respiratory chain
88	180	1.04302248261017	8.36713797967258e-47	<0.001	GO:0071845	cellular component disassembly at cellular level
13	26	1.04120105244121	3.59500355461033e-08	<0.001	GO:0004129	cytochrome-c oxidase activity
13	26	1.04120105244121	3.59500355461033e-08	<0.001	GO:0015002	heme-copper terminal oxidase activity
13	26	1.04120105244121	3.59500355461033e-08	<0.001	GO:0016676	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor
88	181	1.03832508439038	1.50023036663094e-46	<0.001	GO:0022411	cellular component disassembly
50	105	1.01033284389759	2.12881566827084e-26	<0.001	GO:0022904	respiratory electron transport chain
13	27	1.01013943569481	6.3992281019869e-08	<0.001	GO:0016675	oxidoreductase activity, acting on a heme group of donors
13	27	1.01013943569481	6.3992281019869e-08	<0.001	GO:0019843	rRNA binding
22	46	1.00663430266062	2.01941178786577e-12	<0.001	GO:0005747	mitochondrial respiratory chain complex I
22	46	1.00663430266062	2.01941178786577e-12	<0.001	GO:0030964	NADH dehydrogenase complex
22	46	1.00663430266062	2.01941178786577e-12	<0.001	GO:0045271	respiratory chain complex I
13	28	0.981148355026407	1.10254306563049e-07	<0.001	GO:0016469	proton-transporting two-sector ATPase complex
20	44	0.965600146054987	6.60764042765157e-11	<0.001	GO:0003954	NADH dehydrogenase activity
20	44	0.965600146054987	6.60764042765157e-11	<0.001	GO:0008137	NADH dehydrogenase (ubiquinone) activity
20	44	0.965600146054987	6.60764042765157e-11	<0.001	GO:0050136	NADH dehydrogenase (quinone) activity
10	22	0.964605445604187	4.26260478570593e-06	0.009	GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain
19	43	0.94357846286631	3.63592920641863e-10	<0.001	GO:0006120	mitochondrial electron transport, NADH to ubiquinone
56	140	0.877587213657932	1.12610252541087e-24	<0.001	GO:0022900	electron transport chain
20	50	0.870301953062918	1.08177287181172e-09	<0.001	GO:0016655	oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor
126	329	0.864920624044305	7.76265565058296e-52	<0.001	GO:0016032	viral reproduction
19	48	0.862785503901917	3.4843140169824e-09	<0.001	GO:0022613	ribonucleoprotein complex biogenesis
51	133	0.846254921603025	1.32115278867509e-21	<0.001	GO:0044455	mitochondrial membrane part
13	36	0.80019305087457	3.59382041803685e-06	0.008	GO:0042611	MHC protein complex

Continued on next page

Table 3.3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072401	signal transduction involved in DNA integrity checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072413	signal transduction involved in mitotic cell cycle checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072422	signal transduction involved in DNA damage checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072431	signal transduction involved in mitotic cell cycle G1/S transition DNA damage checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072474	signal transduction involved in mitotic cell cycle G1/S checkpoint
12	34	0.785381071713187	1.12064061460209e-05	0.019	GO:0006626	protein targeting to mitochondrion
31	88	0.784090641205952	1.6208634459227e-12	<0.001	GO:0048002	antigen processing and presentation of peptide antigen
138	408	0.782753905363391	3.18149793786453e-49	<0.001	GO:0010467	gene expression
22	63	0.777286220816123	3.33709849401727e-09	<0.001	GO:0072395	signal transduction involved in cell cycle checkpoint
22	63	0.777286220816123	3.33709849401727e-09	<0.001	GO:0072404	signal transduction involved in G1/S transition checkpoint
19	55	0.770122744469575	4.78383728527817e-08	<0.001	GO:0071843	cellular component biogenesis at cellular level
173	530	0.768759824961379	2.97828415371199e-59	<0.001	GO:0030529	ribonucleoprotein complex
28	82	0.763064993637299	4.54473457242155e-11	<0.001	GO:0002474	antigen processing and presentation of peptide antigen via MHC class I
25	75	0.747063491558582	8.85423589248866e-10	<0.001	GO:0006839	mitochondrial transport
23	69	0.746932600935863	4.15847287902749e-09	<0.001	GO:0044085	cellular component biogenesis
32	97	0.741175253071487	5.45910901839781e-12	<0.001	GO:0015078	hydrogen ion transmembrane transporter activity
16	49	0.73399635379054	1.30379324163326e-06	0.003	GO:0051258	protein polymerization
15	47	0.719731376252704	3.8551671841249e-06	0.008	GO:0051084	'de novo' posttranslational protein folding
40	126	0.71784240561212	5.25251868522275e-14	<0.001	GO:0019882	antigen processing and presentation
23	73	0.710984404898217	1.41497713185833e-08	<0.001	GO:0071158	positive regulation of cell cycle arrest
21	67	0.707697276637143	6.62504934644734e-08	<0.001	GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
95	311	0.705569345151931	4.85030590692832e-30	<0.001	GO:0048610	cellular process involved in reproduction
16	52	0.696666038434829	3.16984759196741e-06	0.007	GO:0006458	'de novo' protein folding
22	72	0.691795994094768	5.42170060581127e-08	<0.001	GO:0051352	negative regulation of ligase activity
22	72	0.691795994094768	5.42170060581127e-08	<0.001	GO:0051444	negative regulation of ubiquitin-protein ligase activity
150	525	0.677439350068986	3.32539653823493e-43	<0.001	GO:0034621	cellular macromolecular complex subunit organization
164	584	0.670332348124673	3.54836301881911e-46	<0.001	GO:0016071	mRNA metabolic process
23	78	0.669845547354692	5.69629303883152e-08	<0.001	GO:0000216	M/G1 transition of mitotic cell cycle
21	72	0.663205782638981	2.6219034657749e-07	<0.001	GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
16	55	0.662279533745671	7.14593997684204e-06	0.014	GO:0000502	proteasome complex
16	55	0.662279533745671	7.14593997684204e-06	0.014	GO:0006521	regulation of cellular amino acid metabolic process
25	86	0.66117774103715	2.09408969093921e-08	<0.001	GO:0030330	DNA damage response, signal transduction by p53 class mediator
19	67	0.64634468268949	1.55090946941351e-06	0.003	GO:0033238	regulation of cellular amine metabolic process
26	92	0.644103259424952	2.09762058586999e-08	<0.001	GO:0072331	signal transduction by p53 class mediator
22	78	0.642874216140607	2.62393000307691e-07	<0.001	GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
19	68	0.637453778789019	1.9842690458222e-06	0.005	GO:0060333	interferon-gamma-mediated signaling pathway
22	79	0.635227361820706	3.35087879773373e-07	<0.001	GO:0051443	positive regulation of ubiquitin-protein ligase activity
23	83	0.63224585989498	2.00949230376557e-07	<0.001	GO:0051351	positive regulation of ligase activity
27	98	0.62887220631022	2.04834672373096e-08	<0.001	GO:0006364	rRNA processing
22	80	0.627711881242392	4.25874069739404e-07	<0.001	GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
81	300	0.62621754092227	6.64252358742553e-22	<0.001	GO:0005743	mitochondrial inner membrane
17	63	0.617087551763504	1.13566498370941e-05	0.021	GO:0006200	ATP catabolic process
24	89	0.616024095582501	1.924740199118e-07	<0.001	GO:0031397	negative regulation of protein ubiquitination
23	86	0.61114511085871	4.04971652012907e-07	<0.001	GO:0016651	oxidoreductase activity, acting on NADH or NADPH
27	103	0.599379377993909	6.42976552229558e-08	<0.001	GO:0016072	rRNA metabolic process
84	327	0.597134802233134	3.86637763010753e-21	<0.001	GO:0019866	organelle inner membrane

Continued on next page

Table 3.3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
24	93	0.590170688226641	4.68334780721795e-07	<0.001	GO:0051438	regulation of ubiquitin-protein ligase activity
25	97	0.589412627207871	2.79036812385036e-07	<0.001	GO:0051340	regulation of ligase activity
28	109	0.587562019917053	5.91212587104247e-08	<0.001	GO:0000084	S phase of mitotic cell cycle
20	79	0.579289164068354	5.71262604161276e-06	0.012	GO:0071346	cellular response to interferon-gamma
141	582	0.575501427143487	5.37209201940123e-32	<0.001	GO:0071822	protein complex subunit organization
35	140	0.572325037803156	2.96216165946153e-09	<0.001	GO:0000082	G1/S transition of mitotic cell cycle
27	110	0.561161855618649	2.76881959100789e-07	<0.001	GO:0042770	signal transduction in response to DNA damage
28	116	0.551583814959451	2.4384223848345e-07	<0.001	GO:0051320	S phase
19	79	0.550001609297223	2.1441603106869e-05	0.031	GO:0009206	purine ribonucleoside triphosphate biosynthetic process
100	434	0.537107458537325	3.65671802133964e-21	<0.001	GO:0031966	mitochondrial membrane
31	132	0.536082594042894	1.10062427133423e-07	<0.001	GO:0090068	positive regulation of cell cycle process
22	95	0.528168311977894	9.49580519250266e-06	0.019	GO:0034341	response to interferon-gamma
84	371	0.523771651861378	1.85277099409276e-17	<0.001	GO:0006091	generation of precursor metabolites and energy
130	595	0.513071456988699	5.70494731538591e-25	<0.001	GO:0005198	structural molecule activity
26	116	0.50961644963599	2.91400334562788e-06	0.007	GO:0031398	positive regulation of protein ubiquitination
148	687	0.508962926945213	1.09821996775926e-27	<0.001	GO:0044429	mitochondrial part
37	168	0.500349876744172	4.12597162657402e-08	<0.001	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
37	168	0.500349876744172	4.12597162657402e-08	<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
36	164	0.498410796185536	6.84854979223404e-08	<0.001	GO:0043623	cellular protein complex assembly
38	175	0.492543109861863	4.05723893001672e-08	<0.001	GO:0000375	RNA splicing, via transesterification reactions
27	126	0.484584970787029	4.69193568014889e-06	0.009	GO:0046034	ATP metabolic process
175	867	0.477792444174725	4.31717391375335e-29	<0.001	GO:0043933	macromolecular complex subunit organization
32	157	0.457121358455519	2.03558979738233e-06	0.005	GO:0000209	protein polyubiquitination
57	282	0.455331541986413	3.35174043037384e-10	<0.001	GO:0008380	RNA splicing
87	437	0.45206487497538	1.75634248577746e-14	<0.001	GO:0055114	oxidation-reduction process
31	154	0.450275487651611	3.84142188009438e-06	0.008	GO:0031396	regulation of protein ubiquitination
32	159	0.450199742241769	2.70953251758628e-06	0.006	GO:0043161	proteasomal ubiquitin-dependent protein catabolic process
155	803	0.44828162002615	1.17127238506759e-23	<0.001	GO:0003723	RNA binding
32	160	0.446779218086837	3.11868915151585e-06	0.007	GO:0010498	proteasomal protein catabolic process
27	135	0.44673045980076	1.79130278352513e-05	0.028	GO:0005774	vacuolar membrane
239	1315	0.43236771896813	1.93513347789966e-32	<0.001	GO:0005739	mitochondrion
39	201	0.43074480120704	6.14144121470793e-07	<0.001	GO:0000075	cell cycle checkpoint
43	226	0.420501515091121	2.92681541907647e-07	<0.001	GO:0071156	regulation of cell cycle arrest
30	158	0.418585946292244	1.82741185723924e-05	0.028	GO:0044437	vacuolar part
57	304	0.414291886997772	6.22443660977404e-09	<0.001	GO:0000278	mitotic cell cycle
73	401	0.400691520310638	1.88837546257743e-10	<0.001	GO:0034622	cellular macromolecular complex assembly
222	1296	0.39523148469997	2.51837746792813e-26	<0.001	GO:0034645	cellular macromolecule biosynthetic process
226	1327	0.392809170439032	1.86450696466837e-26	<0.001	GO:0009059	macromolecule biosynthetic process
334	2084	0.38137706335828	1.33257319443321e-34	<0.001	GO:0044267	cellular protein metabolic process
334	2111	0.373898530461539	1.84879744335921e-33	<0.001	GO:0005829	cytosol
161	961	0.370923240921755	4.01275183856543e-18	<0.001	GO:0032774	RNA biosynthetic process
504	3431	0.370107849871159	4.0461986238786e-44	<0.001	GO:0032991	macromolecular complex
101	602	0.361160597240294	9.47442144063175e-12	<0.001	GO:0006396	RNA processing
766	6058	0.346864258263512	2.31238377684638e-47	<0.001	GO:0044444	cytoplasmic part
58	357	0.337821335867861	7.933880428251e-07	0.001	GO:0010564	regulation of cell cycle process
59	364	0.336711582120088	6.91212093849837e-07	0.001	GO:0006397	mRNA processing
144	911	0.336185023040326	3.85435400725073e-14	<0.001	GO:0005654	nucleoplasm
251	1662	0.330293674534452	1.10940941066799e-21	<0.001	GO:0016070	RNA metabolic process
313	2121	0.329217998390912	1.18114081249402e-25	<0.001	GO:0044249	cellular biosynthetic process
316	2160	0.324822918675263	2.91009865940903e-25	<0.001	GO:0090304	nucleic acid metabolic process
320	2232	0.313971532133542	4.41002855691382e-24	<0.001	GO:0009058	biosynthetic process
1145	11330	0.313948898323752	4.59005189412599e-31	<0.001	GO:0044424	intracellular part
75	485	0.313905177878569	1.5748417548981e-07	<0.001	GO:0022403	cell cycle phase
377	2686	0.312982629840953	4.90694955531477e-27	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
686	5522	0.308771441881537	4.1087928647367e-37	<0.001	GO:0044446	intracellular organelle part
367	2644	0.304397651421782	2.77819227696323e-25	<0.001	GO:0019538	protein metabolic process
529	4046	0.302681908190103	8.82320727696583e-32	<0.001	GO:0044260	cellular macromolecule metabolic process
689	5601	0.302589536132723	8.22904448857063e-36	<0.001	GO:0044422	organelle part
712	5867	0.299367116917043	1.88823085526176e-35	<0.001	GO:0044237	cellular metabolic process
135	945	0.280373762273502	3.82813590337554e-10	<0.001	GO:0022414	reproductive process
74	511	0.279376163393885	2.44628059184637e-06	0.006	GO:0006974	response to DNA damage stimulus
969	9102	0.277956512781225	5.11225805494553e-30	<0.001	GO:0043229	intracellular organelle
969	9117	0.276313895789093	1.11162671743124e-29	<0.001	GO:0043226	organelle
403	3093	0.273819161695768	1.68049396494688e-22	<0.001	GO:0034641	cellular nitrogen compound metabolic process

Continued on next page

Table 3.3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
108	765	0.270319977503456	4.55991397471596e-08	<0.001	GO:0022402	cell cycle process
99	701	0.269327927902864	1.66323414881123e-07	<0.001	GO:0007049	cell cycle
410	3199	0.265219890086891	1.5194034851802e-21	<0.001	GO:0006807	nitrogen compound metabolic process
556	4601	0.25742146969207	1.8068455687858e-24	<0.001	GO:0043170	macromolecule metabolic process
285	2171	0.257317984548616	7.98160838856529e-16	<0.001	GO:0043228	non-membrane-bounded organelle
285	2171	0.257317984548616	7.98160838856529e-16	<0.001	GO:0043232	intracellular non-membrane-bounded organelle
69	497	0.256916076214485	2.21805085904211e-05	0.031	GO:0005730	nucleolus
85	622	0.2504348652183	4.78113517987455e-06	0.01	GO:0006259	DNA metabolic process
744	6632	0.249054934285073	1.58723972093304e-25	<0.001	GO:0008152	metabolic process
293	2272	0.248502561040477	3.01319156248601e-15	<0.001	GO:0071842	cellular component organization at cellular level
295	2299	0.245967084736522	4.57315463821616e-15	<0.001	GO:0071841	cellular component organization or biogenesis at cellular level
98	729	0.243379241039335	1.87437352876668e-06	0.004	GO:0065003	macromolecular complex assembly
93	694	0.241149602052021	3.93342779914606e-06	0.008	GO:0016491	oxidoreductase activity
274	2153	0.238023679234234	1.58478501985632e-13	<0.001	GO:0044428	nuclear part
871	8242	0.235473646680484	5.82916610504252e-23	<0.001	GO:0043227	membrane-bounded organelle
870	8238	0.234562637103022	8.38290144947865e-23	<0.001	GO:0043231	intracellular membrane-bounded organelle
99	751	0.233280218653404	3.94595885765465e-06	0.008	GO:0046907	intracellular transport
665	5992	0.221831084147262	3.46559295245285e-20	<0.001	GO:0044238	primary metabolic process
102	803	0.214295076083374	1.41199427931906e-05	0.022	GO:0033554	cellular response to stress
99	783	0.211573308398039	2.28182618638875e-05	0.031	GO:0071844	cellular component assembly at cellular level
344	2896	0.208867137566652	8.16942218480433e-13	<0.001	GO:0016043	cellular component organization
346	2923	0.20711459632924	1.09506306063704e-12	<0.001	GO:0071840	cellular component organization or biogenesis
329	2821	0.195993854311548	3.24914520945555e-11	<0.001	GO:0043234	protein complex
237	2006	0.191345208630818	1.2625912051491e-08	<0.001	GO:0031090	organelle membrane
1026	10840	0.168964395995212	1.74282758664301e-11	<0.001	GO:0009987	cellular process
485	4933	0.108855314264424	1.27790612421947e-05	0.021	GO:0005737	cytoplasm
Module 2						
5	9	1.4353697859032	1.61126889116162e-05	0.024	GO:0008139	nuclear localization sequence binding
7	20	1.09388232798139	1.28535977541801e-05	0.018	GO:0051983	regulation of chromosome segregation
29	168	0.684180963922458	1.54004341495467e-10	<0.001	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
29	168	0.684180963922458	1.54004341495467e-10	<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
30	175	0.680816140112517	8.87046264555359e-11	<0.001	GO:0000375	RNA splicing, via transesterification reactions
38	282	0.558803234506848	4.52064927189642e-10	<0.001	GO:0008380	RNA splicing
19	141	0.555278310629005	1.0380427172764e-05	0.017	GO:0005681	spliceosomal complex
44	364	0.505813785001266	6.61426439925337e-10	<0.001	GO:0006397	mRNA processing
66	602	0.464516339086547	2.77728663822659e-12	<0.001	GO:0006396	RNA processing
32	288	0.460440562314741	9.75520192342185e-07	0.002	GO:0051301	cell division
62	584	0.446895852037615	5.12932632405468e-11	<0.001	GO:0016071	mRNA metabolic process
52	497	0.436204344115675	3.27009886602057e-09	<0.001	GO:0005730	nucleolus
33	351	0.378532533073338	2.40744755831632e-05	0.041	GO:0044419	interspecies interaction between organisms
172	2153	0.343432744866356	1.75600448691817e-16	<0.001	GO:0044428	nuclear part
42	485	0.340242072283576	1.48503818757099e-05	0.022	GO:0022403	cell cycle phase
40	462	0.339711208044223	2.3766092104457e-05	0.04	GO:0044265	cellular macromolecule catabolic process
60	701	0.339160869975101	3.23571054989503e-07	0.001	GO:0007049	cell cycle
41	475	0.338432406345785	2.02295400933921e-05	0.034	GO:0044427	chromosomal part
68	803	0.336220984744933	7.65671927087538e-08	<0.001	GO:0003723	RNA binding
76	911	0.331218188924209	2.33893486630229e-08	<0.001	GO:0005654	nucleoplasm
43	511	0.326468203963555	2.35721317636639e-05	0.04	GO:0006974	response to DNA damage stimulus
44	530	0.320096499039325	2.63517822318679e-05	0.043	GO:0030529	ribonucleoprotein complex
590	11330	0.307725883035438	1.0960786202461e-16	<0.001	GO:0044424	intracellular part
128	1662	0.307694952226285	4.08803837662614e-11	<0.001	GO:0016070	RNA metabolic process
51	644	0.298647909108822	2.19483688586676e-05	0.035	GO:0051726	regulation of cell cycle
58	765	0.278965516407217	2.16486041316539e-05	0.035	GO:0022402	cell cycle process
152	2160	0.26839415907082	3.64000656801177e-10	<0.001	GO:0090304	nucleic acid metabolic process
151	2171	0.261862930845385	9.73365119250831e-10	<0.001	GO:0043228	non-membrane-bounded organelle
151	2171	0.261862930845385	9.73365119250831e-10	<0.001	GO:0043232	intracellular non-membrane-bounded organelle
340	5601	0.253862278441637	1.31291950589379e-14	<0.001	GO:0044422	organelle part
492	9102	0.250272137858337	4.30276700995675e-14	<0.001	GO:0043229	intracellular organelle
492	9117	0.248695876930274	6.1868252491176e-14	<0.001	GO:0043226	organelle
178	2686	0.244422165955063	9.13870908091538e-10	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
332	5522	0.243714592261868	1.60539211670902e-13	<0.001	GO:0044446	intracellular organelle part
200	3093	0.237932293242852	4.47608150862857e-10	<0.001	GO:0034641	cellular nitrogen compound metabolic process
204	3199	0.231396211719754	9.30860336734763e-10	<0.001	GO:0006807	nitrogen compound metabolic process
251	4046	0.231319523511239	5.73866479299139e-11	<0.001	GO:0044260	cellular macromolecule metabolic process

Continued on next page

Table 3.3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
449	8238	0.229250611380594	1.64011741005672e-12	<0.001	GO:0043231	intracellular membrane-bounded organelle
449	8242	0.228829514942579	1.79770699314044e-12	<0.001	GO:0043227	membrane-bounded organelle
338	5867	0.217174729554922	3.40468652787424e-11	<0.001	GO:0044237	cellular metabolic process
290	4933	0.212021859434929	3.43579358808612e-10	<0.001	GO:0005737	cytoplasm
302	5211	0.207147691965801	5.67529432661151e-10	<0.001	GO:0005634	nucleus
111	1743	0.20458125474222	1.63777858635897e-05	0.024	GO:0043412	macromolecule modification
266	4601	0.19325730820235	1.836736540134e-08	<0.001	GO:0043170	macromolecule metabolic process
363	6632	0.19076943584687	3.42220660572212e-09	<0.001	GO:0008152	metabolic process
328	5992	0.177849130815706	4.97783632875206e-08	<0.001	GO:0044238	primary metabolic process
174	3011	0.167727728807245	1.68644378475076e-05	0.031	GO:0010468	regulation of gene expression
315	6058	0.137883228026376	2.01099934629971e-05	0.034	GO:0044444	cytoplasmic part
Module 3						
8	25	1.12174413109538	1.6195735819603e-06	0.002	GO:0051539	4 iron, 4 sulfur cluster binding
10	50	0.84991907137197	9.0161990136549e-06	0.021	GO:0051536	iron-sulfur cluster binding
10	50	0.84991907137197	9.0161990136549e-06	0.021	GO:0051540	metal cluster binding
25	229	0.538541159788376	7.47720629054225e-07	<0.001	GO:0051186	cofactor metabolic process
271	6058	0.166074173009115	2.78569620584793e-06	0.005	GO:0044444	cytoplasmic part
453	11330	0.161739187745443	1.86883086496289e-05	0.037	GO:0044424	intracellular part
Module 4						
13	97	0.672080308928874	2.01477245308462e-05	0.047	GO:0005741	mitochondrial outer membrane
32	434	0.38218093135728	2.59169277455588e-05	0.05	GO:0031966	mitochondrial membrane
267	6058	0.209635231286387	1.1954907403548e-08	<0.001	GO:0044444	cytoplasmic part
430	11330	0.189685955412112	1.81000117796974e-06	0.003	GO:0044424	intracellular part
266	6401	0.167667578490431	3.97412665691163e-06	0.008	GO:0005515	protein binding
211	4933	0.164971820301435	1.42128334658729e-05	0.024	GO:0005737	cytoplasm
Module 5						
7	33	1.07143232855034	1.02220662682094e-05	0.013	GO:0006695	cholesterol biosynthetic process
34	602	0.418505790853493	2.98488408837491e-06	0.005	GO:0006396	RNA processing
42	803	0.386267510008865	1.45554203385518e-06	0.003	GO:0003723	RNA binding
35	678	0.376246720272173	1.50758252204483e-05	0.044	GO:0044451	nucleoplasm part
74	1662	0.32717492580596	8.08519738698485e-08	<0.001	GO:0016070	RNA metabolic process
92	2160	0.316438423646489	1.22530842680858e-08	<0.001	GO:0090304	nucleic acid metabolic process
109	2686	0.302719794834586	5.16088110052742e-09	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
204	5867	0.290215289462742	2.42264491073003e-11	<0.001	GO:0044237	cellular metabolic process
86	2153	0.279472007299701	7.06167821789315e-07	0.001	GO:0044428	nuclear part
258	8238	0.273394775085421	3.30641000342457e-10	<0.001	GO:0043231	intracellular membrane-bounded organelle
258	8242	0.272981776660256	3.51091215759553e-10	<0.001	GO:0043227	membrane-bounded organelle
320	11330	0.270756828023873	2.59329735790549e-08	<0.001	GO:0044424	intracellular part
117	3093	0.27018336330746	6.56451317805923e-08	<0.001	GO:0034641	cellular nitrogen compound metabolic process
120	3199	0.267574242908932	6.5838967556536e-08	<0.001	GO:0006807	nitrogen compound metabolic process
274	9102	0.259905768442695	3.97303688891244e-09	<0.001	GO:0043229	intracellular organelle
145	4046	0.258957848074746	2.58310912976049e-08	<0.001	GO:0044260	cellular macromolecule metabolic process
274	9117	0.258358056899225	4.89474559380795e-09	<0.001	GO:0043226	organelle
186	5522	0.252387490202399	8.07568575286686e-09	<0.001	GO:0044446	intracellular organelle part
198	5992	0.249866826361352	8.08174185764168e-09	<0.001	GO:0044238	primary metabolic process
187	5601	0.24739223761693	1.47932676168125e-08	<0.001	GO:0044422	organelle part
212	6632	0.239108907629557	2.58709518947142e-08	<0.001	GO:0008152	metabolic process
155	4601	0.229008499074375	4.23085771940228e-07	<0.001	GO:0043170	macromolecule metabolic process
169	5211	0.214851860778109	1.14557794256476e-06	0.002	GO:0005634	nucleus
188	6058	0.198889582472837	3.92901841156595e-06	0.007	GO:0044444	cytoplasmic part
Module 6						
5	17	1.35877762485371	1.27620339057008e-05	0.027	GO:0005680	anaphase-promoting complex
7	27	1.28109582819072	5.64148602718057e-07	0.001	GO:0007094	mitotic cell cycle spindle assembly checkpoint
7	28	1.26038559665177	7.3990905561597e-07	0.004	GO:0045841	negative regulation of mitotic metaphase/anaphase transition
7	28	1.26038559665177	7.3990905561597e-07	0.004	GO:0071173	spindle assembly checkpoint
7	28	1.26038559665177	7.3990905561597e-07	0.004	GO:0071174	mitotic cell cycle spindle checkpoint
7	30	1.22170492439836	1.23095671972835e-06	0.005	GO:0031577	spindle checkpoint
7	33	1.16944999653699	2.45864548304647e-06	0.007	GO:0030071	regulation of mitotic metaphase/anaphase transition
7	33	1.16944999653699	2.45864548304647e-06	0.007	GO:0045839	negative regulation of mitosis
7	33	1.16944999653699	2.45864548304647e-06	0.007	GO:0051784	negative regulation of nuclear division
18	288	0.561314315261963	1.18040781186384e-05	0.026	GO:0051301	cell division
204	8238	0.251051414092351	1.67494198073935e-07	<0.001	GO:0043231	intracellular membrane-bounded organelle
204	8242	0.250640665799907	1.75182061275321e-07	<0.001	GO:0043227	membrane-bounded organelle
216	9102	0.231422137926905	1.80780431583472e-06	0.006	GO:0043229	intracellular organelle
216	9117	0.229882447927919	2.1050968158218e-06	0.007	GO:0043226	organelle
135	5211	0.208102879924293	2.05983907876377e-05	0.037	GO:0005634	nucleus
Module 8						
3	4	2.15674334421143	1.64587987532083e-05	0.032	GO:0048280	vesicle fusion with Golgi apparatus

Continued on next page

Table 3.3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
5	21	1.31441473215979	1.74263922633301e-05	0.037	GO:0032201	telomere maintenance via semi-conservative replication
8	35	1.28611338854416	6.73739625327887e-08	0.001	GO:0006261	DNA-dependent DNA replication
5	23	1.26467580787443	2.80631885844604e-05	0.047	GO:0000722	telomere maintenance via recombination
6	34	1.15088257068433	1.54366414227154e-05	0.024	GO:0010833	telomere maintenance via telomere lengthening
8	69	0.935700200428791	1.48897081315905e-05	0.024	GO:0009411	response to UV
15	170	0.802685314737047	1.12157969185014e-07	0.001	GO:0006260	DNA replication
23	511	0.490955055548057	9.92989139304257e-06	0.019	GO:0006974	response to DNA damage stimulus
42	1327	0.340009071984013	1.98581541366046e-05	0.04	GO:0009059	macromolecule biosynthetic process
41	1296	0.338927033802861	2.55286781292076e-05	0.043	GO:0034645	cellular macromolecule biosynthetic process
63	2160	0.318931266474648	1.75657887022238e-06	0.002	GO:0090304	nucleic acid metabolic process
128	5211	0.298549267749992	1.8238869432011e-08	<0.001	GO:0005634	nucleus
60	2153	0.292881834504383	1.38556689832231e-05	0.023	GO:0044428	nuclear part
100	4046	0.267683010803659	1.63739514172819e-06	0.002	GO:0044260	cellular macromolecule metabolic process
174	8238	0.263181580222692	4.66457450693039e-07	0.002	GO:0043231	intracellular membrane-bounded organelle
174	8242	0.262771810066496	4.85332067450148e-07	0.002	GO:0043227	membrane-bounded organelle
107	4601	0.239108491634204	1.07913454880254e-05	0.02	GO:0043170	macromolecule metabolic process
182	9102	0.22936103280645	1.25445382233209e-05	0.02	GO:0043229	intracellular organelle
182	9117	0.227824532081016	1.42589555343329e-05	0.023	GO:0043226	organelle
Module 10						
190	11330	0.26726174432135	1.71596077751879e-05	0.032	GO:0044424	intracellular part
Module 14						
20	1065	0.528808096440223	1.73407898205248e-05	0.032	GO:0044248	cellular catabolic process
33	2111	0.479914638806195	1.06712795272747e-06	<0.001	GO:0005829	cytosol
39	3093	0.388858388574349	1.71524549827148e-05	0.032	GO:0034641	cellular nitrogen compound metabolic process
42	3431	0.382865420385352	1.42566606441697e-05	0.023	GO:0032991	macromolecular complex
60	5867	0.34446562764228	2.14508227397237e-05	0.032	GO:0044237	cellular metabolic process
61	6058	0.338468119322826	2.84679848054408e-05	0.041	GO:0044444	cytoplasmic part

Table 3.4: Full list of Gene Ontology categories enriched in coexpressed gene modules derived from a mirrored analysis of pool/split datasets. Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al., 2009). A total of 16 modules were detected

N	X	LOD	P	P adj	attrib ID	attrib name
Module 1						
13	272	0.693095067094612	1.02007791233691e-05	0.011	GO:0006511	ubiquitin-dependent protein catabolic process
13	279	0.68135455023532	1.33744095433394e-05	0.018	GO:0019941	modification-dependent protein catabolic process
13	282	0.676415542913035	1.49811431033085e-05	0.019	GO:0043632	modification-dependent macromolecule catabolic process
13	289	0.665098403683509	1.94061972015811e-05	0.024	GO:0051603	proteolysis involved in cellular protein catabolic process
17	462	0.57616161468926	1.43352606692572e-05	0.018	GO:0044265	cellular macromolecule catabolic process
19	545	0.553524223505169	9.47597917000467e-06	0.009	GO:0009057	macromolecule catabolic process
29	1065	0.451649915259144	5.44374777913331e-06	0.007	GO:0044248	cellular catabolic process
50	2153	0.408372755551646	1.54662372619939e-07	¡0.001	GO:0044428	nuclear part
38	1647	0.385284888523943	8.15733567738168e-06	0.007	GO:0006464	protein modification process
39	1743	0.371699071795165	1.23788499774497e-05	0.013	GO:0043412	macromolecule modification
75	4046	0.33429458014633	5.14482381026089e-07	¡0.001	GO:0044260	cellular macromolecule metabolic process
93	5522	0.313764330996431	7.30843785397851e-07	¡0.001	GO:0044446	intracellular organelle part
93	5601	0.304538452915935	1.46302793919708e-06	0.002	GO:0044422	organelle part
79	4601	0.296962579640552	5.07939215329675e-06	0.005	GO:0043170	macromolecule metabolic process
95	5867	0.292386561578076	3.34586773855207e-06	0.003	GO:0044237	cellular metabolic process
104	6632	0.29101237790736	3.17018865849522e-06	0.003	GO:0008152	metabolic process
96	5992	0.287467105557634	4.64261918086881e-06	0.005	GO:0044238	primary metabolic process
Module 3						
5	16	2.01970458115929	9.51160735183834e-09	¡0.001	GO:0016254	preassembly of GPI anchor in ER membrane
6	30	1.76902612001682	5.55484070660725e-09	¡0.001	GO:0018410	C-terminal protein amino acid modification
5	25	1.76842092954302	1.11841352190966e-07	0.001	GO:0006501	C-terminal protein lipidation
5	49	1.43120717990547	3.66720733305651e-06	0.006	GO:0006497	protein lipidation
7	131	1.12827314119116	3.16442237797319e-06	0.005	GO:0043687	post-translational protein modification
22	1743	0.516764340612836	1.71498823237458e-05	0.023	GO:0043412	macromolecule modification
Module 9						
19	2084	0.56988964370998	2.20527274140205e-05	0.031	GO:0044267	cellular protein metabolic process

Table 3.5: Novel splice junctions discovered in spikes. No novel junctions are expected to be discovered in spike-in sequences, therefore such junctions are almost certainly experimental artifacts. Shown is the number of collapsed fragments supporting each junctions in a library

chr	left	right	strand	12517 GM12878 30 cells	12520 10 cells	12820 cell 208	12821 pool/split 5	12823 pool/split 7	12824 pool/split 8	13268 217 pool/split 1	13269 218 pool/split 2	13270 219 pool/split 3	13271 220 pool/split 4	13273 226 pool/split 10	13278 cell 204	13280 232 pool/split
AP2	308	1191	−	0	0	0	0	24	0	0	0	0	0	0	0	0
AP2	357	1205	−	1	0	0	0	0	0	0	0	0	0	0	0	0
AP2	365	1213	+	2	0	0	0	0	0	0	0	0	0	0	0	0
AP2	380	476	+	0	44	0	0	0	0	0	0	0	0	0	0	0
AP2	380	520	+	0	0	0	28	0	0	0	0	0	0	0	0	0
Lambda 9786 clone F	5092	5172	−	0	0	0	0	0	0	0	38	0	0	0	0	0
OBF5	334	466	+	0	0	0	39	0	0	0	0	0	0	0	0	0
OBF5	376	607	+	0	0	0	0	0	0	0	0	0	0	1	0	0
OBF5	381	612	+	0	0	0	0	0	0	0	0	0	0	47	0	0
OBF5	394	893	+	0	0	0	0	0	0	0	0	3	0	0	0	0
OBF5	411	860	+	0	0	1	0	0	0	0	0	0	0	0	0	0
OBF5	413	862	+	0	0	39	0	0	0	0	0	0	0	0	0	0
OBF5	417	526	−	0	0	0	28	0	0	0	0	0	0	0	0	0
OBF5	558	724	−	0	0	0	0	0	0	45	0	0	0	0	0	0
OBF5	558	988	−	0	0	0	0	0	0	0	0	0	44	0	0	0
OBF5	563	1000	+	0	0	0	0	0	0	0	0	0	0	0	0	2
OBF5	572	806	+	0	0	0	0	0	17	0	0	0	0	0	0	0
OBF5	572	848	+	0	0	18	0	0	0	0	0	0	0	0	0	0
OBF5	708	893	+	0	0	0	0	0	31	0	0	0	0	0	0	0
OBF5	708	992	+	0	0	0	0	0	0	0	0	0	0	0	36	0
VATG3	28	277	+	0	0	0	6	0	0	0	0	0	0	0	0	0

Table 3.6: Gene Ontology categories enriched in genes displaying splice site switching between individual cells. Gene Ontology enrichment was assessed using FuncAssociate2.0 (Berriz et al., 2009).

N	X	LOD	P	P adj	attrib ID	attrib name
7	18	1.60434572330591	7.81336494766272e-09	<0.001	GO:0030530	heterogeneous nuclear ribonucleoprotein complex
6	22	1.38371032623861	1.0731443930779e-06	<0.001	GO:0000313	organellar ribosome
6	22	1.38371032623861	1.0731443930779e-06	<0.001	GO:0005761	mitochondrial ribosome
9	62	1.04144125075565	6.88184640563181e-07	<0.001	GO:0015934	large ribosomal subunit
10	79	0.972456465728994	6.10664091150019e-07	<0.001	GO:0071013	catalytic step 2 spliceosome
19	154	0.964236659283489	7.8714230213414e-12	<0.001	GO:0003735	structural constituent of ribosome
20	175	0.927300715975638	9.17354662099906e-12	<0.001	GO:0000375	RNA splicing, via transesterification reactions
19	168	0.92117404141009	3.70476457645963e-11	<0.001	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
19	168	0.92117404141009	3.70476457645963e-11	<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
15	141	0.888071082569232	1.03966880671155e-08	<0.001	GO:0005681	spliceosomal complex
16	159	0.861650703193511	7.48121263599197e-09	<0.001	GO:0005840	ribosome
23	243	0.838249153771732	1.25842265057125e-11	<0.001	GO:0006412	translation
46	530	0.826167157208287	9.02684713922854e-21	<0.001	GO:0030529	ribonucleoprotein complex
23	282	0.766509674540248	2.54718803100937e-10	<0.001	GO:0008380	RNA splicing
26	364	0.706261794294909	3.04918522987691e-10	<0.001	GO:0006397	mRNA processing
28	408	0.689368099675319	1.53848128981224e-10	<0.001	GO:0010467	gene expression
38	584	0.67579720156425	3.18781787872945e-13	<0.001	GO:0016071	mRNA metabolic process
47	803	0.636519902260068	1.87953754853827e-14	<0.001	GO:0003723	RNA binding
36	602	0.632953237795702	1.62794718956973e-11	<0.001	GO:0006396	RNA processing
16	277	0.597979679500433	1.35166171675457e-05	0.022	GO:0034660	ncRNA metabolic process
22	462	0.511423641446678	7.66202769258217e-06	0.015	GO:0044265	cellular macromolecule catabolic process
39	911	0.476845884038679	3.34218894462279e-08	<0.001	GO:0005654	nucleoplasm
24	545	0.47631618686561	1.09283168115262e-05	0.018	GO:0009057	macromolecule catabolic process
118	3431	0.473494407052489	1.56368496887015e-17	<0.001	GO:0032991	macromolecular complex
29	687	0.460527035699242	2.93531833594463e-06	0.005	GO:0044429	mitochondrial part
80	2160	0.454899179204351	5.73798891785767e-13	<0.001	GO:0090304	nucleic acid metabolic process
64	1662	0.453902252858483	4.77393199090586e-11	<0.001	GO:0016070	RNA metabolic process
51	1296	0.449261990101858	3.27437399406931e-09	<0.001	GO:0034645	cellular macromolecule biosynthetic process
79	2153	0.448759965851525	1.38228172140963e-12	<0.001	GO:0044428	nuclear part
52	1327	0.448329514135066	2.57197941829179e-09	<0.001	GO:0009059	macromolecule biosynthetic process
94	2686	0.444930025696524	6.62602821448192e-14	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
77	2111	0.443065370078769	4.00922406264811e-12	<0.001	GO:0005829	cytosol
127	4046	0.437107059343321	7.36675726742068e-16	<0.001	GO:0044260	cellular macromolecule metabolic process
102	3093	0.424687277579968	1.53499538750493e-13	<0.001	GO:0034641	cellular nitrogen compound metabolic process
234	11330	0.410177502418009	4.26279474281887e-11	<0.001	GO:0044424	intracellular part
73	2121	0.408294011011211	2.6476698540441e-10	<0.001	GO:0044249	cellular biosynthetic process
102	3199	0.40626576617879	1.33115703600882e-12	<0.001	GO:0006807	nitrogen compound metabolic process
47	1315	0.398127144349952	2.71558632079032e-07	<0.001	GO:0005739	mitochondrion
157	5867	0.394560459760076	4.17387500074186e-14	<0.001	GO:0044237	cellular metabolic process
150	5522	0.390555837986407	8.85159181992679e-14	<0.001	GO:0044446	intracellular organelle part
74	2232	0.390317428579321	1.04923257025558e-09	<0.001	GO:0009058	biosynthetic process
131	4601	0.386021432185656	5.00037984912871e-13	<0.001	GO:0043170	macromolecule metabolic process
150	5601	0.381254965130809	3.18071089534335e-13	<0.001	GO:0044422	organelle part
157	6058	0.372860438339392	8.46004601459964e-13	<0.001	GO:0044444	cytoplasmic part
151	5992	0.342490468755501	4.80405598147904e-11	<0.001	GO:0044238	primary metabolic process
161	6632	0.334875288021427	1.12251899150943e-10	<0.001	GO:0008152	metabolic process
63	2084	0.331075785878351	7.59282101998268e-07	<0.001	GO:0044267	cellular protein metabolic process
184	8242	0.317619158465355	1.64375270407462e-09	<0.001	GO:0043227	membrane-bounded organelle
183	8238	0.311179434676407	3.27953523717327e-09	<0.001	GO:0043231	intracellular membrane-bounded organelle
194	9102	0.300093429975042	2.09541238702377e-08	<0.001	GO:0043229	intracellular organelle
194	9117	0.298556745854624	2.46813702747217e-08	<0.001	GO:0043226	organelle
70	2644	0.269207945316679	1.84621206416052e-05	0.04	GO:0019538	protein metabolic process
74	2821	0.267942614823057	1.34448891236763e-05	0.021	GO:0043234	protein complex
213	10840	0.265155481178129	2.70868933554443e-06	0.005	GO:0009987	cellular process

Table 3.7: Read mapping statistics. Note that libraries with numbers lower than 12543 used a different spike-in cocktail than other libraries and the correspondence between initial spike-in amounts and final FPKM scores in the sequenced libraries was poor. For this reason, they were excluded from analyses based on estimating absolute transcript abundances in copies per cell.

Library	Read Length	Unique	UniqueSplices	Multi	MultiSplices
12515 100-cell pool A	1x100	17,687,845	3,209,817	2,324,217	87,366
12516 100-cell pool B	1x100	19,196,833	3,613,603	2,472,612	116,124
12517 30-cell pool A	1x100	19,656,269	3,836,281	2,747,606	112,715
12518 30-cell pool B	1x100	15,906,819	3,105,647	2,209,219	107,243
12519 10-cell pool A	1x100	25,589,985	7,716,359	3,942,315	264,713
12520 10-cell pool B	1x100	14,033,035	3,831,207	2,172,320	92,664
12522 cell 183	1x100	13,444,432	4,123,615	1,991,506	151,473
12523 cell 184	1x100	18,553,787	6,282,207	2,753,213	162,393
12524 cell 185	1x100	15,306,477	4,973,962	2,375,825	123,920
12534 cell 186	1x100	9,412,759	1,792,734	1,104,103	92,146
12535 cell 187	1x100	12,021,473	2,593,517	1,762,078	122,152
12536 cell 188	1x100	6,173,793	1,609,714	751,818	35,935
12537 cell 189	1x100	8,900,605	2,552,063	1,195,651	71,165
12538 cell 190	1x100	11,976,901	3,061,070	1,578,373	114,265
12539 cell 191	1x100	4,894,790	990,183	687,469	55,952
12540 cell 192	1x100	8,586,601	2,191,767	1,312,434	70,208
12541 cell 193	1x100	11,615,819	2,810,842	1,636,014	75,938
12542 cell 194	1x100	9,299,741	2,543,984	1,388,630	61,370
12543 cell 195	1x100	9,051,228	1,583,943	1,172,717	52,683
12818 cell 200	1x100	9,465,272	2,903,793	1,338,444	87,282
12819 cell 205	1x100	11,895,334	3,486,064	1,184,543	59,413
12820 cell 208	1x100	13,034,342	2,346,996	1,418,030	120,778
12821 pool/split 5	1x100	9,152,130	2,394,362	1,520,080	76,965
12822 pool/split 6	1x100	13,938,165	3,517,926	2,286,058	113,187
12823 pool/split 7	1x100	11,217,362	1,872,905	1,154,032	73,843
12824 pool/split 8	1x100	11,822,904	2,135,005	1,364,032	70,389
13270 pool/split 3 219	1x100	7,416,424	4,799,669	1,463,631	457,029
13271 pool/split 4 220	1x100	8,421,706	5,262,489	1,644,668	496,781
13272 pool/split 9 225	1x100	12,782,172	4,509,292	1,480,617	427,615
13273 pool/split 10 226	1x100	10,325,385	6,582,179	2,100,134	641,196
13274 10ng RNA rep1	1x100	33,234,882	4,315,401	1,629,950	267,868
13275 10ng RNA rep2	1x100	36,981,036	5,266,981	1,704,651	301,449
13276 100pg RNA rep1	1x100	11,363,854	4,904,470	1,008,244	258,637
13277 100pg RNA rep2	1x100	34,939,583	6,750,980	2,212,161	442,062
13278 cell 204	1x100	20,631,514	11,290,949	3,418,238	921,764
13279 cell 207	1x100	10,926,463	4,949,640	1,664,688	490,150
13280 pool/split 232	1x100	21,240,282	9,537,592	2,722,244	726,943
13281 pool/split 233	1x100	25,425,429	9,576,495	2,510,136	703,065
13282 cell 235	1x100	10,167,950	3,677,729	966,782	191,523
13283 cell 236	1x100	18,782,295	7,784,497	2,210,674	572,837
13284 cell 237	1x100	25,766,827	8,914,958	2,235,457	594,889
13285 cell 238	1x100	16,334,009	6,842,776	2,351,813	602,952
13286 cell 239	1x100	19,717,157	5,801,008	2,473,230	595,738
13287 cell 240	1x100	21,881,195	8,373,245	2,386,125	645,571
13288 cell 242	1x100	19,165,078	6,146,306	1,338,990	330,167
13289 cell 243	1x100	24,802,270	9,575,191	2,885,175	744,245

Continued on next page

Table 3.7 – *Continued from previous page*

Library	Read Length	Unique	UniqueSplices	Multi	MultiSplices
13290 cell 244	1x100	7,400,266	3,086,583	741,408	223,657
13291 cell 245	1x100	21,024,295	7,093,623	2,111,549	519,415
13292 pool/split 246	1x100	17,296,143	8,394,643	2,223,819	572,943
13294 pool/split 248	1x100	14,399,162	6,272,094	1,784,982	459,195
13295 pool/split 249	1x100	22,428,103	10,454,916	2,898,266	815,093
13296 pool/split 250	1x100	19,745,007	8,825,294	2,468,779	697,549
13297 pool/split 251	1x100	21,239,455	9,833,749	2,936,743	724,006
13298 pool/split 252	1x100	4,674,393	2,237,759	591,303	145,488
13299 pool/split 253	1x100	20,948,852	9,729,505	2,726,042	709,672
13300 10-cell pool 254	1x100	29,113,485	8,790,470	2,600,560	702,142
13301 11-cell pool 255	1x100	34,836,093	11,643,761	3,802,039	1,080,518
13302 100-cell pool 256	1x100	18,477,603	4,084,659	1,618,554	329,602
13303 100-cell pool 257	1x100	43,640,710	11,315,061	4,819,940	1,036,363

4

Analysis of RNA-seq data from samples containing a mixture of cells from multiple species

This chapter includes material that was previously published in:

“Raskatov JA, Nickols NG, Hargrove AE, Marinov GK, Wold B, Dervan PB. 2012. Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide. *Proc Natl Acad Sci U S A.* **109**(4):16041-16045. doi: 10.1073/pnas.1214267109

The experimental data in it was generated by Jevgeni Raskatov in the Dervan lab. I contributed the computational approach to xenograft RNA-seq analysis. The paper is reprinted in Appendix D.

4.1 Introduction, Results and Discussion

Multicellular organisms do not exist in isolation but live in association with a very large number of microorganisms (NIH HMP Working Group et al. 2009; Human Microbiome Project Consortium 2012), and they also encounter various pathogens. A consequence of this is that the transcriptome of many organs and tissues is not purely the transcriptome of the host species but a complex mixture of the transcriptomes of the host and all other organisms living in association with it. Understanding the dynamics and interplay of the combined transcriptome is of great interest, and this is especially true about host-pathogen interactions (Westermann et al. 2012). Fortunately, the phylogenetic distance between most pathogens and their multicellular hosts is great, and it is relatively straightforward to analyze their transcriptomes from the same pool of RNA using RNA-seq, as the high level of sequence divergence means no or very few reads map ambiguously to both genomes.

Xenografts, on the other hand, represent a system, which presents much more significant computational challenges to the characterization of its transcriptome. A typical xenograft model involves the grafting of human tumor cells into immunosuppressed mice. Such systems are widely used as models of human cancers in drug development (Sano & Myers 2009; van Weerden et al. 2009; Tentler et al. 2012; Luconi & Mannelli 2012), and accurately assessing transcriptional changes in response to drug treatments is therefore of high importance. The challenge is that host cells often invade the xenograft (although the degree to which this is happening varies depending on the specifics of the tumor and the host tissue) and thus even careful isolation of xenograft tissue from the host results in a mixture of mouse and human cells. Despite some 70 million years of divergence, mouse and human genes retain significant sequence homology (Mouse Genome Sequencing Consortium 2002) and assignment of short reads is not always un-

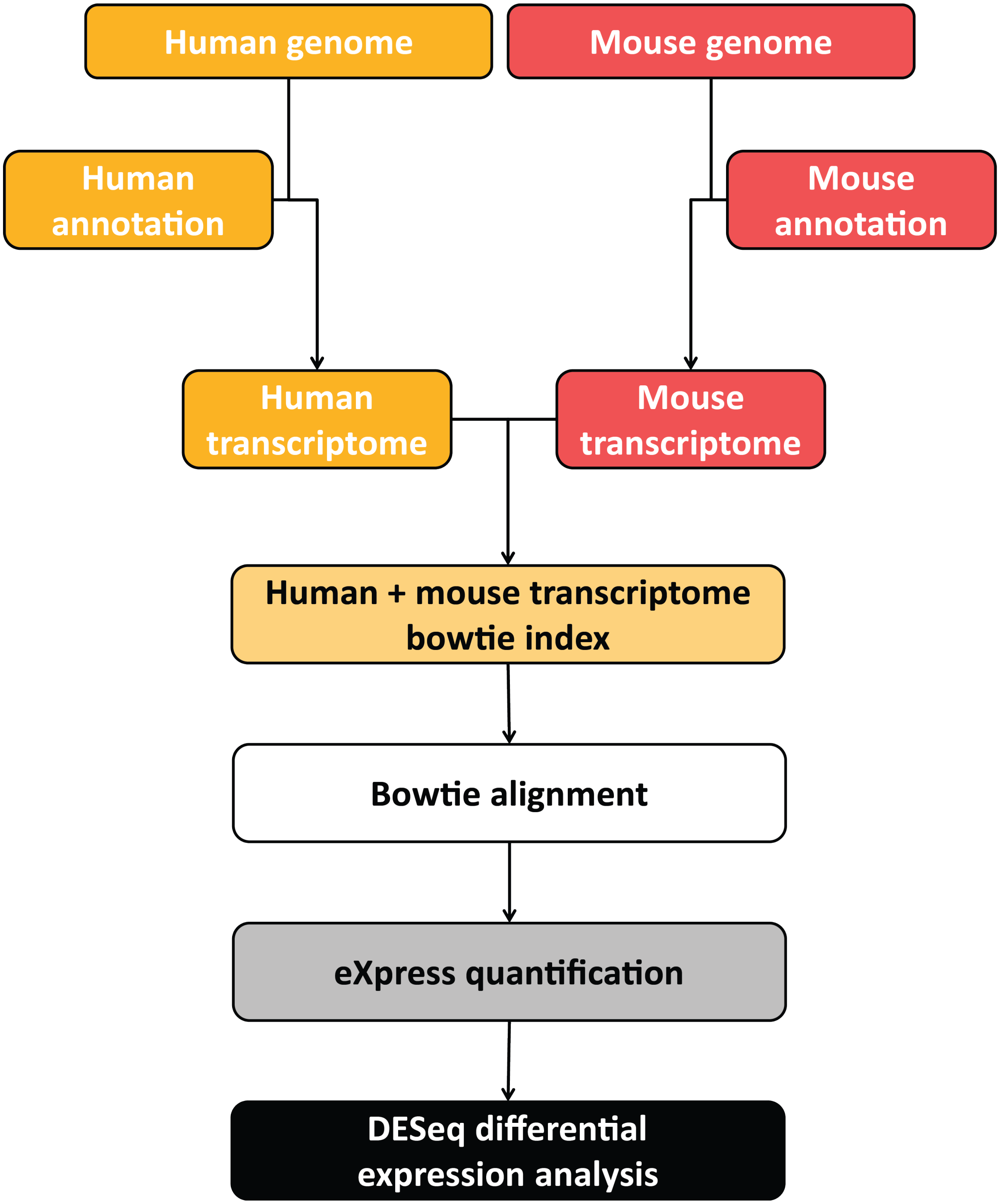


Figure 4.1: Strategy for the simultaneous quantification of dual-species (in this case, mouse and human) RNA-seq data. The sequences of mouse and human transcripts are extracted from the corresponding genomes and annotation files, and a combined Bowtie (Langmead et al. 2009) index for the two transcriptomes is built. In this case, the refSeq annotations were used (Pruitt et al. 2009), downloaded from the UCSC Genome Browser. Sequencing reads from xenograft RNA-seq experiments are then aligned against this combined index, allowing reads to map to an unlimited number of locations. The resulting alignments are used for simultaneous quantification of both transcriptomes using eXpress. The quantification values are then used for downstream analysis, either as FPKMs, or for assessment of differential expression using DESeq (Anders & Huber 2010); in the latter case the "effective counts" values are used.

Table 4.1: Percentage of reads mapping to hg19 or mm9 in the A549-luc-C8 xenograft-derived RNA-seq samples as well as pure mouse and human samples. Read were mapped separately to the mouse (mm9) and human (hg19) genomes using TopHat. The nearly equal number of reads mapping to each genome in the xenograft samples compared to the lower fraction of reads mapping to the other genome in A549 cells and in the mouse spleen sample indicated the presence of a significant fraction of mouse cells in the xenografts.

Sample/Condtion	Number of reads	% mapping to hg19	% mapping to mm9
XenoVehicle 1	35,478,968	59.5	42.9
Xeno Vehicle 2	50,839,514	36.5	27.5
Xeno Vehicle 3	50,150,429	59.2	46.3
Xeno Treated 1	54,437,744	59.5	43.3
Xeno Treated 2	49,087,273	39.7	28.9
Xeno Treated 3	34,553,534	60.4	44.4
A549 in vitro	35,187,689	83.0	11.9
SCID-bg spleen	34,932,537	11.9	95.8

ambiguous, potentially confounding quantification of gene expression changes in the xenograft (Conway et al. 2012; Valdes et al. 2013; Rossello et al. 2013). On the other hand, the presence of mouse cells is also a potential benefit, as in a xenograft containing a large proportion of normal host cells, RNA-seq allows in principle the simultaneous measurements of gene expression changes in both tumor and normal cells. This can be illuminating about the differences between the effects of the drug tested on normal (host) and tumor (xenograft) cells.

Another system where the same problem is encountered are heterokaryons derived from the fusion of cells from two species (for example, mouse emryonic stem cells and human fibroblasts, leading to reprogramming of the latter into a pluripotent stem cell state; Blau et al. 1983; Blau & Blakely 1999), where it is even more pronounced as the proportion of the transcriptome in the sample deriving from each species is approximately equal. Such a system was studied by Brady et al. 2013, whose solution to dealing with reads mapping to both species was to simply discard them.

Several other studies in recent years have also addressed the problem. Bradford et al. 2013 studied tumor and host changes in gene expression after treatment of xenografts the VEGFR tyrosine kinase inhibitor cediranib. Their solution was also to map reads to each species separately and discard the ambiguous ones. This was also the essence of the approach adopted by Rossello et al. 2013, who studies small cell

Table 4.2: Comparison of qRT-PCR and RNA-seq of A549-luc-C8 tumor xenograft gene expression levels. qRT-PCR measurements are normalized to GUSB as the housekeeping gene, with three independent experiments with $N = 5$ animals per treatment condition averaged. Arrows indicated the direction of expression change between the untreated and treated condition ($\Downarrow$ indicates downregulation upon treatment while $\Uparrow$) corresponds to upregulation

Gene	Fold change (qPCR)	Fold change (RNA-seq)
ATM	$\Downarrow 1.5 \pm 0.2$	$\Downarrow 1.5$ ($p > 0.05$)
NPTX1	$\Downarrow 3.3 \pm 0.6$	$\Downarrow 2.9$ ($p < 0.001$)
ROBO1	$\Downarrow 1.5 \pm 0.2$	$\Downarrow 1.7$ ($p > 0.05$)
MMP28	$\Uparrow 1.5 \pm 0.3$	$\Uparrow 2.0$ ($p < 0.05$)
EGFR	$\Downarrow 1.2 \pm 0.2$	$\Downarrow 1.3$ ($p > 0.05$)
CCL2	$\Downarrow 2.3 \pm 0.4$	$\Downarrow 1.7$ ($p < 0.001$)
SERPINE1	$\Downarrow 2.0 \pm 0.2$	$\Downarrow 1.8$ ($p < 0.001$)

lung cancer xenografts, and by Kawahara et al. 2012, who studied the mixed transcriptome of rice (*Oryza sativa*) and the fungal pathogen *Magnaporthe oryzae*.

A tool specifically designed to classify reads from xenograft samples called Xenome has been developed (Conway et al. 2012). It is based on decomposing the transcriptomes of the two species into k-mers (the set of all k-mers in a larger sequence or set of sequences consists the set of all subsequences of length k found in it) and classifying the reads into originating from the host, originating from the xenograft, ambiguous or originating from neither. Once assigned, the reads can then be used for subsequent species-specific mappings and analysis. However, this approach only classifies reads as ambiguous rather than actually attempting to assign them to a given species.

I developed what is in my opinion a much simpler and more elegant solution to the problem, one that uses all reads, assigns them to species, performs proper FPKM normalizations, and does not involve a complex read processing pipeline, by adopting the eXpress tool for isoform-level quantification of RNA-seq data (Roberts & Pachter 2013). I discussed eXpress in a prior chapter so I will not revisit how it works; for the purposes of this chapter it is necessary to note that eXpress was specifically designed to deconvolve the expression levels of transcripts that are highly similar to each other (such as the individual isoforms of a gene and even allelic variants), and to do so in transcriptomic (i.e. only the sequences of spliced transcripts) rather than genomic space. Thus it is ideally suited for the analysis of samples containing a mixture of the transcriptomes of multiple species such as xenografts, and also potentially metagenomes and metatranscriptomes. At the time of writing this text, I am not aware of any study that has actually used it for the analysis of metagenomic/metatranscriptomic samples, thus our results constituted the first proof-of-principle study confirming the utility of the approach to this kind of problems.

We studied the effect of a DNA-binding pyrrole-imidazole polyamides (designed to target 5-WGGWWW-3 sequence motifs) on a xenograft of the A549 cancer cell line onto immunocompromised SCID mice (Raskatov et al., 2012). Such polyamides are of potential therapeutic interest as they can bind to DNA sequences occupied by transcription factors driving the proliferation of cancer cells, outcompete them and antagonize their action (Dervan & Edelson 2003; Chenoweth & Dervan PB 2009; Nickols & Dervan 2007; Muzikar et al. 2009). The goal of the study was to examine the transcriptional changes in the xenograft upon polyamide treatment. However, we faced the problem of the xenograft tissue containing a significant number of host cells, which can potentially confound the quantification of gene expression in human cells (see Table 4.1). To resolve that problem I devised the pipeline outlined in Figure 4.1. It consists of extracting the transcriptome sequences for both the human and mouse genome, merging them together, creating a Bowtie index for the merged set of sequences, then mapping RNA-seq reads to it and quantifying expression levels based on the resulting alignments using eXpress. Subsequently, we used the "effective counts" generated by eXpress and DESeq (Anders & Huber 2010) to assess changes in gene expression in both species. We identified 615 differentially expressed human genes. Notably, we also found 1338 mouse genes that were differentially expressed between the two conditions. We selected several human genes for orthogonal testing of expression changes using qPCR and species-specific human primers and found excellent agreement between the fold changes estimated from RNA-seq data and those calculated using qPCR (Table 4.2), underscoring the usefulness of the eXpress-centered computational approach for analyzing dual- and multi-species RNA-seq data.

4.2 Methods

4.2.1 RNA-seq Sample Preparation

Double polyA-selection was used in order to enrich for mRNA. RNA-seq libraries were prepared using standard Illumina reagents and protocols (Mortazavi & Williams et al. 2008). All experiments were carried out in triplicates and 35-50 $\times 10^6$ single-end sequences of 50 bp were generated for each library.

4.2.2 RNA-seq Data Processing

Sequencing data were mapped to a combined human and mouse transcriptome index (using the hg19 and mm9 refSeq annotations) using

Bowtie version 0.12.7 (Langmead et al. 2009) with the following settings: `-v 2 -a`, i.e. allowing for two mismatches and an unlimited number of locations a read can map to. Alignments were quantified on the transcript level using eXpress, version 1.0.0 (Roberts & Pachter 2013). For each gene the quantification values of all its transcripts were summed and the eXpress-determined "effective counts" were used as input for differential expression analysis using DESeq (Anders & Huber 2010).

Part II

Functional Genomics of Organelles

This part contains four chapters describing the functional genomic studies of proteins associated with organellar DNA. It grew out of a collaboration with the Chan lab at Caltech, in which the goal was to map the occupancy of the TFAM protein over the mitochondrial nucleoid in human cells. The results of it are described in the first chapter here. While working on this problem we noticed that many nuclear transcription factors assayed by ENCODE exhibited strong enrichment over some areas of the mitochondrial genomes. We investigated the phenomenon in depth, which resulted in a study of the binding of nuclear transcription factors to mitochondrial DNA in animal genomes, described in the third chapter in this part. I then carried out a similar analysis on published ChIP-seq data in plants, which contain both a mitochondrial and a plastid genome. The results from it are presented in the fourth chapter. In parallel, we extended our TFAM study to other proteins involved in mitochondrial transcription and replication, the occupancy of which we mapped in a couple of ENCODE cell lines in collaboration with the Myers lab at the HudsonAlpha Institute of Biotechnology, with the results detailed in the second chapter.

5

Genome-Wide Analysis Reveals Coating of the Mitochondrial Genome by TFAM

Most of the material in this chapter was published as:

Wang YE, Marinov GK, Wold BJ, Chan DC. 2013. Genome-wide analysis reveals coating of the mitochondrial genome by TFAM. *PLoS ONE* **8**(8):e74513. doi: 10.1371/journal.pone.0074513

The experimental data in it was generated by Yun Elisabeth Wang in the Chan lab. I contributed the computational analysis. The paper is reprinted in Appendix H.

Abstract

Human mitochondria contain a 16.6 kb circular-mapping genome encoding 13 proteins as well as mitochondrial tRNAs and rRNAs. Copies of the genome are organized into nucleoids containing both DNA and proteins, including the machinery required for mtDNA replication and transcription. The transcription factor TFAM is critical for initiation of transcription and replication of the genome, and is also thought to perform a packaging function. Although specific binding sites required for initiation of transcription have been identified in the D-loop, little is known about the characteristics of TFAM binding in its nonspecific packaging state. In addition, it is unclear whether TFAM also plays a role in the regulation of nuclear gene expression. We investigated these questions by using ChIP-seq to directly localize TFAM binding to DNA in human cells. Our results demonstrated that TFAM uniformly coats the whole mitochondrial genome, with no evidence of robust TFAM binding to the nuclear genome and represent the first direct assessment of TFAM binding on a genome-wide scale.

5.1 Introduction

Mitochondria are essential eukaryotic organelles, serving as the epicenter of ATP production in the cell through oxidative phosphorylation. To perform this bioenergetic function, mitochondria utilize gene products encoded by the mitochondrial genome, a circular DNA that is 16.6 kb long. This genome is organized into DNA/protein structures termed nucleoids (Bogenhagen et al., 2008). Mitochondrial DNA (mtDNA) encodes thirteen components of the electron transport chain, as well as 22 tRNAs and two ribosomal RNA genes. These gene products are essential for the proper function of the respiratory chain, and therefore maintenance of mtDNA levels and sequence fidelity is essential for cellular bioenergetics. In a human cell, there are hundreds to thousands of copies of the mtDNA genome (Bogenhagen & Clayton, 1974; Satoh & Kuroiwa, 1991). Damage or depletion of mtDNA causes numerous inherited disorders, including Alpers Disease, ataxia neuropathy spectrum, and progressive external ophthalmoplegia (Suomalainen et al., 2010; Stumpf

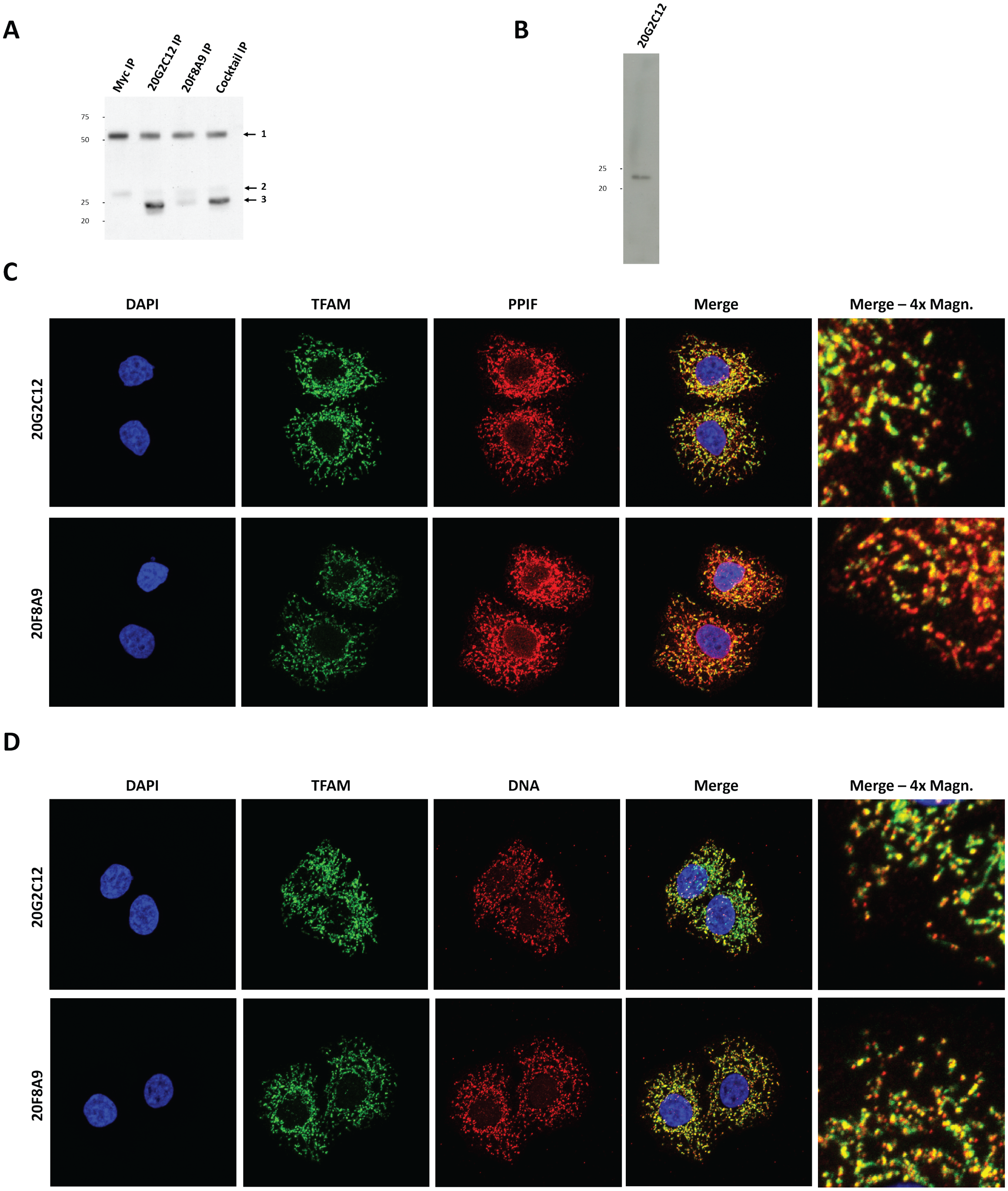


Figure 5.1: Characterization of TFAM monoclonal antibodies. (A) Immunoprecipitation of TFAM from cell lysates. HeLa cell lysate was applied to sheep anti-mouse Dynabeads conjugated to anti-Myc, 20G2C12 TFAM antibody, 20F8A9 TFAM antibody, or a 50/50 mixture of 20G2C12 and 20F8A9 TFAM antibodies The labeled bands are: 1) Antibody heavy chain; 2) antibody light chain; 3) TFAM. (B) Western blot using the 20G2C12 antibody detects a ∼23kDa band. (C and D) Immunocytochemistry showing TFAM localization. Mitochondria were identified by PPIF staining; mtDNA was identified by anti-DNA staining. There was no evidence for nuclear localization of TFAM using either antibody.

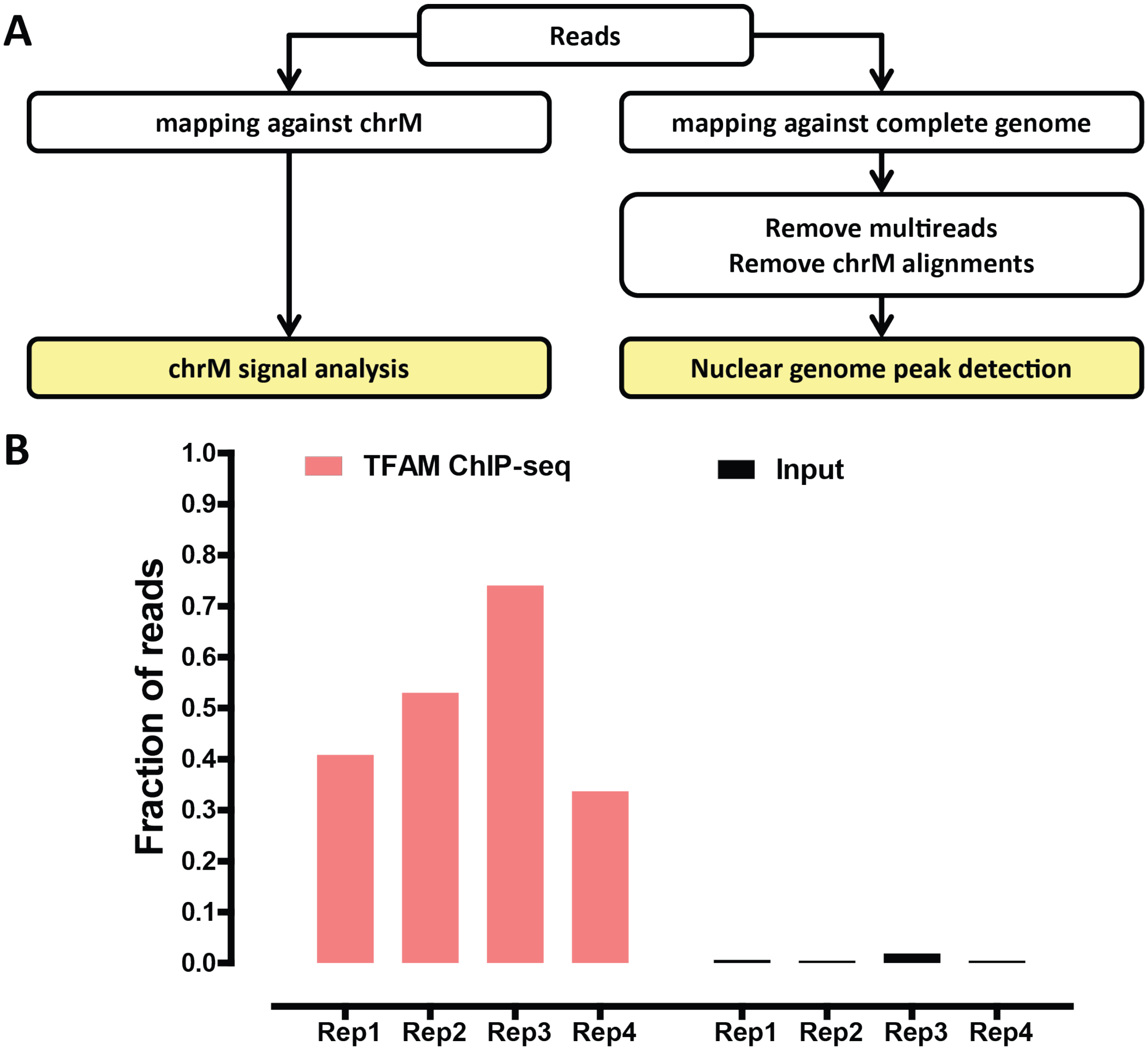


Figure 5.2: ChIP-seq analysis of genome-wide TFAM binding (A) Overview of computational processing of data. Reads were trimmed to 36 bp and then either mapped against the mitochondrial genome (chrM), or the complete hg19 version of the genome. After removing multireads and alignments to the mitochondrial genome, peaks in the nuclear genome were called using MACS2. (B) The proportion of sequencing reads mapping to chrM in ChIP and input datasets. All replicates of the ChIP-seq resulted in at least 30% of reads mapping to the mitochondrial genome, much greater than the 0.4-1.9% of reads mapping to mtDNA in the input datasets. Replicates 1-3 were performed using the 20G2C12 antibody, while Replicate 4 was performed using the 20F8A9 antibody.

et al., 2013). Furthermore, loss and damage to mtDNA has been implicated in cardiovascular disease (Sugiyama et al., 1991; Ide et al., 2001; Karamanlidis et al., 2010; Karamanlidis et al., 2011), diabetes (Maassen et al., 2004; Simmons et al., 2005; Gauthier et al., 2009), neurodegenerative disorders such as Alzheimers (Coskun et al., 2004; Coskun et al., 2012), and aging (Corral-Debrinski et al., 1992; Trifuvonic & Larsson, 2008). Strikingly, increasing mtDNA copy number promotes cell survival or function in many models of disease associated with decreased mtDNA abundance, such as diabetes (Gauthier et al., 2009; Suarez et al., 2008), aging (Hayashi et al., 2008), Alzheimer's (Xu et al., 2009), and Parkinson's (Keeney et al., 2009;

Piao et al., 2012). Thus, it is critical to understand how mtDNA copy number and integrity are maintained.

Mitochondrial transcription factor A (TFAM) is a DNA binding protein that plays multiple roles in regulating mtDNA function. As a sequence-specific transcription factor, it binds upstream of the light strand promoter (LSP) and heavy strand promoter 1 (HSP1) to activate initiation of transcription. At these sites, the footprint of TFAM binding is ∼22 bp long (Fisher & Clayton, 1998; Ngo et al., 2011). As a result, TFAM is essential for production of gene products from the mitochondrial genome. In addition, TFAM is required for normal mtDNA copy number, because RNA primers generated from LSP are used to prime mtDNA replication (Chang & Clayton, 1984; Chang & Clayton, 1985). Mice heterozygous for a knockout of TFAM exhibit not only an expected reduction (22%) in mitochondrial transcript levels in the heart and kidney, but also a universal 34% reduction in mtDNA copy number across all assayed tissues. Furthermore, homozygous knockout mice have no detectable levels of mtDNA and die during embryogenesis (Larsson et al., 1998), highlighting the importance of TFAM in maintenance of mtDNA levels and in cellular and organismal viability.

Apart from its sequence-specific functions, TFAM is thought to organize the mtDNA genome by coating it in a nonspecific manner. Although how TFAM packages mtDNA is not well-understood, it is known to bind nonspecifically to DNA (Fisher et al., 1989) and is estimated to be sufficiently abundant to coat the genome completely (Alam et al., 2003; Ekstrand et al., 2004; Kaufman et al., 2007). One model suggests that nonspecific binding radiates from the TFAM LSP binding site, which acts as a nucleation site for subsequent cooperative binding in a phased pattern to yield an inter-genome homogeneous pattern of binding (Fisher et al., 1992; Ghivizzani et al., 1994). The packaging function of TFAM appears to have important consequences for maintenance of the mtDNA genome. A TFAM variant that is deficient in transcriptional activation but competent in DNA binding is capable of preventing mtDNA depletion (Kanki et al., 2004). Therefore, as a prominent component of mtDNA nucleoids, TFAM appears to coat the mitochondrial genome, perhaps protecting it from turnover or deleterious damage.

Despite the importance of the associations of TFAM with mtDNA in the maintenance of mtDNA integrity and in cellular viability, these interactions have not been characterized in vivo. Therefore, to capture a high-resolution profile of TFAM-mtDNA interactions across the entire mitochondrial genome, we performed chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) (Johnson et al., 2007) for TFAM in human HeLa cells.

5.2 Results

5.2.1 Detection of TFAM-DNA interactions using ChIP-seq

To characterize TFAM binding to both the mitochondrial and nuclear genomes in an unbiased manner, we performed ChIP-seq targeting TFAM in HeLa cells. Because ChIP-seq data is highly dependent on the use of high-quality antibodies, generated two new TFAM monoclonal antibodies (20G2C12 and 20F8A9) that efficiently immunoprecipitated TFAM were generated and characterized (Figure 5.1A). Both of these antibodies gave clean mitochondrial and nucleoid signals in immunofluoresecence experiments with cultured HeLa cells (Figure 5.1C,D). The 20G2C12 antibody also performed well in Western blots of whole-cell lysates, recognizing a single protein band of ∼23 kD (Figure 5.1B).

Given the high efficiency of 20G2C12 in immunoprecipitating TFAM, as well as its high specificity, we used it to capture TFAM-associated DNA fragments for ChIP-seq analysis. DNA was sonicated prior to immunoenrichment and size-selected prior to library building so that the average fragment length of the final library was centered around 200 bp, a fragment distribution allowing for high-resolution deconvolution of binding events. We generated 3 replicates and matching controls. The sequencing depth of all samples was between 18 million and 48 million mappable reads, which is generally sufficient for comprehensive identification of transcription factor binding sites as shown before (Landt, Marinov & Kundaje, 2012).

A common concern with ChIP-seq datasets is the variability of enrichment for true binding events as compared to background. In a typical ChIP-seq experiment, a minority of sequencing reads originates from binding events, with the majority representing random genomic

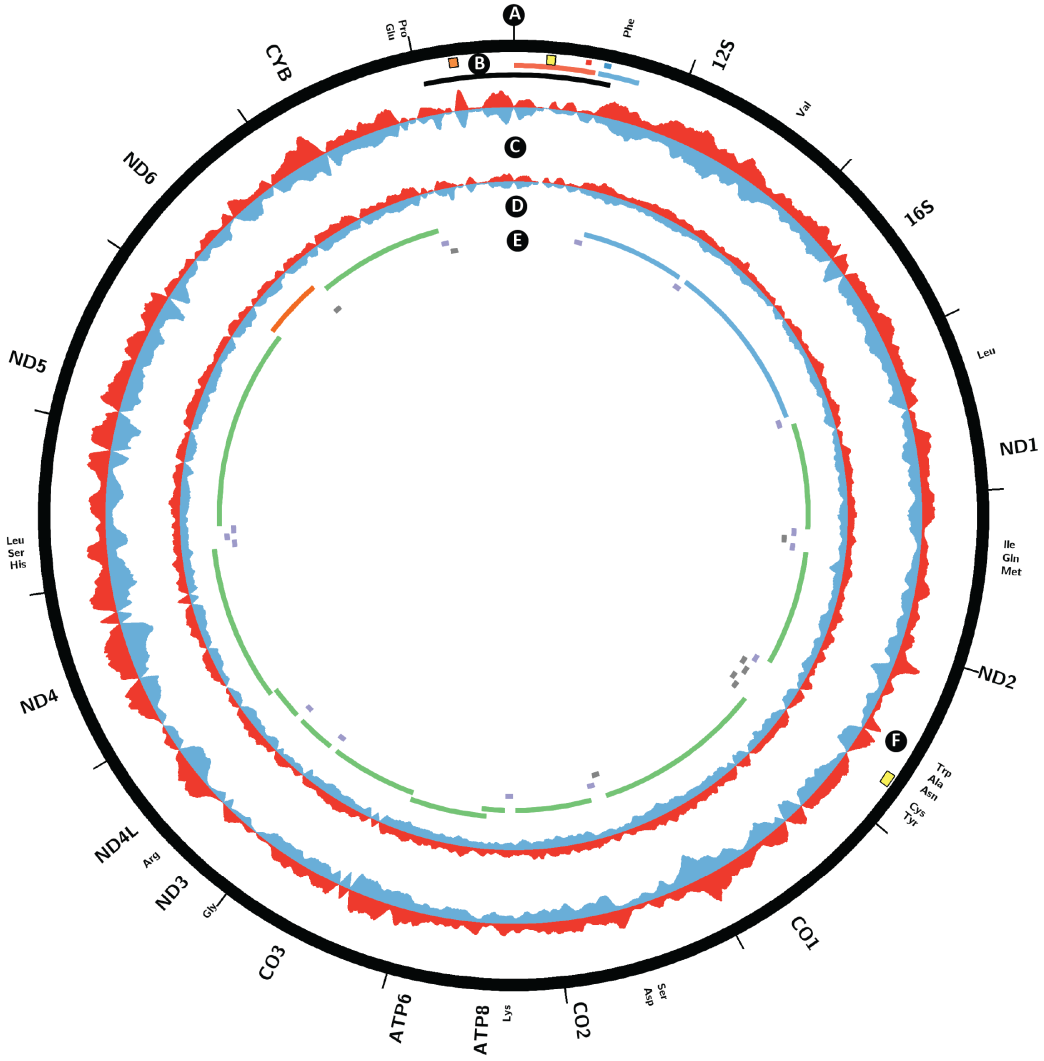


Figure 5.3: Coating of the mitochondrial genome by TFAM in HeLa cells. Circos plot of plus strand and minus strand TFAM ChIP-seq and input read density signal over chrM. (A, E) Annotation of protein coding (green on forward/heavy strand, red on reverse/light strand), ribosomal RNAs (yellow) and tRNAs (blue on forward/heavy strand, grey on reverse/light strand) transcripts. (B) D-loop (black), LSP promoter (large red tile), known LSP TFAM binding site (small red tile), HSP promoter (large blue tile), known HSP TFAM binding site (small blue tile), and origins of heavy strand replication (Ori-b, orange tile; O_H, yellow tile). (C) TFAM ChIP-seq signal on forward (red) and reverse (blue) strands. (D) Input signal on forward (red) and reverse (blue) strands. (F) Origin of light strand replication (yellow tile). Note that the input signal is exaggerated 60-fold relative to the ChIP-seq signal in order to visualize coverage irregularities. The signal from the TFAM ChIP-seq largely follows that of the input, indicating generalized binding across the mitochondrial genome.

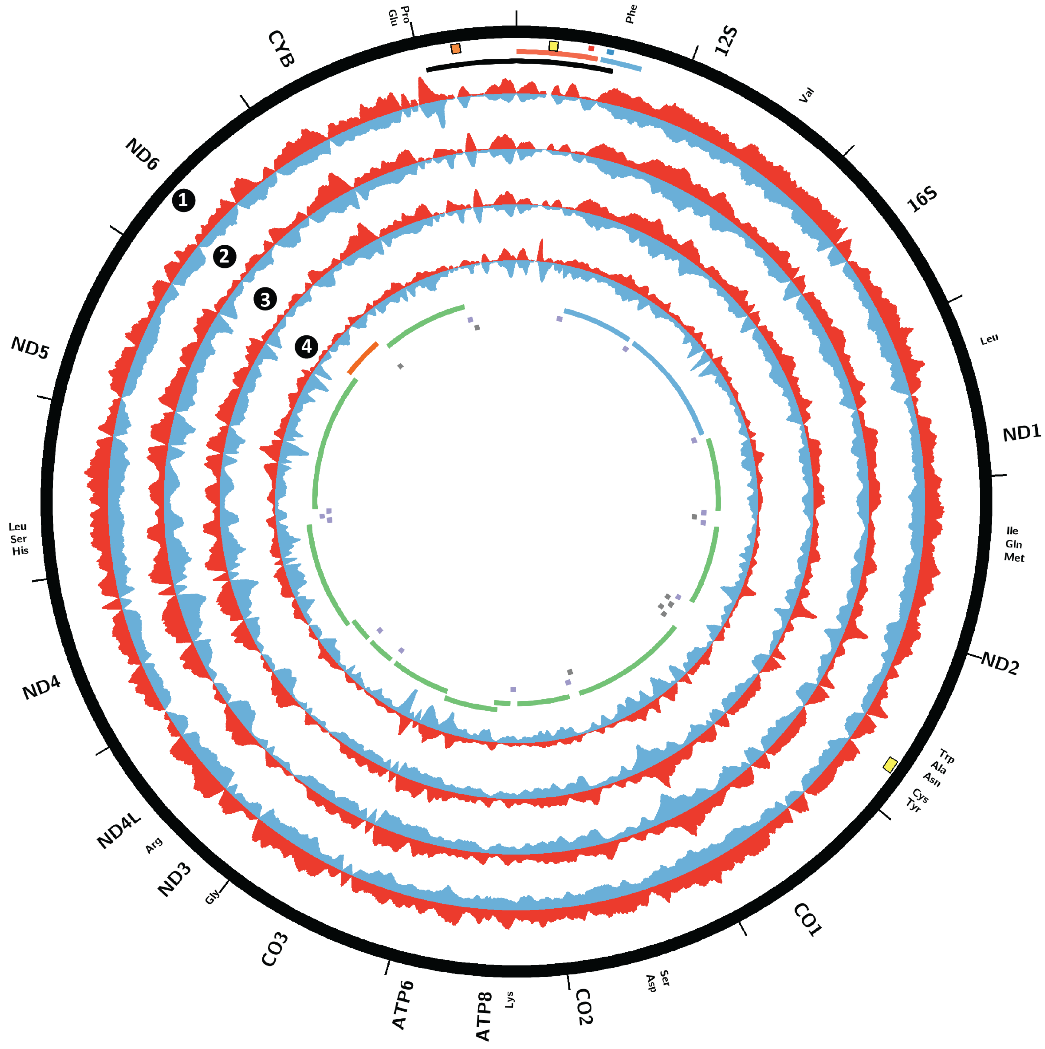


Figure 5.4: Comparison of profiles of TFAM binding to mitochondrial genome. Circos plots (A) of TFAM ChIP-seq experiments: (1) 20F8A9 antibody ChIP-Seq; (2) 20G2C12 replicate 1; (3) 20G2C12 replicate 2; (4) 20G2C12 replicate 3. Read profiles are very similar across replicates and antibodies.

DNA. Even for the same DNA binding factor, large variations in the strength of enrichment can be observed, and therefore it is critical to assess the degree of enrichment before downstream analysis. A number of ChIP-seq quality control metrics have been developed (Landt, Marinov & Kundaje, 2012) for nuclear transcription factors. However, TFAM is expected to bind to the mitochondrial genome, which has very different characteristics from the nuclear genome. In addition, it is predicted to bind both in the classical localized manner (Kharchenko et al, 2008) as well as broadly across the mitochondrial genome. As a result, metrics for evaluating nuclear transcription factors are not well-suited for analysis of TFAM binding data. We therefore examined the fraction of sequencing reads in our libraries mapping to the mitochondria as a proxy for the enrichment of TFAM binding events. Strikingly, between 30% and 75% of TFAM ChIP-seq reads

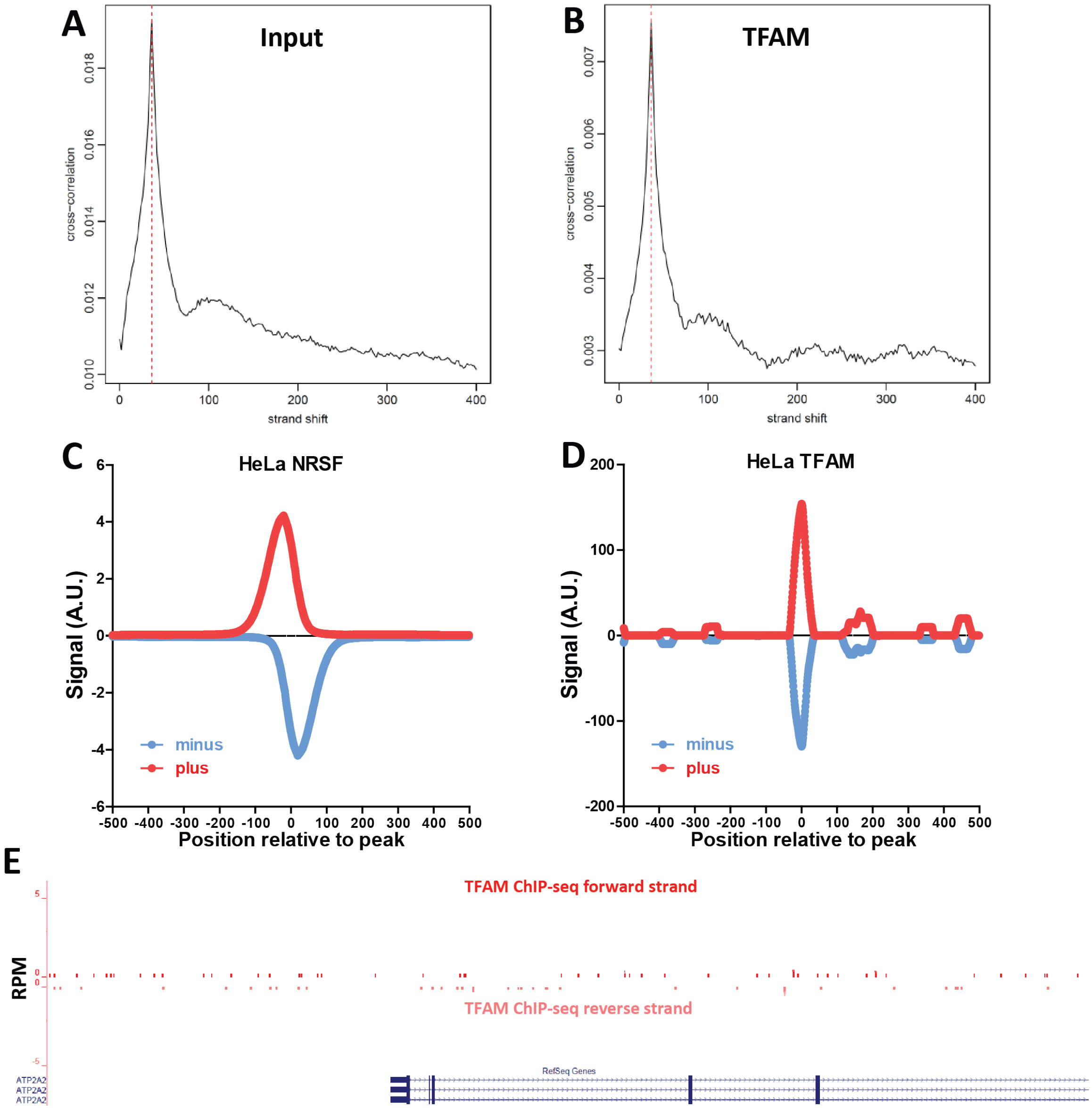


Figure 5.5: Absence of TFAM binding to the nuclear genome. (A) Cross-correlation plot of input DNA computed over the nuclear genome. (B) Cross-correlation plot of TFAM ChIP-seq computed over the nuclear genome. (C) Distribution of ChIP-seq reads mapping to the plus and minus strand around called binding sites in a ChIP-seq dataset for the NRSF transcription factor (Schoenherr & Anderson, 1995) in HeLa cells, generated by the ENCODE consortium (ENCODE Project Consortium, 2011, ENCODE Project Consortium, 2012). (D) Distribution of TFAM ChIP-seq reads mapping to the plus and minus strand around called binding sites indicates lack of real binding sites. (E) No ChIP-seq enrichment around the promoter of the SERCA2/ATP2A2 gene, previously suggested to be a TFAM target.

mapped to the mitochondrial genome, while less than 2% of reads mapped to the mitochondrial genome in the input samples, indicating that our TFAM ChIP-seq datasets are indeed highly enriched for TFAM binding events (Figure 5.1B). We note that 75% ChIP enrichment is extremely high (in fact, practically unprecedented) for any transcription factor dataset (Landt, Marinov & Kundaje, 2012), thus underscoring the high experimental quality of our datasets.

Because partial copies of the mitochondrial genome are also present in the nuclear genome, not all reads originating from mtDNA can be mapped uniquely. Therefore, we characterized TFAM binding to mtDNA and to the nuclear genome separately. We analyzed mitochondrial binding events by aligning sequencing reads to the mitochondrial genome alone (restricting our analysis to reads mapping perfectly without any mismatches to further increase mapping accuracy), and analyzed binding to the nuclear genome by aligning only the reads which did not map to the mitochondrial genome, as outlined in Figure 5.2A. For a standard nuclear transcription factor, this approach may cause some reads originating from the nuclear genome to artificially map to the mitochondrial genome. However, given that TFAM is known to bind to the mitochondrial genome and the extremely high enrichment for TFAM binding to mtDNA in our TFAM ChIP-seq libraries, this should not be a significant confounding factor.

5.2.2 TFAM coats the mitochondrial genome

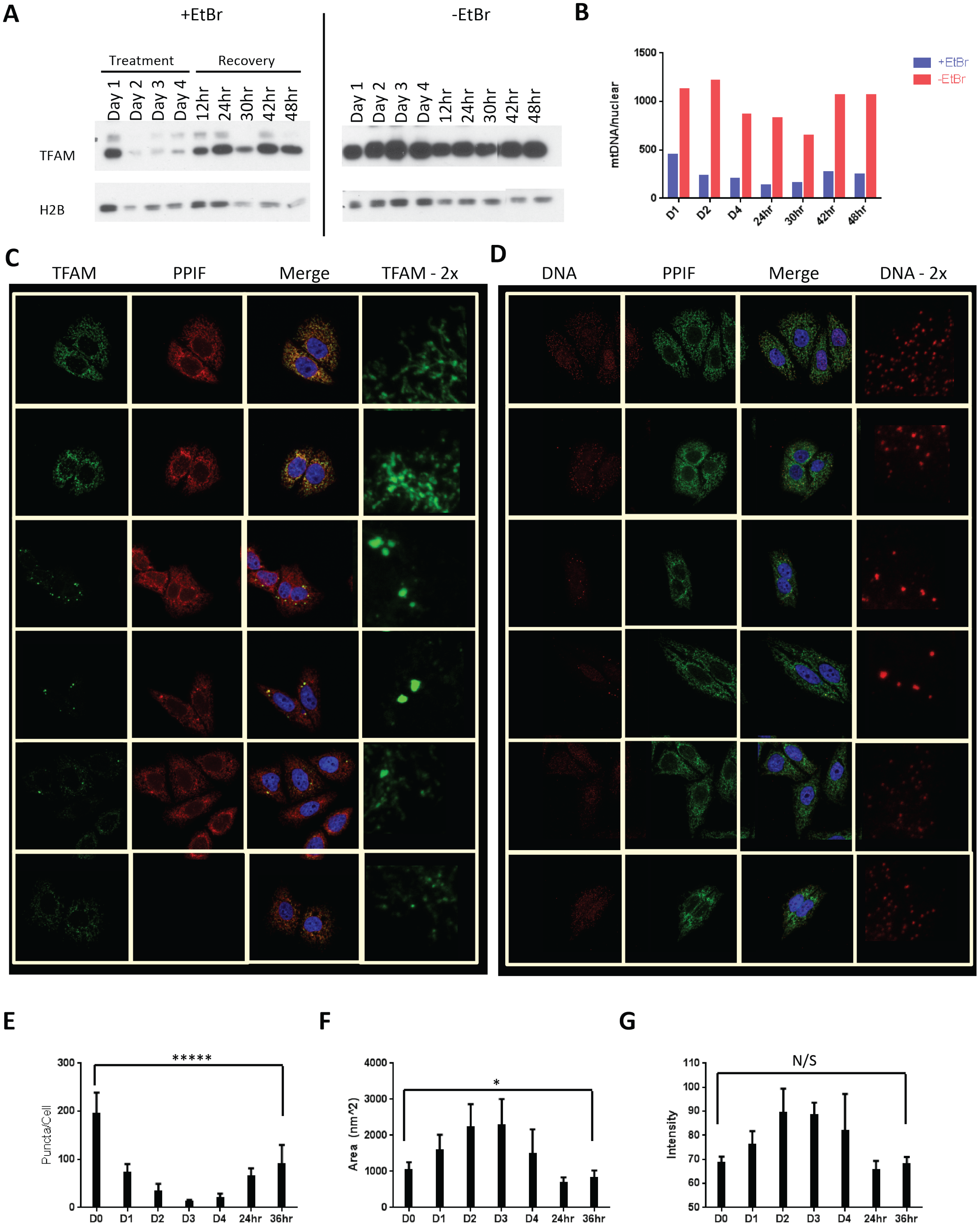
As discussed above, TFAM has not only been proposed to bind specifically to well-defined binding sites in the D-loop, but has also been suggested to play a nonspecific packaging role in the nucleoid that is essential for mtDNA integrity. However, little is known about the pattern of non-specific binding of TFAM to the mitochondrial genome. Localized binding at the D-loop and diffuse binding across the rest of the genome are expected to result in distinct ChIP-seq signal profiles. Localized, "point-source" binding to DNA results in an asymmetric distribution of reads mapping to the forward and reverse strand around the binding site of the protein (Kharchenko et al, 2008, Pepke et al, 2009), while diffuse binding does not produce such strand asymmetry.

To characterize TFAM binding to mtDNA, we examined the forward and reverse strand read distribution after mapping TFAM ChIP-seq and input library reads to the mitochondrial genome. Strikingly, we did not observe regions of obvious enrichment and strand asymmetry in the D-loop; in particular, we did not see specific binding at the predicted HSP1 and LSP sites. On the whole, the TFAM ChIP-seq signal was broadly distributed over the whole mitochondrial chromosome, and while coverage was not perfectly uniform, the amplitude of the non-uniformity was not significant and the signal profile closely tracked that of the input sample (Figure 5.3). The low level of non-uniformity likely results from sequencing biases, which has been documented to skew coverage (Dohm et al., 2008; Ross et al., 2013). Because our libraries were carefully size-selected for fragments in the 200 bp range, discrete TFAM binding sites would be expected to yield discrete signal localizations. Therefore, we interpret these results as evidence for the uniform coating of the whole mitochondrial genome by TFAM. We observed one region of apparent localized enrichment exhibiting strand asymmetry in the ND2 ORF near the origin of light strand replication (O_L) (Figure 5.2F), which we discuss in the Discussion section.

To further verify our results, we carried out ChIP-seq against TFAM with a second TFAM monoclonal antibody, 20F8A9. We obtained similar results (Figure 5.4) and found significant correlation between the 20F8A9 dataset and the three datasets obtained from the 20G2C12 antibody datasets ($p < 0.0001$).

5.2.3 No evidence for binding to the nuclear genome

Previous studies have suggested that TFAM can be found in the nucleus and that it modulates the transcription of nuclear genes. In rat neonatal cardiac myocytes, TFAM was found to bind to the promoter of *SERCA2*, the homolog of human sarco(endo)plasmic reticulum calcium-ATPase 2 (*ATP2A2*), and was implicated in regulating its transcription (Watanabe et al, 2011). Given the extremely high degree of TFAM binding enrichment in our datasets, any robust nuclear TFAM binding events should be readily detectable. To analyze nuclear binding, we excluded all sequencing reads mapping to the mitochondrial genome and used the resulting set of reads to identify putative TFAM binding sites. We first looked for significant global read clustering using cross-correlation between reads mapping to the forward and the reverse DNA strands (Kharchenko et al, 2008, Landt, Marinov & Kundaje, 2012). Cross-correlation plots for input samples and for TFAM ChIP-seq datasets were indistinguishable from each other (Figure 5.5A,B). Next, we called putative TFAM binding sites using MACS2 (Zhang et al, 2008). Using


A
+EtBr
-EtBr
Treatment
Recovery
Day 1
Day 2
Day 3
Day 4
12hr
24hr
30hr
42hr
48hr
TFAM
H2B
B
1500
1000
500
0
mtDNA/nuclear
+EtBr
-EtBr
D1
D2
D4
24hr
30hr
42hr
48hr
C
TFAM
PPIF
Merge
TFAM - 2x
D
DNA
PPIF
Merge
DNA - 2x
E
*****
300
200
100
0
Puncta/Cell
D0
D1
D2
D3
D4
24hr
36hr
F
*
4000
3000
2000
1000
0
Area (nm^2)
G
N/S
110
100
90
80
70
60
50
Intensity

default settings (corresponding to a q-value cutoff of 10^{-2}), we identified 72, 137 and 153 sites respectively for the three replicates generated with antibody 20G2C12, and a single site for the 20F8A9 antibody. However, manual inspection of each of the identified sites revealed that all were likely to represent artifacts, mostly associated with repetitive DNA sequences, as none had the expected strand asymmetry of read distribution around a binding site. Instead, the two strand profiles at each site were identical (summarized in Figure 5.5D, with the classic nuclear transcription factor NRSF shown for comparison in Figure 5.5C), and numerous unmappable regions and repetitive elements were present in the immediate vicinity of many of the called sites. Inspection of the *ATP2A2* gene revealed no TFAM enrichment neither in the promoter region nor anywhere else in the neighborhood of the gene (Figure 5.5E). Furthermore, we did not detect nuclear localization of TFAM in our cells (Figure 5.1C). Therefore, in HeLa cells under normal growth conditions, we find no evidence for specific binding of TFAM to nuclear target genes.

5.3 Discussion

Previous in vitro studies have suggested that TFAM binds specifically to LSP and HSP1, and that it may also bind nonspecifically in a phased manner. Furthermore, evidence has been presented for its nuclear localization and action as a canonical nuclear transcription factor in rat neonatal cardiac myocytes. However, no direct genome-wide measurements of TFAM binding have been previously reported. Our TFAM ChIP-seq data reveal very high enrichment for reads mapping to the mitochondrial genome, but a binding pattern that largely mirrors the read distribution observed in the input DNA, suggesting broad, non-specific binding to mitochondrial genome. This pattern is highly reproducible, indicating that the average population-wide state of TFAM-mtDNA interactions is stable. We found no correlation between irregularities in TFAM signal distribution and characteristics of the mitochondrial genome such as GC content (data not shown). Thus, our conclusion is that TFAM binds to the mitochondrial genome nonspecifically and without bias when cells are grown under typical culture conditions. Although we did not observe the synchronized phased binding seen in in-vitro studies, we cannot rule out a model where individual mtDNAs have such a pattern of binding initiating from a non-universal nucleation site.

Strikingly, we did not observe localized enrichment of binding at the known LSP and HSP1 TFAM binding sites. The ChIP-seq signal pattern mirrored that of the input in these regions, and no ChIP-seq peaks displaying the canonical strand asymmetry in read distribution were observed. This finding can be explained by a model in which the interaction of TFAM with the LSP and HSP1 binding sites is relatively transient and infrequent compared to a more stable non-specific association with the genome in its packaging state.

We did detect one site in the genome exhibiting the characteristics of a specific, localized ChIP-seq peak, centered at 5175 bp in the ND2 ORF. The localized nature of the ChIP signal at this site suggests higher occupancy of TFAM. This peak localizes to 546 bp upstream of the O_L. Strikingly, TFAM has previously been localized 520 bp upstream of the O_L of rat mtDNA (Gadaleta et al., 1996; Cingolani et al., 1997; Pierro et al., 1999). We found no sequence similarity between the rat and human sites, and in general this region of the mtDNA genome shows low homology between the two species. Further work will be required to understand the signifi-

Figure 5.6 *(preceding page)*: Cells treated with 50ng/ml EtBr experience rapid depletion of TFAM levels, as assayed by anti-TFAM Western blot (A). This coincides with a depletion of mtDNA levels to 17% that of wildtype after 4 days of treatment, as determined by relative qPCR quantification (B). Removal of EtBr leads to a rapid increase in TFAM levels and to an increase in mtDNA copy number per cell within 30 to 48 hours. Immunohistochemistry of HeLa cells for TFAM, mitochondrial matrix protein PPIF, and DAPI show that TFAM is mitochondrial under wildtype conditions (C). Treatment with EtBr leads to a remarkable consolidation of both TFAM and mtDNA puncta (C, D), leading to larger, fewer nucleoids (E-G). By 24 to 36 hours post-recovery, nucleoids redistribute uniformly throughout the mitochondrial network, with partial recovery of nucleoid size, intensity, and number per cell.

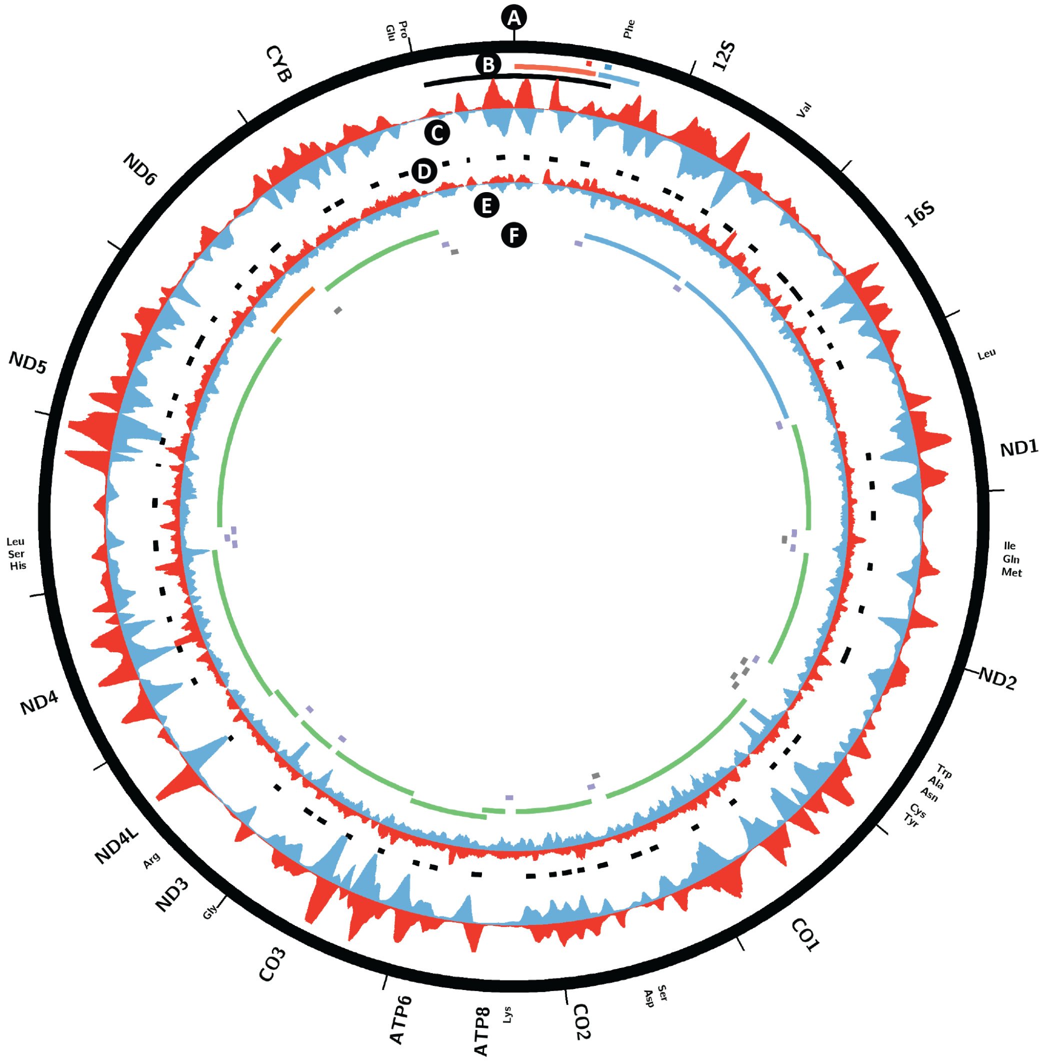


Figure 5.7: Discrete localization of TFAM on the mitochondrial genomes following mtDNA depletion after EtBr treatment in HeLa cells. Circos plot of plus strand and minus strand TFAM ChIP-seq and input read density signal over chrM. (A, F) Annotation of protein coding (green on forward/heavy strand, red on reverse/light strand), ribosomal RNAs (yellow) and tRNAs (blue on forward/heavy strand, grey on reverse/light strand) transcripts. (B) D-loop (black), LSP promoters (large red tile), known LSP TFAM binding site (small red tile), HSP promoter (large blue tile) and known HSP TFAM binding site (large blue tile). (C) TFAM ChIP-seq signal on forward (red) and reverse (blue) strands. (D) Manually determined localized TFAM binding sites (black tiles). (E) Input signal on forward (red) and reverse (blue) strands. Note that the input signal is greatly exaggerated relative to the ChIP-seq signal (Fig. 1B) in order to visualize coverage irregularities.

cance of this putative TFAM binding site.

Finally, analysis of all datasets for TFAM binding to the nuclear genome yielded no hits distinguishable from common ChIP-seq artifacts. Although Watanabe et al. observed regulation of the *SERCA2* gene in rat myocytes, we

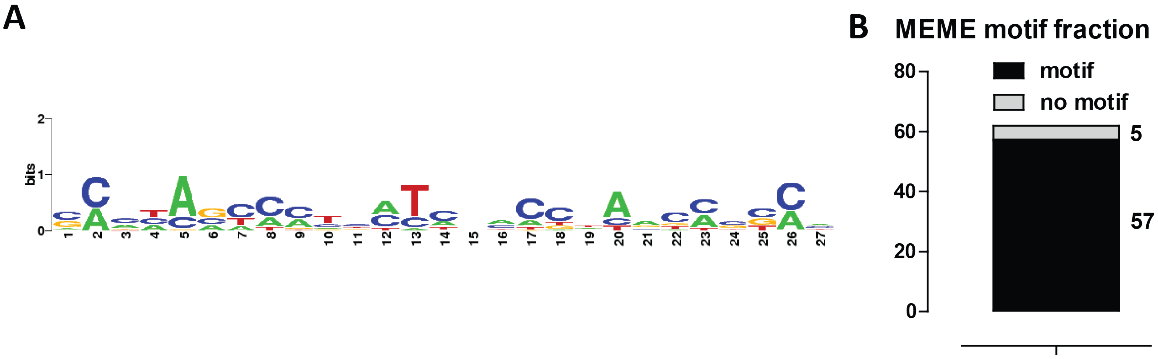


Figure 5.8: MEME-derived motif for TFAM enriched sites. (A) Sequence logo. (B) Fraction of sites the motif is found in.

did not detect TFAM binding at the promoter of its ortholog in humans. Previous studies have shown nuclear localization of TFAM in rat hepatoma cells (Dong et al., 2002), as well as an alternate isoform of TFAM in mouse testis nuclei (Larsson et al., 1996). We have thus far been unable to detect nuclear TFAM localization in HeLa cells (Figure 5.1C), suggesting that nuclear localization and transcriptional regulation may be cell type or perhaps species-dependent. ChIP-seq in different cell lines may be able to detect such nuclear interactions.

In this study, we presented the first in vivo ChIP-seq analysis of TFAM binding to the mitochondrial genome. Aside from generalized, largely non-specific binding across the mitochondrial genome, we detected a putative specific binding site upstream of the origin of light strand replication. We did not observe the expected binding at the known HSP1 and LSP sites, nor did we identify any nuclear binding sites. An area that remains to be explored is the dynamic nature of TFAM-DNA interactions with respect to both the nuclear and mitochondrial genomes. ChIP-chip on the yeast mitochondrial genome has shown that metabolic changes can lead to differential binding of the yeast TFAM homolog, Abf2p (Kucej et al., 2008). It is possible that such remodeling also occurs in the mammalian system, and further studies will provide insight into the dynamic nature of the mtDNA-protein interactions within the nucleoid that serve to protect its integrity. Some speculative results we obtained from experiments addressing these issues that are also relevant to later chapters are presented as an additional section at the end of this chapter.

5.4 Materials and Methods

5.4.1 Cell growth and treatment

HeLaS3 cells were cultured in Dulbeco's modified Eagle's medium (DMEM, Invitrogen #11995) containing 10% bovine serum (Invitrogen #16170), penicillin and streptomycin, and additional L-glutamine (2mM). Cells were fed 24 hours before harvest for ChIP-seq, which was performed at 80-90% confluency.

5.4.2 Antibody Production and characterization

Antibodies were produced by the Caltech Monoclonal Antibody Facility and raised against the full-length TFAM protein in mouse. Immunoprecipitation with 20G2C12 and 20F8A9 TFAM antibodies and Myc antibody (Santa Cruz #sc-40) was performed according to established protocols using M-280 sheep anti-mouse Dynabeads (Invitrogen #11201D). Immunoblotting of IP products was performed using a monoclonal TFAM 18G102B2E11 antibody, also custom generated, at 1:2000, with goat anti-mouse HRP antibody (1:10,000, Jackson ImmunoResearch #115-056-003). Immunoblotting of HeLa whole cell lysate with 20G2C12 was performed at a 1:200 dilution and with goat anti-mouse HRP antibody.

5.4.3 Immunocytochemistry

HeLa cells cultured as described above were plated onto poly-lysine coated glass coverslips 48 hours prior to fixation in formaldehyde and permeabilization with 0.1% Triton X-100. For colo-

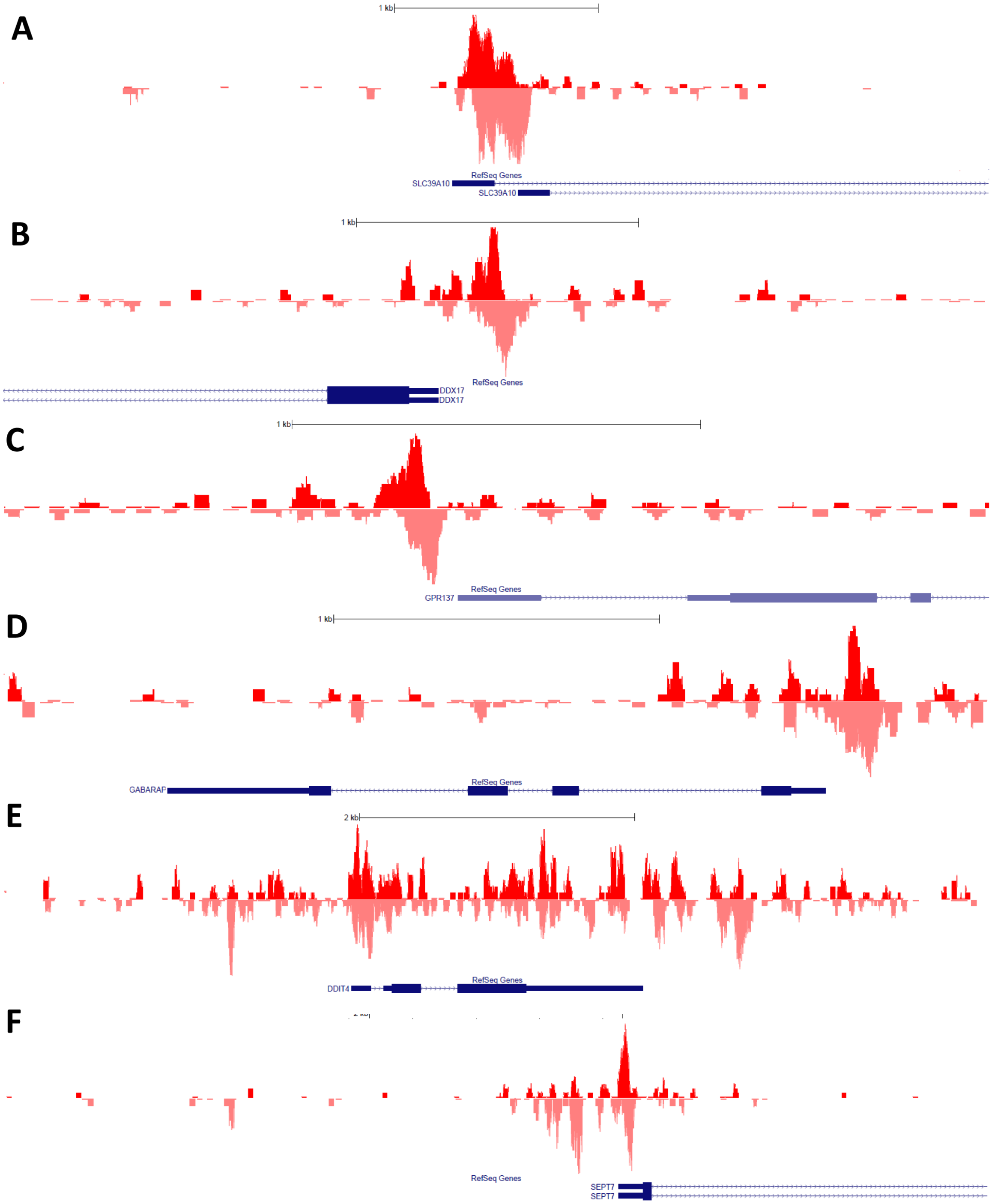


Figure 5.9: The few nuclear genome locations with TFAM ChIP-seq signal characteristics similar to those of robust ChIP-seq peaks. (A) *SLC39A10* (B) *DDX17* (C) *GPR137* (D) *GABARAP* (E) *DDIT4* (F) *SEPT17*.

calization of TFAM to mitochondria, 20G2C12 or 20F8A9 antibodies were used at 1:10 in conjunction with PPIF at 1:200 (ProteinTech #18466-1-AP). Secondary antibodies were goat anti-mouse AF488 (1:500, Invitrogen #A11001) and donkey anti-rabbit AF546 (1:500, Invitrogen

#A10040). Cells were also stained with DAPI to visualize nuclei. Immunocytochemistry to visualize colocalization of mitochondrial nucleoids and TFAM was performed sequentially due to both antibodies being raised in mouse. Sequential immunostaining yielded no background fluorescence due to cross-antibody reactivity (data not shown). Order was as follows: anti-TFAM antibody (1:10); goat anti-mouse AF488 (1:500, Invitrogen #A11001); anti-DNA antibody (1:25, Millipore #CBL186); goat anti-mouse AF555 (1:500, Invitrogen #A21426), DAPI. Images were acquired with a Zeiss LSM 710 confocal microscope with PlanApochromat 63X/1.4 oil objective. Z-stack acquisitions were converted to maximum z-projections using ImageJ software.

5.4.4 Chromatin immunoprecipitation and sequencing

ChIP experiments and preparation of DNA for sequencing were performed following standard procedures (Johnson & Mortazavi et al, 2007) with some modifications. Cells were fixed for 10min at RT in 1% formaldehyde, harvested using a cell scraper, washed once in ice-cold PBS, and resuspended in RIPA buffer with protease inhibitor. The sample was then sonicated using a 3.2mm microtip (QSonica Sonicator 4000) at 30s on/30s off intervals and 40% amplitude for 180min while in a −30 °C 3:1 isopropanol and water bath containing dry ice. Subsequent steps were performed as per the standard protocol. DNA was size-selected during library building to an average fragment size of 200bp. Libraries were sequenced using Illumina GAIIx and Illumina HiSeq 2000. Sequencing data is available under GEO accession record GSE48176.

5.4.5 Sequencing data processing and analysis

Sequencing reads were trimmed down to 36 bp and then mapped against either the female set of human chromosomes (excluding the Y chromosome and all random chromosomes and haplotypes) or the mitochondrial genome alone, using the hg19 version of the human genome as a reference. Bowtie 0.12.7 (Langmead et al. 2009) was used for aligning reads, not allowing for any mismatches between the reads and the reference. ChIP-seq peaks were called using MACS2 (Zhang et al. 2008) with default settings except for the mfold parameter which was lowered to (2,30). Circos plots were generated using Circos version 0.60 (Krzywinski et al 2009). Additional data processing was carried out using custom-written python scripts. ENCODE data was downloaded from the UCSC browser (`http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs`) and its use here complies with its terms of usage. Pearson correlation coefficient, t-test, and p values were calculated using embedded and custom Microsoft Excel functions.

5.5 Possible discrete localization of TFAM to the mitochondrial genome following mtDNA depletion and recovery after EtBr treatment

Previous studies have demonstrated that treatment of cells with ethidium bromide (EtBr) selectively causes depletion of mtDNA (Zylber et al. 1969; Desjardins et al. 1985; King & Attardi 1989; Micol et al. 1997; Herzberg et al. 1993). Furthermore, we found that EtBr also results in marked depletion of TFAM at the protein level as measured by Western blot (Figure 5.6A,B). We investigated this phenomenon further by examining the dynamics of mitochondrial nucleoid morphology during an EtBr treatment time course. We observed a remarkable consolidation of both mtDNA and TFAM localization within the first 24 hours, resulting in the formation of large, bright nucleoids (Figure 5.6C-G). These changes were concomitant with depletion of TFAM and mtDNA and stabilized by the 4th day of treatment. Upon withdrawal of EtBr, nucleoid morphology and TFAM and mtDNA levels were partially restored. At 36 hours after withdrawal, mtDNA and TFAM levels are still significantly lower than in untreated cells but mitochondrial replication is observed again.

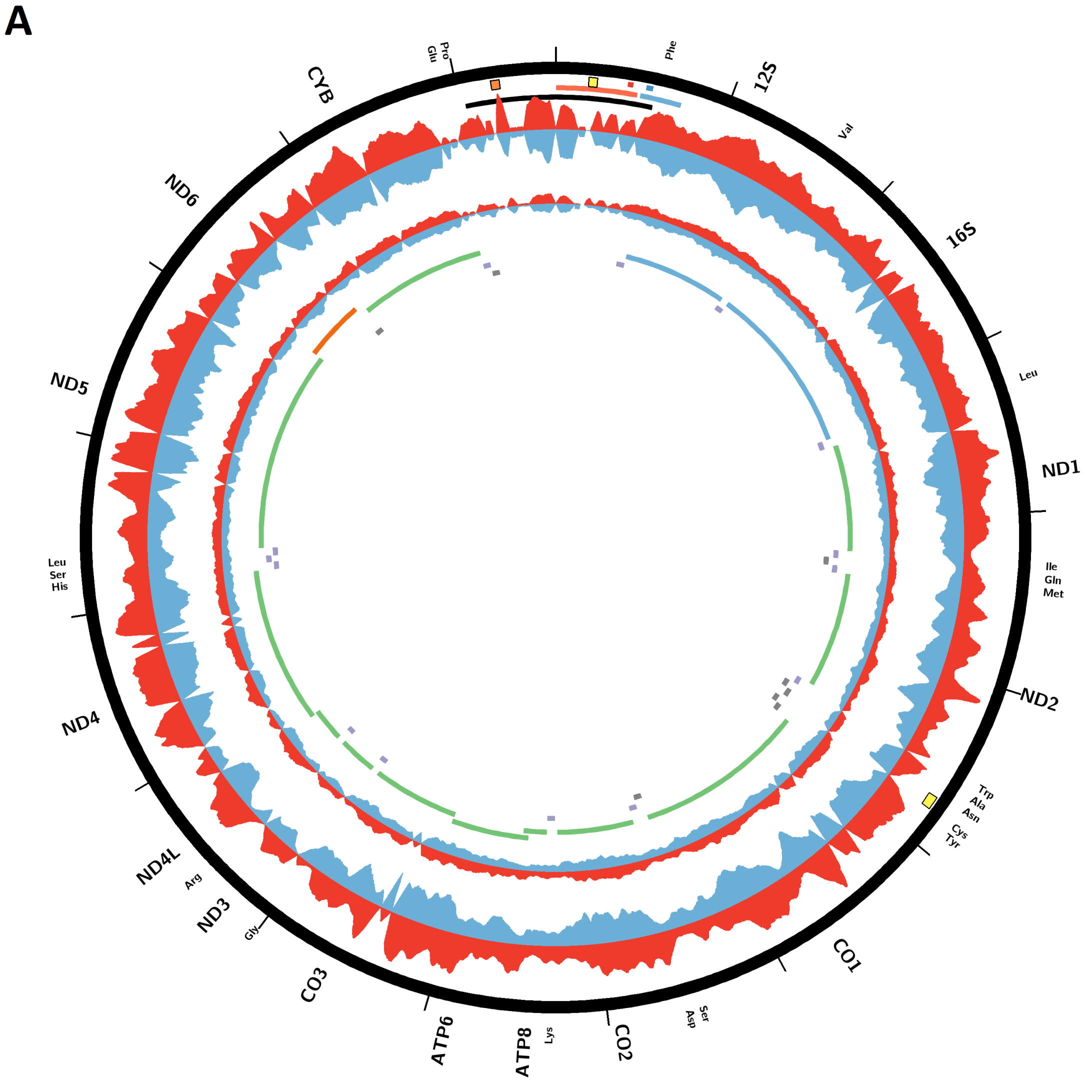
To elucidate the dynamics of TFAM binding to mtDNA concomittant with these changes, we performed TFAM ChIP-seq at 36 hours upon withdrawal of EtBr after 4 days of treatment. In contrast with the uniform TFAM distribution over the mitochondrial genome in untreated cells, we observed clear foci of TFAM localization, with the strand asymmetry characteristic of robust ChIP-seq transcription-factor binding peaks (Figure 5.7C,D). Due to the extremely high coverage of the mitochondrial genome by sequencing reads, peak calling using standard publicly available packages such as MACS, ERANGE (Johnson & Mortazavi et al. 2007), and SPP (Kharchenko et al, 2008) was not successful in resolving individual peaks, so we curated TFAM binding foci manually following the criteria that forward and reversed strand signal peaks should be separated by a distance related to the input fragment distribution (determined using cross-correlation analysis and the BioAnalyzer profiles for the libraries). We identified 66 high confidence enrichment foci in this manner (Figure 5.7D).

These observations raise the question of why and how TFAM would localize to narrowly defined loci after mtDNA depletion. The changes in TFAM binding could be due to a combination of sequence specificity of binding, the local structure of DNA, the global state of the mitochondrial nucleoid, and possible TFAM protein interactors with other protein with DNA binding affinity on their own. We attempted to derive enriched DNA sequence motifs from the 100bp regions flanking the summit of TFAM binding sites using MEME (Bailey et al, 2009). A loosely constrained, long (26bp), almost symmetric, C- and A-rich motif emerged from this analysis (Figure 5.8), and was found in the majority (57 of 66) of sites. However, we were not able to detect differential affinity of TFAM for the best and worst matches for this motif among the sites using fluorescence anisotropy (data not shown).

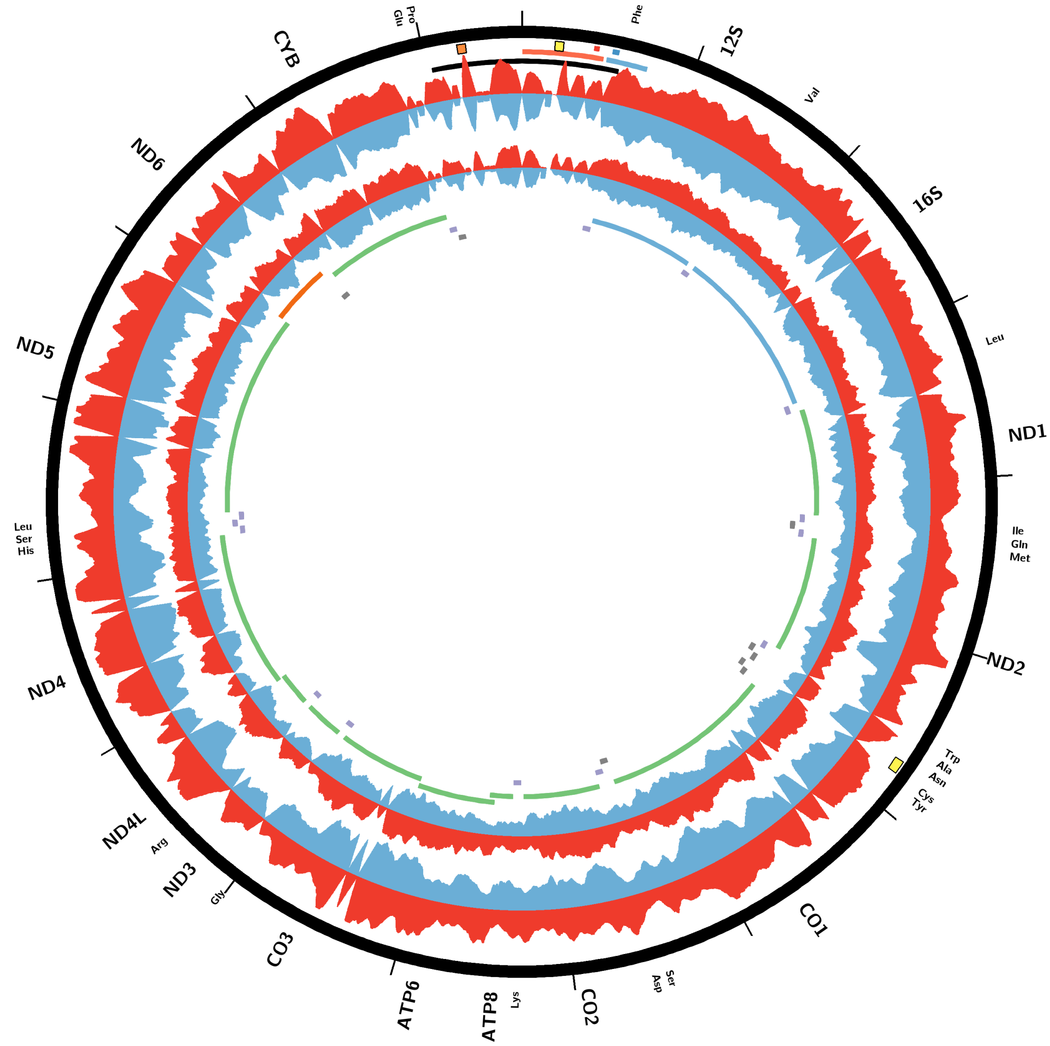
We also reexamined the question of whether high-confidence TFAM ChIP-seq peaks can be identified in the nuclear genome, this time using the EtBr-treated cell dataset. Several peaks much stronger than those seen in untreated cells were identified (Figure 5.9) and were associated with the promoters of the *SLC39A10*, *DDX17*, *GPR137*, *GABARAP*, *DDIT4* and *SEPT17* genes. As these genes are not obviously related to mitochondrial function and because these peaks are not extremely strong compared to conventional transcription factor ChIP-seq peaks, they cannot be confidently considered instances of functional binding before more studies are performed.

The observation of highly localized TFAM binding to mtDNA following EtBr treatment is highly intriguing, and potentially of great significance, but it was based on a single ChIP-seq experiment, not the multiple replicates one would like to have. Naturally, we invested a lot of effort in replicating the result, repeating the experiment three times, but in all three cases the resulting TFAM ChIP-seq pattern was much closer to the one observed in resting cells than to the initial EtBr replicate (Figure 5.10), even though the extent of TFAM localization to discrete loci varied between different replicates (compare Figure 5.10B with Figure 5.10C).

It is unlikely that the original EtBr observation was an artifact given the quality characteristics of that dataset. TFAM in those cells indeed localized to discrete loci after EtBr treatment and this was measured by the assay. Our explanation for the failure to replicate the result is that the discrete localization happens only at particular times during the EtBr time course and the dynamics of the time course varies between experiments, i.e. the same sequence of events happens each time the experiment is done but at different times. As a result the point in time at which TFAM is highly localized shifts in time relative to the cell harvesting time point. We were unfortunate not to capture that moment when we repeated the experiments, and to eventually run out of time and resources before we could succeed. However, more recent completely orthogonal observations that we made quite strongly suggest that TFAM localization to those sites of the mitochondrial genome is indeed real. They are discussed in more detail in a later chapter.


A
Phe
12S
Val
16S
Leu
ND1
Ile
Gln
Met
ND2
Trp
Ala
Asn
Cys
Tyr
CO1
Ser
Asp
CO2
Lys
ATP8
ATP6
CO3
Gly
ND3
Arg
ND4L
ND4
Leu
Ser
His
ND5
ND6
CYB
Pro
Glu

B


Phe
12S
Val
16S
Leu
ND1
Ile
Gln
Met
ND2
Trp
Ala
Asn
Cys
Tyr
CO1
Ser
Asp
CO2
Lys
ATP8
ATP6
CO3
Gly
ND3
Arg
ND4L
ND4
Leu
Ser
His
ND5
ND6
CYB
Glu
Pro

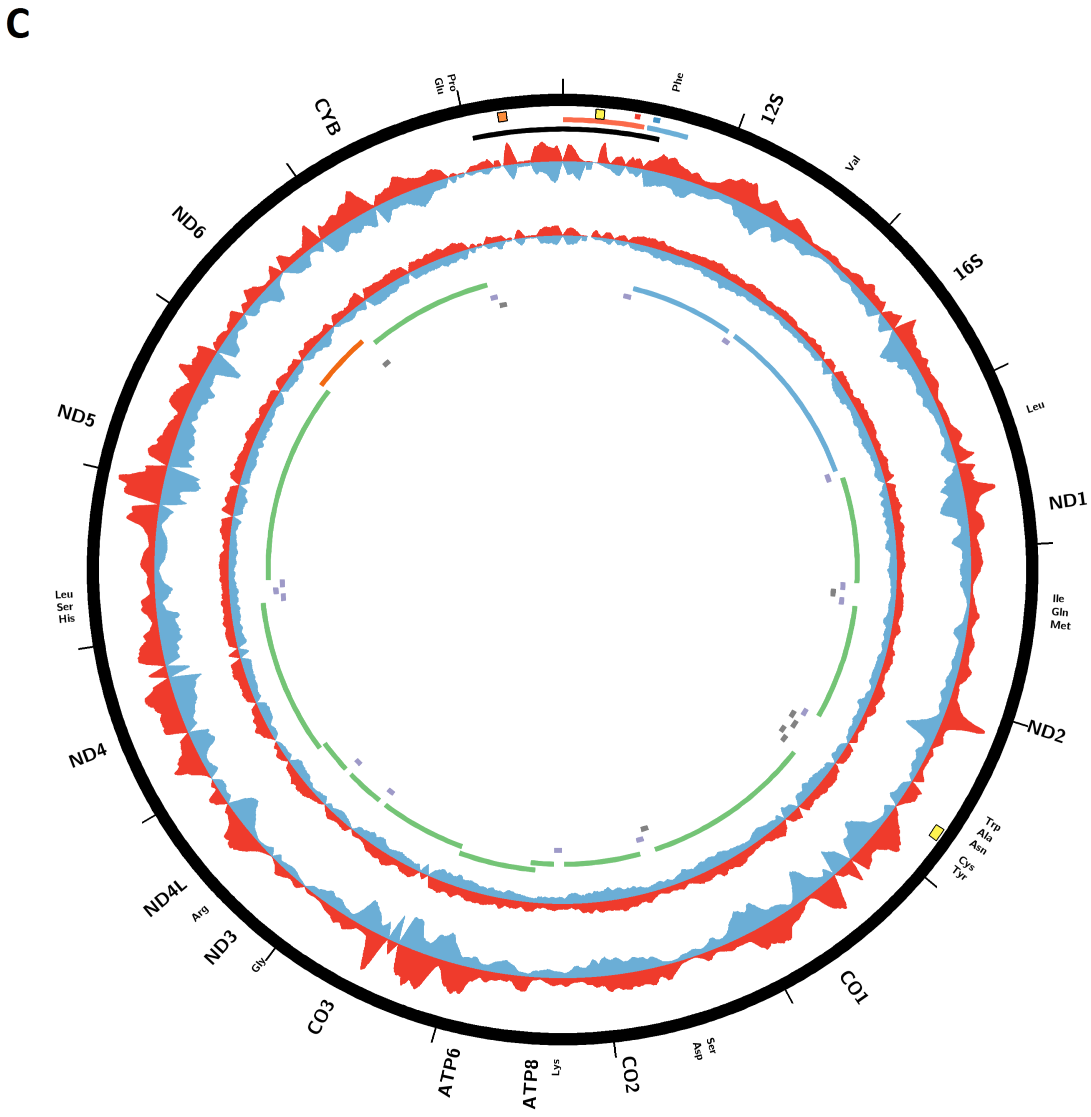


Figure 5.10: Replication of TFAM EtBr ChIP-seq results. (A,B,C) Three independent biological replicate TFAM ChIP-seq datasets generated after treatment of HeLa cells with EtBr.

6

Genome-Wide Analysis of the Human Mitochondrial Transcription and Replication Machineries

The material in this chapter is based on data generated in the Myers lab at the HudsonAlpha Institute of Biotechnology, and is a preliminary version of what is intended for publication in the future as:

Marinov GK, Wang YE, Pauli-Behn F, Newberry K, Chan DC, Myers RN, Wold BJ. Genome-Wide Analysis of the Human Mitochondrial Transcription and Replication Machineries

Abstract

Mitochondria are vital to eukaryote biology organelles of endosymbiotic origin containing their own (albeit highly reduced) genome. Mitochondrial transcription and replication are regulated and carried out by a dedicated set of proteins. These processes have been studied in most detail in mammalian mitochondria; however, the genome-wide occupancy of most of the factors involved has so far not been characterized. Here, we report global maps of the distribution of the transcriptional regulators TFB2M and TFAM, the transcription termination factor MTERF, and the mitochondrial RNA and DNA polymerases POLRMT and POLG. These data allow us to evaluate the mechanistic models of replication and transcriptional regulation in mitochondria.

6.1 Introduction

Mitochondria are the primary site of oxidative phosphorylation in most eukaryotic cells, and in addition to that play a vital role in a long list of other important cellular processes (Williams et al. 2013; Andersen & Kornbluth 2013; Miller 2011; Lill et al. 2012). They possess their own genome (Nass et al. 1965; Schatz 1963), which in mammals is circular mapping and ∼16 kilobases long (16,571 bp in humans) (Anderson et al. 1981; Bibb et al. 1981; Satoh & Kuroiwa 1991). Mitochondria originated very early in eukaryote evolution, when their most likely α-proteobacterial ancestor became an endosymbiont to the ancestor of modern eukaryotes (Yang et al. 1985). The mitochondrial genome is the remnant of the genome of that prokaryotic endosymbiont, which, as a result of the loss of genes and the transfer of genes from the organellar to the nuclear genome (Kleine et al. 2009), has been greatly reduced in size and gene content. In humans, it encodes 13 proteins (components of the electron transport chains), 2 rRNAs and 22 tRNAs. It has only one significant stretch of non-coding DNA – the so called control, or D-loop (Arnberg et al. 1971; ter Schegget et al. 1971) regulatory region (or non-coding region, NCR), which is approximately 1kb long and plays an important role in the processes of transcription and replication.

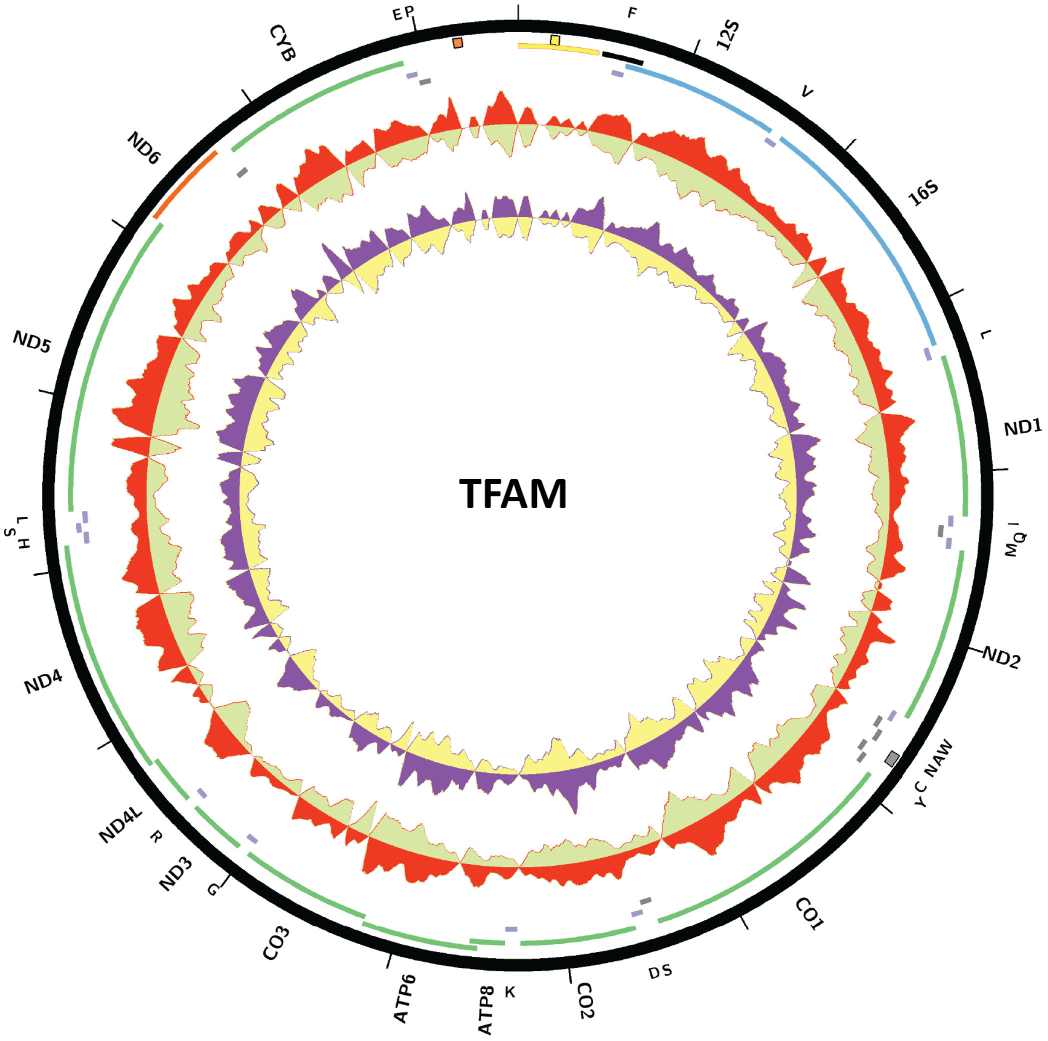


Figure 6.1: TFAM occupancy over the human mitochondrial genome. ChIP-seq against TFAM was carried out in HepG2 cells. Shown is the plus and minus strand distribution of mapped reads in ChIP (outer tracks, red and green) and control input datasets (inner track, blue and yellow). Also shown are the rRNA (blue tiles), tRNA (purple tiles), and heavy and light strand protein coding genes (green and red tiles) as well as the LSP promoter (yellow tile), HSP promoter (black tile), and the origins of heavy strand (*Ori-b*, orange square, and O_H, yellow square) and light strand (O_L, gray square) replication. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

The mitochondrial genome is expressed and replicated by dedicated transcription and replication machineries separate from those acting in the nucleus. Transcription initiates from three different promoters located in the D-loop – two promoters transcribing the "heavy" or H-strand (HSP1 and HSP2) and one "light"-strand promoter (LSP). Mitochondrial transcripts are polycistronic and the mature mRNAs are produced by posttranscriptional processing mediated by the excision of the tRNAs that are found between all genes (Ojala et al. 1981). It is carried out by POLRMT, an RNA polymerase of apparent phage origin (Masters et al. 1987; Shutt & Gray 2002). Initiation of transcription requires the activities of the TFAM (mitochondrial transcription factor A) and TFB2M (mitochondrial transcription factor B2; Falkenberg et al. 2002; Metodiev et al. 2009; Sologub et al. 2009) proteins. TFAM also plays a structural and packaging role in the mitochondrial nucleoid and is necessary for mtDNA replication and maintenance (Alam et al. 2003; Ekstrand et al. 2004; Kaufman).

Transcription from the HSP1 promoter is thought to generate a transcript that includes the two ribosomal RNAs and terminates shortly after. The same site is also where termination of transcription in the other direction, originating from the LSP promoter and containing the ND6 gene, occurs. Both termination events are triggered by the presence of the DNA binding protein MTERF, which acts as a termination factor (Christianson & Clayton 1988; Kruse et al. 1989; Fernandez-Silva et al. 1997; Shang & Clayton 1994). The polycistronic transcript originating from the HSP2 promoter includes all other protein coding genes and reaches all the way to the other end of the D-loop (Montoya et al. 1983; Asin-Cayuela & Gustafsson 2007).

Replication of mitochondrial DNA is carried out by DNA Polymerase γ, which consists of two subunits, the catalytic POLG and the accessory POLG2 (Ropp & Copeland 1996; Yakubovskaya et al. 2006; Chan & Copeland 2009; Wanrooij & Falkenberg 2010); in addition, the mitochondrial single-strand binding protein (mtSSB) and the helicase TWINKLE also play an important role. Multiple models for how the process of replication occurs have been proposed (Holt & Reyes 2012). The classic asynchronous strand displacement model (SDM) of replication involves the initiation of replication of the heavy strand from replication origins within the D-loop. Leading strand replication then proceeds for about two thirds of the length of the mitochondrial genome until the origin of light strand replication is encountered (O_L; Martens & Clayton 1979), upon which replication of the light strand begins (Kasamatsu & Vinograd 1973; Robberson & Clayton 1972; Clayton 1982). Two origins of heavy strand replication have been mapped: O_H and *Ori-b* (Kang et al. 1997; Pham et al. 2006; Crews et al. 1979; Fish et al. 2004). DNA replication is primed by POLRMT transcription initiating from the LSP promoter (Chang et al. 1985; Chang & Clayton 1985; Kang et al. 1997; Pham et al. 2006); some 600bp downstream of the O_H, near the end of the D-loop region, replication often arrests, and a triple-stranded D-loop structure forms in the NCR (Arnberg et al. 1971; Kasamatsu et al. 1971; ter Schegget et al. 1971).

In the last decade, evidence for different models of replication has been accumulating. These include the RITOLS (**R**ibonucleotide **I**ncorporation **T**hrough**O**ut the **L**agging **S**trand) model (Yasukawa et al. 2005; Yasukawa et al. 2006; Pohjoismäki et al. 2010; Holt & Reyes 2012) and the strand-coupled model (Holt et al. 2000). The RITOLS model is somewhat similar in its mechanism to the SDM model in that both the D-loop and the O_L replication origins play a role; however, in contrast to SDM, it features the incorporation of RNA on the lagging strand while the leading strand is being synthesized. Under the strand-coupled model, replication is bidirectional and can initiate from regions outside of the NCR (Bowmaker et al. 2003).

These models have been developed using traditional molecular biology approaches, which have proven highly useful in understanding the biology of mitochondria. However, the global distribution of these proteins over the mitochondrial genome has so far not been systematically characterized. Here, we used ChIP-seq (Chromatin Immunoprecipitation coupled with high-throughput sequencing; Johnson et al. 2007) to generate such maps for TFAM, TFB2M, MTERF, POLRMT and POLG and to further elucidate the role of these proteins in the processes of mitochondrial transcription and replication.

6.2 Results

6.2.1 Measuring the genome-wide occupancy of mitochondrial proteins using ChIP-seq

In order to characterize the genome-wide occupancy of mitochondrial proteins, we applied ChIP-seq using standard, previously described, protocols (Johnson et al. 2007) in two ENCODE cell lines: the lymphoblastoid GM12878 and the liver carcinoma HepG2 cells. The results were generally very similar between the two lines, and for this reason only a single dataset is used for the visualization of the occupancy of each protein throughout the manuscript. We gen-

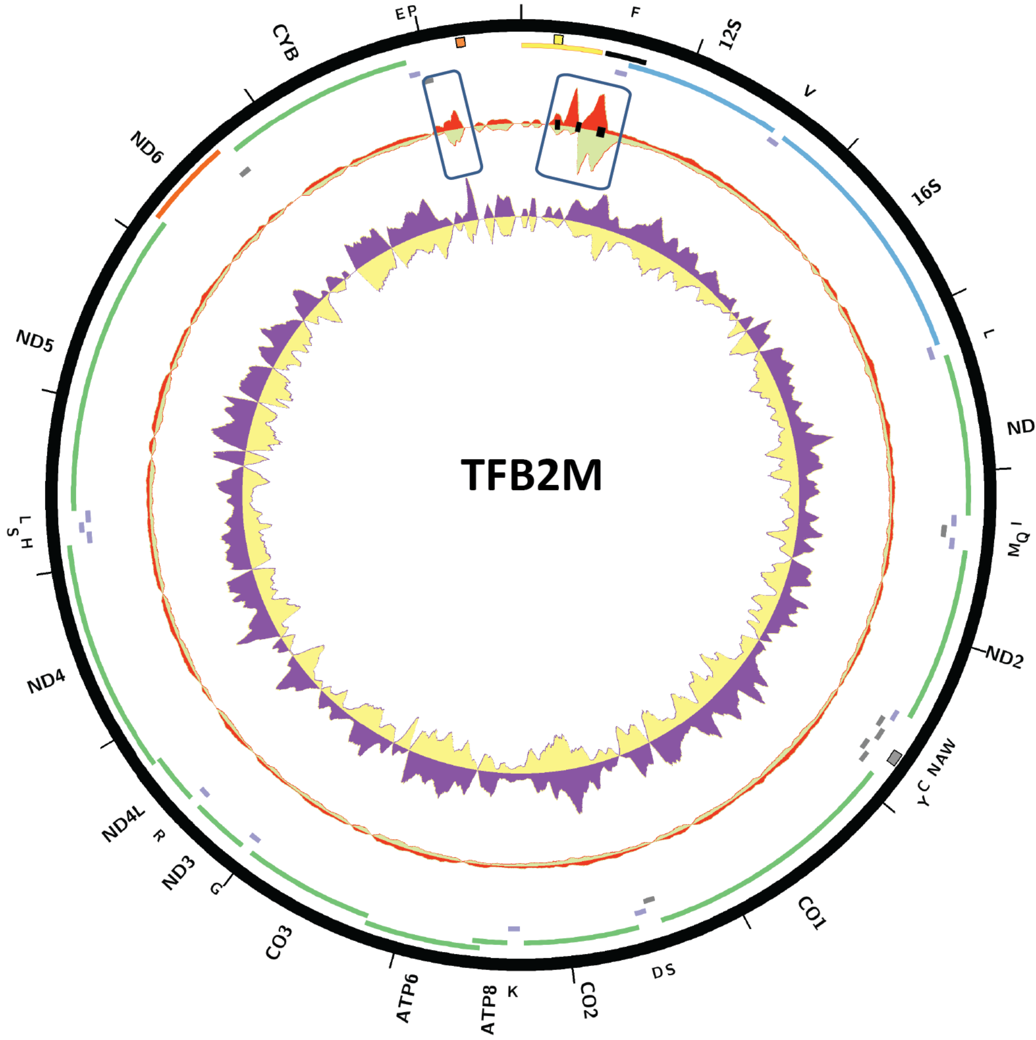


Figure 6.2: TFB2M binding over the human mitochondrial genome. ChIP-seq against TFB2M was carried out in HepG2 cells. Shown is the plus and minus strand distribution of mapped reads in ChIP (outer tracks, red and green) and control input datasets (inner track, blue and yellow). Also shown are the rRNA (blue tiles), tRNA (purple tiles), and heavy and light strand protein coding genes (green and red tiles) as well as the LSP promoter (yellow tile), HSP promoter (black tile), and the origins of heavy strand (*Ori-b*, orange square, and O_H, yellow square) and light strand (O_L, gray square) replication. The black rectangles indicate the putative TFB2M binding sites within the LSP and the HSP. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

erated 50bp long sequencing reads and aligned them against the mitochondrial genome (version hg19 of the human genome, see the Methods section for more details). As the size of the mitochondrial genome is small and makes this approach feasible (in contrast to the ≥3Gb nuclear genome), manual inspection of the resulting ChIP and control dataset read profiles was used to identify regions of enrichment. In addition, we also mapped reads against the nuclear genome in order to examine the possible association of mitochondrial proteins such as POLRMT

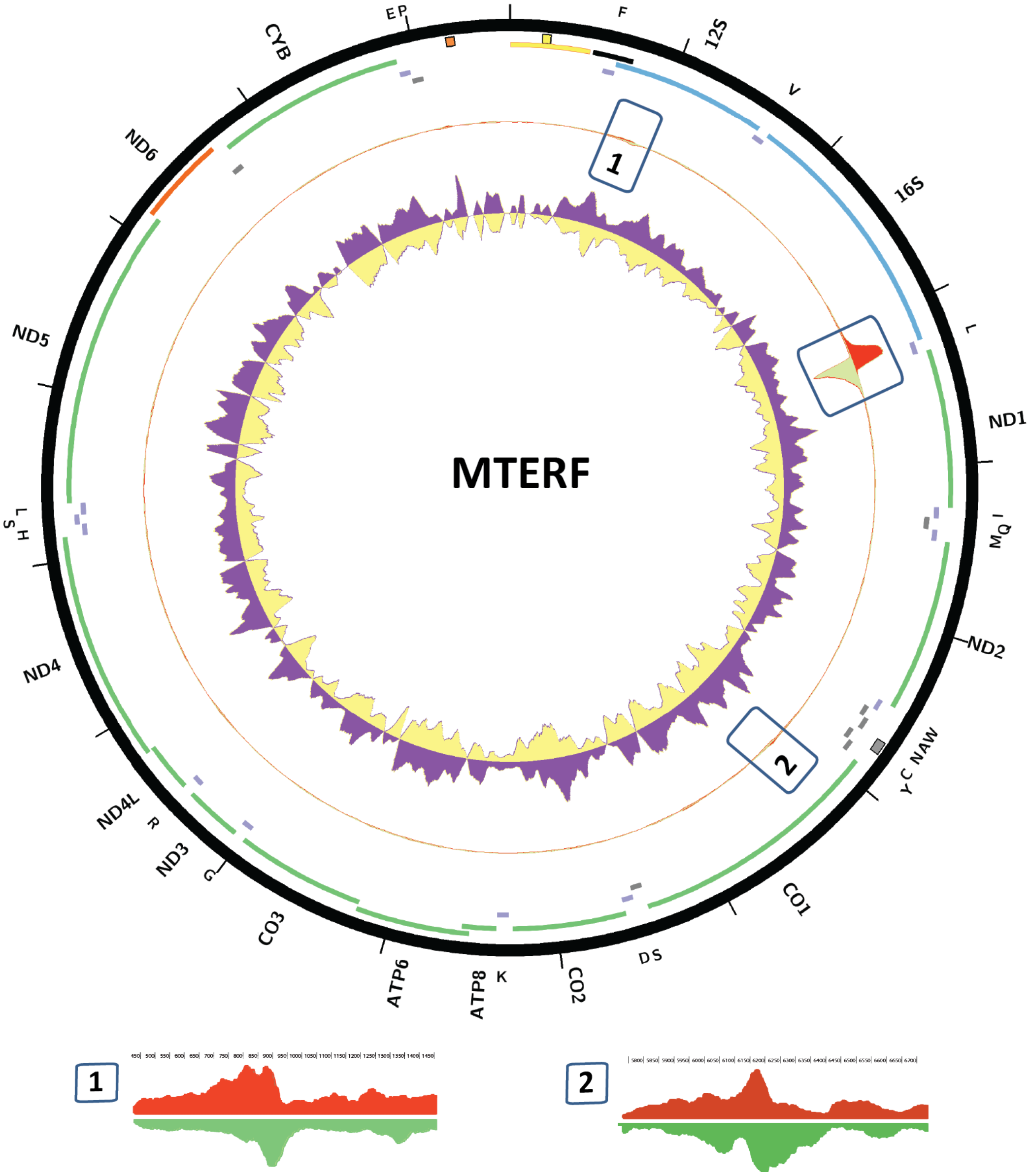


Figure 6.3: MTERF occupancy over the human mitochondrial genome. ChIP-seq against MTERF was carried out in HepG2 cells. Shown is the plus and minus strand distribution of mapped reads in ChIP (outer tracks, red and green) and control input datasets (inner track, blue and yellow). Also shown are the rRNA (blue tiles), tRNA (purple tiles), and heavy and light strand protein coding genes (green and red tiles) as well as the LSP promoter (yellow tile), HSP promoter (black tile), and the origins of heavy strand (*Ori-b*, orange square, and O_H, yellow square) and light strand (O_L, gray square) replication. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009). Two sites of putative MTERF occupancy of lower intensity are also shown separately as insets.

with the nuclear genome.

6.2.2 TFAM

We previously characterized the occupancy of the mitochondrial genome by TFAM using

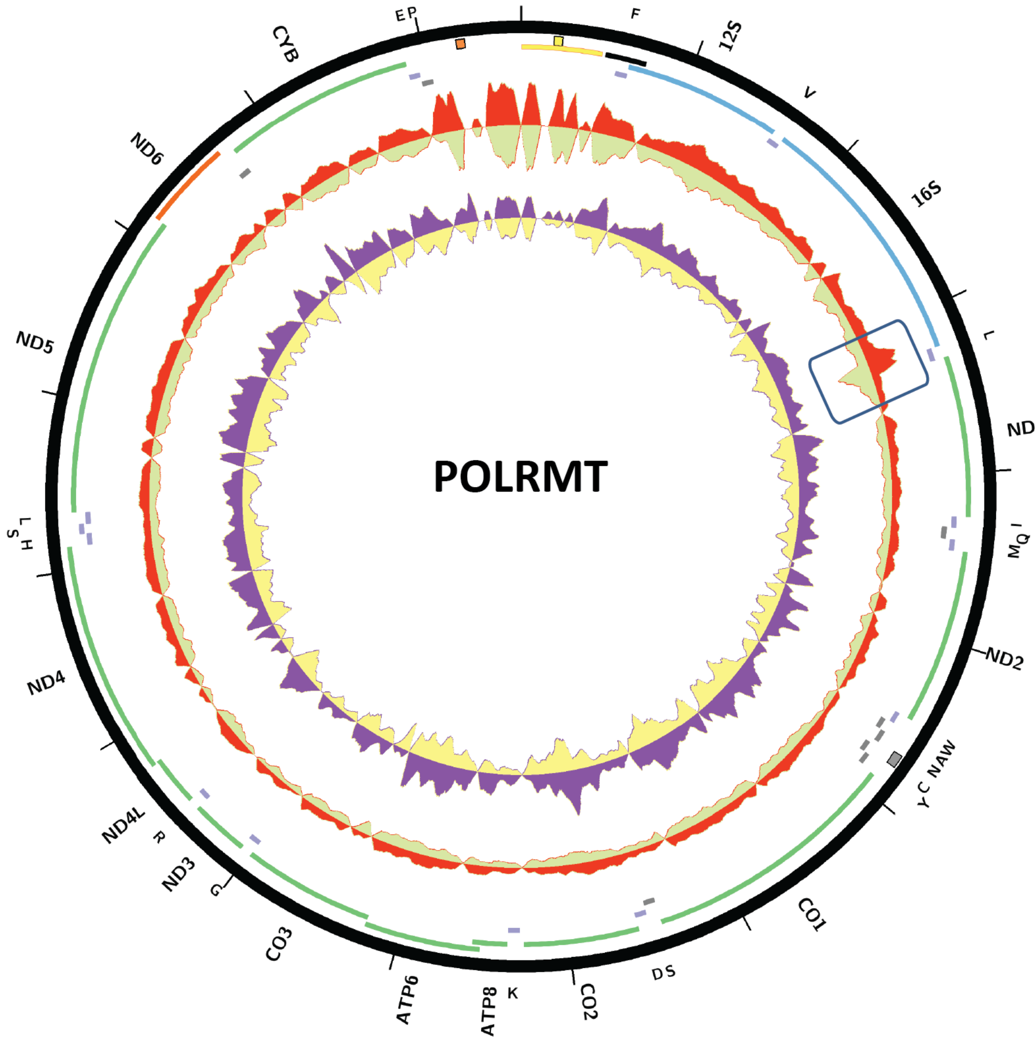


Figure 6.4: POLRMT occupancy over the human mitochondrial genome. ChIP-seq against POLRMT was carried out in HepG2 cells. Shown is the plus and minus strand distribution of mapped reads in ChIP (outer tracks, red and green) and control input datasets (inner track, blue and yellow). Also shown are the rRNA (blue tiles), tRNA (purple tiles), and heavy and light strand protein coding genes (green and red tiles) as well as the LSP promoter (yellow tile), HSP promoter (black tile), and the origins of heavy strand (*Ori-b*, orange square, and O_H, yellow square) and light strand (O_L, gray square) replication. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

ChIP-seq in HeLa cells (Wang et al. 2013). TFAM was found to fully coat mtDNA with little evidence for strong site-specific occupancy over the promoter regions in the D-loop (although such occupancy is by no means incompatible with the data – TFAM plays a well-characterized role in transcriptional initiation, but its association with DNA might be transient compared to the steady-state packaging role it also plays, and as a result generates ChIP-seq signal that is too weak in comparison to stand out). Here we also carried out TFAM ChIP-seq in HepG2 cells and

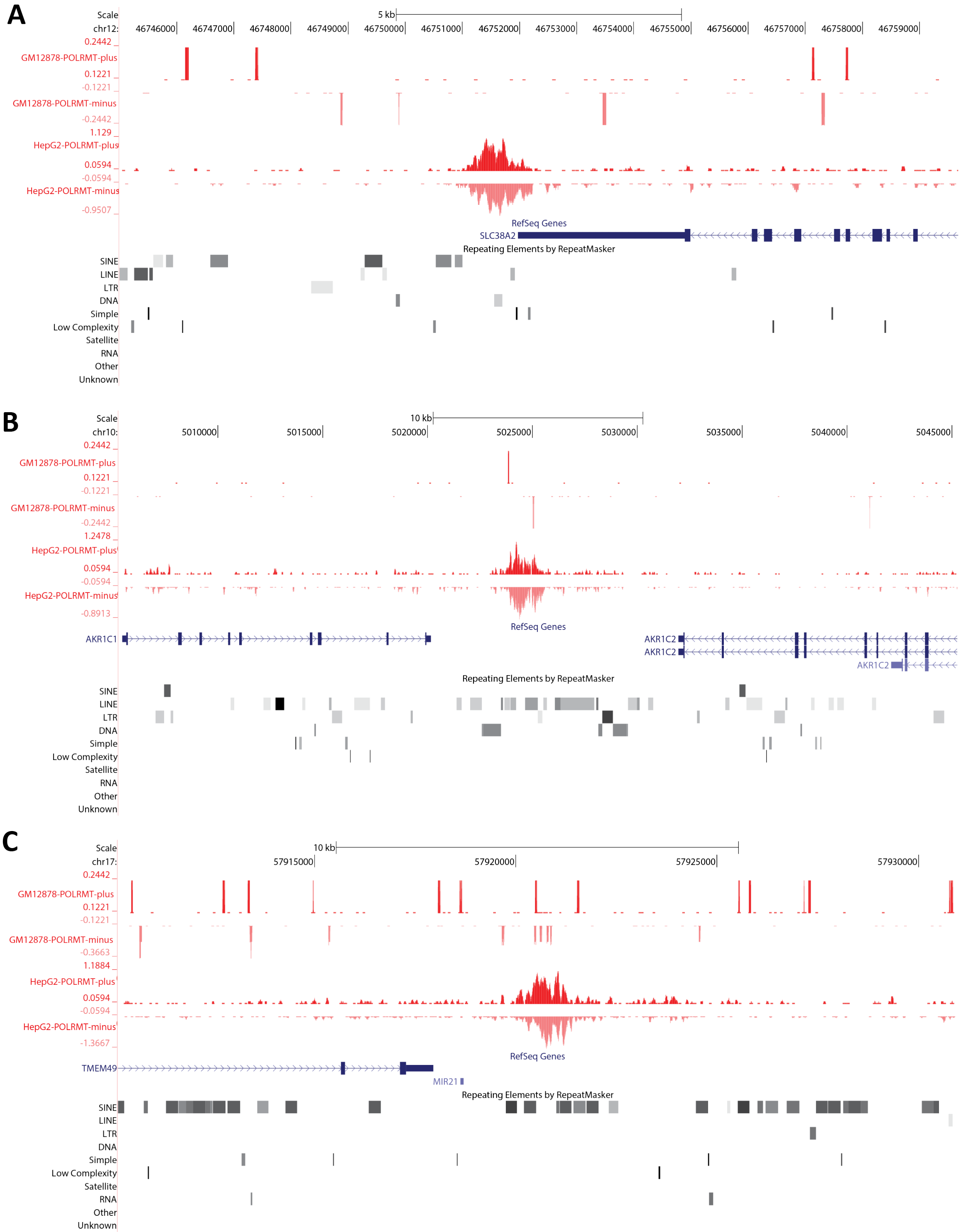


Figure 6.5: Examples of nuclear loci displaying evidence for physical association with either POLRMT or its short nuclear isoform spRNAP-IV.

obtained very similar results. The fraction of all mapped reads (to the nuclear chromosomes or to chrM) mapping to the mitochondrial genome was 68%. In the same time the TFAM ChIP-seq

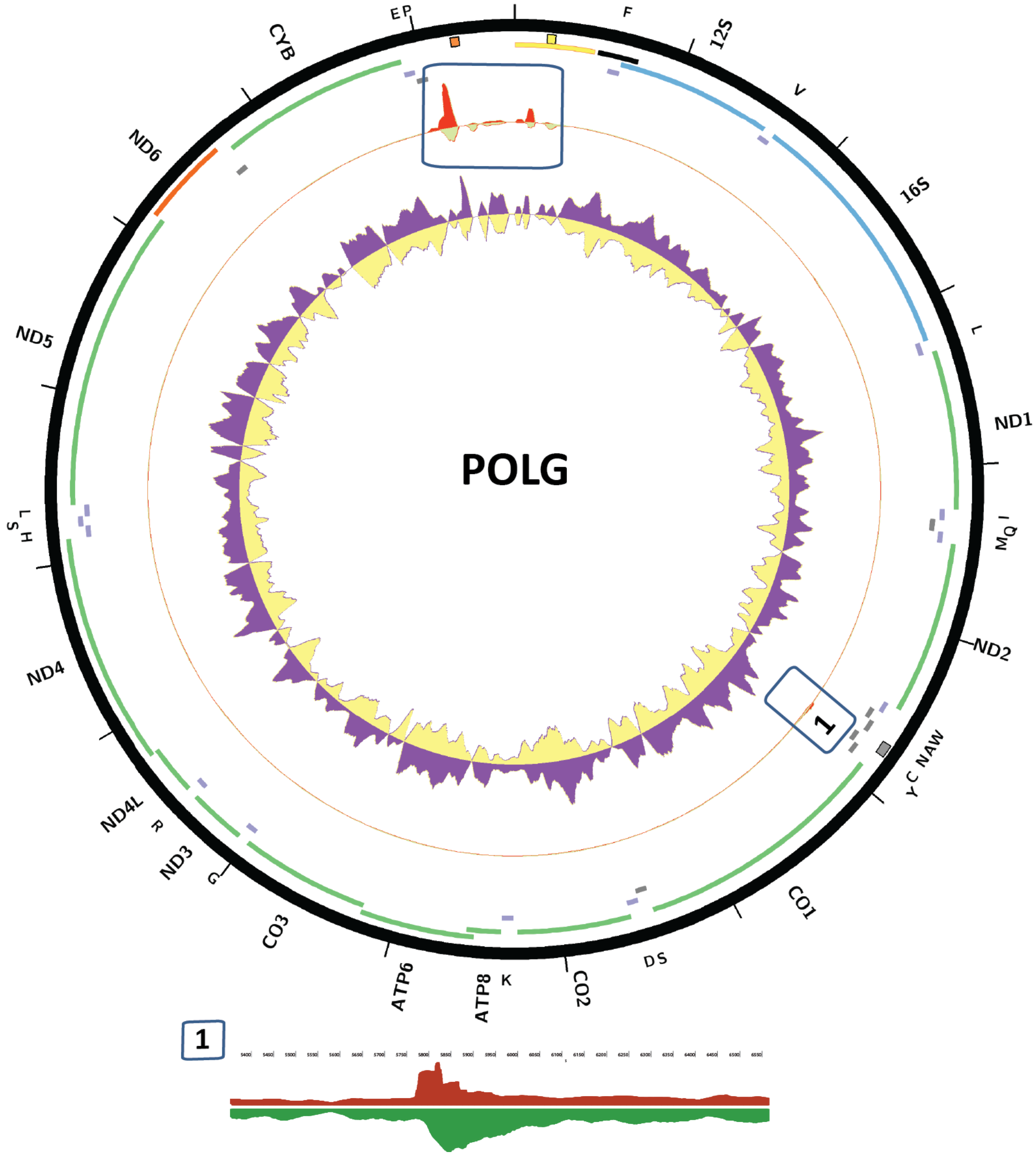


Figure 6.6: Occupancy of the DNA polymerase γ catalytic subunit POLG over the human mitochondrial genome. ChIP-seq against POLG was carried out in HepG2 cells. Shown is the plus and minus strand distribution of mapped reads in ChIP (outer tracks, red and green) and control input datasets (inner track, blue and yellow). Also shown are the rRNA (blue tiles), tRNA (purple tiles), and heavy and light strand protein coding genes (green and red tiles) as well as the LSP promoter (yellow tile), HSP promoter (black tile), and the origins of heavy strand (*Ori-b*, orange square, and O_H, yellow square) and light strand (O_L, gray square) replication. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009). A site of putative POLG occupancy of near the OriL is shown separately as an inset.

signal profile was highly similar to that of the control sonicated input sample (Figure 6.1). As discussed before (Wang et al. 2013), these ob-

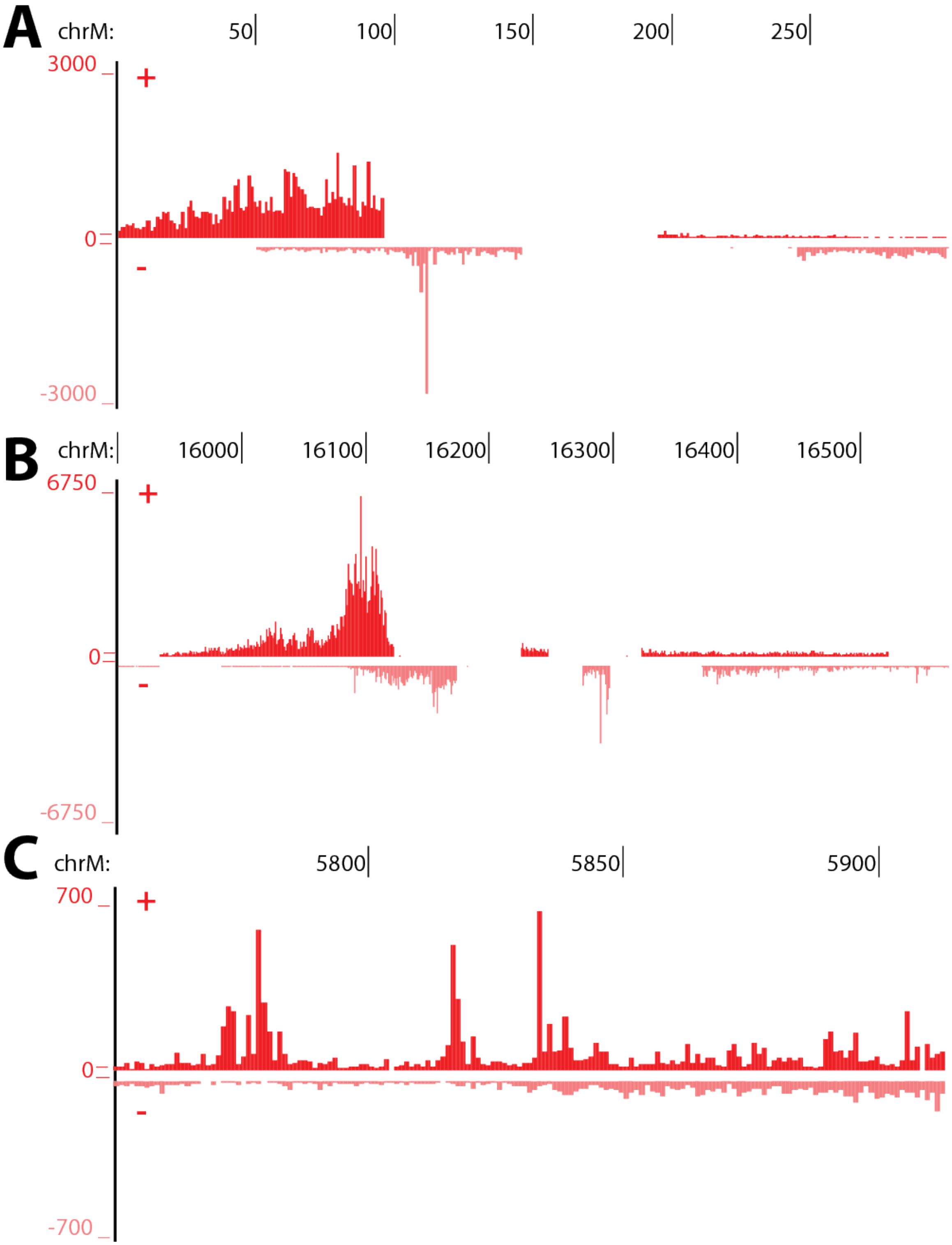


Figure 6.7: Distribution of the 5' ends of forward and minus strand ChIP-seq reads around the known origins of mitochondrial replication. Shown are UCSC Genome Browser snapshots of POLG ChIP-seq tracks created by assigning non-zero scores only to the positions to which the 5' ends of reads map to. (A) Positions 1 to 300, containing the O_H replication origin. (B) Positions 15,900 to 16,571, containing the *Ori-b* replication origin. (C) Positions 5,751 to 5,913 around the O_L replication origin.

servations argue for the uniform coating of the mitochondrial genome by TFAM, with the non-uniformities in coverage being due to a combination of sonication, library preparation and sequencing biases.

6.2.3 TFB2M

We next characterized the genome-wide occupancy of TFB2M. TFB2M is required for transcription of mitochondrial genes and is known to directly interact with DNA as part of the POLRMT transcription initiation complex (Sologub et al. 2009; Litonin 2010). As such it is expected to physically localize to the HSP and LSP promoters. ChIP-seq data confirms this expectation but also reveals a potentially more complicated picture (Figure 6.2). Within the HSP promoter region, we observed a single large occupancy site, which displayed the characteristic for ChIP-seq peaks asymmetry between the read distribution on the plus and minus strands (Kharchenko et al. 2008; Landt et al. 2012). We were not able to distinguish separate binding sites for the HSP1 and HSP2 promoters; however, this is likely due to the fact that they are too closely spaced relative to the resolution of the ChIP-seq assay. The ChIP profile in the LSP region was more complex, with at least two putative occupancy sites observed – a stronger one exhibiting very strong read asymmetry and located in the general region of LSP transcription initiation, and a weaker one located close to it downstream in the direction of LSP transcription. The significance of the second site is at present unknown. Even more surprising is the observation of a region of elevated TFB2M occupancy (but with difficult to identify specific binding sites) at the very end of the D-loop downstream of the LSP and the origins of replication. To the best of the authors' knowledge, association of TFB2M with this region has previously not been reported and it is not clear what role it might be playing there; future studies will be needed to elucidate it.

6.2.4 MTERF

We then profiled the genome-wide occupancy of the MTERF protein. As expected, we observed a single extremely strong occupancy site, exhibiting very notable strand asymmetry at the LSP and HSP1 termination site (Figure 6.3). These results are also in agreement with recently published data in HeLaS3 cells (Terzioglu et al. 2013). However, detailed examination of the read profiles identified two more putative occupancy sites for MTERF, which, although much weaker in terms of ChIP-seq signal strength, are nevertheless above background and display a read asymmetry suggesting they correspond to real biochemical events. The first one (site "1" in Figure 6.3) is located immediately downstream of the HSP1/2 promoter, and is of interest as it has been reported that MTERF interacts with this region of mtDNA and a loops is formed between the two MTERF binding sites, with this interaction being important for the activation of transcription (Martin et al. 2005). The second one is located in the vicinity (but not within) the O_L region and its functional significance is at present unclear.

6.2.5 Mitochondrial RNA Polymerase (POLRMT)

Next we studied the association of the mitochondrial RNA Polymerase POLRMT with the mitochondrial genome (Figure 6.4). We found one site of strong POLRMT localization in the mitochondrial genome, and it coincided with the MTERF termination sites between the 16S rRNA and the *ND1* genes. This suggests that termination of transcription is associated with pausing and/or an appreciable increase in the residence time of the polymerase around this site. In addition, we observed generally elevated read coverage over the D-loop but without obvious distinct sites of localized enrichment.

In addition to its well characterized role in mitochondrial transcription, the *POLRMT* gene has been suggested to also participate in nuclear transcription, through the production of an alternative isoform called spRNAP-IV that lacks the N-terminal 262 amino acids (Kravchenko et al. 2005; Lee et al. 2011). The antibody we used to carry out ChIP-seq against POLRMT has been raised against amino acids 841-1140 located near its C-terminus, and should therefore also react with spRNAP-IV. We examined the possible nuclear role of spRNAP-IV by calling ChIP-seq peaks in the nuclear genome using MACS2 (Feng et al. 2012). After manual curation of the resulting peaks (in order to filter out obvious artifacts), we identified 47 loci with significant enrichment in POLRMT ChIP-seq in HepG2 cells. Notably, only a small minority of these peaks were also called in GM12878 cells suggesting that if spRNAP-IV indeed transcribes through these regions, it may do so in cell-type specific manner. Sites of POLRMT enrichment displayed a preferential localization in the immediate upstream and downstream regions of protein coding genes, and were often associated with transposable elements (representative examples are shown in Figure 6.5).

6.2.6 Mitochondrial DNA Polymerase (Pol γ)

Finally, we analyzed the mitochondrial genome occupancy of the catalytic subunit (POLG2) of the mitochondrial DNA polymerase (Pol γ). We observed several regions of significant enrichment (Figure 6.6). First, a site of very strong and localized read density is observed at the end of the D-loop in the direction of LSP replication and about a 100bp downstream of the $\mathrm{O_H}$ replication origin. Second, a similar but lower-intensity region of occupancy is found about a 100bp downstream of the *Ori-b* replication origin. Third, a weaker but detectable occupancy site is seen at the $\mathrm{O_L}$ replicaiton origin. Of note, only the the $\mathrm{O_L}$ site exhibited the typical for a ChIP-seq peak strand asymmetry; in contrast, the two strong signal peaks in the D-loop displayed forward and reverse strand profiles that were similar to each other, with little shift between the peaks on the two strand, and with a markedly higher signal on the forward strand than on the reverse one. The features of the site at the end of the D-loop would indicate these sites to be a possible experimental artifact in other settings, and indeed this is how they were interpreted previously when they were observed in the ChIP-seq read profiles around the D-loop of nuclear transcription factors (Marinov et al. 2014). However, given that the DNA polymerase is known to pause at this site, that the strength of the signal compared to the background level is much stronger than it is for nuclear transcription factors, and that the DNA-polymerase is also observed at other sites where its occupancy is expected, in this case it is more likely that a significant portion of the observed signal corresponds to true physical association events of POLG with mtDNA. Also, the properties of DNA polymerase are expected to be somewhat different from those of ChIP-seq against conventional double-stranded DNA binding proteins. The typical asymmetric, strand-shifted ChIP-seq profile arises in the context of long double-stranded DNA molecules within which transcription factor binding sites are embedded and occupied, but the replicating DNA polymerase is associated with free 3' ends, and also with free 5' ends near the initiation sites. In the latter case, the structure also contains an RNA-DNA hybrid; depending on whether the RNA portion of it survives the process of fixation and immunoprecipitation, it is expected that if the DNA polymerase spends significant time around the region of initiation, the process of end repair and ChIP-seq library construction will produce highly phased 5' ends corresponding most likely to the RNA-to-DNA transition positions, and possibly to the initiation of transcription. To examine this in depth, we generated forward and reverse strand coverage tracks showing only the 5' ends of ChIP-seq reads (Figure 6.7). Remarkably, we indeed observed highly phased clustering of 5' ends on the reverse strand (and not on the forward strand) in the region of the D-loop around the $\mathrm{O_H}$ replication origin (Figure 6.7A), located around position 110. This position is somewhat different from the locations suggested by previous efforts to map the heavy strand origin of replication (position 191 according to Crews et al. 1979, position 57 according to Fish et al. 2004, position 300 according to Pham et al. 2006; Holt & Reyes 2012), but is in the same region; of note, it is also considerably downstream of the LSP promoter suggesting it corresponds to the site of initiation of replication from the RNA primer. A region of phasing of reverse strand reads was also observed around position 16,280 near the *Ori-b* replication origin (Figure 6.7B), although it has been previously suggested to be precisely located at nucleotide 16,197 (Yasukawa et al. 2005). Finally, in the region around the $\mathrm{O_L}$ replication origin, we observed several positions with phased 5' ends, all on the forward strand (in contrast to the D-loop origins, where phasing is on the reverse strand).

These observations support several features of the existing models of mitochondrial DNA replication. First, the $\mathrm{O_L}$ origin is definitely used, as suggested by both the increased occupancy of POLG there, and the detection of phased 5' ends of ChIP-seq reads only on the forward strand (likely corresponding to the 5' end of newly synthesized DNA strands). Second, the $\mathrm{O_H}$ origin is also used, as evidenced by the strong phasing of reads on the reverse strand nearby, and the *Ori-b* is likely used too, for similar reasons (although it does not exhibit the same strong phasing of 5' ends). Third, as suggested previously, replication pauses at the end of the D-loop. The significant asymmetry in the number of reads on the forward strand (corresponding to the template strand for replication originating from $\mathrm{O_H}$ or *Ori-b* in the direction of this end of the NCR) and reverse strand, with the forward strand displaying larger number of reads compared to the reverse strand, and with little shift between the two profiles, remains puzzling. It is at present not clear what the reason for this pattern is, as it might be due to a complex combination of multiple poorly understood factors having to do with crosslinking, sonication, size selection and library generation. However, one attractive possibility is that it is related to the direction of replication; of note, lower in magnitude but detectable peaks which display the opposite pattern of asymmetry between read density on the forward and reverse strands are observed in the opposite direction of both the $\mathrm{O_H}$ and *Ori-b* origins, which is intriguing as there have reports that replication initiation from these promoters is bidirectional (Holt & Reyes 2012).

6.3 Conclusions

In this work, we generated comprehensive genome-wide maps of the physical occupancy over the mitochondrial genome of the main proteins involved in mitochondrial transcription (TFAM, TFB2M, MTERF and POLRMT) and replication (POLG). Consistent with previous work (Wang et al. 2013), we found TFAM to fully coat the mitochondrial genome, with no outstanding localized sites of enrichment in the cells studied here. We found TFB2M to occupy the HSP and LSP promoters, in line with previous observations. However, its occupancy seems to be more complex than previously thought, with at least three occupancy cites in the region; in addition to this, we also observed it localizing to the opposite end of the D-loop, where its role is at present unclear. We observed very strong MTERF localization at the known transcription termination site. In addition, we also found that MTERF can be crosslinked (though weakly and likely due to an indirect interaction with mtDNA) to the region immediately downstream of the HSP1/2 promoter, consistent with the previously suggested model in which MTERF mediates looping between the termination sites and the promoter region to activate transcription (Martin et al. 2005). The termination site was also the most notable region of strongly localized read enrichment for POLRMT, suggesting it pauses there while termination takes place. We also examined the previously proposed nuclear role of the POLRMT genes, through its alternative isoform, spRNAP-IV, and found it to indeed associate with a limited number of nuclear loci, but its binding patterns do not clearly reveal its possible functional roles.

The POLG ChIP-seq datasets presented strong evidence in support of existing models in which all three known origins of replication feature prominently (although the data is consistent with more frequent usage of the $\mathrm{O_H}$ than of the *Ori-b*) origin. We did not find direct evidence for replication initiation elsewhere in the mitochondrial genome, as suggested by some versions of the strand-coupled replication model (Bowmaker et al. 2003); however, that such modes of replication are used cannot be ruled out by the data as it is possible that they do not generate strongly localized POLG occupancy. The data is consistent with both the SDM and the RITOLS models, as the behavior of the DNA polymerase is very similar under both, but the SDM and RITOLS models can be distinguished using functional genomics tools: under RITOLS, the lagging strand is covered by RNA, while in the SDM model, the lagging strand is associated with the mtSSB protein. It will therefore be highly informative to apply strand-specific ChIP-seq protocols (Zhou et al. 2013) to mtSSB (and also, to the polymerase itself).

In addition, occupancy maps of other proteins involved in the biology of mtDNA (for example, the other three members of the MTERF family) should prove highly valuable for understanding their functional role, as in contrast to the well-known proteins studied here, much less is known about them at present.

6.4 Methods

6.4.1 Cell growth, chromatin immunoprecipitation and sequencing

Cells were grown under standard ENCODE protocols, which can be found at `http://genome.ucsc.edu/ENCODE/protocols/cell/human/`. ChIP experiments and preparation of DNA for sequencing were performed following standard procedures (Johnson & Mortazavi et al. 2007; Gasper et al., in press). The following antibodies were used: mouse polyclonal α-TFAM (Sigma-Aldrich, SAB1401382), mouse monoclonal α-TFB2M (Novus Biologicals, H00064216-M01), goat polyclonal α-MTERF (Santa Cruz, sc-160543), goat polyclonal α-POLG (Santa Cruz, sc-5930) and mouse monoclonal α-POLRMT (Santa Cruz, sc-365082). Libraries were sequenced using the Illumina HiSeq 2000.

6.4.2 Data processing and analysis

Reads were aligned as described previously (Wang et al. 2010) using Bowtie (Langmead et al. 2009), version 0.12.7. Two sets of alignments were generated. Firs, reads were mapped against either the female or male hg19 version of the human genome (excluding all random chromosomes and haplotypes; assembly downloaded from the UCSC genome browser) depending on the sex of the cell line (male for HepG2, female for GM12878) with the following settings: `‘‘-v 2 -t -k 2 -m 1 --best --strata’’`, which allow for two mismatches relative to the reference and only retain unique alignments. These alignments exclude all reads mapping ambiguously to both the nuclear and mitochondrial genomes, and were used for calling peaks in the nuclear genome with MACS2 (Feng et al. 2012), version 2.0.9. Second, reads were mapped against chrM alone, with the following settings: `‘‘-v 0 -t -k 2 -m 1 --best --strata’’`, i.e. allowing for zero mismatches to the reference. These alignments were used for visualization and evaluation of ChIP enrichment over the mitochondrial genome. Circos plots were generated using Circos version 0.60 (Krzywinski et al 2009). Additional data processing was carried out using custom-written python scripts.

7

Evidence for Site-Specific Occupancy of the Mitochondrial Genome by Nuclear Transcription Factors

The material in this chapter was published as:

Marinov GK*, Wang YE*, Chan DC, Wold BJ. 2014. Evidence for site-specific occupancy of the mitochondrial genome by nuclear transcription factors. *PLoS ONE* **9**(1):e84713. doi: 10.1371/journal.pone.0084713
The experimental data in it was generated by Yun Elisabeth Wang in the Chan lab. The paper is reprinted in Appendix L

Abstract

Mitochondria contain their own circular genome, with mitochondria-specific transcription and replication systems and corresponding regulatory proteins. All of these proteins are encoded in the nuclear genome and are post-translationally imported into mitochondria. In addition, several nuclear transcription factors have been reported to act in mitochondria, but there has been no comprehensive mapping of their occupancy patterns and it is not clear how many other factors may also be found in mitochondria. We addressed these questions by analyzing ChIP-seq data from the ENCODE, mouseENCODE and modENCODE consortia for 151 human, 31 mouse and 35 *C. elegans* factors. We identified 8 human and 3 mouse transcription factors with strong localized enrichment over the mitochondrial genome that was usually associated with the corresponding recognition sequence motif. Notably, these sites of occupancy are often the sites with highest ChIP-seq signal intensity within both the nuclear and mitochondrial genomes and are thus best explained as true binding events to mitochondrial DNA, which exists in high copy numbers in each cell. We corroborated these findings by immunocytochemical staining evidence for mitochondrial localization. However, we were unable to find clear evidence for mitochondrial binding in ENCODE and other publicly available ChIP-seq data for most factors previously reported to localize there. As the first global analysis of nuclear transcription factors binding in mitochondria, this work opens the door to future studies that probe the functional significance of the phenomenon.

In the course of our study of the association of TFAM with the mitochondrial nucleoid, we made the accidental but very intriguing observation that a number of transcription factors for which ChIP-seq data was available from the ENCODE Consortium exhibited high levels of localized signal enrichment over the mitochondrial genome. We followed these observations and investigated the phenomenon in depth. It turned out this was not an entirely new observations and the physical localization of nuclear trancsription factors to the mitochondria had been reported in the past. However, the power of the resolution and comprehensiveness of coverage of ChIP-seq had not been utilized in none of those

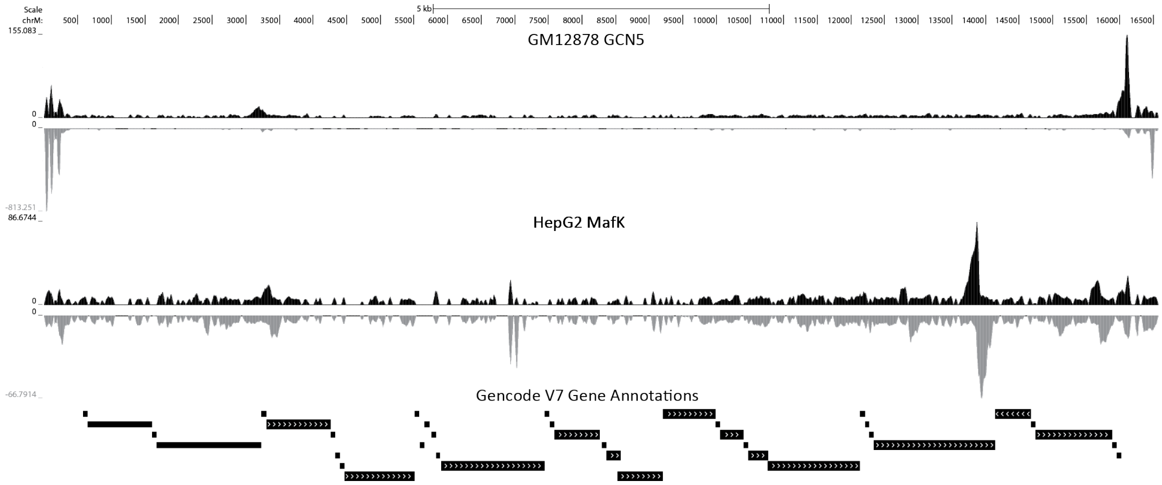


Figure 7.1: Representative USCS Genome Browser snapshots of nuclear transcription factor ChIP-seq datasets exhibiting strong enrichment in the mitochondrial genome. (A) GM12878 GCN5 shows high signal intensity in the D-loop (the region between coordinates 16030 and 580, i.e. the non-coding regions on the left and right ends of the snapshot) representative of the D-loop enrichment observed for a large number of transcription factors (B) In contrast, a large MafK peak is observed in a coding region outside of the D-loop in HepG2 cells. Upper track (black) shows reads aligning to the forward strand, lower track (gray) shows read aligning to the reverse strand

studies, in fact there was very little direct biochemical evidence for the binding of those factors to mtDNA. Our study, the results of which I present in this chapter, was the first global survey of these events, both in terms of covering the whole mitochondrial genome in multiple species, and the number of transcription factors included in it.

7.1 Introduction

In addition to the well-characterized regulators of mitochondrial transcription, multiple reports have suggested that transcription factors that typically act in the nucleus might also have regulatory functions in mitochondrial transcription (Leigh-Brown et al. 2010; Szczepanek et al. 2012b). The glucocorticoid receptor (GR) was the first such factor reported to localize to mitochondria and to interact with mtDNA (Demonacos et al. 1993; Demonacos et al. 1995; Koufali et al. 2003; Psarra et al. 2006). A 43kDa isoform of the thyroid hormone T_3 receptor $T_3R\alpha1$ called p43 has been found to directly control mitochondrial transcription (Casas et al. 1999; Enríquez et al. 1999a; Enríquez et al. 1999b; Wrutniak et al. 1995). Cyclic-AMP Response element Binding protein (CREB) has been shown to localize to mitochondria and suggested to bind to the D-loop (Lee et al. 2005; Ryu et al. 2005; Cammarota et al. 1999; De Rasmo et al. 2009). The tumor suppressor transcription factor p53 has been implicated in mtDNA repair and regulation of gene expression through interactions with TFAM (Marchenko et al. 2000; Marchenko et al. 2007; Achanta et al. 2005; Heyne et al. 2004; Yoshida et al. 2003). It has also been proposed to play a proapoptotic role through association with the outer mitochondrial membrane (Vaseva & Moll 2009). A similar role has been also ascribed to the IRF3 transcription factor (Liu et al. 2010; Chattopadhyay et al. 2010). The mitochondrial localization of the estrogen receptor (ER) is also well established, for both its ERα and ERβ isoforms, and it too has been suggested to bind to the D-loop (Chen et al. 2004; Monje & Boland 2001). NFκB and IκBα have been found in mitochondria and have been proposed to regulate mitochondrial gene expression (Cogswell et al. 2003; Johnson et al. 2011). The AP-1 and PPARγ2 transcription factors have been proposed to localize to mitochondria and bind to the genome. (Casas et al. 2000; Ogita et al. 2003; Ogita et al. 2002) and the MEF2D transcription factor was found

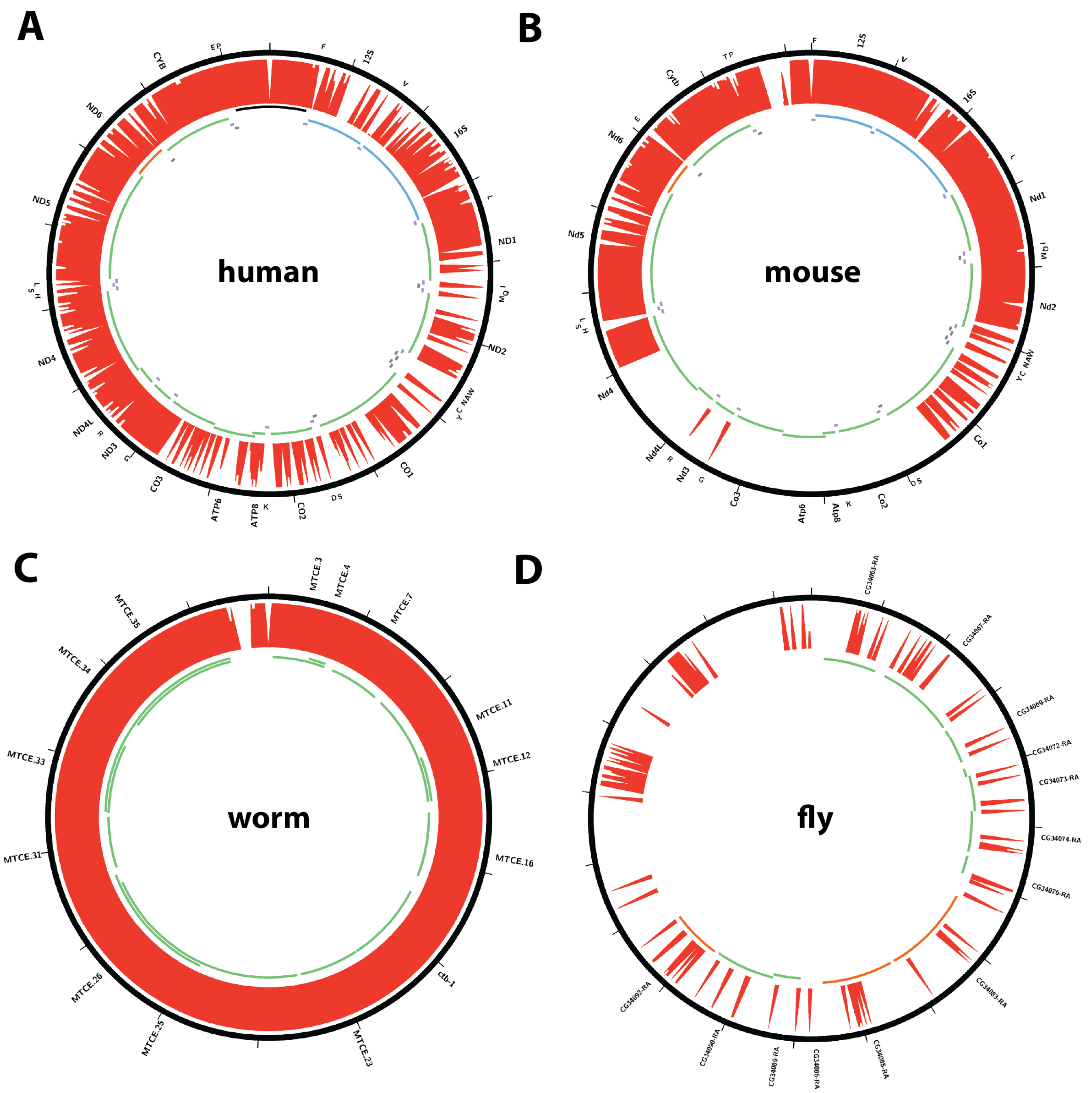
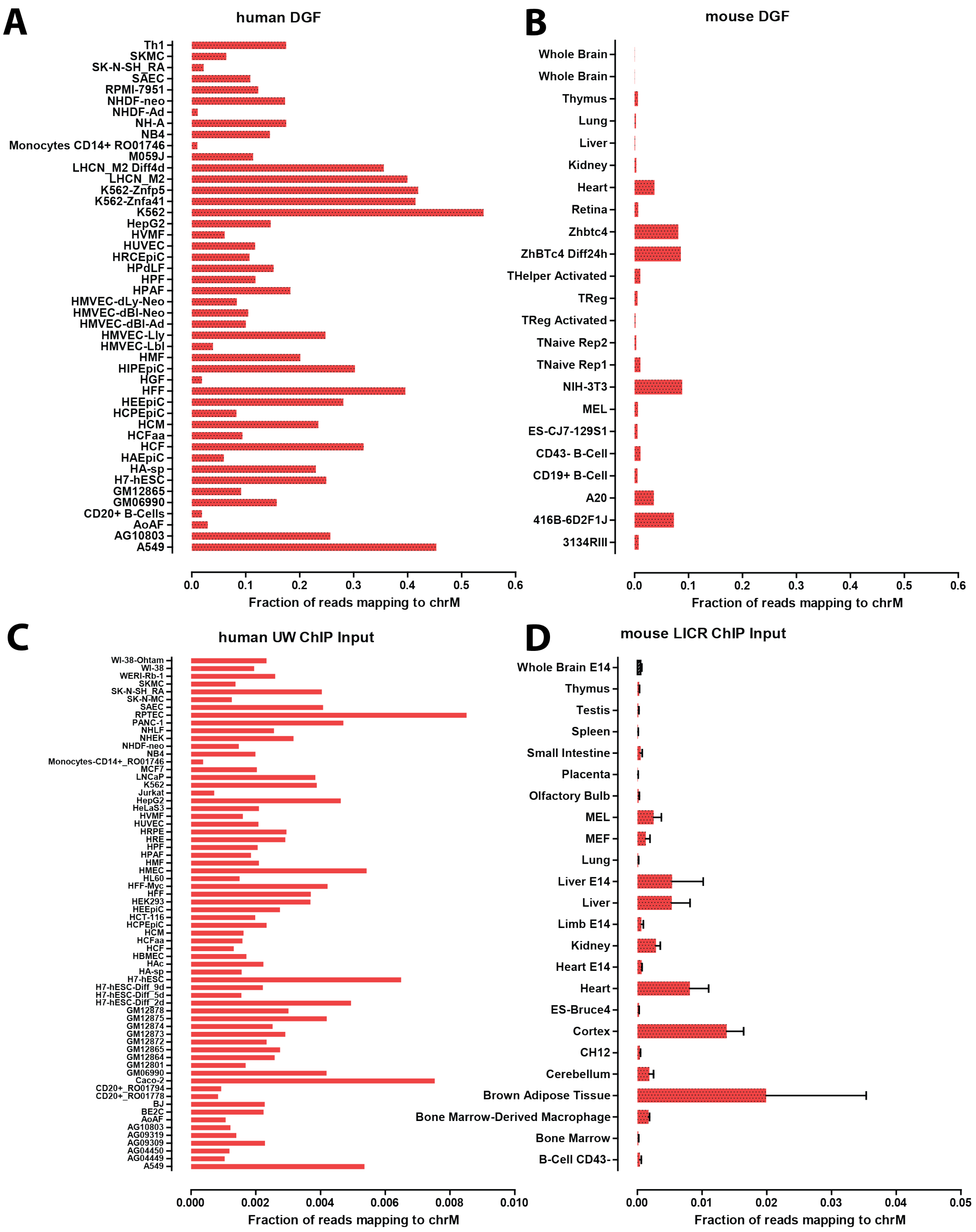


Figure 7.2: Unique mappability of the mitochondrial genome (chrM) in ENCODE and modENCODE species. (A) human; (B) mouse; (C) *C. elegans*; (D) *D. melanogaster*. The 36bp mappability track (see Methods for details) is shown. The annotated protein coding and rRNA and tRNA genes are shown in the inner circles as follows: forward-strand genes are shown as green lines, while reverse-strand genes are shown as red lines, with the exception of mouse and human rRNA and tRNAs (blue). The D-loop region in human is shown in black. Gene annotations were obtained from ENSEMBL (version 66). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

to regulate the expression of the ND6 gene by binding to a consensus sequence recognition motif within it (She et al. 2011). Finally, the presence of STAT3 in mitochondria has been found to be important for the function of the electron transport chains and also to be necessary for TNF-induced necroptosis (Szczepanek et al. 2011; Szczepanek et al. 2012a; Szczepanek et al. 2012b; Wegrzyn et al. 2009; Shulga & Pastorino 2012), although direct mtDNA binding has not been established. Mitochondrial localization has also been reported for STAT1 and STAT5 (Bo-


A
human DGF
Th1
SKMC
SK-N-SH_RA
SAEC
RPMI-7951
NHDF-neo
NHDF-Ad
NH-A
NB4
Monocytes CD14+ RO01746
M059J
LHCN_M2 Diff4d
LHCN_M2
K562-Znfp5
K562-Znfa41
K562
HepG2
HVMF
HUVEC
HRCEpiC
HPdLF
HPF
HPAF
HMVEC-dLy-Neo
HMVEC-dBl-Neo
HMVEC-dBl-Ad
HMVEC-Lly
HMVEC-Lbl
HMF
HIPEpiC
HGF
HFF
HEEpiC
HCPEpiC
HCM
HCFaa
HCF
HAEpiC
HA-sp
H7-hESC
GM12865
GM06990
CD20+ B-Cells
AoAF
AG10803
A549
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Fraction of reads mapping to chrM
B
mouse DGF
Whole Brain
Whole Brain
Thymus
Lung
Liver
Kidney
Heart
Retina
Zhbtc4
ZhBTc4 Diff24h
THelper Activated
TReg
TReg Activated
TNaive Rep2
TNaive Rep1
NIH-3T3
MEL
ES-CJ7-129S1
CD43- B-Cell
CD19+ B-Cell
A20
416B-6D2F1J
3134RIII
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Fraction of reads mapping to chrM
C
human UW ChIP Input
WI-38-Ohtam
WI-38
WERI-Rb-1
SKMC
SK-N-SH_RA
SK-N-MC
SAEC
RPTEC
PANC-1
NHLF
NHEK
NHDF-neo
NB4
Monocytes-CD14+_RO01746
MCF7
LNCaP
K562
Jurkat
HepG2
HeLaS3
HVMF
HUVEC
HRPE
HRE
HPF
HPAF
HMF
HMEC
HL60
HFF-Myc
HFF
HEK293
HEEpiC
HCT-116
HCPEpiC
HCM
HCFaa
HCF
HBMEC
HAc
HA-sp
H7-hESC
H7-hESC-Diff_9d
H7-hESC-Diff_5d
H7-hESC-Diff_2d
GM12878
GM12875
GM12874
GM12873
GM12872
GM12865
GM12864
GM12801
GM06990
Caco-2
CD20+_RO01794
CD20+_RO01778
BJ
BE2C
AoAF
AG10803
AG09319
AG09309
AG04450
AG04449
A549
0.000
0.002
0.004
0.006
0.008
0.010
Fraction of reads mapping to chrM
D
mouse LICR ChIP Input
Whole Brain E14
Thymus
Testis
Spleen
Small Intestine
Placenta
Olfactory Bulb
MEL
MEF
Lung
Liver E14
Liver
Limb E14
Kidney
Heart E14
Heart
ES-Bruce4
Cortex
CH12
Cerebellum
Brown Adipose Tissue
Bone Marrow-Derived Macrophage
Bone Marrow
B-Cell CD43-
0.00
0.01
0.02
0.03
0.04
0.05
Fraction of reads mapping to chrM

engler et al. 2010; Chueh et al. 2010).

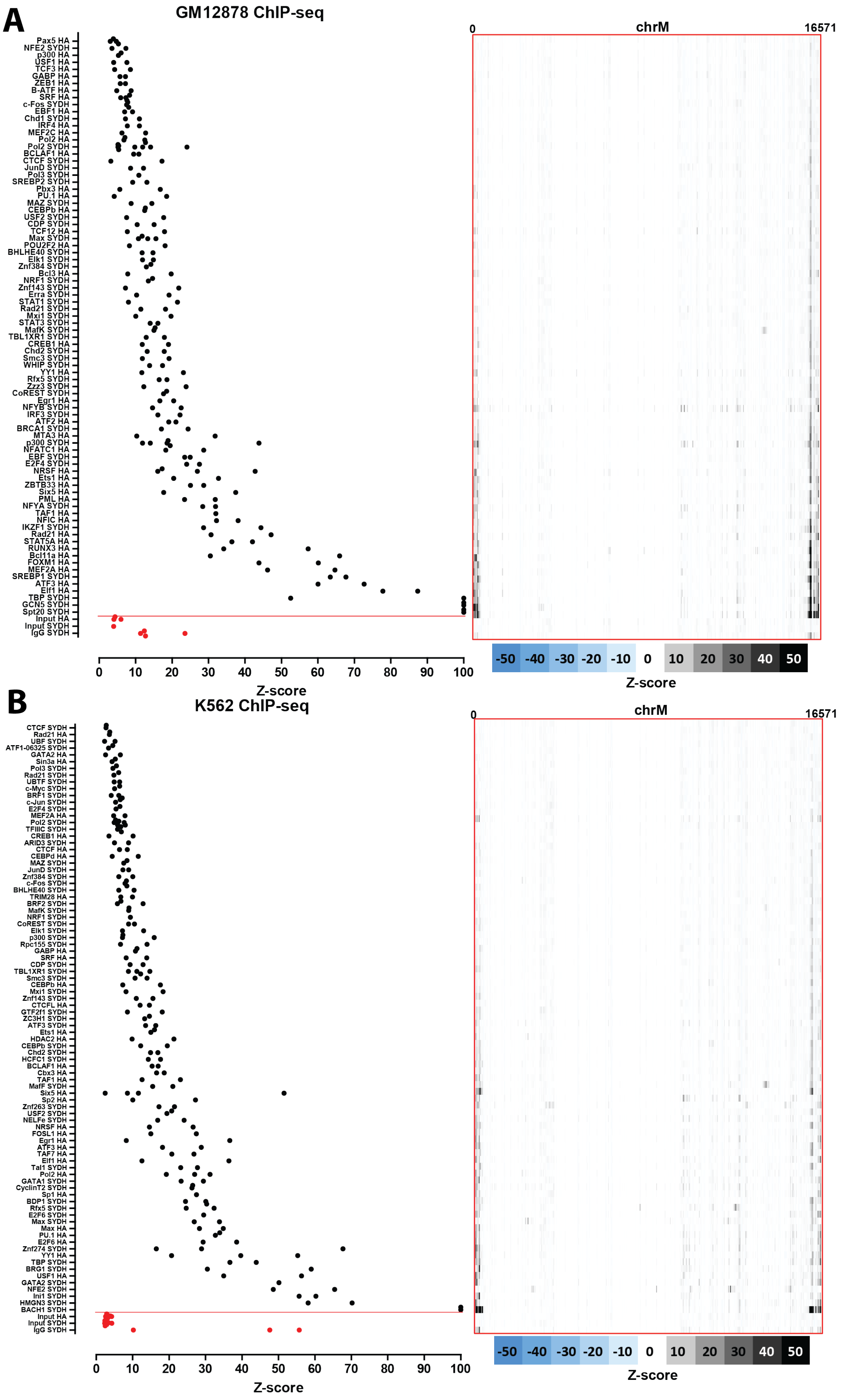
However, direct *in vivo* chromatin immunoprecipitation evidence for the binding of these factors to mtDNA exists only for CREB (Lee et al. 2005), p53 (Achanta et al. 2005) and MEF2D (She et al. 2011), and with the exception of MEF2D characterization is limited to the D-loop region. No prior studies have assayed transcription factor occupancy across the entire mitochondrial genome in vivo with modern high resolution techniques such as ChIP-seq (Chromatin Immunoprecipitation coupled with deep sequencing, (Johnson & Mortazavi et al. 2007). As a result, the precise nature, and in many instances the existence, of the proposed binding events remains unknown. The limited sampling of transcription factors in previous studies also leaves uncertain how common or rare localization to mitochondria and binding to mtDNA is for nuclear transcription factors in general.

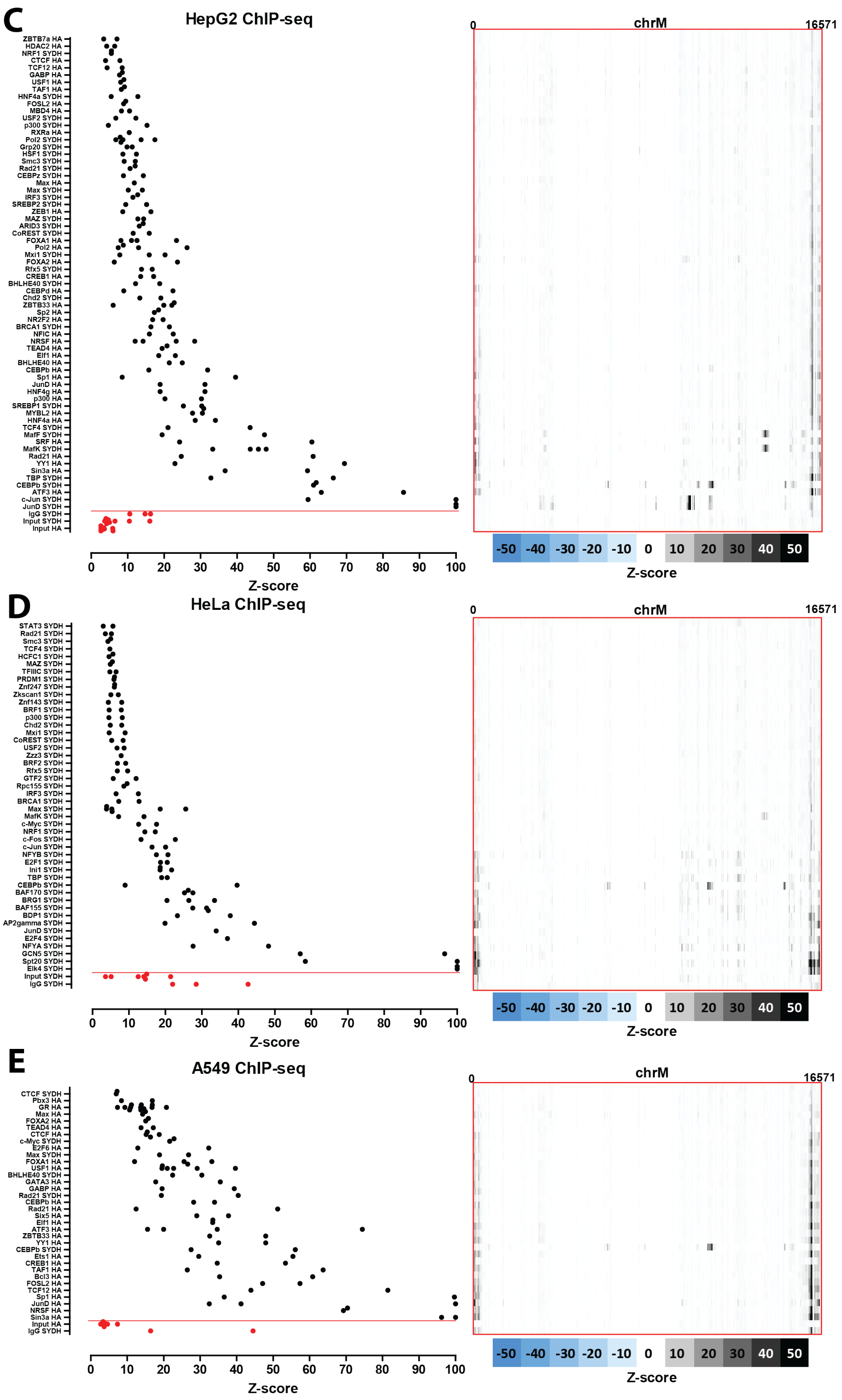
To address these questions, I surveyed the large compendium of ChIP-seq and other functional genomic data made publicly available by the ENCODE, mouseENCODE and modENCODE Consortia (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012; Gerstein et al. 2010; modENCODE Consortium 2010; Mouse ENCODE Consortium 2012) to identify transcription factors that associate directly with mtDNA and to characterize the nature of these interactions. This resulted in the identification of eight human and three mouse transcription factors for which robust evidence of site-specific occupancy in the mitochondrial genome exists. These sites exhibit the strand asymmetry typical of nuclear transcription factor binding sites, usually contain the recognition motifs for the factors in question, and are typically the strongest (as measured by ChIP-seq signal strength) binding sites found in both the nuclear and mitochondrial genome by a wide margin. Notably, these interactions are all found outside of the non-coding D-loop region. The D-loop region itself exhibits widespread sequencing read enrichment for dozens of transcription factors. However, it does not show the aforementioned feature characteristics of true binding events. Though not observed in control datasets generated from sonicated input DNA, the high ChIP-seq signal over the D-loop is frequently seen in control datasets generated using mock immunoprecipitation, suggesting that it is likely to represent an experimental artifact. Examination of available ChIP-seq data for the transcription factors previously proposed to play a role in mitochondria (GR, ERα, CREB, STAT3, p53) revealed no robust binding sites except for enrichment in the D-loop. Resolving the functional significance of the identified occupancy sites in future studies should provide exciting insights into the biology of both mitochondrial and nuclear transcriptional regulation.

7.2 Results

In the course of a study of TFAM occupancy in the mitochondrial and nuclear genomes (Wang et al. 2013), we noticed that a number of nuclear transcription factors exhibit localized enrichment in certain areas of the mitochondrial genome in ChIP-seq data (Figure 7.1). These events could be divided in two classes: high ChIP-seq signal over the NCR, and localized high read density over regions outside of it. Given prior reports suggesting that nuclear transcription factors might act in mitochondria, this prompted me to determine the general prevalence of the phenomenon among transcription factors and investigate evidence of occupancy in detail, as the power and resolution of ChIP-seq have not previously been brought to bear on this somewhat mysterious phenomenon. We took advantage of the wide compendium of human, mouse, fly and worm functional genomics data generated by the ENCODE (ENCODE Project

Figure 7.3 ***(preceding page)*****: Variation in mitochondrial DNA copy number in cell lines and tissues**. The fraction of reads mapping to the mitochondrial genome (chrM) is shown. (A,B) UW human (A) and mouse (B) UW ENCODE digital genomic footprinting (DGF) data; (C) UW human ChIP input datasets; (D) LICR mouse ChIP input datasets. "UW" and "LICR" refers to the ENCODE production groups that generated the data. Inputs from the UW and LICR groups were chosen because they are the largest ENCODE sets in terms of number of cell lines/tissues assayed by the same production groups, thus avoiding possible variation between different laboratories. A general positive correlation between the expected metabolic demand of the tissue type and the relative amount of reads mapping to chrM is observed.


A
GM12878 ChIP-seq
Pax5 HA
NFE2 SYDH
p300 HA
USF1 HA
TCF3 HA
GABP HA
ZEB1 HA
B-ATF HA
SRF HA
c-Fos SYDH
EBF1 HA
Chd1 SYDH
IRF4 HA
MEF2C HA
Pol2 HA
Pol2 SYDH
BCLAF1 HA
CTCF SYDH
JunD SYDH
Pol3 SYDH
SREBP2 SYDH
Pbx3 HA
PU.1 HA
MAZ SYDH
CEBPb HA
USF2 SYDH
CDP SYDH
TCF12 HA
Max SYDH
POU2F2 HA
BHLHE40 SYDH
Elk1 SYDH
Znf384 SYDH
Bcl3 HA
NRF1 SYDH
Znf143 SYDH
Erra SYDH
STAT1 SYDH
Rad21 SYDH
Mxi1 SYDH
STAT3 SYDH
MafK SYDH
TBL1XR1 SYDH
CREB1 HA
Chd2 SYDH
Smc3 SYDH
WHIP SYDH
YY1 HA
Rfx5 SYDH
Zzz3 SYDH
CoREST SYDH
Egr1 HA
NFYB SYDH
IRF3 SYDH
ATF2 HA
BRCA1 SYDH
MTA3 HA
p300 SYDH
NFATC1 HA
EBF SYDH
E2F4 SYDH
NRSF HA
Ets1 HA
ZBTB33 HA
Six5 HA
PML HA
NFYA SYDH
TAF1 HA
NFIC HA
IKZF1 SYDH
Rad21 HA
STAT5A HA
RUNX3 HA
Bcl11a HA
FOXM1 HA
MEF2A HA
SREBP1 SYDH
ATF3 HA
Elf1 HA
TBP SYDH
GCN5 SYDH
Spt20 SYDH
Input HA
Input SYDH
IgG SYDH
0
10
20
30
40
50
60
70
80
90
100
Z-score
0
chrM
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score
B
K562 ChIP-seq
CTCF SYDH
Rad21 HA
UBF SYDH
ATF1-06325 SYDH
GATA2 HA
Sin3a HA
Pol3 SYDH
Rad21 SYDH
UBTF SYDH
c-Myc SYDH
BRF1 SYDH
c-Jun SYDH
E2F4 SYDH
MEF2A HA
Pol2 SYDH
TFIIIC SYDH
CREB1 HA
ARID3 SYDH
CTCF HA
CEBPd HA
MAZ SYDH
JunD SYDH
Znf384 SYDH
c-Fos SYDH
BHLHE40 SYDH
TRIM28 HA
BRF2 SYDH
MafK SYDH
NRF1 SYDH
CoREST SYDH
Elk1 SYDH
p300 SYDH
Rpc155 SYDH
GABP HA
SRF HA
CDP SYDH
TBL1XR1 SYDH
Smc3 SYDH
CEBPb HA
Mxi1 SYDH
Znf143 SYDH
CTCFL HA
GTF2f1 SYDH
ZC3H1 SYDH
ATF3 SYDH
Ets1 HA
HDAC2 HA
CEBPb SYDH
Chd2 SYDH
HCFC1 SYDH
BCLAF1 HA
Cbx3 HA
TAF1 HA
MafF SYDH
Six5 HA
Sp2 HA
Znf263 SYDH
USF2 SYDH
NELFe SYDH
NRSF HA
FOSL1 HA
Egr1 HA
ATF3 HA
TAF7 HA
Elf1 HA
Tal1 SYDH
Pol2 HA
GATA1 SYDH
CyclinT2 SYDH
Sp1 HA
BDP1 SYDH
Rfx5 SYDH
E2F6 SYDH
Max SYDH
Max HA
PU.1 HA
E2F6 HA
Znf274 SYDH
YY1 HA
TBP SYDH
BRG1 SYDH
USF1 HA
GATA2 SYDH
NFE2 SYDH
Ini1 SYDH
HMGN3 SYDH
BACH1 SYDH
Input HA
Input SYDH
IgG SYDH
0
10
20
30
40
50
60
70
80
90
100
Z-score
0
chrM
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score


C
HepG2 ChIP-seq
ZBTB7a HA
HDAC2 HA
NRF1 SYDH
CTCF HA
TCF12 HA
GABP HA
USF1 HA
TAF1 HA
HNF4a SYDH
FOSL2 HA
MBD4 HA
USF2 SYDH
p300 SYDH
RXRa HA
Pol2 SYDH
Grp20 SYDH
HSF1 SYDH
Smc3 SYDH
Rad21 SYDH
CEBPz SYDH
Max HA
Max SYDH
IRF3 SYDH
SREBP2 SYDH
ZEB1 HA
MAZ SYDH
ARID3 SYDH
CoREST SYDH
FOXA1 HA
Pol2 HA
Mxi1 SYDH
FOXA2 HA
Rfx5 SYDH
CREB1 HA
BHLHE40 SYDH
CEBPd HA
Chd2 SYDH
ZBTB33 HA
Sp2 HA
NR2F2 HA
BRCA1 SYDH
NFIC HA
NRSF HA
TEAD4 HA
Elf1 HA
BHLHE40 HA
CEBPb HA
Sp1 HA
JunD HA
HNF4g HA
p300 HA
SREBP1 SYDH
MYBL2 HA
HNF4a HA
TCF4 SYDH
MafF SYDH
SRF HA
MafK SYDH
Rad21 HA
YY1 HA
Sin3a HA
TBP SYDH
CEBPb SYDH
ATF3 HA
c-Jun SYDH
JunD SYDH
IgG SYDH
Input SYDH
Input HA
0
10
20
30
40
50
60
70
80
90
100
Z-score
chrM
0
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score
D
HeLa ChIP-seq
STAT3 SYDH
Rad21 SYDH
Smc3 SYDH
TCF4 SYDH
HCFC1 SYDH
MAZ SYDH
TFIIIC SYDH
PRDM1 SYDH
Znf247 SYDH
Zkscan1 SYDH
Znf143 SYDH
BRF1 SYDH
p300 SYDH
Chd2 SYDH
Mxi1 SYDH
CoREST SYDH
USF2 SYDH
Zzz3 SYDH
BRF2 SYDH
Rfx5 SYDH
GTF2 SYDH
Rpc155 SYDH
IRF3 SYDH
BRCA1 SYDH
Max SYDH
MafK SYDH
c-Myc SYDH
NRF1 SYDH
c-Fos SYDH
c-Jun SYDH
NFYB SYDH
E2F1 SYDH
Ini1 SYDH
TBP SYDH
CEBPb SYDH
BAF170 SYDH
BRG1 SYDH
BAF155 SYDH
BDP1 SYDH
AP2gamma SYDH
JunD SYDH
E2F4 SYDH
NFYA SYDH
GCN5 SYDH
Spt20 SYDH
Elk4 SYDH
Input SYDH
IgG SYDH
0
10
20
30
40
50
60
70
80
90
100
Z-score
chrM
0
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score
E
A549 ChIP-seq
CTCF SYDH
Pbx3 HA
GR HA
Max HA
FOXA2 HA
TEAD4 HA
CTCF HA
c-Myc SYDH
E2F6 HA
Max SYDH
FOXA1 HA
USF1 HA
BHLHE40 SYDH
GATA3 HA
GABP HA
Rad21 SYDH
CEBPb HA
Rad21 HA
Six5 HA
Elf1 HA
ATF3 HA
ZBTB33 HA
YY1 HA
CEBPb SYDH
Ets1 HA
CREB1 HA
TAF1 HA
Bcl3 HA
FOSL2 HA
TCF12 HA
Sp1 HA
JunD HA
NRSF HA
Sin3a HA
Input HA
IgG SYDH
0
10
20
30
40
50
60
70
80
90
100
Z-score
chrM
0
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score

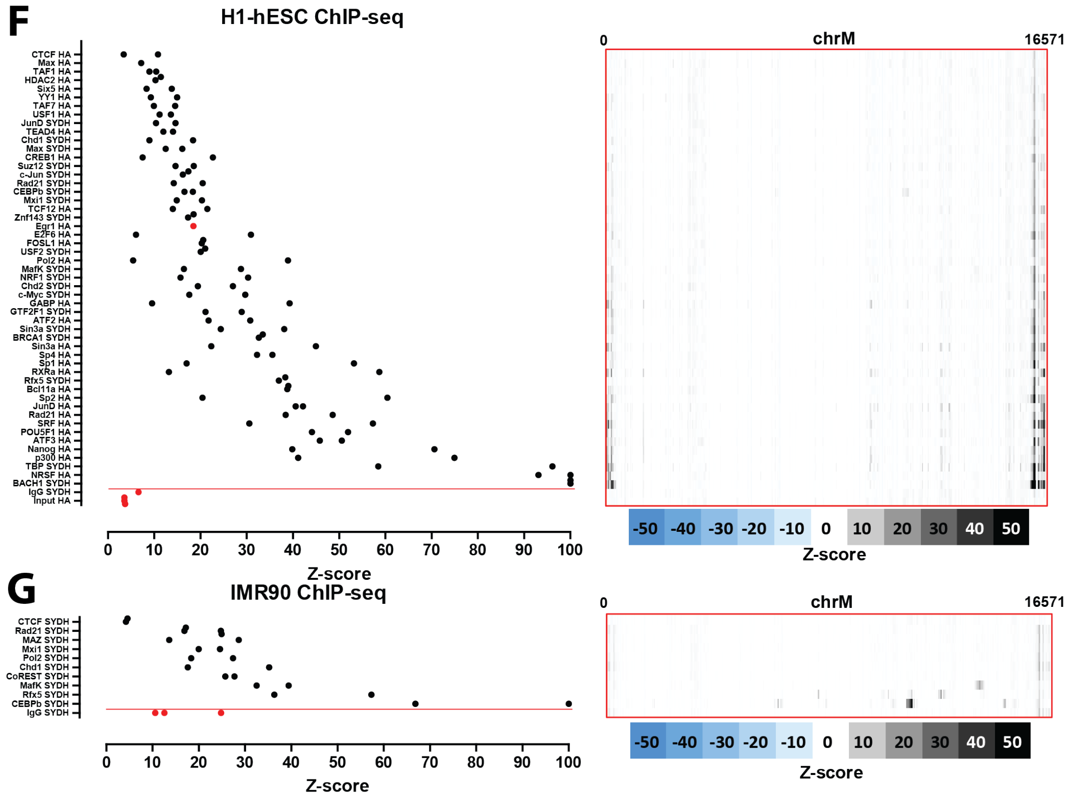


Figure 7.4: Signal distribution over the mitochondrial genome in human ChIP-seq datasets. The maximum z-score for each individual TF ChIP-seq replicate in each cell line is shown on the left (factors are sorted by average z-score, with control datasets always shown on the bottom in red, below the red horizontal line). The z-score profile along the mitochondrial chromosome for the replicate with the highest z-score is shown on the right. "SYDH" and "HA" refer to the ENCODE production groups which generated the data. Z-scores ≥100 are shown as equal to 100. (A) GM12878 cells; (B) K562 cells; (C) HepG2 cells; (D) HeLa cells; (E) A549 cells; (F) H1-hESC cells; (G) IMR90.

Consortium 2011; ENCODE Project Consortium 2012), mouseENCODE (Mouse ENCODE Consortium 2012) and modENCODE (Gerstein et al. 2010; modENCODE Consortium 2010) consortia.

7.2.1 Identifying transcription factor binding events in the mitochondrial genome

I downloaded publicly available (as of February 2012) ENCODE and mouseENCODE ChIP-seq and control data from the UCSC Genome Browser and modENCODE data from `ftp://ftp.modencode.org`, including ChIP-seq data for 151 transcription factors in human cell lines (Wang et al. 2012), 31 in mouse and 35 in *C.elegans* (see discussion on *D. melanogaster* below). I also downloaded DNase hypersensitvity (both DNase-seq (Thurman et al. 2012) and DGF (Neph et al. 2012)), FAIRE-seq (Song et al. 2012) and MNase-seq data as these datasets provide valuable orthogonal information about potentially artifactual patterns of read enrichment over the mitochondrial genome.

It is well known that the nuclear genome contains partial copies of the mitochondrial genome (NUMTs) (du Buy & Riley 1967; Hazkani-Covo E et al. 2010). Depending on their levels of divergence from the mitochondrial sequence, they can present an informatics challenge for distinguishing binding events to the true mitochondrial genome from binding events to NUMTs. For this reason, I aligned reads simultaneously

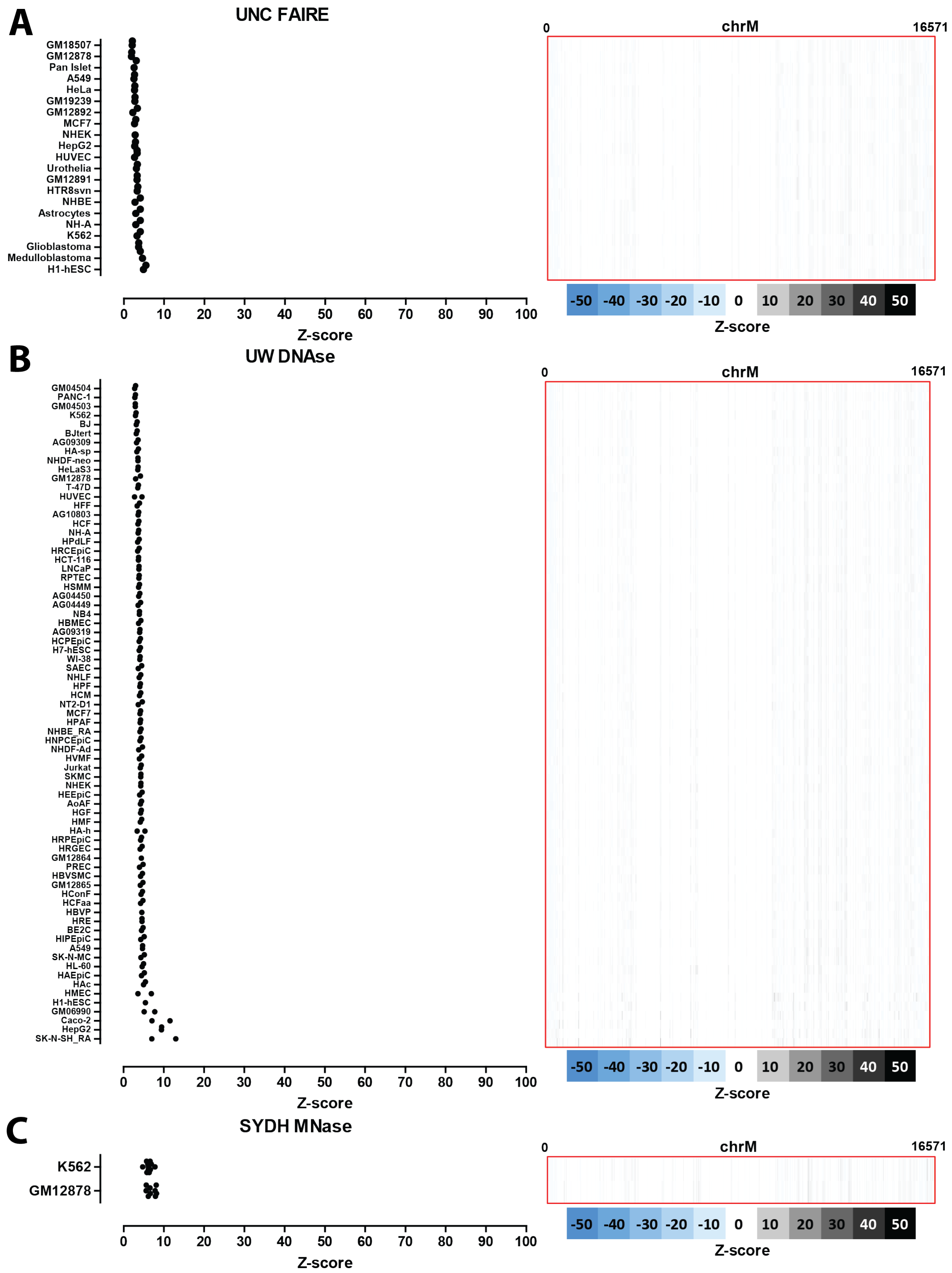


Figure 7.5: Signal distribution over the mitochondrial genome in human FAIRE-seq, DNAse-seq and MNAse-seq datasets. Shown is the maximum z-score for each individual replicate for each cell line (left) and the z-score profile along the mitochondrial chromosome for the replicate with the highest z-score (right). (A) FAIRE data; (B) DNAse data; (C) MNAse data. "UNC", "UW" and "SYDH" refer to the ENCODE production groups which generated the data. Z-scores larger than 100 are shown as 100. No read enrichment over the D-loop is observed, suggesting that the D-loop signal found in TF ChIP-seq datasets is not due to sequencing biases but is a result of the immunoprecipitation process.

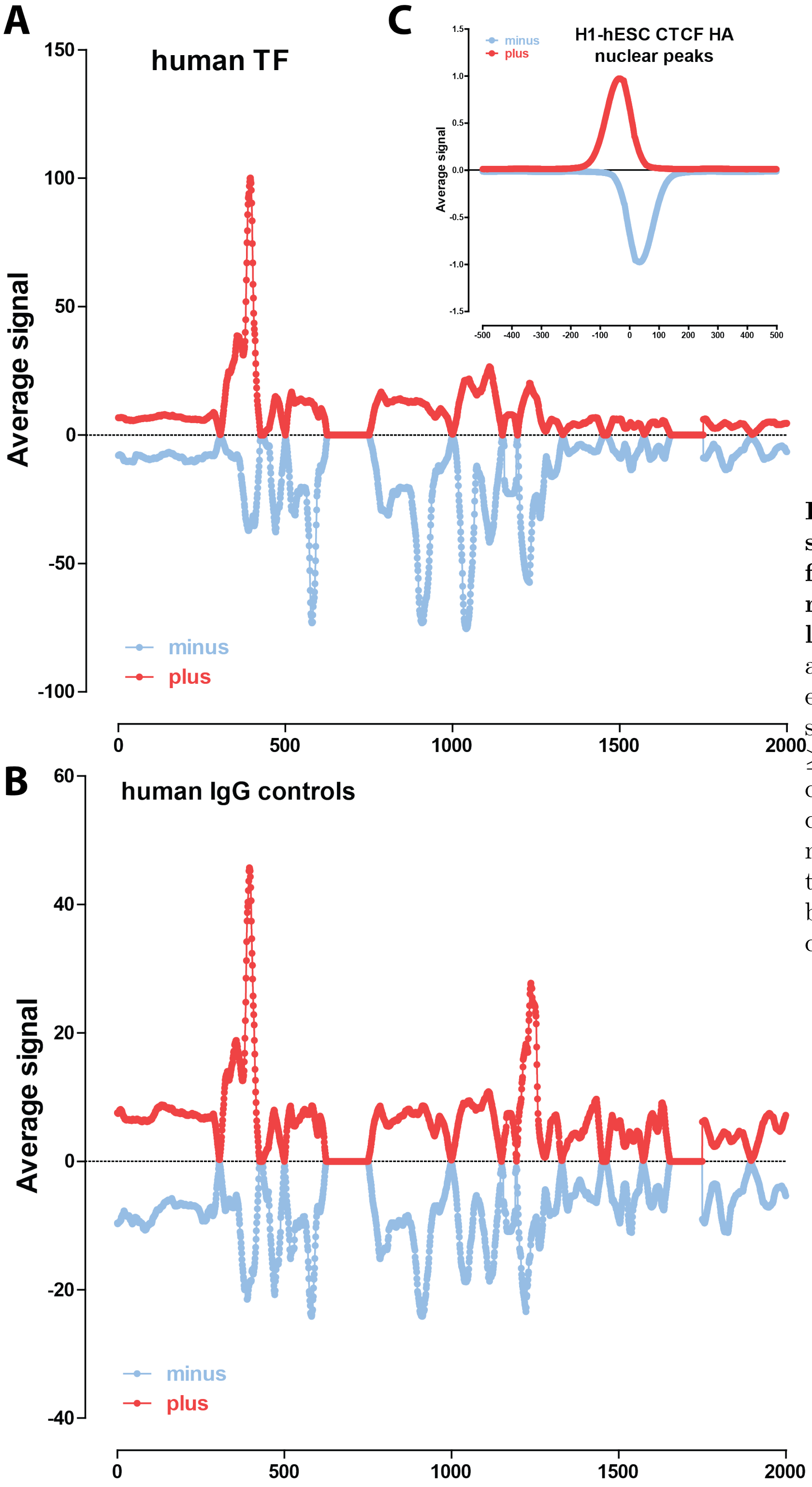
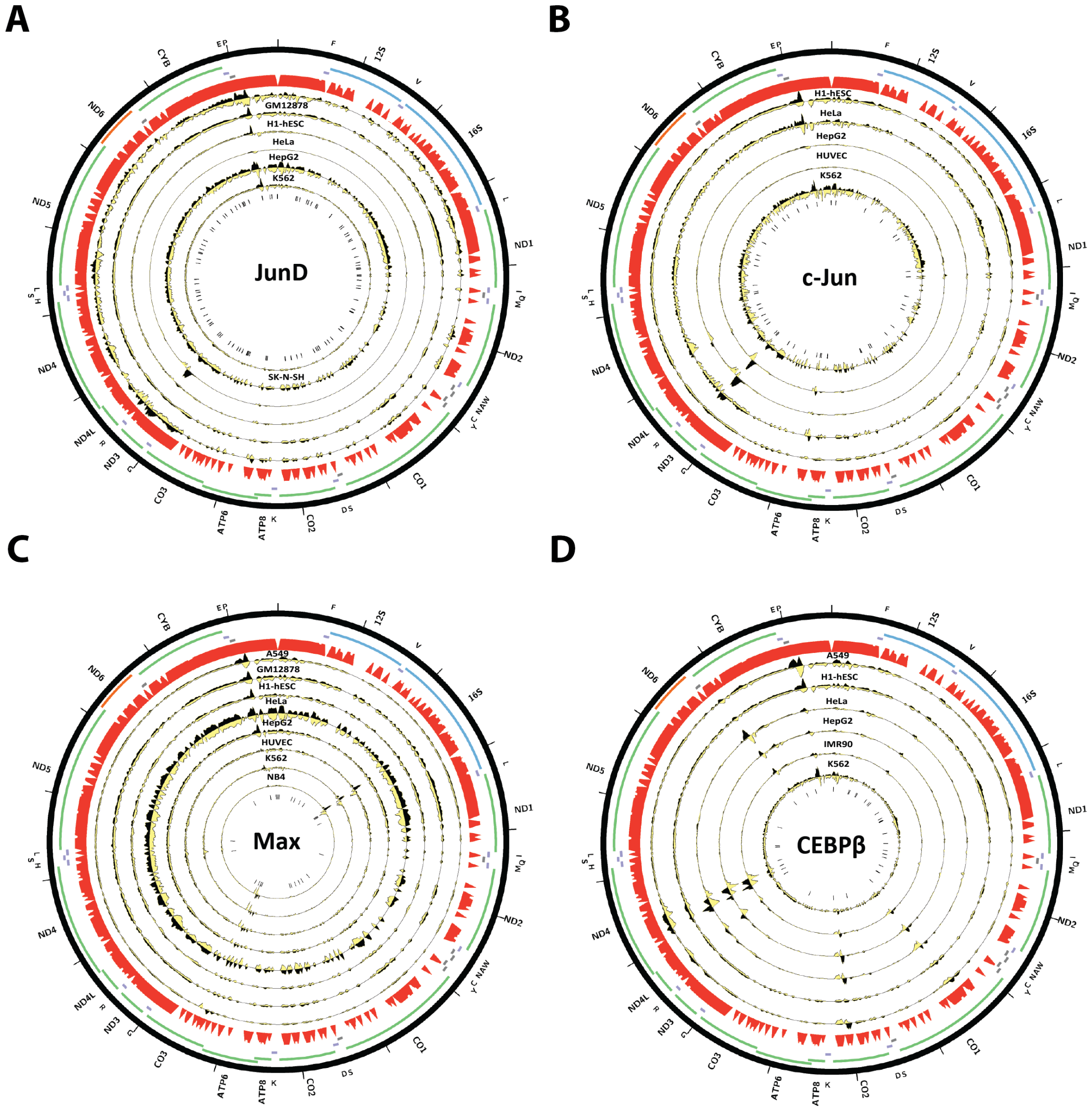


Figure 7.6: Combined signal distribution profile for the forward and reverse strand in the D-loop region. Shown is the average signal (in RPM) for each strand in human ChIP-seq datasets with z-scores ≥ 20 (A) and human IgG controls (B). Also shown for comparison is the plus and minus strand read distribution around nuclear CTCF binding sites in H1-hESC cells (C)

against the nuclear and mitochondrial genomes. I then retained only reads that map uniquely, and with no mismatches, relative to the reference for further analysis (see Methods for details). As a consequence this stringent mapping strategy, regions of the mitochondrial genome that are also present as perfectly identical copies in the nuclear genome are "invisible" to analysis; this was a necessary compromise in order to focus only on a maximally stringent set of putative mitochondrial binding events. However, before proceeding, I examined how widely affected the mitochondrial genome is by this treatment in the four relevant species by generating mappability tracks (shown in Figure 7.2). The human mitochondrial genome contains numerous small islands of unmappable sequence, particularly concentrated between the ND1 and CO3 genes, but it displays no large completely unmappable segments (Figure 7.2A). The mouse genome contains a large unmappable stretch between the CO1 and ND4 genes (Figure 7.2B). The *C. elegans* mitochondrial genome is almost completely uniquely mappable (Figure 7.2C). In contrast, the *D. melanogaster* genome is almost completely unmappable, indicating the presence of very recent insertions into the nuclear genome with high sequence similarity. Fly datasets were therefore excluded from further analysis and I focused on human, mouse and worm data.

Mammalian cells typically contain hundreds to thousands of copies of mtDNA, with the pre-



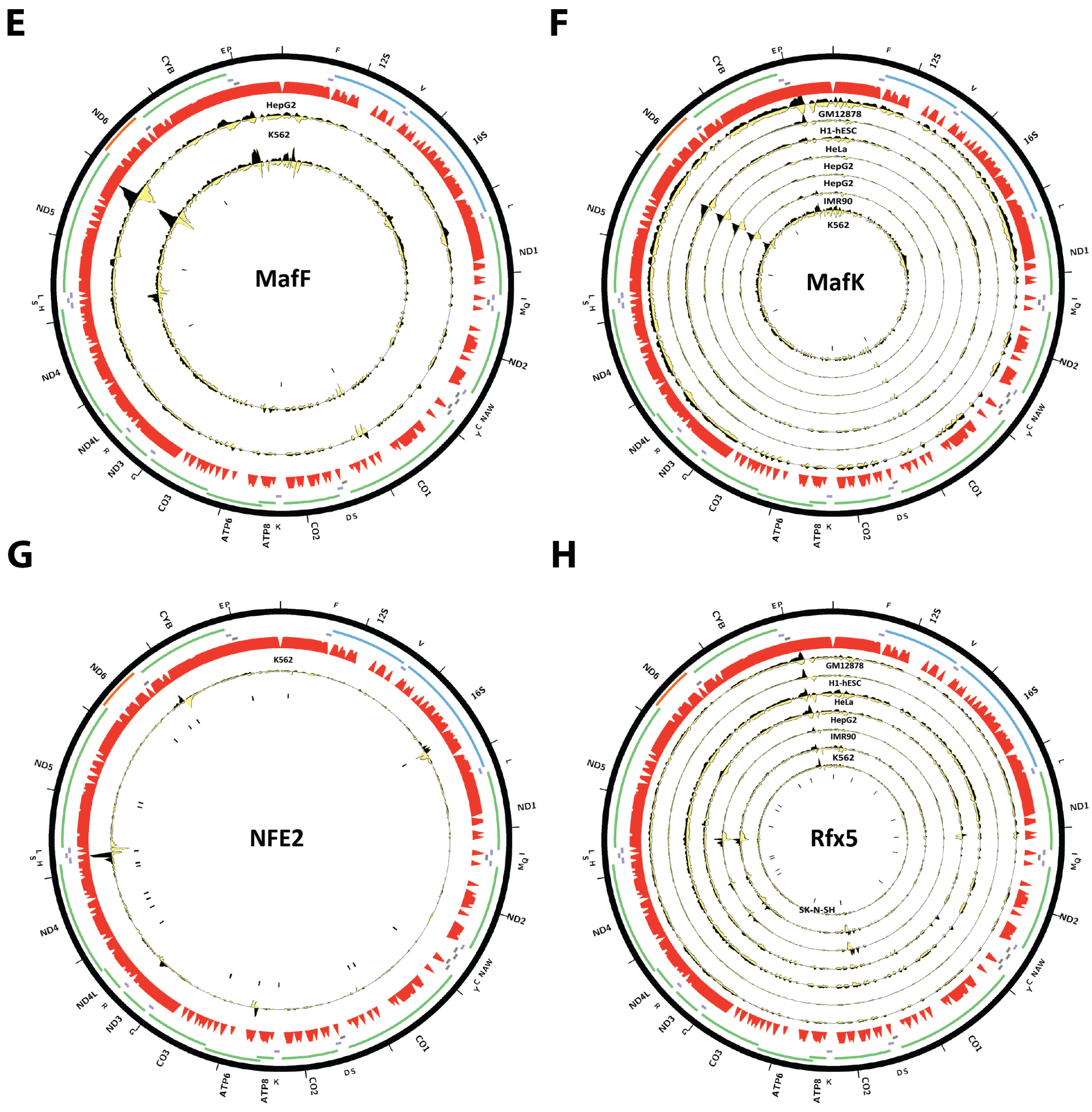


Figure 7.7: Human transcription factors with canonical ChIP-seq peaks (displaying the typical strand asymmetry in read distribution around the putative binding site) outside of the D-loop. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) CEBPβ; (B) c-Jun; (C) MafF; (D) MafK (note that MafK has been assayed using two different antibodies in HepG2, both of which are shown); (E) NFE2; (F) Rfx5. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

cise number varying depending on the metabolic needs of the particular cell type (Bogenhagen & Clayton 1974; Williams 1986; Satoh & Kuroiwa 1991). This variation is relevant to analysis because the relative read density over the mitochondrial genome is expected to scale with

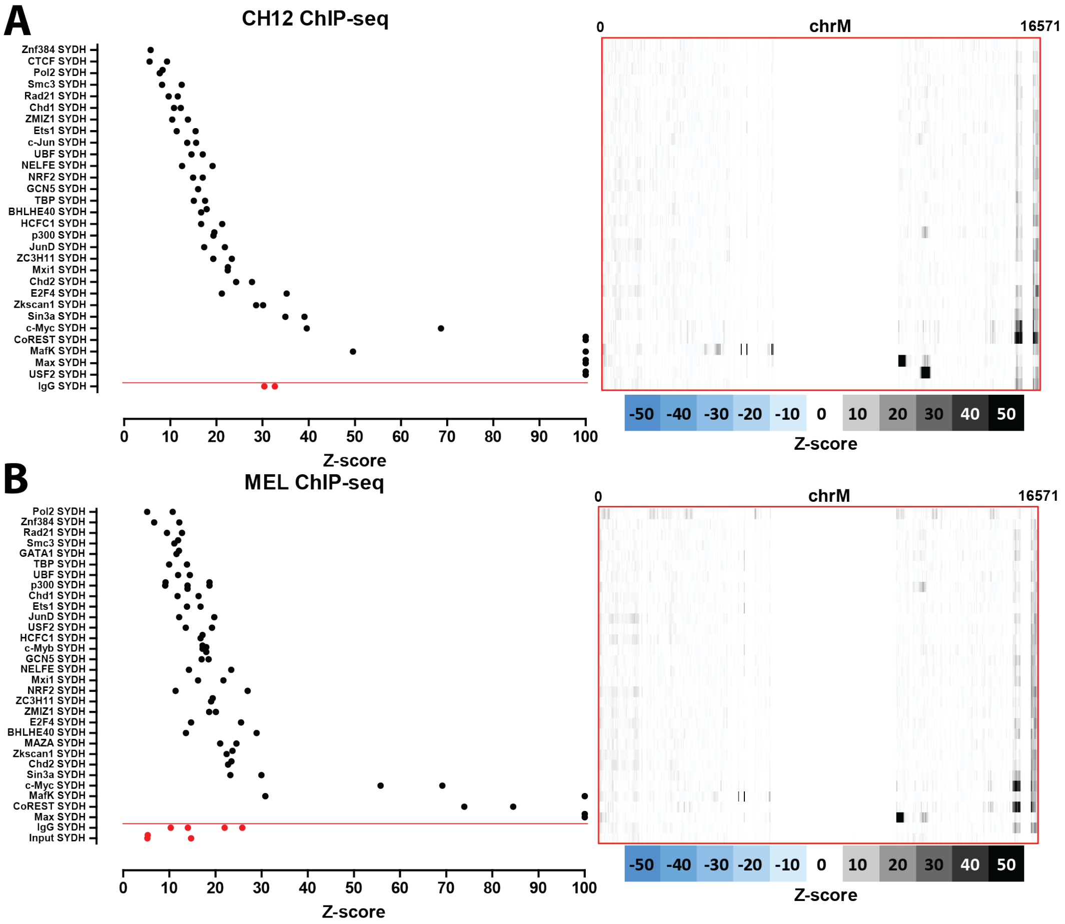


Figure 7.8: Signal distribution over the mitochondrial genome in mouse ChIP-seq datasets. Shown is the maximum z-score for each individual replicate for each cell line (left) and the z-score profile along the mitochondrial chromosome for the replicate with the highest z-score (right). Control datasets are shown in red on the bottom, below the red horizontal line. (A) CH12 cells; (B) MEL cells.

the mtDNA:nuclear DNA ratio for a given cell. Thus, cell types with very high mtDNA copy number are expected to display correspondingly elevated background read density over the mitochondrial genome. Several types of ENCODE data provide a rough proxy for the relative mitochondrial genome copy number per cell. In particular, the fraction of reads originating from the mitochondrial genome in DNase hypersensitivity and ChIP control datasets is expected to scale accordingly. I examined the distribution of this fraction in ENCODE and mouseENCODE DGF datasets and observed very large differences between different cell lines and tissues (Figure 7.3). For example, about half of reads in K562 DGF data originated from mitochondria, while the fraction was less than 2% in CD20+ B-cells (Figure 7.3A). Notably, these differences are in many cases (though not always) consistent with what is known about the cell lines, with certain cancer cell lines (such as K562 and A549) and muscle cells (LHCN) showing the largest number of mitochondrial reads, while primary cells with small volumes of cytoplasm such as B-cells showed the least.

Mouse DGF data was available mostly for tissues, and the fraction of mitochondrial reads in these was much smaller compared to both the human cell lines and the few mouse cell lines assayed (Figure 7.3B). This is consistent with a significant proportion of cells in tissues being in a less active metabolic state than cell lines in culture. Still, some expected differences between tissues were observed. For example, one of the tissues that was most enriched for reads mapping to the mitochondrial genome was the heart. Similarly large differences were observed in ChIP control datasets (Figure 7.3CD), although the absolute number of reads was much lower than it was in DGF data. Again, the mouse tissues with the highest number of mitochondrial reads were the more metabolically active ones, such as

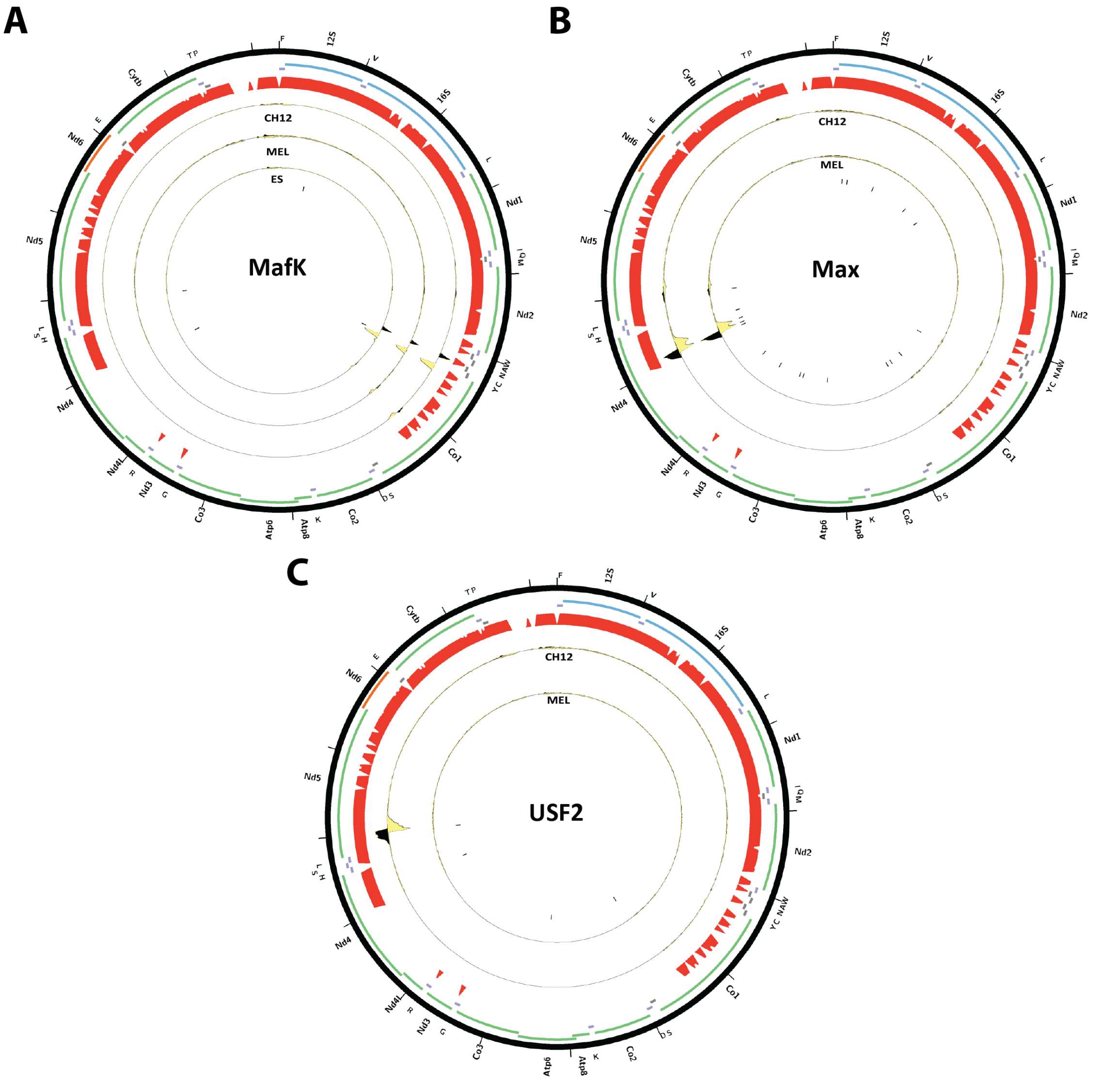


Figure 7.9: Mouse transcription factors with canonical ChIP-seq peaks (displaying the typical strand asymmetry in read distribution around the putative binding site) outside of the D-loop. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) MafK (note that the putative binding site is found in a region that is not completely mappable, thus the read profiles loses the canonical shape but the strand asymmetry is nevertheless apparent and a motif is present); (B) Max; (C) USF2. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

brown adipose tissue, cortex, and heart.

These large differences in background read coverage between different cells lines/tissues have two consequences for the analysis of putative transcription factor binding to the mitochondrial genome. First, peak calling algorithms usually used to identify transcription factor binding sites from ChIP-seq data may not work equally well in different cell lines due to the highly variable background read density. Second, these differences render comparing the strength of binding across cell lines difficult.

I therefore devised a normalization procedure (described in Methods) to convert read coverage to signal intensity z-scores reflecting how strongly regions of enrichment stand out compared to the average background read density along the mitochondrial genome for each dataset. I then used the maximum z-scores for each dataset to identify datasets with very strong such enrichment, which I then examined manually in detail.

7.2.2 Nuclear transcription factor binding to the mitochondrial genome in human cell lines

The distribution of read density z-scores for transcription factor ChIP-seq and control datasets in seven ENCODE human cell lines (GM1278, K562, HepG2, HeLa, H1-hESC, IMR90 and A549) is shown in Figure 7.4. A wide range in the values of the maximum z-score is observed, from less than 5, to more than 100. Strikingly, most factors exhibit high read density in the NCR. One obvious explanation for this observation is that it represents an experimental artifact. This is likely, as the NCR contains the D-loop (Shadel & Clayton 1997), the unique triple-strand structure of which could conceivably either cause overrepresentation of DNA fragments originating from it in sequencing libraries or it could be non-specifically bound by antibodies during the immunoprecipitation process. To distinguish between these possibilities, I carried out the same analysis on DNase, FAIRE and MNase data. As these assays do not involve an immunoprecipitation step, they are a proper control for sequencing artifacts. I did not observe significant localized read enrichment in these datasets (Figure 7.5), suggesting that the observed read enrichment over the D-loop is not due to sequencing biases or overrepresentation of D-loop fragments in ChIP libraries. Similarly, I did not observe enrichment in the matched sonicated input ChIP-seq control datasets. However, a number of mock-immunoprecipitation (IgG) control datasets did exhibit high z-scores (up to >50 in K562 cells) and closely matched the signal profile over the D-loop of ChIP-seq datasets (Figure 7.6B). We also examined the forward and reverse strand read distribution in the NCR (Figure 7.6). Site-specific transcription factor binding events display a characteristic asymmetry in the distribution of reads mapping to the forward and reverse strands, with reads on the forward strand showing a peak to the left of the binding site and reads on the reverse strand showing a peak to the right of it (Kharchenko et al. 2008) (Figure 7.6C). Such read asymmetry was not observed in the D-loop region (average profile shown in Figure 7.6A, individual dataset profile shown in Figure 7.1, and also in Figures 7.7 and 7.13).

These results suggest that while immunoprecipitation is necessary for high enrichment over the D-loop, the enrichment might not be mediated by the proteins targeted by the primary antibody. This does not explain why a large number of factors show little enrichment over the D-loop (Figure 7.4) and why some factors show enrichment that is much higher than that observed in K562 IgG controls, with z-scores of up to 300 (compared to a maximum of 50 for the most highly enriched IgG controls). Still, given the lack of clear hallmarks of site-specific occupancy, and the IgG control results, enrichment over the D-loop has to be provisionally considered to be primarily the result of an experimental artifact, even if it cannot be ruled that at least in some cases it is the result of real biochemical association with nuclear transcriptional regulators.

In contrast to the widespread, but likely artifactual, read enrichment over the D-loop, I observed strong enrichment, exhibiting the canonical characteristics of a ChIP-seq peak over a true transcription factor binding site, in other regions of the human mitochondrial genome for eight of the examined transcription factors using a minimum z-score threshold of 20: CEBPβ, c-Jun, JunD, MafF, MafK, Max, NFE2 and Rfx5. Figure 7.7 shows the forward and reverse strand read distribution for representative replicates of each factor in each assayed cell line, as well as the occurrences of the corresponding explanatory motifs (identified from the top 500 ChIP-seq peaks in the nuclear genome, see Methods for details). The putative binding sites outside of the D-loop are characterized by an asymmetric forward and reverse strand read distribution, and in most cases, the presence of the explanatory motif in a position consistent with binding by the factor. I identified multiple binding sites for CEBPβ: a strong site of enrichment around the 5' end of the CYB gene, what seems to be two closely clustered sites in the ND4 gene, a weaker site in the ND4L gene, and two other regions of enrichment over CO2 and CO1 (Figure 7.7D). A single very strong binding site over the ND3 gene was observed for c-Jun, as well as two weaker sites, one coinciding with the ND4 CEBPβ sites and one near the 5' end of ATP6 (Figure 7.7B); the strong ND3 site was also observed for JunD in HepG2 cells. Max exhibited two putative binding sites: one in the middle of the 16S rRNA gene, containing a cluster of Max motifs, and another one around the 5' end of CO3, which also contains a cluster of Max motifs but is in a region of poor mappability. A common and very strong MafK and MafF binding site is present near the 3' end of ND5, though it does not contain the common explanatory motif for both factors (Figure 7.7CD). Several putative binding sites were identified for NFE2: one close to the CEBPβ site in the 5'end of CYB, one over the tRNA cluster between ND4 and ND5, one in the 5' end of ATP6 and one in the 16S rRNA gene (Figure 7.7D). Finally, two putative binding sites ar observed for Rfx5, at the 5' end of ND5 and in the middle of CO2 (Figure 7.7E). Intriguingly, these binding events are not always present in all cell lines. For example, CEBPβ binding around CYB was absent in K562, A549 and H1-hESC cells, while the MafK ND5 binding site was absent in GM18278 and H1-hESC cells, but present in the other cell lines for which data is available.

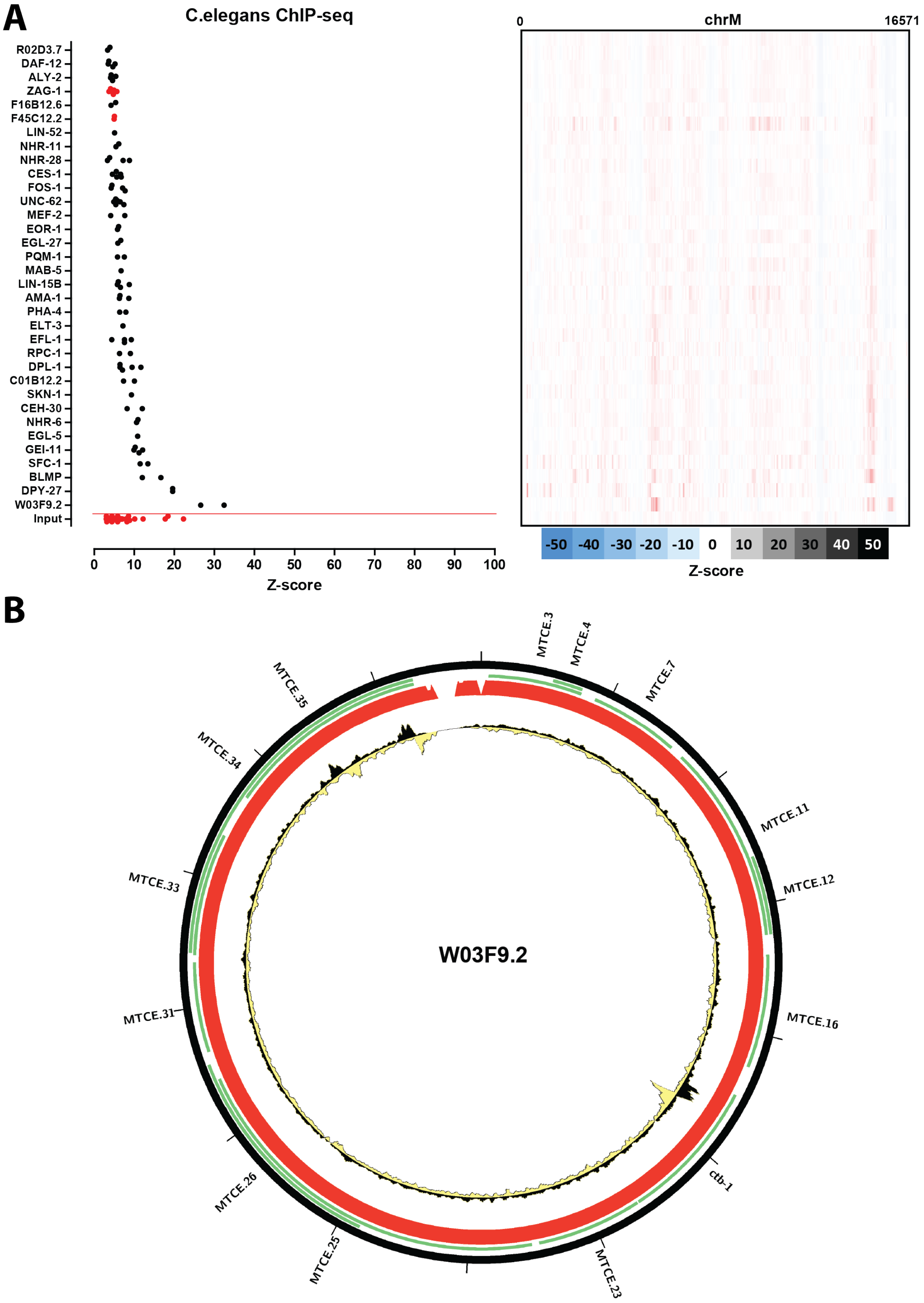
7.2.3 Nuclear transcription factor occupancy to the mitochondrial genome in model organisms

I carried out the same analysis as described above on mouse and *C. elegans* ChIP-seq datasets. Figure 7.8 shows the distribution of read density z-scores in mouse CH12 and MEL cells. Similarly to the human data, I observed widespread but probably artifactual read enrichment over the D-loop. In addition to that, we saw that three transcription factors (Max, MafK, and USF2) also exhibit strong enrichment elsewhere in the mitochondrial genome (Figure 7.9). I observed a single MafK binding site, containing the explanatory motif and situated over the tRNA cluster between the ND2 and CO1 genes (Figure 7.9A). Max displayed a strong binding site (possibly a cluster of closely spaced binding sites) in the ND4 gene, and a weaker binding site near the 5' end of ND5; both sites contained the explanatory motif (Figure 7.9B). Finally, a single site, also containing the explanatory motif for the factor and situated near the ND5 Max site, was present in CH12 USF2 datasets (but not in MEL cells) (Figure 7.9C). MafK and Max were also assayed in human cells, and, as discussed above, putative mitochondrial sites were identified there for both, though not at obviously orthologous to those found in the mouse data positions in the genome. I also analyzed available ChIP-seq data for the mouse orthologs of c-Jun and JunD, which in human cells exhibited putative mitochondrial binding sites. In contrast to observation in human, I did not detect strong sites for either protein in mouse.

Unlike the mouse and human datasets, most *C. elegans* ChIP-seq datasets did not show very strong enrichment over the mitochondrial genome (Figure 7.10A), with the exception of DPY-27 and W03F9.2. Of these, only W03F9.2 exhibited regions of enrichment with the characteristics of transcription factor binding sites (Figure 7.10B); however, very little is known about this protein and the significance of its binding to the mitochondrial genome is unclear.

7.2.4 ChIP-seq signal is significantly stronger over mitochondrial occupancy sites than it is over nucleus sites

The occupancy observations reported above for human and mouse mitochondria do not formally rule out the possibility that there are unannotated NUMTs in the genomes of the cell lines in which binding is detected in our analysis and the observed binding is in fact nuclear. Such an explanation is superficially likely, given that binding to the mitochondrial genome was observed in some cell lines and not in others. However, closer examination reveals that this hypothesis would require different NUMTs in different cell lines as the cell lines that lack binding are not the same for all factors. For example, MafF and MafK binding is very prominent in K562 cells but CEBPβ and c-Jun seem not to bind


A
C.elegans ChIP-seq
R02D3.7
DAF-12
ALY-2
ZAG-1
F16B12.6
F45C12.2
LIN-52
NHR-11
NHR-28
CES-1
FOS-1
UNC-62
MEF-2
EOR-1
EGL-27
PQM-1
MAB-5
LIN-15B
AMA-1
PHA-4
ELT-3
EFL-1
RPC-1
DPL-1
C01B12.2
SKN-1
CEH-30
NHR-6
EGL-5
GEI-11
SFC-1
BLMP
DPY-27
W03F9.2
Input
0
10
20
30
40
50
60
70
80
90
100
Z-score
0
chrM
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score
B
W03F9.2
MTCE.3
MTCE.4
MTCE.7
MTCE.11
MTCE.12
MTCE.16
ctb-1
MTCE.23
MTCE.25
MTCE.26
MTCE.31
MTCE.33
MTCE.34
MTCE.35

to mtDNA in those cells. While still possible, we consider the independent insertion of multiple partial NUMTs in different cell lines to be an unlikely explanation for the observed binding patterns.

Each chromosome in the nuclear genome exists as only two copies in diploid cells, as compared to the hundreds of mitochondria, each of which may contain multiple copies of the mitochondrial genome (Satoh & Kuroiwa 1991; Bogenhagen & Clayton 1974), and although cancer cells may exhibit various aneuploidies and copy number variants, the number of mtDNA copies is still expected to be much higher. Thus, higher read density over mitochondrial transcription factor binding sites than over nuclear ones is expected, assuming similar occupancy rates. I therefore used the strength of ChIP-seq signal over mitochondrial occupancy sites in order to test the hypothesis that they are in fact nuclear, and not mitochondrial in origin. I compared the peak height (in RPM) of the top 10 nuclear peaks (peak calls generated by the ENCODE consortium were downloaded from the UCSC Genome Browser) with that of the putatively mitochondrial binding sites (Figure 7.11). I found that the mitochondrial binding sites are usually the strongest binding sites by a wide margin, or at least within the top three of all peaks. For example, while the strongest nuclear MafK peak in mouse CH12 cells has a peak height of 14.5 RPM, the mitochondrial binding site has a peak height of 290 RPM. These observations are difficult to explain as being the result of binding to unannotated NUMTs in the nuclear genome, but are entirely consistent with the hypothesis that these factors indeed bind to the large number of copies of the mitochondrial genome present in each cell.

7.2.5 Evidence for localization of transcription factors to mitochondria

If the observed binding sites in ChIP-seq data are the result of actual association of nuclear transcription factors with mtDNA, then these transcription factors should exhibit mitochondrial localization. We directly tested this by performing immunocytochemistry (ICC) for MafK in HepG2 cells (Figure 7.12). It is important to note that such an assay for localization to mitochondria is potentially difficult to interpret if binding is the result of only a few protein molecules entering mitochondria, which would not yield sufficient signal for interpretation via ICC. However, strikingly, we observe clear colocalization of MafK to mitochondira in 60% of cells ($n = 124$). These observations provide independent corroboration for the mtDNA binding events identified through ChIP-seq.

7.2.6 No robust mitochondrial occupancy in ChIP-seq data for most previously reported mitochondrially targeted nuclear factors

I note that none of the factors previously reported to be localized to mitochondria and to bind to mtDNA was retrieved by our analysis, even though CREB, GR, ERα, IRF3, NFκB, STAT1, STAT5A and STAT3 were assayed by the ENCODE Consortium. This failure could be attributed to the use of too stringent a z-score threshold when selecting datasets with significant enrichment. I therefore examined available ChIP-seq data against these factors more carefully (Figure 7.13, Figure 7.14). I also performed the same analysis on published mouse and human p53 ChIP-seq data (Kenzelmann Broz et al. 2013; Li et al. 2012; Aksoy et al. 2012)

Figure 7.10 *(preceding page)*: Signal distribution over the mitochondrial genome in *C.elegans* ChIP-seq datasets. (A) Shown is the maximum z-score for each individual replicate for each cell line (left) and the z-score profile along the mitochondrial chromosome for the replicate with the highest z-score (right). Control datasets are shown in red on the bottom, below the red horizontal line; (B) Forward and reverse strand read distribution over the *C.elegans* mitochondrial genome for W03F9.2 (“Young Adult” stage). Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Plots generated using Circos version 0.60 (Krzywinski et al. 2009).

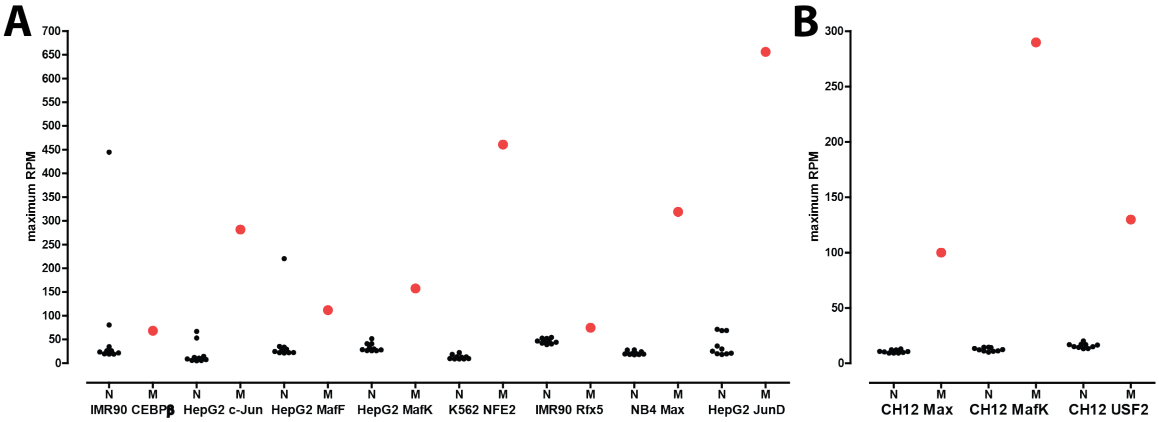


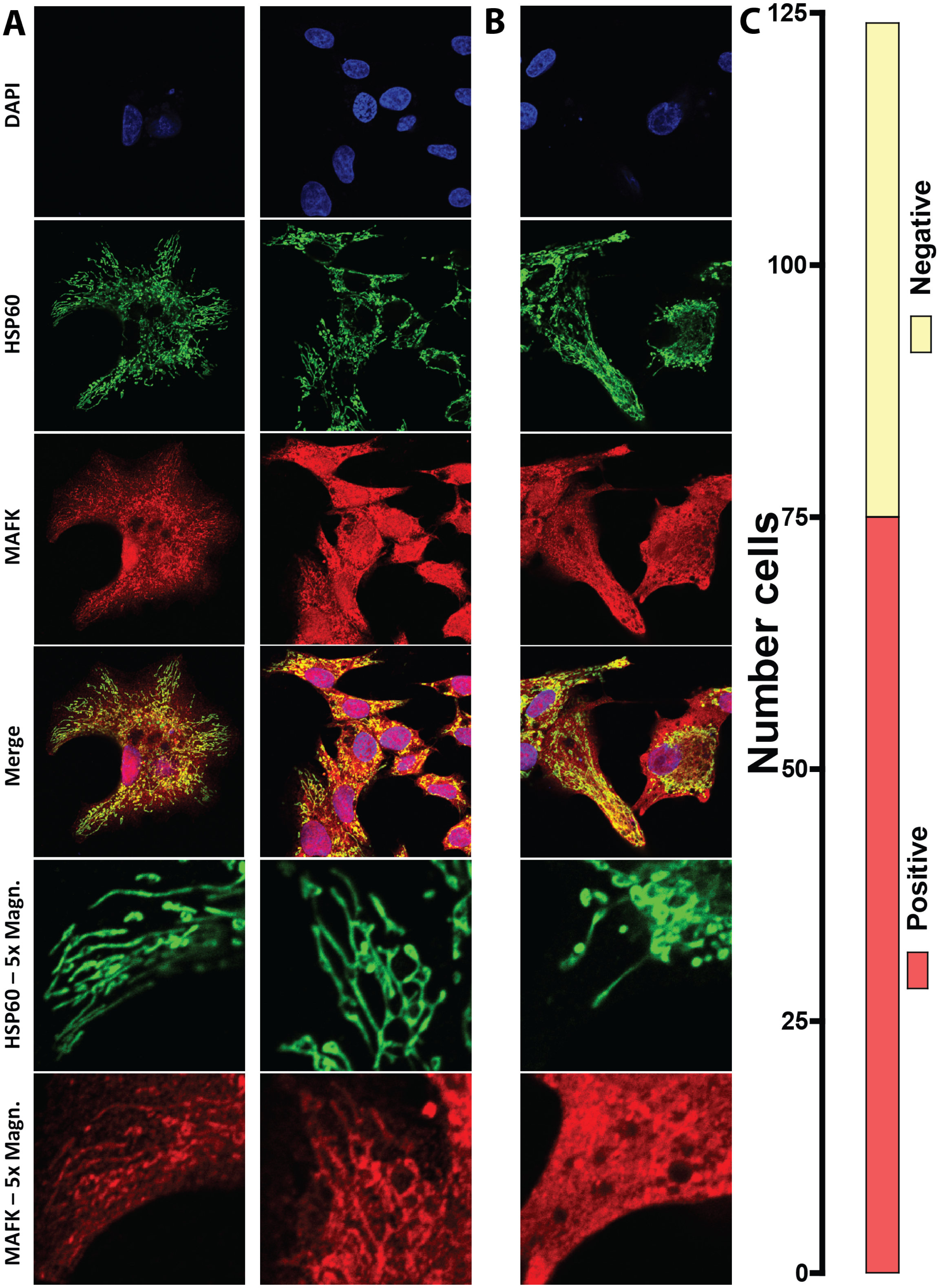
Figure 7.11: Mitochondrial ChIP-seq peaks are generally significantly stronger than nuclear peaks. Shown is the maximum signal (in RPM) for the top 10 nuclear peaks ("N", smaller black dots), and the maximum signal intensity (also in RPM) in the mitochondrial genome ("M", larger red dot) for representative ChIP-seq datasets for each factor. (A) Mouse datasets (B) Human datasets.

(Figure 7.15). Again, I did not observe any major sites of enrichment outside of the D-loop. For these factors, the D-loop region exhibits the same putatively artifactual pattern discussed previously. And for STAT3 and p53, even the enrichment over the D-loop was low. The one factor for which binding to mtDNA is confirmed by ChIP-seq is MEF2D, data for two of the isoforms of which in mouse C2C12 myoblasts was recently published (Sebastian et al. 2013) (Figure 7.16). It exhibits a very complex binding pattern over large portions of the mouse mitochondrial genome, which is not straightforward to interpet, but nevertheless a number of locations exhibit strand asymmetry and contain the MEF2 sequence recognition motif. Notably, most of these are outside the ND6 gene.

It is at present not clear how to interpret these discrepancies. It is not surprising that some of these factors do not exhibit binding to mtDNA, as they were reported to play a role in mitochondrial biology through mechanisms other than regulating gene expression (for example, IRF3 and STAT3). However, this is not the case for all of them. One possibility is that many prior studies reporting physical association of transcription factors with the D-loop suffered from the same artifactual read enrichment over that region that we observe, but this would not have been noticeable using the methods of the time. This would not be surprising, as it is only apparent that D-loop enrichment is likely to be artifactual when the high spatial resolution of ChIP-seq is combined with the joint analysis of input and mock immunoprecipitation controls. However, the mitochondrial localization of these factors has been carefully documented in a number of cases (Cammarota et al. 1999; Casas et al. 1999; De Rasmo et al. 2009). Another possiblity is that binding to mtDNA only occurs under certain physiological conditions and the factors were assayed using ChIP-seq only in cellular states not matching those. Further analysis of ChIP-seq data collected over a wide range of conditions should help resolve these issues.

7.3 Discussion

I present here the first large-scale characterization of the association of nuclear transcription factors along the entire mitochondrial genome by utilizing the vast ChIP-seq data resource made publicly available by the ENCODE and modENCODE consortia. I find two classes of signal enrichment events, neither of which is detected in high-throughput sequencing datasets that do not involve immunoprecipitation and therefore they are not due to sequencing biases. First, the majority of factors for which we detect strong read enrichment over the mitochondrial genome display high ChIP-seq signal only over the D-loop non-coding region in both human and mouse datasets. However, these signals do not have the characteristics of sequence


A
B
C
DAPI
HSP60
MAFK
Merge
HSP60 – 5x Magn.
MAFK – 5x Magn.
Number cells
125
100
75
50
25
0
Negative
Positive

specific occupancy and are present in a number of mock-immunoprecipitation control datasets. They are thus best explained as experimental artifacts, although it remains possible that they represent real non-canonical association with the D-loop for some factors. Second, for a subset of factors, specific ChIP-seq peaks are observed outside of the D-loop, and these display the additional hallmark characteristics of sequence specific occupancy.

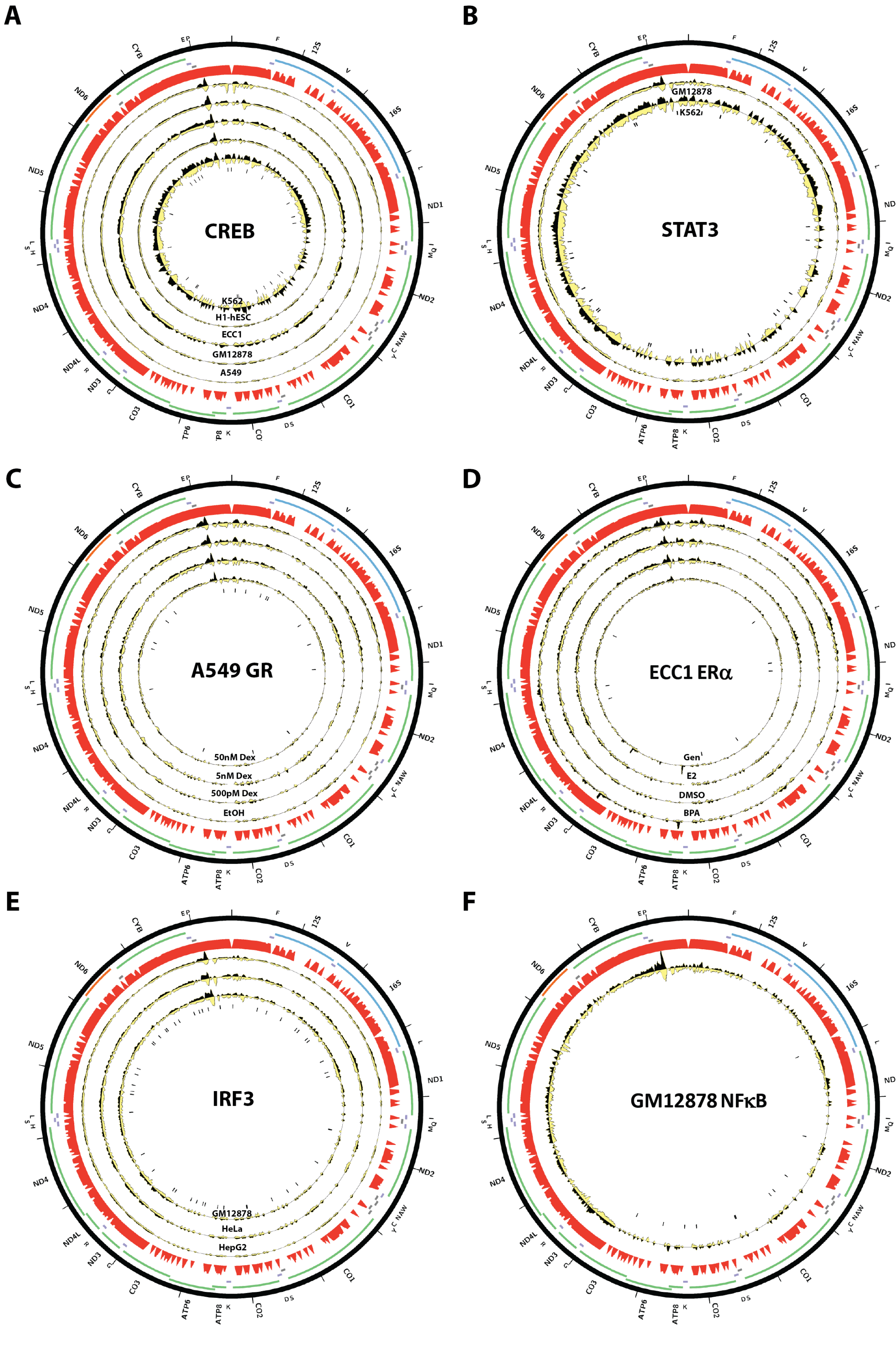
Nuclear transcription factors previously reported to localize to mitochondria either did not exhibit significant enrichment in the available ChIP-seq datasets or, when they did, it was over the D-loop region with similar non-specific read distribution shape as other factors. In contrast, applying conservative thresholds I found eight human and three mouse transcription factors (two in common between the two species) that strongly occupy sites outside of the D-loop. They display the strand asymmetry pattern around the putative binding site that typifies true nuclear ChIP-seq peaks. Even more convincing is the fact that the explanatory motif for the factor is usually found under the observed enrichment peaks, further suggesting that they correspond to true in vivo biochemical events.

There are three main explanations for these observations. First, it is possible that despite our considerable bioinformatic precautions the observed binding events are in fact nuclear, originating from NUMTs present in the genomes of the cell lines assayed, but absent from the reference genome sequence. I believe that this is very unlikely. An experimental argument against unknown NUMTs comes from the strength of the ChIP-seq signal that is seen in the mitochondrial genome. These signals are much higher than even the strongest peaks in the nuclear genome for the same factor in the same dataset. This is expected for true mitochondrial genome binding because of the presence of many copies of the mitochondrial genome per cell, in contrast to the presence of only two copies of the nuclear genome. Second, it is possible that mitochondria are sometimes lysed in vivo, with mitochondrial DNA spilling into the cytoplasm where transcription factors could then bind. This cannot be ruled out based on the ChIP data alone but we consider it unlikely, as this would need to happen with a sufficient frequency to explain the remarkable strength of mitochondrial occupancy sites. The third and most plausible interpretation is that these nuclear transcription factors indeed translocate to the mitochondria and interact with the genome, as has been observed for the D-loop in some previous studies for other factors. Indeed, immunocytochemistry experiments in our study confirm the presence of MafK in mitochondria in a majority of HepG2 cells.

Several major questions are raised by these results. First, it is not clear how these nuclear transcription factors are targeted to the mitochondria. Mitochondrial proteins are typically imported into the mitochondrial matrix through the TIM/TOM protein translocator complex, and are targeted to the organelle by a mitochondrial localization sequence, which is cleaved upon import. We scanned both human and mouse versions of our factors for mitochondrial target sequences (MTS) with both Mitoprot (Claros & Vincens 1996) and TargetP (Emanuelsson et al. 2007) (using default settings), but we were unable to identify significant matches using either. This seems to be a common feature of nuclear transcription factors previously found to localize to mitochondria, most of which lack import sequences and are instead imported through other means (Casas et al. 1999; Szczepanek et al. 2012b). Posttranslational modifications may be important for import, as has been demonstrated for STAT3 in TNF-induced necroptosis (Shulga et al. 2012).

Second, it is unclear why the same factor binds detectably to the mitochondrial genome in some cell types but not in others. It is certainly possible that different splice isoforms or post-translationally modified proteins are present in different cell types, with only some capable of being imported into mitochondria, or that import into mitochondria only happens under cer-

Figure 7.12 *(preceding page)*: Localization of MafK to the mitochondria (A) Immunocytochemistry showing MafK localization in HepG2 cells. Mitochondria were identified by HSP60 staining. Shown are two representative images of cells showing that MAFK localizes strongly to the nucleus and mitochondria, and exhibits diffuse staining in the cytoplasm. In 60% of cells (C), there is colocalization of HSP60 with MAFK staining at an intensity higher than that of the surrounding cytoplasm. (B) An example of a cell exhibiting only nuclear and cytoplasmic MAFK localization.


A
CREB
K562
H1-hESC
ECC1
GM12878
A549
B
STAT3
GM12878
K562
C
A549 GR
50nM Dex
5nM Dex
500pM Dex
EtOH
D
ECC1 ERα
Gen
E2
DMSO
BPA
E
IRF3
GM12878
HeLa
HepG2
F
GM12878 NFκB
E P
F
12S
V
16S
L
ND1
I
M Q
ND2
Y C N A W
CO1
D S
CO2
K
ATP8
ATP6
CO3
G
ND3
R
ND4L
ND4
S H L
ND5
ND6
CYB

tain physiological conditions only met in some cell lines.

Third, the question of the biochemical reality of transcription factor binding at the D-loop remains open. Previous studies understandably focused on the D-loop, given its well-appreciated importance in regulating mitochondrial transcription. As a consequence, the literature supporting a role for some nuclear factors in mitochondria suggests that they do so through binding to the D-loop. Our analysis of ChIP-seq data, which was carried out in an agnostic manner, revealed that dozens of transcription factors – many more than had been studied locally at the D-loop alone – also show high level of enrichment over the D-loop. However, the observed enrichment has characteristics suggesting that these signals are mainly due to experimental artifacts. In support of this judgment, the explanatory motifs for most of these factors were generally not found under the area of strongest enrichment in the D-loop. Therefore a conservative interpretation is that enrichment over the D-loop is an artifact in most cases.

Finally, and most importantly, the functional significance of factor occupancy observed by ChIP-seq remains unknown. It is entirely possible that it represents biochemical noise, with transcription factors entering the mitochondria because they have the right biochemical properties necessary to be imported, then binding to mtDNA but with little functional consequence. Alternatively, nuclear transcription factors may in fact be playing a regulatory role in mtDNA. It is difficult to imagine the exact mechanisms through which they might be acting, aside from interactions with the regulatory D-loop. While I do observe pairs of related factor such as c-Jun and JunD, and MafK and MafF binding to the same sites, binding events are overall widely dispersed over the mitochondrial genome and are found outside of the known regulatory regions. Plausible regulatory relationships are therefore not obvious and our results suggest that biological noise should be the working null hypothesis explaining the data. The functional regulatory role of these nuclear transcription factors in mitochondria is a very exciting possibility but it will have to be demonstrated in subsequent studies. Direct functional tests are the golden standard for establishing regulatory relationships, using gain and loss of function experiments and genetic manipulation of putative regulatory sites. The latter is at present not possible for mitochondria while the former are difficult to interpret in the case of the role of nuclear transcription factors in mitochondrial gene regulation, as it is not easy to separate the direct effects of binding to mtDNA from the indirect effects of transcriptional changes in the nucleus. Thus, it may be some time before definitive answers to these questions are obtained. In the meantime, larger compendia of transcription factor ChIP-seq data such as those expected to be generated by the next phase of the ENCODE project will be a primary source of further insight by providing binding data for additional nuclear transcription factors that will clarify allowed or preferred occupancy patterns across the mitochondrial genome.

7.4 Materials and Methods

7.4.1 Sequencing read alignment

Raw sequencing reads were downloaded from the UCSC genome browser for ENCODE and

Figure 7.13 *(preceding page)*: Distribution of reads over the human mitochondrial genome for factors previously reported to bind to mitochondria in ENCODE ChIP-seq data. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) CREB; (B) STAT3; (C) GR in A549 cells treated with different concentrations of dexamethasone (Dex) (Reddy et al. 2009; Reddy et al. 2012); (D) ERα in untreated (DMSO) ECC1 cells and ECC1 cells treated with bisphenol A (BPA), genistein (Gen) or 17β-estradiol (E2) (Gertz et al. 2012); (E) IRF3; (F) NFκB in GM12878 cells treated with TNFα (Kasowski et al. 2010). The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

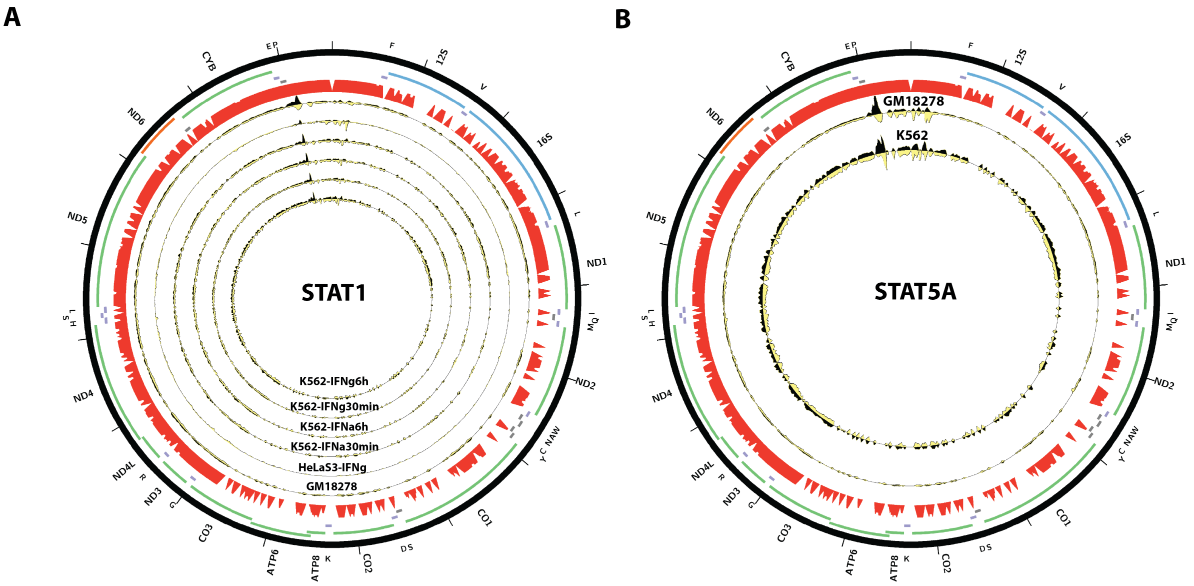


Figure 7.14: Distribution of reads over the human mitochondrial genome for STAT1 and STAT5A in ENCODE ChIP-seq data. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) STAT1; (B) STAT5A; The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

mouseENCODE (Mouse ENCODE Consortium 2012) data, and from `ftp://ftp.modencode.org` for modENCODE data (Gerstein et al. 2010; modENCODE Consortium 2010) (data current as of February 2012). ChIP-seq data for p53 was obtained rom GEO series GSE26361 (Li et al. 2012), GSE46240 (Kenzelmann Broz et al. 2013) and GSE42728 (Aksoy et al. 2012). Reads were aligned using Bowtie (Langmead et al. 2009), version 0.12.7. Human data was mapped against either the female or the male set of human chromosomes (excluding the Y chromosome and/or all random chromosomes and haplotypes) depending on the sex of the cell line (where the sex was known, otherwise the Y chromosome was included), genome version `hg19`. Mouse data was mapped against the `mm9` version of the mouse genome. modENCODE *D. melanogaster* data was mapped against the `dm3` version of the fly genome. modENCODE data for *C. elegans* was mapped against the `ce10` version of the worm genome. Reads were mapped with the following settings: ``‘‘-v 2 -k 2 -m 1 -t --best --strata’’``, which allow for two mismatches relative to the reference, however for all downstream analysis only reads mapping uniquely and with zero mismatches were considered, to eliminate any possible mapping artifacts.

7.4.2 Mappability track generation

Mappability was assessed as follows. Sequences of length N bases were generated starting at each position in the mitochondrial genome. The resulting set of “reads” was then mapped against the same bowtie index used for mapping real data. Positions covered by N reads were considered fully mappable. In this case, $N = 36$ as this is the read length for most of the sequencing data analyzed in this study.

7.4.3 Signal normalization of ChIP-seq data over the mitochondrial genome

Because the number of mitochondria per cell varies from one cell line/tissue to another, direct comparisons between datasets based on the absolute magnitude of the signal in RPM are not entirely valid. For this reason, we normalized the signal as follows. For each dataset, we

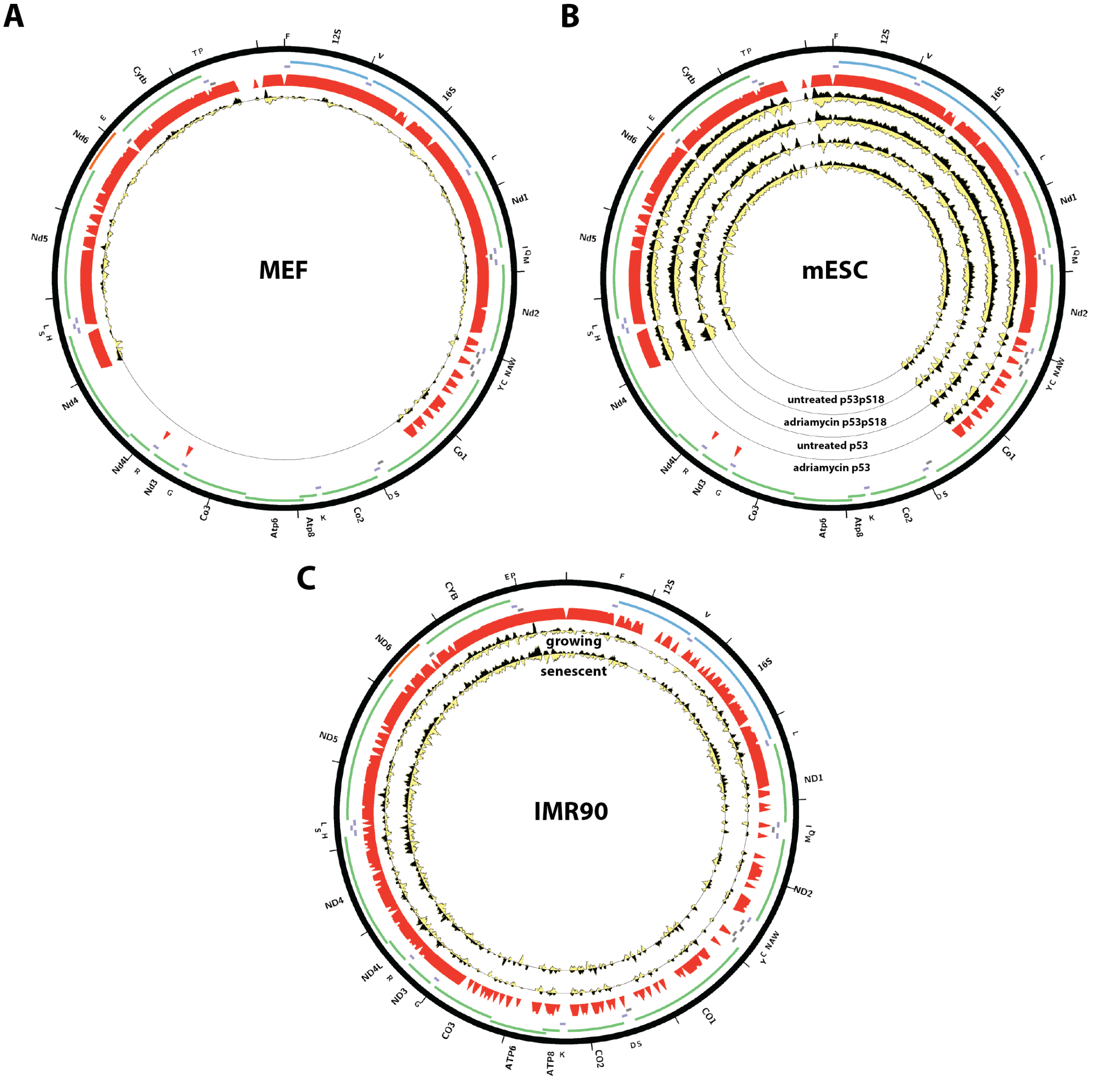


Figure 7.15: Distribution of reads over the human and mouse mitochondrial genome for p53 in publicly available ChIP-seq datasets. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) p53 in mouse embryionic fibroblasts (MEFs), data from (Kenzelmann Broz et al. 2013), GSE46240. (B) p53 in mouse embryonic stem cells (mESC), data from (Li et al. 2012), GSE26361; (C) p53 in human IMR90 cells, data from (Aksoy et al. 2012), GSE42728. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

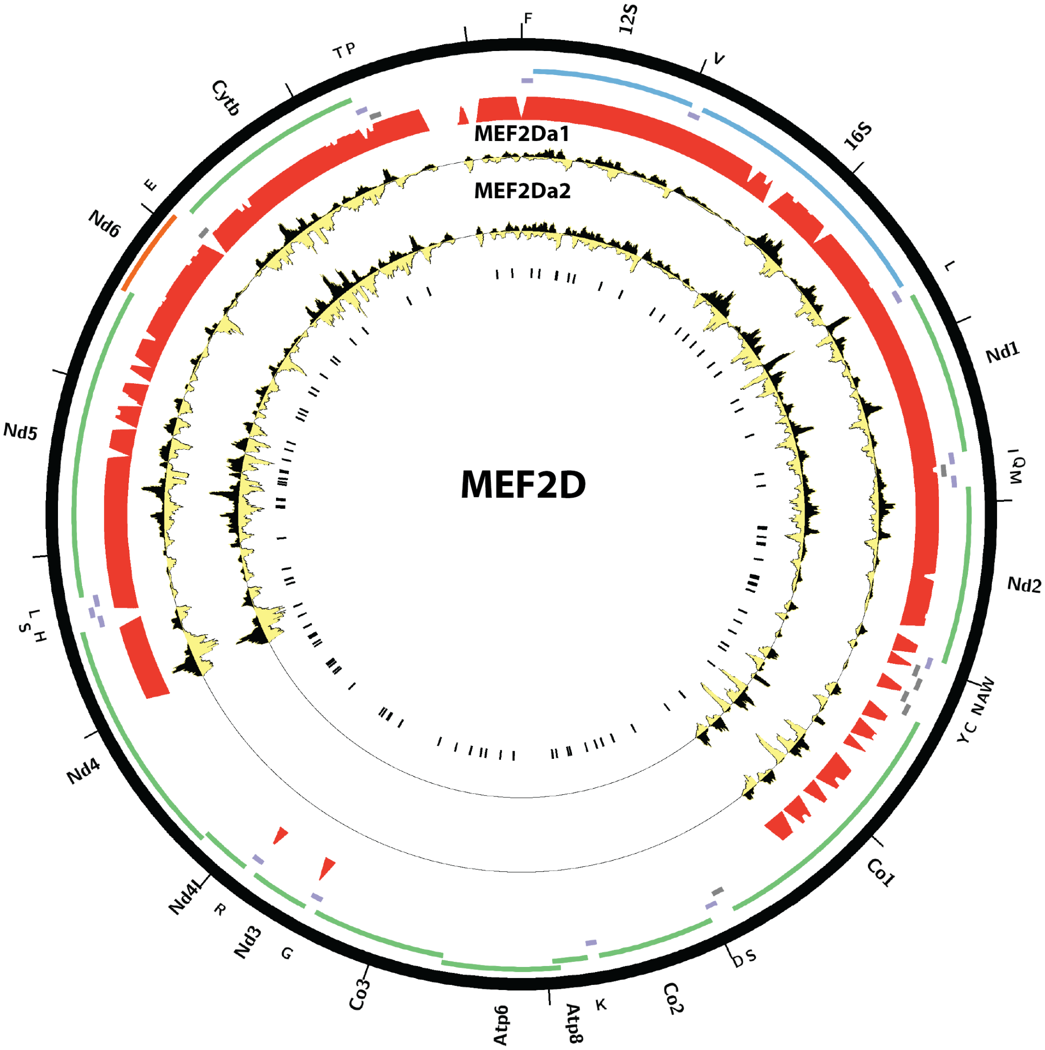


Figure 7.16: Distribution of reads over the mouse mitochondrial genome for MEF2D isoforms MEF2Da1 and MEF2Da2 in C2C12 myoblasts. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the MEF2D motif occurrences in the mitochondrial genome as black vertical bars. Data was obtained from (Sebastian et al. 2013), GSE43223. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

fit a Gamma distribution over the RPM coverage scores for the bottom F_b percentile of fully mappable position on the mitochondrial chromosome. The estimated parameters were then used to rescale the raw signal over all position to a z-score. This results in datasets with strong peaks receiving low z-scores over most of the mappable mitochondrial genome, and very high z-scores over the regions with highly localized enrichment. We used $F_b = 0.8$ for our analysis. As this procedure is sensitive to datasets with very low total read coverage over the mitochondrial genome, we restricted our analysis to datasets with at least 5000 uniquely mappable reads (and with no mismatches to the reference), i.e. $\geq 10x$ coverage. We used a z-score cutoff of 20 to select datasets with high enrichment over the mitochondrial genome, as it was the highest z-score observed in sonicated input samples

7.4.4 Motif analysis

The peak calls for human and mouse ENCODE data available from the USCS Genome Browser were used to find de novo motifs for transcription factors from ChIP-seq data. The sequence around the peak summit (using a 50bp radius) was retrieved for the top 500 called peaks for each factor in each cell line and motifs were called using the MEME program in the MEME SUITE, version 4.6.1 (Bailey et al. 2009). The MEME-defined position weight matrix was then used to scan the mitochondrial genome for motif matches following the approach described in (Mortazavi et al. 2006).

7.4.5 Cell growth and immunocytochemistry

HepG2 cells were grown following the standard ENCODE protocol (DMEM media, 4mM L-glutamine, 4.5g/L glucose, without sodium pyruvate, with 10% FBS (Invitrogen 10091-148) and penicillin-streptomycin). Cells were fixed in 10% formalin (Sigma-Aldrich HT501128-4L) for 10 min, permeabilized with 0.1% Triton X-100, and blocked in 5% FBS. Primary antibodies used were MafK (1:100, Abcam, ab50322) and Hsp60 (1:125, Santa Cruz, sc-1052). Secondary antibodies used were donkey anti-goat AF488 (Invitrogen A11055) and donkey anti-rabbit AF546 (Invitrogen A10040). Imaging on a Zeiss LSM 710 confocal microscope with PlanApochromat 63X/1.4 oil objective, and 0.7μm optical sections were acquired.

8

Physical association of nuclear trancription factors with organellar DNA in plants

This chapter generalizes the observation that nuclear transcription factors associate with mitochondrial DNA to plants, and also suggest this might also be happening in chloroplasts. It is based on a still very limited set of ChIP-seq datasets, thus the results are still preliminary; it will be extended in the future when more data become available.

Abstract

Plants contain two organelles of endosymbiotic origin, mitochondria and plastids, each of them containing their own genome of bacterial origin. These genomes are greatly reduced in terms of their gene content due to the transfer of genes to the nucleus. However, the organellar proteomes are not straightforward derivatives of the ancestral bacterial genome but are in fact a complex mixture of the products of genes that also originate from the nuclear genome, from the other organelle and from additional sources. Nuclear transcription factors have been detected in mammalian mitochondria for many years. Recently, the compendium of ChIP-seq (Chromatin Immunoprecipitation coupled with sequencing) data for a wide diversity of metazoan transcription factors generated by the ENCODE and modENCODE consortia was used to test their association with mitochondrial DNA (mtDNA), which was observed for between 5 and 10% of them. Here, publicly available ChIP-seq datasets for nuclear transcription factors in *Arabidopsis thaliana* and *Zea mays* were examined to determine whether the same phenomenon is also observed in plant genomes. Evidence for physical association with the mitochondrial genome was found for 2 of 21 transcription factors in *Arabidopsis*, and putative such association with the plastid genome was detected for 1 of the 3 maize factors for which ChIP-seq data was available. While the sampling of plant transcription factors assayed by ChIP-seq is still very limited, these results suggest that the phenomenon of nuclear transcription factors localizing to organelles and physically interacting with their genomes may be widespread across eukaryotes.

8.1 Introduction

The evolution of eukaryotes is marked by two profoundly significant primary endosymbiotic events. All known extant eukaryotes share ancestrally a mitochondrion, an organelle vitally important for oxidative phosphorylation (as well as numerous other functions it has acquired during its evolution), which arose as a result of the endosymbiosis of the common ancestor of

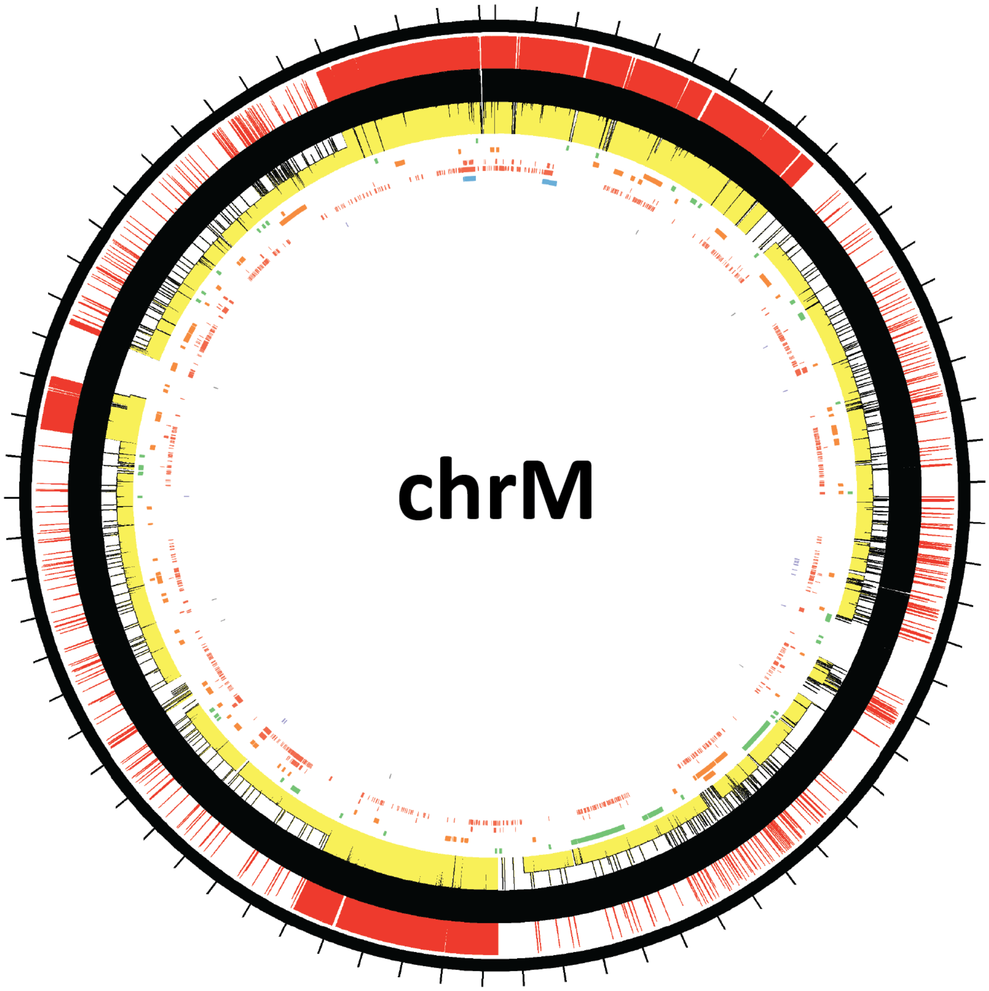


Figure 8.1: Mappability of the *A. thaliana* mitochondrial (chrM) genome (36bp reads). The outer track (red) shows mappability evaluated against the whole genome allowing only for unique reads (colored regions are uniquely mappable). The middle track (black) shows mappability evaluated against the two organellar genomes (chrM and chrP) allowing for up to 2 locations to which a read can map to. The inner track (yellow) shows mappability evaluated against the whole genome allowing for up to 2 locations to which a read can map to (note that the regions where the track is at half of its full height denote regions for which there is a single integration copy in the nuclear genome). The innermost tracks show the mitochondrial genome annotation as follows: forward-strand protein coding genes (green), reverse-strand protein coding genes (orange), repeats (red), rRNAs (blue), forward-strand tRNAs (purple), and forward-strand tRNAs (grey). Genome annotation was obtained from ENSEMBL plants (version 19). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

modern eukaryotes and a member of the α-proteobacteria clade (Yang et al. 1985). A hallmark of mitochondria is the presence of their own genome (Nass et al. 1965), derived from

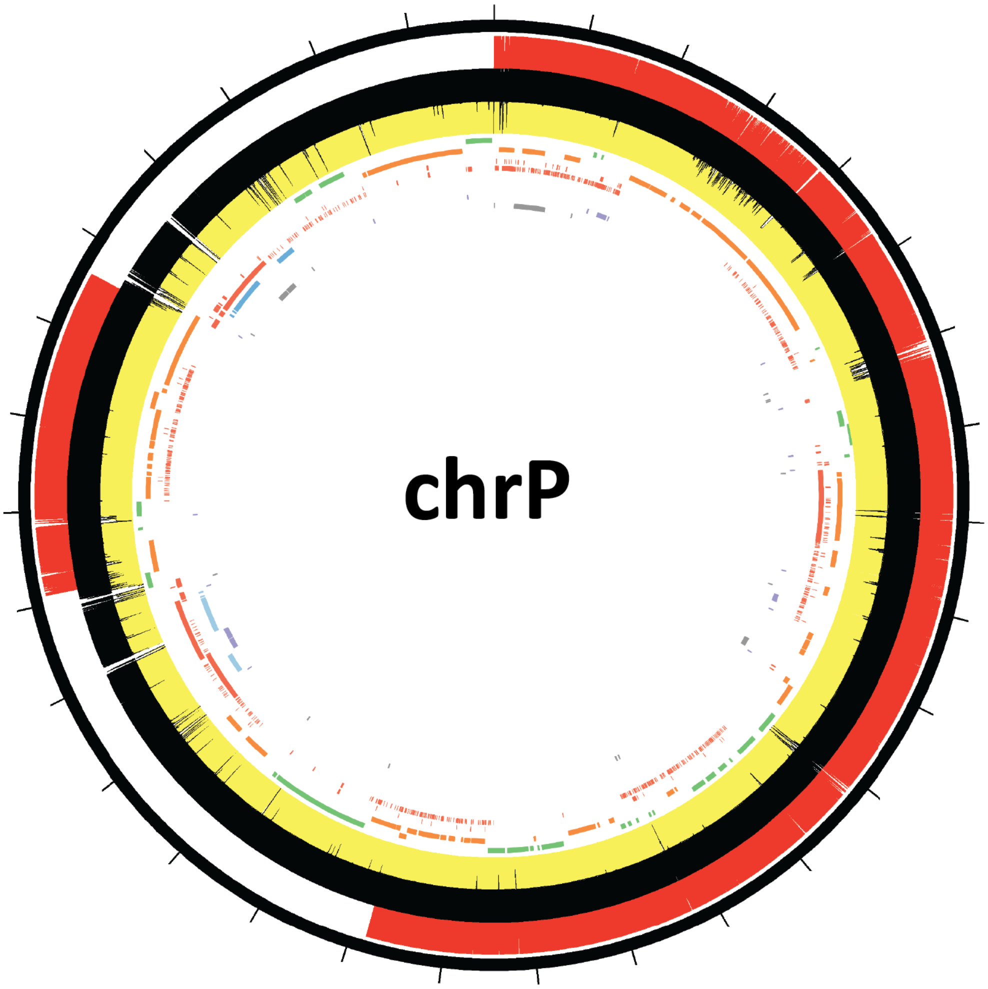


Figure 8.2: Mappability of the *A. thaliana* plastid (chrP) genome (36bp reads). The outer track (red) shows mappability evaluated against the whole genome allowing only for unique reads (colored regions are uniquely mappable). The inner track (black) shows mappability evaluated against the two organellar genomes (chrM and chrP) allowing for up to 2 locations to which a read can map to. The inner track (yellow) shows mappability evaluated against the whole genome allowing for up to 2 locations to which a read can map to. The innermost tracks show the plastid genome annotation as follows: forward-strand protein coding genes (green), reverse-strand protein coding genes (orange), repeats (red), reverse-strand rRNAs (blue), forward-strand rRNAs (light blue), forward-strand tRNAs (purple), and forward-strand tRNAs (grey). Genome annotation was obtained from ENSEMBL plants (version 19). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

their bacterial ancestor, although in some lineages mitochondria have been subsequently reduced to hydrogenosomes (Lindmark & Müller 1973) and mitosomes (Tovar et al. 1999; Tovar et al. 2003; Williams et al. 2002), in which the

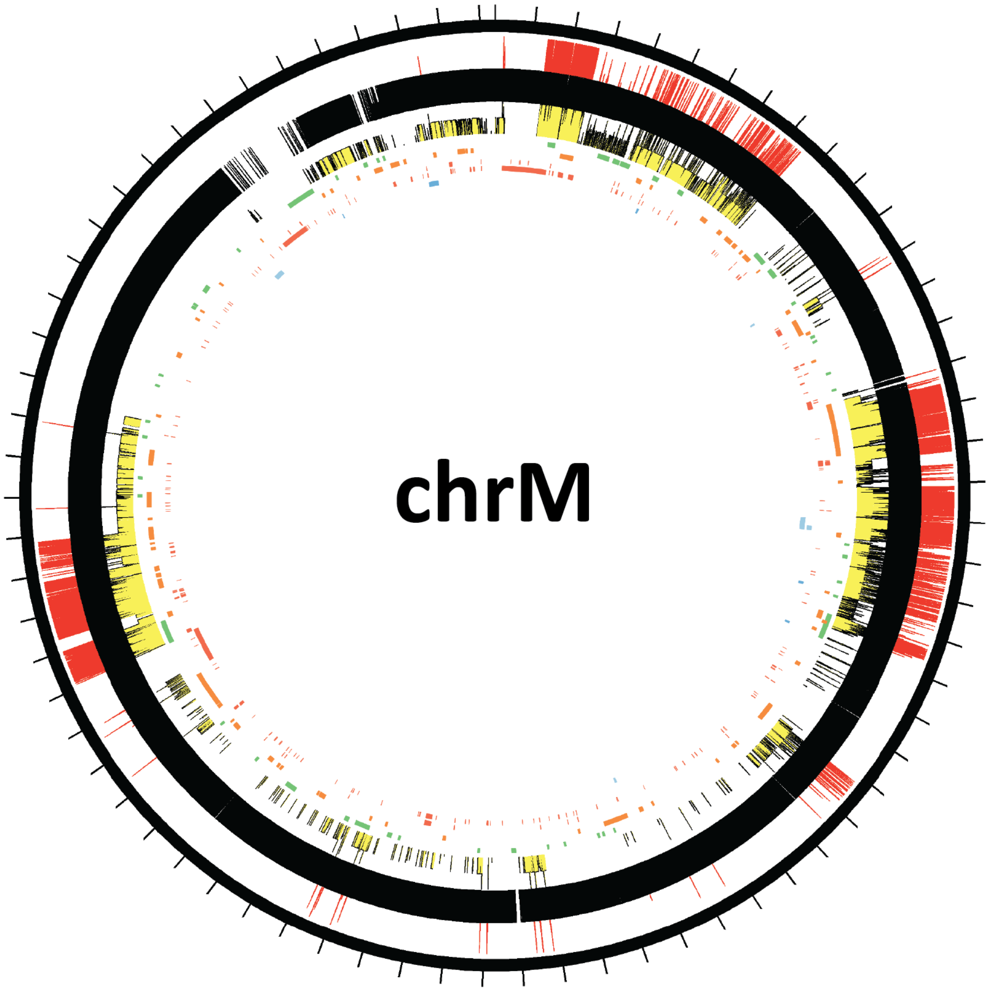


Figure 8.3: Mappability of the *Zea mays* mitochondrial (chrM) genome (36bp reads). The outer track (red) shows mappability evaluated against the whole genome allowing only for unique reads (colored regions are uniquely mappable). The middle track (black) shows mappability evaluated against the two organellar genomes (chrM and chrP) allowing for up to 2 locations to which a read can map to. The inner track (yellow) shows mappability evaluated against the whole genome allowing for up to 2 locations to which a read can map to (note that the regions where the track is at half of its full height denote regions for which there is a single integration copy in the nuclear genome). The innermost tracks show the mitochondrial genome annotation as follows: forward-strand protein coding genes (green), reverse-strand protein coding genes (orange), repeats (red), forward-strand pseudogenes (blue), and reverse-strand pseudogenes (purple). Genome annotation was obtained from ENSEMBL plants (version 19). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

genome has been lost.

A second endosymbiotic event occurred in the lineage to which modern green plants and red algae belong, and involved the acquisi-

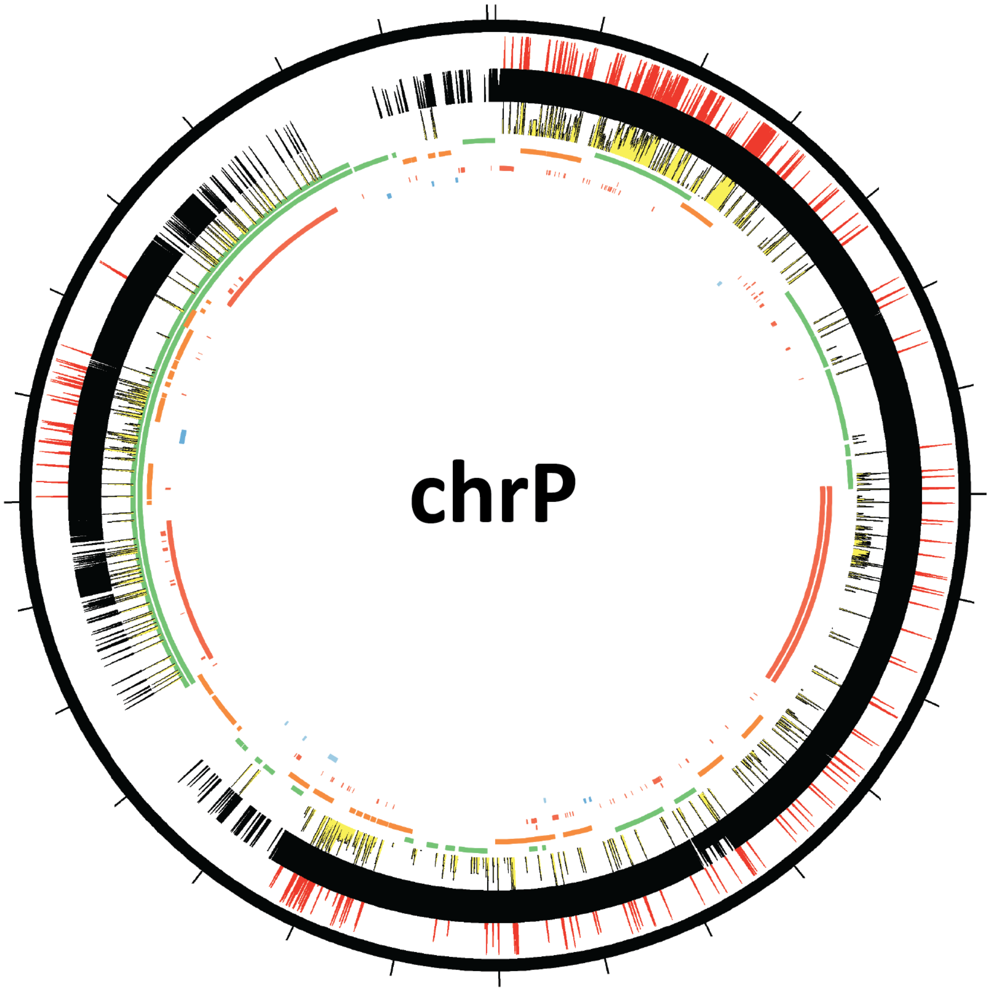


Figure 8.4: Mappability of the *Zea mays* plastid (chrP) genome (36bp reads). The outer track (red) shows mappability evaluated against the whole genome allowing only for unique reads (colored regions are uniquely mappable). The inner track (black) shows mappability evaluated against the two organellar genomes (chrM and chrP) allowing for up to 2 locations to which a read can map to. The inner track (yellow) shows mappability evaluated against the whole genome allowing for up to 2 locations to which a read can map to. The innermost tracks show the plastid genome annotation as follows: forward-strand protein coding genes (green), reverse-strand protein coding genes (orange), repeats (red), forward-strand pseudogenes (blue), and reverse-strand pseudogenes (purple). Genome annotation was obtained from ENSEMBL plants (version 19). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

tion of a photosynthetic cyanobacterial prokaryote, which eventually became the chloroplast. Subsequently, on multiple occasions, nonphotosynthetic eukaryotes established secondary endosymbiosis with photosynthetic eukaryotes (Archibald & Keeling 2002; Keeling 2004; Keeling 2010; Keeling 2013). Plastids also contain their own genome derived from their prokaryotic

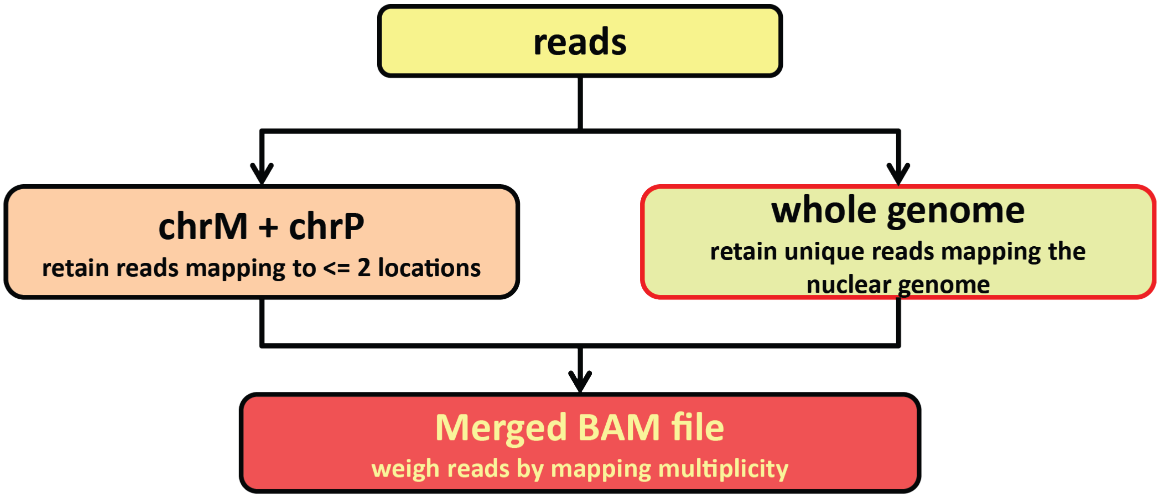

reads
chrM + chrP
retain reads mapping to <= 2 locations
whole genome
retain unique reads mapping the nuclear genome
Merged BAM file
weigh reads by mapping multiplicity

Figure 8.5: Read mapping strategy. The IR regions of the plastid genome are not uniquely mappable within it, but are generally uniquely mappable compared to the nuclear genome (see Figure 8.2). The question of whether detected read enrichment within them is specific to the plastid can therefore be answered conclusively even if it is not possible to distinguish between the two IR copies. For these reasons, a two-step alignment procedure was implemented. First, reads were mapped to the union of the two organellar genomes allowing for up to 2 locations a read can map to. Second, reads were mapped against the union of all three genomes retaining uniquely mappable reads only. Read aligning to the nuclear genomes from the second step were combined with the reads from the first step and subsequent analysis was carried out on the resulting set of alignments while weighing all multireads by the number of locations they map to (i.e. a read that maps to two locations is counted as half a read at each).

ancestor.

Both mitochondrial and plastid genomes are greatly reduced in terms of their gene content, as a result of the transfer of genes from the organellar genome to the nucleus. DNA fragments from degraded organelles can enter the nucleus and integrate into the nuclear genome (a constantly ongoing process, which can be observed even today; Ayliffe et al. 1998; Huang et al. 2003; Hazkani-Covo et al. 2010), the genes they contain can then evolve the ability to be targeted back to the organelle, at which point the organellar copy is not under selective pressure anymore and can be lost. However, the present-day organellar proteomes are not simply a subset of the ancestral prokaryotic proteomes, but instead contain the products of numerous genes originally from the nucleus, the other organelle (in the case of plants), or even external sources (Suzuki & Miyagishima 2010).

A representative example of the latter are the polymerases that transcribe organellar DNA. Organellar genomes possess dedicated machineries that regulate and carry out the process of transcription of their genomes, even though, with one notable exception, their components are encoded in the nucleus. The mitochondrial RNA polymerase of most eukaryotes is of bacteriophage origin (Shutt & Gray 2006; Barbrook et al. 2010), while two separate polymerases operate in plastids, one of them also of phage origin and encoded in then nucleus (NEP), and another one of cyanobacterial origin, encoded in the plastid genome (PEP) (Hess & Börner 1999).

The organization of plastid genomes is relatively consistent between different lineages, with some notable exceptions (Zhang et al. 1999). They are typically circular mapping, between 100 and 200kb long, contain between ∼100 and ∼250 genes, and usually feature two large inverted repeats (Barbrook et al. 2010). The *Arabidopsis thaliana* plastid genome is 154,478bp long and contains 88 protein coding genes (Sato et al. 1999); the *Zea mays* plastid genome is 140,387 bp in size (Maier et al. 1995). The genes are transcribed as polycistronic units from multiple promoters by the NEP or the PEP.

In contrast a wide diversity of topology, or-

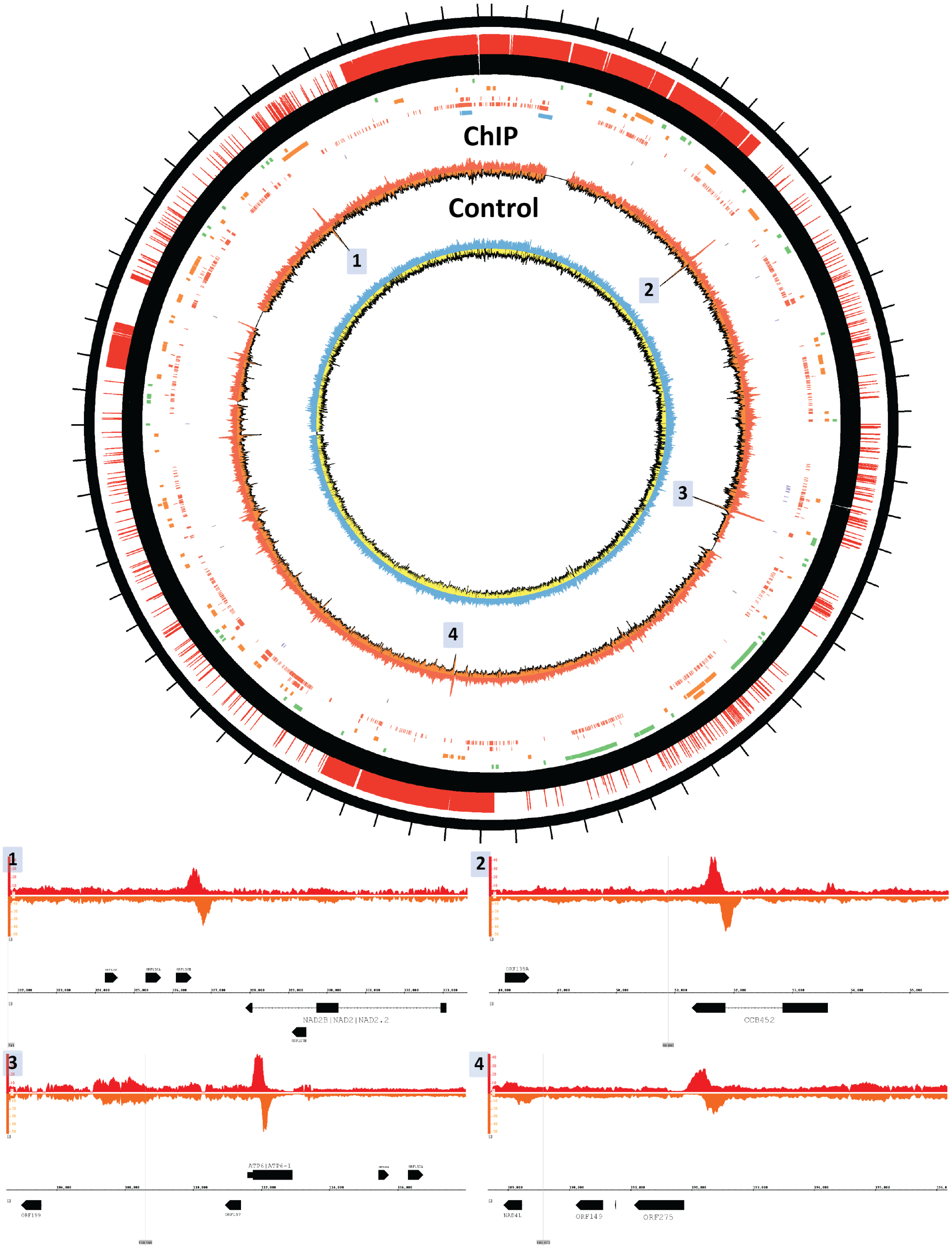


Figure 8.6: Association of APETALA3 with the *Arabidopsis thaliana* mitochondrial genome. ChIP and control datasets are drawn to the same scale, set to be the maximum signal level within all four tracks (forward and reverse strand, ChIP and control). IGB browser plots of the putative occupancy sites are shown zoomed in below. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009) and the Integrated Genome Browser (Nicol et al. 2009).

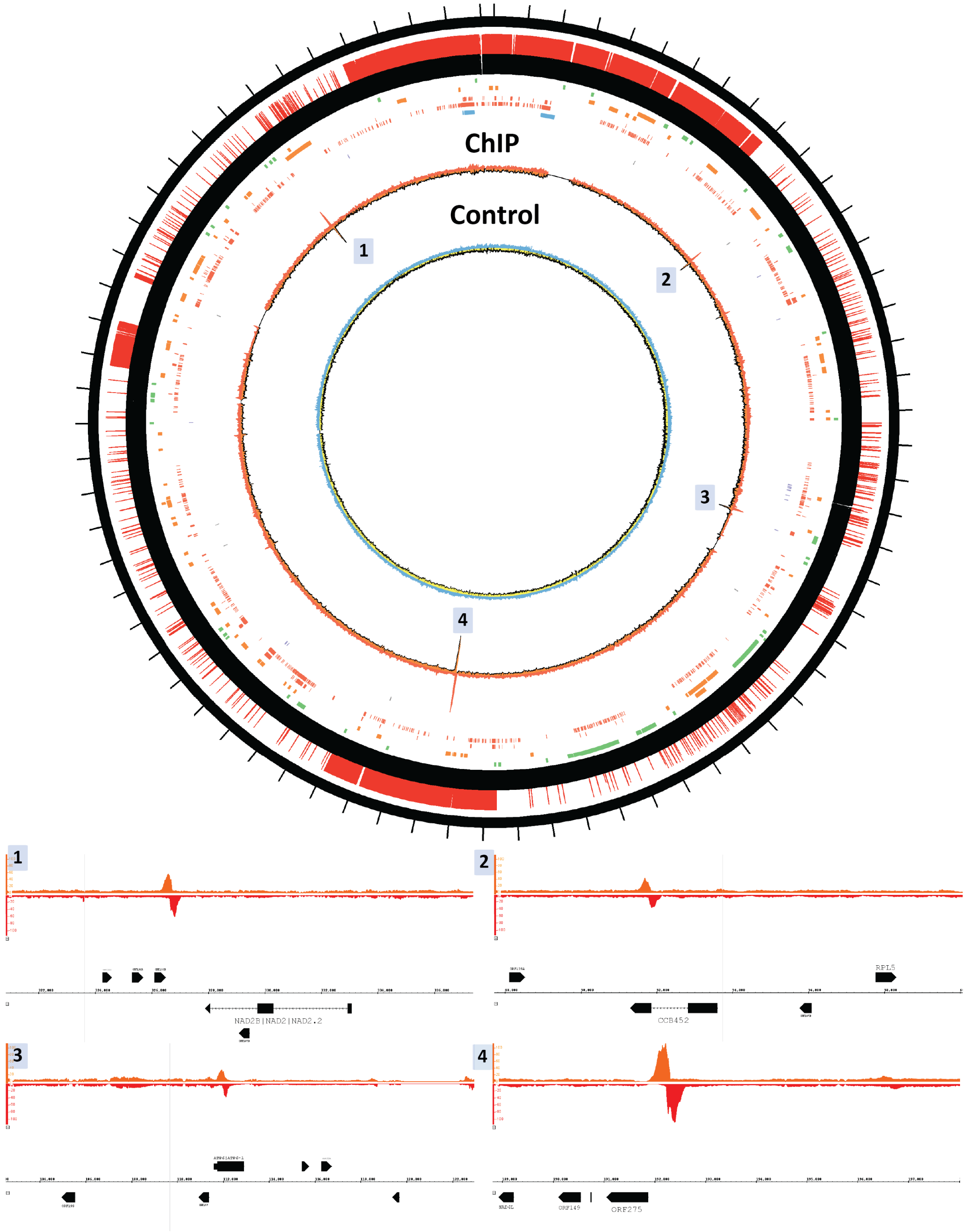


Figure 8.7: Association of PISTILLATA with the *Arabidopsis thaliana* mitochondrial genome. ChIP and control datasets are drawn to the same scale, set to be the maximum signal level within all four tracks (forward and reverse strand, ChIP and control). IGB browser plots of the putative occupancy sites are shown zoomed in below. Plots were generated using Circos version 0.60 (Krzywinski et al. 2009) and the Integrated Genome Browser (Nicol et al. 2009).

ganization, and sizes is observed in mitochondrial genomes. The best known mitochondrial genomes are those of mammals. The human mitochondrial genome is highly reduced, only 16,571bp long, and contains 13 protein coding genes, 22 tRNAs and 2 rRNAs (Anderson et al. 1981; Bibb et al. 1981). It features only one significant noncoding region, the so-called D-loop, from which transcription originates bidirectionally generating long polycistronic messages, which are then processed to generate the mature mRNAs, tRNAs and rRNAs (Montoya et al. 1982; Shutt et al. 2011). Plant mitochondrial genomes are strikingly different (Lynch et al. 2006), being hundreds of kilobases (and sometimes megabases; Alverson et al. 2011; Ward et al. 1981; Sloan et al. 2012) long, which is primarily due to the presence of introns and of very large amounts of repetitive DNA. The *Arabidopsis thaliana* mitochondrial genome is 366,924bp long (Unseld et al. 1997); the *Zea mays* plastid genome is 569,630bp long (Clifton et al. 1995). Not much is known about the details of transcription and its regulation in these genomes, but given how widely dispersed genes are within them they are most likely transcribed into multiple independent units.

These processes are best understood in mammalian systems, where a curious phenomenon has also been observed: the presence of nuclear transcription factors in mitochondria (Leigh-Brown et al. 2010). The functional significance of this localization has been conclusively demonstrated only in a few cases (Casas et al. 1999; Enríquez et al. 1999a; Enríquez et al. 1999b; Wrutniak et al. 1995); the direct physical association of these factors with mtDNA had similarly not been directly shown. Recently, these issues were addressed by utilizing the vast ChIP-seq (Johnson et al. 2007) resource generated by the ENCODE and modENCODE consortia (Celniker et al. 2009; ENCODE Project Consortium 2012); occupancy of mtDNA by nuclear transcription factors was conclusively demonstrated by the direct biochemical evidence for it provided by ChIP-seq (Marinov et al. 2014). However, all reliably observed occupancy events were located in regions of mammalian mitochondrial genomes distant of the regulatory D-loop, making their functional significance difficult to interpret.

To gain further insight into the phenomenon, publicly available ChIP-seq datasets for *Arabidopsis thaliana* and *Zea mays* transcription factors were examined. As plants possess two organelles with proteomes of complex history, the genomes of which are not compact but instead contain dispersed genes organized into multiple transcriptional units, it is of great interest whether nuclear transcription factors localize to these organelles, and where their occupancy sites are in their genomes. Two such factors (out of 21 tested) were found to associate with mtDNA *Arabidopsis thaliana* and one (out of 3) factor might be associating with plastid DNA (ptDNA) in *Zea mays*, though the evidence is not entirely conclusive in the latter case. Similarly to nuclear transcription factors in human and mouse mitochondria, these occupancy sites were mostly not located in immediately obvious regulatory regions (with the caveat that transcriptional units and regulatory elements are still to be precisely defined in plant organelles).

8.2 Results

8.2.1 ChIP-seq datasets and data processing

Publicly available in the Gene Expression Omnibus (GEO) as of April 1st 2014 ChIP-seq data for *Arabidopsis thaliana* and *Zea mays* transcription factors was downloaded and processed as described below and in the Methods section. Data for the following transcription factors was included in the final collection: FLM, AGAMOUS, SHORT VEGETATIVE PHASE, SOC1, PIF, APETALA1 and SEPALLATA3, JAGGED, FLOWERING LOCUS C (FLC), PIF4, APETALA3 and PISTILLATA, REVOLUTA, PIF5, TOC1, FAR-RED ELONGATED HYPOCOTYL3 (FHY3), LEAFY, ABORTED MICROSPORES (AMS), APETALA2, APETALA1, SEPALLATA3, and KANADI1 (in *Arabidopsis thaliana*) and RAMOSA1, Pericarp Color 1 (P1), and KNOTTED1 in *Zea mays*. In addition, *Arabidopsis* ChIP-seq datasets against histone marks, the H3 and H3.3 histones, AGO4, RNA polymerase IV and RNA polymerase V (NRPE1), the DNA methyltransferase CMT3, the polyadenylation factor PCFS4, as well as MNAse-seq and DNAse-seq datasets, and *Zea mays* ChIP-seq data for centromere histone variants were also examined, as potential negative controls.

As already mentioned, DNA from mitochondrial and plastid genomes is continuously transferred to the nucleus, which means that fragments of organellar DNA can be present in the nuclear genome. This poses a challenge when distinguishing true physical occupancy of organellar DNA from occupancy of organellar-derived DNA in the nucleus (Marinov et al. 2014). Therefore, the mappability of the organellar genomes in *Arabidopsis thaliana* and *Zea mays* was first examined before a data processing strategy was designed accordingly.

Unlike the mitochondrial genomes of mammals, plant organellar genomes are large and contain repetitive elements. Thus two different mappings are relevant to the question of how uniquely mappable they are and how that affects data analysis and interpretation: mapping reads against the union of the two organellar genomes, and mapping reads to all three genomes (including the nuclear one). In addition, the plastid genome contains two large inverted repeats, which may or may not be unique to it but are highly similar to each other, and would not be "visible" to downstream analysis if only uniquely mappable reads are considered, even if only the plastid genome is used as a reference during the mapping step. For these reasons, mappability was evaluated as follows:

1. **Full genome unique mappability**, using all three genomes as a reference and considering only unique alignments (Bowtie 0.12.7 was used; Langmead et al. 2009)
2. **Combined organellar mappability with maximum read multiplicity of 2**, using both chrP and chrM as referenc and considering reads mapping to up to 2 locations.
3. **Full genome mappability with maximum read multiplicity of 2**, using all three genomes as a reference and considering reads mapping to up to 2 locations.

Figures 8.1 and 8.2 show the mappability of the *Arabidopsis thaliana* mitochondrial (chrM) and plastid (chrP) genomes, respectively. There are large portions of the mitochondrial genome that are not uniquely mappable in the full-genome unique mappability track, while the plastid genome is largely uniquely mappable with the exception of the two inverted repeat regions. In contrast, both organellar genomes are almost completely fully mappable in the track representing the combined organellar mappability with maximum read multiplicity of 2. When reads are mapped with the same settings (maximum read multiplicity of 2) but including the nuclear genome, the plastid genome is still mostly fully mappable. Interestingly, the mitochondrial genome is not fully mappable but the shape of the track indicates that only a single copy of it is present in the nuclear genome (Figure 8.1). This means that it is in principle possible to distinguish organellar from nuclear occupancy of organellar-derived DNA as the ChIP-seq signal strength in the organellar genomes should be significantly higher than what is observed in nuclear genomes due to the larger number of copies of these organelles relative to the two copies of the nuclear genome that exist in each cell. This criterion was successfully used to confirm the reality of mitochondrial occupancy by nuclear transcription in mammalian cells (Marinov et al. 2014).

Taking these considerations into account, the data processing strategy outlined in Figure 8.5 was adopted. Reads were mapped independently against the nuclear genome, retaining only unique alignments, and against the combined organellar genomes, allowing for reads to map to up to two locations (in order to make inverted repeats "visible" to subsequent analysis). The two sets of alignments were then combined for each sample, and read coverage was calculated as described in the Methods section, normalizing for both total sequencing depth across all three genomes and for read multiplicity.

This strategy works well for *Arabidopsis thaliana* and its compact genome with relatively low repetitive element content. However, the maize genome is much larger and composed largely of repeats (∼85%; Schnable et al. 2009). The same analysis of mappability was carried out for the *Zea mays* nuclear, mitochondrial and plastid genome assemblies, and it revealed that the large repetitive portions of the maize genome apparently also contain multiple copies of both organellar genomes. The majority of both chrM and chrP is not mappable in mappings including the nuclear genome, even when the maximum read multiplicity is relaxed to 2, and significant portions of the plastid genome are not fully mappable even in organelle-only mappings (Figures 8.3 and 8.4). For consistency, the same analysis pipeline was adopted for maize as for *Arabidopsis*; results were subsequently interpreted with caution.

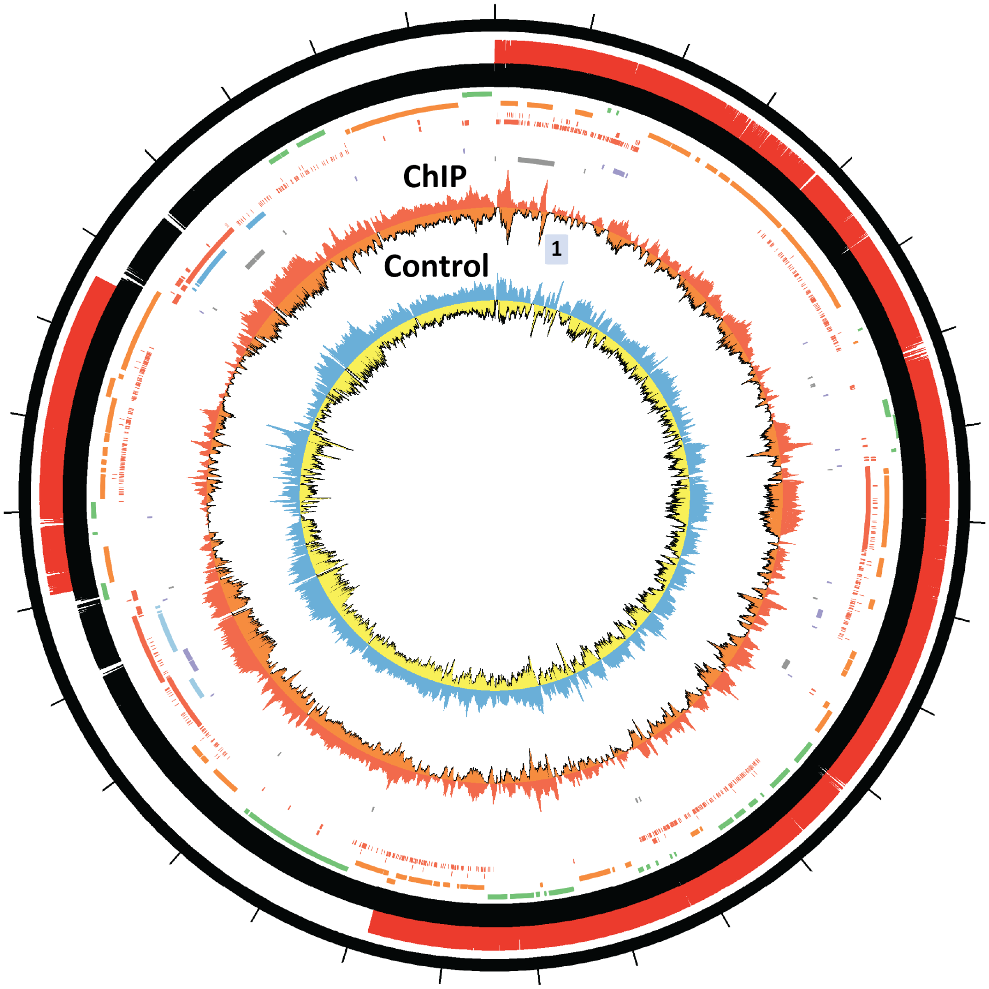


Figure 8.8: APETALA3 ChIP-seq profile over the *Arabidopsis thaliana* plastid genome. ChIP and control datasets are drawn to the same scale, set to be the maximum signal level within all four tracks (forward and reverse strand, ChIP and control). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

8.2.2 Physical association of nuclear transcription factors with organellar genomes in *Arabidopsis thaliana*

After examining the organellar genomes signal profiles of the 21 transcription factors included in this survey (see the Methods section for details), two of them were found to display evidence of physical association with organellar DNA. Figures 8.6 and 8.7 show the forward and reverse strand ChIP-seq read distribution over chrM for APETALA3 and PISTILLATA, respectively. The two datasets were generated as part of the same study and from the same source material, and the two profiles are very similar to each other. Four major putative occupancy sites were observed. They were characterized by the typical asymmetry of read distribution on the two strands around the bind-

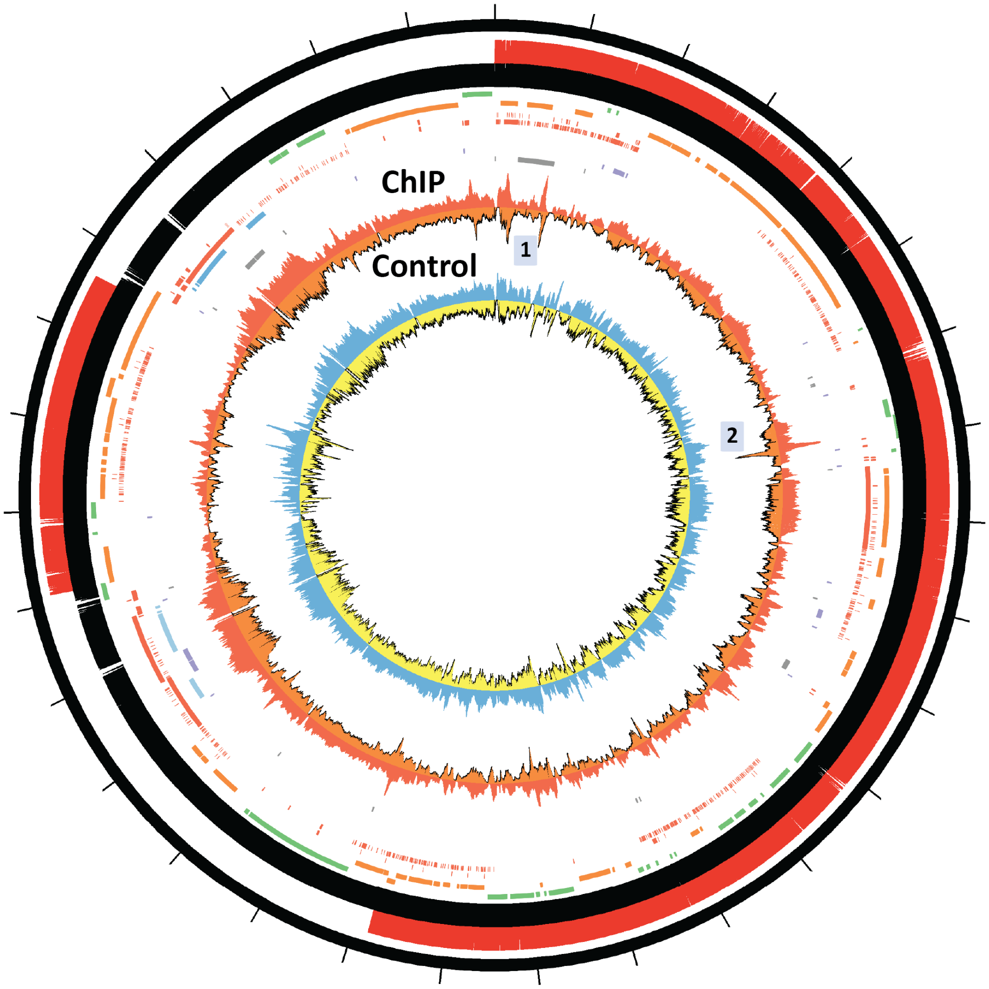


Figure 8.9: PISTILLATA ChIP-seq profile over the *Arabidopsis thaliana* plastid genome. ChIP and control datasets are drawn to the same scale, set to be the maximum signal level within all four tracks (forward and reverse strand, ChIP and control). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009).

ing site (Kharchenko et al. 2008; Landt et al. 2012; Marinov et al. 2014), suggesting that they are not sequencing artifacts. They are located outside of the uniquely mappable portions of the *Arabidopsis thaliana* mitochondrial genome, however, their signal strength suggests they are unlikely to be instances of occupancy over the mitochondrial-derived sequence in the nuclear genome. The three strongest nuclear peaks (out of 12,440 identified using MACS2; Feng et al. 2012) for APETALA3 have maximum peak heights of 123, 109 and 54 RPM (Reads Per Million), while the strongest mitochondrial peak has a height of ∼100 RPM. The three strongest PISTILLATA nuclear peaks (out of 8414) have maximum peak heights of 56, 25 and 24 RPM, while the mitochondrial peaks have a peak height of ∼230 RPM. It is not impossible that the nuclear copies of the mitochondrial genome contain the strongest binding sites for

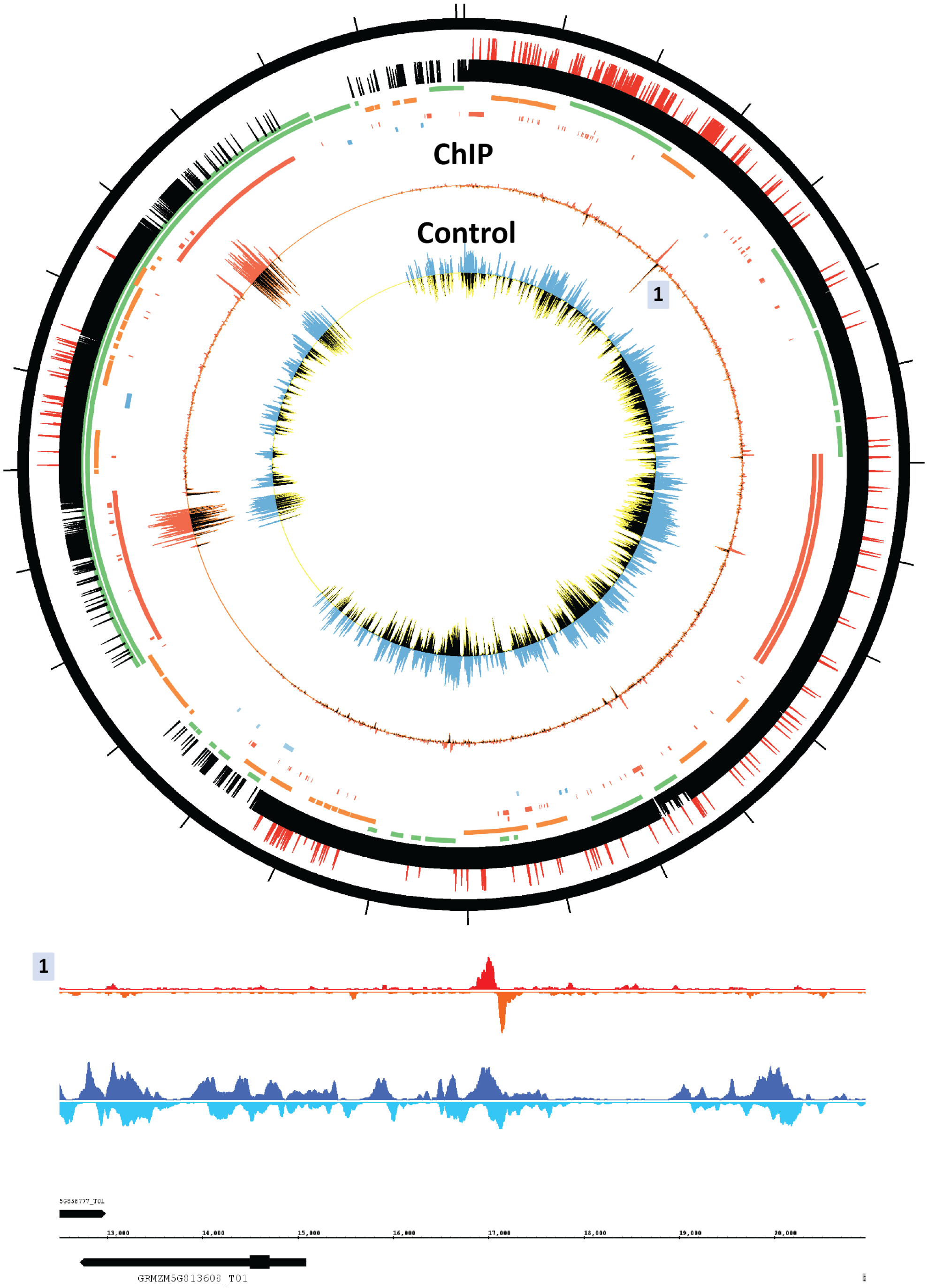


Figure 8.10: ChIP-seq profile over the *Zea mays* plastid genome for RAMOSA1 in ear primordia. ChIP and control datasets are drawn to the same scale, set to be the maximum signal level within all four tracks (forward and reverse strand, ChIP and control). Plots were generated using Circos version 0.60 (Krzywinski et al. 2009). Note that the putative occupancy sites were only observed in RAMOSA1 ChIP-seq data from ear primordia, but not in tassel primordia (data not shown).

these factors in the whole genome, but the more parsimonious explanation is that these factors indeed bind to the mitochondrial genome.

The location of the four peaks identified did not suggest immediately obvious regulatory roles they might be playing. The first of them is located just downstream of and partially overlaps the short uncharacterized ORF196B putative protein coding gene. The second one is located in the 3' portion of CCB452. The third one is within the 5' end of the ATP6 gene, and the fourth one (which is also the strongest) is located a few hundred base pairs upstream of ORF275. The last two might be playing a role in regulating the transcription of their proximal genes as they are in the vicinity (but right on top) of their 5' ends, but it is less clear how the first two might have a regulatory influence on gene expression. However, as precise delineation of the transcriptional units, the sites of transcriptional initiation and its regulation in the organellar genomes of plants is at present lacking, it is not yet possible to draw conclusions regarding these questions.

The APETALA3 and PISTILLATA transcription factors do not display strong evidence for binding to the plastid genome. Peaks exhibiting a asymmetric read distribution profile were observed (Figures 8.8 and 8.9) but these were of relatively small absolute magnitude, both within the plastid genome and compared to the input, thus they cannot be confidently concluded to be true instances of physical association.

No peaks displaying the characteristics of true ChIP-seq occupancy peaks were observed in DNAse-seq, MNAse-seq, Polymerase IV and V, histone and histone mark datasets in *Arabidopsis*.

8.2.3 Putative physical association of nuclear transcription factors with organellar genomes in *Zea mays*

The same analysis was also carried out on the available maize ChIP-seq datasets. Their number is at present very limited (only 3) but one of them, RAMOSA1, did exhibit what might be physical association with the plastid genome. Figure 8.10) shows the RAMOSA1 signal profile over chrP in ear primordia, displaying one relatively strong putative occupancy site with very clear read distribution asymmetry, and multiple other smaller peaks. Interestingly, the same profile was observed in both RAMOSA1 ChIP-seq replicates available from ear primordia but the putative occupancy sites were completely absent in the two replicates from tassel primordia (data not shown). However, some read asymmetry around the same site was also observed in the read profiles of the input control datasets (Figure 8.10)), thus it cannot be ruled out that these observations are due to an experimental artifact.

8.3 Discussion

The results presented here extend the observation that nuclear transcription factors localize to mitochondria and associate with mtDNA to plants, and also suggest that the same phenomenon might also occur in plastids. As is the case with mammalian mitochondria, the functional significance of the physical association of these transcription factors with mtDNA is at present unknown; as previously discussed (Marinov et al. 2014), it is entirely possible that it represents biochemical noise, with transcription factors being transported to mitochondria without their presence there having regulatory influence on mitochondrial gene expression. Such understanding is not inconsistent with what we know about the proteome content of organelles in plants. The main theme in the evolution of organellar genomes has been the transfer of genes to the nucleus, with the products of those essential to the organelle's function acquiring the capacity to be targeted to it. However, their modern proteomes are not exclusively derived from the ancestral prokaryotes genome – for example, less than half of the plastid proteins in Arabidopsis are of direct cyanobacterial ancestry (Bogorad 2008; Abdallah et al. 2000; Martin et al. 2002), with the rest originating from the host genome or from other external sources. It is possible that the translocation of nuclear transcription factors to organelles represents intermediate steps in such transitions, with at present neutral adaptive and functional significance – these proteins have biochemical properties that make it possible for them to be imported into organelles but they have not yet acquired specific regulatory roles there.

The more exciting possibility is that they do in fact regulate gene expression there, but it is currently difficult to say how and in what capacity they might be doing that. This is at least in part because the organellar genomes of plants are very poorly functionally annotated complicating the interpretation of protein occupancy measured by ChIP-seq. First steps towards filling these gaps in our knowledge have been made in the form of mapping the transcriptome and the genome-wide localization of PEP in plastids (Fujii et al. 2011; Finster et al. 2013), however, these efforts have been mostly array-based and not coordinated with each other to derive a unified picture clearly delineating transcriptional units. Much additional work remains to be done in this area. In addition, whether these occupancy events are functionally important or not, the question why these factors bind to only a limited set of sites within genomes that are hundreds of bases long and contain numerous instances of their recognition motifs will remain open.

Another gap in our knowledge that has to be mentioned is the still very small number of existing ChIP-seq datasets in plants. Here, all publicly available transcription factor ChIP-seq datasets were surveyed, yet this only amounted to 21 factors in *Arabidopsis thaliana* and 3 in *Zea mays*, and these datasets were generated in a wide diversity of labs using different protocols making direct comparisons between datasets and the exclusion of experimental artifacts as explanation for certain observations less than straightforward (for example, the APETALA3 and PISTILLATA datasets that do display strong evidence for association with mtDNA were obtained from the same study; the two factors are functionally related so it is not entirely surprising both of them would localize to the same sites in mtDNA, but the observation of the same phenomenon for the same and for other factors in datasets generated from other labs would definitely be encouraging). This situation is in marked contrast with the vast resources that are at this point available in mammalian, fly and worm systems through the efforts of the ENCODE, mouse ENCODE and modENCODE projects, and many other individual labs.

The future should bring a significant expansion in the number of available plant transcription factor ChIP-seq datasets, which should enable the much broader generalization of the findings present here. Even more exciting would be the generation of ChIP-seq datasets in systems, in which mitochondrial and plastid genomes display unusual organizations. There are such numerous such examples, especially in protists (Burger et al. 2003; Gray et al. 2004; Gray MW. 2012), and include mitochondrial genomes organized into multiple small and large circles, mitochondrial genomes existing in the form of multiple linear chromosomes, plastid genomes organized into minicircles, genes existing in split from on multiple separate minichromosomes and many other variations of these themes. Importantly, in system where large numbers of minicircles or small linear chromosomes exist in mitochondria, transcriptional units are by necessity numerous and relatively well defined; it would be therefore illuminative to know whether and where nuclear transcription factors bind to mtDNA in such systems. The generation of large numbers of transcription factor ChIP-seq datasets in a large number of diverse systems in the future should provide answers to many of these questions.

8.4 Methods

8.4.1 Data processing and analysis

Data was downloaded from the following GEO series or SRA accession numbers and their associated publications: GSE48082 (Posé et al. 2013), GSE45939 and GSE45938 (ÓMaoiléidighet al. 2013), GSE45368 (Law et al. 2013), GSE33120 (Gregis et al. 2013), GSE45846 (Immink et al. 2012), GSE39215 (Zhang et al. 2013), GSE39097 (Du et al. 2012), GSE35381 (Zheng et al. 2012), GSE39247 (Zhong et al. 2012), GSE35315, GSE38358 (Wuest et al. 2012), GSE26722 (Brandt et al. 2012), GSE36629 (Wollmann et al. 2012), GSE35059, GSE35952 (Huang et al. 2012), GSE34840 (Stroud et al. 2012), GSE30711 (Ouyang et al. 2011), GSE24568 (Moyroud et al. 2011), GSE22276 (Ha et al. 2011), GSE16940 (Wang et al. 2010), GSE21301 (Yant et al. 2010), GSE20176 (Kaufmann et al. 2010), GSE14600 (Kaufmann et al. 2009), GSE48081 (Merelo et al. 2013), GSE51048 and GSE51050 (Eveland et al. 2014), GSE47342 (Wang et al. 2014), GSE38587 (Morohashi et al. 2012), GSE39161 (Bolduc et al. 2012), GSE48793 (Heyman et al. 2013), GSE46894 and GSE46986 (Pajoro et al. 2014), GSE51537 (Schiessl et al. 2014), SRP005412 (Deng et al. 2011), SRA060798 (Xing et al. 2013).

Reads were trimmed to 36bp and aligned using Bowtie (Langmead et al. 2009), version 0.12.7, in two stages. First, an alignment against the combined mitochondrial and plastid organellar genomes was carried out with the following settings: `‘‘-v 2 -t -k 3 -m 2 --best --strata’’`, i.e. allowing for two mismatches relative to the reference and for up to 2 locations to which a read could map to. This was done in order to retain reads aligning to the inverted repeats in the plastid genomes. Second, reads were aligned against all three genomes with the following settings: `‘‘-v 2 -t -k 2 -m 1 --best --strata’’`, retaining unique alignments only. The TAIR10 version of the *Arabidopsis thaliana* genome and the AGPv3 assembly of the *Zea mays* genome were used, downloaded from ENSEMBL. Reads mapping to the nuclear genome from the second mapping were combined with the reads mapping to the organellar genomes into a single BAM file and subsequent analysis was carried out on this set of alignments. Read coverage was calculated by weighing reads according to the following rule:

$$S_{c,i} = \frac{\displaystyle\sum_{R \in R_{c,i}} \frac{1}{NH_R}}{\dfrac{|R|}{10^6}} \tag{8.1}$$

Where $S_{c,i}$ is the signal score for position i on chromosome c, $|R|$ is the total number of aligned reads, $|R_{c,i}|$ is the number of reads covering position i on chromosome c, and NH_R is the number of locations in the genome a given read maps to. This has the effect of counting multireads that align to each inverted repeat as “half-reads”, thus making read coverage across those regions comparable with that of the rest of the genome. Only reads aligning with zero mismatches were considered for the organellar genomes.

Nuclear peaks were called using MACS, version 2.0.9 (Feng et al. 2012). Regions of enrichment over the organellar genomes were determined by manual curation. This was feasible thanks to the small size of these genomes and necessary as several peak callers were tried – MACS version 2.0.9., GEM (Guo et al. 2012), and SPP (Kharchenko et al. 2008) – but each produced significant numbers of obvious false negatives and/or false positives.

Part III

Quality Assessment and Analysis of Chromatin Immunoprecipitation Data

The three chapters in this part contain the work on developing and applying metrics for assessing the success and quality of ChIP-seq experiments that I have been involved in, with an eye towards automating this critically important step in working with data of this type by reducing it to a simple set of numbers that can be rapidly scanned by humans or machines. This goal has not quite been achieved as reality has turned out to be a little bit too complex for such an approach to be always applicable without any human input, but in the process we have learned a tremendous amount about the ChIP-seq itself. I should perhaps also note that intellectual honesty requires to admit that initially the motivation behind this work was a bit different - a large number of datasets of obviously poor quality were apparent within the ENCODE project and elsewhere and the frustration with that state of affairs is what prompted the development of standardized ways of measuring quality, in which I played some role. This should be particularly noticeable in the second chapter in this part.

I have also included a chapter on the development of a robotic ChIP protocol, in which the quality-control metrics described in prior chapters played a major role, and I carried out the computational analysis. This was not a project in which I had the leading role, but it is important for my vision for the future laid out in the last chapter of the thesis, thus its inclusion was important for the self-consistency of the text as a whole.

The three chapters as originally written as individual papers contain a lot of redundant material as they focus on different aspects of the same issue. I have retained the redundant material for the sake of each chapter being as much a self-contained entry as possible.

9

ChIP-seq Quality Evaluation Metrics of the ENCODE Consortium

The major part of the material contain in this chapter was published as part of the ENCODE Project Consortium paper on ChiP-seq quality evaluation and standardization:

Landt SG*, **Marinov GK***, Kundaje A*, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, Desalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**(9):1813–1831. doi: 10.1101/gr.136184.111.

The paper is reprinted in Appendix C. I have omitted some portions of the part of it that concerns the characterization and validation of antibodies (which was contributed by Steven Landt and the Snyder lab at Stanford). I have also omitted the several sets of specific guidelines for doing certain things that were provided in the paper while including some material that was part of its earlier version but did not make the final cut.

The NRF and FRiP metrics described here were developed based on work from Ali Mortazavi (Johnson & Mortazavi et al. 2008. Cross-correlation metrics were developed by Anshul Kundaje (A. Kundaje et al. 2014, unpublished) based on Kharchenko et al. 2008. IDR was developed and described by Li et al. 2011. The IDR pipeline was developed by Anshul Kundaje and others.)

Abstract

Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of transcription-factor binding and histone modifications in living cells. Despite its widespread use, there are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use. These practices affect the quality and utility of any global ChIP experiment. Based on the extensive experience the ENCODE and modENCODE consortia have accumulated working with ChIP-seq, a set of working standards and metrics for the quality evaluation of ChIP experiments were developed. The standards and metrics, as well as how ChIP quality, assessed in these ways, affects different uses of ChIP-seq data, are discussed here.

9.1 Introduction

Methods for mapping transcription factor occupancy across the genome by chromatin immunoprecipitation (ChIP) were developed more than

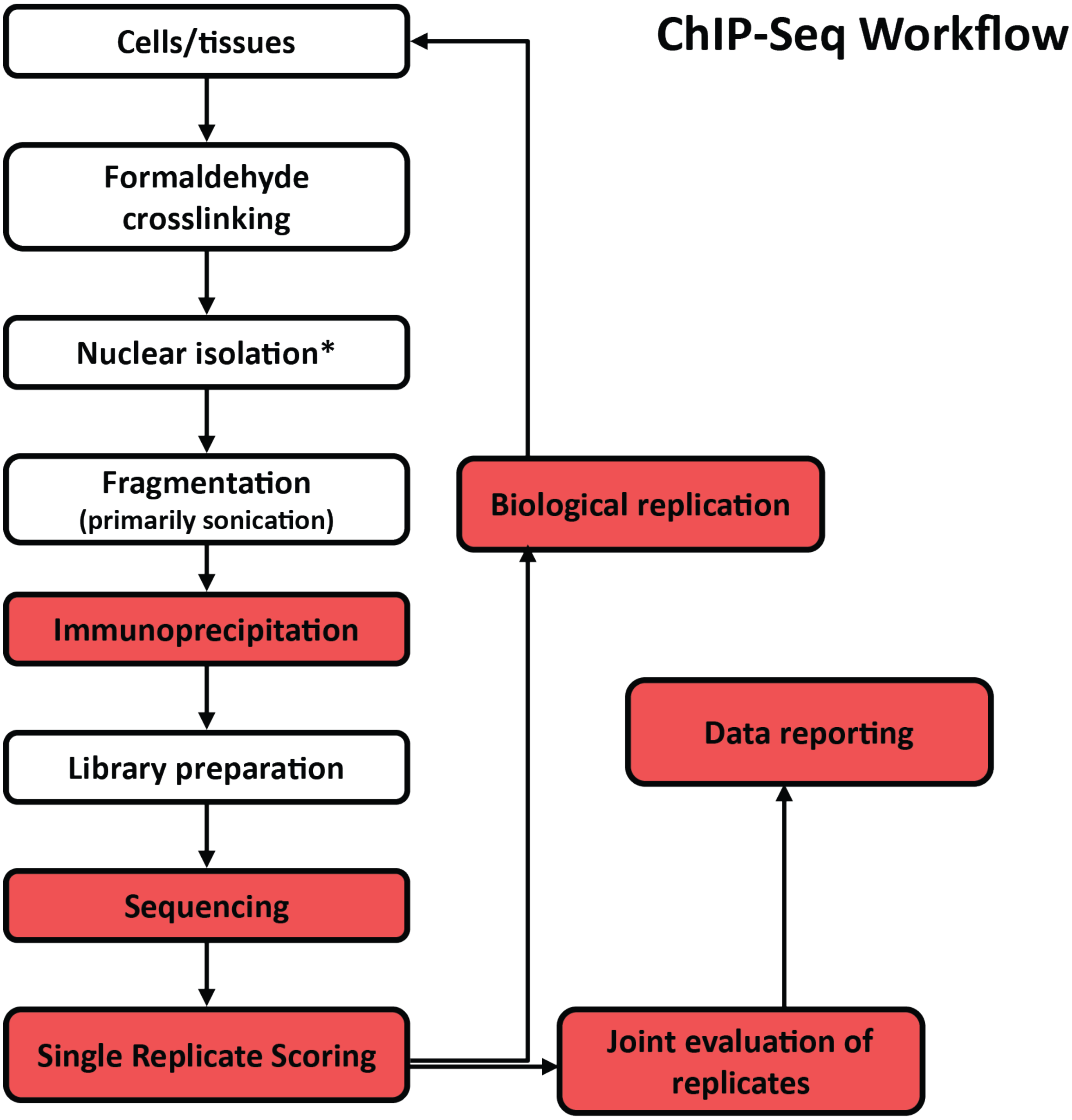
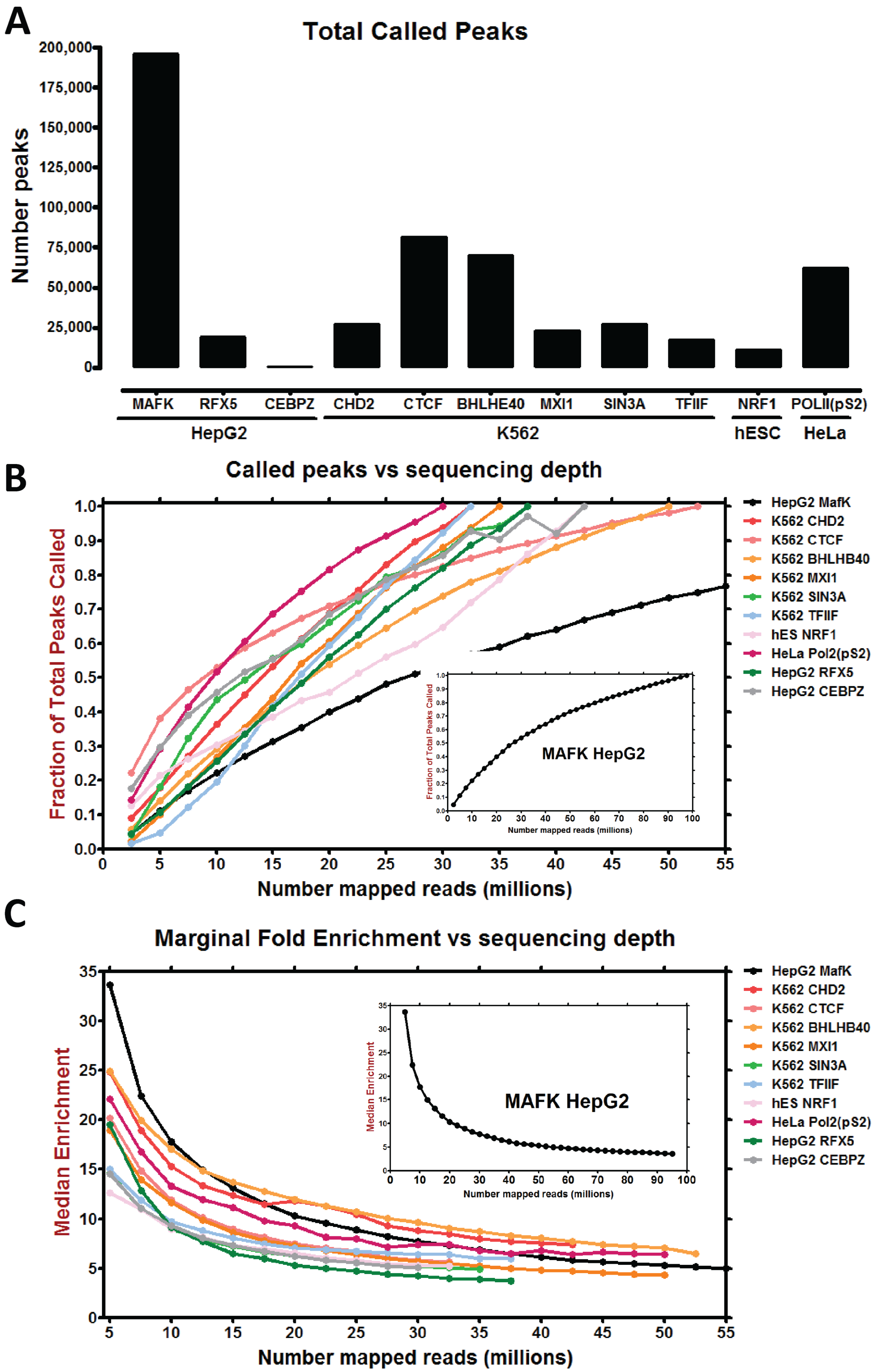


Figure 9.1: Overview of ChIP-seq workflow and antibody characterization procedures. Steps for which specific ENCODE guidelines were established are indicated in red. (*) indicates a commonly used but optional step.

a decade ago (Ren et al. 2000; Lieb et al. 2001; Iyer et al. 2001; Horak and Snyder 2002; Weinmann et al. 2002). In ChIP assays, a transcription factor, co-factor, or other chromatin protein of interest is enriched by immunoprecipitation from crosslinked cells (Gilmour & Lis 1984; Gilmour & Lis 1985; Hecht et al. 1996; Solomon et al. 1988), along with its associated DNA. Genomic DNA sites enriched in this manner were initially identified by qPCR, later by DNA hybridization to a microarray (ChIP-chip) (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak and Snyder 2002, Weinmann et al. 2002), and more recently by DNA sequencing (ChIP-seq) (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007). ChIP-seq has now been widely used for many transcription factors, histone modifications, chromatin modifying com-


A
Total Called Peaks
Number peaks
200,000
175,000
150,000
125,000
100,000
75,000
50,000
25,000
0
MAFK
RFX5
CEBPZ
CHD2
CTCF
BHLHE40
MXI1
SIN3A
TFIIF
NRF1
POLII(pS2)
HepG2
K562
hESC
HeLa
B
Called peaks vs sequencing depth
Fraction of Total Peaks Called
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
35
40
45
50
55
Number mapped reads (millions)
HepG2 MafK
K562 CHD2
K562 CTCF
K562 BHLHB40
K562 MXI1
K562 SIN3A
K562 TFIIF
hES NRF1
HeLa Pol2(pS2)
HepG2 RFX5
HepG2 CEBPZ
MAFK HepG2
Fraction of Total Peaks Called
Number mapped reads (millions)
0
10
20
30
40
50
60
70
80
90
100
C
Marginal Fold Enrichment vs sequencing depth
Median Enrichment
35
30
25
20
15
10
5
0
5
10
15
20
25
30
35
40
45
50
55
Number mapped reads (millions)
HepG2 MafK
K562 CHD2
K562 CTCF
K562 BHLHB40
K562 MXI1
K562 SIN3A
K562 TFIIF
hES NRF1
HeLa Pol2(pS2)
HepG2 RFX5
HepG2 CEBPZ
MAFK HepG2
Median Enrichment
Number mapped reads (millions)
0
10
20
30
40
50
60
70
80
90
100

plexes, and other chromatin-associated proteins in a wide variety of organisms. There is, however, much diversity in the way ChIP-seq experiments are designed, executed, scored and reported. The resulting variability and data quality issues affect not only primary measurements, but also the ability to compare data from multiple studies or to perform integrative analyses across multiple data-types.

The ENCODE and modENCODE Consortia performed more than a thousand individual ChIP-seq experiments for more than 140 different factors and histone modifications in more than 100 cell types in four different organisms (*Drosophila melanogaster*, *Caenorhabditis elegans*, mouse and human), using multiple independent data production and processing pipelines (ENCODE Project Consortium 2004; Celniker et al. 2009, ENCODE Project Consortium 2011). During this work, guidelines, practices, and quality metrics for ensuring the high quality of datasets used for analysis were developed and applied to all ChIP-seq work done by the Consortium (Park 2009). Here they are described, together with supporting data and illustrative examples. Issues common to all ChIP-seq studies are emphasized: immunoprecipitation quality, impact of DNA sequencing depth, scoring and evaluation of datasets, appropriate control experiments, biological replication, and data reporting.

9.2 ChIP Overview

The goals of a genome-wide ChIP experiment are to map the binding sites of a target protein with maximal signal-to-noise ratio and completeness across the genome. The basic ChIP-seq procedure is outlined in Figure 9.1. Cells or tissues are treated with a chemical agent, usually formaldehyde, to crosslink proteins covalently to DNA. This is followed by cell disruption and sonication, or, in some cases, enzymatic digestion, to shear the chromatin to a target size of 100-300 base pairs (bp) (Iyer et al. 2001; Ren et al. 2000). The protein of interest (transcription factor, modified histone, RNA polymerase, etc.) with its bound DNA is then enriched relative to the starting chromatin by purification with an antibody specific for the factor. Alternatively, cell lines expressing an epitope-tagged factor can be generated and the fusion protein immunoprecipitated via the epitope tag.

After immuno-enrichment, crosslinks are reversed, and the enriched DNA is purified and prepared for analysis. In ChIP-chip, the DNA is fluorescently labeled and hybridized to a DNA microarray, along with differentially labeled reference DNA (Ren et al. 2000; Iyer et al. 2001). In ChIP-seq, the DNA is analyzed by high-throughput DNA sequencing. The ENCODE Consortium chose ChIP-seq for human and mouse experiments because it permits comprehensive coverage of large genomes and increases site resolution (Johnson et al., 2007; Robertson et al. 2007). For organisms with small genomes, the modENCODE Consortium has used both ChIP-chip and ChIP-seq, as the arrays available at the time provided high-resolution coverage of small genomes (Gerstein et al. 2010; Roy et al. 2010). In all formats, putatively enriched genomic regions are identified by comparing ChIP signals in the experimental sample with a similarly processed reference sample prepared from appropriate control chromatin or a control immunoprecipitation.

Different protein classes have distinct modes of interaction with the genome that necessitate different analytical approaches (Pepke et al. 2009):

1. Point-source factors and certain chromatin modifications are localized at specific posi-

Figure 9.2 *(preceding page)*: Peak counts depend on sequencing depth. (A) Number of peaks called with Peak-seq (0.01% FDR cut-off) for 11 ENCODE ChIP-seq data sets. (B) Called peak numbers for 11 ChIP-seq data sets as a function of the number of uniquely mapped reads used for peak calling. (Inset) Called peak data for the MAFK data set from HepG2 cells, currently the most deeply sequenced ENCODE ChIP-seq data set (displayed separately due to the significantly larger number of reads relative to the other data sets). Data sets are indicated by cell line and transcription factor (e.g., cell line HepG2, transcription factor MAFK). (C) Fold-enrichment for newly called peaks as a function of sequencing depth. For each incremental addition of 2.5 million uniquely mapped reads, the median fold-enrichment for newly called peaks as compared with an IgG control data set sequenced to identical depth is plotted.

tions that generate highly localized ChIP-seq signals. This class includes most sequence specific transcription factors, their co-factors, and, with some caveats, transcription start site or enhancer-associated histone marks. These comprise the majority of ENCODE and modENCODE determinations and are therefore the primary focus of this paper.

2. Broad-source factors are associated with large genomic domains. Examples include certain chromatin marks (H3K9me3, H3K36me3, etc.) and chromatin proteins associated with transcriptional elongation or repression (e.g. ZNF217) (Krig et al. 2007).

3. Mixed-source factors can bind in point-source fashion to some locations of the genome but form broader domains of binding in others. RNA polymerase II, as well as some chromatin modifying proteins (e.g. SUZ12) behave in this way (Squazzo et al. 2006).

9.3 ChIP-seq Experimental Design Considerations

9.3.1 Antibody and immunoprecipitation specificity

The quality of any ChIP experiment is governed by the specificity of the antibody and the degree of enrichment achieved in the affinity precipitation step. The majority of ENCODE ChIP experiments in human cells and in *Drosophila* embryos have been performed with antibodies directed against individual factors and histone modifications. 145 polyclonal and 43 monoclonal antibodies were used to successfully generate ChIP-seq data as of October 2011. As also discussed below, the majority of antibodies tested for ChIP performance either did not perform well in ChIP or presented concerns about specificity. In the case of polyclonal reagents, lot-to-lot variation can also be significant and confounding. For these reasons, it is necessary that the specificity of antibodies be assessed experimentally separately from the ChIP reaction, through immunoblotting, immunofluorescene, IP coupled with mass-spectrometry, or other means. A detailed description of the ENCODE procedures for carrying out this assessment can be found in Appendix C.

9.3.2 Immunoprecipitation using epitope tagged constructs

Given the challenges in obtaining antibodies for suitable ChIP, an attractive alternative is to epitope tag the factor followed by immunopurification with a well-characterized monoclonal reagent specific for the tag. Epitope-tagging addresses the problems of antibody variation and cross-reaction with different members of multigene families by using a highly specific reagent that can be used for many different factors. However, this introduces new issues relating to how the tagged factor is introduced into cells, whether expression levels are near-physiological, and whether tagging alters the activity of the factor. The level of expression is currently addressed by using large clones carrying as much regulatory information as possible to make the level of expression nearly physiological (Hua et al. 2009; Poser et al. 2008). Higher expression is known to result in occupancy of sites not necessarily occupied at physiological levels (DeKoter and Singh 2000; Fernandez et al. 2003). Within ENCODE, tagged factors have been used most extensively thus far for *C. elegans* studies, where factors have been tagged with GFP and shown to complement null mutants (Zhong et al. 2010). In some cases, information regarding expression is not available and expression from an exogenous promoter has been used. More recently, endogenous knock-in of GFP using CRISP-mediated genome editing (Jinek et al. 2012) has been reported in various systems (Dickinson et al. 2013; Chen et al. 2013; Auer et al. 2014); such approached hold a lot of promise for alleviating some of these issues.

9.3.3 Sequencing depth, library complexity and site discovery

For ChIP-seq performed for a typical point-source DNA binding factor, the number of target sites identified by any contemporary peak calling algorithm typically increases as the number of sequenced reads increases (ENCODE Project Consortium 2011) until the curve ultimately becomes shallower and begins to plateau. This pattern is now generally expected, partly because studies of numerous factors by ENCODE and by other groups have repeatedly found a continuum of ChIP signal strength, rather than a sharply bounded and discrete set of positive sites (ENCODE Project Consortium 2011; Rozowsky et al. 2009). In addition, sites with lower ChIP signal strength are now detected more readily and with greater confidence because of increases in statistical power afforded by more reads. Figure 9.2 shows an analysis of peak calls for eleven human ENCODE ChIP-seq datasets for which deep sequence data (50-100 $\times$ 10^6 mapped reads) were obtained. Clear saturation of peak counts was observed for one factor with few binding sites, but counts continued to increase at varying rates for all other factors, including a case in which >150,000 peaks were called using 100×10^6 mapped reads. Examination of peak signals reveals that the signal enrichments plateau at greater sequencing depths. At 20 $\times$ 10^6 mapped reads, 5-13 fold enrichments are still attainable. The strongest peaks have been identified at this read depth, with new peaks identified after 20 $\times$ 10^6 reads giving enrichments that are, on average, ∼20% of the maximum enrichments identified (Figure 9.2). Interestingly, many additional significant peaks, with enrichment values of 3-7-fold, can still be found by sequencing to much greater depths, indicating that many regions of the genome are enriched in a ChIP-experiment. It is likely that many of these regions correspond to low affinity sites and/or regions of open chromatin that bind factors of interest less specifically.

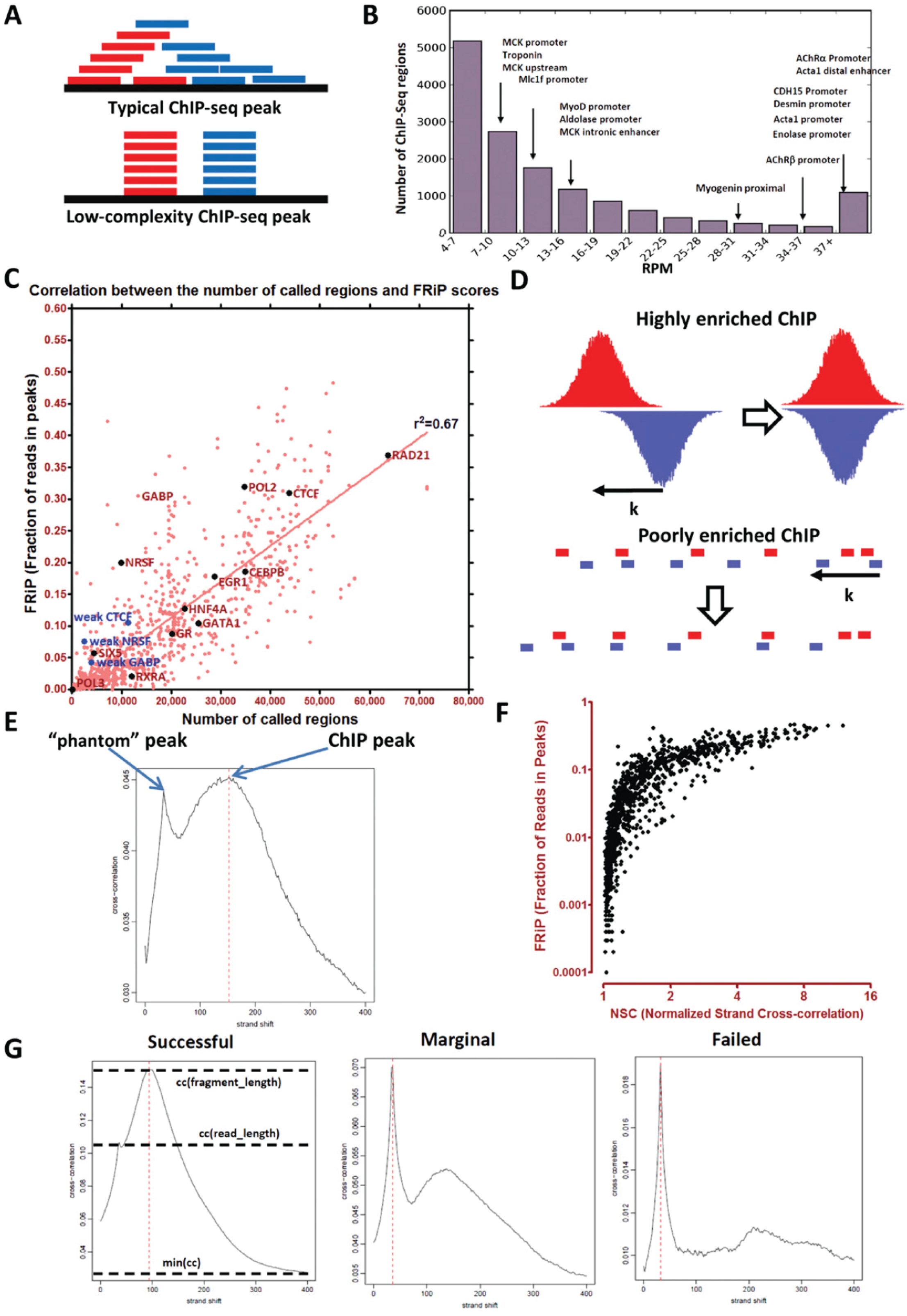
The relationship of ChIP signal strength to biological regulatory activity is a current area of active investigation. A pertinent observation is that biological activity of enhancers, defined in the literature independently of ChIP data, is distributed quite broadly relative to ChIP-seq signal strength. Some highly active transcriptional enhancers reproducibly display modest ChIP signals. This means that one cannot a priori set a specific target threshold for ChIP site number or ChIP signal strength that will assure inclusion of all functional sites (see Discussion section of this chapter). Therefore, a practical goal is to maximize site discovery by optimizing immunoprecipitation and sequencing deeply, within expense constraints. For point-source factors in mammalian cells, a minimum of 10×10^6 uniquely mapped reads are recommended for each biological replicate (providing a minimum of 20×10^6 uniquely mapped reads per factor); for worms and flies a minimum of 4×10^6 uniquely mapped reads per replicate is recommended. For broad areas of enrichment, the appropriate number of uniquely mapped reads is currently under investigation, but at least 20×10^6 uniquely mapped reads per replicate for mammalian cells and 5×10^6 uniquely mapped reads per replicate for worms and flies was produced for most experiments in ENCODE and modENCODE. [1]

Another factor affecting site discovery and reproducibility is the complexity of a ChIP-seq sequencing library. Library complexity is defined as the number of non-redundant DNA fragments. With increasing sequencing of a library, a point is eventually reached where the complexity will be exhausted and the same PCR-amplified DNA fragments will be sequenced repeatedly (Figure 9.3A). Library complexity can vary dramatically, depending on the number of starting nuclei, the efficiency of DNA shearing and size selection range, the efficiency of the immunoprecipitation, and genome size. The objective is to create a library that is sufficiently complex that it does not interfere with the ability of modern peak callers to identify legitimate signals or become the limiting variable in discovering additional sites. Low complexity of libraries is indicative of very low amounts of DNA isolated during the IP or library construction failure. In most cases, the complete repertoire of binding sites for a factor cannot be identified using such datasets.

A useful working metric for complexity is calculated as the fraction of non-redundant mapped reads in a ChIP-seq dataset (non-redundant fraction or NRF), which is similar to a redundancy metric in Heinz et al. 2010[2]. NRF decreases with sequencing depth, and for point source TFs, a reasonable target in mammalian genomes is NRF > 0.8 for 10×10^6 uniquely mapped reads. As sequencing technology improves and read numbers in the hundreds of millions per

[1] More recently, the optimal sequencing depth for histone mark ChIP-seq datasets was examined in detail (Jung et al. 2014), and it was found that $< 20 \times 10^6$ reads is sufficient for fly, and 40-50 $\times$ 10^6 reads is a practical minimum for the human genome.

[2] More recently, computational approaches for the estimation of the total number of unique fragments in a library, which is of course, the quantity one would most like to measure, have been developed (for example the Pre-Seq package, Daley & Smith 2013). NRF is still a highly useful metric for mammalian-sized genomes though, because of the extensive experience in working with them that has accumulated, and the corresponding calibration curve for the metric that exists



$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

lane become feasible, it is expected that even complex libraries from point-source factor can be sequenced at depths greater than necessary. To maximize information that can be obtained for each DNA sequencing run and to prevent oversequencing, barcoding and pooling strategies can be used [3].

9.3.4 Control sample

An appropriate control dataset is critical for analysis of any ChIP-seq experiment because DNA breakage during sonication is not uniform. In particular, some regions of open chromatin are preferentially represented in the sonicated sample, creating a non-uniform background (Auerbach et al. 2009). There are also platform-specific sequencing efficiency biases that contribute to non-uniformity (Dohm et al. 2008). There are two basic methods to produce control DNA samples:

1. DNA is isolated from cells that have been crosslinked and fragmented under the same conditions as the immunoprecipitated DNA and is referred to as "Input" DNA,

2. A "mock" ChIP reaction is performed using a control antibody that reacts with an irrelevant, non-nuclear antigen (often called an IgG control).

For both types of controls, sequencing is performed to a depth at least equal to, and preferably larger than, that of the ChIP sample. For the IgG control, care must be taken that sufficient DNA is recovered to build a high complexity library [4]

Regardless of the type of control used, a separate control is required for each cell line, developmental stage and different condition/treatment because of known and unknown differences in ploidy, genotype and epigenetic features that affect chromatin preparation. To serve as a valid control, the protocol used to build ChIP and control sequencing libraries must be identical (i.e. the number of PCR amplification cycles, fragment size, etc.).

Although rare in our experience, control libraries with a particularly strong sonication biases have been observed and they can adversely affect peak calling (see the following chapter for an extensive discussion of the phenomenon). Although it not always feasible, the optimal study design is to produce a matching control chromatin library for each cell growth, fixation and sonication condition used to prepare chromatin for a ChIP-seq experiment.

Figure 9.3 *(preceding page)*: Criteria for assessing the quality of a ChIP-seq experiment. (A) Library complexity. Individual reads mapping to the plus (red) or minus strand (blue) are represented. (B) Distribution of functional regulatory elements with respect to the strength of the ChIP-seq signal. ChIP-seq was performed against myogenin, a major regulator of muscle differentiation, in differentiated mouse myocytes. While many extensively characterized muscle regulatory elements exhibit strong myogenin binding, a large number of known functional sites are at the low end of the binding strength continuum. (C) Number of called peaks vs. ChIP enrichment. Except in special cases, successful experiments identify thousands to tens of thousands of peaks for most TFs and, depending on the peak finder used, numbers in the hundreds or low thousands indicate a failure. Peaks were called using MACS with default thresholds. (D) Generation of a cross-correlation plot. Reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Pearson correlation between the per-position read count vectors for each strand is calculated. Read coverage as wigglegram is represented, not to the same scale in the top and bottom panels.) (E) Two cross-correlation peaks are usually observed in a ChIP experiment, one corresponding to the read length ("phantom" peak) and one to the average fragment length of the library. (F) Correlation between the fraction of reads within called regions and the relative cross-correlation coefficient for 1052 human ChIP-seq experiments. (G) The absolute and relative height of the two peaks are useful determinants of the success of a ChIP-seq experiment. A high-quality IP is characterized by a ChIP peak that is much higher than the "phantom" peak, while often very small or no such peak is seen in failed experiments.

[3] This has indeed become a common practice since the writing of this text

[4] See the following chapter for more on this issue

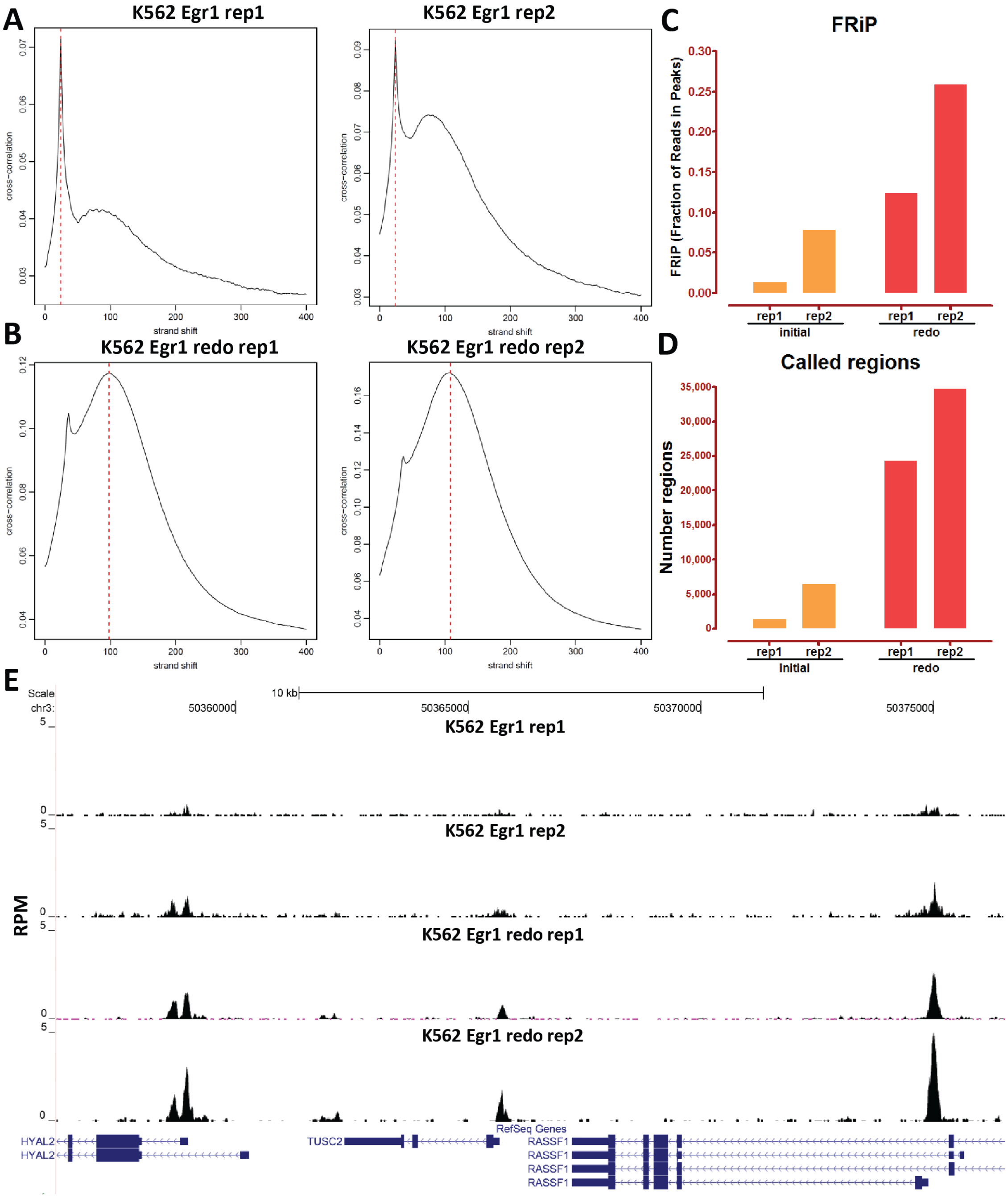


Figure 9.4: Quality control of ChIP-seq data sets in practice. EGR1 ChIP-seq was performed in K562 cells in two replicates. ChIP enriched regions were identified using MACS. However, the cross-correlation plot profiles (A) indicated that both IPs were suboptimal, with one being unacceptable. In agreement with this judgment, ChIP enrichment (C) and peak number (D) also indicated failure. The ChIP-seq assays were repeated (B), with all quality control metrics improving significantly (B,D), and many additional EGR1 peaks were identified as a result. (E) Representative browser snapshot of the four EGR1 ChIP-seq experiments, showing the much stronger peaks obtained with the second set of replicates.

9.3.5 Peak Calling

After mapping reads to the genome, software is used to identify regions of enriched by the ChIP experiment. To identify point-source binding regions from ChIP-seq data, a very large number of peak calling algorithms and corresponding software packages have been developed (MACS/MACS2, Zhang et al. 2008; Feng et al. 2012; ZINBA, Rashid et al. 2011; SISSRs, Jothi et al. 2008; cisGenome, Ji et al. 2008; SICER, Zang et al. 2009; HPeak, Qin et al. 2010; GPS, Guo et al. 2010; USeq, Nix et al. 2008; QUEST, Valouev et al. 2008; PeakSeq, Rozowsky et al. 2009; GLITR, Tuteja et al. 2009; F-Seq, Boyle et al. 2008; FindPeaks, Fejes et al. 2008; CS-Deconv, Lun et al. 2009; PeakRanger, Feng et al. 2011; Sole-Search, Blahnik et al. 2010; CHANCE, Diaz et al. 2012a, Diaz et al. 2012b; NCIS, Liang & Keleş 2012; MAnorm, Shao et al. 2012; CSAR, Muiño et al. 2011; Taslim et al. 2009; PICS, Zhang et al. 2011; and others). The output of these algorithms generally ranks called regions by absolute signal (read counts) or by computed significance of enrichment (e.g. p-values and false discovery rates). Because ChIP signal strength is a continuum with more weak sites than strong ones (Figure 9.3B), the composition of the final peak list depends on specifics of parameter settings and the algorithm used, as well as the quality of the experiment itself. Relaxed thresholds lead to overcalling and a high proportion of false positives. However, moderate overcalling can be useful when there are biological replicates, which can help determine which of the peaks are reproducibly identified (see IDR analysis below). When using standard peak calling thresholds, successful experiments generally identify thousands to tens of thousands of peaks for most TFs (although some exceptions are known; Frietze et al. 2010; Raha et al. 2010) and, depending on the peak finder used, numbers in the hundreds or low thousands indicate an experimental failure. In all study designs, an appropriate control experiment should be performed and should be accounted for in the peak calling, either within the peak calling algorithm employed or by means of direct comparison to the experimental sample. It should be noted that results from different algorithms use different approaches to calculate p-values and false discovery rates (FDR), which means that these values will not be directly comparable between packages.

Calling discrete regions of enrichment for Broad-source factors or Mixed-source factors is more challenging. Methods to identify such regions are emerging (for example ZINBA; Rashid et al. 2011), and MACS2, an updated version of MACS that is specifically designed to process mixed signal types. However, these methods are not as mature as point-source signal processing algorithms. Therefore, statistical and biological metrics for evaluating their performance remain under development, and standards for identifying broad regions of enrichment are not yet in place. [5]

9.3.6 Number of replicates

To ensure that ChIP experiments are reproducible, biological replicate experiments using independent cell cultures, embryo pools, or tissue samples are prepared for ChIP analysis. Initial experiments for RNA Polymerase II indicated that more than two replicates did not significantly improve site discovery (Rozowsky et al. 2009). Initially, either of the following two criteria were applied in order to ensure that a high level of reproducibility is maintained:

1. 80% of the top 40% of the targets identified from one replicate using an acceptable scoring method should overlap the list of targets from the other replicate. This standard was chosen based on the experiences of the ENCODE production groups to allow an achievable threshold of reproducibility for most validated antibodies while generally producing high-quality target lists.

2. Target lists scored using all available reads from each replicate should share more than 75% of targets in common. Reads from replicate experiments that meet either of the above criteria are usually pooled for final peak calling.

However, these were *ad hoc* criteria with not much statistical justification, and were later replaced by the Irreproducible Discovery Rate (IDR) analysis methodology (Li et al. 2011), which has been employed to assess replicate agreement and set thresholds; IDR is discussed in detail below. Examples of replicate experiments that pass IDR are shown in Figure 9.5. It

[5] This statement is still true two years later.

should be noted that the analysis of replicates is generally sensitive to the presence of a weak replicate, and if this is the case, it is desirable that a third replicate be performed to ensure that a comprehensive and reproducible set the binding regions is identified.

9.4 Evaluation of ChIP-seq data

The quality of individual ChIP-seq experiments varies considerably and can be difficult to evaluate, especially when new antibodies are being tested and when little is known about the factor and its binding motif. The first question most experimenters want to answer is: How well did this immunoprecipitation “work”?

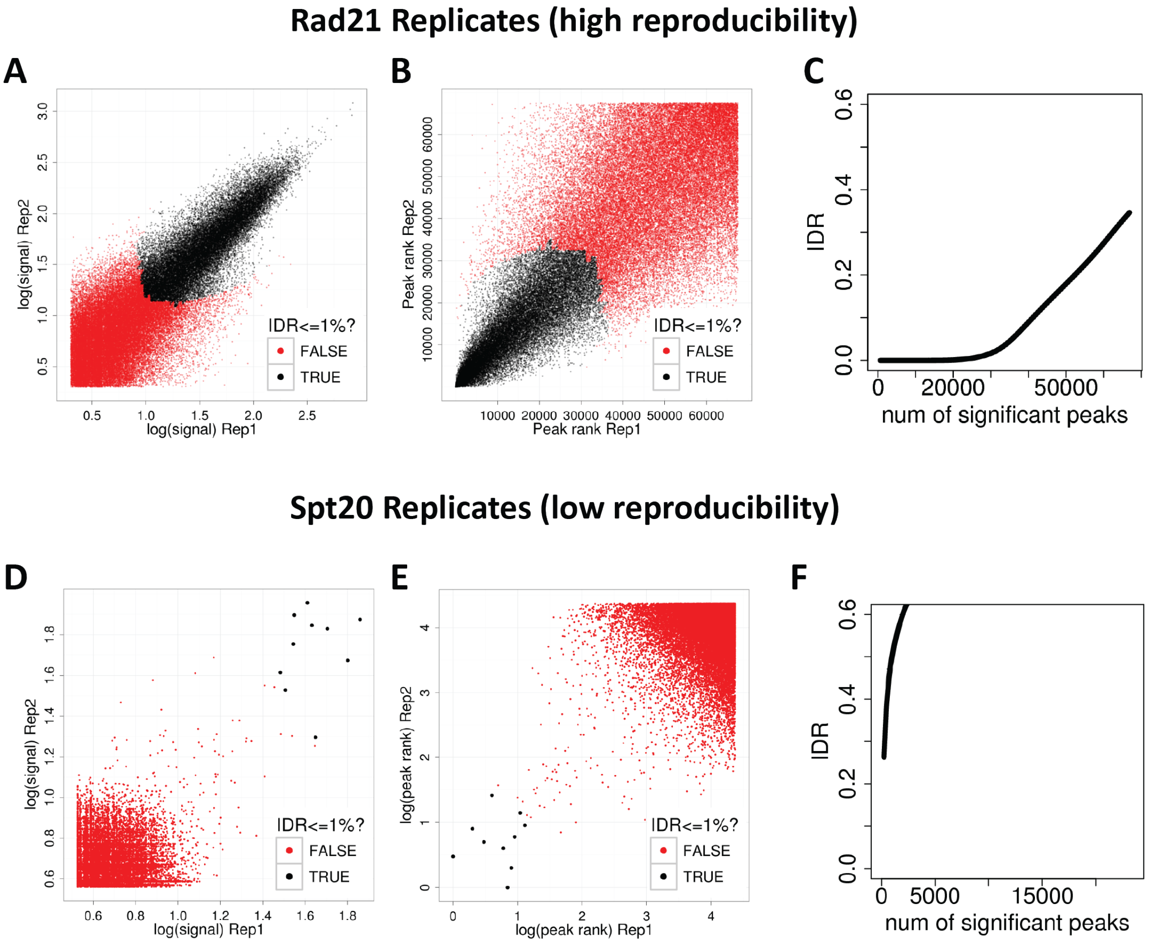


Figure 9.5: The irreproducible discovery rate (IDR) framework for assessing reproducibility of ChIP-seq data sets. (AC) Reproducibility analysis for a pair of high-quality RAD21 ChIP-seq replicates. (D,E) The same analysis for a pair of low quality SPT20 ChIP-seq replicates. (A,D) Scatter plots of signal scores of peaks that overlap in each pair of replicates. (B,E) Scatter plots of ranks of peaks that overlap in each pair of replicates. Note that low ranks correspond to high signal and vice versa. (C,F) The estimated IDR as a function of different rank thresholds. (A,B,D,E) Black data points represent pairs of peaks that pass an IDR threshold of 1%, whereas the red data points represent pairs of peaks that do not pass the IDR threshold of 1%. The RAD21 replicates show high reproducibility with ∼30,000 peaks passing an IDR threshold of 1%, whereas the SPT20 replicates show poor reproducibility with only six peaks passing the 1% IDR threshold.

The ENCODE consortium developed metrics for assessing ChIP-seq quality that are described and applied below, together with traditional inspection-based evaluation. It is worth noting that for each metric, there are some datasets for which it is not ideally suited. However, when they are applied in totality and interpreted as a group, they provide a useful overall assessment of experimental success and data quality.

9.4.1 Browser inspection and previously known sites

A first impression about ChIP-seq quality can be obtained by local inspection of mapped sequence reads on a genome browser, and this remains invaluable. When there is prior biological knowledge of binding at a given genomic location, this site can be examined manually by using the shape and signal strength relative to control reads to gain a sense of ChIP quality. The number and pattern of read tags can give confidence that the known true site has been detected within the large-scale experiment. A true signal is expected to show a clear asymmetrical distribution of reads mapping to the forward and reverse strand around the midpoint (peak) of accumulated reads. This signal should be large compared to the same region for the control library. An example of a set of experiments displaying these characteristics is shown in Figure 9.3C. Of course it is not feasible to inspect the whole genome in this manner, and evaluating a limited number of the strongest sites can misleadingly overestimate the quality of the entire dataset. In addition, it is not possible to compare many different datasets to each other by visual inspection. For these reasons the genome-wide metrics discussed below were developed.

9.4.2 Measuring global ChIP enrichment (FRiP)

For point-source datasets, a first global metric is calculated as the fraction of all mapped reads that fall into peak regions identified by a peak calling algorithm. Typically, only a minority of reads in ChIP-seq experiments come from the read-enriched regions caused by factor occupancy. The remainder is background. Because of this, the fraction of reads falling within peak regions is a useful first-cut metric for the success of the immunoprecipitation, and is called FRiP (Fraction of Reads in Peaks). In general, FRiP values correlates positively with the number of called regions, although there are exceptions, such as NRSF and GABP, which always yield a more limited number of called regions but very high enrichment (Figure 9.3C). In practice, most (787 out of 1052) ENCODE datasets had a FRiP enrichment of 1% or more when peaks were called using MACS with default parameters. When FRiP falls below 1%, the experiment should be further scrutinized.

The FRiP guideline works well when there are thousands to tens of thousands of called occupancy sites in a large mammalian genome. However, passing this threshold does not automatically mean that an experiment is successful and a FRiP below the threshold does not automatically mean failure. For example, in cases such as human RNA Polymerase III where there are very few true binding sites (Frietze et al. 2010; Raha et al. 2010), a FRiP value of less than 1% is obtained. At the other extreme, lesser-quality ChIP experiments using combinations of antibodies and factors that usually have very high enrichment and/or numbers of binding sites can still result in FRiP scores that exceed those generally obtained for most factors (Figure 9.3C). Thus, FRiP is very useful as a quality control measure for comparing results for broadly expressed factors using the same antibody across cell lines or using different antibodies. A caveat is that FRiP is sensitive to the specifics of peak calling, including the way the algorithm defines regions of enrichment and the parameters and thresholds used. This means that all FRiP values that are compared should be derived from peaks uniformly called by one algorithm and parameter set.

9.4.3 Cross-correlation analysis

A very useful ChIP-seq quality metric that is independent of peak calling is cross-correlation. It is based on the fact that a high-quality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered around the binding site. As illustrated in Figure 9.3D, these “true signal” sequence tags are positioned at a distance from the binding site center that depends on the fragment size distribution (Kharchenko et al. 2008). A control experiment, such as sequenced input DNA, lacks this pattern of shifted stranded tag

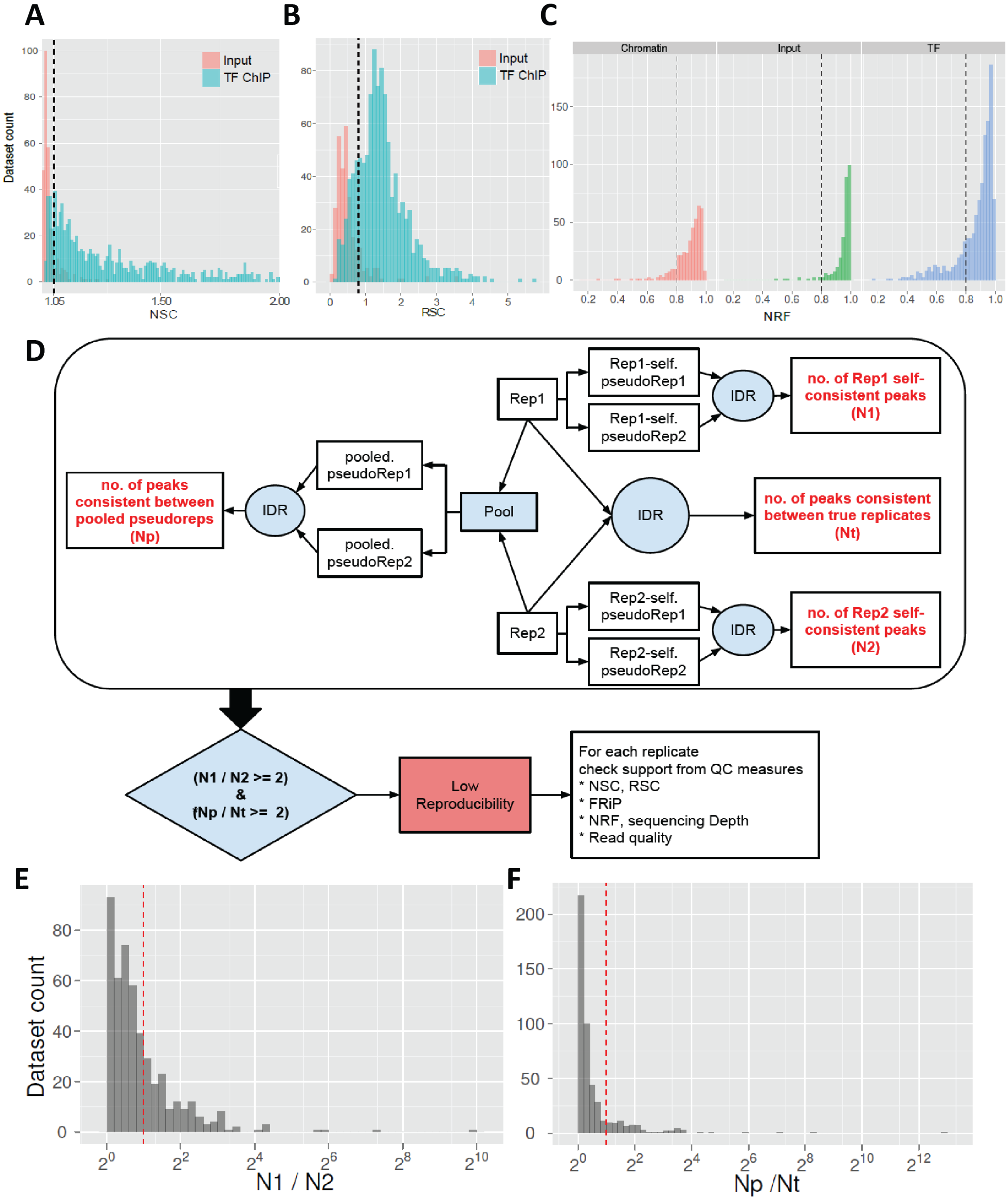


Figure 9.6: Analysis of ENCODE data sets using the quality control guidelines. (AC) Thresholds and distribution of quality control metric values in human ENCODE transcription-factor ChIP-seq data sets. (A) NSC, (B) RSC, (C) NRF. (D) IDR pipeline for assessing ChIP-seq quality using replicate data sets. (E,F) Thresholds and distribution of IDR pipeline quality control metrics in human ENCODE transcription factor ChIP-seq data sets. (Dashed lines) Current ENCODE thresholds for the given metric, which are NSC > 1.05 (A); RSC > 0.8 (B); NRF > 0.8, $N_1/N_2 \geq 2$ (where N_1 refers to the replicate with higher N) (E); $N_p/N_t \geq 2$ (F).

densities. This made it possible to develop a metric that quantifies fragment clustering (IP enrichment) based on the correlation between genome-wide stranded tag densities. It is computed as the correlation of the Crick strand to the Watson strand, after shifting Watson by k base pairs (Figure 9.3D). This typically produces two peaks when cross-correlation is plotted against the shift value: A peak of enrichment corresponding to the fragment length and a peak of short fragments corresponding to the read length (Figure 9.3E).

The normalized ratio between the enrichment peaks and the background correlation (NSC):

$$NSC = \frac{\text{cross-correlation(fragment length)}}{\text{min(cross-correlation)}} \tag{9.1}$$

and the ratio between the read length peak and the enrichment peak, called RSC:

$$RSC = \frac{\text{cross-correlation(fragment length)} - \text{min(cross-correlation)}}{\text{cross-correlation(read length)} - \text{min(cross-correlation)}} \tag{9.2}$$

are useful metrics for assessing ChIP-enrichment. High-quality ChIP-seq datasets have larger enrichment peaks compared to the read-length peak, whereas failed ones and inputs have little or no such peak (Figure 9.3G). Most (797 of 1052) ENCODE datasets had an NSC ratio greater than 1.1 An example of a result from a failed experiment is shown in Figure 9.4. In general, a continuum between the two extremes is observed, and broad-source datasets are expected to have flatter cross-correlation profiles than point-source ones, even when they are of very high quality. As expected, the NSC/RSC and FRiP metrics are strongly and positively correlated for the vast majority of experiments (Figure 9.3F). As with the other quality metrics, even high quality datasets generated for factors with few genuine binding sites tend to produce relatively low NSCs.

9.4.4 Consistency of replicates: Analysis using IDR

To identify high-confidence data and to eliminate biologically unstable measurements, the ENCODE Consortium made its goal a minimum of two successful independent biological replicates, with each experiment passing the basic quality control filters described above. To take advantage of the reproducibility information provided by replicates, the IDR (irreproducible discovery rate) statistic was developed for ChIP-seq (Li et al. 2011).

Given a set of peak calls for a pair of replicate datasets, the peaks can be ranked based on a criterion of significance, such as the p-value, the q-value, the ChIP to input enrichment, or the read coverage for each peak (Figure 9.5A, 9.5B, 9.5D, and 9.5E). If two replicates measure the same underlying biology, the most significant peaks, which are likely to be genuine signals, are expected to have high consistency between replicates, whereas peaks with low significance, which are more likely to be noise, are expected to have low consistency. The latter peaks exhibit higher variability in their ranks and begin to appear at the noise level. If the consistency between a pair of rank lists that contains both significant and insignificant findings is plotted, a change (discontinuity) in consistency is expected [Figures 9.5C and 9.5F]. This self-consistency discontinuity provides an internal indicator of the transition from signal-to-noise and suggests how many peaks have been reliably detected.

The IDR statistic quantifies the above expectations of consistent and inconsistent groups by modeling all pairs of peaks present in both replicates as belonging to one of two groups: a reproducible group and an irreproducible group (Li et al. 2011). In general, the signals in the reproducible group are more consistent (i.e., have a larger correlation coefficient) and are ranked higher than the irreproducible group. The proportion of identifications that belong to the noise component and the correlation of the significant component are estimated adaptively from the data. The IDR provides a score for each peak, which reflects the posterior probability that the peak belongs to the irreproducible group.

A major advantage of IDR is that it can be used to establish a stable threshold for called peaks that is more consistent across laboratories, antibodies, and analysis protocols (e.g., peak callers) than are FDR measures (Li, et al. 2011). Increased consistency comes from the fact that IDR uses information from replicates, whereas the FDR is computed on each replicate independently. The application of IDR to real-life data is shown in Figure 9.5. A pair of high quality Rad21 ChIP-seq replicates display good consistency between IDR ranks for a large number (∼28,000) of highly reproducible peaks (Figures 9.5AB), with a clear inflection between the signal and noise populations near the 1% IDR value (Figure 9.5C). In contrast, a pair of Spt20 replicates, which had already been flagged as low-quality based on the individual FRiP and NSC/RSC metrics, display very low reproducibility as shown by IDR, with very few significant peaks, and they show no visible inflection in the IDR curve (Figure 9.5F). It is important that the peak-calling threshold used as input to IDR analysis not be so stringent that the noise component is entirely unrepresented in the data, because the algorithm requires sampling of both signal and noise distributions to separate them. A caution in applying IDR is that it is dominated by the weakest replicate. That is, the IDR is a conservative statistical approach, and hence if one replicate is quite poor, many "good" peaks from the higher quality replicate will be rejected by IDR analysis.

9.4.5 Metrics Applied In Practice

The application of the ChIP-seq quality metrics to a failed experiment is shown in Figure 9.4. Initially, two EGR1 ChIP-seq replicates were generated in the K562 cell line. Based on the cross-correlation profiles, the number of called regions, and the FRiP score, these initial replicates were flagged as marginal in quality. The experiments were then repeated, with all quality control metrics improving considerably. On this basis, the superior measurements replaced the initial ones in the ENCODE database. A summary of the distribution of the values of the different metrics and of the IDR pipeline used for the joint assessment of replicates is shown in Figure 9.6.

9.5 Discussion

As part of the ENCODE Project, we and others developed a set of working best practices and guidelines for ChIP-seq experiments based on more than 1,000 experiments as of October 2011 (and many more since then). They addressed the central issues of immune reagent specificity and performance by establishing a menu of primary and secondary methods for antibody characterization, and the development and application of global metrics to assess the quality of several aspects of an individual ChIP-seq experiment: library complexity (which can be measured by the non-redundant fraction (NRF)), immunoenrichment (which can be measured by the fraction of reads in called peaks, FRiP, and by cross-correlation analysis). How different aspects of data quality interact with specific uses of ChIP-seq data is discussed below.

9.5.1 Challenges in obtaining high quality affinity reagents

Certainly one of the major challenges in ChIP is the availability of high quality affinity reagents. There are approximately 1500 transcription factors in humans (Vaquerizas et al. 2009), but fewer than 200 antibodies against different transcription factors have passed ENCODE characterization criteria. Because only 25% of antibodies, on average, pass quality controls, it is likely that over 6000 antibodies will need to be examined to complete the analysis of all human transcription factors. The use of epitope-tagged constructs (especially knock-ins into the enodgenous loci that are expressed at correspondingly endogenous levels) will help generate data for many factors, but they will still not be suitable for introduction into human tissues and may not work well for all cell lines of relevance. Thus, significant effort is needed to expand our antibody repertoire. The use of renewable reagents (such as monoclonal antibodies) will be particularly valuable so that well-characterized and plentiful reagents can be distributed to and used by the entire scientific community.

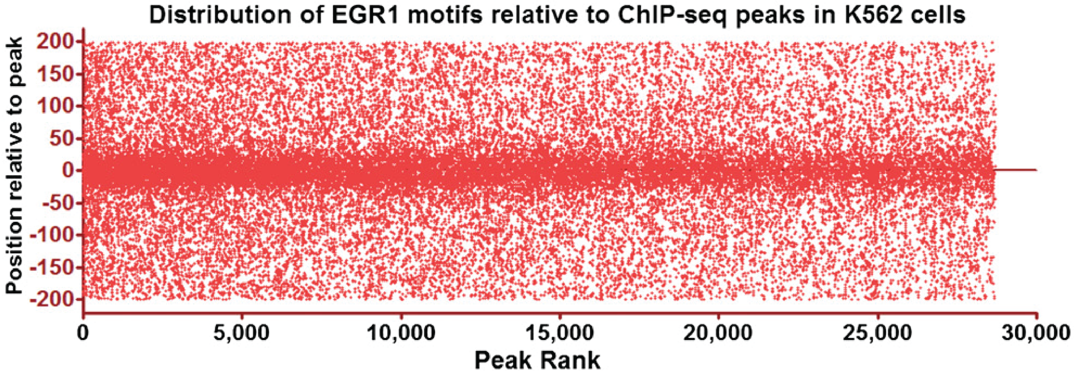


Figure 9.7: Distribution of EGR1 motifs relative to the bioinformatically defined peak position of EGR1-occupied regions derived from ChIP-seq data in K562 cells. Regions are ranked by their confidence scores as called by SPP. Motifs were called using MEME (Bailey et al., 2009; version 4.6.1), based on the top 500 regions. The top motif was used and its instances in all called peaks identified using the approach described in Mortazavi et al. 2006. The position of each motif instance relative to the peak summit is plotted.

9.5.2 How good can a ChIP-seq experiment be?

Thus far, the most successful point-source factor experiments for ENCODE have FRiP values of 0.2 to 0.5 (factors such as NRSF, GABP, and CTCF; Figure 9.3C) and NSC/RSC values of 5 to 12. This implies very high biochemical enrichment. These experiments produced different site numbers: the peak caller SPP reported 50,000 for CTCF but only a few thousand for NRSF, arguing that different point-source factors vary considerably in the number of occupied regions, even when technical quality issues are minimal. Although these quality scores and characteristics can be routinely obtained for the best-performing factor/antibody combinations, they are not the rule. For most transcription factors, the ChIP quality metrics obtained are substantially lower and more variable. There are likely multiple determinants of successful enrichment, and they are not all controllable or easy to measure. The quality of antibody (affinity and specificity) is certainly very important, but epitope availability within fixed chromatin, sensitivity of the antibody to post-translational modification of the antigen, the nuclear levels of protein, and other physical characteristics of the protein-DNA interaction can also contribute.

It is common for a lower-quality replicate, by the criteria of FRiP, NSC/RSC and track inspection, to identify thousands fewer sites than the best available replicate. Are sites detected in only the best ChIP replicate “real” in the sense of reflecting in vivo occupancy? Motif analysis suggests that many are. A representative example is shown in Figure 9.7, where the position of Egr1 motifs relative to Egr1 ChIP-seq peaks is shown. The known binding motif is prominent and concentrated centrally under the ChIP peaks, as expected if the motif mediates occupancy; importantly, the central location of the motif is observed even in the low ranking peaks and this trend seems to continue below the peak calling cut-off, suggesting the existence of additional true sites. This means that the true number of binding sites and how exhaustively a ChIP-seq experiment identifies them is rarely clear, especially when a factor is assayed for the first time.

9.5.3 How good does a ChIP-seq experiment need to be?

It would be ideal if every ChIP experiment mapped all occupancy sites in the genome that are biologically meaningful with minimal false positives. The main impediment to this result is that the field has not learned to determine, a priori, the level of ChIP signal above which all biologically functional sites have been identified, or even if this is a valid concept. We have observed that some biologically important sites can have modest ChIP-seq signals while some sites with very high enrichment fail to give positive functional readouts (Figure 9.3B). Until more biologically-informed thresholds are established, the best practical guidance for thresholds of sensitivity, specificity and replicability will depend on how the data is used. Below, four different common uses for ChIP-seq data are outline, ranging from relaxed to stringent in quality requirements.

1. **Motif analysis**. Deriving DNA sequence motifs for a ChIP-assayed factor is relatively simple and can be performed successfully with most ENCODE ChIP-seq datasets. Experiments that pass suggested thresholds for NRF, FRiP, and NSC/RSC typically produce thousands of regions, a sub-sample of which can be readily used to deduce the recognition motif, assuming that the protein bound is a sequence-specific factor. Causal motifs are typically centrally positioned and this can be used as a confirming diagnostic. Motif finding can also be successful from marginal quality data that fall below recommended quality metric thresholds (especially if only the top-ranked peaks are analyzed). However, the risk of artifacts increases if lower quality data is used and results from such analyses should be interpreted and validated with special care.

2. **Discovering regions to test for biological function** (such as transcriptional enhancement, silencing, or insulation). Biologists often use ChIP-seq data to identify candidate regulatory regions at loci of interest. When the goal is to find a set of representative regulatory domains, data of modest quality can be effective. In general, inspection of ChIP signals is strongly advised before investing deeply in functional and/or mutagenesis studies, especially if the criteria for selecting regions of interest are computational. However, when the aspiration is to identify and sample all regulatory regions bound by a factor, weaker datasets are not adequate.

3. **Deducing and mapping combinatoric occupancy**. Typical *cis*-acting regulatory modules (CRM) are occupied by multiple factors (Ghisletti et al. 2010; He et al. 2011a; He et al. 2011b; Lin et al. 2010; Tijssen et al. 2011; Wilson et al. 2010) and histones present at these elements are modified with multiple marks (Barski et al. 2007; Mikkelsen et al. 2007; Wang et al. 2008). A frequent goal of ChIP-seq studies is to deduce a combination of factors that mediate a common regulatory action at multiple sites in the genome. The presence of one or more weak datasets that fail to identify significant fractions of the true occupancy sites can seriously confound such an analysis. Therefore, only high quality datasets should be used for such studies.

4. **Integrative analysis**. A new frontier of whole genome analysis is the integration of data from many (hundreds or thousands) experiments with the goal of uncovering complex relationships. These endeavours typically use sophisticated machine learning methods (Ernst & Kellis 2010; Ernst et al. 2011; Mortazavi et al. 2013) with complex and varying sensitivity to ChIP strength; such efforts can be significantly affected by data quality. Again, only high quality datasets are recommended to be used for such studies.

9.5.4 Uncertainties in distinguishing high quality from low quality datasets

Evaluating ChIP-seq data quality includes the challenge of distinguishing technical versus biological sources of noise or error. I use the TAF1 subunit of the TFIID complex, part of the transcription initiation machinery, as an example. Given the known biological functions of TAF1, one might expect that the set of genomic locations occupied by TAF1 would reflect the number and identity of active promoters in each cell type. Based on RNA-seq measurements of gene expression, it can be concluded that the number of active transcriptional start sites is similar in most cell types (Figure 9.9). Yet we have observed substantial differences in the number of identified TAF1-bound regions that appear to depend on cell type (Figure 9.8). One explanation is that this is entirely due to technical variability in the quality of the TAF1 ChIP experiments. However, it has been suggested that TAF1 does not play a role in transcription initiation in certain cell types (Deato & Tjian 2007); ChIP-seq experiments against TAF1 in such cell lines would appear very similar to technical failures. Additional experiments will be required to discriminate between these possibilities.

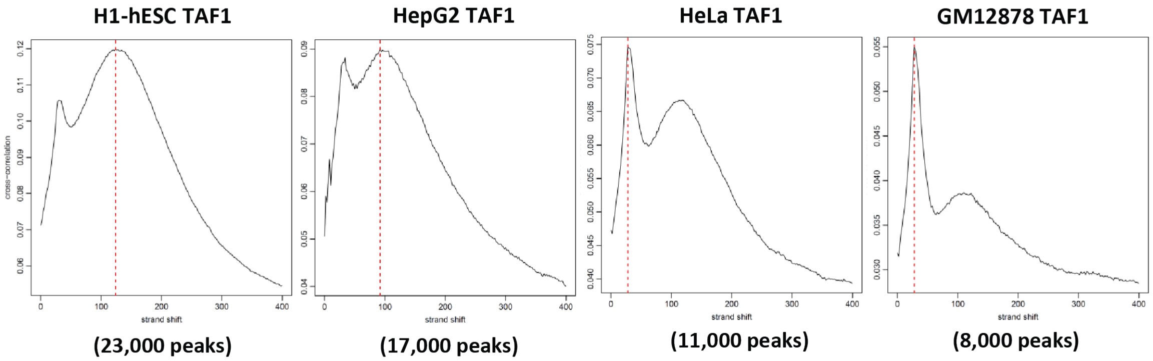


Figure 9.8: Cross-correlation profiles and number of TAF1 ChIP-seq peaks in different ENCODE cell lines. Regions are called from ChIP-seq data using MACS (version 1.4). The best replicates for each cell line are shown, i.e. the low number of peaks in GM12878 cells was consistently observed in multiple replicates ($n > 5$ for GM12878).

9.6 Conclusion

ChIP experiments that map the genomic distribution of transcription factor and modified histone binding sites have proven to be an important tool across a wide range of organisms and in different tissues and cell types. The quality control metrics described above should provide assistance to the scientific community with the goal of ensuring that high quality data are pro-

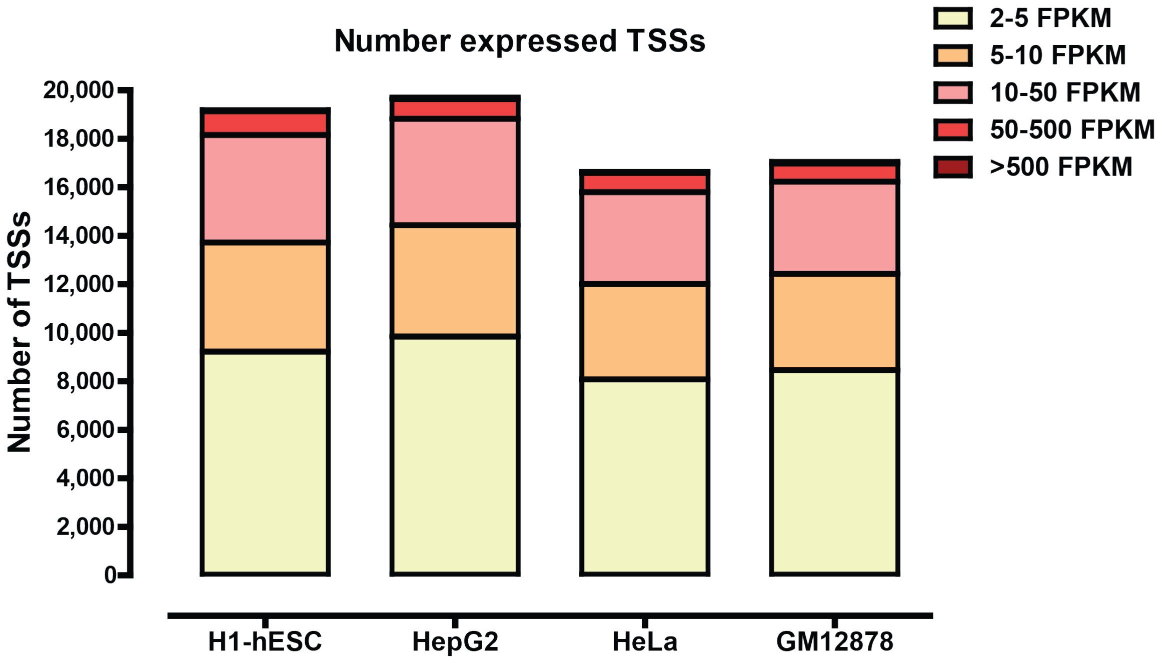


Figure 9.9: Number of expressed transcription start sites in four ENCODE cell lines. Show is the number of expressed transcription start sites (TSSs) at the indicated FPKM levels, based on RNA-seq data for each cell line. The FPKM for each TSS was calculated as the sum of the FPKMs of all transcripts containing that TSS: $\text{FPKM}_{TSS} = \sum_{T \ni TSS} \text{FPKM}_T$. Transcript-level quantification was carried out on GENCODE version 7 (Harrow et al. 2012) using Cufflinks version 0.9.3 (Trapnell et al. 2010)

duced and reported, which will not only enable the mapping of regulatory information and networks, but will also be critical in elucidating the effects of genomic variation in mediating human traits and diseases.

10

Large-scale quality analysis of published ChIP-seq data

This chapter contains material that was published as:

Marinov GK, Kundaje A, Park PJ, Wold B. 2014. Large-scale quality analysis of published ChIP-seq data. *G3* **4**(2):209-23. doi: 10.1534/g3.113.008680.

The paper is reprinted in Appendix K

Abstract

ChIP-seq has become the primary method for identifying in vivo protein-DNA interactions on a genome-wide scale, with nearly 800 publications involving the technique in PubMed as of December 2012. Individually and in aggregate these data are an important and information-rich resource. However, uncertainties about data quality confound their use by the wider research community. In the previous chapter, I described the metrics developed and applied by the ENCODE Project Consortium to objectively measure ChIP-seq data quality (which are also reviewed here, in some cases more extensively). The ENCODE quality analysis was useful for flagging datasets for closer inspection, eliminating or replacing poor data, and for driving changes in experimental pipelines. However, there had been no similarly systematic quality analysis of the large and disparate body of published ChIP-seq profiles. To address this question, I carried a uniform analysis of vertebrate transcription factor ChIP-seq datasets in the Gene Expression Omnibus (GEO) repository as of April 1st 2012. The majority (55%) of datasets scored as highly successful, but a substantial minority (20%) were of apparently poor quality, and another ∼25% were of intermediate quality. I discuss how different uses of ChIP-Seq data are affected by specific aspects of data quality, and highlight exceptional instances for which the metric values should not be taken at face value. Unexpectedly, I discovered that a significant subset of control datasets (i.e. no-immunoprecipitation and mock-immunoprecipitation samples) display an enrichment structure similar to successful ChIP-seq data. This can, in turn, affect peak calling and data interpretation. In the future, ChIP-seq quality assessment similar to that used here could guide experimentalists at early stages in a study, provide useful input in the publication process, and be used to stratify ChIP-seq data for different community-wide uses.

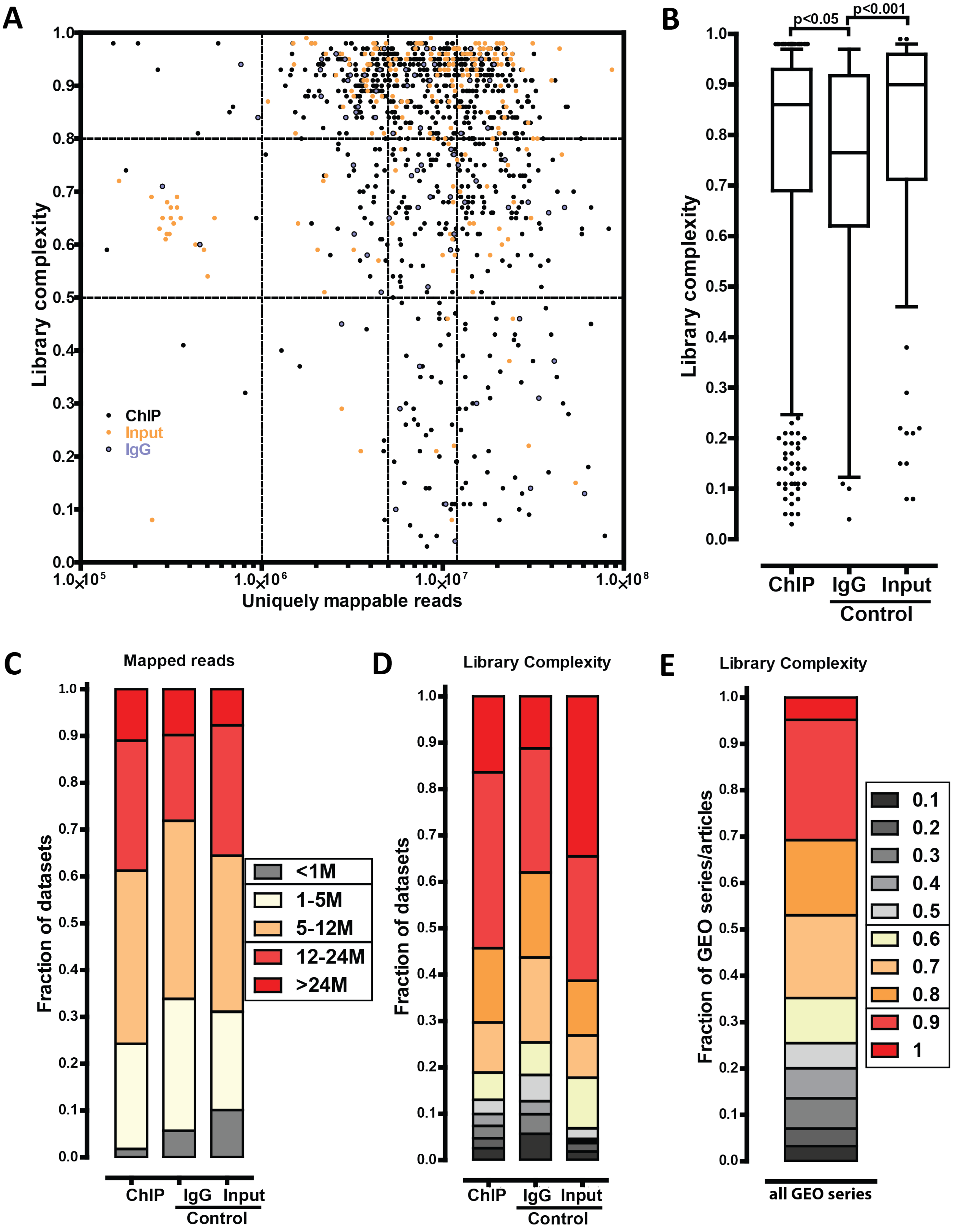
10.1 Introduction

Chromatin immunoprecipitation (ChIP) (Gilmour and Lis 1984; Gilmour and Lis 1985; Solomon et al. 1988) experiments identify sites of occupancy by specific transcription factors, cofactors, and other chromatin-associated proteins as well as histone modifications. Such proteins are concentrated at specific loci via direct binding to DNA or by indirect binding mediated by other proteins or RNA molecules. In most ChIP protocols, proteins are first crosslinked to DNA, most often using formaldehyde. The fixed chromatin is sheared, and an antibody specific for the protein or histone modification of interest is used to retrieve protein:DNA complexes from which the DNA segments are released and then assayed. The assay was first applied to individual transcription factor/promoter complexes by using qPCR to detect enrichment over specific DNA segments (Hecht et al. 1996). Subsequent adaptations extended it to large sets of promoters or other genomic regions by using microarrays (ChIP-on-Chip/ChIP-Chip) (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak and Snyder 2002; Weinmann et al. 2002). Ultimately, the entire genome became accessible with the advent of high-throughput sequencing and the development of ChIP-seq (Johnson et al. 2007; Barski et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007).

In all cases, preferential enrichment of a given immunoprecipitated DNA segment is detected and quantified by comparing it with a control experiment, in which there is no specific antibody enrichment step. These controls can be generated from sonicated DNA prior to immunoprecipitation (Input) or a mock immunoprecipitation with an unrelated antibody (IgG). Sequencing-based ChIP has become the method of choice because it enables genome-wide coverage, even for large genomes, and because of its superior signal-to-noise characteristics compared to alternative methods. Since its initial development, ChIP-seq has been used in hundreds of publications (778 in PubMed as of December 18th, 2012), including by the ENCODE consortium (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012), to map occupancy over a hundred human transcription factors and cofactors in a diverse collection of cell lines. (Gerstein et al. 2012; Wang et al. 2012).

A basic question for any ChIP-seq experiment is how successful it has been. It has taken several years for the field to develop objective ways to quantify key aspects of success in immunoprecipitation enrichment, library building and final sequencing. Poor datasets that have high false negative rates in peak calling are a predictable pitfall that has significant downstream consequences for some kinds of biological and computational analyses. For example, when lower quality data-sets are used for integrative analyses that are sensitive to false negatives, incorrect inferences and conclusions become likely (see Discussion). In estimating data quality, the traditional approach of visual inspection at a limited number of sites (often previously well-characterized using low-throughput approaches) is inefficient, subjective and can ultimately be deceptive. It is possible (and commonly observed in practice) that sites, the biological importance of which has been defined by independent functional assays, can fall below the sensitivity threshold of a poor or mediocre ChIP-seq experiment. Moreover, there is no current way to predict, a priori, the number of sites in the genome that should be detected for a given factor and cell type. Most transcription factors studied thus far reproducibly occupy thousands to tens of thousands of sites (ENCODE Project Consortium 2012; Landt et al. 2012); Thus a dataset, for which several thousand sites have been called, might in fact be capturing a minority of true positive interactions, or it might encompass virtually all biologically pertinent sites. To help address the problem of data assessment, as part of the ENCODE project, we and others developed a comprehensive set of ChIP-seq quality control metrics and guidelines (Landt et al. 2012), which were adopted by the Consortium and applied to all of its datasets. Substandard datasets were consequently replaced, flagged as substandard, and/or removed from analysis (ENCODE Project Consortium 2012; Landt et al. 2012).

Incorporating published datasets into an ongoing study can bring new biological insights and avoid unnecessary duplication of work. Variable quality of published data can be a significant barrier to these uses of existing data. They are the product of work from many different labs, with invaluable expertise is specific biological systems, but also using many variations in ChIP-seq experimental protocols and bioinformatics treatments. The extent and nature of the variations has not been assessed globally and systematically. In this work, I examined the GEO submission series containing vertebrate transcription factor ChIP-seq datasets and found that ∼20% of datasets score as being of low-quality with additional ∼25% exhibiting intermediate ChIP enrichment. I also noticed that roughly a third of studies have control datasets with a high degree of read clustering that is nor-


A
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Library complexity
ChIP
Input
IgG
1.0x10⁵
1.0x10⁶
1.0x10⁷
1.0x10⁸
Uniquely mappable reads
B
p<0.05
p<0.001
Library complexity
ChIP
IgG
Input
Control
C
Mapped reads
Fraction of datasets
<1M
1-5M
5-12M
12-24M
>24M
ChIP
IgG
Input
Control
D
Library Complexity
Fraction of datasets
ChIP
IgG
Input
Control
E
Library Complexity
Fraction of GEO series/articles
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
all GEO series

mally expected only in ChIP-seq datasets. This was observed more often for the IgG control design than for input DNA controls. These and related observations suggest that routine characterization and reporting of the quality of ChIP-seq data be applied.

10.2 Results

10.2.1 Dataset collection, data processing and quality metrics

I downloaded all GEO series containing ChIP-seq datasets for vertebrate transcription factors or chromatin modifying and remodeling proteins, along with their corresponding control libraries, submitted prior to April 1st 2012. I excluded ENCODE datasets as they have previously been subjected to this quality assessment (ENCODE Project Consortium 2012) and the results were summarized in (Landt et al. 2012), although I also provide here a summary of ENCODE transcription factor (TF) ChIP-seq data from the two main production groups in Figures 10.13 and 10.14.

A different logic led to the exclusion of histone modification and RNA Polymerase II datasets. First, in our experience, ChIP-seq against these targets is robust to experimental variation and the success rate is reliably high (provided the antibody reagents used are of high quality). Second, an especially large proportion of published data are for histone marks. The effect of including all of these in the survey would have been to obscure or skew the trends for transcription factors and cofactors. Finally, the currently available quality control metrics were designed and are best suited for transcription factor data that produce highly localized "point-source" occupancy (as they quantify the extent of read clustering in the genome). This means that the metrics themselves need to be interpreted differently if they are applied to, for example, repressive histone marks such as H3K9me3 and H3K27me3, which form large "broad-source" regions of enrichment (Pepke et al. 2009). Arguably, these data will need their own metrics and this will be a challenge for the future.

The final collection of datasets contained 191 GEO series containing a total of 917 ChIP-seq and 292 control libraries. Except for a limited number of cases in which a GEO series was associated with multiple publications, two or three GEO series were associated with the same publication, or a GEO series has not yet been used in a publication, there is a one-to-one relationship between GEO series and published articles in the literature (An et al. 2011; Ang et al. 2011; Avvakumov et al. 2012; Barish et al. 2010; Barish et al. 2012; Bergsland et al. 2011; Bernt et al. 2011; Bilodeau et al. 2009; Blow et al. 2010; Boergesen et al. 2012; Botcheva et al. 2011; Brown et al. 2011; Bugge et al. 2012; Canella et al. 2012; Cao et al. 2010; Cardamone et al. 2012; Ceol et al. 2011; Ceschin et al. 2011; Chen et al. 2008; Cheng et al. 2009; Cheng et al. 2012; Chi et al. 2010; Chia et al. 2010; Chicas et al. 2010; Chlon et al. 2012; Cho et al. 2012; Corbo et al. 2010; Costessi et al. 2011; Cuddapah et al. 2009; De Santa et al. 2009; Doré et al. 2012; Durant et al. 2010; Ebert et al. 2011; Fan et al. 2012; Fang et al. 2011; Feng et al. 2012; Fong et al. 2012; Fortschegger et al. 2010; Gao et al. 2012; Gotea et al. 2010; Gowher et al. 2012; Gu et al. 2010; Han et al. 2010; Handoko et al. 2011; He et al. 2011; Heikkinen et al. 2011; Heinz et al. 2010; Heng et al. 2010; Ho et al. 2010; Hollenhorst et al. 2010; Holmstrom et al. 2011; Horiuchi et al. 2011; Hu et al. 2010; Hu et al. 2011; Hunkapiller et al. 2012; Hutchins et al. 2012; Johannes et al. 2010; Joseph et al. 2011; Jung et al. 2010; Kagey et al. 2010; Kassouf et al. 2010; Kim et al. 2010; Kim et al. 2011; Klisch et al. 2011; Koeppel et

Figure 10.1 ***(preceding page)*****: Sequencing library characteristics.** (A) Joint distribution of library complexity and sequencing depth for all datasets examined. Vertical lines are drawn at 1 million, 5 million and 12 million reads. Horizontal and vertical lines indicate quality classes discussed in the text. The upper right domain (number of uniquely mappable reads $\geq$12 million and library complexity $\geq$0.8) passes current quality thresholds. (B) Distribution of library complexity for ChIP-seq datasets, IgG controls and Inputs; (C) Distribution of sequencing depth for ChIP-seq datasets, IgG controls and sonicated inputs; (D) Fraction of ChIP-seq, IgG and Input datasets exhibiting high, medium and low complexity; (E) Fraction of studies containing libraries of high, medium and low complexity (the distribution of the minimum library complexity observed is shown)

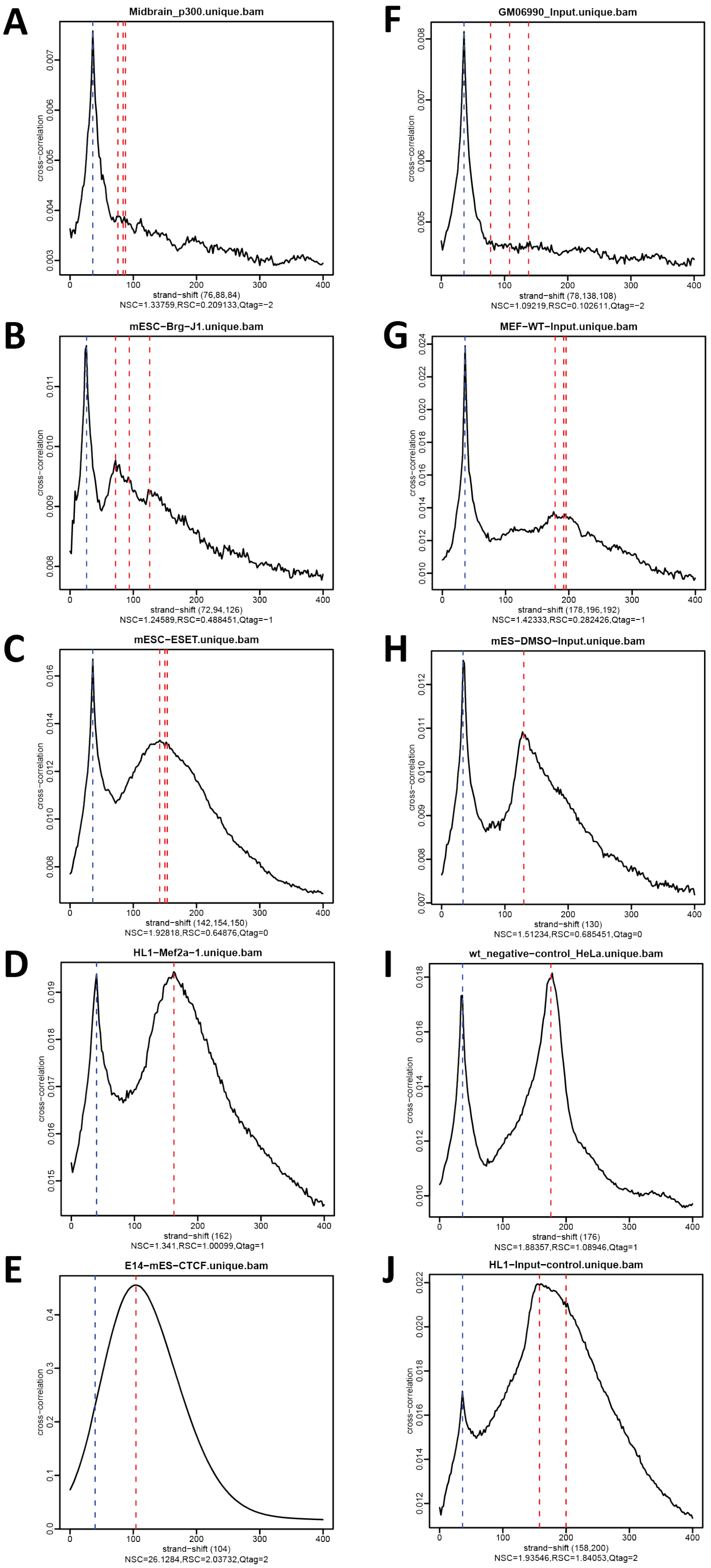


Figure 10.2: Examples of cross-correlation plots and QC score assignments for both ChIP-seq and control datasets. Successful ChIP-seq is expected to show a very high cross-correlation peak relative to the read length ”phantom peak”. Failed ChIP-seq experiments lack such a peak. Control libraries (sonicated inputs or IgG input) are also expected to lack this peak; the presence of a high cross-correlation peak is most likely due to a very strong Sono-seq effect (Auerbach et al. 2009). (A). Example of a ChIP-seq dataset with QC score of -2 (from Visel et al. 2009; Gotea et al. 2010; Blow et al. 2010). (B). Example of a ChIP-seq dataset with QC score of -1 (from Ho et al. 2009). (C). Example of a ChIP-seq dataset with QC score of 0 (from Yuan et al. 2009). (D). Example of a dataset with QC score of 1 (from He et al. 2011). (E). Example of a ChIP-seq dataset with QC score of 2 (from Handoko et al. 2011). (F). Example of a control dataset with QC score of -2 (from Lee et al. 2010). (G). Example of a control dataset with QC score of -1 (from GSE15844). (H). Example of a control dataset with QC score of 0 (from GSE23581). (I). Example of a dataset with QC score of 1 (from Vermeulen et al. 2010). (J). Example of a control dataset with QC score of 2 (from He et al. 2011).

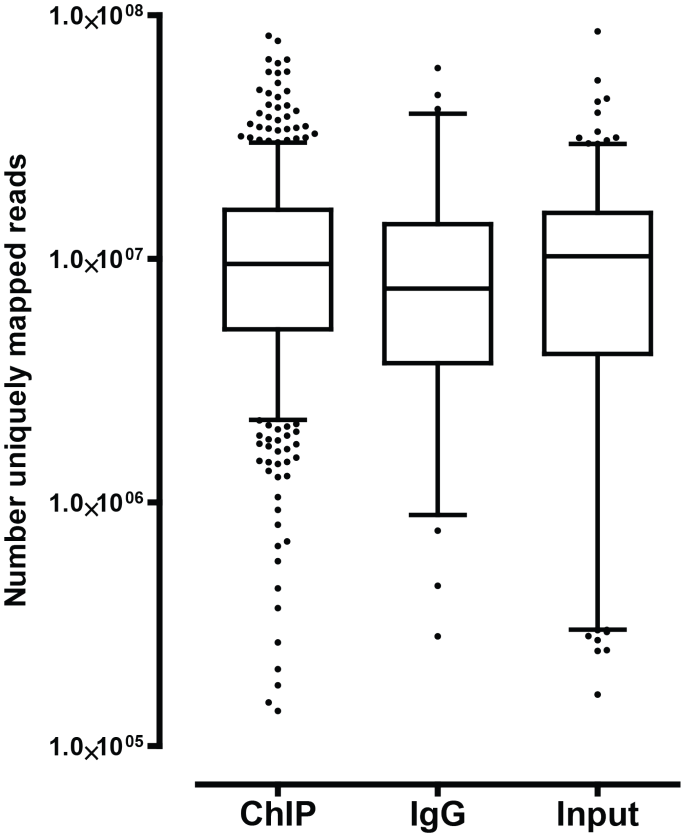


Figure 10.3: Sequencing depth distribution for ChIP-seq and IgG and Input control datasets.

al. 2011; Kong et al. 2011; Kouwenhoven et al. 2010; Krebs et al. 2010; Kunarso et al. 2010; Kwon et al. 2010; Law et al. 2010; Lee et al. 2010; Lefterova et al. 2010; Li et al. 2010; Li et al. 2012; Lin et al. 2010; Lister et al. 2009; Little et al. 2011; Liu et al. 2010; Liu et al. 2011; Lo et al. 2011; Lu et al. 2012; Ma et al. 2010; MacIsaac et al. 2010; Mahony et al. 2010; Marban et al. 2011; Marson et al. 2008; Martinez et al. 2010; Mazzoni et al. 2011; McManus et al. 2011; Mendoza-Parra et al. 2011; Meyer et al. 2012; Miller et al. 2012; Miyazaki et al. 2011; Mullen et al. 2011; Mullican et al. 2011; Nakayamada et al. 2011; Nishiyama et al. 2009; Nitzsche et al. 2011; Norton et al. 2011; Novershtern et al. 2011; Ntziachristos et al. 2012; Palii et al. 2010; Pehkonen et al. 2012; Ptasinska et al. 2012; Qi et al. 2010; Quenneville et al. 2011; Rada-Iglesias et al. 2010; Rahl et al. 2010; Ramagopalan et al. 2010; Ramos et al. 2010; Rao et al. 2011; Remeseiro et al. 2012; Rey et al. 2011; Robertson et al. 2007; Sadasivam et al. 2012; Sahu et al. 2011; Sakabe et al. 2012; Schödel et al. 2012; Schlesinger et al. 2010; Schmitz et al. 2011; Schnetz et al. 2010; Sehat et al. 2010; Seitz et al. 2011; Shen et al. 2011; Shukla et al. 2011; Siersbæk et al. 2011; Smeenk et al. 2011; Smith et al. 2011; Soccio et al. 2011; Stadler et al. 2011; Steger et al. 2010; Sun et al. 2011; Tallack et al. 2010; Tan et al. 2011a; Tan et al. 2011b; Tang et al. 2010; Teo et al. 2011; Tijssen et al. 2011; Tiwari et al. 2011a; Tiwari et al. 2011b; Trompouki et al. 2011; Trowbridge et al. 2012; van Heeringen et al. 2011; Vermeulen et al. 2010; Verzi et al. 2010; Verzi et al. 2011; Vilagos et al. 2012; Visel et al. 2009; Vivar et al. 2010; Wang et al. 2011a; Wang et al. 2011b; Wei et al. 2010; Wei et al. 2011; Welboren et al. 2009; Whyte et al. 2011; Wilson et al. 2009; Woodfield et al. 2010; Wu et al. 2011a; Wu et al. 2011b; Wu et al. 2012; Xiao et al. 2012; Xu et al. 2011; Yang et al. 2010; Yang et al. 2011; Yao et al. 2010; Yildirim et al. 2011; Yoon et al. 2011; Yu et al. 2009; Yu et al. 2010; Yu et al. 2012;; Yuan et al. 2009; Zhang et al. 2011; Zhao et al. 2011a; Zhao et al. 2011b; unpublished at the time of completion of the manuscript GEO accession numbers: GSE33346, GSE33850, GSE36561, GSE30919, GSE33128, GSE35109, GSE25426, GSE31951, GSE26711, GSE23581,

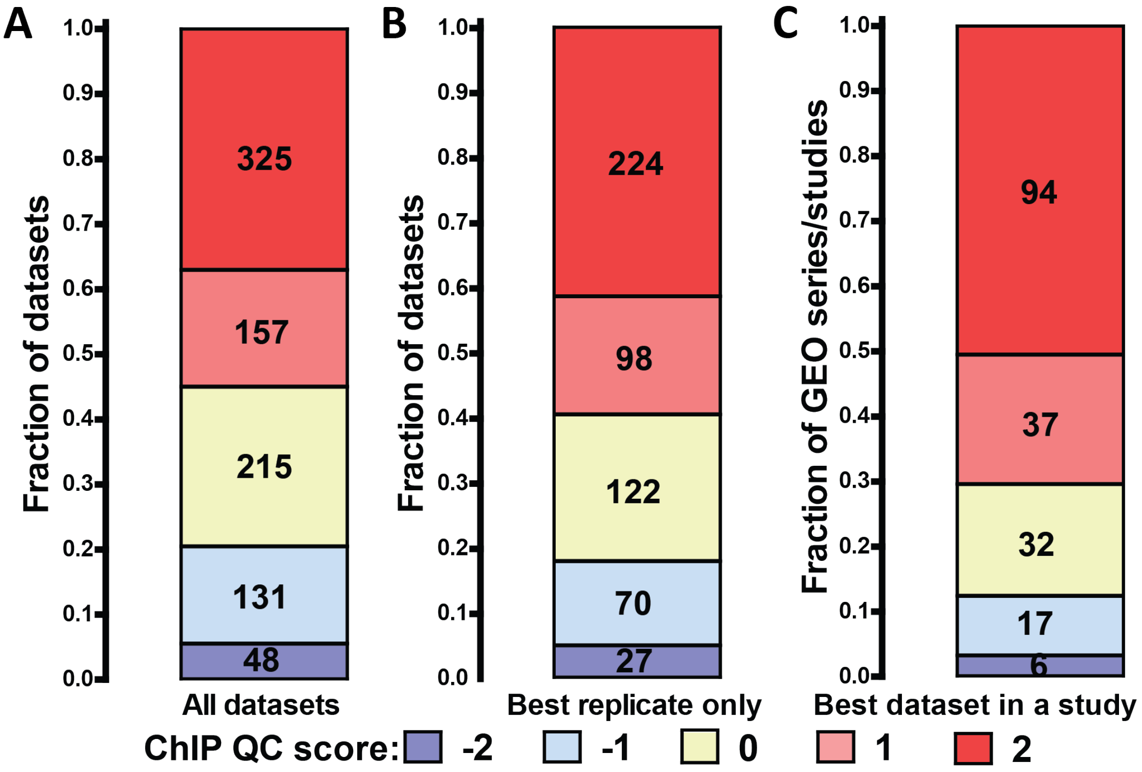


Figure 10.4: ChIP QC assessment summary. The numbers in each box indicate the total number of datasets/studies belonging to it. SPP QC scores of +1 and +2 indicate a high degree of read clustering in a dataset. (A) Distribution of SPP QC scores for all ChIP-seq datasets examined; (B) Distribution of SPP QC scores for the best replicates for a factor/condition combination in each study; (C) Distribution of the maximum SPP QC scores for all ChIP-seq datasets in a study.

GSE26136, GSE26680, GSE15844, GSE21916, GSE22303 and GSE29180; direct links to all GEO series can be found in Supplementary Table 1).

I discuss IgG and Input controls separately as, to the best of my knowledge, any potential general differences between the two types of controls have not been investigated systematically in the context of ChIP-seq (Peng et al. 2007 addressed these questions for ChIP-Chip data, however, the nature of the background is substantially different for microarrays)

I mapped all reads with uniform settings (see the Methods section for details) and examined library and ChIP quality control metrics for each dataset. These criteria have already been discussed in Landt et al. 2012, and a detailed treatment of cross-correlation is presented elsewhere (Kundaje et al., submitted). Here we provide a brief overview of each.

1. **Sequencing depth.** If a ChIP-Seq experiment achieves successful immune-enrichment and the resulting llibrary adequately represents the sample, greater sequencing depth will produce a more complete map of transcription factor occupancy (Landt et al. 2012). At greater depth, the measurement will identify a larger number of reproducible sites containing the corresponding DNA binding sequence motif. Undersequencing of an otherwise successful library will lead to false negatives. It has been difficult to establish a universal minimal sequencing depth due to differences between factors. Any threshold is going to be somewhat arbitrary, but in general, the major cost/benefit tradeoff is between sequencing one sample more deeply and generating additional replicates: for most contemporary purposes, an independent duplicate measurement of 12 million reads arguably adds greater overall

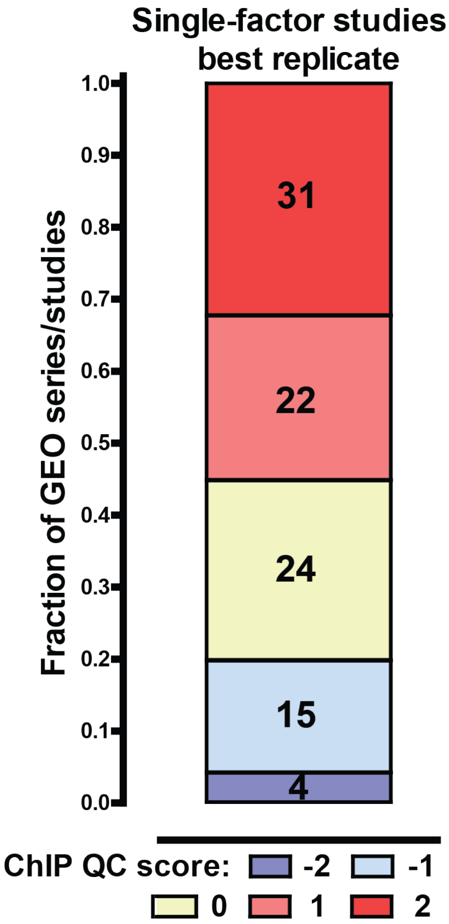


Figure 10.5: Distribution of the maximum SPP QC scores for studies in which only a single transcription factor was assayed.

value than a single determination with 24 million reads, even though the higher number of reads will increase sensitivity. Numbers of mapped reads below 1-2 million for a typical transcription factor, will usually be inadequate for capturing the complexity of an interactome for a mammalian-sized genome. Many datasets now in the public domain were generated when sequencing throughput was lower than it is now and costs were higher (between 2007 and 2013, sequencing throughput has increased by about two orders of magnitude). As a consequence, many early ChIP-seq libraries were sequenced to a depth of only a few million reads. I therefore divided datasets into sequencing bins by using thresholds of 1,5,12 and 24 million uniquely mapped reads (taking into account sequencing depths recommended in the past by the ENCODE consortium for transcription factors). Libraries having less than a million reads are considered severely undersequenced, and those with above 12 million reasonably deeply sequenced.

2. **Library complexity**. A second characteristic that influences the quality of a ChIP-seq measurement is the sequence fragment diversity of the sequencing library. This is generally referred to as library complexity and low complexity is undesirable with current technology, though I note that much better IP enrichment than what is now obtained could in the future lead to high-quality datasets with low library complexity. Currently, low-complexity libraries mainly result from experimental deficiencies: either too few starting molecules at the end of the immunoprecipitation step or inefficient steps in subsequent library building. As a result, the same starting molecules are sequenced repeatedly. Very low-complexity

libraries will not contain enough information to effectively sample the true positive binding sites and they distort the signal position and intensity. This can confuse peak callers (especially if the algorithm does not collapse presumptive PCR duplicates), leading to peak calling artifacts (Landt et al. 2012). We use the following metric as an indicator of library complexity (Landt et al. 2012):

$$\text{Library complexity} = \frac{\text{Number positions in the genome that uniquely mappable reads map to}}{\text{Number uniquely mappable reads}} \tag{10.1}$$

- Estimated in this way, library complexity is expected to decrease eventually with increased sequencing depth because even highly complex libraries become exhausted by very deep sequencing. Reduced apparent complexity would also be observed with extremely successful ChIP-seq experiments for transcription factors that bind to the genome in a highly discriminative fashion and to a limited number of locations. In such libraries, the majority of reads would originate from the limited genomic subspace around binding sites resulting in low apparent library complexity. With current methods, this is, a largely theoretical consideration; in practice, in most ChIP-seq libraries only a minority of reads originate from factor-bound sites, with the rest (the majority) representing genomic background. As the vast majority of libraries examined fell in the sequencing depth range over which these values represent library complexity reasonably well (Figure 10.1A, Table 10.1), I split datasets in the following complexity groups: high complexity (apparent library complexity ≥ 0.8), medium to low complexity (apparent library complexity between 0.5 and 0.8), and very low complexity (apparent library complexity ≤ 0.5). Finally, I note that in substantially smaller genomes the apparent library complexity is expected to be lower as the number of positions from which sequencing library fragments can originate is smaller.

3. **Cross-correlation analysis of read clustering and ChIP enrichment.** Since the majority of sequencing reads in a ChIP-seq library represents non-specific genomic background, these reads are expected to be randomly distributed over the genome. In contrast, reads originating from specific occupancy events cluster around the sites of protein-DNA interactions, where they are distributed in characteristic asymmetric pattern on the plus and minus strand (Kharchenko et al. 2008). Cross-correlation analysis is an effective way of measuring the extent of this clustering. It also captures additional global features of the data such as the average fragment length and fragment length distribution (Kharchenko et al. 2008; Landt et al. 2012). Specifically, the read coverage profiles on the two strands are shifted relative to the other over a range of shift values and the correlation between the profiles is calculated at each shift (Kharchenko et al. 2008). The resulting plot has one ("phantom") peak corresponding to the read length and another peak corresponding to the average fragment length; the height of the fragment-length peak is highly informative of the extent of read clustering in the library and in turn of the success of a ChIP-seq experiment. This feature is best captured by the normalized and relative strand correlation (NSC and RSC) metrics discussed in (Landt et al. 2012). I applied `SPP` (Kharchenko et al. 2008) to carry out cross-correlation analysis for all libraries in this survey. I then used the RSC cross-correlation metric to assign integer quality control tag values in the $\{-2, 2\}$ range to datasets, with QC values of 2 corresponding to very highly clustered (and mosty likely, also successful) datasets and QC values of -2 to datasets exhibiting no to minimal read clustering; negative values are expected for input datasets.

> The RSC metric captures well the extent of read enrichment in vertebrate genomes similar in size to humans, which this study focuses on. I provide representative examples of cross-correlation plots for each of the five QC categories in Figure 10.2A) and use these tags as convenient general proxies for ChIP quality throughout the following analysis. I note that the discretization thresholds are not meant to be absolute determinants of quality, but they enable one to rapidly scan very large numbers of datasets. In practice, examining the cross-correlation plots and the continuously-distributed NSC and RSC values is always more informative and provides more nuance in understanding specific datasets.

An additional major component of the ChIP-seq quality control pipeline developed by the ENCODE consortium is reproducibility analysis of replicates, based on the irreproducible discovery rate (IDR) statistic (Li et al. 2011). However, since many of the studies surveyed did not have replicates, I only evaluated datasets on the level of individual experiments. Single dataset evaluation is also almost always a valuable precursor to evaluation of replicates, as typically a second replicate is generated following a successful first one.

The full list of datasets, mapping and quality control statistics is provided in Table 10.1.

10.2.2 Sequencing depth and library complexity

Figure 10.1A shows the distribution of sequencing depth and library complexity for ChIP-seq and control datasets. The upper right domain, bounded by 12 million reads per sample and a complexity value of 0.8 is an arbitrary but useful definition of high quality according to these measures. A majority of datasets had reasonably good complexity and severely undersequenced libraries were rare (Figure 10.1C). A minority (38.8%) of datasets had more than 12 million mapped reads; however, as discussed above, this is not unexpected, as a large fraction of the datasets we surveyed were generated in times of significantly higher sequencing cost and lower throughput. Strikingly, the median complexity of IgG control datasets was below 0.8 and considerably lower than that of either ChIP-seq or sonicated Input libraries (Figure 10.1B). This is not a result of IgG datasets having been sequenced much more deeply than the other two groups; in fact the median sequencing depth of IgG controls is lower (Figure 10.3). The concern that individual IgG inputs might provide insufficient DNA mass to build highly complex libraries has been raised before (Landt et al. 2012) and our observations are consistent with this concern, although it is not a uniform problem for all IgG controls.

Slightly more than half (54.3%) of ChIP-seq datasets had library complexity higher than 0.8 while very low-complexity (< 0.5) libraries comprised 12.9% of datasets; the fraction of very low-complexity libraries was higher and lower for IgG and Input datasets, respectively (Figure 10.1D). As most GEO series contained multiple libraries, I also asked how common is the presence of low-complexity libraries in individual studies. Figure 10.3E shows the distribution of the minimum library complexity in each such series (for all types of datasets). A quarter (25.4%) of all studies contained very low-complexity libraries.

10.2.3 Cross-correlation quality assessment of ChIP-seq datasets

Next, I examined the distribution of `SPP` QC scores for ChIP-seq datasets. Before doing this, I excluded a minority of datasets for which there was a good reason to think high ChIP enrichment should not be expected. For example, experiments executed in knock-outs, knock-downs, or settings, in which the factor is not expressed, are not expected to produce a high-scoring measurement. And in a few cases, the factor in question might be known to bind to only a small number of sites in the genome; this has been proposed, for example, for some ZNF transcription factors and Pol3 and its associated factors (Landt et al., 2012). The detailed criteria for inclusion are described in the Methods section.

Figure 10.4A shows the QC score distribution for all ChIP-seq datasets we retained. Strikingly, only 55% (482 out of 876) of datasets had QC scores of 1 or 2, i.e. they are likely to be highly successful. Additional 24.5% (215 out of 876) had a score of 0, marking them as of intermediate quality, and 20.4% (179 out of 876) had low-quality scores of -1 and -2. Sometimes multiple replicates for a factor were submitted but only one fails, so I also compiled a second set of ChIP-seq experiments that only included the best available replicate for each factor and

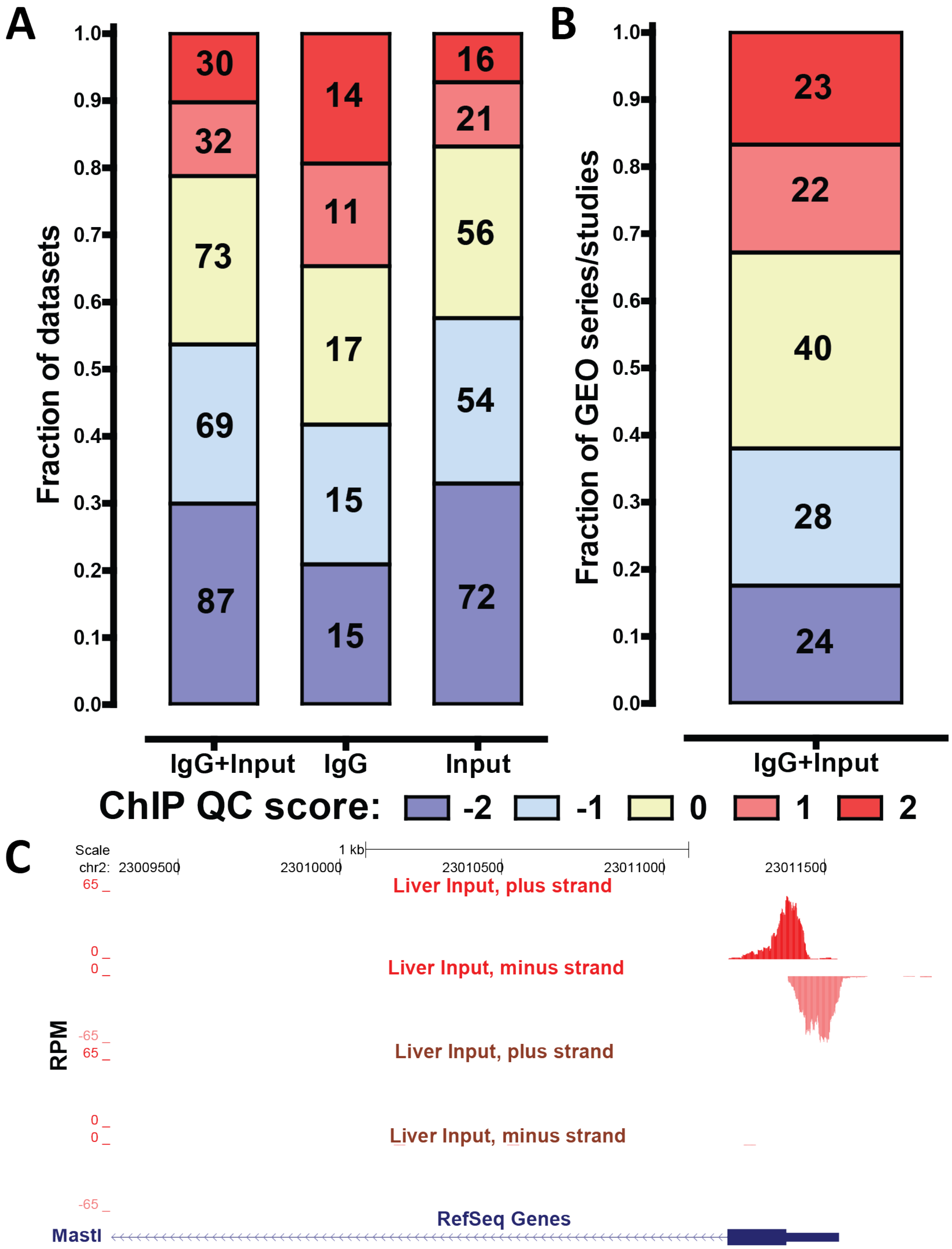


Figure 10.6: Assessment of read clustering in control datasets. The numbers in each box indicate the total number of datasets/studies belonging to it. SPP QC scores of 1 and 2 indicate a high degree of read clustering in a dataset. (A) Distribution of SPP QC scores for all control datasets (IgG+Input), IgG/mock IP controls (IgG) and sonicated inputs (Inputs); (B) Fraction of studies containing highly clustered inputs. The distribution of the maximum SPP QC score for all inputs in a dataset is shown. (C) Examples of a highly clustered input (mouse liver, upper two tracks, from MacIsaac et al. 2010, QC score of 2) and an input that does not show high extent of read clustering (also mouse liver, lower two tracks, from Soccio et al. 2011, QC score of -1). The promoter of the *MASTL* gene is shown. All tracks are shown to the same scale and reads mapping to the plus and minus strands and displayed separately for better visualization of the cross-correlation between the two.

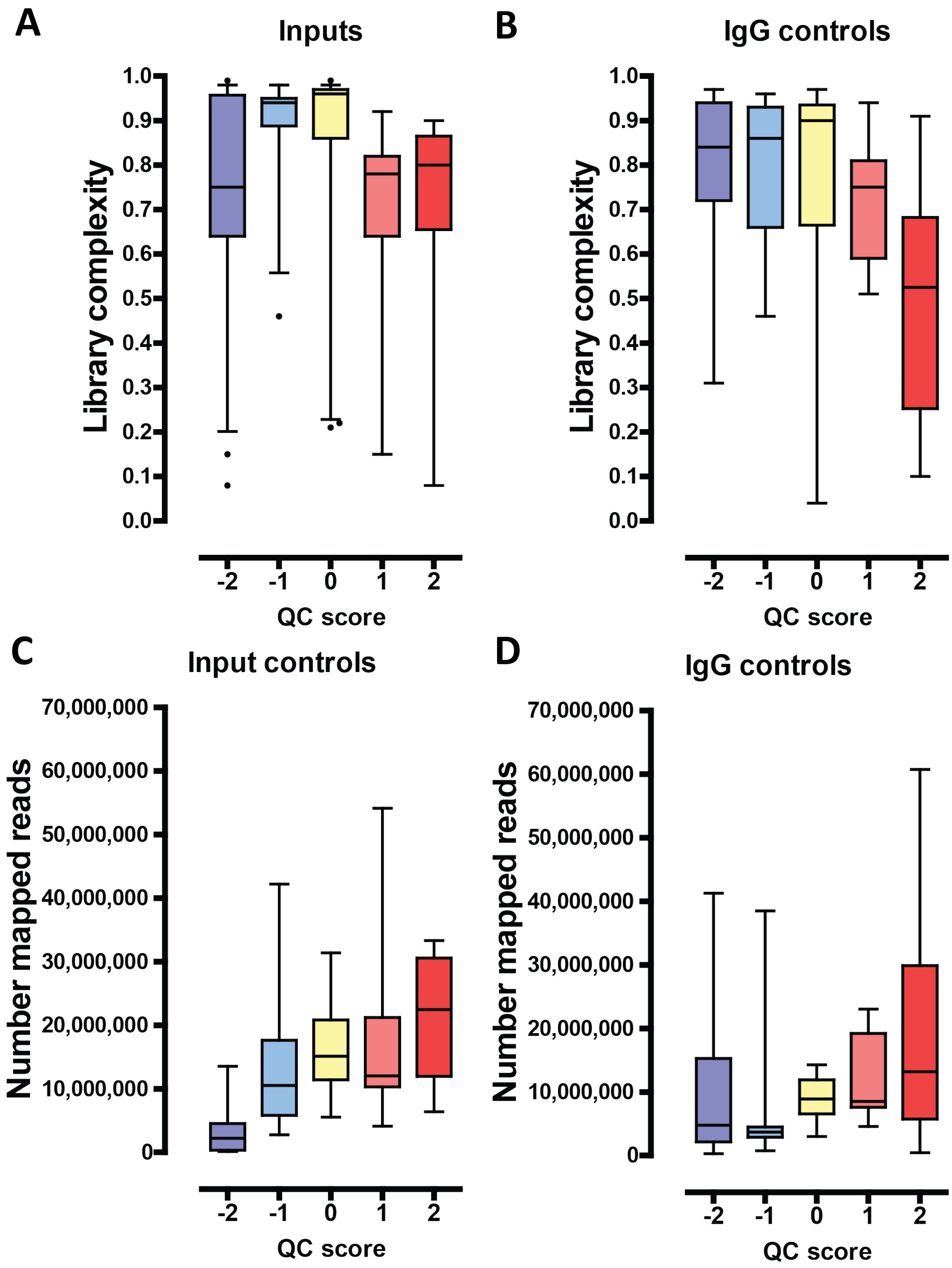


Figure 10.7: Distribution of library complexity values and sequencing depth for Input and IgG control datasets divided by QC scores. (A,B) Library complexity. (C,D) Sequencing depth.

condition (Figure 10.4B). This set includes 322 datasets (59%) with QC scores of 2 or 1. The fraction of intermediate-quality and failing datasets in this set decreased as expected; however the decrease was relatively small - 18% (97 out of 541) of the best available replicates still had scores of -1 or -2, and 22.5% (122 out of 541) had a score of 0.

I then examined the distribution of the maximum QC score for each study, regardless of which target it was for (Figure 10.4C). The fraction of failing scores decreased further, yet still only 70.4% of studies (131 out of 186) had a score of 1 or 2 for their best experiment. I also compiled a list of the best datasets from all studies that only assayed a single transcription factor; 19.7% (19 out of 96) such studies had scores of -1 or -2, 25% (24 of 96) had a score of 0, and 55.2% (53 of 96) were marked as likely to be successful with scores of 1 and 2 (Figure 10.5).

10.2.4 Read clustering in control datasets

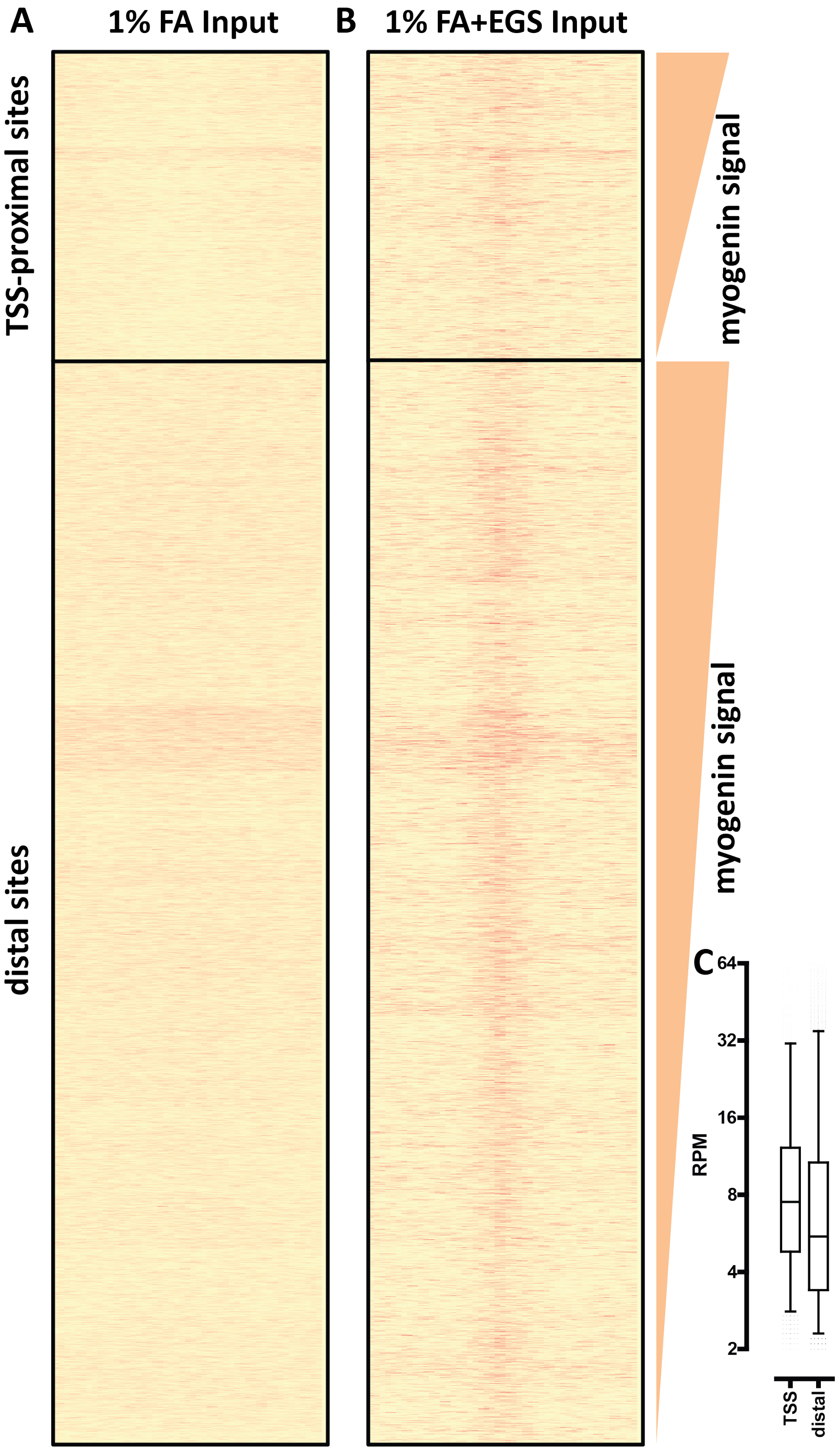
Control datasets serve the important purpose of helping to distinguish read enrichment due to the immunoprecipitation step from artifactual read clustering due to other experimental factors, both known and unknown. It is, for example, well appreciated that differential chromatin shearing efficiency can lead to the overrepresentation of areas of open chromatin (usually immediately surrounding transcribed promoters) in sequencing libraries. This is termed "Sono-seq" effect when attributed to sonication (Auerbach et al. 2009). In addition, unknown copy number variants and sequence composition biases may give false positive putative occupancy. In particular, specifics of the amplification step in most sequencing platforms can introduce significant bias for GC content (Ho et al. 2011).

In general, control datasets are not expected to exhibit a pattern of significant read clustering similar in strength to that of successful ChIP-seq datasets. In our own practice, under standard crosslinking protocols, most do not. However, we have noticed that a minority of control datasets show positive ChIP QC scores along with prominent cross-correlation peaks. Figure 10.2B shows examples of cross-correlation plots for individual control datasets with all possible QC scores, from -2 to 2 and Figure 10.6C shows a browser snapshot of a region with strong read enrichment in a highly clustered (QC score of 2) input library and no such enrichment in a library from a similar biological source with a QC score of -1.

I asked how general this phenomenon is by examining the distribution of QC scores of both IgG and Input control datasets (Figure 10.6A). Surprisingly, only 53.6% (156 out of 291) of control datasets had QC scores of -2 or -1 and 25% (73 of 291) had a score of 0, while 21.3% (62 of 291) exhibited very high degree of read clustering and received scores of 1 or 2. The highly clustered inputs were notably more common among IgG controls than among Input chromatin controls (Figure 10.6A). Moreover, high read clustering was more often found in low-complexity libraries (which are themselves more common among IgG controls) (Figure 10.7A and 10.7B).

I also examined how widespread input clustering is on the level of GEO series/studies to see if the phenomenon is restricted to a few larger studies. Figure 10.6B shows the distribution of the maximal QC score for all control datasets in a study. Of the studies for which control datasets were available, 32.8% (45 of 123) contained at least one highly clustered control with a score of 1 or 2 and 29.2% (40 of 123) contained a control with a score of 0. Thus control datasets surprisingly often exhibit a high extent of read clustering similar to that of ChIP-seq datasets. This is even more striking considering that FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) data (an assay that is based on the preferential enrichment of open chromatin in sonicated DNA and aims at achieving high read clustering) from ENCODE usually has QC scores between -2 and 0, and that the Sono-seq datasets published by Auerbach et al. all have scores of -2.

I note that unless this effect is very strong and is associated with notable genomic features such as promoters of genes, it can be difficult to detect by the usual methods of visual inspection of signal tracks on a genome browser. It is, however, readily apparent in cross-correlation analysis and these results raise awareness of its existence. As mentioned above, one candidate explanation for this phenomenon is the previously described "Sono-seq" effect. Using standard experimental protocols this effect has been rare in our experience, but under more aggressive crosslinking conditions, we have observed increased read clustering (Figure 10.8). Notably, the original "Sono-seq" description focused on promoter regions, but we have also observed it over distal regulatory elements, where its strength was even


A
1% FA Input
B
1% FA+EGS Input
TSS-proximal sites
distal sites
myogenin signal
myogenin signal
C
64
32
16
8
4
2
RPM
TSS
distal

higher than at promoters (Figure 10.8). Thus variation in the extent of fixation, as well as sonication, might be a substantial contributor to variation in read clustering across the broader data collection. Another potential contributing factor is sequencing depth – "Sono-seq" effect of the same magnitude can result in a more prominent cross-correlation profile with increasing sequencing depth as more and more reads can be found in proximity to each other (the same correlation is observed in ChIP-seq datasets). The average sequencing depth for highly clustered IgG and Input controls is higher than that of controls with negative QC scores (Figure 10.7C and 10.7D); however, this by no means explains all the clustering observed in controls as there are plenty of examples of deeply sequenced Input and IgG libraries with no significant cross-correlation peaks. Finally, "Sono-seq" need not be the only explanation. Other, not yet identified, causes may be behind the phenomenon, and the cause might not be the same in all cases. Indeed, while a number of control datasets with QC scores of 2 exhibited higher read coverage around promoters, others did not (Figure 10.9), suggesting at least one other source of read enrichment over regions located elsewhere in the genome. As rich annotation of functional genomic elements outside of promoter regions is not available for many cell types in our study, this phenomenon is a subject for future studies.

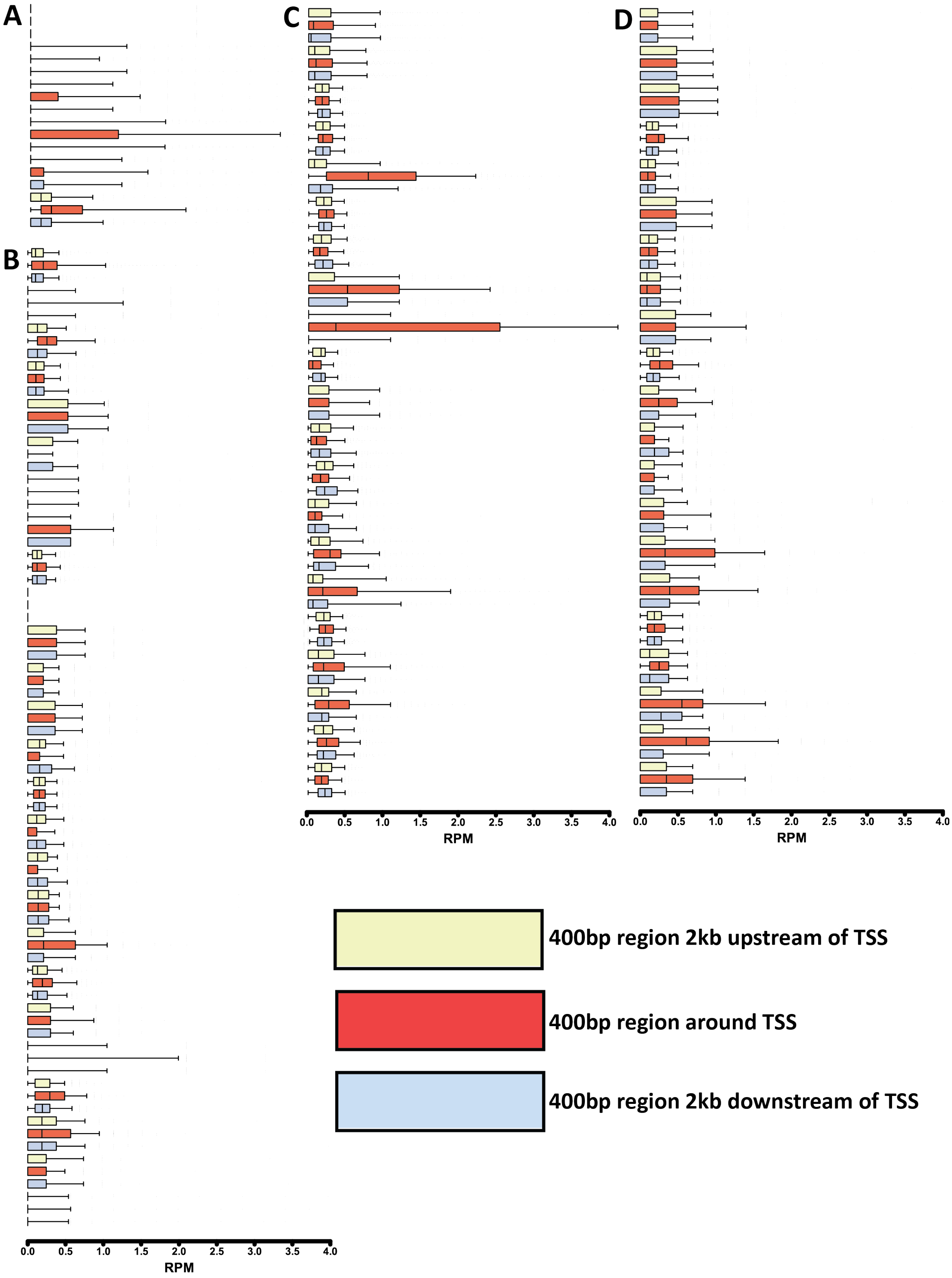
10.3 Discussion

In this study I carried out a systematic survey of ChIP quality for publicly available vertebrate ChIP-seq datasets. Over half of these datasets were found by our measures to be of high quality. This group comprises a set that can be used with confidence for integrative analyses. This conclusion carries the important caveat that I did not assess the specificity of the immune reagents used to carry out the experiments, which is obviously a critical concern of a different kind.

A substantial minority of published datasets (between 20% and 45% of those examined) were of low or intermediate quality by our metrics. This was true not only for individual libraries, but was also true when only the best replicate from each study was examined. In addition, I observed a substantial number of low-complexity datasets and an unexpected group of highly clustered control datasets. These observations underscore the widespread variability in ChIP-seq data. They also raised questions about which kinds of conclusions in primary publications are more or less sensitive to data quality. Global quality analysis is especially useful to guide subsequent re-uses of published data that require higher quality than was needed or achieved in the source study.

Dataset quality issues appeared in publications across impact levels. I separated datasets into groups according to the 2011 Thomson

Figure 10.8 *(preceding page)*: Relation between a well defined set of promoter-proximal and promoter-distal transcription factor binding sites and input datasets with minimal and significant read clustering. The high-quality C2C12 myogenin dataset shown in Figure 4 was used, ERANGE3.2 binding sites were separated into promoter promoter-proximal (sites for which the peak position, defined by the peak caller was within 1kb of a TSS present in the ENSEMBL63 annotation of the mm9 version of the mouse genome) and promoter-distal (sites for which the peak position was more than 1kb away from TSSs) groups, each group was ranked by decreasing myogenin signal and the distribution of input signal was plotted for the 1kb region around the peak position. (A) A C2C12 input dataset generated from cells fixed with the usually used 1% concentration of formaldehyde (FA) for 15 minutes, and showing little read clustering genome-wide (QC score of -1). (B) A C2C12 input dataset generated from cells fixed with a combination of 1% formaldehyde (for 10 minutes) and subsequent additional fixation with the long-arm crosslinker ethylene glycol-bis(succinimidylsuccinate) (EGS) (Abdella et al. 1979) in order to enhance crosslinking between proteins and capture the interactions of factors more loosely associated with chromatin (Zeng et al. 2006). There are reason to expect that such more aggressive crosslinking conditions will results in a stronger Sono-seq effect and indeed this dataset exhibits significant amount of read clustering (QC score of 2). The 1%FA+EGS input signal around myogenin binding sites is considerably higher than the 1% FA input signal. Notably, the 1%FA+EGS signal signal is stronger for promoter-distal sites than it is for promoter-proximal sites even though promoter-proximal sites are generally stronger (C).


A
B
C
D
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
RPM
400bp region 2kb upstream of TSS
400bp region around TSS
400bp region 2kb downstream of TSS

Reuters Impact Factor for the journal in which the corresponding article was published, and examined the distribution of QC scores in each group (Figure 10.11). The group with highest impact factor (≥ 25) contained the largest fraction of datasets with a QC score of -2 or -1. I also examined the distribution of QC scores with respect to the year of publication (Figure 10.10). A higher proportion of low-quality datasets were seen in earlier publications (2008-2009), although this might be due to the smaller number of datasets from early years. It is encouraging that the fraction has stabilized in the last three years at around 20% (Figure 10.10).

It is important to recognize that datasets scoring as poor quality by the metrics used here can, nevertheless, make important biological discoveries. For this reason, it would be an error to set a fixed "standard" that every published dataset of the future would have to meet. Instead, routine QC analysis would make it easy to see when there is reason for concern about a given dataset. It would also provide a first tier of uniform guidance about what uses are likely to be appropriate for a given dataset. As discussed previously, the appropriate level of quality control stringency depends on the specific goals of the experiment and methods of analysis (Landt et al. 2012). In particular, some analyses that are sensitive to false negatives are particularly vulnerable to inclusion of low-scoring datasets. For example, trying to derive combinatorial transcription factor occupancy rules is seriously compromised and even misleading, if a subset of the datasets included are suboptimal.

I illustrate this with a simple example from our own practice in Figure 10.12. The MyoD and myogenin transcription factors are well known regulators of muscle differentiation (Yun & Wold 1996) and C2C12 cells (Yaffe & Saxel 1977) have been widely used to study the process as they can be propagated in an undifferentiated myoblast state and easily induced to differentiate into myocytes and myotubes. We have done several ChIP-seq experiments with these factors in differentiated and undifferentiated C2C12 cells (G. DeSalvo et al., in prep.; A. Kirilusha et al., in prep., K. Fisher-Aylor et al., in prep.), some of which have been highly successful, while others were of poor or intermediate quality. Here, I examined the effect of weaker ChIP-seq datasets on combinatorial occupancy analysis, using a MyoD ChIP-seq dataset with very high QC metrics, and three myogenin datasets with very high, moderately good, and very low such metrics (Figure 10.12A). Using the best myogenin dataset, we find a high degree of overlap between the binding sites of the two factors (Figure 10.12B). When the medium-quality myogenin dataset is used instead, a sizable group of MyoD-only sites emerges (Figure 10.12C) and the erroneous conclusion that a substantial number of MyoD sites lack myogenin binding could be reached if this was the only dataset available for analysis. Finally, the poor-quality myogenin dataset contains very few called peaks and as a result almost all MyoD sites show no myogenin binding when it is used for analysis (Figure 10.12D).

Recently, IDR analysis of replicate datasets (Li et al 2011; ENCODE Project Consortium 2012; Landt et al. 2012) emerged as a robust method for deriving lists of reproducible occupancy sites from ChIP-seq datasets. IDR is based on differences in the consistency of ranking (usually by signal strength as measured by read enrichment or by statistical significance) for all identified peaks in a pair of ChIP-seq replicates. A virtue of this approach is that it allows a statistically robust and reproducible set of binding sites to be derived largely independent of thresholds and settings specific to a particular peak-calling algorithm. Ideally, IDR would be used in conjunction with the quality metrics used here (ENCODE Project Consortium 2012; Landt et al. 2012). However, replicate measurements do not exist for many of the datasets in this survey, so it was not part of the pipeline

Figure 10.9 ***(preceding page)*****: Distribution of signal around TSSs in control datasets.** Each group of three blue, red and yellow boxplots represents to one dataset, with blue corresponding to a region 2kb upstream of TSSs, red to the region immediately surrounding the TSS, and yellow to a region 2kb downstream of TSS. Datasets in which signal over TSSs is considerably higher than the signal over flanking regions imply a possible "Sono-seq" overrepresentation effect; this, however, is not evident (at least over TSSs) in all highly clustered datasets. (A) Human control datasets with a QC score of +2. (B) Human control datasets with a QC score of -2. (C) Mouse control datasets with a QC score of +2. (D) Mouse control datasets with a QC score of -2.

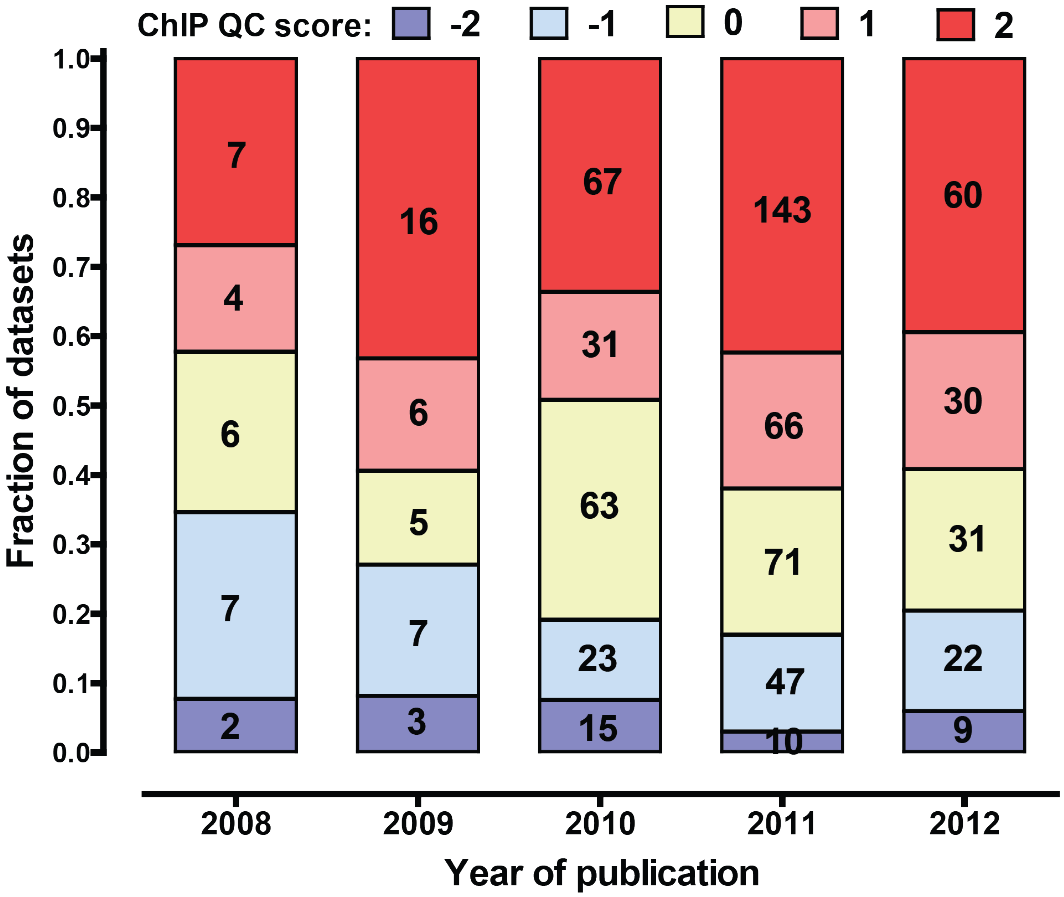


Figure 10.10: Distribution of dataset quality relative to year of publication.

in this study. IDR will likely become common practice, as sequencing costs drop. Even when that happens, measuring of the quality of individual datasets will remain important because IDR analysis is sensitive to the presence of poor-quality replicates. An asymmetric pair, consisting of one high-quality and one poor-quality dataset, is dominated in IDR by the weaker replicate, resulting in a shorter list of sites and a high false-negative rate. Care should be exercised in such cases. The best approach is to obtain a second high-quality replicate but if this is not possible, special strategies for treating asymmetric replicates have been devised (Landt et al. 2012).

The most perplexing observation made in this survey was that a subset of control datasets have extensive read clustering in the same range as successful ChIP-seq experiments. In our own practice, we have rarely encountered such libraries, and, to the best of my knowledge, there has been no extensive treatment of this issue or its influence on data analysis in the literature. The phenomenon occurred more frequently in IgG controls than in Input chromatin controls, although it is by no means limited to the former. In theory, an IgG control should be a superior representation of the true background noise in a ChIP-seq sample, as it incorporates biases introduced by the entire immunoprecipitation process, in addition to any enrichments or biases created by chromatin shearing. Following this logic, a simple interpretation is that high read clustering in these controls correctly identifies background that depends on something other than the factor-specificity of the antibody. However, I also observed a large number of IgG controls (Figure 10.6A) that show no such clustering, meaning that this is not a general feature. In addition, the overall lower complexity of IgG libraries introduces higher signal stochasticity, and that may offset the benefit of providing a better approximation of the immunoprecipita-

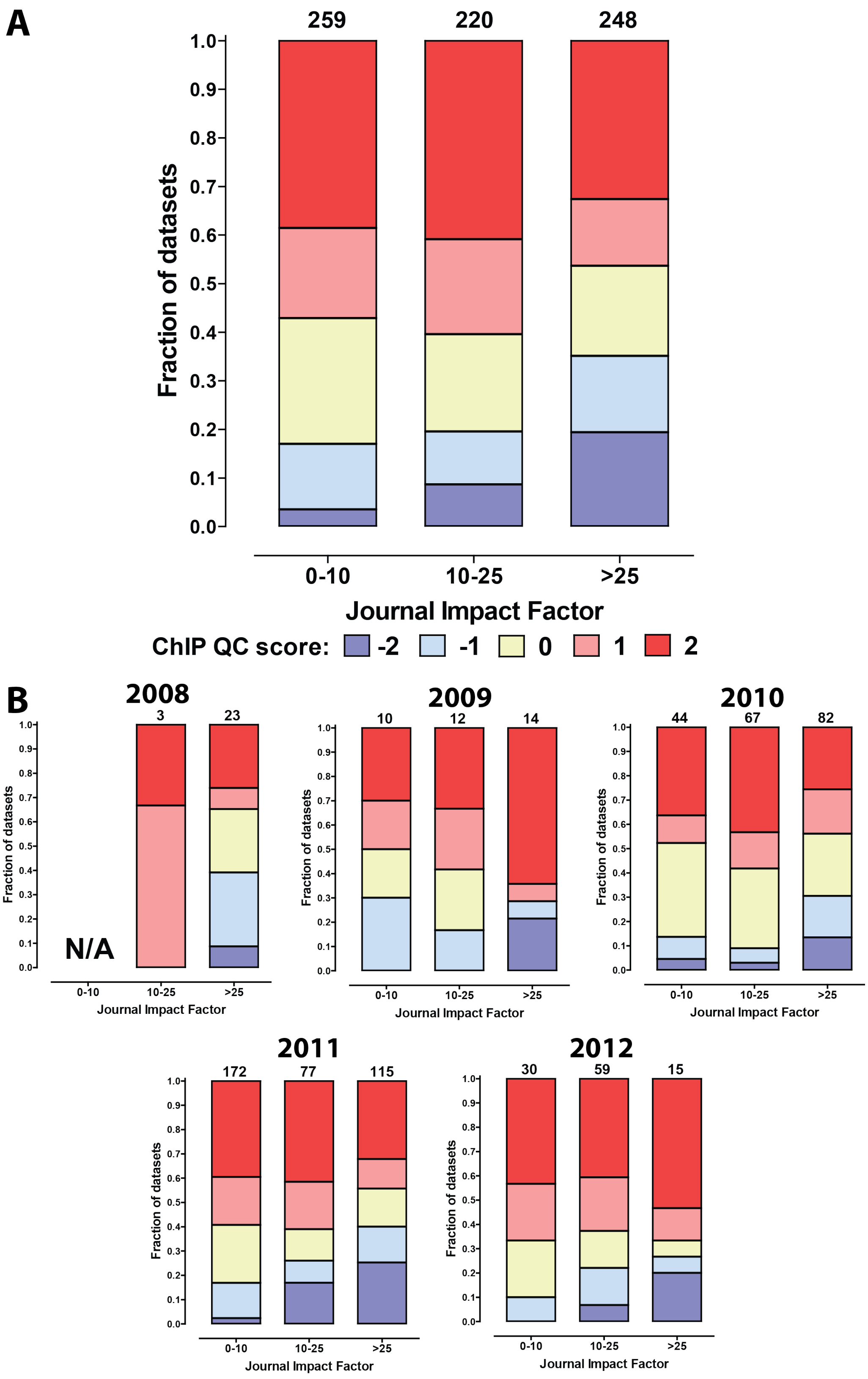


Figure 10.11: Distribution of dataset quality relative to the impact factor of the journal where an article was published. Shown are the 2011 Thompson-Reuters impact factor scores for the journals in which ChIP-seq datasets were published in. (A) All datasets. (B). Breakdown by year of publication.

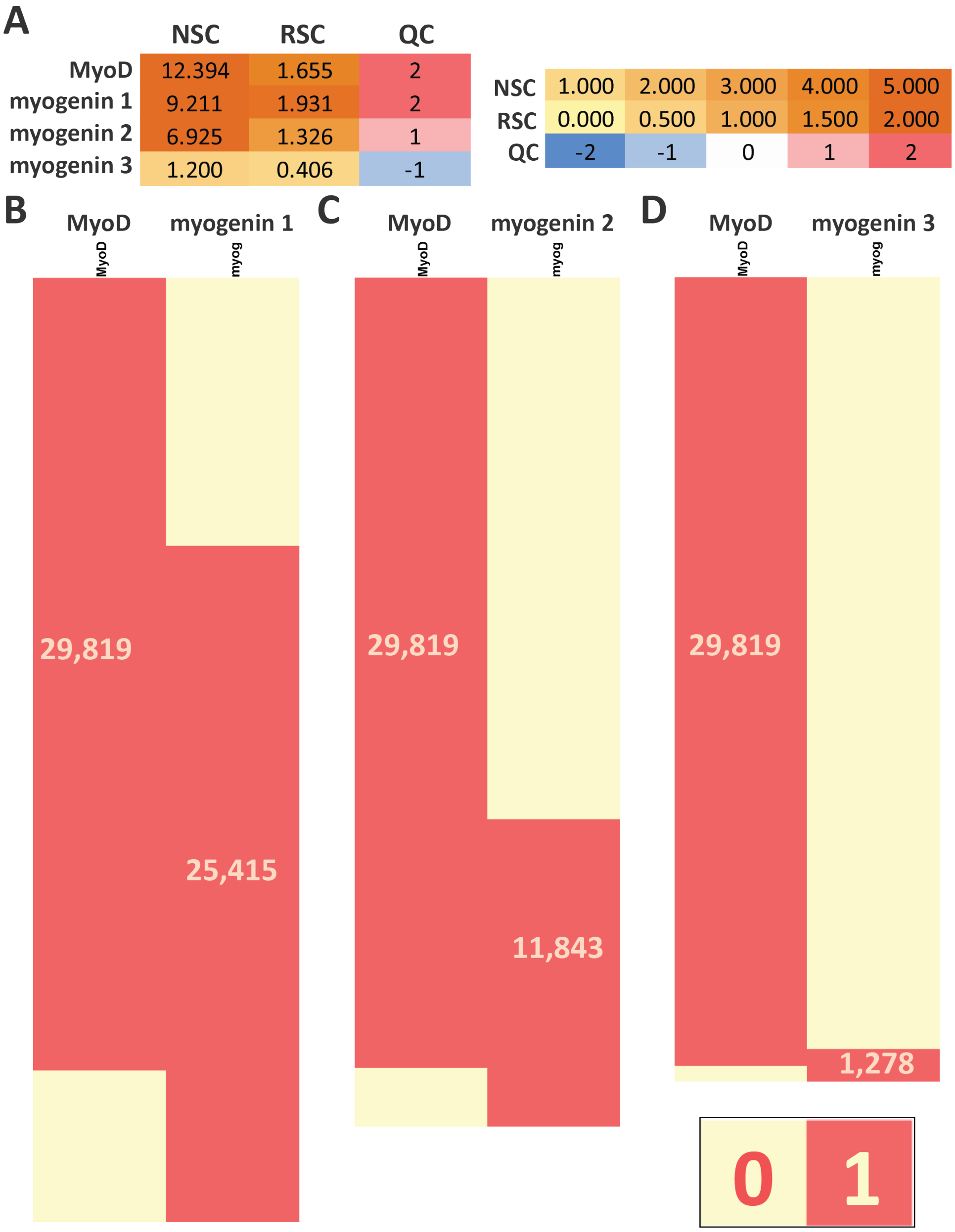


Figure 10.12: Effect of suboptimal datasets on combinatorial occupancy analysis. The muscle regulatory factors MyoD and myogenin were assayed in C2C12 myocytes at 60h after differentiation. Shown are a single, highly successful, MyoD ChIP-seq dataset and three myogenin ChIP-seq datasets, one of which is similarly highly successful ("myogenin 1"), a second, weaker one ("myogenin 2"), and a third, one which is an experimental failure ("myogenin 3"). (A) Quality control metrics. (B,C,D) The extent of overlap of MyoD and myogenin binding sites as determined using each of the three myogenin datasets (See Methods for data processing details). MyoD and myogenin are mostly found to bind to the same sites when interactome determinations of comparable strength are used (B). A sizable group of apparently MyoD-only sites emerges when the medium-strength myogenin dataset is used due to a large number of false negative myogenin calls (C). Finally, the unsuccessful myogenin ChIP reveals most MyoD sites to not be shared by myogenin (D). Numbers listed in the red blocks corresponding to the each set of peak calls indicates its size.

tion process.

A crucial issue is the extent to which clustering in controls is also present as experimental noise in ChIP libraries from the same material. For example, a very strong Sono-seq effect in a control sample is expected to give ChIP-seq libraries with high read clustering in a combination of true ChIP (antibody-specific) signal, plus non-specific promoter and enhancer Sono-seq noise. While most contemporary peak callers normalize for enrichment in controls, very strong background noise will diminish the signal-to-noise ratio and ultimately affect sensitivity. How much this affects the results will depend on the overlap between true factor occupancy sites assayed and the regions of artifactual read enrichment (for some factors this overlap may be negligible as they do not bind to such regions), on the magnitude of the Sono-seq effect, and on the strength of the ChIP itself (sufficiently strong determinations will not be affected greatly by this issue). Conversely, if a ChIP-seq library contains a strong Sono-seq enrichment component, but peak calling is done against a control sample in which the Sono-seq effect is of significantly lower magnitude, the rate of false positive peak calls is expected to increase. Unfortunately, in practice such cases can be difficult to detect especially when little is already known about the expected true positives. Similar reasoning applies if the noise source is something other than Sono-seq, and the same increased caution about reproducibility, sensitivity and attribution to the factor of interest will apply.

Uniform retrospective quality assessment is resource-intensive and will not be practically feasible as the number of ChIP-seq datasets is growing exponentially. Retrospective analysis also comes too late to influence the experiments themselves or to contribute to the review process. A reasonable path forward would be to incorporate routine quality assessment into experimental analysis, review for publication and submission to public repositories, as a matter of community practice. However, the results presented here also strongly caution against the blind application of our metrics or others, in the absence of experimental and biological context. We have seen that it is possible for good datasets to receive low QC scores in certain special situations. It is also possible for some poor or mediocre datasets to receive high QC scores. For example, this can happen in the presence in the IP of very strongly clustered background of the kind we found in some control datasets. It can also happen for factors that ChIP so well, and receive such high scores, that even data that are substantially suboptimal score highly (for example, CTCF ChIP-seq datasets routinely identify 35-40,000 reproducible binding sites and have QC scores of 2; a dataset that identifies only 15,000 sites is clearly suboptimal given that knowledge, yet it can still contain enough read clustering to receive a positive QC score and to match the maximum extent of read clustering observed for many other transcription factors). For these reasons, the quality metrics should be applied and interpreted in the context of what is known about the factor, the system, and the questions under study. Despite the important nuances of interpretation, using metrics and making the results readily accessible for every dataset would facilitate better informed data use by the wider community. An important adjunct to public QC annotation would be the ability, in major public data repositories, to flag and explain the exceptional cases for which QC scores should not be taken at face value. Finally, quality metrics themselves will continue to improve as the field's understanding of data structure, experimental artifacts, and the underlying biology all become more sophisticated. Provisions will be needed for incorporating such advances into routine dataset annotation, while still achieving comparability through time.

10.4 Methods

10.4.1 Sequencing read alignment

Raw sequencing reads for all non-ENCODE GEO series containing ChIP-seq datasets against transcription factors and chromatin modifying proteins (submitted prior to April 1st 2012) were downloaded from GEO in SRA format and converted to FASTQ format using the `fastq-dump` program in the `sratoolkit`, version 2.1.9. Reads were aligned using Bowtie (Langmead et al. 2009), version 0.12.7, with the following settings: `''-v 2 -t -k 2 -m 1 --best --strata''`, which allow for two mismatches relative to the reference and only retain unique alignments. Human datasets were mapped against the male set of chromosomes (excluding all random chromosomes and haplotypes) for version `hg19` of the human genome; the `mm9` version of the mouse genome was used for mouse data, `rn5` for rat, `danRer7` for zebrafish, `susScr2` for pig, and `xenTro3` for the clawed frog *Xaenopus tropicalis*; all assemblies were downloaded from the UCSC genome browser (Kent at al. 2002).

10.4.2 ChIP quality assessment

ChIP quality assessment was carried out on both ChIP and input datasets using the general strategy described in (Landt et al. 2012). Because a library may not score as a successful ChIP for reasons other than the IP itself failing (such as it being carried out in a knockout background, in si/shRNA-treated cells or in conditions under which the factor is not expressed or not bound to DNA), the following additional criteria were used to determine whether each library is expected to score positively in the QC assessment:

1. All experiments claimed to be successful by authors are expected to exhibit high level of read clustering

2. All inputs (sonicated DNA and IgG mock IPs) are expected to exhibit minimal read clustering (QC tag of -2 or -1)

3. All ChIP-seq experiments carried out in a knock-out background for the factor are expected to exhibit minimal read clustering (QC tag of -2 or -1)

4. As knock-down efficiency varies and it is unknown what protein levels would be sufficiently high for the factor to be successfully ChIP-ed, ChIP-seq experiments carried out in cells treated with si/shRNAs targeting the factor are set aside as "unknown" and assessed for library complexity and sequencing depth but not for ChIP quality.

5. Experiments against factors known to bind to DNA upon some stimulus carried out in unstimulated cells are also tagged as "unknown" as lower level binding in unstimulated cells cannot be ruled out (and is in fact often observed).

6. Experiments carried out in conditions which may result in the factor not binding to DNA (time courses, knock-downs or knock-outs for other factors that may affect binding of the targeted factor, etc.) are also tagged as "unknown"

7. Other experiments not matching any of these categories are expected to exhibit high level of read clustering

Cross-correlation analysis was performed using version 1.10.1 of `SPP` (Kharchenko et al. 2008) and the following parameters: `'-s=0:2:400'`. QC scores were assigned based on the RSC values (integers ranging from -2 to -2, $RSC \in \{0, 0.25\} \Rightarrow QC \leftarrow -2$, $RSC \in \{0.25, 0.50\} \Rightarrow QC \leftarrow -1$, $RSC \in \{0.50, 1.00\} \Rightarrow QC \leftarrow 0$, $RSC \in \{1, 1.50\} \Rightarrow QC \leftarrow +1$, $RSC \geq 1.5 \Rightarrow QC \leftarrow +2$, with -2 corresponding to minimal read clustering and 2 to a highly clustered library;) and used as a measure of ChIP quality. These scores capture the extent of read clustering in a ChIP-seq experiment in organisms whose genomes have similar size and structure to those of mammals. We point out that these scores may not be appropriate in genomes with very different size and/or structure. This motivated us to discard data from non-vertebrate model organisms for this analysis). Different values of RSC or NSC coefficients may be more informative for such genomes and is a topic for future investigation. Cross-correlation plots were manually examined in order to ensure no artifactual QC scores were included due to size selection issues (such as, for example, a library being fragmented to an average size close to the read length and confusing the automated fragment peak assignment). The code for running SPP and assigning QC scores is available at `https://code.google.com/p/phantompeakqualtools/`

10.4.3 MyoD and myogenin ChIP-seq peak calling

MyoD and myogenin datasets were generated by the Wold lab and are available under GEO accession number GSE44824. We note that the apparent weakness of the "myogenin 2" ChIP dataset is most likely due to undersequencing and would be elevated to high quality status if sequenced deeper; undersequencing is one possible reason for suboptimal quality metrics (Kundaje et al, submitted). Reads were mapped as described above and peaks called using ERANGE3.2 (Johnson et al. 2007) with the following settings: `'-minimum 2 -ratio 3 -shift learn -revbackground -listPeak'`. ChIP-seq peak calls were counted as overlapping if their summits were within 200bp of each other. Read mapping statistics and QC metrics for these datasets can be found in Supplementary Table 2.

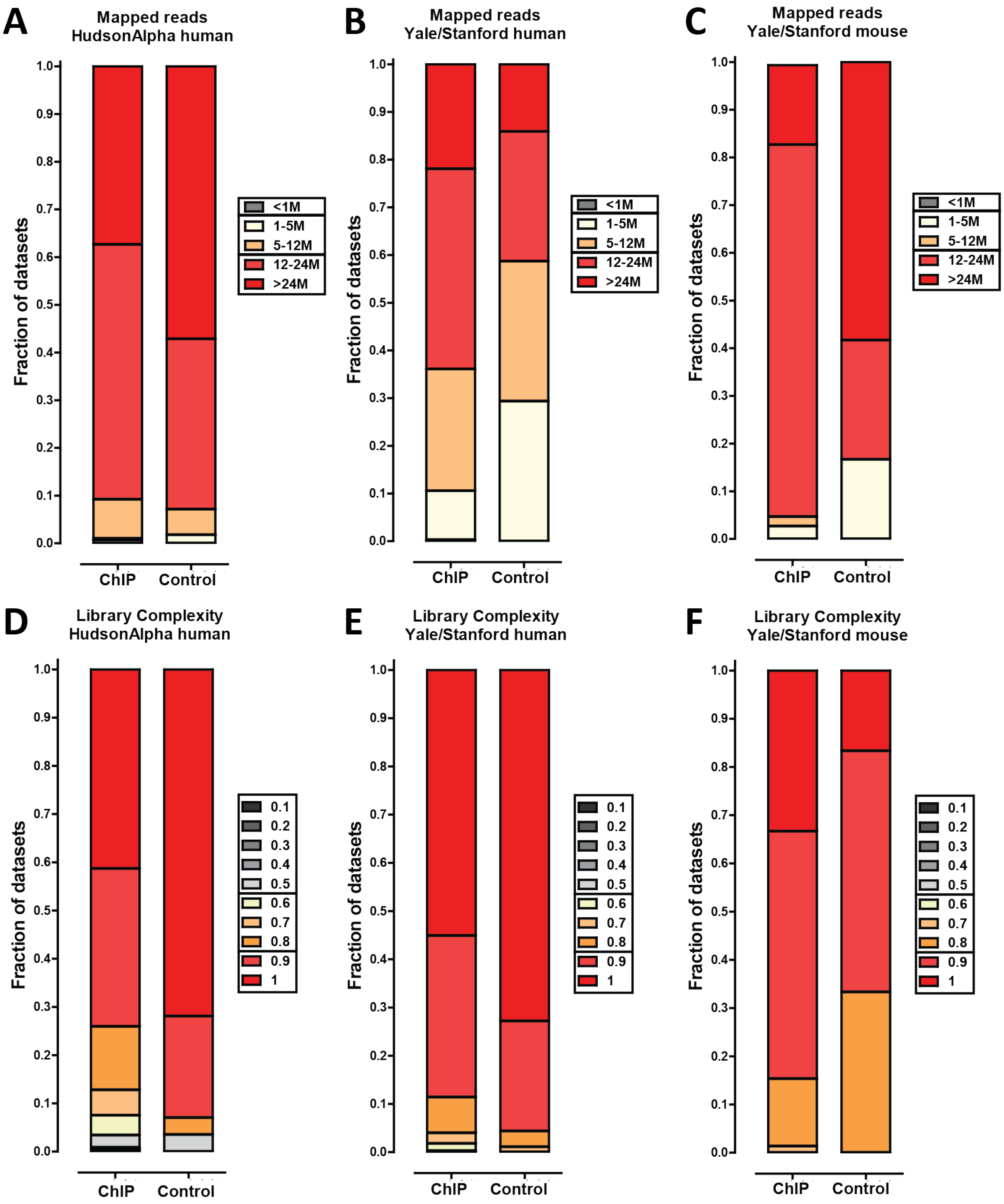


Figure 10.13: Distribution of the number of mapped reads and library complexity for data from the main two TF ChIP-seq production groups in ENCODE. (A,B,C) Number mapped reads. (D,E,F). Library complexity. Note that the same filters on the dataset inclusions that were used on publicly available data (see Methods section) were also applied to ENCODE datasets.

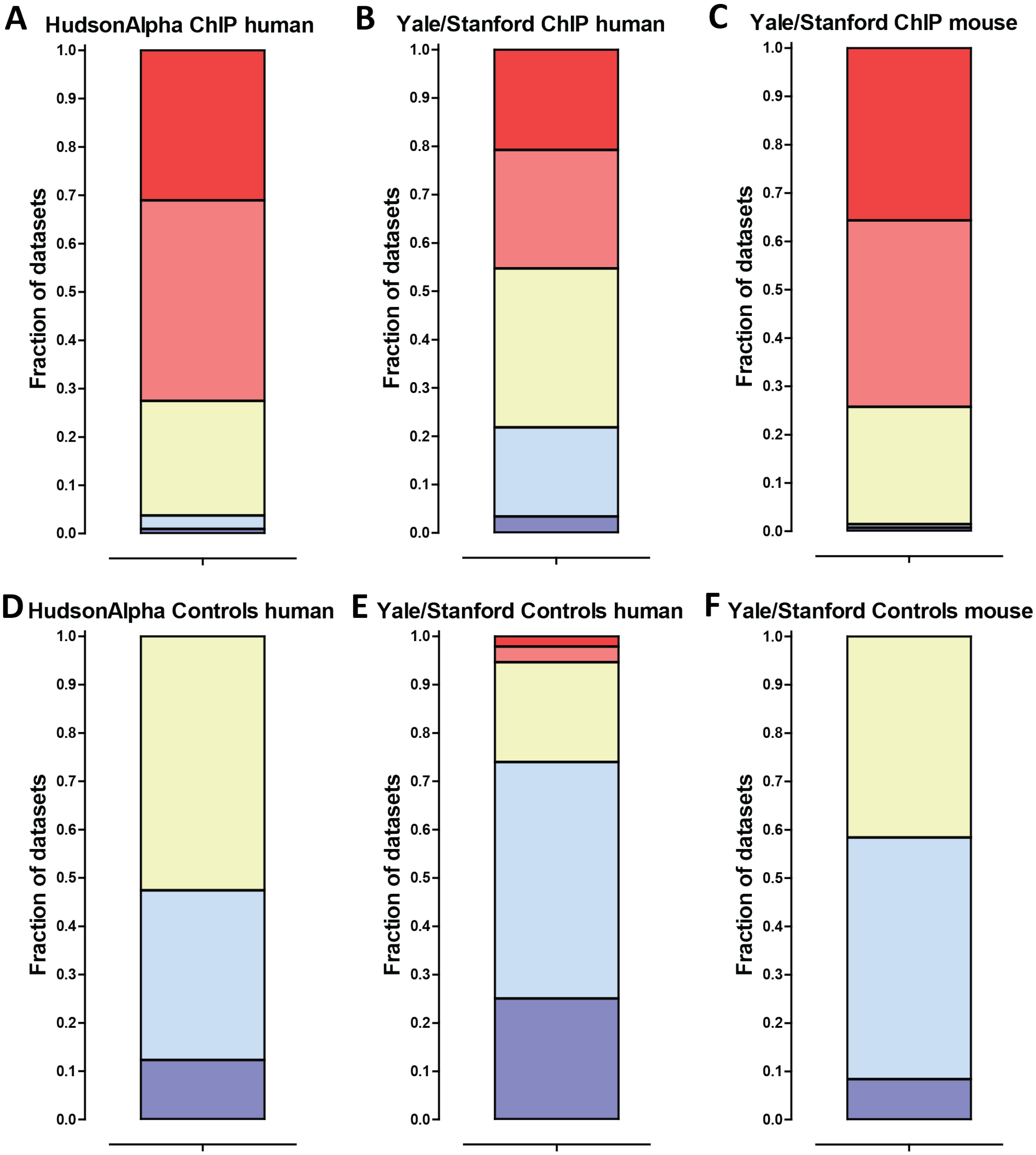


Figure 10.14: Distribution of the discretized RSC QC scores for data from the main two TF ChIP-seq production groups in ENCODE. (A,B,C) Transcription factor ChIP-seq data. (D,E,F). Control datasets (Input and IgG).

Table 10.1: Dataset QC evaluation and mapping statistics. A direct link to the GEO entry is provided in the “Source” field

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Marson et al. 2008	mouse	Nanog-mES-rep1	0.94	4.37	1.67	2	26	26	26	4,305,381	ChIP	yes
Marson et al. 2008	mouse	Nanog-mES-rep2	0.94	4.32	1.67	2	26	26	26	4,396,374	ChIP	yes
Marson et al. 2008	mouse	oct4-mES-rep1	0.95	6.54	0.34	-1	26	26	26	4,341,147	ChIP	yes
Marson et al. 2008	mouse	sox2-mES-rep1	0.96	4.03	1.21	1	26	26	26	4,033,101	ChIP	yes
Marson et al. 2008	mouse	sox2-mES-rep2	0.97	4.07	1.19	1	26	26	26	3,287,130	ChIP	yes
Marson et al. 2008	mouse	suz12-mES-repl	0.97	1.63	0.15	-2	26	26	26	3,624,473	ChIP	yes
Marson et al. 2008	mouse	suz12-rep2-1	0.98	1.24	0.41	-1	26	26	26	207,308	ChIP	yes
Marson et al. 2008	mouse	Tcf3-mES-rep1	0.96	3	0.72	0	26	26	26	5,247,274	ChIP	yes
Marson et al. 2008	mouse	Tcf3-mES-rep2	0.96	2.96	0.66	0	26	26	26	5,388,916	ChIP	yes
Marson et al. 2008	mouse	WCE-mES-rep1	0.94	1.37	0.2	-2	26	26	26	1,507,157	Input	no
Marson et al. 2008	mouse	WCE-mES-rep2	0.9	1.35	0.21	-2	26	26	26	3,770,502	Input	no
Chen et al. 2008	mouse	ES-c-Myc	0.86	1.51	0.48	-1	26	26	26	11,714,595	ChIP	yes
Chen et al. 2008	mouse	ES-E2f1	0.84	1.43	0.92	0	26	26	26	13,374,901	ChIP	yes
Chen et al. 2008	mouse	ES-Esrrb	0.88	4.5	1.69	2	26	26	26	7,982,162	ChIP	yes
Chen et al. 2008	mouse	ES-GFP	0.82	1.26	0.15	-2	26	26	26	7,520,858	IgG	no
Chen et al. 2008	mouse	ES-Klf4	0.41	1.96	0.62	0	36	36	36	368,908	ChIP	yes
Chen et al. 2008	mouse	ES-Nanog	0.81	3158	356	2	26	26	26	9,166,834	ChIP	yes
Chen et al. 2008	mouse	ES-n-Myc	0.79	1.74	0.41	-1	26	26	26	10,099,160	ChIP	yes
Chen et al. 2008	mouse	ES-Oct4	0.59	1.61	0.46	-1	36	36	36	139,512	ChIP	yes
Chen et al. 2008	mouse	ES-p300	0.77	1.26	0.23	-2	26	26	26	9,396,456	ChIP	yes
Chen et al. 2008	mouse	ES-Smad1	0.96	3074	298	2	26	26	26	9,681,328	ChIP	yes
Chen et al. 2008	mouse	ES-Sox2	0.86	1.94	0.62	0	26	26	26	12,489,175	ChIP	yes
Chen et al. 2008	mouse	ES-STAT3	0.77	1.68	0.38	-1	26	26	26	8,384,452	ChIP	yes
Chen et al. 2008	mouse	ES-Suz12	0.87	1.21	0.27	-1	26	26	26	12,378,715	ChIP	yes
Chen et al. 2008	mouse	ES-Tcfcp2I1	0.81	13.53	2.42	2	26	26	26	8,800,970	ChIP	yes
Chen et al. 2008	mouse	ES-Zfx	0.92	1.9	0.71	0	31	31	31	9,543,774	ChIP	yes
Kwon et al. 2009	mouse	GIgG-post-IL21-in-B-cells	0.84	1.4	0.47	-1	25	25	25	2,915,090	IgG	no
Kwon et al. 2009	mouse	GIgG-post-IL6	0.88	1.49	0.36	-1	25	25	25	2,129,448	IgG	no
Kwon et al. 2009	mouse	IgG-post-IL21-dup	0.89	1.26	0.34	-1	25	25	25	4,286,349	IgG	no
Kwon et al. 2009	mouse	IgG-post-IL21-in-B-cells	0.91	1.33	0.52	0	25	25	25	2,993,063	IgG	no
Kwon et al. 2009	mouse	IgG-post-IL21-in-WT-quar	0.74	1.89	0.7	0	25	25	25	7,228,784	IgG	no
Kwon et al. 2009	mouse	IgG-post-IL21-ter	0.92	1.35	0.11	-2	25	25	25	3,080,974	IgG	no
Kwon et al. 2009	mouse	IgG-post-IL2-dup	0.86	1.31	0.4	-1	25	25	25	3,406,531	IgG	no
Kwon et al. 2009	mouse	IgG-post-IL6	0.84	1.31	0.35	-1	25	25	25	4,247,181	IgG	no
Kwon et al. 2009	mouse	IRF4-post-IL21-dup-seq-1	0.86	2.08	0.89	0	25	25	25	3,295,774	ChIP	yes
Kwon et al. 2009	mouse	IRF4-post-IL21-dup-seq-2	0.66	2.21	0.92	0	25	25	25	5,496,827	ChIP	yes
Kwon et al. 2009	mouse	IRF4-post-IL21-in-IRF4-KO-mice	0.94	1.9	0.66	0	25	25	25	3,613,839	ChIP	no
Kwon et al. 2009	mouse	IRF4-post-IL21-seq-1	0.9	1.88	0.72	0	25	25	25	3,560,575	ChIP	yes
Kwon et al. 2009	mouse	IRF4-post-IL21-seq-2	0.93	2.14	0.73	0	25	25	25	1,273,441	ChIP	yes
Kwon et al. 2009	mouse	IRF4-pre-IL21-dup-seq-1	0.89	1.76	0.7	0	25	25	25	3,996,341	ChIP	yes
Kwon et al. 2009	mouse	IRF4-pre-IL21-dup-seq-2	0.9	1.83	0.72	0	25	25	25	3,308,064	ChIP	yes
Kwon et al. 2009	mouse	IRF4-pre-IL21-in-IRF4-KO-mice	0.81	1.5	0.42	-1	25	25	25	5,741,089	ChIP	no
Kwon et al. 2009	mouse	IRF4-pre-IL21-seq-1	0.89	2.19	0.9	0	25	25	25	2,882,656	ChIP	yes
Kwon et al. 2009	mouse	IRF4-pre-IL21-seq-2	0.73	2.44	0.84	0	25	25	25	4,304,206	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-dup-seq-1	0.93	1.47	0.44	-1	25	25	25	266,685	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-dup-seq-2	0.89	1.47	0.46	-1	25	25	25	1,818,900	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-in-IRF4-KO-mice	0.79	1.42	0.28	-1	25	25	25	5,334,084	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-post-IL21-in-IRF4-KO-mice-second-exp	0.81	1.52	0.52	0	25	25	25	5,744,414	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-post-IL21-in-IRF4-KO-mice-third-exp	0.82	1.5	0.48	-1	25	25	25	4,041,874	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-post-IL21-in-WT-cinq	0.77	1.87	0.58	0	25	25	25	5,281,605	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-in-WT-quar	0.76	1.88	0.63	0	25	25	25	5,450,390	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-in-WT-ter	0.85	1.46	0.43	-1	25	25	25	3,416,726	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-seq-1	0.88	2.75	0.86	0	25	25	25	3,446,457	ChIP	yes
Kwon et al. 2009	mouse	STAT3-post-IL21-seq-2	0.84	2.89	0.81	0	25	25	25	3,340,925	ChIP	yes
Kwon et al. 2009	mouse	STAT3-pre-IL21-dup-seq-1	0.82	1.37	0.41	-1	25	25	25	3,736,863	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-dup-seq-2	0.86	1.42	0.4	-1	25	25	25	2,709,584	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-in-IRF4-KO-mice-first-exp	0.85	1.56	0.45	-1	25	25	25	3,709,343	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-in-IRF4-KO-mice-second-exp	0.79	1.83	0.71	0	25	25	25	6,924,787	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-in-WT-4th-experiment	0.82	1.51	0.37	-1	25	25	25	4,257,111	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-in-WT-ter	0.81	1.41	0.14	-2	25	25	25	3,395,506	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-seq-1	0.78	3.1	0.76	0	25	25	25	3,700,560	ChIP	unknown
Kwon et al. 2009	mouse	STAT3-pre-IL21-seq-2	0.76	3.1	0.67	0	25	25	25	3,667,278	ChIP	unknown

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Hollenhorst et al. 2009	human	Jurkat-CBP-1	0.86	2.22	1.38	1	36	36	36	9,275,556	ChIP	yes
Hollenhorst et al. 2009	human	Jurkat-ETS1-1	0.85	2.41	2.58	2	26	26	26	7,562,377	ChIP	yes
Hollenhorst et al. 2009	human	Jurkat-Input-1	0.93	1.16	0.15	-2	26	26	26	15,389,799	Input	no
Hollenhorst et al. 2009	human	Jurkat-RUNX-1	0.63	2.55	0.61	0	36	36	36	10,337,694	ChIP	yes
Han et al. 2010	mouse	mESC-Input	0.96	1.17	0.18	-2	37	37	37	9,567,449	ChIP	no
Han et al. 2010	mouse	mESC-Tbx3	0.92	1.46	0.29	-1	35	35	35	7,526,549	ChIP	yes
Yu et al. 2009	mouse	MEL86-GATA1	0.94	1.34	0.28	-1	33.87	36	28	5,866,520	ChIP	yes
De et al. 2009	mouse	Macrophages-JMJD3	0.95	1.32	0.36	-1	36	36	36	8,731,417	ChIP	yes
Yuan et al. 2009	mouse	mESC-ESET	0.92	1.93	0.65	0	36	36	36	11,607,868	ChIP	yes
Bilodeau et al. 2009	mouse	mESC-SetDB1-rep-1	0.92	1.68	0.54	0	36	36	36	3,620,404	ChIP	yes
Bilodeau et al. 2009	mouse	mESC-SetDB1-rep-2	0.93	1.64	0.42	-1	36	36	36	3,301,043	ChIP	yes
Bilodeau et al. 2009	mouse	mESC-SetDB1-rep-3	0.92	1.65	0.51	0	36	36	36	4,251,421	ChIP	yes
Bilodeau et al. 2009	mouse	mESC-WCE-mES-rep-1	0.94	1.38	0.31	-1	36	36	36	3,966,359	Input	no
Lister et al. 2009	human	hESC-NANOG-1a	0.65	18.86	4.47	2	36	36	36	3,701,686	ChIP	yes
Lister et al. 2009	human	hESC-NANOG-1b	0.61	17.44	4.43	2	36	36	36	4,523,040	ChIP	yes
Lister et al. 2009	human	hESC-SOX2-1a	0.82	9.01	4.94	2	36	36	36	4,591,769	ChIP	yes
Lister et al. 2009	human	hESC-KLF4-1a	0.32	46.06	24.39	2	36	36	36	810,796	ChIP	yes
Lister et al. 2009	human	hESC-MYC-1a	0.58	4.15	2.02	2	36	36	36	2,391,782	ChIP	yes
Lister et al. 2009	human	hESC-Oct4-1a	0.98	2.46	1.04	1	36	36	36	574,662	ChIP	yes
Lister et al. 2009	human	hESC-Oct4-2a	0.98	4.37	1.81	2	36	36	36	151,346	ChIP	yes
Lister et al. 2009	human	hESC-P300-1a	0.57	7.51	2.52	2	36	36	36	3,490,165	ChIP	yes
Lister et al. 2009	human	hESC-TAFII-1a	0.64	2.93	1.96	2	36	36	36	4,031,316	ChIP	yes
Lister et al. 2009	human	hESC-TAFII-1b	0.67	2.9	1.72	2	36	36	36	3,507,401	ChIP	yes
Nishiyama et al. 2009	mouse	mESC-Cdx2	0.94	1.14	0.32	-1	36	36	36	7,347,351	ChIP	yes
Cheng et al. 2009	mouse	G1E-ER4-GATA1	0.96	1.75	1.2	1	36	36	36	24,281,091	ChIP	yes
Cheng et al. 2009	mouse	G1E-ER4-Input	0.97	1.28	0.58	0	36	36	36	15,990,494	Input	no
Wilson et al. 2009	mouse	HPC-7-Scl-1	0.95	1.86	0.88	0	45	45	45	5,563,933	ChIP	yes
Wilson et al. 2009	mouse	HPC-7-Scl-2	0.96	1.83	0.47	-1	36	36	36	3,637,766	ChIP	yes
Robertson et al. 2007	human	HeLaS3-IFNgamma-1-STAT1	0.86	4.41	1.1	1	27	27	27	693,473	ChIP	yes
Robertson et al. 2007	human	HeLaS3-IFNgamma-2-STAT1	0.85	3.67	1.04	1	27	27	27	663,874	ChIP	yes
Robertson et al. 2007	human	HeLaS3-IFNgamma-3-STAT1	0.94	4.07	1.58	2	36	36	36	3,079,284	ChIP	yes
Robertson et al. 2007	human	HeLaS3-IFNgamma-4-STAT1	0.94	3.97	1.59	2	27	27	27	2,176,985	ChIP	yes
Robertson et al. 2007	human	HeLaS3-IFNgamma-5-STAT1	0.95	4	1.96	2	27	27	27	2,808,038	ChIP	yes
Robertson et al. 2007	human	HeLaS3-IFNgamma-6-STAT1	0.95	4.17	1.95	2	27	27	27	2,718,185	ChIP	yes
Robertson et al. 2007	human	HeLaS3-unstimulated-1-STAT1	0.86	1.98	0.33	-1	27	27	27	478,619	ChIP	unknown
Robertson et al. 2007	human	HeLaS3-unstimulated-2-STAT1	0.84	2.6	0.34	-1	27	27	27	500,638	ChIP	unknown
Robertson et al. 2007	human	HeLaS3-unstimulated-3-STAT1	0.82	2.2	0.29	-1	27	27	27	496,979	ChIP	unknown
Robertson et al. 2007	human	HeLaS3-unstimulated-4-STAT1	0.92	1.45	0.23	-2	36	36	36	2,746,723	ChIP	unknown
Robertson et al. 2007	human	HeLaS3-unstimulated-5-STAT1	0.93	2.17	0.3	-1	27	27	27	1,447,320	ChIP	unknown
Robertson et al. 2007	human	HeLaS3-unstimulated-6-STAT1	0.93	2.05	0.28	-1	27	27	27	1,425,741	ChIP	unknown
Robertson et al. 2007	human	HeLaS3-unstimulated-7-STAT1	0.95	1.47	0.34	-1	27	27	27	2,452,058	ChIP	unknown
Welboren et al. 2009	human	MCF7-E2-ERa	0.7	12.72	1.74	2	32	32	32	9,428,987	ChIP	yes
Welboren et al. 2009	human	MCF7-Fulvestrant-ERa	0.83	5.17	1.22	1	32	32	32	6,243,484	ChIP	yes
Welboren et al. 2009	human	MCF7-mock-treated-ERa	0.67	6.52	2.9	2	32	32	32	1,722,599	ChIP	unknown
Welboren et al. 2009	human	MCF7-Tamoxifen-ERa	0.82	8.86	1.55	2	32	32	32	5,836,314	ChIP	yes
Visel et al. 2009; Gotea et al. 2010; Blow et al. 2010	mouse	Forebrain-p300	0.57	1.72	0.15	-2	36.32	38	36	4,842,793	ChIP	yes
Visel et al. 2009; Gotea et al. 2010; Blow et al. 2010	mouse	Limb-p300	0.73	2.16	0.15	-2	36	36	35	2,209,017	ChIP	yes
Visel et al. 2009; Gotea et al. 2010; Blow et al. 2010	mouse	Midbrain-p300	0.49	1.97	0.18	-2	36.25	38	36	5,942,773	ChIP	yes
Ho et al. 2009	mouse	mESC-Brg-J1	0.86	1.25	0.49	-1	25	25	25	12,146,582	ChIP	yes
Ho et al. 2009	mouse	mESC-IgG	0.91	1.17	0.51	0	25	25	25	14,118,667	IgG	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Cuddapah et al. 2009	human	CD4+-CTCF	0.88	31.05	2.29	2	24	24	24	2,942,119	ChIP	yes
Cuddapah et al. 2009	human	HeLa-CTCF	0.93	5.22	1.29	1	24	24	24	3,294,793	ChIP	yes
Cuddapah et al. 2009	human	Jurkat-CTCF	0.91	4.22	1.01	1	25	25	25	4,367,791	ChIP	yes
Krebs et al. 2010	mouse	mESC-LUZP1	0.62	16.78	4.02	2	36	36	36	7,021,192	ChIP	yes
Krebs et al. 2010	mouse	mESC-mock	0.37	16.17	3.94	2	36	36	36	7,446,009	IgG	no
Corbo et al. 2010	mouse	NRL-KO-Crx-Rep1	0.69	1.81	1.18	1	36	36	36	12,527,332	ChIP	yes
Corbo et al. 2010	mouse	NRL-KO-Crx-Rep2	0.8	1.8	1.24	1	36	36	36	10,488,445	ChIP	yes
Corbo et al. 2010	mouse	NRL-KO-IgG-Rep1	0.75	1.58	0.9	0	36	36	36	12,160,830	IgG	no
Corbo et al. 2010	mouse	NRL-KO-IgG-Rep2	0.69	2.43	1.51	2	36	36	36	11,005,528	IgG	no
Corbo et al. 2010	mouse	WT-Crx-Rep1	0.89	3.69	0.63	0	36	36	36	4,302,798	ChIP	yes
Corbo et al. 2010	mouse	WT-Crx-Rep2	0.9	4.07	0.89	0	36	36	36	4,308,655	ChIP	yes
Corbo et al. 2010	mouse	WT-IgG-Rep1	0.92	2.53	0.42	-1	36	36	36	3,707,696	IgG	no
Ramagopalan et al. 2010	human	GM10855-Input	0.94	1.16	0.27	-1	36	36	36	11,412,903	Input	no
Ramagopalan et al. 2010	human	GM10855-unstimulated-rep1	0.87	1.41	0.56	0	36	36	36	13,520,376	ChIP	unknown
Ramagopalan et al. 2010	human	GM10855-unstimulated-rep2	0.88	1.47	0.55	0	36	36	36	10,791,763	ChIP	unknown
Ramagopalan et al. 2010	human	GM10855-vitaminD-rep1	0.89	1.76	0.83	0	36	36	36	13,970,589	ChIP	yes
Ramagopalan et al. 2010	human	GM10855-vitaminD-rep2	0.89	1.67	0.82	0	36	36	36	14,642,572	ChIP	yes
Ramagopalan et al. 2010	human	GM10861-Input	0.95	1.19	0.35	-1	36	36	36	11,404,257	Input	no
Ramagopalan et al. 2010	human	GM10861-unstimulated-rep1	0.93	1.39	0.56	0	36	36	36	10,157,583	ChIP	unknown
Ramagopalan et al. 2010	human	GM10861-unstimulated-rep2	0.93	1.52	0.67	0	36	36	36	7,922,208	ChIP	unknown
Ramagopalan et al. 2010	human	GM10861-vitaminD-rep1	0.92	1.88	0.95	0	36	36	36	10,649,722	ChIP	yes
Ramagopalan et al. 2010	human	GM10861-vitaminD-rep2	0.93	1.88	0.95	0	36	36	36	11,754,302	ChIP	yes
Wei et al. 2010	mouse	Th1-STAT4-KO-STAT4	0.84	2.05	1.25	1	36	36	36	9,339,036	ChIP	no
Wei et al. 2010	mouse	Th1-WT-STAT4	0.88	7.69	2.19	2	36	36	36	10,525,607	ChIP	yes
Wei et al. 2010	mouse	Th2-Normal-Rabbit-Serum	0.75	2.42	1.31	1	36	36	36	7,610,146	IgG	no
Wei et al. 2010	mouse	Th2-STAT6-KO-STAT6	0.89	2.38	1.37	1	36	36	36	9,734,600	ChIP	no
Wei et al. 2010	mouse	Th2-WT-STAT6	0.83	6.45	1.62	2	36	36	36	9,139,067	ChIP	yes
Schnetz et al. 2010	mouse	mES-CHD7	0.91	1.56	0.55	0	37	37	37	8,269,486	ChIP	yes
Schnetz et al. 2010	mouse	mES-p300	0.96	1.35	0.73	0	37	37	37	17,677,307	ChIP	yes
GSE22303	mouse	mES-B2-TBP	0.92	2.43	1.21	1	36	36	36	18,683,322	ChIP	yes
GSE22303	mouse	mES-B6-TBP	0.91	1.94	0.73	0	26	26	26	3,617,586	ChIP	yes
Lin et al. 2010	mouse	A12-E2A-6h-E47ER	0.93	3.41	0.61	0	36	36	36	2,776,323	ChIP	yes
Lin et al. 2010	mouse	E2AKO-E2A-1h-E47ER	0.68	9.12	1.13	1	36	36	36	5,948,823	ChIP	yes
Lin et al. 2010	mouse	E2AKO-E2A-6h-E47ER	0.96	4.34	0.59	0	36	36	36	2,196,108	ChIP	yes
Lin et al. 2010	mouse	EBFKO-E2A	0.92	1.72	0.53	0	36	36	36	9,159,853	ChIP	yes
Lin et al. 2010	mouse	Input2	0.93	1.15	0.11	-2	36	36	36	10,675,120	Input	no
Lin et al. 2010	mouse	RAG1KO-CTCF	0.91	15.22	2.31	2	36	36	36	4,804,275	ChIP	yes
Lin et al. 2010	mouse	RAG1KO-E2A	0.85	4.13	1.39	1	30	36	25	7,601,861	ChIP	yes
Lin et al. 2010	mouse	RAG1KO-EBF	0.81	10.08	1.23	1	36	36	36	2,935,481	ChIP	yes
Lin et al. 2010	mouse	RAG1KO-FOXO1-1	0.91	6.82	1.06	1	36	36	36	15,561,578	ChIP	yes
Durant et al. 2010	mouse	Th17-Stat3fl-fl-FoxP3-GFP-STAT3	0.81	2.61	1.36	1	36	36	36	12,871,479	ChIP	yes
Heinz et al. 2010	mouse	Bcell-input-ChIP-Seq	0.68	1.75	0.12	-2	36	36	36	11,410,688	Input	no
Heinz et al. 2010	mouse	Bcell-Oct2-ChIP-Seq	0.95	2.16	0.17	-2	36	36	36	2,296,228	ChIP	yes
Heinz et al. 2010	mouse	Bcell-PU.1-ChIP-Seq	0.92	5.56	4.62	2	36	36	36	8,207,220	ChIP	yes
Heinz et al. 2010	mouse	BirA-input-GW-ChIP-Seq	0.96	1.24	0.54	0	23	23	23	2,263,641	Input	no
Heinz et al. 2010	mouse	BLRP-LXRb-GW-ChIP-Seq	0.8	2.02	0.68	0	22.45	25	22	9,426,604	ChIP	yes
Heinz et al. 2010	mouse	BMDM.LXRDKO-PU.1-ChIP-Seq	0.97	9.26	2.72	2	23	23	23	2,410,527	ChIP	yes
Heinz et al. 2010	mouse	BMDM-PU.1-ChIP-Seq	0.93	10.78	1.97	2	36	36	36	9,617,221	ChIP	yes
Heinz et al. 2010	mouse	E2AKO-PU.1-bHLH-ER-ChIP-Seq	0.91	7.77	2.76	2	23	23	23	5,093,144	ChIP	yes
Heinz et al. 2010	mouse	E2AKO-PU.1-ChIP-Seq	0.89	10.21	2.94	2	23	23	23	3,615,197	ChIP	yes
Heinz et al. 2010	mouse	E2AKO-PU.1-E2A-ER-ChIP-Seq	0.88	8.28	2.67	2	23	23	23	4,724,664	ChIP	yes
Heinz et al. 2010	mouse	EBFKO-PU.1-ChIP-Seq	0.94	11.94	2.4	2	23	23	23	3,058,714	ChIP	yes
Heinz et al. 2010	mouse	PU.1KO-CEBPb-ChIP-Seq	0.89	4.52	1.42	1	23	23	23	4,179,430	ChIP	yes
Heinz et al. 2010	mouse	PU.1KO-PU.1-ChIP-Seq	0.88	2.79	0.53	0	23	23	23	4,615,899	ChIP	no
Heinz et al. 2010	mouse	PUER-CEBPb-0h-ChIP-Seq	0.92	8.61	2.14	2	23	23	23	4,672,159	ChIP	yes
Heinz et al. 2010	mouse	PUER-CEBPb-1h-ChIP-Seq	0.92	9.02	2.1	2	23	23	23	3,790,612	ChIP	yes
Heinz et al. 2010	mouse	PUER-CEBPb-24h-ChIP-Seq	0.89	10.77	2.54	2	23	23	23	4,625,986	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Heinz et al. 2010	mouse	PUER-CEBPb-48h-ChIP-Seq	0.9	8.09	1.95	2	23	23	23	5,022,074	ChIP	yes
Heinz et al. 2010	mouse	PUER-CEBPb-6h-ChIP-Seq	0.89	9.43	2.06	2	23	23	23	4,417,004	ChIP	yes
Heinz et al. 2010	mouse	PUER-PU.1-0h-ChIP-Seq	0.94	4.68	0.57	0	23	23	23	2,053,953	ChIP	yes
Heinz et al. 2010	mouse	PUER-PU.1-1h-ChIP-Seq	0.92	12.06	2.58	2	23	23	23	2,541,096	ChIP	yes
Heinz et al. 2010	mouse	PUER-PU.1-24h-ChIP-Seq	0.9	18.69	2.85	2	23	23	23	3,403,839	ChIP	yes
Heinz et al. 2010	mouse	PUER-PU.1-48h-ChIP-Seq	0.9	14.68	3.42	2	23	23	23	4,138,465	ChIP	yes
Heinz et al. 2010	mouse	PUER-PU.1-6h-ChIP-Seq	0.9	16.62	2.83	2	23	23	23	3,477,600	ChIP	yes
Heinz et al. 2010	mouse	RAG1KO-PU.1-ChIP-Seq	0.86	12.16	2.26	2	23	23	23	6,302,473	ChIP	yes
Heinz et al. 2010	mouse	ThioMac-CEBPa-ChIP-Seq	0.92	5.55	2.96	2	23	23	23	7,067,160	ChIP	yes
Heinz et al. 2010	mouse	ThioMac-input-ChIP-Seq	0.93	1.25	0.11	-2	23.67	25	22	5,491,097	Input	no
Heinz et al. 2010	mouse	ThioMac-PU.1-ChIP-Seq	0.97	5.1	3.4	2	23	23	23	5,289,667	ChIP	yes
Steger et al. 2010	mouse	3T3-L1-0hr-CEBPb	0.51	5.34	1.6	2	36	36	36	11,295,935	ChIP	yes
Steger et al. 2010	mouse	3T3-L1-0hr-Input	0.94	3.23	0.46	-1	36	36	36	5,129,801	Input	no
Steger et al. 2010	mouse	3T3-L1-240hr-Input	0.95	3.48	0.55	0	36	36	36	5,019,654	Input	no
Steger et al. 2010	mouse	3T3-L1-24hr-Input	0.81	6.84	1.26	1	36	36	36	4,731,402	Input	no
Steger et al. 2010	mouse	3T3-L1-6hr-CEBPb	0.87	3.05	1.09	1	36	36	36	10,746,117	ChIP	yes
Steger et al. 2010	mouse	3T3-L1-6hr-GR	0.86	1.79	0.75	0	36	36	36	10,761,593	ChIP	yes
Steger et al. 2010	mouse	3T3-L1-6hr-Input	0.9	1.4	0.56	0	36	36	36	11,352,790	Input	no
GSE21916	human	H9-IgG	0.93	1.53	0.33	-1	26	26	26	4,499,095	IgG	no
GSE21916	human	H9-Oct4-replicate-2	0.97	1.97	0.83	0	36	36	36	4,556,649	ChIP	yes
GSE21916	human	H9-Oct4-technical-replicate-1	0.92	1.81	0.48	-1	26	26	26	4,187,685	ChIP	yes
GSE21916	human	H9-Oct4-technical-replicate-2	0.95	1.85	0.6	0	36	36	36	4,119,980	ChIP	yes
Kassouf et al. 2010	mouse	RER-SCL	0.72	1.94	0.91	0	36	36	36	5,208,895	ChIP	no
Kassouf et al. 2010	mouse	RER-SCL-no-AB	0.51	9.5	1.38	1	36	36	36	4,571,728	IgG	no
Kassouf et al. 2010	mouse	WT-no-AB	0.81	5.94	1.05	1	36	36	36	5,312,397	IgG	no
Kassouf et al. 2010	mouse	WT-SCL	0.68	2.77	1.4	1	36	36	36	4,154,252	ChIP	yes
MacIsaac et al. 2010	mouse	CEBPa-3T3-L1	0.93	2.43	0.75	0	35	35	35	4,326,509	ChIP	yes
MacIsaac et al. 2010	mouse	CEBPa-liver	0.9	14.08	2.03	2	35	35	35	4,595,713	ChIP	yes
MacIsaac et al. 2010	mouse	E2F4-liver	0.92	16.44	2.26	2	35	35	35	2,214,727	ChIP	yes
MacIsaac et al. 2010	mouse	FOXA1-liver	0.54	12.93	3.19	2	35	35	35	3,968,403	ChIP	yes
MacIsaac et al. 2010	mouse	FOXA2-liver	0.95	6.49	1.99	2	35	35	35	6,593,622	ChIP	yes
MacIsaac et al. 2010	mouse	p300-3T3-L1	0.96	1.74	0.77	0	35	35	35	3,575,940	ChIP	yes
MacIsaac et al. 2010	mouse	p300-liver	0.95	6.6	2.79	2	35	35	35	4,718,264	ChIP	yes
MacIsaac et al. 2010	mouse	Sample-control-reads-3T3-L1	0.29	1.46	0.47	-1	35	35	35	2,767,084	Input	no
MacIsaac et al. 2010	mouse	Sample-control-reads-cerebellum	0.93	1.9	0.52	0	35	35	35	5,139,906	Input	no
MacIsaac et al. 2010	mouse	Sample-control-reads-liver	0.64	23.8	2.7	2	35	35	35	5,270,015	Input	no
Vivar et al. 2010	human	U2OS-ERb-Doxy-nonspecificAntibodyIgG-rep1	0.96	1.33	0.26	-1	26	26	26	2,576,564	IgG	no
Vivar et al. 2010	human	U2OS-ERb-Doxy-specificAntibody-rep1	0.95	2.8	0.7	0	26	26	26	2,749,749	ChIP	yes
Vivar et al. 2010	human	U2OS-ERb-DoxyE2-nonspecificAntibodyIgG-rep1	0.95	1.43	0.31	-1	26	26	26	2,880,960	IgG	no
Vivar et al. 2010	human	U2OS-ERb-DoxyE2-specificAntibody-rep1	0.95	6.1	1.13	1	26	26	26	2,638,244	ChIP	yes
Fortschegger et al. 2010	human	Input-DNA-Hs68+FBS	0.97	1.32	0.27	-1	40	40	40	8,279,525	Input	no
Fortschegger et al. 2010	human	Input-DNA-Hs68-FBS	0.97	1.34	0.3	-1	40	40	40	7,059,465	Input	no
Fortschegger et al. 2010	human	Normal-IgG-293T	0.94	1.21	0.22	-2	50	50	50	7,860,447	IgG	no
Fortschegger et al. 2010	human	Normal-IgG-HeLa	0.92	1.76	0.55	0	50	50	50	7,000,514	IgG	no
Fortschegger et al. 2010	human	PHF8-293T	0.96	1.71	0.71	0	50	50	50	7,015,757	ChIP	yes
Fortschegger et al. 2010	human	PHF8-HeLa	0.95	2.74	1.13	1	50	50	50	6,982,792	ChIP	yes
Fortschegger et al. 2010	human	PHF8-Hs68+FBS	0.94	1.89	0.69	0	40	40	40	7,339,329	ChIP	yes
Fortschegger et al. 2010	human	PHF8-Hs68-FBS	0.9	1.75	0.51	0	35	35	35	11,313,461	ChIP	yes
GSE15844	mouse	MEF-NFIC-KO-NFI	0.29	7.55	1.85	2	35	35	35	10,794,407	ChIP	no
GSE15844	mouse	MEF-WT-Input	0.74	1.42	0.28	-1	36	36	36	5,483,670	Input	no
GSE15844	mouse	MEF-WT-NFI	0.34	5.81	1.72	2	35	35	35	9,746,594	ChIP	yes
Kim et al. 2010	mouse	ChIP-CBP-ab2832-KCl-b1	0.88	2.46	0.88	0	33	33	33	1,742,367	ChIP	yes
Kim et al. 2010	mouse	ChIP-CBP-ab2832-KCl-b2	0.9	1.9	0.31	-1	33	33	33	1,350,494	ChIP	yes
Kim et al. 2010	mouse	ChIP-CBP-ab2832-un-b1	0.11	2.19	0.23	-2	33	33	33	13,062,901	ChIP	yes
Kim et al. 2010	mouse	ChIP-CBP-Millipore-KCl-b1	0.14	2.55	0.19	-2	33	33	33	21,475,816	ChIP	yes
Kim et al. 2010	mouse	ChIP-CBP-Millipore-un-b1	0.14	3.57	0.22	-2	33	33	33	12,767,854	ChIP	yes
Kim et al. 2010	mouse	ChIP-CREB-SC-KCl-b1	0.25	1.6	0.2	-2	33	33	33	12,606,497	ChIP	yes
Kim et al. 2010	mouse	ChIP-CREB-SC-KCl-b2	0.11	2.71	0.25	-1	33	33	33	14,186,880	ChIP	yes
Kim et al. 2010	mouse	ChIP-CREB-SC-un-b1	0.47	1.39	0.25	-2	33	33	33	11,723,416	ChIP	yes
Kim et al. 2010	mouse	ChIP-CREB-SC-un-b2	0.11	3.54	0.43	-1	33	33	33	11,668,187	ChIP	yes
Kim et al. 2010	mouse	ChIP-input-KCl-b1	0.22	2.19	0.19	-2	33	33	33	29,829,497	Input	no
Kim et al. 2010	mouse	ChIP-input-KCl-b2	0.55	1.29	0.3	-1	33	33	33	11,407,302	Input	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Kim et al. 2010	mouse	ChIP-input-un-b1	0.57	1.86	0.16	-2	33	33	33	4,413,802	Input	no
Kim et al. 2010	mouse	ChIP-input-un-b2	0.59	1.28	0.39	-1	33	33	33	2,034,854	Input	no
Kim et al. 2010	mouse	ChIP-Npas4-KCl-b1	0.3	3.33	1.38	1	33	33	33	6,262,184	ChIP	yes
Kim et al. 2010	mouse	ChIP-Npas4-KCl-b2	0.7	2.39	0.92	0	33	33	33	3,474,756	ChIP	yes
Kim et al. 2010	mouse	ChIP-Npas4-un-b1	0.39	1.84	0.21	-2	33	33	33	12,918,805	ChIP	yes
Kim et al. 2010	mouse	ChIP-SRF-SC-KCl-b1	0.88	3.77	0.28	-1	33	33	33	1,953,844	ChIP	yes
Kim et al. 2010	mouse	ChIP-SRF-SC-KCl-b2	0.86	2.72	0.46	-1	33	33	33	7,001,063	ChIP	yes
Kim et al. 2010	mouse	ChIP-SRF-SC-un-b1	0.89	2.58	0.47	-1	33	33	33	2,076,216	ChIP	yes
Kim et al. 2010	mouse	ChIP-SRF-SC-un-b2	0.87	2.2	0.98	0	33	33	33	8,797,223	ChIP	yes
Lefterova et al. 2010	mouse	Lefterova-ad-PPARg	0.85	2.18	0.73	0	36	36	36	5,258,157	ChIP	yes
Lefterova et al. 2010	mouse	Lefterova-mac-CEBPb	0.67	11.03	1.5	2	36	36	36	5,717,739	ChIP	yes
Lefterova et al. 2010	mouse	Lefterova-mac-PPARg-1	0.87	1.59	0.69	0	36	36	36	10,646,239	ChIP	yes
Lefterova et al. 2010	mouse	Lefterova-mac-PU.1	0.86	13.27	1.64	2	36	36	36	6,261,063	ChIP	yes
Tallack et al. 2010	mouse	KLF1-Input-2	0.96	9.63	0.58	0	48	48	48	10,405,126	Input	no
Tallack et al. 2010	mouse	KLF1-2	0.68	1.28	0.18	-2	33	33	33	10,757,339	ChIP	yes
Tallack et al. 2010	mouse	KLF1-3	0.55	1.39	0.47	-1	48	48	48	17,728,355	ChIP	yes
Tallack et al. 2010	mouse	KLF1-Input-3	0.65	1.23	0.4	-1	33	33	33	548,382	Input	no
Rahl et al. 2010	mouse	mES-Ctr9	0.96	1.37	0.97	0	26	26	26	5,468,214	ChIP	yes
Rahl et al. 2010	mouse	mES-NelfA	0.7	2.42	1.42	1	36	36	36	3,643,555	ChIP	yes
Rahl et al. 2010	mouse	mES-Spt5	0.95	1.59	0.94	0	26	26	26	5,595,215	ChIP	yes
Ramos et al. 2010	human	T98G-CBP-T0	0.94	1.55	0.31	-1	32	32	32	4,047,183	ChIP	yes
Ramos et al. 2010	human	T98G-CBP-T30-1	0.94	1.63	0.4	-1	32	32	32	4,885,700	ChIP	yes
Ramos et al. 2010	human	T98G-CBP-T30-2	0.84	1.75	2.14	2	32	32	32	5,034,834	ChIP	yes
Ramos et al. 2010	human	T98G-p300-T0	0.96	1.79	0.85	0	32	32	32	5,119,057	ChIP	yes
Ramos et al. 2010	human	T98G-p300-T30-1	0.97	2.22	1.19	1	32	32	32	5,191,684	ChIP	yes
Ramos et al. 2010	human	T98G-p300-T30-2	0.87	1.95	3.44	2	32	32	32	5,159,200	ChIP	yes
Kunarso et al. 2010	human	hESC-CTCF	0.92	15.2	1.62	2	37	37	37	10,828,759	ChIP	yes
Kunarso et al. 2010	human	hESC-NANOG	0.94	3.91	1.42	1	36.18	37	36	10,240,400	ChIP	yes
Kunarso et al. 2010	human	hESC-Nanog-and-CTCF-control	0.96	1.25	0.19	-2	37	37	37	8,641,430	Input	no
Kunarso et al. 2010	human	hESC-OCT4	0.98	1.94	0.42	-1	30.07	36	26	11,288,800	ChIP	yes
Kunarso et al. 2010	human	hESC-Oct4-control	0.95	1.26	0.4	-1	36	36	36	8,560,581	Input	no
Johannes et al. 2010	human	HeLa-BTAF	0.79	13.25	2.7	2	33	33	33	2,654,681	ChIP	yes
Johannes et al. 2010	human	HeLa-GAPDH	0.84	2.51	0.02	-2	33	33	33	953,719	IgG	no
Hu et al. 2010	human	MCF7-E2-ER	0.8	10.66	1.28	1	36	36	36	1,656,740	ChIP	yes
Hu et al. 2010	human	MCF7-ethl-ER	0.81	3.76	0.87	0	36	36	36	2,857,720	ChIP	unknown
Heng et al. 2010	mouse	mESC-HA-1	0.97	1.25	0.83	0	35	35	35	14,266,600	IgG	no
Heng et al. 2010	mouse	mESC-HA-Nr5a2-1	0.85	1.54	0.36	-1	35	35	35	9,395,231	ChIP	yes
Chicas et al. 2010	human	IMR90-Growing-cells-pRb-experiment-1-1	0.78	2.94	0.41	-1	36	36	36	6,181,869	ChIP	yes
Chicas et al. 2010	human	IMR90-Mock-1	0.85	2.18	0.13	-2	36	36	36	3,317,485	IgG	no
Chicas et al. 2010	human	IMR90-Quiescent-cells-p130	0.92	4.12	0.95	0	36	36	36	3,753,591	ChIP	yes
Chicas et al. 2010	human	IMR90-Quiescent-cells-pRb-experiment-1-1	0.92	3.3	0.38	-1	36	36	36	1,441,212	ChIP	yes
Chicas et al. 2010	human	IMR90-Quiescent-cells-pRb-experiment-2	0.93	2.43	0.05	-2	36	36	36	4,608,677	ChIP	yes
Chicas et al. 2010	human	IMR90-Quiescent-cells-Rb-shRNA-p130	0.78	2.2	0.24	-2	36	36	36	5,348,557	ChIP	yes
Chicas et al. 2010	human	IMR90-Quiescent-cells-Rb-shRNA-Rb	0.85	2.42	0.38	-1	36	36	36	1,114,921	ChIP	no
Chicas et al. 2010	human	IMR90-Senescent-cells-p130	0.94	4.21	1.21	1	36	36	36	4,388,261	ChIP	yes
Chicas et al. 2010	human	IMR90-Senescent-cells-pRb-experiment-1-1	0.93	3.63	0.39	-1	36	36	36	3,867,162	ChIP	yes
Chicas et al. 2010	human	IMR90-Senescent-cells-pRb-experiment-2	0.91	2.29	0.22	-2	36	36	36	4,109,281	ChIP	yes
Chicas et al. 2010	human	IMR90-Senescent-cells-Rb-shRNA-p130	0.78	3.98	1.26	1	36	36	36	4,232,255	ChIP	yes
Chicas et al. 2010	human	IMR90-Senescent-cells-Rb-shRNA-pRb	0.89	2.02	0.23	-2	36	36	36	2,813,407	ChIP	no
Martinez et al. 2010	mouse	Input	0.51	2.26	2.59	2	36	36	36	14,617,059	Input	no
Martinez et al. 2010	mouse	RAP1-ko1-RAP1	0.09	20.55	2.72	2	36	36	36	11,542,127	ChIP	no
Martinez et al. 2010	mouse	RAP1-ko2-RAP1	0.18	21.54	3.88	2	36	36	36	7,585,528	ChIP	no
Martinez et al. 2010	mouse	WT-1-RAP1	0.26	11.02	2.45	2	36	36	36	11,249,746	ChIP	yes
Martinez et al. 2010	mouse	WT-2-RAP1	0.19	11.05	2.88	2	36	36	36	11,856,303	ChIP	yes
Qi et al. 2010	human	HeLa-PHF8	0.9	1.66	0.78	0	25	25	25	9,252,893	ChIP	yes
Woodfield et al. 2010	human	MCF7-IgG-control	0.97	1.38	0.63	0	40	40	40	8,158,903	IgG	no
Woodfield et al. 2010	human	MCF7-TFAP2C	0.93	7.1	1.74	2	40	40	40	8,188,674	ChIP	yes
Kagey et al. 2010	mouse	MEF-Med12-Rep1	0.85	1.58	0.46	-1	36	36	36	8,167,440	ChIP	yes
Kagey et al. 2010	mouse	MEF-Med1-Rep1	0.94	2.03	0.82	0	36	36	36	7,326,311	ChIP	yes
Kagey et al. 2010	mouse	MEF-Smc1-Rep1	0.62	6.04	2.96	2	36	36	36	9,601,525	ChIP	yes
Kagey et al. 2010	mouse	MEF-Smc1-Rep2	0.94	1.34	0.9	0	36	36	36	22,977,719	ChIP	yes
Kagey et al. 2010	mouse	mESC-CTCF-Rep1	0.94	7.28	1.29	1	36	36	36	3,966,359	ChIP	yes
Kagey et al. 2010	mouse	mESC-CTCF-Rep2	0.85	1.73	1.4	1	36	36	36	4,953,685	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Kagey et al. 2010	mouse	mESC-Med12-051809-ChipSeq	0.84	7.16	1.9	2	36	36	36	22,763,608	ChIP	yes
Kagey et al. 2010	mouse	mESC-Med12-Rep2	0.63	1.8	1.12	1	36	36	36	12,861,074	ChIP	yes
Kagey et al. 2010	mouse	mESC-Med1-Rep1	0.92	2.27	1.68	2	36	36	36	18,346,720	ChIP	yes
Kagey et al. 2010	mouse	mESC-Med1-Rep2	0.94	1.73	1.04	1	36	36	36	18,725,724	ChIP	yes
Kagey et al. 2010	mouse	mESC-Nipbl-Rep1	0.26	1.33	0.54	0	36	36	36	6,241,538	ChIP	yes
Kagey et al. 2010	mouse	mESC-Nipbl-Rep2	0.96	1.54	0.99	0	36	36	36	12,668,428	ChIP	yes
Kagey et al. 2010	mouse	mESC-Smc1-Rep1	0.96	3.37	1.88	2	36	36	36	21,733,223	ChIP	yes
Kagey et al. 2010	mouse	mESC-Smc1-Rep2	0.95	3.29	1.57	2	36	36	36	4,936,893	ChIP	yes
Kagey et al. 2010	mouse	mESC-Smc3-Rep3	0.89	3.86	1.81	2	36	36	36	21,491,459	ChIP	yes
Kagey et al. 2010	mouse	mESC-Smc3-Rep4	0.89	4.34	2.26	2	36	36	36	21,522,393	ChIP	yes
Kagey et al. 2010	mouse	mES-WCE	0.93	1.38	0.31	-1	36	36	36	3,669,758	Input	no
Kouwenhoven et al. 2010	human	Keratinocytes-p63-1	0.96	19.62	2.46	2	32	32	32	2,722,489	ChIP	yes
Kouwenhoven et al. 2010	human	Keratinocytes-p63-2	0.96	4.51	2.51	2	32	32	32	5,588,217	ChIP	yes
Kouwenhoven et al. 2010	human	Keratinocytes-p63-3	0.81	9.05	5.19	2	35	35	35	20,435,516	ChIP	yes
Cao et al. 2010	human	RD-Input	0.82	1.34	0.33	-1	40	40	40	6,587,573	Input	no
Cao et al. 2010	human	RD-pFM2-1	0.76	1.9	0.63	0	40	40	40	8,593,218	ChIP	yes
Cao et al. 2010	human	Rh4-Input-1	0.82	1.35	0.36	-1	38.7	40	36	20,270,400	Input	no
Cao et al. 2010	human	Rh4-pFM2-1	0.74	2.33	0.9	0	38.83	40	36	20,920,563	ChIP	yes
Blow et al. 2010	mouse	Heart-p300	0.85	1.77	0.16	-2	36	36	36	1,531,274	ChIP	yes
Blow et al. 2010	mouse	Midbrain-p300	0.87	1.34	0.21	-2	36	36	36	6,406,542	ChIP	yes
Sehat et al. 2010	human	DFB-IGF1R	0.9	1.56	0.13	-2	36	36	36	3,664,071	ChIP	yes
Liu et al. 2010	human	E2F1-HeLa	0.96	1.91	0.81	0	36	36	36	8,595,301	ChIP	yes
Liu et al. 2010	human	PHF8-HeLa-unsyn	0.97	2.21	0.84	0	36	36	36	3,841,047	ChIP	yes
Liu et al. 2010	human	SMC4-HeLa-M	0.62	2.43	0.65	0	36	36	36	9,809,944	ChIP	yes
Tang et al. 2010	human	K562-PMA-Egr1	0.84	1.62	0.21	-2	33	33	33	3,581,558	ChIP	yes
Jung et al. 2010	mouse	iHoxc9-Day5	0.97	1.35	0.79	0	36	36	36	10,149,860	ChIP	yes
Jung et al. 2010	mouse	WCE-Day5	0.94	1.37	0.82	0	36	36	36	15,043,390	Input	no
Vermeulen et al. 2010	human	BAP18-GFP-HeLa-rep1	0.8	2.1	1.6	2	35	35	35	11,153,198	ChIP	yes
Vermeulen et al. 2010	human	BAP18-GFP-HeLa-rep2	0.83	1.91	2.1	2	35	35	35	28,580,771	ChIP	yes
Vermeulen et al. 2010	human	GATAD1-GFP-HeLa-rep1	0.9	1.71	2.69	2	35	35	35	5,413,596	ChIP	yes
Vermeulen et al. 2010	human	GATAD1-GFP-HeLa-rep2	0.81	2.4	1.83	2	35	35	35	12,596,319	ChIP	yes
Vermeulen et al. 2010	human	LRWD1-GFP-HeLa	0.88	2.29	0.98	0	35	35	35	11,634,470	ChIP	yes
Vermeulen et al. 2010	human	N-PAC-GFP-HeLa-rep1	0.85	2.01	2.46	2	35	35	35	5,436,726	ChIP	yes
Vermeulen et al. 2010	human	N-PAC-GFP-HeLa-rep2	0.77	2.78	2.81	2	35	35	35	12,669,139	ChIP	yes
Vermeulen et al. 2010	human	PHF8-GFP-HeLa-rep1	0.88	1.7	2.13	2	35	35	35	4,896,779	ChIP	yes
Vermeulen et al. 2010	human	PHF8-GFP-HeLa-rep2	0.83	1.79	2.2	2	35	35	35	29,180,126	ChIP	yes
Vermeulen et al. 2010	human	Sgf29-GFP-HeLa-rep1	0.87	2.36	1.45	1	35	35	35	12,636,931	ChIP	yes
Vermeulen et al. 2010	human	Sgf29-GFP-HeLa-rep2	0.87	1.84	1.98	2	35	35	35	29,275,648	ChIP	yes
Vermeulen et al. 2010	human	TRRAP-GFP-HeLa-rep1	0.74	3.43	3.94	2	35	35	35	7,851,229	ChIP	yes
Vermeulen et al. 2010	human	TRRAP-GFP-HeLa-rep2	0.87	1.71	1.76	2	35	35	35	29,410,330	ChIP	yes
Vermeulen et al. 2010	human	wt-negative-control-HeLa	0.87	1.88	1.09	1	35	35	35	10,851,096	Input	no
Chi et al. 2010	human	GIST48-ETV1	0.95	2.79	0.97	0	36	36	36	10,740,357	ChIP	yes
Chi et al. 2010	human	GIST48-Input	0.98	1.12	0.27	-1	36	36	36	15,177,140	Input	no
Chia et al. 2010	human	hESC-Input	0.98	1.48	0.82	0	35	35	35	17,097,337	Input	no
Chia et al. 2010	human	hESC-PRDM14	0.95	1.87	0.67	0	35	35	35	14,268,098	ChIP	no
Palii et al. 2010	human	Erythroid-TAL1	0.94	6.12	0.9	0	37	37	37	6,882,358	ChIP	yes
Palii et al. 2010	human	Jurkat-IgG	0.97	1.48	0.16	-2	37	37	37	4,760,148	IgG	no
Palii et al. 2010	human	Jurkat-TAL1	0.94	2.58	0.45	-1	37	37	37	6,151,678	ChIP	yes
Lee et al. 2010	human	GM06990-E2F4	0.95	1.99	0.5	0	36	36	36	2,845,819	ChIP	yes
Lee et al. 2010	human	GM06990-Input	0.98	1.09	0.1	-2	36	36	36	7,164,483	Input	no
Law et al. 2010	human	ATRX-Human-Erythroid	0.98	1.83	0.31	-1	36	36	36	1,481,778	ChIP	yes
Law et al. 2010	human	ATRX-Human-Erythroid-Input	0.86	1.69	0.44	-1	36	36	36	2,815,016	Input	no
Law et al. 2010	mouse	ATRX-Mouse-ES	0.86	1.52	0.39	-1	51	51	51	47,903,467	ChIP	yes
Law et al. 2010	mouse	ATRX-Mouse-ES-Input	0.46	1.61	2.14	2	51	51	51	24,366,842	Input	no
Yao et al. 2010	human	HeLa-Input	0.97	1.17	0.51	0	36	36	36	44,239,692	Input	no
Yao et al. 2010	human	HeLa-p68	0.88	1.39	0.53	0	36	36	36	29,417,892	ChIP	yes
Verzi et al. 2010	human	Caco2-differentiated-CDX2	0.86	3.41	1.53	2	40	40	40	12,916,083	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Verzi et al. 2010	human	Caco2-differentiated-GATA6	0.86	2.75	0.73	0	40	40	40	14,079,635	ChIP	unknown
Verzi et al. 2010	human	Caco2-differentiated-HNF4A	0.9	6.15	1.42	1	40	40	40	5,599,576	ChIP	yes
Verzi et al. 2010	human	Caco2-Input	0.79	2.68	1.27	1	40	40	40	10,777,726	Input	no
Verzi et al. 2010	human	Caco2-proliferating-CDX2	0.87	3.04	1.24	1	40	40	40	11,527,010	ChIP	unknown
Verzi et al. 2010	human	Caco2-proliferating-GATA6	0.77	12.01	1.6	2	40	40	40	7,337,182	ChIP	yes
Verzi et al. 2010	human	Caco2-proliferating-HNF4A	0.76	6.64	1.55	2	40	40	40	9,186,141	ChIP	yes
Barish et al. 2010	mouse	macrophage-BCL6	0.73	3.1	1.36	1	42	42	42	14,741,775	ChIP	yes
Barish et al. 2010	mouse	macrophage-BCL6-LPS-1	0.59	2.47	1.38	1	36	36	36	12,161,935	ChIP	unknown
Barish et al. 2010	mouse	macrophage-BCL6-LPS-2	0.52	1.99	1.02	1	42	42	42	19,613,630	ChIP	unknown
Barish et al. 2010	mouse	macrophage-Bcl6-REP2	0.43	8.05	3.11	2	36	36	36	10,772,781	ChIP	yes
Barish et al. 2010	mouse	macrophage-IgG	0.59	3.32	1.14	1	36	36	36	11,046,455	IgG	no
Barish et al. 2010	mouse	macrophage-Input	0.96	1.33	0.5	-1	36	36	36	14,265,664	Input	no
Barish et al. 2010	mouse	macrophage-p65	0.85	3.3	2.23	2	42	42	42	13,878,454	ChIP	no
Barish et al. 2010	mouse	macrophage-p65-LPS-1	0.71	2.6	1.23	1	43	43	43	12,731,143	ChIP	yes
Barish et al. 2010	mouse	macrophage-p65-LPS-2	0.75	1.95	1.2	1	42	42	42	10,819,755	ChIP	yes
Mahony et al. 2010	mouse	HBG3-RAR-Day2+8hrsRA-1	0.67	1.6	0.6	0	26	26	26	16,947,890	ChIP	yes
Mahony et al. 2010	mouse	HBG3-RAR-Day2-1	0.7	1.87	0.76	0	26	26	26	19,693,750	ChIP	yes
Mahony et al. 2010	mouse	HBG3-WCE-Day2	0.92	1.52	0.15	-2	26	26	26	2,570,671	Input	no
Mahony et al. 2010	mouse	HBG3-WCE-Day3	0.94	1.44	0.2	-2	26	26	26	3,038,741	Input	no
Yu et al. 2010	human	HPC-GABPa	0.73	7	1.7	2	24	24	24	3,036,253	ChIP	yes
Yu et al. 2010	human	HPC-IgG	0.45	15.8	2.15	2	25	25	25	2,762,252	IgG	no
Rada-Iglesias et al. 2010	human	ESC-BRG1	0.95	1.66	1.37	1	36	36	36	16,085,353	ChIP	yes
Rada-Iglesias et al. 2010	human	ESC-input	0.95	1.33	0.98	0	36	36	36	14,508,164	Input	no
Rada-Iglesias et al. 2010	human	ESC-p300	0.92	2.01	2.76	2	36	36	36	12,822,655	ChIP	yes
Rada-Iglesias et al. 2010	human	NEC-input	0.97	1.34	0.65	0	36	36	36	21,774,646	Input	no
Rada-Iglesias et al. 2010	human	NEC-p300	0.94	1.71	0.58	0	36	36	36	13,264,013	ChIP	yes
Gu et al. 2010	human	MCF7-control-ERa	0.79	1.57	0.24	-2	36	36	36	4,385,795	ChIP	no
Gu et al. 2010	human	MCF7-E2-ERa	0.8	1.81	0.28	-1	36	36	36	5,785,635	ChIP	yes
Ma et al. 2010	mouse	mESC-FLAG-HA	0.93	1.59	0.59	0	36	36	36	6,257,485	IgG	no
Ma et al. 2010	mouse	mESC-Input	0.9	1.36	1.23	1	36	36	36	8,480,128	Input	no
Ma et al. 2010	mouse	mESC-Prdm14	0.84	5.1	2.59	2	36	36	36	10,899,040	ChIP	yes
Schlesinger et al. 2010	mouse	HL1-SRF	0.93	1.53	1.12	1	36	36	36	5,086,170	ChIP	yes
Li et al. 2010	mouse	Lin–Gata2	0.85	1.89	0.88	0	25	25	25	7,512,398	ChIP	yes
Li et al. 2010	mouse	Lin–IgG	0.75	1.87	0.2	-2	25	25	25	3,211,969	IgG	no
Li et al. 2010	mouse	Lin–Ldb1	0.71	5.7	1.95	2	25	25	25	4,251,705	ChIP	yes
Li et al. 2010	mouse	Lin–Tal1	0.81	4.15	1.77	2	36	36	36	11,482,776	ChIP	yes
Kong et al. 2010	human	ECC1-E2-ERa	0.95	1.65	0.37	-1	31.26	36	26	7,178,094	ChIP	yes
Kong et al. 2010	human	ECC1-EtOH-ERa	0.94	1.35	0.21	-2	26	26	26	11,049,926	ChIP	no
Kong et al. 2010	human	ECC1-Input	0.98	1.19	0.19	-2	30.16	36	26	7,631,501	Input	no
Kong et al. 2010	human	Ishikawa-E2-ERa	0.97	1.42	0.41	-1	30.73	36	26	10,438,320	ChIP	yes
Kong et al. 2010	human	Ishikawa-EtOH-ERa	0.97	1.38	0.36	-1	30.55	36	26	10,175,702	ChIP	no
Kong et al. 2010	human	Ishikawa-Input	0.98	1.19	0.37	-1	26	26	26	21,437,974	Input	no
Kong et al. 2010	human	MCF7-E2-ERa	0.95	5.96	1.63	2	26	26	26	9,652,711	ChIP	yes
Kong et al. 2010	human	MCF7-EtOH-ERa	0.97	1.23	0.4	-1	33.21	36	26	14,488,769	ChIP	no
Kong et al. 2010	human	MCF7-Input	0.97	1.17	0.18	-2	26	26	26	8,379,328	Input	no
Kong et al. 2010	human	T47D-E2-ERa	0.97	2.84	1.13	1	26	26	26	10,608,916	ChIP	yes
Kong et al. 2010	human	T47D-EtOH-ERa	0.97	1.21	0.29	-1	33.66	36	26	13,430,557	ChIP	no
Kong et al. 2010	human	T47D-Input	0.96	1.18	0.19	-2	26	26	26	12,933,672	Input	no
Yang et al. 2010	human	MCF7-IgG	0.1	14.93	1.71	2	36	36	36	5,500,498	IgG	no
Yang et al. 2010	human	MCF7-TDRD3	0.34	3.12	1.25	1	36	36	36	27,147,620	ChIP	yes
Fang et al. 2011	human	LN229-IgG	0.6	2.01	1.53	2	40	40	40	455,630	IgG	no
Fang et al. 2011	human	LN229-Sox2	0.65	1.8	1.04	1	40	40	40	932,166	ChIP	yes
van Heeringen et al. 2011	xaenopus	TBP-ChIPSeq	0.83	N/A	N/A	N/A	32	32	32	6,569,902	ChIP	yes
GSE26680	mouse	mES-MCAF1	0.86	1.63	1.04	1	36	36	36	13,907,040	ChIP	yes
GSE26680	mouse	mES-REST	0.86	2.61	1.01	1	26	26	26	4,159,486	ChIP	yes
GSE26680	mouse	mES-Ring1b	0.92	1.23	0.43	-1	26	26	26	3,785,138	ChIP	yes
Teo et al. 2011	human	hESC-48h-endodiff-EOMES-XL-eps1and2	0.74	4.08	1.67	2	36	36	36	33,687,700	ChIP	yes
Teo et al. 2011	human	hESC-Input-XL	0.98	1.37	0.44	-1	36	36	36	7,422,963	Input	no
Joseph et al. 2011; Kong et al. 2010	human	MCF7-DMSO-cFos-1	0.93	1.31	0.43	-1	36	36	36	18,781,755	ChIP	yes
Joseph et al. 2011; Kong et al. 2010	human	MCF7-DMSO-cJun-1	0.96	1.31	0.37	-1	36	36	36	14,827,454	ChIP	yes
Joseph et al. 2011; Kong et al. 2010	human	MCF7-DMSO-FOXA1-1	0.95	1.51	0.63	0	36	36	36	14,414,733	ChIP	yes
Joseph et al. 2011; Kong et al. 2010	human	MCF7-E2-cFos-1	0.95	1.48	0.68	0	31	31	31	12,684,762	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Joseph et al. 2011; Kong et al. 2010	human	MCF7-E2-cJun-1	0.95	1.31	0.4	-1	36	36	36	18,012,142	ChIP	yes
Joseph et al. 2011; Kong et al. 2010	human	MCF7-E2-FOXA1-1	0.93	2.52	1.07	1	36	36	36	15,884,461	ChIP	yes
Joseph et al. 2011; Kong et al. 2010	human	T47D-DMSO-FOXA1	0.94	5.13	1.78	2	36	36	36	14,981,282	ChIP	yes
Joseph et al. 2011; Kong et al. 2010	human	T47D-E2-FOXA1-1	0.94	2.27	0.92	0	36	36	36	11,819,434	ChIP	yes
Novershtern et al. 2011	human	HSPC-Ikaros-rep1	0.74	2.04	0.36	-1	36	36	36	3,228,102	ChIP	yes
Novershtern et al. 2011	human	HSPC-Ikaros-rep2	0.78	1.98	0.31	-1	36	36	36	2,635,528	ChIP	yes
Novershtern et al. 2011	human	HSPC-MEIS1-rep1	0.46	1.85	0.34	-1	36	36	36	9,565,937	ChIP	yes
Novershtern et al. 2011	human	HSPC-MEIS1-rep2	0.29	1.84	0.36	-1	36	36	36	12,342,658	ChIP	yes
Novershtern et al. 2011	human	HSPC-MEIS1-rep3	0.46	1.78	0.34	-1	36	36	36	10,465,042	ChIP	yes
Novershtern et al. 2011	human	HSPC-Pu.1-rep1	0.6	3.44	1.05	1	36	36	36	4,940,474	ChIP	yes
Novershtern et al. 2011	human	HSPC-Pu.1-rep2	0.61	3.19	1.03	1	36	36	36	4,617,421	ChIP	yes
Novershtern et al. 2011	human	HSPC-TAL1-rep1	0.69	1.31	0.18	-2	36	36	36	8,788,837	ChIP	yes
Novershtern et al. 2011	human	HSPC-TAL1-rep2	0.72	1.36	0.19	-2	36	36	36	7,439,145	ChIP	yes
Novershtern et al. 2011	human	HSPC-WCE	0.95	1.23	0.24	-2	36	36	36	6,321,189	Input	no
GSE23581	mouse	mES-Acitvin-Input	0.93	1.44	0.56	0	35	35	35	9,674,331	Input	no
GSE23581	mouse	mES-Acitvin-pSmad2	0.81	1.98	1.31	1	35	35	35	11,730,560	ChIP	yes
GSE23581	mouse	mES-DMSO-Input	0.93	1.51	0.69	0	35	35	35	10,750,428	Input	no
GSE23581	mouse	mES-DMSO-pSmad2	0.77	2.26	1.47	1	35	35	35	11,288,314	ChIP	no
GSE23581	mouse	mES-SP-Input	0.92	1.65	0.73	0	35	35	35	9,325,370	Input	no
GSE23581	mouse	mES-SP-pSmad2	0.8	2.16	1.08	1	35	35	35	9,079,108	ChIP	no
GSE26136	mouse	mES-Dpy-30	0.69	1.69	1.54	2	36	36	36	24,620,668	ChIP	yes
Klisch et al. 2011	mouse	Cerebella-Atoh1.control	0.95	2.28	0.97	0	35	35	35	10,310,101	Input	no
Klisch et al. 2011	mouse	Cerebella-Atoh1.rep1	0.89	7.47	1.86	2	35	35	35	2,649,698	ChIP	yes
Klisch et al. 2011	mouse	Cerebella-Atoh1.rep2	0.92	3.36	0.96	0	35	35	35	7,166,233	ChIP	yes
Klisch et al. 2011	mouse	Cerebella-IgG.s-5	0.69	2.33	1.44	1	36	36	36	8,514,915	IgG	no
Yang et al. 2011	mouse	WTTh17STAT3	0.73	4.58	1.63	2	25	25	25	28,501,100	ChIP	yes
Yang et al. 2011	mouse	WTTh17STAT5	0.58	6.08	1.65	2	25	25	25	30,799,471	ChIP	yes
Ebert et al. 2011; McManus et al. 2011	mouse	DP-Tcell-CTCF	0.35	5.94	5.08	2	36	36	36	13,326,337	ChIP	yes
Ebert et al. 2011; McManus et al. 2011	mouse	Mature-Bcell-CTCF	0.66	4.24	6.97	2	36	36	36	14,505,107	ChIP	yes
Ebert et al. 2011; McManus et al. 2011	mouse	Pro-Bcell-Rad21	0.84	6.82	3.35	2	36	36	36	25,074,201	ChIP	yes
Ebert et al. 2011; McManus et al. 2011	mouse	Pro-Bcell-Rag2KO-CTCF	0.62	4.52	3.98	2	36	36	36	15,641,228	ChIP	yes
Zhao et al. 2011	mouse	Myb-activated-B1T1	0.81	3.23	0.98	0	36	36	36	7,467,313	ChIP	yes
Zhao et al. 2011	mouse	Myb-activated-B1T2	0.83	3.33	0.88	0	36	36	36	5,454,019	ChIP	yes
Zhao et al. 2011	mouse	Myb-activated-B2	0.86	3.03	1.07	1	36	36	36	7,331,382	ChIP	yes
Zhao et al. 2011	mouse	Myb-activated-IgG	0.73	1.98	0.58	0	36	36	36	6,724,529	IgG	no
Zhao et al. 2011	mouse	Myb-inactivated-B1	0.83	1.79	0.5	-1	36	36	36	5,729,128	ChIP	no
Zhao et al. 2011	mouse	Myb-inactivated-B2	0.87	1.79	0.51	0	36	36	36	5,734,826	ChIP	no
Zhao et al. 2011	mouse	Myb-inactivated-IgG	0.94	4.16	0.34	-1	36	36	36	766,809	IgG	no
Rey et al. 2011	mouse	BMAL1-ZT02-rep1	0.8	2.4	1.15	1	37	37	37	9,023,818	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT02-rep2	0.74	1.76	1.5	1	37	37	37	24,294,126	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT06-rep1	0.66	2.1	1.71	2	37	37	37	20,808,528	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT06-rep2	0.25	5.49	1.89	2	37	37	37	20,234,777	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT10-rep1	0.9	1.75	0.92	0	37	37	37	9,220,495	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT10-rep2	0.84	1.74	1.53	2	37	37	37	22,892,744	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT14-rep1	0.88	1.61	0.84	0	37	37	37	12,447,404	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT14-rep2	0.64	2.91	2.58	2	37	37	37	20,961,930	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT18-rep1	0.7	2.03	1.77	2	37	37	37	22,079,073	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT18-rep2	0.21	4.51	1.94	2	37	37	37	30,803,372	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT22-rep1	0.67	2.77	1.33	1	37	37	37	10,194,650	ChIP	unknown
Rey et al. 2011	mouse	BMAL1-ZT22-rep2	0.88	1.4	0.85	0	37	37	37	21,473,568	ChIP	unknown
Rey et al. 2011	mouse	Input-DNA	0.84	1.76	1.44	1	37	37	37	21,940,254	Input	no
Koeppel et al. 2011	human	Saos-2-ChIP-Input-control	0.96	1.2	0.44	-1	35	35	35	15,967,510	Input	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Koeppel et al. 2011	human	Saos-2-p53-replicate1-1	0.86	2.41	1	1	35	35	35	14,932,313	ChIP	yes
Koeppel et al. 2011	human	Saos-2-p53-replicate2	0.95	2.26	1.68	2	35	35	35	14,969,104	ChIP	yes
Koeppel et al. 2011	human	Saos-2-TAp73alpha-replicate1	0.84	2.57	2.6	2	35	35	35	14,905,593	ChIP	yes
Koeppel et al. 2011	human	Saos-2-TAp73alpha-replicate2	0.93	2.05	1.34	1	35	35	35	14,626,232	ChIP	yes
Koeppel et al. 2011	human	Saos-2-TAp73beta-replicate1	0.96	5.1	4.12	2	32	32	32	4,927,558	ChIP	yes
Koeppel et al. 2011	human	Saos-2-TAp73beta-replicate2	0.94	6.37	2.61	2	35	35	35	16,272,496	ChIP	yes
He et al. 2011	mouse	HL1-BirA-control-1	0.3	8.52	9.14	2	40	40	40	15,388,943	IgG	no
He et al. 2011	mouse	HL1-Gata4-1	0.57	2.11	2.92	2	37.77	40	35	21,352,298	ChIP	yes
He et al. 2011	mouse	HL1-Input-control	0.92	1.94	1.84	2	36	36	36	13,770,246	Input	no
He et al. 2011	mouse	HL1-Mef2a-1	0.92	1.34	1	1	38.18	40	35	20,160,274	ChIP	yes
He et al. 2011	mouse	HL1-Nkx2-5-1	0.91	1.69	3.09	2	38.59	40	36	24,181,076	ChIP	yes
He et al. 2011	mouse	HL1-P300	0.9	1.83	1.28	1	36	36	36	15,446,431	ChIP	yes
He et al. 2011	mouse	HL1-Srf-1	0.8	1.81	2.37	2	38.29	40	36	25,881,877	ChIP	yes
He et al. 2011	mouse	HL1-Tbx5-1	0.9	2.1	2.13	2	37.91	40	36	11,074,980	ChIP	yes
Bugge et al. 2011	mouse	Liver-HDAC3-ZT10	0.49	2.56	1.67	2	36.93	38	36	38,211,882	ChIP	unknown
Bugge et al. 2011	mouse	Liver-HDAC3-ZT22	0.76	1.34	1.09	1	39.03	40	38	38,822,996	ChIP	unknown
Bugge et al. 2011	mouse	Liver-input-Mnase-ZT10	0.81	4.86	1.96	2	38	38	38	20,627,184	Input	no
Bugge et al. 2011	mouse	Liver-input-Mnase-ZT22	0.8	4.49	1.52	2	38	38	38	18,828,586	Input	no
Bugge et al. 2011	mouse	Liver-input-ZT10	0.58	1.35	1.13	1	40	40	40	18,254,032	Input	no
Bugge et al. 2011	mouse	Liver-input-ZT22	0.64	2.1	2.06	2	40	40	40	14,072,057	Input	no
Bugge et al. 2011	mouse	Liver-NCoR-ZT10	0.72	2.5	1.24	1	38	38	38	10,955,647	ChIP	unknown
Bugge et al. 2011	mouse	Liver-NCoR-ZT22	0.8	1.4	0.86	0	38	38	38	18,218,400	ChIP	unknown
Bugge et al. 2011	mouse	Liver-Rev-erba-ZT10	0.54	3.5	1.52	2	36	36	36	23,266,910	ChIP	unknown
Bugge et al. 2011	mouse	Liver-Rev-erba-ZT22	0.38	1.8	0.87	0	36	36	36	26,701,376	ChIP	unknown
Siersbæk et al. 2011	mouse	CEBPbeta-2-hours	0.86	4.2	4.51	2	36	36	36	13,391,765	ChIP	yes
Siersbæk et al. 2011	mouse	CEBPbeta-4-hours	0.82	3.94	5.44	2	36	36	36	14,184,719	ChIP	yes
Siersbæk et al. 2011	mouse	CEBPbeta-day-0	0.69	4.74	5.95	2	36	36	36	13,823,228	ChIP	yes
Siersbæk et al. 2011	mouse	CEBPbeta-day-2	0.77	3.59	2.97	2	36	36	36	11,535,365	ChIP	yes
Siersbæk et al. 2011	mouse	CEBPdelta-4-hours	0.63	6.24	3.44	2	40	40	40	11,803,122	ChIP	yes
Siersbæk et al. 2011	mouse	CEBPdelta-day-0	0.59	5.05	4.73	2	40	40	40	12,036,027	ChIP	yes
Siersbæk et al. 2011	mouse	GR-4-hours	0.41	3.12	1.97	2	24	24	24	9,694,597	ChIP	yes
Siersbæk et al. 2011	mouse	Input	0.92	1.27	0.71	0	36	36	36	12,904,842	Input	no
Siersbæk et al. 2011	mouse	PPARgamma-day-2	0.29	2.31	0.96	0	40	40	40	13,429,961	ChIP	yes
Siersbæk et al. 2011	mouse	PPARgamma-day-6	0.3	2.28	1.64	2	40	40	40	14,620,856	ChIP	yes
Siersbæk et al. 2011	mouse	RXR-4-hours	0.7	3.12	2.85	2	40	40	40	12,219,467	ChIP	yes
Siersbæk et al. 2011	mouse	Stat5a-4-hours	0.62	4.33	5.62	2	36	36	36	13,644,334	ChIP	yes
Smeenk et al. 2011	human	U2OS-p53-ActD	0.97	5.72	1.81	2	32	32	32	6,940,755	ChIP	yes
Smeenk et al. 2011	human	U2OS-p53-Eto	0.97	4.28	1.27	1	32	32	32	7,272,634	ChIP	yes
Smeenk et al. 2011	human	U2OS-p53-pS15-ActD	0.96	1.74	0.54	0	32	32	32	4,742,221	ChIP	yes
Smeenk et al. 2011	human	U2OS-p53-pS15-Eto	0.94	1.8	0.66	0	32	32	32	6,590,995	ChIP	yes
Smeenk et al. 2011	human	U2OS-p53-pS46-ActD	0.92	1.63	0.29	-1	32	32	32	5,408,031	ChIP	yes
Smeenk et al. 2011	human	U2OS-p53-pS46-Eto	0.94	1.84	0.47	-1	32	32	32	5,748,594	ChIP	yes
Ceol et al. 2011	human	WM262-MCAF1	0.45	3.9	1.5	1	36	36	36	13,346,938	ChIP	yes
Ceol et al. 2011	human	WM262-SetDB1	0.54	1.49	0.62	0	36	36	36	5,307,748	ChIP	yes
Ceol et al. 2011	human	Wm451-lu-SetDB1	0.39	1.39	0.67	0	36	36	36	6,295,121	ChIP	yes
Wu et al. 2011a; Wu et al. 2011b	mouse	mES-IgG-exp1-no-KD	0.67	1.62	1.07	1	25	25	25	23,036,303	IgG	no
Wu et al. 2011a; Wu et al. 2011b	mouse	mES-IgG-exp2-mock-KD	0.52	2.04	1.4	1	36	36	36	8,283,677	IgG	no
Wu et al. 2011a; Wu et al. 2011b	mouse	mES-Tet1-exp1-no-KD	0.91	1.33	1.25	1	25	25	25	28,536,436	ChIP	yes
Wu et al. 2011a; Wu et al. 2011b	mouse	mES-Tet1-exp2-mock-KD	0.89	1.54	1	0	36	36	36	9,521,384	ChIP	yes
Wu et al. 2011a; Wu et al. 2011b	mouse	mES-Tet1-exp3-Tet1-KD	0.9	1.33	1.06	1	25	25	25	28,297,858	ChIP	no
Wu et al. 2011a; Wu et al. 2011b	mouse	mES-Tet1-exp4-Tet1-KD	0.56	1.71	0.5	0	25	25	25	4,286,921	ChIP	no
Horiuchi et al. 2011	mouse	Th1-1-Input	0.84	1.4	0.36	-1	36	36	36	5,981,126	Input	no
Horiuchi et al. 2011	mouse	Th1-2-Input	0.88	2.25	1.22	1	36	36	36	4,609,331	Input	no
Horiuchi et al. 2011	mouse	Th1-GATA3	0.94	1.81	0.62	0	36	36	36	7,219,267	ChIP	no
Horiuchi et al. 2011	mouse	Th1-IgG	0.78	2.37	1.02	1	36	36	36	11,136,690	IgG	no
Horiuchi et al. 2011	mouse	Th2-1-Input	0.81	1.58	0.42	-1	36	36	36	5,428,949	Input	no
Horiuchi et al. 2011	mouse	Th2-2-Input	0.88	2.28	1.33	1	36	36	36	4,095,787	Input	no
Horiuchi et al. 2011	mouse	Th2-GATA3	0.91	1.49	0.41	-1	36	36	36	6,332,390	ChIP	yes
Horiuchi et al. 2011	mouse	Th2-IgG	0.77	1.26	0.25	-2	36	36	36	11,532,524	IgG	no
Soccio et al. 2011	human	Human-Adipocytes-Input-rep1-GAII	0.38	1.43	0.56	0	38	38	38	23,331,240	Input	no
Soccio et al. 2011	human	Human-Adipocytes-PPARg-rep1-GAII	0.21	2.69	0.65	0	38	38	38	17,934,158	ChIP	yes
Soccio et al. 2011	human	Human-Adipocytes-PPARg-rep2-GAII	0.71	1.95	0.56	0	40	40	40	19,418,441	ChIP	yes
Soccio et al. 2011	human	Human-Liver-FOXA2-rep1-GAI-1	0.89	2.46	0.26	-1	32	32	32	3,597,158	ChIP	yes
Soccio et al. 2011	human	Human-Liver-FOXA2-rep1-GAII	0.87	1.84	0.29	-1	36	36	36	6,591,761	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Soccio et al. 2011	human	Human-Liver-FOXA2-rep2-GAI-1	0.89	2.41	0.26	-1	32	32	32	3,559,308	ChIP	yes
Soccio et al. 2011	human	Human-Liver-FOXA2-rep2-GAII	0.8	2.73	0.44	-1	36	36	36	6,415,023	ChIP	yes
Soccio et al. 2011	human	Human-Liver-Input-rep1-GAI	0.98	1.57	0.13	-2	36	36	36	4,853,927	Input	no
Soccio et al. 2011	human	Human-Liver-Input-rep1-GAII	0.96	1.27	0.11	-2	32	32	32	2,775,576	Input	no
Soccio et al. 2011	human	Human-Liver-Input-rep2-GAI	0.96	1.49	0.11	-2	32	32	32	2,636,496	Input	no
Soccio et al. 2011	mouse	Mouse-Adipocytes-PPARg-rep2-GAII	0.68	1.69	0.91	0	36	36	36	16,907,011	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-FOXA2-rep1-GAI	0.87	3.47	0.35	-1	36	36	36	2,288,906	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-FOXA2-rep2-GAI	0.77	4.67	0.6	0	36	36	36	2,986,172	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-FOXA2-rep3-GAI	0.56	3.79	0.46	-1	36	36	36	7,770,167	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-FOXA2-rep3-GAII	0.81	3.35	0.83	0	36	36	36	2,686,815	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-FOXA2-rep4-GAI	0.46	3.39	0.32	-1	36	36	36	7,311,631	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-FOXA2-rep4-GAII	0.84	5.04	1.06	1	36	36	36	1,701,117	ChIP	yes
Soccio et al. 2011	mouse	Mouse-Liver-Input-rep1-GAI	0.91	2.15	0.32	-1	36	36	36	3,658,469	Input	no
Soccio et al. 2011	mouse	Mouse-Liver-Input-rep2-GAI	0.91	2.34	0.4	-1	36	36	36	3,808,896	Input	no
Soccio et al. 2011	mouse	Mouse-Liver-Input-rep3-GAI	0.89	2.36	0.4	-1	36	36	36	3,849,533	Input	no
Ang et al. 2011	mouse	CCE-mES-Input	0.94	1.18	0.37	-1	36	36	36	20,085,978	Input	no
Ang et al. 2011	mouse	CCE-mES-Negative	0.78	4.01	1.06	1	36	36	36	5,894,488	Input	no
Ang et al. 2011	mouse	CCE-mES-Oct4	0.97	1.58	0.55	0	36	36	36	4,368,039	ChIP	yes
Ang et al. 2011	mouse	CCE-mES-Rbbp5	0.85	1.1	0.14	-2	36	36	36	20,687,485	ChIP	yes
Ang et al. 2011	mouse	CCE-mES-Wdr5	0.39	2.95	1.05	1	36	36	36	18,192,088	ChIP	yes
Ang et al. 2011	mouse	CCE-mES-WDR5-FL	0.93	2.09	0.89	0	36	36	36	9,435,450	ChIP	yes
Verzi et al. 2011	mouse	Jejunum-villus-cells-Cdx2	0.87	3.18	1.4	1	40	40	40	5,335,016	ChIP	yes
Verzi et al. 2011	mouse	Jejunum-villus-cells-Input	0.91	1.92	0.76	0	40	40	40	5,579,736	Input	no
Wang et al. 2011	human	LNCaP-AR-dht-siCTRL	0.94	1.31	0.3	-1	25	25	25	12,537,593	ChIP	yes
Wang et al. 2011	human	LNCaP-AR-dht-siFoxA1	0.96	2.25	1.64	2	25	25	25	7,690,074	ChIP	unknown
Wang et al. 2011	human	LNCaP-FoxA1-dht-siCTRL	0.95	5.46	5.19	2	22	22	22	7,796,027	ChIP	yes
Wang et al. 2011	human	LNCaP-FoxA1-vehicle-siCTRL	0.95	3.88	3	2	22	22	22	7,780,805	ChIP	yes
Wang et al. 2011	human	LNCaP-input-dht-1	0.98	1.48	0.58	0	36	36	36	4,211,736	Input	no
Wang et al. 2011	human	LNCaP-MED12-dht-siCTRL	0.98	1.53	0.45	-1	36	36	36	4,305,257	ChIP	yes
Wang et al. 2011	human	LNCaP-MED12-dht-siFoxA1	0.96	1.55	0.46	-1	28	28	28	17,506,375	ChIP	yes
Wang et al. 2011	human	LNCaP-p300-dht-siCTRL	0.98	1.59	0.51	0	36	36	36	3,133,925	ChIP	yes
Wang et al. 2011	human	LNCaP-p300-dht-siFoxA1	0.98	1.62	0.53	0	36	36	36	3,120,380	ChIP	yes
Nitzsche et al. 2011	mouse	mESC-CTCF-GFP	0.95	2.74	2.78	2	35	35	35	9,433,929	ChIP	yes
Nitzsche et al. 2011	mouse	mESC-IgG	0.94	1.92	0.8	0	35	35	35	9,008,251	IgG	no
Nitzsche et al. 2011	mouse	EB-Rad21-GFP	0.93	2.51	1.93	2	35	35	35	9,039,705	ChIP	yes
Nitzsche et al. 2011	mouse	EB-Rad21-GFP-IgG	0.91	2.25	1.19	1	35	35	35	8,488,336	IgG	no
Nitzsche et al. 2011	mouse	mESC-Rad21-GFP	0.92	2.38	2.86	2	35	35	35	20,118,696	ChIP	yes
Nitzsche et al. 2011	mouse	mESC-Rad21-GFP-IgG	0.91	2.12	1.86	2	35	35	35	18,171,398	IgG	no
Kim et al. 2011	human	Endoderm-FOXH1-pool	0.95	6.88	1.72	2	36	36	36	11,630,871	ChIP	yes
Kim et al. 2011	human	Endoderm-Input	0.97	1.45	0.69	0	36	36	36	16,775,681	Input	no
Kim et al. 2011	human	Endoderm-SMAD2-3-A-pool	0.98	1.69	0.71	0	36	36	36	10,591,855	ChIP	yes
Kim et al. 2011	human	Endoderm-SMAD2-3-B-rep1	0.98	2.65	1.02	1	36	36	36	6,467,438	ChIP	yes
Kim et al. 2011	human	Endoderm-SMAD3-rep1	0.98	1.86	0.63	0	36	36	36	6,664,422	ChIP	yes
Kim et al. 2011	human	Endoderm-SMAD4-rep1	0.98	2.4	0.98	0	36	36	36	6,664,039	ChIP	yes
Kim et al. 2011	human	hESC-FOXH1-pool-1	0.97	3.21	1.37	1	36	36	36	11,570,426	ChIP	yes
Kim et al. 2011	human	hESC-Input-1	0.65	1.53	0.51	0	36	36	36	30,699,298	Input	no
Kim et al. 2011	human	hESC-SMAD2-3-A-pool	0.98	1.88	0.9	0	36	36	36	11,364,210	ChIP	yes
Kim et al. 2011	human	hESC-SMAD2-3-B-rep1	0.97	2.3	0.93	0	36	36	36	9,667,298	ChIP	yes
Kim et al. 2011	human	hESC-SMAD3-rep1	0.98	1.78	0.81	0	36	36	36	7,743,314	ChIP	yes
Kim et al. 2011	human	hESC-SMAD4-rep1	0.96	1.89	0.75	0	36	36	36	10,007,703	ChIP	yes
Lo et al. 2011	human	Adipocytes-CEBPa	0.4	15.56	0.55	0	35	35	35	1,285,131	ChIP	yes
Lo et al. 2011	human	Adipocytes-E2F4	0.89	13.68	0.2	-2	35	35	35	64,667	ChIP	yes
Lo et al. 2011	human	Adipocytes-HSF1	0.74	6.03	0.05	-2	35	35	35	177,695	ChIP	yes
Lo et al. 2011	human	Adipocytes-IgG	0.71	21.81	0.12	-2	35	35	35	282,753	IgG	no
Tijssen et al. 2011	human	Megakaryocytes-FLI1	0.95	2.17	0.9	0	54	54	54	12,154,848	ChIP	yes
Tijssen et al. 2011	human	Megakaryocytes-GATA1	0.92	2.75	1.05	1	37	37	37	12,848,211	ChIP	yes
Tijssen et al. 2011	human	Megakaryocytes-GATA2	0.95	2.3	0.83	0	54	54	54	8,984,141	ChIP	yes
Tijssen et al. 2011	human	Megakaryocytes-rIgG	0.68	2.11	0.89	0	37	37	37	13,241,658	IgG	no
Tijssen et al. 2011	human	Megakaryocytes-RUNX1	0.97	8.42	2.6	2	54	54	54	10,822,021	ChIP	yes
Tijssen et al. 2011	human	Megakaryocytes-SCL	0.96	1.34	0.26	-1	54	54	54	11,782,604	ChIP	yes
Tan et al. 2011	human	MCF7-E2-AP2g	0.94	3.36	1.94	2	36	36	36	13,328,869	ChIP	yes
Tan et al. 2011	human	MCF7-E2-FoxA1	0.93	6.24	2.32	2	36	36	36	14,308,936	ChIP	yes
Tan et al. 2011	human	MCF7-EtOH-AP2g	0.95	3.31	2.03	2	36	36	36	13,306,339	ChIP	yes
Tan et al. 2011	human	MCF7-EtOH-FoxA1	0.92	6.76	2.13	2	36	36	36	17,586,631	ChIP	yes
Handoko et al. 2011	mouse	E14-mES-CTCF	0.8	26.13	2.04	2	37	37	37	14,006,006	ChIP	yes
Handoko et al. 2011	mouse	E14-mES-Input	0.96	1.17	0.18	-2	37	37	37	9,567,449	Input	no
Handoko et al. 2011	mouse	E14-mES-LaminB	0.89	1.45	0.8	0	36	36	36	15,336,482	ChIP	yes
Handoko et al. 2011	mouse	E14-mES-p300	0.96	1.35	0.73	0	37	37	37	17,677,307	ChIP	yes
Hu et al. 2011	human	CD34-WT-Brg1-1	0.9	2	0.77	0	25	25	25	6,821,309	ChIP	yes
Hu et al. 2011	human	CD34-WT-CTCF	0.93	2.76	0.96	0	25	25	25	6,413,538	ChIP	yes
Hu et al. 2011	human	CD34-WT-input	0.94	1.85	0.44	-1	25	25	25	3,838,343	Input	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Hu et al. 2011	human	CD34-WT-TAL1	0.87	1.69	0.23	-2	25	25	25	3,089,084	ChIP	yes
Hu et al. 2011	human	CD36-shBrg1-CTCF	0.9	12.03	2.57	2	24.63	25	24	10,427,559	ChIP	yes
Hu et al. 2011	human	CD36-shbrg1-GATA1	0.92	7.21	1.78	2	32.67	35	25	10,380,913	ChIP	yes
Hu et al. 2011	human	CD36-shBrg1-input	0.8	1.34	0.46	-1	25	25	25	8,880,654	Input	no
Hu et al. 2011	human	CD36-shbrg1-TAL1	0.97	10.44	1.8	2	25	25	25	10,119,729	ChIP	yes
Hu et al. 2011	human	CD36-shLuc-CTCF	0.81	12.76	2.53	2	25	25	25	9,434,898	ChIP	yes
Hu et al. 2011	human	CD36-shLuc-GATA1	0.89	6.8	1.84	2	25	25	25	13,602,919	ChIP	yes
Hu et al. 2011	human	CD36-shLuc-input	0.62	1.36	0.46	-1	25	25	25	10,984,175	Input	no
Hu et al. 2011	human	CD36-shLuc-TAL1	0.97	12.67	2	2	25	25	25	10,455,880	ChIP	yes
Hu et al. 2011	human	CD36-WT-Brg1-1	0.85	1.66	0.56	0	25	25	25	13,673,639	ChIP	yes
Hu et al. 2011	human	CD36-WT-input-1	0.8	2.53	1.05	1	25	25	25	10,309,351	Input	no
Zhao et al. 2011	human	IB4-EBNA2-rep1	0.94	1.91	0.3	-1	36	36	36	5,803,658	ChIP	yes
Zhao et al. 2011	human	IB4-EBNA2-rep2	0.95	2.87	1.03	1	40	40	40	5,536,068	ChIP	yes
Zhao et al. 2011	human	IB4-Input-rep1	0.98	1.38	0.3	-1	40	40	40	4,144,311	Input	no
Zhao et al. 2011	human	IB4-Input-rep2	0.96	1.27	0.3	-1	40	40	40	10,404,527	Input	no
Zhao et al. 2011	human	IB4-RBPJ-rep1	0.98	2.48	0.6	0	36	36	36	2,919,539	ChIP	yes
Zhao et al. 2011	human	IB4-RBPJ-rep2	0.93	3.03	1.26	1	40	40	40	7,475,552	ChIP	yes
Rao et al. 2011	human	HeLaB2-GR-DMSO-GRKD	0.39	8.06	6.32	2	35	35	35	25,313,813	ChIP	no
Rao et al. 2011	human	HeLaB2-GR-DMSO-p65KD	0.96	4.84	4.97	2	35	35	35	12,286,932	ChIP	unknown
Rao et al. 2011	human	HeLaB2-GR-DMSO-WT	0.6	1.35	0.51	0	35	35	35	26,883,356	ChIP	unknown
Rao et al. 2011	human	HeLaB2-GR-TA-WT	0.95	1.82	1.14	1	35	35	35	13,061,670	ChIP	unknown
Rao et al. 2011	human	HeLaB2-GR-TA+TNFa-GRKD	0.55	3.84	3.05	2	35	35	35	23,851,932	ChIP	unknown
Rao et al. 2011	human	HeLaB2-GR-TA+TNFa-p65KD	0.96	5.27	2.96	2	35	35	35	13,570,984	ChIP	unknown
Rao et al. 2011	human	HeLaB2-GR-TA+TNFa-WT	0.66	1.63	0.99	0	35	35	35	27,313,718	ChIP	unknown
Rao et al. 2011	human	HeLaB2-GR-TNFa-WT	0.96	1.29	0.47	-1	35	35	35	13,022,367	ChIP	unknown
Rao et al. 2011	human	HeLaB2-p65-DMSO-GRKD	0.4	7.86	7.04	2	35	35	35	25,556,594	ChIP	no
Rao et al. 2011	human	HeLaB2-p65-DMSO-p65KD	0.92	5.12	3.33	2	35	35	35	15,380,858	ChIP	no
Rao et al. 2011	human	HeLaB2-p65-DMSO-WT	0.52	1.73	0.99	0	35	35	35	17,693,337	ChIP	no
Rao et al. 2011	human	HeLaB2-p65-TA-WT	0.93	1.66	1.03	1	35	35	35	16,120,222	ChIP	yes
Rao et al. 2011	human	HeLaB2-p65-TA+TNFa-GRKD	0.58	4.86	4.49	2	35	35	35	25,972,505	ChIP	yes
Rao et al. 2011	human	HeLaB2-p65-TA+TNFa-p65KD	0.93	3.97	4.16	2	35	35	35	16,624,445	ChIP	no
Rao et al. 2011	human	HeLaB2-p65-TA+TNFa-WT	0.67	2.29	1.67	2	35	35	35	26,290,176	ChIP	yes
Rao et al. 2011	human	HeLaB2-p65-TNFa-WT	0.93	2.31	1.83	2	35	35	35	16,380,803	ChIP	yes
Wang et al. 2011	human	CUTLL-Input-1	0.98	1.24	0.42	-1	40	40	40	19,896,199	Input	no
Wang et al. 2011	human	CUTLL-Input-2	0.98	1.29	0.51	0	40	40	40	20,712,816	Input	no
Wang et al. 2011	human	CUTLL-Notch1-1	0.97	2.44	1.03	1	40	40	40	19,820,660	ChIP	yes
Wang et al. 2011	human	CUTLL-Notch1-2	0.93	4.98	1.27	1	40	40	40	15,252,998	ChIP	yes
Wang et al. 2011	human	CUTLL-RBPJ-1	0.97	1.57	0.69	0	40	40	40	20,226,038	ChIP	yes
Wang et al. 2011	human	CUTLL-RBPJ-2	0.9	3.1	1.06	1	40	40	40	17,569,147	ChIP	yes
Wang et al. 2011	human	CUTLL-ZNF143	0.8	7.23	1.9	2	40	40	40	25,444,869	ChIP	yes
Wang et al. 2011	mouse	G4A2-Input	0.6	1.99	1.14	1	39	39	39	21,212,246	Input	no
Wang et al. 2011	mouse	G4A2-Notch1	0.73	2.24	1.62	2	39	39	39	27,613,376	ChIP	yes
Wang et al. 2011	mouse	G4A2-RBPJ	0.89	1.53	1.03	1	40	40	40	12,929,417	ChIP	yes
Wang et al. 2011	mouse	T6E-Input	0.96	1.24	0.76	0	38	38	38	24,179,307	Input	no
Wang et al. 2011	mouse	T6E-Notch1	0.92	2.22	1.15	1	38	38	38	21,336,323	ChIP	yes
Wang et al. 2011	mouse	T6E-RBPJ	0.93	1.74	0.87	0	38	38	38	16,046,706	ChIP	yes
Costessi et al. 2011	human	K562-NFYA	0.78	4.8	6.12	2	35	35	35	11,661,523	ChIP	yes
Costessi et al. 2011	human	K562-NFYB	0.58	5.4	6.13	2	35	35	35	15,460,623	ChIP	yes
Costessi et al. 2011	human	K562-PRAME	0.87	1.76	1.32	1	35	35	35	6,685,161	ChIP	yes
Costessi et al. 2011	human	K562-Preimmune	0.9	1.67	0.84	0	35	35	35	6,366,475	IgG	no
Miyazaki et al. 2011	mouse	E2A-Day0	0.94	1.39	0.26	-1	36	36	36	9,650,009	ChIP	yes
Miyazaki et al. 2011	mouse	E2A-Day2	0.94	1.34	0.19	-2	36	36	36	8,529,512	ChIP	unknown
Miyazaki et al. 2011	mouse	Input	0.89	1.4	0.24	-2	36	36	36	11,673,268	Input	no
GSE26711	mouse	C2C12-FLAG	0.95	1.76	0.19	-2	26	26	26	2,144,135	IgG	no
GSE26711	mouse	C2C12-FLAG-Msx1	0.74	2.6	1.35	1	32.58	26	36	4,769,291	ChIP	yes
Sun et al. 2011	mouse	MEF-Input	0.85	1.37	0.8	0	36	36	36	17,709,015	Input	no
Sun et al. 2011	mouse	MEF-NelfB	0.75	1.99	1.52	2	36	36	36	16,971,968	ChIP	yes
Heikkinen et al. 2011	human	THP1-calcitriol-VDR	0.41	1.63	0.58	0	36	36	36	26,125,837	ChIP	yes
Heikkinen et al. 2011	human	THP1-IgG	0.46	1.4	0.28	-1	36	36	36	26,578,895	IgG	no
Heikkinen et al. 2011	human	THP1-unstimulated-VDR	0.4	1.37	0.32	-1	36	36	36	22,822,851	ChIP	no
Yoon et al. 2011	xaenopus	Input	0.59	N/A	N/A	N/A	36	36	36	3,219,500	Input	no
Yoon et al. 2011	xaenopus	Smad2-3	0.93	N/A	N/A	N/A	36	36	36	8,168,342	ChIP	yes
Mullen et al. 2011	human	BGO3-Oct4	0.94	2.13	1.51	2	36	36	36	7,835,807	ChIP	yes
Mullen et al. 2011	human	BGO3-Smad3	0.9	2.21	0.73	0	36	36	36	10,206,400	ChIP	yes
Mullen et al. 2011	human	BGO3-WCE	0.99	1.41	0.55	0	36	36	36	8,589,186	Input	no
Mullen et al. 2011	mouse	ESC-Activin-Smad3	0.85	1.98	0.36	-1	36	36	36	3,469,014	ChIP	yes
Mullen et al. 2011	mouse	ESC-Smad2-3-Activin	0.85	1.97	0.35	-1	36	36	36	3,521,351	ChIP	yes
Mullen et al. 2011	mouse	ESC-Smad3	0.92	1.86	0.3	-1	26	26	26	3,650,000	ChIP	unknown
Mullen et al. 2011	mouse	mESC-NoMyod1-Day2-Smad3	0.84	1.67	1.74	2	36	36	36	8,780,818	ChIP	yes
Mullen et al. 2011	mouse	mESC-NoMyod1-Day5-Smad3	0.64	2.84	1.43	1	36	36	36	7,935,259	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Mullen et al. 2011	mouse	mESC-PlusMyod1-Day2-Smad3	0.71	1.89	1.48	1	36	36	36	13,783,301	ChIP	yes
Mullen et al. 2011	mouse	mESC-PlusMyod1-Day5-MyoD-H2Flag	0.67	2.13	1.19	1	36	36	36	12,790,865	ChIP	yes
Mullen et al. 2011	mouse	mESC-PlusMyod1-Day5-Smad3	0.75	1.73	1	0	36	36	36	8,585,103	ChIP	yes
Mullen et al. 2011	mouse	Myotubes-IgG	0.65	1.98	0.75	0	36	36	36	5,056,829	IgG	no
Mullen et al. 2011	mouse	Myotubes-MyoD1-Rep1	0.78	6.25	2.23	2	36	36	36	4,485,416	ChIP	yes
Mullen et al. 2011	mouse	Myotubes-MyoD1-Rep2	0.54	3.75	1.62	2	36	36	36	14,493,250	ChIP	yes
Mullen et al. 2011	mouse	Myotubes-Smad3-Rep1	0.68	3.39	2.15	2	36	36	36	14,630,938	ChIP	yes
Mullen et al. 2011	mouse	Myotubes-Smad3-Rep2	0.19	2.5	2.03	2	36	36	36	11,953,645	ChIP	yes
Mullen et al. 2011	mouse	Pro-Bcells-IgG	0.76	3.51	1.09	1	36	36	36	22,066,974	IgG	no
Mullen et al. 2011	mouse	Pro-Bcells-PU.1-Rep1	0.61	4	1.75	2	36	36	36	11,557,346	ChIP	yes
Mullen et al. 2011	mouse	Pro-Bcells-PU.1-Rep2	0.76	6.98	1.47	1	36	36	36	21,066,565	ChIP	yes
Mullen et al. 2011	mouse	Pro-Bcells-Smad3-Rep1	0.68	1.84	1.26	1	36	36	36	13,801,014	ChIP	yes
Mullen et al. 2011	mouse	Pro-Bcells-Smad3-Rep2	0.74	4.61	2.45	2	36	36	36	13,745,867	ChIP	yes
Wei et al. 2011	mouse	CD4-Gata3	0.5	3.55	1.68	2	25	26	25	5,311,260	ChIP	yes
Wei et al. 2011	mouse	CD8-Fli1	0.64	1.78	0.91	0	25	25	25	4,267,162	ChIP	unknown
Wei et al. 2011	mouse	CD8-Gata3	0.8	1.9	1.08	1	25	25	25	4,827,087	ChIP	yes
Wei et al. 2011	mouse	CD8-Gata3-KO-Fli1	0.95	1.49	0.69	0	25	25	25	3,152,001	ChIP	unknown
Wei et al. 2011	mouse	CD8-Gata3-KO-Gata3	0.92	1.32	0.4	-1	25	25	25	1,997,286	ChIP	yes
Wei et al. 2011	mouse	DN-Gata3	0.66	2.87	1.68	2	25	25	25	6,301,966	ChIP	yes
Wei et al. 2011	mouse	DP-Gata3	0.76	1.86	1.14	1	25	25	25	6,402,211	ChIP	yes
Wei et al. 2011	mouse	DP-Gata3-replicate	0.08	14.33	7.09	2	25	25	25	20,563,880	ChIP	yes
Wei et al. 2011	mouse	iTreg-Gata3	0.26	2.42	1.28	1	25	25	25	7,299,209	ChIP	yes
Wei et al. 2011	mouse	NKT-Gata3	0.21	16.16	2.81	2	25	25	25	4,716,486	ChIP	yes
Wei et al. 2011	mouse	nTreg-Gata3	0.69	5.83	2.01	2	25	25	25	4,163,536	ChIP	yes
Wei et al. 2011	mouse	Th17-Gata3	0.32	1.64	0.68	0	25	25	25	5,051,835	ChIP	unknown
Wei et al. 2011	mouse	Th1-Gata3	0.67	2.79	1.53	2	25	25	25	6,296,541	ChIP	yes
Wei et al. 2011	mouse	Th2-Ets1	0.37	3.54	1.87	2	25	25	25	1,620,989	ChIP	yes
Wei et al. 2011	mouse	Th2-Fli1	0.81	4.76	0.02	-2	24	24	24	444,327	ChIP	yes
Wei et al. 2011	mouse	Th2-Gata3	0.86	2.86	2.47	2	25	25	25	7,514,211	ChIP	yes
Wei et al. 2011	mouse	Th2-Gata3-replicate	0.86	2.86	2.47	2	25	25	25	7,514,211	ChIP	yes
Liu et al. 2011	mouse	mES-TAF1	0.7	1.04	0.11	-2	36	36	36	42,959,794	ChIP	yes
Liu et al. 2011	mouse	mES-TAF1-IgG	0.66	1.1	0.27	-1	36	36	36	38,486,238	IgG	no
Liu et al. 2011	mouse	mES-TAF3	0.48	1.76	0.94	0	36	36	36	37,109,895	ChIP	yes
Liu et al. 2011	mouse	mES-TAF3-IgG	0.38	1.11	0.2	-2	36	36	36	41,265,618	IgG	no
Liu et al. 2011	mouse	mES-TBP	0.64	2.1	0.93	0	36	36	36	34,110,153	ChIP	yes
Liu et al. 2011	mouse	mES-TBP-IgG	0.31	1.21	0.17	-2	36	36	36	33,960,211	IgG	no
Kong et al. 2011	human	MCF7-DMSO-GATA3	0.92	2.27	1.02	1	36	36	36	16,110,797	ChIP	yes
Kong et al. 2011	human	MCF7-DMSO-p300	0.94	1.49	0.53	0	36	36	36	16,598,044	ChIP	yes
Kong et al. 2011	human	MCF7-E2-GATA3	0.94	3.42	1.43	1	36	36	36	22,771,157	ChIP	yes
Kong et al. 2011	human	MCF7-E2-p300	0.92	1.54	0.46	-1	36	36	36	12,820,747	ChIP	yes
GSE31951	mouse	0hrKCl-Input-sampleB1	0.87	2.26	0.91	0	33	33	33	21,405,879	Input	no
GSE31951	mouse	0hrKCl-Input-sampleB2	0.61	1.21	0.17	-2	33	33	33	11,303,008	Input	no
GSE31951	mouse	0hrKCl-MeCP2IP-sampleB1	0.84	3.88	2.11	2	33	33	33	34,260,253	ChIP	yes
GSE31951	mouse	0hrKCl-MeCP2IP-sampleB2	0.8	1.97	1.68	2	33	33	33	14,827,886	ChIP	yes
GSE31951	mouse	2hrKcl-Input-sampleB1	0.88	1.97	0.32	-1	33	33	33	8,725,472	Input	no
GSE31951	mouse	2hrKCl-Input-sampleB2	0.77	1.15	0.16	-2	33	33	33	45,501,766	Input	no
GSE31951	mouse	2hrKCl-MeCP2IP-sampleB1	0.8	4.79	2.65	2	33	33	33	13,050,973	ChIP	yes
GSE31951	mouse	2hrKCl-MeCP2IP-sampleB2	0.47	1.9	1.33	1	33	33	33	9,573,633	ChIP	yes
GSE31951	mouse	2hrKCl-pS421MeCP2IP-sampleB2	0.89	1.86	0.67	0	33	33	33	3,733,245	ChIP	yes
Norton et al. 2011	rat	H4IIE-input	0.75	3.28	1.41	1	40	40	40	22,889,534	Input	no
Norton et al. 2011	rat	H4IIE-TCF7L2	0.81	3.08	1.14	1	40	40	40	21,999,570	ChIP	yes
Bernt et al. 2011	mouse	MLL-AF9	0.76	2.2	2.05	2	36	36	36	20,979,495	ChIP	yes
Sahu et al. 2011	human	LNCaP-AR-rep1	0.96	2.15	0.84	0	30	30	30	13,178,048	ChIP	yes
Sahu et al. 2011	human	LNCaP-AR-rep2	0.97	1.71	0.69	0	30	30	30	13,295,369	ChIP	yes
Sahu et al. 2011	human	LNCaP-AR-siFoxA1-rep1	0.97	2.45	1.13	1	30	30	30	16,070,383	ChIP	yes
Sahu et al. 2011	human	LNCaP-AR-siFoxA1-rep2	0.96	2.64	1.18	1	30	30	30	16,077,043	ChIP	yes
Sahu et al. 2011	human	LNCaP-FoxA1-rep1	0.98	3.05	1.14	1	30	30	30	7,592,193	ChIP	yes
Sahu et al. 2011	human	LNCaP-FoxA1-rep2	0.98	3.56	1.38	1	30	30	30	8,058,879	ChIP	yes
Sahu et al. 2011	human	LNCaP-FoxA1-siFoxA1-rep1	0.97	1.77	0.44	-1	30	30	30	5,946,745	ChIP	no
Sahu et al. 2011	human	LNCaP-FoxA1-siFoxA1-rep2	0.97	1.76	0.42	-1	30	30	30	5,835,884	ChIP	no
Sahu et al. 2011	human	LNCaP-GR	0.97	1.78	0.83	0	36	36	36	22,124,446	ChIP	yes
Sahu et al. 2011	human	LNCaP-GR-siFoxA1	0.93	1.68	0.91	0	36	36	36	17,943,158	ChIP	yes
Sahu et al. 2011	human	LNCaP-rIgG	0.95	1.16	0.21	-2	30	30	30	16,327,209	IgG	no
An et al. 2011	mouse	C2C12-Input-rep1	0.95	1.13	0.48	-1	37	37	37	17,130,843	Input	no
An et al. 2011	mouse	C2C12-Input-rep2	0.94	1.1	0.5	-1	40	40	40	24,457,563	Input	no
An et al. 2011	mouse	C2C12-Sox6-rep1	0.92	2.16	0.39	-1	40	40	40	2,989,595	ChIP	yes
An et al. 2011	mouse	C2C12-Sox6-rep2	0.96	2.37	0.26	-1	40	40	40	1,470,144	ChIP	yes
Shukla et al. 2011	human	BJAB-CTCF	0.85	3.53	1.88	2	35	35	35	20,488,614	ChIP	yes
Shukla et al. 2011	human	BJAB-Rabbit-IgG	0.82	3.24	2.1	2	35	35	35	17,746,364	IgG	no
Shukla et al. 2011	human	BL41-CTCF	0.81	3.63	2.27	2	35	35	35	27,623,415	ChIP	yes
Shukla et al. 2011	human	BL41-Rabbit-IgG	0.68	4.44	3.71	2	35	35	35	29,655,822	IgG	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Trompouki et al. 2011*	mouse	Gata1-G1ERbmp-r1-100914-4	0.97	1.5	0.99	0	36	36	36	18,435,160	ChIP	yes
Trompouki et al. 2011*	mouse	Gata2-G1Ebmp-r1-101201-3	0.82	12.68	1.59	2	36	36	36	8,484,282	ChIP	yes
Trompouki et al. 2011*	mouse	Smad1-G1Ebmp-r1-100914-6	0.8	4.78	1.52	2	36	36	36	14,561,496	ChIP	yes
Trompouki et al. 2011*	mouse	Smad1-G1ERbmp-r1-100914-5	0.85	4.16	1.38	1	36	36	36	16,186,687	ChIP	yes
Trompouki et al. 2011*	mouse	WCE-G1Ebmp-r1-101201-2	0.97	1.43	0.92	0	36	36	36	14,429,966	Input	no
Trompouki et al. 2011*	mouse	WCE-G1ERbio-r1-100914-1	0.97	1.53	0.98	0	36	36	36	17,835,267	Input	no
Trompouki et al. 2011*	human	GATA1-CD34eryth-bio-r1-101103-6	0.73	4.25	0.05	-2	36	36	36	94,232	ChIP	yes
Trompouki et al. 2011*	human	GATA1-CD34eryth-bio-r2-101103-7	0.22	21.87	0.14	-2	36	36	36	744,924	ChIP	yes
Trompouki et al. 2011*	human	GATA1-CD34eryth-bmp-r1-100922-4	0.57	10.86	0.14	-2	36	36	36	667,864	ChIP	yes
Trompouki et al. 2011*	human	GATA1-CD34eryth-bmp-r2-101105-1	0.49	10.81	0.15	-2	36	36	36	900,792	ChIP	yes
Trompouki et al. 2011*	human	GATA2-CD34prog-bmp-r1-101201-1	0.59	2.91	0.05	-2	36	36	36	479,725	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-CD34eryth-bmp-r1-100922-5	0.65	6.9	0.07	-2	36	36	36	634,638	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-CD34eryth-bmp-r2-101103-7	0.62	9.05	0.07	-2	36	36	36	730,479	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-CD34prog-bmp-r1-100901-1	0.68	4.26	0.06	-2	36	36	36	322,324	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-CD34prog-bmp-r2-101105-3	0.68	3.35	0.06	-2	36	36	36	620,959	ChIP	yes
Trompouki et al. 2011*	human	TCF7L2-CD34prog-bio-r1-100826-7	0.72	3.01	0.06	-2	36	36	36	339,258	ChIP	yes
Trompouki et al. 2011*	human	TCF7L2-CD34prog-bio-r2-101105-4	0.64	2.96	0.07	-2	36	36	36	529,093	ChIP	yes
Trompouki et al. 2011*	human	WCE-CD34eryth-bio-r1-101103-4	0.59	19.41	0.28	-1	36	36	36	79,783	Input	no
Trompouki et al. 2011*	human	WCE-CD34eryth-bio-r1-101201-4	0.54	3.54	0.05	-2	36	36	36	502,346	Input	no
Trompouki et al. 2011*	human	WCE-CD34eryth-bio-r2-101103-5	0.69	8.49	0.04	-2	36	36	36	340,786	Input	no
Trompouki et al. 2011*	human	WCE-CD34eryth-bmp-r1-100922-3	0.69	11.24	0.05	-2	36	36	36	246,281	Input	no
Trompouki et al. 2011*	human	WCE-CD34eryth-bmp-r2-101105-2	0.68	10.7	0.05	-2	36	36	36	300,293	Input	no
Trompouki et al. 2011*	human	WCE-CD34prog-bio-r1-100826-6	0.65	2.88	0.04	-2	36	36	36	356,819	Input	no
Trompouki et al. 2011*	human	WCE-CD34prog-bio-r1-101201-1	0.59	2.91	0.05	-2	36	36	36	479,725	Input	no
Trompouki et al. 2011*	human	WCE-CD34prog-bio-r2-101105-5	0.65	2.75	0.04	-2	36	36	36	283,167	Input	no
Trompouki et al. 2011*	human	WCE-CD34prog-bmp-r1-101201-7	0.6	3.13	0.05	-2	36	36	36	430,773	Input	no
Trompouki et al. 2011*	human	CEBPA-U937bio-r1-100709-5	0.58	26.28	0.3	-1	35	35	35	4,430,334	ChIP	yes
Trompouki et al. 2011*	human	CEBPA-U937dmso-r1-100505-5	0.41	23.29	0.18	-2	36	36	36	151,538	ChIP	yes
Trompouki et al. 2011*	human	CEBPA-K562-CEBPA-bmp4	0.46	10.99	0.34	-1	35	35	35	2,662,588	ChIP	yes
Trompouki et al. 2011*	human	CEBPA-U937-bmp4	0.64	3.16	0.44	-1	36	36	36	228,810	ChIP	yes
Trompouki et al. 2011*	human	GATA1-K562bio-r1-110325-6	0.49	11.55	0.09	-2	39	39	39	245,220	ChIP	yes
Trompouki et al. 2011*	human	GATA1-K562bmp-r1-110325-4	0.67	11.66	0.16	-2	36	36	36	335,062	ChIP	yes
Trompouki et al. 2011*	human	GATA1-K562	0.57	3.41	0.11	-2	36	36	36	371,785	ChIP	yes
Trompouki et al. 2011*	human	GATA2-K562bio-r1-110325-5	0.51	5.76	0.13	-2	39	39	39	190,367	ChIP	yes
Trompouki et al. 2011*	human	GATA2-K562bmp-r1-110325-3	0.27	5.97	0.14	-2	36	36	36	405,703	ChIP	yes
Trompouki et al. 2011*	human	GATA2-K562	0.47	10.43	0.02	-2	36	36	36	451,795	ChIP	yes
Trompouki et al. 2011*	human	Input-K562-CEBPA-bmp4	0.08	27.04	0.2	-2	39	39	39	248,035	Input	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Trompouki et al. 2011*	human	SMAD1-K562bmp4-r1-100608-2	0.75	8	0.09	-2	35	35	35	834,331	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-K562campk-r1-110323-2	0.71	2.48	0.14	-2	36	36	36	645,936	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-U937bmp4-r1-100608-1	0.69	12.93	0.24	-2	36	36	36	1,184,890	ChIP	yes
Trompouki et al. 2011*	human	SMAD1-K562-CEBPA-bmp4	0.83	16.08	0.21	-2	36	36	36	3,126,161	ChIP	yes
Trompouki et al. 2011*	human	TCF7L2-K562bio-r1-100106-7	0.75	10.1	0.05	-2	36	36	36	88,763	ChIP	yes
Trompouki et al. 2011*	human	TCF7L2-K562bio-r2-Childrens	0.73	5.78	0.07	-2	40	40	40	116,187	ChIP	yes
Trompouki et al. 2011*	human	TCF7L2-U937bio-r1-100505-7	0.39	11.03	0.18	-2	36	36	36	145,311	ChIP	yes
Trompouki et al. 2011*	human	WCE-K562bio-r1-100106-5	0.72	8.07	0.02	-2	36	36	36	163,001	Input	no
Trompouki et al. 2011*	human	WCE-K562bio-r1-100608-2	0.67	5.23	0.02	-2	36	36	36	314,395	Input	no
Trompouki et al. 2011*	human	WCE-K562bmp4-r1-100608-1	0.67	8.04	0.05	-2	36	36	36	340,757	Input	no
Trompouki et al. 2011*	human	WCE-U937bio-r1-100505-6	0.61	4.26	0.07	-2	36	36	36	294,327	Input	no
Trompouki et al. 2011*	human	WCE-U937bio-r1-100608-5	0.64	3.52	0.06	-2	36	36	36	326,001	Input	no
Trompouki et al. 2011*	human	WCE-U937bio-r1-100709-4	0.62	3.02	0.06	-2	36	36	36	308,988	Input	no
Trompouki et al. 2011*	human	WCE-U937bio-r1-100709-6	0.62	2.64	0.06	-2	36	36	36	299,190	Input	no
Trompouki et al. 2011*	human	WCE-U937bmp4-r1-100608-3	0.65	3.73	0.04	-2	36	36	36	314,568	Input	no
Trompouki et al. 2011*	human	WCE-U937dmso-r1-100505-3	0.63	4.64	0.07	-2	36	36	36	272,327	Input	no
Ceschin et al. 2011	human	H3396-CARM1-E2	0.88	2.05	0.19	-2	49	49	49	99,711	ChIP	yes
Ceschin et al. 2011	human	H3396-CARM1-EtOH	0.88	2.15	0.2	-2	49	49	49	105,031	ChIP	yes
Ceschin et al. 2011	human	H3396-CBP-E2	0.51	4.35	0.02	-2	36	36	36	152,170	ChIP	yes
Ceschin et al. 2011	human	H3396-CBP-EtOH-1	0.33	8.51	0.06	-2	36	36	36	289,897	ChIP	yes
Ceschin et al. 2011	human	H3396-CBPR2151m-E2	0.44	7.62	0.03	-2	36	36	36	168,286	ChIP	yes
Ceschin et al. 2011	human	H3396-CBPR2151m-EtOH	0.46	8.58	0.03	-2	36	36	36	149,173	ChIP	yes
Ceschin et al. 2011	human	H3396-CBPR742m-E2-1	0.39	1.93	0.07	-2	36	36	36	383,127	ChIP	yes
Ceschin et al. 2011	human	H3396-CBPR768m-E2	0.41	10.17	0.04	-2	40	40	40	168,599	ChIP	yes
Ceschin et al. 2011	human	H3396-CBPR768m-EtOH	0.4	3.59	0.12	-2	40	40	40	157,261	ChIP	yes
Ceschin et al. 2011	human	H3396-ERa-E2	0.47	8.8	0.01	-2	36	36	36	105,766	ChIP	yes
Ceschin et al. 2011	human	H3396-ERa-EtOH	0.51	10.95	0.02	-2	36	36	36	73,181	ChIP	no
Ceschin et al. 2011	human	H3396-Input-E2-rep1-1	0.71	9.46	0.03	-2	36	36	36	92,932	Input	no
Ceschin et al. 2011	human	H3396-RAC3-E2	0.52	6.03	0.02	-2	36	36	36	110,731	ChIP	yes
Mendoza-Parra et al. 2011	mouse	mouse-F9-WCE	0.87	8.28	2.79	2	36	36	36	6,377,439	Input	no
Mendoza-Parra et al. 2011	mouse	RARg-24h-ATRA	0.87	4.27	1.95	2	36	36	36	5,864,836	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RARg-2h-ATRA	0.87	4.27	1.92	2	36	36	36	6,545,542	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RARg-48h-ATRA	0.91	3.46	1.82	2	36	36	36	3,543,638	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RARg-48h-EtOH	0.8	4.55	0.93	0	36	36	36	6,281,297	ChIP	unknown
Mendoza-Parra et al. 2011	mouse	RARg-6h-ATRA	0.65	5.31	1.93	2	36	36	36	6,353,453	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RXRa-24h-ATRA	0.67	4.42	1.29	1	36	36	36	6,444,150	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RXRa-2h-ATRA	0.56	9.77	3.79	2	36	36	36	6,676,769	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RXRa-48h-ATRA	0.6	11.1	3.89	2	36	36	36	5,869,783	ChIP	yes
Mendoza-Parra et al. 2011	mouse	RXRa-48h-EtOH	0.7	5.14	1.32	1	36	36	36	6,631,973	ChIP	unknown
Mendoza-Parra et al. 2011	mouse	RXRa-6h-ATRA	0.54	7.61	3.08	2	36	36	36	5,834,436	ChIP	yes
Mendoza-Parra et al. 2011	mouse	rxra-ko-RXRa-48h-ATRA	0.89	2.86	0.88	0	36	36	36	4,573,205	ChIP	yes
Schmitz et al. 2011	mouse	mESC-Jarid1b-1	0.87	1.25	0.34	-1	34	34	34	3,996,359	ChIP	yes
Schmitz et al. 2011	mouse	mESC-Jarid1b-2	0.88	1.24	0.36	-1	26	26	26	3,488,817	ChIP	yes
Bergsland et al. 2011	mouse	C2C12-Sox3-transfected-Sox3	0.77	1.44	0.85	0	53	53	53	29,894,751	ChIP	yes
Bergsland et al. 2011	mouse	Early-formed-neurons-IgG	0.93	2.05	0.2	-2	33	33	33	2,107,025	IgG	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Bergsland et al. 2011	mouse	Early-formed-neurons-Sox11-rep1	0.94	2.15	0.27	-1	33	33	33	2,103,532	ChIP	yes
Bergsland et al. 2011	mouse	Early-formed-neurons-Sox11-rep2	0.95	1.99	0.27	-1	33	33	33	2,328,712	ChIP	yes
Bergsland et al. 2011	mouse	Early-formed-neurons-Sox11-rep3	0.96	1.67	0.34	-1	33	33	33	2,668,012	ChIP	yes
Bergsland et al. 2011	mouse	NPC-Sox2-rep1	0.9	1.38	0.29	-1	38	38	38	6,840,926	ChIP	yes
Bergsland et al. 2011	mouse	NPC-Sox2-rep2	0.74	1.62	0.69	0	38	38	38	12,391,326	ChIP	yes
Bergsland et al. 2011	mouse	NPC-Sox2-rep3	0.79	1.9	1.49	1	38	38	38	15,894,900	ChIP	yes
Bergsland et al. 2011	mouse	NPC-Sox3-rep1	0.88	2.68	1.34	1	34	34	34	3,339,224	ChIP	yes
Bergsland et al. 2011	mouse	NPC-Sox3-rep2	0.93	2.47	0.5	0	34	34	34	1,464,673	ChIP	yes
Bergsland et al. 2011	mouse	NPC-Sox3-rep3	0.87	2.87	2.24	2	34	34	34	3,496,087	ChIP	yes
Marban et al. 2011	human	Jurkat-Input	0.96	3.11	0.77	0	76	76	76	15,973,065	Input	no
Marban et al. 2011	human	Jurkat-Tat	0.94	4.07	1.04	1	76	76	76	18,900,158	ChIP	yes
Quenneville et al. 2011	mouse	mESC-HA	0.67	7.09	10.49	2	37.45	38	37	47,077,818	IgG	no
Quenneville et al. 2011	mouse	mESC-HAZFP57-HA	0.76	4.65	6.75	2	37.35	38	37	40,511,425	ChIP	yes
Quenneville et al. 2011	mouse	mESC-KAP1	0.63	4.31	7.86	2	49.63	76	38	58,793,249	ChIP	yes
Mullican et al. 2011	mouse	Macrophage-BSA-HDAC3	0.84	1.7	1.31	1	38	38	38	18,260,410	ChIP	yes
Mullican et al. 2011	mouse	Macrophage-IL4-HDAC3	0.89	1.64	1.18	1	38	38	38	17,042,856	ChIP	yes
Mullican et al. 2011	mouse	Macrophage-Input	0.95	1.09	0.36	-1	36	36	36	19,136,736	Input	no
Brown et al. 2011	human	hESC-D0-Smad-XL-rep1	0.95	1.67	0.42	-1	38	38	38	5,323,799	ChIP	yes
Brown et al. 2011	human	hESC-D0-Smad-XL-rep2	0.72	1.44	0.51	0	36	36	36	30,063,231	ChIP	yes
Brown et al. 2011	human	hESC-D3-Smad-XL-rep1	0.97	1.62	0.36	-1	38	38	38	6,844,734	ChIP	yes
Brown et al. 2011	human	hESC-D3-Smad-XL-rep2	0.75	1.44	0.42	-1	36	36	36	29,936,111	ChIP	yes
Brown et al. 2011	human	hESC-Input-XL	0.98	1.37	0.44	-1	36	36	36	7,422,963	Input	no
Mazzoni et al. 2011	mouse	Progenitor-Motor-Neurons-Day4-iOlig2-V5	0.92	3.13	1.85	2	36	36	36	3,330,651	ChIP	yes
Mazzoni et al. 2011	mouse	Progenitor-Motor-Neurons-Day4-Olig2	0.9	5	1.52	2	36	36	36	8,348,180	ChIP	yes
Mazzoni et al. 2011	mouse	Progenitor-Motor-Neurons-Day4-V5-control	0.93	1.48	0.48	-1	36	36	36	13,581,601	Input	no
Mazzoni et al. 2011	mouse	Progenitor-Motor-Neurons-Day5-iFlag-Hoxc9	0.87	3.68	2.59	2	36	36	36	29,775,081	ChIP	yes
Mazzoni et al. 2011	mouse	Progenitor-Motor-Neurons-Day5-iHoxc9-V5	0.71	2.48	2.42	2	69.05	76	36	28,150,488	ChIP	yes
Tan et al. 2011	human	LNCap-DHT-AR-1	0.83	11.17	1.68	2	36	36	36	13,158,813	ChIP	yes
Tan et al. 2011	human	LNCap-DHT-FoxA1-1	0.89	9.94	2.58	2	36	36	36	18,910,797	ChIP	yes
Tan et al. 2011	human	LNCap-DHT-NKX3-1	0.93	1.98	0.62	0	36	36	36	11,840,488	ChIP	yes
Tan et al. 2011	human	LNCap-EtOH-AR-1	0.92	2.71	0.92	0	36	36	36	10,786,161	ChIP	unknown
Tan et al. 2011	human	LNCap-EtOH-FoxA1	0.96	9.35	2.52	2	36	36	36	5,367,267	ChIP	yes
Tan et al. 2011	human	LNCap-EtOH-NKX3-1	0.91	1.59	0.51	0	36	36	36	16,850,974	ChIP	yes
Tan et al. 2011	human	LNCaP-Genomic-Input-1	0.95	1.54	0.51	0	36	36	36	10,550,285	Input	no
Shen et al. 2011	mouse	Heart-input1	0.87	1.86	0.47	-1	36	36	36	5,928,909	Input	no
Shen et al. 2011	mouse	Heart-input2	0.95	1.38	0.41	-1	36	36	36	6,264,090	Input	no
Shen et al. 2011	mouse	Heart-input3	0.94	1.21	0.48	-1	36	36	36	10,837,874	Input	no
Shen et al. 2011	mouse	Heart-Tbx20-GFP	0.95	1.9	0.63	0	36	36	36	23,754,878	ChIP	yes
Seitz et al. 2011	human	BL41-Input	0.98	1.19	0.1	-2	31	31	31	1,972,404	Input	no
Seitz et al. 2011	human	BL41-Myc	0.86	3.51	1.22	1	33	33	33	2,719,977	ChIP	yes
Seitz et al. 2011	human	Blue1-Input	0.99	1.27	0.12	-2	31	31	31	1,765,339	Input	no
Seitz et al. 2011	human	Blue1-Myc	0.98	2.46	0.66	0	31	31	31	1,884,244	ChIP	yes
Seitz et al. 2011	human	CA46-Input	0.98	1.2	0.13	-2	31	31	31	1,492,644	Input	no
Seitz et al. 2011	human	CA46-Myc	0.93	1.4	0.14	-2	71	71	71	1,734,564	ChIP	yes
Seitz et al. 2011	human	Raji-Input	0.95	1.26	0.16	-2	34	34	34	3,027,850	Input	no
Seitz et al. 2011	human	Raji-Myc	0.82	1.22	0.25	-2	34	34	34	2,186,015	ChIP	yes
Seitz et al. 2011	human	Ramos-Input	0.98	1.2	0.14	-2	31	31	31	1,880,856	Input	no
Seitz et al. 2011	human	Ramos-Myc	0.94	2.09	0.61	0	33	33	33	3,293,975	ChIP	yes
Little et al. 2011	human	C4-2B-Input	0.93	1.05	0.33	-1	50	50	50	85,985,363	Input	no
Little et al. 2011	human	C4-2B-Runx2	0.18	2.9	3.54	2	50	50	50	63,645,646	ChIP	yes
Whyte et al. 2011	mouse	mES-CoREST	0.94	1.37	0.39	-1	36	36	36	9,515,699	ChIP	yes
Whyte et al. 2011	mouse	mES-HDAC1	0.33	4.48	2.65	2	36	36	36	17,775,205	ChIP	yes
Whyte et al. 2011	mouse	mES-HDAC1-rep2	0.13	2.33	1.22	1	36	36	36	27,399,530	ChIP	yes
Whyte et al. 2011	mouse	mES-HDAC2	0.69	2.49	1.65	2	36	36	36	14,740,848	ChIP	yes
Whyte et al. 2011	mouse	mES-HDAC2-rep2	0.16	2.29	1.74	2	36	36	36	25,056,680	ChIP	yes
Whyte et al. 2011	mouse	mES-LSD1	0.94	1.54	0.79	0	36	36	36	3,907,159	ChIP	yes
Whyte et al. 2011	mouse	mES-LSD1-rep2	0.93	2.25	1.23	1	36	36	36	24,506,916	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Whyte et al. 2011	mouse	mES-Mi-2	0.42	1.54	0.56	0	36	36	36	24,712,531	ChIP	yes
Whyte et al. 2011	mouse	mES-Mi-2b	0.95	1.27	0.45	-1	36	36	36	10,665,386	ChIP	yes
Whyte et al. 2011	mouse	mES-REST	0.73	3.31	1.57	2	36	36	36	24,569,235	ChIP	yes
Whyte et al. 2011	mouse	WCE-DMSO-t0	0.71	2.01	0.9	0	36	36	36	11,409,350	Input	no
Whyte et al. 2011	mouse	WCE-DMSO-t48	0.77	2.31	1.4	1	36	36	36	13,324,722	Input	no
Whyte et al. 2011	mouse	WCE-TCP-48	0.76	2.37	1.41	1	36	36	36	12,021,728	Input	no
GSE25426	human	THP-1-Control	0.94	1.29	0.36	-1	36	36	36	21,074,660	Input	no
GSE25426	human	THP-1-PPARg	0.96	1.82	0.63	0	36	36	36	14,473,006	ChIP	yes
GSE25426	human	THP-1-PU.1	0.95	4.78	2.12	2	35	35	35	13,571,315	ChIP	yes
GSE25426	human	THP-1-RXR	0.98	1.46	0.29	-1	35	35	35	6,999,922	ChIP	yes
Yildirim et al. 2011	mouse	mESC-Brg1-KD-Mbd3	0.96	2.73	0.15	-2	36	36	36	1,656,511	ChIP	unknown
Yildirim et al. 2011	mouse	mESC-Mbd3-rep1	0.94	2.21	0.33	-1	36	36	36	2,189,692	ChIP	yes
Yildirim et al. 2011	mouse	mESC-Mbd3-rep2	0.85	1.23	0.61	0	36	36	36	15,055,944	ChIP	yes
Yildirim et al. 2011	mouse	mESC-Tet1-KD-Mbd3	0.97	1.56	0.25	-1	36	36	36	3,626,622	ChIP	unknown
Botcheva et al. 2011	human	IMR90-Input	0.87	1.27	0.21	-2	36	36	36	9,286,134	Input	no
Botcheva et al. 2011	human	IMR90-p53	0.7	2.66	0.61	0	36	36	36	5,285,892	ChIP	yes
Stadler et al. 2011	mouse	ES-CTCF-rep1	0.64	28.28	1.67	2	37	37	37	10,466,451	ChIP	yes
Stadler et al. 2011	mouse	ES-CTCF-rep2	0.44	26.16	2.87	2	38	38	38	13,296,384	ChIP	yes
Stadler et al. 2011	mouse	ES-CTCF-rep3	0.49	9.15	8.28	2	38	38	38	9,587,128	ChIP	yes
Stadler et al. 2011	mouse	ES-Input-rep1	0.82	1.77	1.05	1	38	38	38	11,095,374	Input	no
Stadler et al. 2011	mouse	ES-Input-rep2	0.83	1.94	2.94	2	38	38	38	29,650,665	Input	no
Stadler et al. 2011	mouse	TKO-CTCF-rep1	0.63	10.83	3.72	2	36	36	36	34,828,958	ChIP	yes
Stadler et al. 2011	mouse	TKO-CTCF-rep2	0.91	5.48	3.61	2	36	36	36	2,836,169	ChIP	yes
Holmstrom et al. 2011	mouse	Pancreas-Input	0.97	1.46	0.88	0	36	36	36	11,479,285	Input	no
Holmstrom et al. 2011	mouse	Pancreas-Lrh1	0.84	4.4	1.89	2	36	36	36	13,587,564	ChIP	yes
Xu et al. 2011	zebrafish	Mxtx2-4.5hpf	0.77	1.7	1.39	1	36	36	36	11,341,093	ChIP	yes
Xu et al. 2011	zebrafish	Nanog-like-3.5hpf	0.63	1.89	1.44	1	36	36	36	7,535,238	ChIP	yes
Xu et al. 2011	zebrafish	Nanog-like-4.5hpf	0.84	3.94	1.22	1	36	36	36	10,103,194	ChIP	yes
Xu et al. 2011	zebrafish	WCE-Mxtx2-4.5hpf	0.97	1.24	0.59	0	36	36	36	18,309,687	Input	no
Xu et al. 2011	zebrafish	WCE-Nanog-like-3.5hpf	0.91	1.63	0.9	0	36	36	36	11,252,453	Input	no
Xu et al. 2011	zebrafish	WCE-Nanog-like-4.5hpf	0.98	1.36	0.67	0	36	36	36	15,831,173	Input	no
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	ES-JNK13-biological-replicate-a	0.82	4.27	2.79	2	38	38	38	8,462,462	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	ES-JNK13-biological-replicate-b	0.51	10.19	3.89	2	38	38	38	8,175,875	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	ES-NFYA-biological-replicate-a	0.79	2.98	5.24	2	38	38	38	19,929,924	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	ES-NFYA-biological-replicate-b	0.66	3.93	7.62	2	38	38	38	24,051,713	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	Input	0.82	1.77	1.05	1	38	38	38	11,095,374	Input	no
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	NP-JNK13-biological-replicate-a	0.69	8.11	5.07	2	38	38	38	8,802,240	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	NP-JNK13-biological-replicate-b	0.92	2.07	1.37	1	38	38	38	9,691,977	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	NP-NFYA-biological-replicate-a	0.85	2.18	3.96	2	38	38	38	23,674,653	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	NP-NFYA-biological-replicate-b	0.86	1.97	3.33	2	38	38	38	21,717,487	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	TN-DMSO-JNK1-3	0.17	11.6	12.92	2	36	36	36	38,425,945	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	TN-JNK1-3-biological-replicate-a	0.35	17.31	5.6	2	38	38	38	8,678,605	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	TN-JNK1-3-biological-replicate-b	0.29	20.52	7.67	2	38	38	38	6,897,900	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	TN-JNKi-JNK1-3	0.35	5.86	12.97	2	36	36	36	42,637,275	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	TN-NFYA-biological-replicate-a	0.68	3.06	5.64	2	38	38	38	27,386,709	ChIP	yes
Tiwari et al. 2011a; Tiwari et al. 2011b	mouse	TN-NFYA-biological-replicate-b	0.84	2.15	3.33	2	38	38	38	25,748,779	ChIP	yes
Zhang et al. 2011	mouse	F-Bcl6-rep1-G51	0.91	2.05	1.74	2	36	36	36	7,810,319	ChIP	yes
Zhang et al. 2011	mouse	F-Bcl6-rep2-G65-M1	0.9	6.17	0.58	0	36	36	36	6,073,003	ChIP	yes
Zhang et al. 2011	mouse	F-Bcl6-rep3-G65-M2	0.75	2.83	1.16	1	36	36	36	6,266,286	ChIP	yes
Zhang et al. 2011	mouse	F-Bcl6-rep4-G65-M3	0.93	3.23	0.42	-1	36	36	36	12,764,985	ChIP	yes
Zhang et al. 2011	mouse	FH-STAT5-rep1-G66-M1	0.83	6.19	3.78	2	36	36	36	6,691,463	ChIP	yes
Zhang et al. 2011	mouse	FH-STAT5-rep2-G66-M2	0.69	5.59	2.97	2	36	36	36	6,110,031	ChIP	yes
Zhang et al. 2011	mouse	FH-STAT5-rep3-G66-M3	0.65	9.48	4.33	2	36	36	36	13,444,170	ChIP	yes
Zhang et al. 2011	mouse	FL-STAT5-rep1-G52	0.19	2.48	1.94	2	36	36	36	5,389,553	ChIP	yes
Zhang et al. 2011	mouse	FL-STAT5-rep2-G70-M3	0.52	3.27	1.47	1	36	36	36	5,884,969	ChIP	yes
Zhang et al. 2011	mouse	FL-STAT5-rep3-G72-M1	0.63	2.55	0.92	0	36	36	36	2,627,103	ChIP	yes
Zhang et al. 2011	mouse	IgG-control	0.62	1.66	0.81	0	35	35	35	11,562,651	IgG	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Zhang et al. 2011	mouse	M-Bcl6-rep1-G49	0.49	3.13	1.72	2	35	35	35	18,985,967	ChIP	yes
Zhang et al. 2011	mouse	M-Bcl6-rep2-G50	0.8	2.86	2.36	2	35	35	35	14,452,480	ChIP	yes
Zhang et al. 2011	mouse	M-Bcl6-rep3-G71-M2	0.76	2.33	0.74	0	35	35	35	14,149,114	ChIP	yes
Zhang et al. 2011	mouse	MH-STAT5-rep1-G36	0.62	4.29	2.69	2	35	35	35	15,997,841	ChIP	yes
Zhang et al. 2011	mouse	MH-STAT5-rep2-G41	0.76	3.64	2.47	2	35	35	35	12,841,332	ChIP	yes
Zhang et al. 2011	mouse	MH-STAT5-rep3-G42	0.64	3.39	2.67	2	36	36	36	9,528,779	ChIP	yes
Zhang et al. 2011	mouse	ML-STAT5-rep1-G35	0.71	2.18	2.17	2	36	36	36	16,024,096	ChIP	yes
Zhang et al. 2011	mouse	ML-STAT5-rep2-G40	0.84	1.69	1.42	1	36	36	36	5,688,929	ChIP	yes
Smith et al. 2011	mouse	mES-ELL	0.79	1.71	0.79	0	40	40	40	16,754,758	ChIP	yes
Smith et al. 2011	mouse	mES-Input	0.96	1.38	0.76	0	40	40	40	19,454,353	Input	no
Nakayamada et al. 2011	mouse	CD4+-Tbet	0.56	3.09	1.78	2	36	36	36	23,421,318	ChIP	yes
Lu et al. 2012	human	IgG-1-BCBL1	0.94	2.23	1.07	1	36	36	36	19,209,840	IgG	no
Lu et al. 2012	human	IgG-2-BCBL1	0.81	1.44	0.3	-1	36	36	36	12,604,075	IgG	no
Lu et al. 2012	human	LANA-1-BCBL1	0.46	1.6	0.53	0	36	36	36	19,777,228	ChIP	yes
Lu et al. 2012	human	LANA-2-BCBL1	0.86	1.44	0.27	-1	36	36	36	11,880,205	ChIP	yes
Meyer et al. 2012	human	LS180-bCat-125-1	0.43	13.11	1.87	2	35.43	36	35	39,712,224	ChIP	yes
Meyer et al. 2012	human	LS180-bCat-Veh-1	0.43	6.85	2.83	2	35.1	36	35	25,024,509	ChIP	yes
Meyer et al. 2012	human	LS180-CDX2-125-1	0.57	4.36	1.81	2	35.59	36	35	38,267,118	ChIP	yes
Meyer et al. 2012	human	LS180-CDX2-Veh-1	0.56	4.57	1.71	2	35.6	36	35	34,581,066	ChIP	yes
Meyer et al. 2012	human	LS180-CEBPb-125-1	0.81	8.48	1.8	2	35	35	35	24,978,947	ChIP	yes
Meyer et al. 2012	human	LS180-CEBPb-Veh-1	0.05	8.15	1.82	2	35.75	36	35	78,542,681	ChIP	yes
Meyer et al. 2012	human	LS180-Input-1	0.15	1.83	1.02	1	35.66	36	35	54,134,263	Input	no
Meyer et al. 2012	human	LS180-RXR-125-1	0.09	7.72	1.64	2	36	36	36	29,948,896	ChIP	yes
Meyer et al. 2012	human	LS180-RXR-Veh-1	0.1	7.3	1.79	2	36	36	36	26,448,441	ChIP	yes
Meyer et al. 2012	human	LS180-TCF4-125-2	0.28	10.15	1.78	2	45.01	50	35	49,453,419	ChIP	yes
Meyer et al. 2012	human	LS180-TCF4-Veh-2	0.36	10.26	1.9	2	42.43	50	35	20,780,670	ChIP	yes
Meyer et al. 2012	human	LS180-VDR-125-1	0.23	5.74	3.37	2	36	36	36	4,734,750	ChIP	yes
Meyer et al. 2012	human	LS180-VDR-Veh-1	0.18	11.61	5.58	2	35.79	36	35	72,061,937	ChIP	unknown
Ntziachristos et al. 2012	mouse	DP-mnase-input-replicate-1	0.92	2.74	1.13	1	34	34	34	15,457,880	Input	no
Ntziachristos et al. 2012	mouse	DP-mnase-input-replicate-2	0.9	6.08	3.06	2	34	34	34	12,676,911	Input	no
Ntziachristos et al. 2012	mouse	T-ALL-mnase-input-replicate-1	0.58	1.7	0.18	-2	34	34	34	9,970,383	Input	no
Ntziachristos et al. 2012	mouse	T-ALL-mnase-input-replicate-2	0.86	2.17	0.68	0	34	34	34	12,351,316	Input	no
Ntziachristos et al. 2012	mouse	T-ALL-Notch1	0.75	2.23	1.97	2	34	34	34	15,248,670	ChIP	yes
Ntziachristos et al. 2012	mouse	T-ALL-sonicated-input	0.7	1.28	0.17	-2	34	34	34	12,479,110	Input	no
Cheng et al. 2012	human	Gdown1-Control	0.97	1.69	0.66	0	36	36	36	3,798,010	ChIP	yes
Cheng et al. 2012	human	Gdown1-Flavo	0.93	1.77	0.74	0	36	36	36	7,869,560	ChIP	yes
GSE33128	human	Gdown1-IMR90	0.67	2.83	1.46	1	36	36	36	13,781,340	ChIP	yes
GSE33128	human	IgG-IMR90	0.67	7.4	2.05	2	36	36	36	7,308,478	IgG	no
GSE33128	human	Input-IMR90	0.96	1.47	0.69	0	36	36	36	14,239,395	Input	no
GSE35109	human	ERa-ChIP-seq-1	0.86	1.46	1.99	2	51	51	51	48,891,564	ChIP	yes
GSE35109	human	ERa-ChIP-seq-2	0.7	2.02	3.17	2	51	51	51	52,808,583	ChIP	yes
GSE35109	human	ERa-ChIP-seq-3	0.3	5.75	5.72	2	51	51	51	46,155,863	ChIP	yes
GSE35109	human	ERa-ChIP-seq-4	0.8	1.64	2.76	2	51	51	51	57,965,746	ChIP	yes
Canella et al. 2012	mouse	INPUT-Rep1	0.8	1.37	1.93	2	75	75	75	31,537,710	Input	no
Canella et al. 2012	mouse	INPUT-Rep2	0.8	1.38	1.96	2	75	75	75	33,328,402	Input	no
Canella et al. 2012	mouse	RPB2-Rep1	0.8	1.55	1.74	2	75	75	75	35,847,372	ChIP	yes
Canella et al. 2012	mouse	RPB2-Rep2	0.83	1.94	1.75	2	75	75	75	30,551,646	ChIP	yes
Canella et al. 2012	mouse	RPC1-Rep1	0.9	1.9	1.48	1	75	75	75	23,033,105	ChIP	yes
Canella et al. 2012	mouse	RPC1-Rep2	0.91	1.72	1.31	1	75	75	75	22,145,329	ChIP	yes
Canella et al. 2012	mouse	RPC4-Rep1	0.87	1.9	1.49	1	75	75	75	25,973,018	ChIP	yes
Canella et al. 2012	mouse	RPC4-Rep2	0.86	1.83	1.6	2	75	75	75	31,517,301	ChIP	yes
Sadasivam et al. 2012	human	BMyb-HeLa-Rep1	0.88	1.36	0.32	-1	36	36	36	15,389,344	ChIP	yes
Sadasivam et al. 2012	human	BMyb-HeLa-Rep2	0.77	1.81	0.07	-2	36	36	36	1,052,761	ChIP	yes
Sadasivam et al. 2012	human	Input-HeLa-Rep1	0.71	1.79	0.69	0	36	36	36	17,569,472	Input	no
Sadasivam et al. 2012	human	Input-HeLa-Rep2	0.64	1.56	0.08	-2	36	36	36	1,586,928	Input	no
Sadasivam et al. 2012	human	LIN9-HeLa-Rep1	0.92	1.39	0.49	-1	36	36	36	17,000,309	ChIP	yes
Sadasivam et al. 2012	human	LIN9-HeLa-Rep2	0.88	1.41	0.05	-2	36	36	36	1,798,161	ChIP	yes
Boergesen et al. 2012	mouse	LXR-WT-Bexarotene	0.93	1.94	2.05	2	35	35	35	6,469,307	ChIP	yes
Boergesen et al. 2012	mouse	LXR-WT-Control	0.9	1.91	1.45	1	35	35	35	6,086,575	ChIP	unknown

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Boergesen et al. 2012	mouse	LXR-WT-T0901317	0.93	2.21	3.81	2	35	35	35	6,773,502	ChIP	unknown
Boergesen et al. 2012	mouse	PPARalpha-LXRdKO-Control	0.69	4.14	4.68	2	35	35	35	12,603,632	ChIP	yes
Boergesen et al. 2012	mouse	PPARalpha-WT-Control	0.66	4.02	10.05	2	35	35	35	13,493,293	ChIP	yes
Boergesen et al. 2012	mouse	RXR-LXRdKO-Bexarotene	0.96	2.71	2	2	34	34	34	4,499,835	ChIP	yes
Boergesen et al. 2012	mouse	RXR-LXRdKO-Control	0.92	2.7	1.99	2	32	32	32	5,011,146	ChIP	yes
Boergesen et al. 2012	mouse	RXR-LXRdKO-T0901317	0.94	2.14	1.9	2	32	32	32	5,048,268	ChIP	unknown
Boergesen et al. 2012	mouse	RXR-WT-Bexarotene	0.94	2.57	1.82	2	34	34	34	4,819,549	ChIP	yes
Boergesen et al. 2012	mouse	RXR-WT-Control	0.95	1.96	1.33	1	32	32	32	5,847,078	ChIP	yes
Boergesen et al. 2012	mouse	RXR-WT-T0901317	0.93	3.08	2.13	2	32	32	32	5,510,973	ChIP	unknown
Schödel et al. 2012	human	HIF-1beta	0.37	3.41	1.44	1	51	51	51	7,729,167	ChIP	yes
Schödel et al. 2012	human	HIF-2alpha	0.64	3.11	1.21	1	51	51	51	1,885,345	ChIP	yes
Schödel et al. 2012	human	Pre-immune-control	0.29	4.94	1.87	2	51	51	51	5,806,061	IgG	no
Pehkonen et al. 2012	human	IgG-control	0.72	1.41	0.24	-2	36	36	36	15,281,888	IgG	no
Pehkonen et al. 2012	human	LXR-T09	0.87	1.47	0.29	-1	36	36	36	14,265,491	ChIP	yes
Pehkonen et al. 2012	human	LXR-vehicle	0.9	1.42	0.27	-1	36	36	36	14,289,777	ChIP	unknown
GSE30919	mouse	CapH2-Ab1-DMSO-NOT-NORMALIZED-mES-MM8	0.69	1.61	0.91	0	36	36	36	16,534,945	ChIP	yes
GSE30919	mouse	CapH2-Ab1-FLAVO-NOT-NORMALIZED-mES-MM8	0.71	1.61	0.83	0	36	36	36	15,830,789	ChIP	yes
GSE30919	mouse	CapH2-Ab1-WT-mES-MM8	0.66	1.73	0.92	0	36	36	36	16,607,056	ChIP	yes
GSE30919	mouse	CapH2-Ab2-WT-mES-MM8	0.9	1.41	0.78	0	36	36	36	17,717,075	ChIP	yes
GSE30919	mouse	Smc1-DMSO-NOT-NORMALIZED-mES-MM8	0.84	5.22	1.99	2	36	36	36	19,206,320	ChIP	yes
GSE30919	mouse	Smc1-FLAVO-NOT-NORMALIZED-mES-MM8	0.78	4.43	1.88	2	36	36	36	19,650,774	ChIP	yes
Gao et al. 2012	human	CBX2	0.83	1.51	0.52	0	46	46	46	11,796,622	ChIP	yes
Gao et al. 2012	human	FH-CBX2.HA	0.89	1.22	0.37	-1	36	36	36	20,303,587	ChIP	yes
Gao et al. 2012	human	FH-PCGF1.HA	0.82	1.35	0.35	-1	36	36	36	18,667,442	ChIP	yes
Gao et al. 2012	human	FH-PCGF2.HA	0.66	1.9	0.67	0	36	36	36	18,549,373	ChIP	yes
Gao et al. 2012	human	FH-PCGF4.HA	0.31	2.44	0.81	0	36	36	36	18,274,491	ChIP	yes
Gao et al. 2012	human	FH-PCGF5.HA	0.66	1.78	0.65	0	36	36	36	18,930,930	ChIP	yes
Gao et al. 2012	human	FH-PCGF6.HA	0.8	1.48	0.43	-1	36	36	36	19,548,786	ChIP	yes
Gao et al. 2012	human	FH-RING1B.HA	0.43	1.94	1.58	2	36	36	36	19,398,688	ChIP	yes
Gao et al. 2012	human	FH-RYBP.HA	0.83	1.32	0.36	-1	36	36	36	16,950,286	ChIP	yes
Gao et al. 2012	human	input	0.78	1.23	0.22	-2	36	36	36	19,426,459	Input	no
Gao et al. 2012	human	PCGF4	0.93	1.16	0.19	-2	46	46	46	14,654,954	ChIP	yes
Gao et al. 2012	human	RING1B	0.9	1.18	0.24	-2	46	46	46	19,431,342	ChIP	yes
Gao et al. 2012	human	RYBP	0.91	1.36	0.44	-1	46	46	46	15,442,467	ChIP	yes
Yu et al. 2012	mouse	CBFb-induced-1	0.62	1.62	2.88	2	39.24	40	36	58,627,013	ChIP	yes
Yu et al. 2012	mouse	CBFb-thymocyte-control	0.25	1.63	0.62	0	40	40	40	8,637,405	ChIP	yes
Yu et al. 2012	mouse	CBFb-thymocyte-Runx1KO	0.35	1.4	0.54	0	40	40	40	7,518,656	ChIP	yes
Yu et al. 2012	mouse	CBFb-uninduced-1	0.73	1.58	2.28	2	36	36	36	26,748,905	ChIP	yes
Yu et al. 2012	mouse	IgG-induced-1	0.13	4.73	3.85	2	39.39	40	36	60,744,963	IgG	no
Yu et al. 2012	mouse	IgG-thymocyte-control	0.11	7.03	0.93	0	40	40	40	10,387,710	IgG	no
Yu et al. 2012	mouse	IgG-thymocyte-Runx1KO	0.04	9.33	0.8	0	40	40	40	11,696,369	IgG	no
Yu et al. 2012	mouse	IgG-uninduced-1	0.14	3.57	1.51	2	36	36	36	30,535,688	IgG	no
Yu et al. 2012	mouse	Ring1b-alt-ab	0.11	8.57	1.74	2	40	40	40	17,104,492	ChIP	yes
Yu et al. 2012	mouse	Ring1b-induced-1	0.67	1.5	2.59	2	39.31	40	36	65,911,236	ChIP	yes
Yu et al. 2012	mouse	Ring1b-thymocyte-control	0.21	3.38	0.58	0	40	40	40	7,476,406	ChIP	yes
Yu et al. 2012	mouse	Ring1b-thymocyte-Runx1KO	0.1	8.31	1.1	1	40	40	40	11,067,615	ChIP	yes
Yu et al. 2012	mouse	Ring1b-uninduced-1	0.38	2.1	3.58	2	36	36	36	31,481,625	ChIP	yes
Yu et al. 2012	mouse	Runx1-for-Ring1b-alt-ab	0.2	6.16	2.13	2	40	40	40	14,979,699	ChIP	yes
Yu et al. 2012	mouse	Runx1-induced-1	0.45	1.72	3.1	2	39.26	40	36	65,873,746	ChIP	yes
Yu et al. 2012	mouse	Runx1-thymocyte-control	0.24	5.51	1.23	1	40	40	40	8,075,699	ChIP	yes
Yu et al. 2012	mouse	Runx1-thymocyte-Runx1KO	0.17	4.65	0.84	0	40	40	40	9,234,112	ChIP	no
Yu et al. 2012	mouse	Runx1-uninduced-1	0.65	1.55	1.73	2	36	36	36	23,915,032	ChIP	yes
GSE29180	human	Jurkat-GATA3	0.75	4.85	0.85	0	36	36	36	4,308,315	ChIP	yes
GSE29180	human	Jurkat-Input-Rep1	0.97	1.19	0.42	-1	36	36	36	12,308,677	Input	no
GSE29180	human	Jurkat-RUNX1-Rep1	0.61	2.83	0.82	0	36	36	36	5,791,954	ChIP	yes
GSE29180	human	Jurkat-RUNX1-Rep2	0.53	2.44	0.53	0	36	36	36	5,800,696	ChIP	yes
GSE29180	human	Jurkat-RUNX1-Rep3	0.44	4.33	0.89	0	36	36	36	3,795,121	ChIP	yes
GSE29180	human	Jurkat-TAL1-Rep1	0.92	1.62	0.38	-1	36	36	36	5,485,444	ChIP	yes

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
GSE29180	human	Jurkat-TAL1-Rep2	0.86	2.59	0.68	0	36	36	36	6,350,195	ChIP	yes
GSE29180	human	Jurkat-TCF12	0.81	1.53	0.38	-1	36	36	36	8,594,457	ChIP	yes
GSE29180	human	Jurkat-TCF3	0.84	1.51	0.24	-2	36	36	36	5,398,758	ChIP	yes
Sakabe et al. 2012	mouse	input-1	0.95	1.86	0.47	-1	36	36	36	6,264,090	Input	no
Sakabe et al. 2012	mouse	input-2	0.94	1.38	0.41	-1	36	36	36	10,837,874	Input	no
Sakabe et al. 2012	mouse	input-3	0.95	1.21	0.48	-1	36	36	36	23,754,878	Input	no
Sakabe et al. 2012	mouse	Tbx20-GFP	0.87	1.9	0.63	0	36	36	36	5,928,909	ChIP	yes
Miller et al. 2012	human	HCC-1428-LTED-ER	0.92	1.2	0.36	-1	43	43	43	23,589,680	ChIP	yes
Miller et al. 2012	human	MCF-7-LTED-ER	0.93	1.15	0.55	0	43	43	43	27,118,853	ChIP	yes
Hutchins et al. 2012	mouse	PEC-IL10-treated-Input	0.97	1.39	0.25	-1	49	49	49	4,244,316	Input	no
Hutchins et al. 2012	mouse	PEC-IL10-treated-STAT3	0.71	5.06	1.09	1	49	49	49	3,841,121	ChIP	yes
Hutchins et al. 2012	mouse	PEC-Untreated-Input	0.97	1.39	0.17	-2	49	49	49	4,321,159	Input	no
Hutchins et al. 2012	mouse	PEC-Untreated-STAT3	0.83	3.12	0.73	0	49	49	49	4,189,247	ChIP	unknown
Trowbridge et al. 2012	mouse	MLL1	0.96	1.44	0.35	-1	36	36	36	4,933,023	ChIP	yes
Xiao et al. 2012	mouse	E14-IgG	0.58	2.13	0.42	-1	100	100	100	3,823,799	IgG	no
Xiao et al. 2012	mouse	E14-TAF1	0.95	1.07	0.37	-1	75	75	75	22,675,646	ChIP	yes
Xiao et al. 2012	pig	piPSC-IgG	0.62	2.33	0.45	-1	75	75	75	3,237,532	IgG	no
Xiao et al. 2012	pig	piPSC-NANOG	0.79	1.06	0.2	-2	75	75	75	15,130,135	ChIP	yes
Xiao et al. 2012	pig	piPSC-OCT4	0.84	1.58	0.63	0	75	75	75	4,150,813	ChIP	yes
Xiao et al. 2012	pig	piPSC-p300	0.86	1.07	0.16	-2	75	75	75	27,328,401	ChIP	yes
Xiao et al. 2012	pig	piPSC-TAF1	0.76	1.06	0.23	-2	75	75	75	8,822,964	ChIP	yes
Doré et al. 2012; Chlon et al. 2012	mouse	G1ME-ETS1	0.91	1.37	1.33	1	36	36	36	31,187,821	ChIP	yes
Doré et al. 2012; Chlon et al. 2012	mouse	G1ME-GATA1	0.52	1.77	2.14	2	36	36	36	35,032,324	ChIP	yes
Doré et al. 2012; Chlon et al. 2012	mouse	G1ME-GATA2	0.96	1.8	1.22	1	36	36	36	10,496,766	ChIP	yes
Doré et al. 2012; Chlon et al. 2012	mouse	G1ME-INPUT-GAII	0.93	1.23	0.19	-2	36	36	36	10,209,628	Input	no
Doré et al. 2012; Chlon et al. 2012	mouse	G1ME-INPUT-GAIIx	0.63	1.42	1.59	2	36	36	36	20,517,340	Input	no
Li et al. 2012	mouse	Input-seq-Adr8h	0.86	1.98	1.75	2	35	35	35	22,456,496	Input	no
Li et al. 2012	mouse	Input-seq-untreated	0.68	3.73	4.81	2	35	35	35	24,631,682	Input	no
Li et al. 2012	mouse	p53-Adr8h	0.74	8.66	2.17	2	35	35	35	22,316,127	ChIP	yes
Li et al. 2012	mouse	p53-untreated	0.95	5.29	1.57	2	35	35	35	9,544,532	ChIP	unknown
Li et al. 2012	mouse	p53S18P-Adr8h	0.92	14.22	1.65	2	35	35	35	9,487,356	ChIP	yes
Li et al. 2012	mouse	p53S18P-untreated	0.91	2.65	1.59	2	35	35	35	15,417,989	ChIP	unknown
Bugge et al. 2012; Feng et al. 2012	mouse	Reverb-alpha-null-5pm	0.63	1.91	1.73	2	50	50	50	82,551,235	ChIP	yes
Bugge et al. 2012; Feng et al. 2012	mouse	Reverb-beta-5am	0.83	1.7	0.73	0	36	36	36	7,098,042	ChIP	unknown
Bugge et al. 2012; Feng et al. 2012	mouse	Reverb-beta-5pm	0.18	2.08	1.77	2	36	36	36	39,165,327	ChIP	unknown
Gowher et al. 2012	human	HA-flag-Vezf1-Rep1	0.91	1.22	0.5	0	36	36	36	41,807,364	ChIP	yes
Gowher et al. 2012	human	HA-flag-Vezf1-Rep2	0.94	1.22	0.38	-1	36	36	36	10,730,653	ChIP	yes
Gowher et al. 2012	human	Input-HELA-Rep1	0.91	1.34	0.96	0	36	36	36	39,886,595	Input	no
Gowher et al. 2012	human	Input-HELA-Rep2	0.96	1.32	0.58	0	36	36	36	10,704,869	Input	no
Gowher et al. 2012	mouse	Input-mm9ES-wt	0.96	1.31	0.86	0	36	36	36	14,112,421	Input	no
Gowher et al. 2012	mouse	Input-Vezf1-ko	0.96	1.35	0.83	0	36	36	36	13,596,489	Input	no
GSE33346	mouse	CapD3-Nocodazole-mES	0.66	1.78	0.96	0	36	36	36	23,506,234	ChIP	unknown
GSE33346	mouse	CapD3-WT-mES	0.77	2.75	1.33	1	36	36	36	20,944,575	ChIP	yes
GSE33346	mouse	CapG-Nocodazole-mES	0.72	1.45	0.63	0	36	36	36	22,267,698	ChIP	unknown
GSE33346	mouse	CapG-WT-mES	0.73	1.62	1.11	1	36	36	36	23,314,867	ChIP	yes
GSE33346	mouse	CapH2-Nocodazole-mES	0.42	2.65	1.94	2	36	36	36	19,469,725	ChIP	unknown
GSE33346	mouse	CapH2-shGFP-mES	0.81	1.57	1.18	1	36	36	36	22,027,077	ChIP	yes
GSE33346	mouse	CapH2-shNipbl-mES	0.49	2.23	1.31	1	36	36	36	21,121,437	ChIP	unknown
GSE33346	mouse	Rad21-rep1-WT-mES	0.93	12.57	1.37	1	36	36	36	14,695,398	ChIP	yes
GSE33346	mouse	Rad21-rep2-WT-mES	0.85	13.29	2.19	2	36	36	36	20,290,096	ChIP	yes
GSE33346	mouse	WCE-Nocodazole-mES	0.61	1.54	0.97	0	36	36	36	22,934,718	Input	no
GSE33346	mouse	WCE-shGFP-mES	0.9	1.32	0.93	0	36	36	36	20,882,926	Input	no
GSE33346	mouse	WCE-shNipbl-mES	0.81	1.47	0.68	0	36	36	36	8,493,397	Input	no
GSE33850	human	E2A-CCRF-CEM	0.2	3.3	0.46	-1	40	40	40	9,580,539	ChIP	yes
GSE33850	human	GATA3-CCRF-CEM	0.14	4.27	0.54	0	40	40	40	8,433,815	ChIP	yes
GSE33850	human	HEB-CCRF-CEM-Rep1	0.11	5.07	0.79	0	39	39	39	11,256,332	ChIP	yes
GSE33850	human	HEB-CCRF-CEM-Rep2	0.15	4.66	0.66	0	40	40	40	13,394,868	ChIP	yes
GSE33850	human	Input-WCE-CCRF-CEM-Rep1	0.21	3.28	0.24	-2	39	39	39	3,524,267	Input	no
GSE33850	human	Input-WCE-CCRF-CEM-Rep2	0.81	1.33	0.08	-2	40	40	40	4,060,683	Input	no
GSE33850	human	Input-WCE-Prima2-T-ALL-Rep1	0.08	2.97	1.63	2	39	39	39	11,209,256	Input	no
GSE33850	human	Input-WCE-Prima2-T-ALL-Rep2	0.22	1.72	0.53	0	40	40	40	11,519,126	Input	no
GSE33850	human	Input-WCE-Prima5-T-ALL-Rep1	0.21	2.32	0.69	0	39	39	39	9,218,972	Input	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
GSE33850	human	Input-WCE-Prima5-T-ALL-Rep2	0.46	1.54	0.46	-1	40	40	40	10,635,018	Input	no
GSE33850	human	LMO1-Jurkat-Rep1	0.27	1.67	0.12	-2	40	40	40	6,940,746	ChIP	yes
GSE33850	human	LMO1-Jurkat-Rep2	0.57	1.79	0.27	-1	36	36	36	5,951,620	ChIP	yes
GSE33850	human	LMO2-CCRF-CEM-Rep1	0.11	2.74	0.28	-1	39	39	39	10,129,558	ChIP	yes
GSE33850	human	LMO2-CCRF-CEM-Rep2	0.15	2.72	0.25	-1	40	40	40	6,136,649	ChIP	yes
GSE33850	human	RUNX1-CCRF-CEM-Rep1	0.03	82.73	4.18	2	39	39	39	8,181,063	ChIP	yes
GSE33850	human	RUNX1-CCRF-CEM-Rep2	0.38	2.56	0.46	-1	40	40	40	12,118,147	ChIP	yes
GSE33850	human	TAL1-CCRF-CEM-Rep1	0.14	5.99	1.07	1	39	39	39	8,072,878	ChIP	yes
GSE33850	human	TAL1-CCRF-CEM-Rep2	0.17	3.12	0.68	0	40	40	40	17,651,204	ChIP	yes
GSE33850	human	TAL1-Prima2-T-ALL-Rep1	0.08	10.81	1.54	2	40	40	40	4,774,060	ChIP	yes
GSE33850	human	TAL1-Prima2-T-ALL-Rep2	0.05	10.97	1.65	2	39	39	39	7,554,079	ChIP	yes
GSE33850	human	TAL1-Prima5-T-ALL-Rep1	0.07	6.54	1.44	1	40	40	40	6,603,228	ChIP	yes
GSE33850	human	TAL1-Prima5-T-ALL-Rep2	0.05	7.04	1.56	2	39	39	39	9,252,579	ChIP	yes
Avvakumov et al. 2012	human	HBO1	0.96	1.24	1	0	36	36	36	31,901,032	ChIP	yes
Avvakumov et al. 2012	human	input	0.98	1.19	0.55	0	36	36	36	31,414,277	Input	no
Hunkapiller et al. 2012	mouse	InputDNA-Pcl3-shRNA	0.64	4.68	1.18	1	30	30	30	14,811,561	Input	no
Hunkapiller et al. 2012	mouse	InputDNA-Pcl3-shRNA6	0.95	1.2	0.4	-1	36	36	36	15,249,656	Input	no
Hunkapiller et al. 2012	mouse	InputDNA-Pcl3-shRNA7	0.95	1.22	0.55	0	36	36	36	19,965,283	Input	no
Hunkapiller et al. 2012	mouse	InputDNA-scramble	0.58	5.36	1.45	1	30	30	30	11,650,029	Input	no
Hunkapiller et al. 2012	mouse	Pcl3-shRNA6	0.88	1.28	1.38	1	36	36	36	14,295,321	ChIP	no
Hunkapiller et al. 2012	mouse	Pcl3-shRNA7	0.86	1.41	0.8	0	36	36	36	10,534,049	ChIP	no
Hunkapiller et al. 2012	mouse	Suz12-Pcl3-shRNA	0.78	1.41	1.43	1	30	30	30	13,893,316	ChIP	unknown
Hunkapiller et al. 2012	mouse	Suz12-scramble	0.78	1.77	1.38	1	30	30	30	11,020,925	ChIP	yes
Remeseiro et al. 2012	mouse	Input	0.89	1.22	0.56	0	40	40	40	25,401,900	Input	no
Remeseiro et al. 2012	mouse	InputMEFs	0.93	1.14	0.44	-1	40	40	40	27,631,354	Input	no
Remeseiro et al. 2012	mouse	KO-SA1	0.86	1.35	0.64	0	40	40	40	20,865,198	ChIP	no
Remeseiro et al. 2012	mouse	KO-SA2	0.79	1.49	0.95	0	40	40	40	26,737,423	ChIP	unknown
Remeseiro et al. 2012	mouse	SMC1-KO.R1	0.78	4.13	1.66	2	40	40	40	9,276,356	ChIP	unknown
Remeseiro et al. 2012	mouse	SMC1-KO.R2	0.82	3.98	1.93	2	40	40	40	12,183,058	ChIP	unknown
Remeseiro et al. 2012	mouse	SMC1-WT	0.91	1.91	1.38	1	40	40	40	22,390,032	ChIP	yes
Remeseiro et al. 2012	mouse	SMC3-KO	0.88	2.54	2.01	2	40	40	40	27,111,387	ChIP	unknown
Remeseiro et al. 2012	mouse	SMC3-WT	0.9	1.48	1.03	1	40	40	40	25,310,295	ChIP	yes
Remeseiro et al. 2012	mouse	WT-SA1	0.78	4.43	2.46	2	40	40	40	26,143,843	ChIP	yes
Remeseiro et al. 2012	mouse	WT-SA2	0.65	2.12	1.59	2	40	40	40	25,387,005	ChIP	yes
GSE36561	mouse	Brd4-mES	0.94	1.47	1.14	1	36	36	36	18,715,973	ChIP	yes
GSE36561	mouse	Brg1-mES	0.92	1.62	0.42	-1	36	36	36	4,204,507	ChIP	yes
GSE36561	mouse	SA1-mES-Rep1	0.95	13.34	2.23	2	36	36	36	6,935,496	ChIP	yes
GSE36561	mouse	SA1-mES-Rep2	0.84	21.27	2.29	2	36	36	36	18,853,613	ChIP	yes
GSE36561	mouse	SA2-mES-Rep1	0.94	16.15	2.13	2	36	36	36	7,883,128	ChIP	yes
GSE36561	mouse	SA2-mES-Rep2	0.84	15.88	2.29	2	36	36	36	18,512,023	ChIP	yes
Vilagos et al. 2012	mouse	EBF1-8246.2	0.96	3.03	1.04	1	36	36	36	5,435,592	ChIP	yes
Vilagos et al. 2012	mouse	EBF1-8246.6	0.96	3.1	1.13	1	36	36	36	7,748,856	ChIP	yes
Vilagos et al. 2012	mouse	EBF1-mature-B-8271	0.94	1.89	0.41	-1	36	36	36	5,327,224	ChIP	yes
Vilagos et al. 2012	mouse	EBF1-mature-B-9842	0.51	2.34	1.05	1	36	36	36	16,361,104	ChIP	yes
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8091.5	0.72	2.43	0.08	-2	36	36	36	2,188,795	Input	no
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8091.6	0.73	2.22	0.08	-2	36	36	36	2,267,935	Input	no
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8112.1	0.97	1.39	0.17	-2	36	36	36	4,627,018	Input	no
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8112.6	0.96	1.28	0.19	-2	36	36	36	6,424,234	Input	no
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8123.2	0.51	2.4	0.08	-2	36	36	36	2,220,931	Input	no
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8149.8.301DTAAXX	0.81	3.37	0.19	-2	36	36	36	1,534,780	Input	no

Continued on next page

Table 10.1 – *Continued from previous page*

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Max. Read Length	Min. Read Length	Mapped reads	Type	Should exhibit read clustering
Vilagos et al. 2012	mouse	Rag2.Pro-B.input-8149.8.30222AAXX	0.86	2.37	0.57	0	36	36	36	5,533,095	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8042.5.208KBAAXX	0.87	1.32	0.1	-2	34	34	34	4,297,854	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8042.7.207JYAAXX	0.89	1.57	0.12	-2	36	36	36	3,133,666	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8042.7.20CUYAAXX	0.88	1.66	0.13	-2	36	36	36	3,220,120	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8042.8.208KDAAXX	0.87	3.9	0.2	-2	36	36	36	1,082,634	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8087	0.97	1.6	0.18	-2	32	32	32	4,201,759	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8089	0.96	1.59	0.22	-2	36	36	36	4,101,017	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8094	0.96	1.53	0.16	-2	36	36	36	3,717,876	Input	no
Vilagos et al. 2012	mouse	WT.Mature-B.Input-8096	0.95	1.33	0.17	-2	36	36	36	5,480,836	Input	no
Cardamone et al. 2012	human	GPS2	0.89	1.77	1.2	1	76	76	76	8,251,524	ChIP	yes
Cardamone et al. 2012	human	NCOR-siCTL	0.67	3	2.11	2	36.01	44	36	6,572,892	ChIP	yes
Cardamone et al. 2012	human	NCOR-siGPS2	0.72	2.5	0.78	0	36.01	44	36	5,121,903	ChIP	unknown
Cardamone et al. 2012	human	TBL1	0.87	1.87	1.93	2	36	36	36	9,798,221	ChIP	yes
Fan et al. 2012	mouse	HoxB4-day-16	0.95	2.1	1.08	1	41	41	41	8,877,542	ChIP	yes
Fan et al. 2012	mouse	HoxB4-day-26	0.94	4.72	1.79	2	41	41	41	10,871,546	ChIP	yes
Fan et al. 2012	mouse	HoxB4-day-6	0.92	2.36	1.63	2	36	36	36	6,336,688	ChIP	yes
Fan et al. 2012	mouse	Input-day-16	0.97	1.41	0.72	0	41	41	41	12,098,959	Input	no
Fan et al. 2012	mouse	Input-day-26	0.97	1.45	0.79	0	41	41	41	11,607,750	Input	no
Fan et al. 2012	mouse	Input-day-6	0.97	1.22	0.3	-1	36	36	36	8,817,894	Input	no
Fong et al. 2012	mouse	MM-MyoD	0.84	8.25	1.83	2	39	39	39	21,182,386	ChIP	yes
Fong et al. 2012	mouse	MM-NeuroD2	0.92	5.14	1.67	2	39	39	39	13,996,908	ChIP	yes
Fong et al. 2012	mouse	P19-control	0.97	1.42	0.56	0	38	39	37	8,903,023	IgG	no
Fong et al. 2012	mouse	P19-MyoD	0.92	12.89	1.94	2	39	39	39	12,117,729	ChIP	yes
Fong et al. 2012	mouse	P19-NeuroD2	0.94	7.18	1.67	2	39	39	39	14,558,083	ChIP	yes
Ptasinska et al. 2012	human	Input	0.88	1.35	0.2	-2	40	40	40	5,280,044	Input	no
Ptasinska et al. 2012	human	RUNX1-Kasumi-1	0.97	1.37	0.83	0	43.34	80	36	17,904,797	ChIP	yes
Ptasinska et al. 2012	human	RUNX1-non-t-8-21	0.91	3.67	1.81	2	36	36	36	30,747,325	ChIP	yes
Ptasinska et al. 2012	human	RUNX1ETO-control	0.95	1.79	0.97	0	75.95	80	40	7,462,090	ChIP	yes
Ptasinska et al. 2012	human	RUNX1ETO-siMM	0.94	1.65	0.97	0	73.57	80	40	12,843,591	ChIP	yes
Ptasinska et al. 2012	human	RUNX1ETO-siRE	0.82	2.82	1.2	1	67.36	80	40	5,525,324	ChIP	no
Cho et al. 2012	mouse	liver-input	0.78	1.54	1.25	1	42	42	42	29,085,894	Input	no
Cho et al. 2012	mouse	REV-ERBalpha	0.89	2.05	1.69	2	42	42	42	32,677,790	ChIP	yes
Cho et al. 2012	mouse	REV-ERBbeta	0.65	2.15	2.84	2	42	42	42	28,812,418	ChIP	yes
Wu et al. 2012	mouse	input-RUNX1	0.97	1.26	0.58	0	34	34	34	11,771,941	Input	no
Wu et al. 2012	mouse	input-TCF7	0.96	1.2	0.82	0	36	36	36	22,172,123	Input	no
Wu et al. 2012	mouse	RUNX1-Rep1	0.71	3.8	2.2	2	34	34	34	9,285,076	ChIP	yes
Wu et al. 2012	mouse	RUNX1-Rep2	0.68	4.01	2.32	2	34	34	34	10,064,029	ChIP	yes
Wu et al. 2012	mouse	TCF7	0.83	1.85	1	1	36	36	36	13,877,190	ChIP	yes
Barish et al. 2012	mouse	Bcl6-KO-macrophage-NCoR	0.66	1.75	1.37	1	42	42	42	25,491,046	ChIP	yes
Barish et al. 2012	mouse	Bcl6-KO-macrophage-SMRT	0.81	1.51	1.14	1	42	42	42	25,610,348	ChIP	yes
Barish et al. 2012	mouse	WT-macrophage-NCoR	0.84	1.81	1.79	2	43	43	43	24,281,787	ChIP	yes
Barish et al. 2012	mouse	WT-macrophage-SMRT	0.62	2.05	2.21	2	43	43	43	27,456,911	ChIP	yes

* Note: datasets from Trompouki et al. 2011 were excluded from figures as the vast majority of them had a very low number of mapped reads.

Table 10.2: Dataset QC evaluation and mapping statistics for MyoD and myogenin datasets

Source	Species	Library	Complexity	NSC	RSC	QC	Ave. Read Length	Min. Read Length	Max. Read Length	Mapped reads	Type	Should exhibit read clustering
Wold Lab	mouse	C2C12 60h MyoD	0.90	12.39	1.65	2	36	36	36	6,771,837	ChIP	yes
Wold Lab	mouse	C2C12 60h myogenin 1	0.88	9.21	1.93	2	36	36	36	10,385,089	ChIP	yes
Wold Lab	mouse	C2C12 60h myogenin 2	0.97	6.95	1.32	1	36	36	36	1,198,656	ChIP	yes
Wold Lab	mouse	C2C12 60h myogenin 3	0.93	1.20	0.40	-1	36	36	36	19,600,577	ChIP	yes
Wold Lab	mouse	C2C12 60h 1%FA Input 3	0.94	1.22	0.46	-1	36	36	36	17,856,564	ChIP	no
Wold Lab	mouse	C2C12 60h 1%FA+EGS Input 3	0.87	4.88	1.52	2	36	36	36	9,092,000	ChIP	no

11

High-throughput robotic chromatin immunoprecipitation for ChIP-seq (R-ChIP)

The material in this chapter has, at the time of writing this, prepared for publication as:

Gasper WG, Marinov GK, Pauli-Behn F, Scott MT, Newberry K, deSalvo G, Ou S, Myers RM, Vielmetter J, Wold BJ. Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: Identifying ChIP-quality p300 monoclonal antibodies.

The R-ChIP protocol was developed by Clarke Gasper and Jost Vielmetter. My contribution was in analyzing the data and writing the manuscript.

Abstract

Chromatin immunoprecipitation coupled with DNA sequencing (ChIP-seq) is the major contemporary method for mapping in vivo protein-DNA interactions in the genome. It identifies sites of transcription factor, cofactor and polymerase occupancy, as well as the distribution of histone marks. Consortia such as the ENCyclopedia Of DNA Elements (ENCODE) and the NIH Roadmap Epigenomics Mapping Consortium have produced large datasets over a period of several years using manual protocols. However, future measurements of hundreds of additional factors in many cell types and physiological states call for the higher throughput and uniformity afforded by automation. The immunoprecipitation process has become rate-limiting, and is, in addition, a source of substantial variability when performed manually. Here we report a fully automated robotic ChIP (R-ChIP) pipeline that allows up to 96 reactions, with high consistency and limited human involvement. A second bottleneck is the dearth of renewable ChIP-competent immune reagents, which do not yet exist for the majority of known mouse and human transcription factors and co-factors. We used R-ChIP to screen new mouse monoclonal antibodies raised against p300, a histone acetylase protein well-known as a marker of active transcriptional enhancer elements. Despite its importance, ChIP-competent monoclonal reagents for p300 have been lacking. We identified and validated for ChIP-seq a monoclonal reagent called ENCITp300-1.

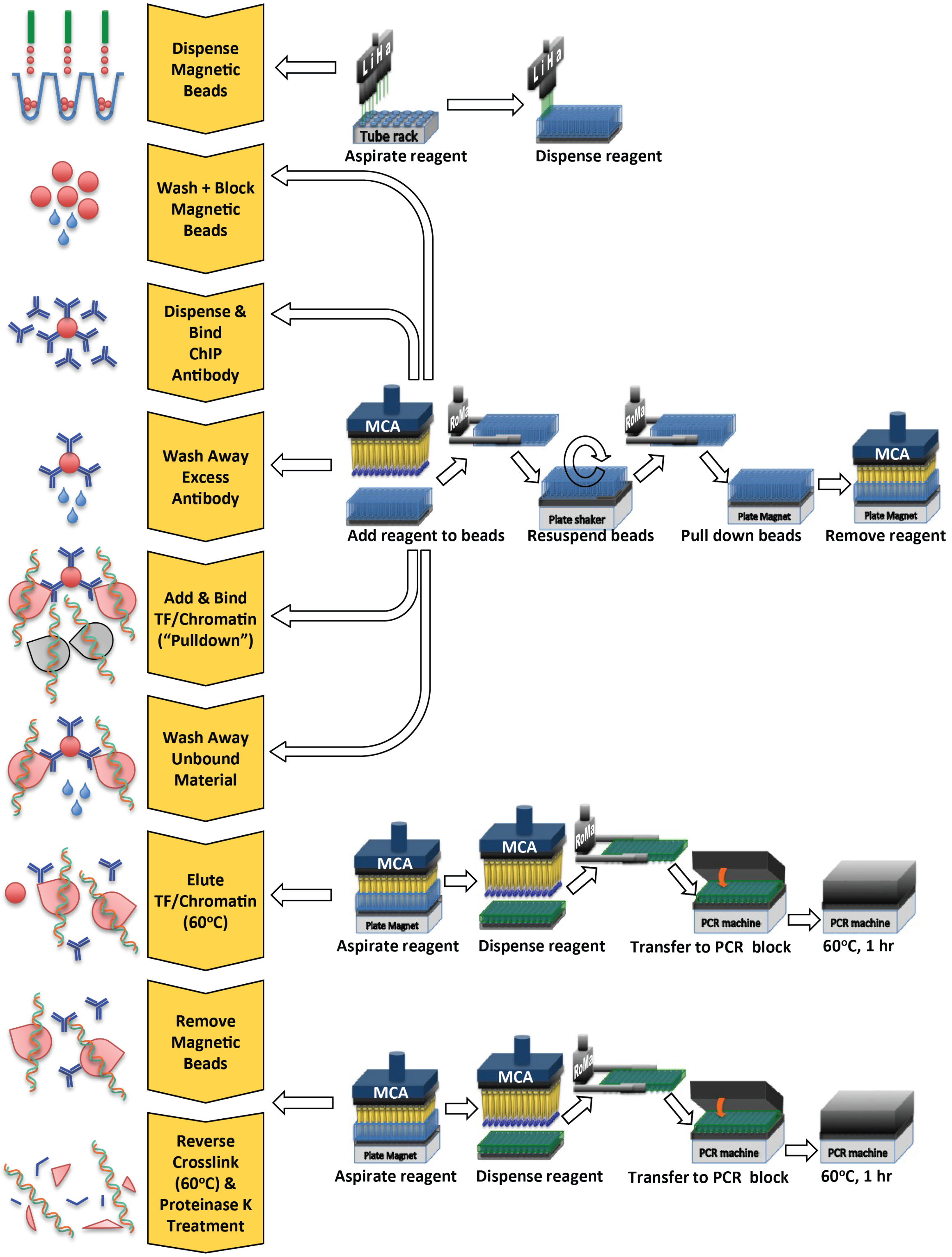
11.1 Introduction

Contemporary studies of gene regulation are often based, at least in part, on learning the patterns of chromatin mark distribution and the locations of specific transcription factor occupancy in the genome. The chromatin Immunoprecipitation (ChIP) assay, in several variations, provides this information (Gilmour & Lis 1984; Gilmour & Lis 1985; Solomon et al. 1988). ChIP protocols typically begin by crosslinking proteins to DNA (usually using formaldehyde); then selectively retrieving DNA fragments associated with a protein of interest by immunoprecipitation; and finally analyzing the enriched DNA. Originally, ChIP-enrichment was analyzed using qPCR at predefined genomic regions (Hecht et al. 1996). Later, it was coupled with microarray readouts (ChIP-Chip/ChIP-on-Chip) which allowed many selected regions to be assayed in parallel (e.g. all promoters) or even whole genomes, especially in organisms with small genomes. (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak & Snyder 2002; Weinmann et al. 2002). Eventually, high-throughput sequencing enabled truly genome-wide mapping of protein-DNA interactions, with high resolution, in the form of ChIP-seq (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007; Wold & Myers 2008).

ChIP-seq has become the workhorse for mapping the whole-genome occupancy of hundreds of transcription factors in human, mouse, fly and worm by the ENCODE (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012; Gerstein et al. 2012; Wang et al. 2012), mouse ENCODE (Mouse ENCODE Consortium 2012) and modENCODE consortia (Gerstein et al. 2010; modENCODE Consortium 2010) and to profile the genomic distribution of numerous histone modifications in a wide variety of human cell lines and tissues by the NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al. 2010). Despite the large number of datasets generated thus far, they are a small fraction of the expected future ChIP-seq experiments from individual laboratories as well as consortia. For example transcription factors assayed by ENCODE through 2013 represent only about 10% of the total number of transcription factors in the genome (Vaquerizas et al. 2009), and this has been done in a very limited number of cell types. Initially, DNA sequencing capacity and cost were major barriers to very large scale ChIP-seq, but sequencing capacity has increased by several orders of magnitude and costs per ChIP have dropped significantly. Notably, the ChIP step has emerged as rate-limiting. It is tedious and, in practice it is often variable from one practitioner to another, from experiment to experiment and even among replicates in a single experiment. This suggested that a robust robotic ChIP protocol could stabilize and improve data quality, reproducibility, manpower use, and overall costs and efficiency per experiment.

A second independent challenge for ChIP-seq experiments is that the supply of high-quality sustainable immune reagents experimentally validated for ChIP remains very limited. Many antibodies, including some marketed as “ChIP-grade” have failed in the ENCODE pipeline, and many that succeed are polyclonal, which means that different lots can vary radically in how well they perform in ChIP. At present, monoclonal antibodies are the most reliable renewable ChIP reagents, although they do not account for the majority of characterized reagents, and there are no ChIP-competent reagents for the majority of human and mouse transcription factors. Validated polyclonal reagents have been shown to vary substantially from one lot to another (Egelhofer et al. 2010). The field therefore faces the twin challenges of generating large quantities of ChIP-seq data in reliable high-throughput manner for factors with extant affinity reagents, and screening and characterizing new sustainable immune reagents.

In this work we developed a fully automated robotic pipeline for the chromatin immunoprecipitation reaction (R-ChIP). High-throughput 96-well plate methods for performing ChIP have been described before (Garber et al. 2012; Blecher-Gonen et al. 2013). However, those methods require substantial hands-on time and are subject to variability inherent in experiments done by humans. The R-ChIP reported here employs a widely used, multipurpose programmable liquid handling robotic platform (Tecan EVO 200), which can be used for a multitude of other purposes, such as robotic plasmid cloning or automated ELISA screenings when it is not being used for automated ChIP. We tested our protocol on factors that had previously been characterized in multiple ENCODE cell lines and show that it performs comparably to manual ChIP-seq in enrichment and in producing high quality ChIP-seq libraries that are consistent within and between experiments. We then applied R-ChIP to screen candidate monoclonal antibodies against the transcriptional co-activator p300, a protein for which monoclonal ChIP-competent reagents have until now not been


Dispense Magnetic Beads
LiHa
Tube rack
Aspirate reagent
LiHa
Dispense reagent
Wash + Block Magnetic Beads
Dispense & Bind ChIP Antibody
Wash Away Excess Antibody
MCA
RoMa
RoMa
MCA
Plate shaker
Plate Magnet
Plate Magnet
Add reagent to beads
Resuspend beads
Pull down beads
Remove reagent
Add & Bind TF/Chromatin ("Pulldown")
Wash Away Unbound Material
Elute TF/Chromatin (60°C)
MCA
Plate Magnet
MCA
RoMa
PCR machine
PCR machine
Aspirate reagent
Dispense reagent
Transfer to PCR block
60°C, 1 hr
Remove Magnetic Beads
Reverse Crosslink (60°C) & Proteinase K Treatment
MCA
Plate Magnet
MCA
RoMa
PCR machine
PCR machine
Aspirate reagent
Dispense reagent
Transfer to PCR block
60°C, 1 hr

available, and for which polyclonal reagent lots have been highly variable.

11.2 Results

11.2.1 Automated ChIP protocol adaptations

The primary goal of this work was to fully automate ChIP without compromising yield and quality. Our design approach was to develop automation that mimics as closely as possible the established manual process, using the ENCODE ChIP protocol as the starting point (the current manual ENCODE ChIP protocol is provided in supplementary appendix). Where process changes were made to accommodate automation, we benchmarked the new process against results from the established protocol.

We configured a Tecan EVO Freedom 200 robot as detailed in Figure 11.1, and programmed it for running a 96-well format of automated chromatin immunoprecipitation reactions (the program itself is supplied as supplementary file). Major considerations for automating ChIP revolved around magnetic bead-handling to achieve successful incubation, washing, and recovery of immunoprecipitated material, while effectively eliminating unbound chromatin. In the manual version of ChIP, bead agitation is achieved by tumbling the reaction mix in standard Eppendorf 1.5-mL micro tubes on a tumbler wheel. The agitation device available on the robot is an orbital shaker with a 2-mm shake radius and adjustable speeds ranging from 100 rpm to 1600 rpm. An alternate method for automated bead agitation mixes by repeated pipetting (trituration). We reasoned that pipetting would have inevitable bead losses to the pipette surface, especially as multiple tip changes would be required. We therefore focused on the orbital shaker. The second automation constraint comes from the 96-well plate format compared with individual microtubes in the manual protocol. This change requires effective robotic washing without cross-contamination between wells or sample spillage. Finally, the 96-well format requires a plate magnet strong enough to efficiently pull down all beads. Several vendors offer plate magnet compatible with the robot platform, but most are designed for standard low profile micro plates, while our automated ChIP protocol requires deep well plates for effective bead washing. A magnet designed specifically for deep well plates (SPRIPlate Super Magnet Plate from Agencourt, Beckman Coulter) proved effective. Its success in our hands was optimal with a round well deep well plate with U-bottom wells (catalog # 278752, Nunc). A summary of major differences for the robotic protocol is below and both protocols are given in detail in Supplemental Methods:

1. Bead agitation was changed from a tumbling motion in the manual protocol to rapid orbital shaking. The shake speed was optimized to keep beads fully suspended without spillage (1400 rpm).

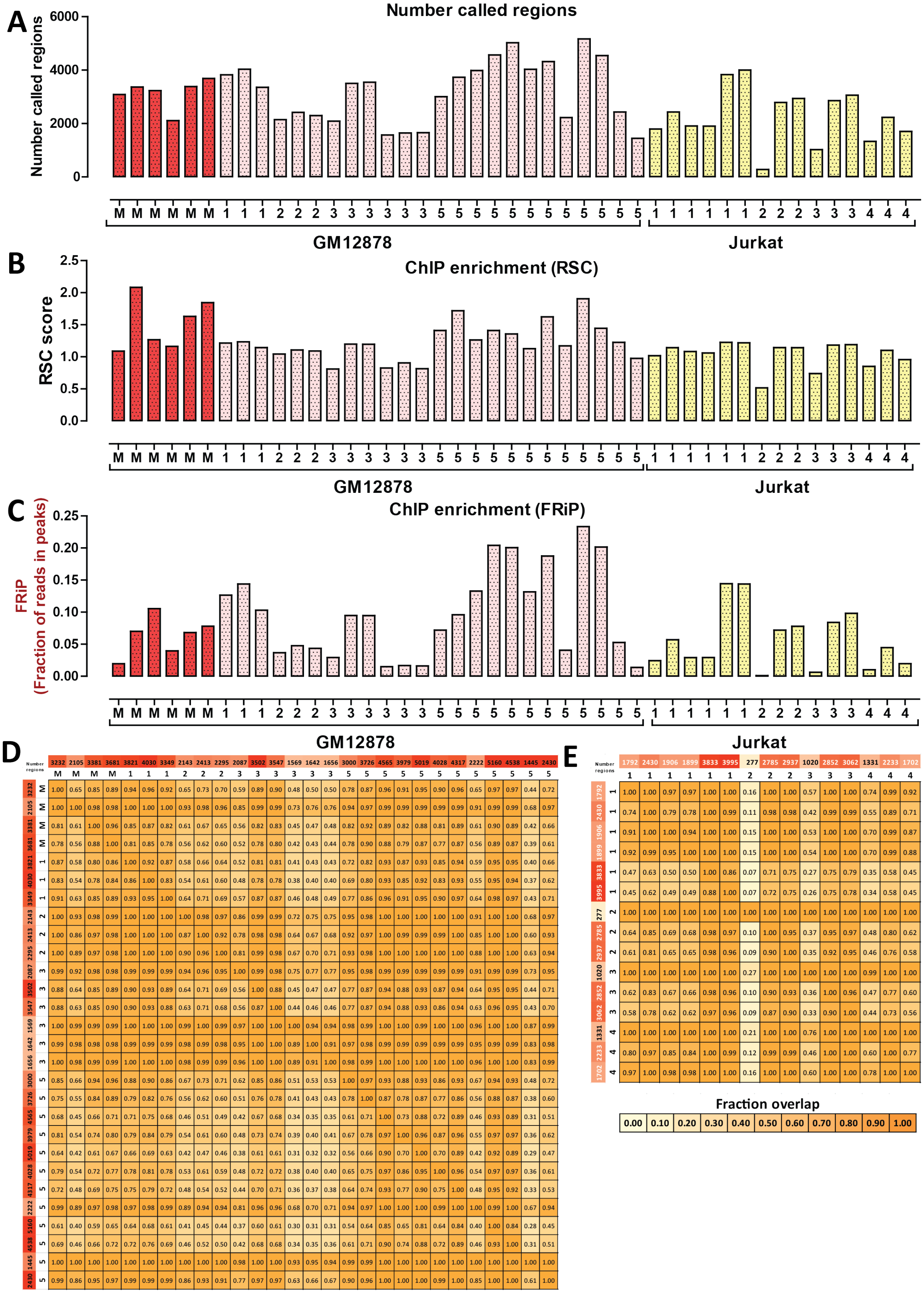
2. The sample volume was reduced from 1000 μL to 500 μL to prevent spillage.

3. Wash steps after antibody and chromatin binding where increased in number from 3 to 4 to compensate for the smaller wash volume.

4. Bead recovery time on the magnet was extended to 7 min on the robot, a condition determined empirically using the criterion that no detectable beads were left behind in the supernatant upon microscopic inspection.

The R-ChIP protocol incubates the ChIP antibody with the magnetic beads in the conjugation step for 1 hour at room temperature, and incubates chromatin with the antibody-conjugated

Figure 11.1 ***(preceding page)*****: Illustration of individual automated ChIP protocol steps.** A Tecan Freedom EVO 200 robot equipped with a Liquid Handling arm (LiHa), a Multi Channel Arm (MCA) and Robotic Manipulator arm (RoMa) is used for all steps. Additional devices integrated into the robot are standard size plate carriers, magnet plate, orbital plate shaker and PCR machine. The cartoons in the left column illustrate each protocol step, described in the flow diagram in the second column. The cartoon sequences on the right illustrate the robotic process step sequences used for each protocol step. The white arrows pointing to the protocol steps indicate which robot sequences apply to each protocol step.



GM12878

Number regions		3232	2105	3381	3681	3821	4030	3349	2143	2413	2295	2087	3502	3547	1569	1642	1656	3000	3726	4565	3979	5019	4028	4317	2222	5160	4538	1445	2430
		M	M	M	M	1	1	1	2	2	2	3	3	3	3	3	3	5	5	5	5	5	5	5	5	5	5	5	5
3232	M	1.00	0.65	0.85	0.89	0.94	0.96	0.92	0.65	0.73	0.70	0.59	0.89	0.90	0.48	0.50	0.50	0.78	0.87	0.96	0.91	0.95	0.90	0.96	0.65	0.97	0.97	0.44	0.72
2105	M	1.00	1.00	0.98	0.98	1.00	1.00	1.00	0.93	0.98	0.96	0.85	0.99	0.99	0.73	0.76	0.76	0.94	0.97	0.99	0.99	0.99	0.99	0.99	0.91	0.99	0.99	0.68	0.97
3381	M	0.81	0.61	1.00	0.96	0.85	0.87	0.82	0.61	0.67	0.65	0.56	0.82	0.83	0.45	0.47	0.48	0.82	0.92	0.89	0.82	0.88	0.81	0.89	0.61	0.90	0.89	0.42	0.66
3681	M	0.78	0.56	0.88	1.00	0.81	0.85	0.78	0.56	0.62	0.60	0.52	0.78	0.80	0.42	0.43	0.44	0.78	0.90	0.88	0.79	0.87	0.77	0.87	0.56	0.89	0.87	0.39	0.61
3821	1	0.87	0.58	0.80	0.86	1.00	0.92	0.87	0.58	0.66	0.64	0.52	0.81	0.81	0.41	0.43	0.43	0.72	0.82	0.93	0.87	0.93	0.85	0.94	0.59	0.95	0.95	0.40	0.66
4030	1	0.83	0.54	0.78	0.84	0.86	1.00	0.83	0.54	0.61	0.60	0.48	0.78	0.78	0.38	0.40	0.40	0.69	0.80	0.93	0.85	0.92	0.83	0.93	0.55	0.95	0.94	0.37	0.62
3349	1	0.91	0.63	0.85	0.89	0.93	0.95	1.00	0.64	0.71	0.69	0.57	0.87	0.87	0.46	0.48	0.49	0.77	0.86	0.96	0.91	0.95	0.90	0.97	0.64	0.98	0.97	0.43	0.71
2143	2	1.00	0.93	0.98	0.99	1.00	1.00	1.00	1.00	0.98	0.97	0.86	0.99	0.99	0.72	0.75	0.75	0.95	0.98	1.00	1.00	1.00	1.00	1.00	0.91	1.00	1.00	0.68	0.97
2413	2	1.00	0.86	0.97	0.98	1.00	1.00	1.00	0.87	1.00	0.92	0.78	0.98	0.98	0.64	0.67	0.68	0.92	0.97	1.00	0.99	0.99	0.99	1.00	0.85	1.00	1.00	0.60	0.93
2295	2	1.00	0.89	0.97	0.98	1.00	1.00	1.00	0.90	0.96	1.00	0.81	0.99	0.98	0.67	0.70	0.71	0.93	0.98	1.00	1.00	1.00	1.00	1.00	0.88	1.00	1.00	0.63	0.94
2087	3	0.99	0.92	0.98	0.98	0.99	0.99	0.99	0.94	0.96	0.95	1.00	0.99	0.98	0.75	0.77	0.77	0.95	0.98	0.99	0.99	0.99	0.99	0.99	0.91	0.99	0.99	0.73	0.95
3502	3	0.88	0.64	0.85	0.89	0.90	0.93	0.89	0.64	0.71	0.69	0.56	1.00	0.88	0.45	0.47	0.47	0.77	0.87	0.94	0.88	0.93	0.87	0.94	0.64	0.95	0.95	0.44	0.71
3547	3	0.88	0.63	0.85	0.90	0.90	0.93	0.88	0.63	0.71	0.68	0.56	0.87	1.00	0.44	0.46	0.46	0.77	0.87	0.94	0.88	0.93	0.86	0.94	0.63	0.96	0.95	0.43	0.70
1569	3	1.00	0.99	0.99	0.99	1.00	1.00	1.00	0.99	1.00	0.99	0.97	1.00	1.00	1.00	0.94	0.94	0.98	0.99	1.00	1.00	0.99	1.00	1.00	0.96	1.00	1.00	0.87	0.99
1642	3	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.98	0.95	1.00	1.00	0.89	1.00	0.91	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.95	0.99	0.99	0.83	0.98
1656	3	1.00	0.98	0.99	0.99	1.00	1.00	1.00	0.98	0.99	0.99	0.96	1.00	1.00	0.89	0.91	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.83	0.99
3000	5	0.85	0.66	0.94	0.96	0.88	0.90	0.86	0.67	0.73	0.71	0.62	0.85	0.86	0.51	0.53	0.53	1.00	0.97	0.93	0.88	0.93	0.86	0.93	0.67	0.94	0.93	0.48	0.72
3726	5	0.75	0.55	0.84	0.89	0.79	0.82	0.76	0.56	0.62	0.60	0.51	0.76	0.78	0.41	0.43	0.43	0.78	1.00	0.87	0.78	0.87	0.77	0.86	0.56	0.88	0.87	0.38	0.60
4565	5	0.68	0.45	0.66	0.71	0.71	0.75	0.68	0.46	0.51	0.49	0.42	0.67	0.68	0.34	0.35	0.35	0.61	0.71	1.00	0.73	0.88	0.72	0.89	0.46	0.93	0.89	0.31	0.51
3979	5	0.81	0.54	0.74	0.80	0.79	0.84	0.79	0.54	0.61	0.60	0.48	0.73	0.74	0.39	0.40	0.41	0.67	0.78	0.97	1.00	0.96	0.87	0.96	0.55	0.97	0.97	0.36	0.62
5019	5	0.64	0.42	0.61	0.67	0.66	0.69	0.63	0.42	0.47	0.46	0.38	0.61	0.61	0.31	0.32	0.32	0.56	0.66	0.90	0.70	1.00	0.70	0.89	0.42	0.92	0.89	0.29	0.47
4028	5	0.79	0.54	0.72	0.77	0.78	0.81	0.78	0.53	0.61	0.59	0.48	0.72	0.72	0.38	0.40	0.40	0.65	0.75	0.97	0.86	0.95	1.00	0.96	0.54	0.97	0.97	0.36	0.61
4317	5	0.72	0.48	0.69	0.75	0.75	0.79	0.72	0.48	0.54	0.52	0.44	0.70	0.71	0.36	0.37	0.38	0.64	0.74	0.93	0.77	0.90	0.75	1.00	0.48	0.95	0.92	0.33	0.53
2222	5	0.99	0.89	0.97	0.98	0.97	0.98	0.99	0.89	0.94	0.94	0.81	0.96	0.96	0.68	0.70	0.71	0.94	0.97	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	0.67	0.94
5160	5	0.61	0.40	0.59	0.65	0.64	0.68	0.61	0.41	0.45	0.44	0.37	0.60	0.61	0.30	0.31	0.31	0.54	0.64	0.85	0.65	0.81	0.64	0.84	0.40	1.00	0.84	0.28	0.45
4538	5	0.69	0.46	0.66	0.72	0.72	0.76	0.69	0.46	0.52	0.50	0.42	0.68	0.68	0.34	0.35	0.36	0.61	0.71	0.90	0.74	0.88	0.72	0.89	0.46	0.93	1.00	0.31	0.51
1445	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.93	0.95	0.94	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2430	5	0.99	0.86	0.95	0.97	0.99	0.99	0.99	0.86	0.93	0.91	0.77	0.97	0.97	0.63	0.66	0.67	0.90	0.96	1.00	1.00	1.00	0.99	1.00	0.85	1.00	1.00	0.61	1.00

Jurkat

Number regions		1792	2430	1906	1899	3833	3995	277	2785	2937	1020	2852	3062	1331	2233	1702
		1	1	1	1	1	1	2	2	2	3	3	3	4	4	4
1792	1	1.00	1.00	0.97	0.97	1.00	1.00	0.16	1.00	1.00	0.57	1.00	1.00	0.74	0.99	0.92
2430	1	0.74	1.00	0.79	0.78	1.00	0.99	0.11	0.98	0.98	0.42	0.99	0.99	0.55	0.89	0.71
1906	1	0.91	1.00	1.00	0.94	1.00	1.00	0.15	1.00	1.00	0.53	1.00	1.00	0.70	0.99	0.87
1899	1	0.92	0.99	0.95	1.00	1.00	1.00	0.15	1.00	1.00	0.54	1.00	1.00	0.70	0.99	0.88
3833	1	0.47	0.63	0.50	0.50	1.00	0.86	0.07	0.71	0.75	0.27	0.75	0.79	0.35	0.58	0.45
3995	1	0.45	0.62	0.49	0.49	0.88	1.00	0.07	0.72	0.75	0.26	0.75	0.78	0.34	0.58	0.45
277	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2785	2	0.64	0.85	0.69	0.68	0.98	0.97	0.10	1.00	0.95	0.37	0.95	0.97	0.48	0.80	0.62
2937	2	0.61	0.81	0.65	0.65	0.98	0.96	0.09	0.90	1.00	0.35	0.92	0.95	0.46	0.76	0.58
1020	3	1.00	1.00	1.00	1.00	1.00	1.00	0.27	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
2852	3	0.62	0.83	0.67	0.66	0.98	0.96	0.10	0.90	0.93	0.36	1.00	0.96	0.47	0.77	0.60
3062	3	0.58	0.78	0.62	0.62	0.97	0.96	0.09	0.87	0.90	0.33	0.90	1.00	0.44	0.73	0.56
1331	4	1.00	1.00	1.00	1.00	1.00	1.00	0.21	1.00	1.00	0.76	1.00	1.00	1.00	1.00	1.00
2233	4	0.80	0.97	0.85	0.84	1.00	0.99	0.12	0.99	0.99	0.46	1.00	1.00	0.60	1.00	0.77
1702	4	0.97	1.00	0.98	0.98	1.00	1.00	0.16	1.00	1.00	0.60	1.00	1.00	0.78	1.00	1.00

Fraction overlap

0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00

magnetic beads for 1 hour at room temperature plus 1 hour on a 4 °C cooling plate carrier, but with interruptions to resuspend the beads.

The step that dissociates bound chromatin from antibody-magnetic beads is done in the robot's PCR module, thus eliminating bead agitation during the 65 °C 1 hour incubation.

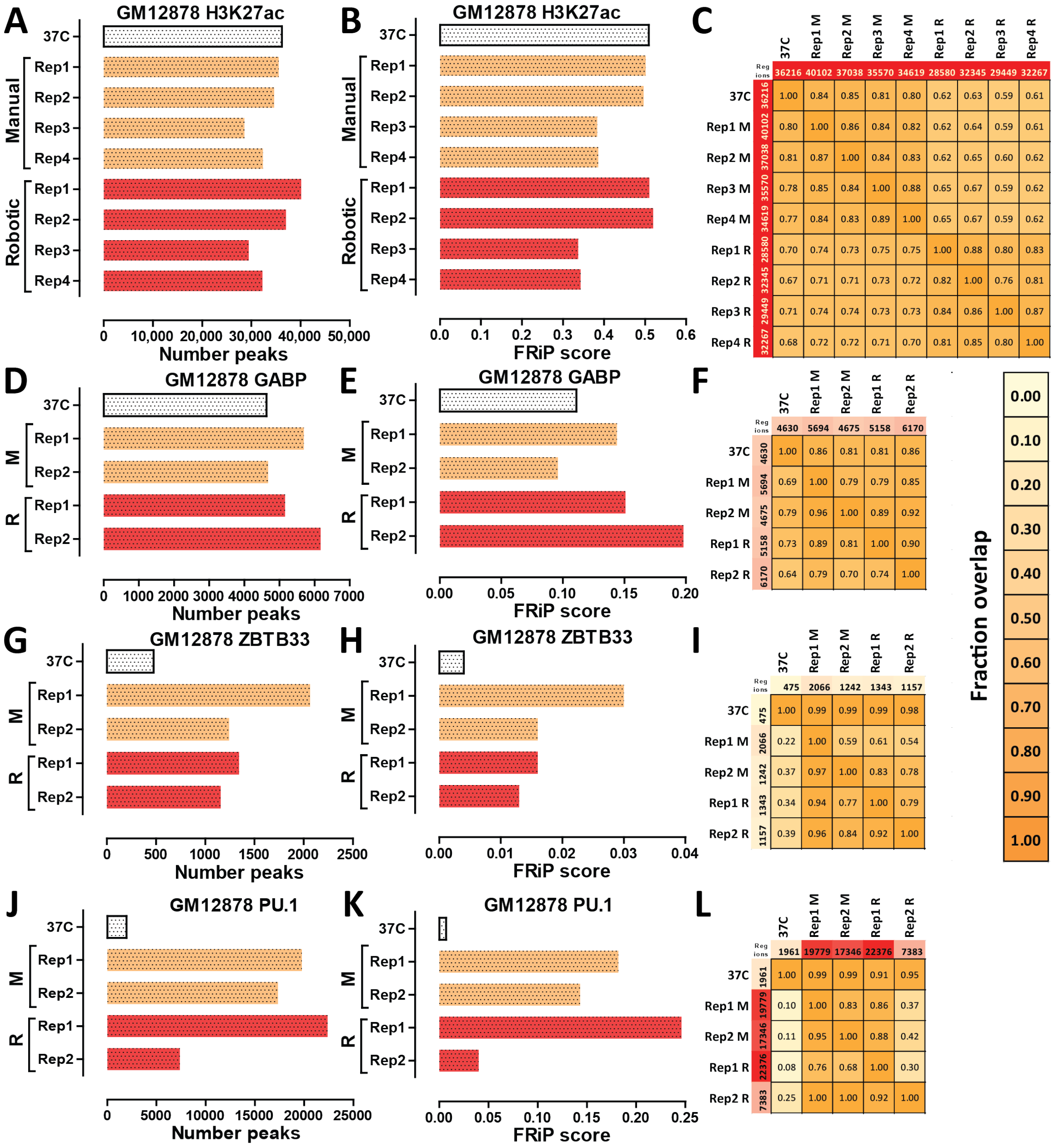
11.2.2 Consistency of robotic ChIP results

We first tested the robustness and reproducibility of our robotic ChIP protocol by carrying out multiple manual and R-ChIP experiments for the NRSF/REST transcription factor. NRSF/REST (Schoenherr & Anderson 1995; Chong et al. 1995) is a negative transcriptional regulator of neuronal genes in non-neuronal cell types. It was the first transcription factor to which ChIP-seq was applied (Johnson et al. 2007), its binding has been extensively mapped in multiple cell lines, and its recognition site (and its binding variants) is well studied. The monoclonal antibody used for NRSF ChIP has been well characterized for specificity and for efficacy in the ChIP-seq format. It is thus an ideal system to characterize our method.

We performed ChIP-seq experiments in two cell lines, GM12878 and Jurkat, producing at least three libraries from four separate plates for GM12878 and from four separate plates for Jurkat. We compared the resulting data to existing manually generated NRSF ChIP-seq datasets for GM18278 cells (ENCODE Project Consortium 2012) and to additional four manual ChIP-seq datasets generated in parallel with the R-ChIP ones. These data are summarized in Figure 2.

To assess ChIP quality, we used library and ChIP QC metrics that were developed previously by us and others as part of the ENCODE Consortium (Landt et al. 2012; Kundaje et al. unpublished; Marinov et al. 2014). The first question regarding ChIP quality is how well the immunoprecipitation step has enriched for DNA fragments attached to the antigen of interest. This can be assessed by calculating the fraction of reads falling within called peaks (FRiP, Landt et al. 2012) or by using cross-correlation (Kharchenko et al. 2008; Landt et al. 2012). Both measures have limitations in some special cases (Marinov et al. 2014), but when both are applied and concur, confidence in the results is high. Figure 11.2A shows the number of called peaks and Figures 11.2B and 11.2C show the RSC (Relative Strand Correlation, Landt et al. 2012; Kundaje et al., unpublished) and FRiP values for manual and robotic NRSF ChIP-seq datasets. R-ChIP data consistently exhibited good RSC values (RSC ≥ 1) and FRiP and peak number values comparable to those of manually generated libraries, with the exception of three Jurkat libraries (the first ChIP on plates 2, 3, and 4, Figure 11.2A, 11.2B and 11.2C) that scored as less successful. We do not presently know the cause of these lower-quality libraries, but their frequency is well within the range of variability of manually generated libraries we have observed over several years, during which sporadic unsuccessful experiments for factor/antibody pairs that are otherwise routinely successful have occurred. Finally, we asked how similar the final sets of called peaks are for the robotic protocol and how they compare with reference manual datasets for the same factor and cell type, by evaluating peaks called after sequencing. Figures 11.2D and 11.2E show the similarity of peak call sets for all libraries measured by calculating the size of the overlap

Figure 11.2 *(preceding page)*: Reproducibility of R-ChIP experiments. Multiple ChIP-seq experiments on multiple plates were generated for the NRSF/REST repressor in GM12878 lymphoblastoid cells ($n = 4$ plates) and Jurkat T-cells ($n = 4$ plates) cell lines. The numbers (1 through 5) refer to the number of the plate a library came from, "M" refers to manually generated datasets. The first two manual GM12878 datasets were previously published as part of the ENCODE project, the next four were generated in parallel with the R-ChIP ones. (A) Number of called regions for each dataset (using ERANGE 4.0, Johnson & Mortazavi et al. 2007) (B) Assessment of ChIP enrichment using RSC (Relative Strand Correlation) cross-correlation scores (Landt et al. 2012); (C) Assessment of ChIP enrichment using FRiP (Fraction of Reads in Peaks) scores (Landt et al. 2012); (D) Overlap between called peaks in robotic and manual ChIP libraries in GM12878 cells; (E) Overlap between called peaks in robotic and manual ChIP libraries in Jurkat cells. The overlap score (O_{XY}) shown in each box indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{XY} = |X \cap Y|/|Y|$.



C

		37C	Rep1 M	Rep2 M	Rep3 M	Rep4 M	Rep1 R	Rep2 R	Rep3 R	Rep4 R
	Reg ions	36216	40102	37038	35570	34619	28580	32345	29449	32267
37C	36216	1.00	0.84	0.85	0.81	0.80	0.62	0.63	0.59	0.61
Rep1 M	40102	0.80	1.00	0.86	0.84	0.82	0.62	0.64	0.59	0.61
Rep2 M	37038	0.81	0.87	1.00	0.84	0.83	0.62	0.65	0.60	0.62
Rep3 M	35570	0.78	0.85	0.84	1.00	0.88	0.65	0.67	0.59	0.62
Rep4 M	34619	0.77	0.84	0.83	0.89	1.00	0.65	0.67	0.59	0.62
Rep1 R	28580	0.70	0.74	0.73	0.75	0.75	1.00	0.88	0.80	0.83
Rep2 R	32345	0.67	0.71	0.71	0.73	0.72	0.82	1.00	0.76	0.81
Rep3 R	29449	0.71	0.74	0.74	0.73	0.73	0.84	0.86	1.00	0.87
Rep4 R	32267	0.68	0.72	0.72	0.71	0.70	0.81	0.85	0.80	1.00

F

		37C	Rep1 M	Rep2 M	Rep1 R	Rep2 R
	Reg ions	4630	5694	4675	5158	6170
37C	4630	1.00	0.86	0.81	0.81	0.86
Rep1 M	5694	0.69	1.00	0.79	0.79	0.85
Rep2 M	4675	0.79	0.96	1.00	0.89	0.92
Rep1 R	5158	0.73	0.89	0.81	1.00	0.90
Rep2 R	6170	0.64	0.79	0.70	0.74	1.00

I

		37C	Rep1 M	Rep2 M	Rep1 R	Rep2 R
	Reg ions	475	2066	1242	1343	1157
37C	475	1.00	0.99	0.99	0.99	0.98
Rep1 M	2066	0.22	1.00	0.59	0.61	0.54
Rep2 M	1242	0.37	0.97	1.00	0.83	0.78
Rep1 R	1343	0.34	0.94	0.77	1.00	0.79
Rep2 R	1157	0.39	0.96	0.84	0.92	1.00

L

		37C	Rep1 M	Rep2 M	Rep1 R	Rep2 R
	Reg ions	1961	19779	17346	22376	7383
37C	1961	1.00	0.99	0.99	0.91	0.95
Rep1 M	19779	0.10	1.00	0.83	0.86	0.37
Rep2 M	17346	0.11	0.95	1.00	0.88	0.42
Rep1 R	22376	0.08	0.76	0.68	1.00	0.30
Rep2 R	7383	0.25	1.00	1.00	0.92	1.00

Figure 11.3: Comparison between manual and robotic ChIP-seq resullts and between ChIP-seq results on GM12878 chromatin fixed under standard fixation conditions and chromatin fixed at 37 °C for additional targets. (A,B,C) ChIP-seq against H3K27ac in GM12878 cells. (A) FRiP score, (B) number of peaks called, (C) overlap between the sets of peaks; (D,E,F) ChIP-seq against GABP in GM12878 cells. (D) FRiP score, (E) number of peaks called, (F) overlap between the sets of peaks; (G,H,I) ChIP-seq against ZBTB33 in GM12878 cells. (G) FRiP score, (H) number of peaks called, (I) overlap between the sets of peaks; (J,K,L) ChIP-seq against PU.1 in GM12878 cells. (J) FRiP score, (K) number of peaks called, (L) overlap between the sets of peaks. The overlap score (O_{XY}) shown in each box indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{XY} = |X \cap Y|/|Y|$

between each pair of libraries. We observed consistently high overlap scores and thus high reproducibility between all libraries. These observations applied both within and between plates, underscoring the consistency and robustness of the R-ChIP protocol.

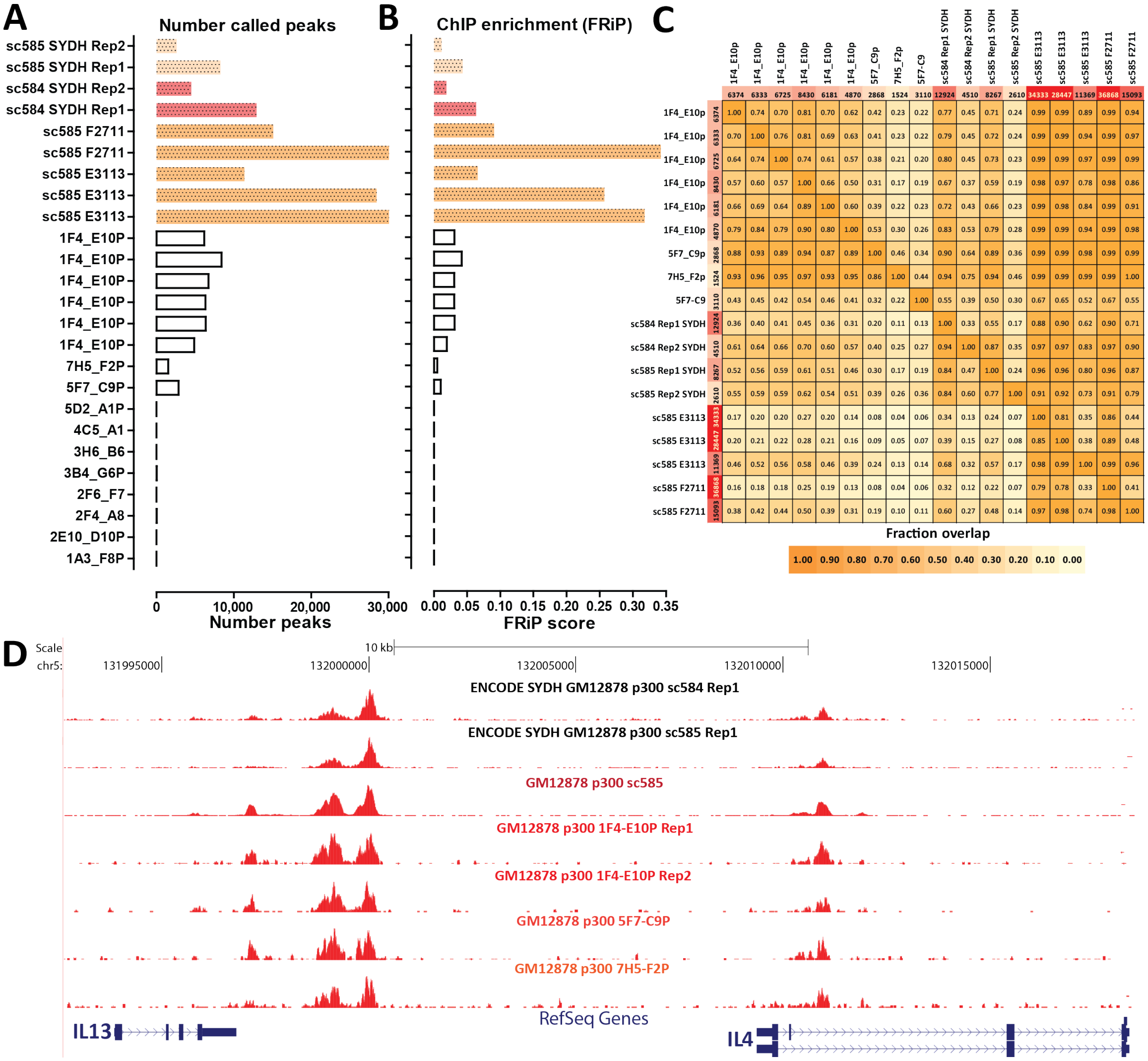
To further characterize the consistency between the results R-ChIP and manual ChIP experiments, we generated paired manual and robotic ChIP-seq datasets using matched chromatin samples for several additional targets (Figure 11.3). These included the H3K27ac histone modification (Figure 11.3A, 11.3B and 11.3C), the GABP transcription factor (Watanabe et al. 1990; Thompson et al. 1991; Collins et al. 2007; Figure 11.3D, 11.3E and 11.3F), the ZBTB33/Kaiso zing-finger protein known for its preferential binding to methylated DNA (Prokhortchouk et al. 2001; Figure 11.3G, 11.3H and 11.3I), and the important regulator of hematopoiesis SPI1/PU.1 (Klemsz et al. 1990; Burda et al. 2010; Figure 11.3J, 11.3K and 11.3L). We observed comparable results between the manual and robotic datasets (with the exception of one not very successful robotic PU.1 libraries, although it should be noted that PU.1 does not perform consistently well in ChIP-seq even though it often produced very strong datasets), further confirming the applicability of R-ChIP to large-scale ChIP-seq production.

11.2.3 Using R-ChIP to characterize new monoclonal p300 antibodies

Having established the R-ChIP protocol, we next applied it to characterize a set of monoclonal antibodies raised against the p300 transcriptional coactivator in the Beckman Institute Protein Expression Center at Caltech. The p300 protein is a histone acetyltransferase (Eckner et al. 1994; Arany et al. 1994; Lundblad et al. 1995; Ogryzko et al. 1996), best known for its role in the acetylation of histones. It is used as a marker of active transcriptional enhancers in mammalian genomes (Visel at al. 2009; Blow et al. 2010; May et al. 2011; Visel et al. 2013). Commercially available antibodies used to generate published p300 data are from a series of polyclonal reagents and are neither identical from lot to lot, historically, nor are they renewable.

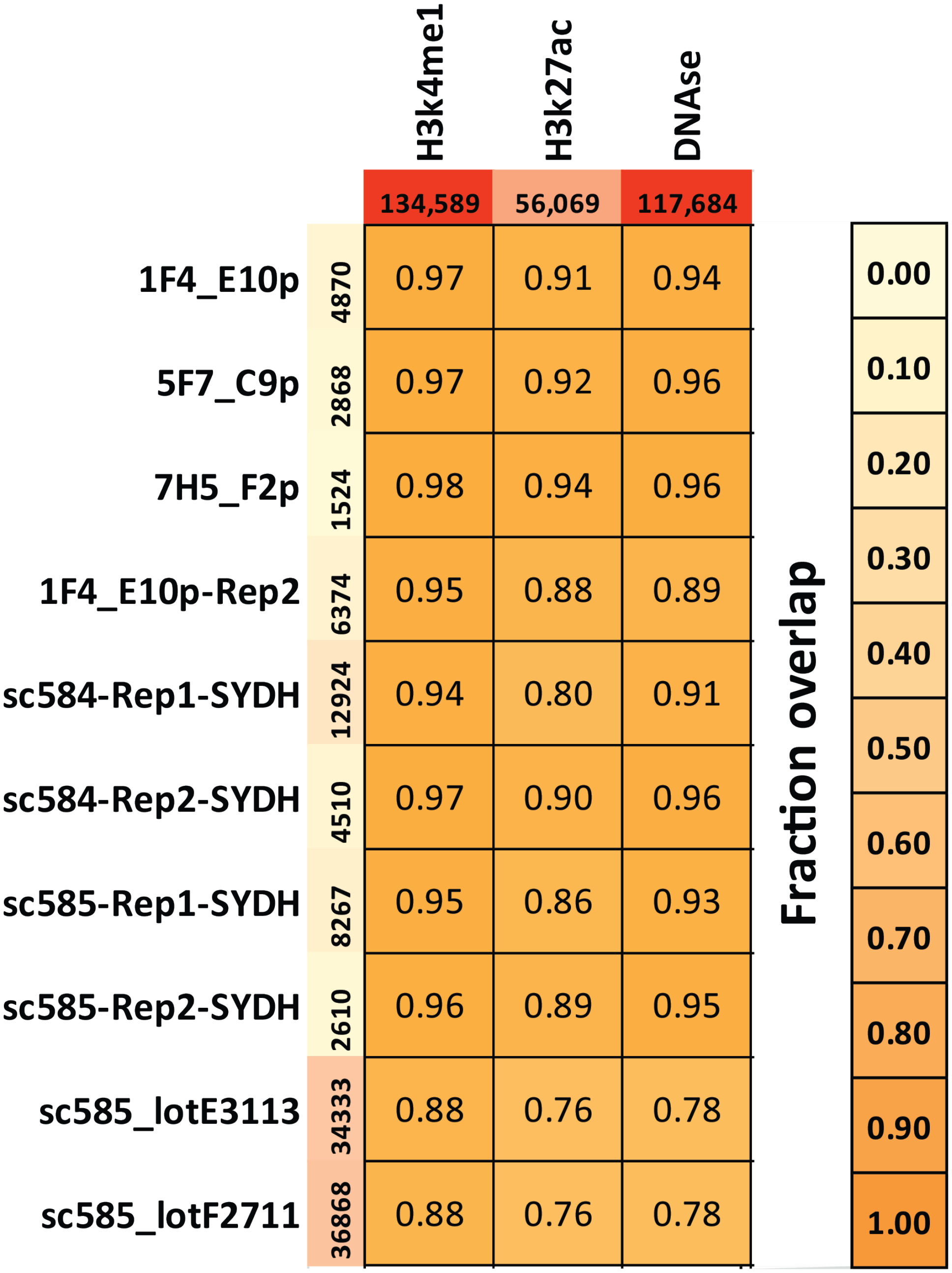
We generated 11 α-p300 mouse monoclonal antibodies which were initially screened, cloned and then rescreened using a plate based ELISA assay. We tested hundreds of individual hybridoma B-cells isolated from spleens of mice injected with a GST-tagged p300 protein fragment (aa 152-213) or a synthetic KLH-coupled peptide (aa 1526-1545). The GST-tagged preparations were subjected to formaldehyde "fix" conditions (1% FA for 10 min) with the goal of increasing the likelihood of reactivity in ChIP. The resulting 11 p300 monoclonal antibodies were tested for ChIP together with two lots of rabbit polyclonal p300 antibodies (Santa Cruz sc-585, lot numbers F2711 and E3113,) on chromatin from GM12878 cells. The resulting datasets were compared to each other and to publicly available ENCODE p300 data from the same cells (using two commercially available rabbit polyclonal antibodies, Santa Cruz sc-585 and sc-584) (Figure 11.4). Three of the mouse monoclonal antibodies raised against and N-terminal p300 synthetic peptide scored positive by ChIP-seq, identifying between 1,524 and 4,870 peaks (Figure 11.4A and 11.4B). We sequenced multiple additional replicates for the best-scoring one, 1F4-E10P and identified and even higher number of peaks in some of the datasets, up to 8,430, with the typical number being ~6,000. The peaks called in the monoclonal antibody dataset are a subset of those found in the polyclonal data (Figure 11.4C) confirming the specificity of the antibodies towards p300. While the monoclonal numbers are lower than the two most successful polyclonal datasets, they are within the range of what was previously observed in ENCODE data, and also within the range of published p300 datasets.

It was not our purpose in this study to characterize new polyclonal reagent lots, but the ones used previously by ENCODE were no longer available. We therefore used two additional rabbit polyclonal antibodies in R-ChIP (Santa Cruz sc-585, lot numbers E3113 and F2711), and they identified up to ~30,000 peaks. This number greatly exceeds previously published p300 datasets, including currently available ENCODE data for the same GM12878 B-cell line (for which between 2,610 and 12,924 peaks were called previously) (Figure 11.4A and 11.4B). This increase has two likely causes, and they are not mutually exclusive. The first well-appreciated variable is different performance by polyclonal antibody lots. In principle, individual lots can vary in the number and identity of epitopes recognized, in effective antibody concentration and in non-specific reactivity. A second difference



		1F4_E10p	1F4_E10p	1F4_E10p	1F4_E10p	1F4_E10p	1F4_E10p	5F7_C9p	7H5_F2p	5F7-C9	sc584 Rep1 SYDH	sc584 Rep2 SYDH	sc585 Rep1 SYDH	sc585 Rep2 SYDH	sc585 E3113	sc585 E3113	sc585 E3113	sc585 F2711	sc585 F2711
		6374	6333	6725	8430	6181	4870	2868	1524	3110	12924	4510	8267	2610	34333	28447	11369	36868	15093
1F4_E10p	6374	1.00	0.74	0.70	0.81	0.70	0.62	0.42	0.23	0.22	0.77	0.45	0.71	0.24	0.99	0.99	0.89	0.99	0.94
1F4_E10p	6333	0.70	1.00	0.76	0.81	0.69	0.63	0.41	0.23	0.22	0.79	0.45	0.72	0.24	0.99	0.99	0.94	0.99	0.97
1F4_E10p	6725	0.64	0.74	1.00	0.74	0.61	0.57	0.38	0.21	0.20	0.80	0.45	0.73	0.23	0.99	0.99	0.97	0.99	0.99
1F4_E10p	8430	0.57	0.60	0.57	1.00	0.66	0.50	0.31	0.17	0.19	0.67	0.37	0.59	0.19	0.98	0.97	0.78	0.98	0.86
1F4_E10p	6181	0.66	0.69	0.64	0.89	1.00	0.60	0.39	0.22	0.23	0.72	0.43	0.66	0.23	0.99	0.98	0.84	0.99	0.91
1F4_E10p	4870	0.79	0.84	0.79	0.90	0.80	1.00	0.53	0.30	0.26	0.83	0.53	0.79	0.28	0.99	0.99	0.94	0.99	0.98
5F7_C9p	2868	0.88	0.93	0.89	0.94	0.87	0.89	1.00	0.46	0.34	0.90	0.64	0.89	0.36	0.99	0.99	0.98	0.99	0.99
7H5_F2p	1524	0.93	0.96	0.95	0.97	0.93	0.95	0.86	1.00	0.44	0.94	0.75	0.94	0.46	0.99	0.99	0.99	0.99	1.00
5F7-C9	3110	0.43	0.45	0.42	0.54	0.46	0.41	0.32	0.22	1.00	0.55	0.39	0.50	0.30	0.67	0.65	0.52	0.67	0.55
sc584 Rep1 SYDH	12924	0.36	0.40	0.41	0.45	0.36	0.31	0.20	0.11	0.13	1.00	0.33	0.55	0.17	0.88	0.90	0.62	0.90	0.71
sc584 Rep2 SYDH	4510	0.61	0.64	0.66	0.70	0.60	0.57	0.40	0.25	0.27	0.94	1.00	0.87	0.35	0.97	0.97	0.83	0.97	0.90
sc585 Rep1 SYDH	8267	0.52	0.56	0.59	0.61	0.51	0.46	0.30	0.17	0.19	0.84	0.47	1.00	0.24	0.96	0.96	0.80	0.96	0.87
sc585 Rep2 SYDH	2610	0.55	0.59	0.59	0.62	0.54	0.51	0.39	0.26	0.36	0.84	0.60	0.77	1.00	0.91	0.92	0.73	0.91	0.79
sc585 E3113	34333	0.17	0.20	0.20	0.27	0.20	0.14	0.08	0.04	0.06	0.34	0.13	0.24	0.07	1.00	0.81	0.35	0.86	0.44
sc585 E3113	28447	0.20	0.21	0.22	0.28	0.21	0.16	0.09	0.05	0.07	0.39	0.15	0.27	0.08	0.85	1.00	0.38	0.89	0.48
sc585 E3113	11369	0.46	0.52	0.56	0.58	0.46	0.39	0.24	0.13	0.14	0.68	0.32	0.57	0.17	0.98	0.99	1.00	0.99	0.96
sc585 F2711	36868	0.16	0.18	0.18	0.25	0.19	0.13	0.08	0.04	0.06	0.32	0.12	0.22	0.07	0.79	0.78	0.33	1.00	0.41
sc585 F2711	15093	0.38	0.42	0.44	0.50	0.39	0.31	0.19	0.10	0.11	0.60	0.27	0.48	0.14	0.97	0.98	0.74	0.98	1.00

Figure 11.4: Characterization of novel monoclonal p300 antibodies using robotic ChIP. ChIP-seq against p300 was carried out in GM12878 cells and prior ENCODE data for it in that cell line (from the "SYDH" production group) was used as a reference. ENCODE data was generated using two different rabbit polyclonal antibodies from Santa Cruz (sc-584 and sc-585). We carried out robotic ChIP testing of two different lots of the sc-585 antibody and 11 different monoclonals we raised. The 1F4-E10P clone scored best and additional replicate were generated in subsequent experiments. (A) Number of called regions; (B) ChIP enrichment as measured by FRiP scores (Landt et al. 2012); (C) Overlap between called peaks with different antibodies. The overlap score (O_{XY}) shown in each box indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{XY} = |X \cap Y|/|Y|$; (D) Representative browser snapshot of p300 ChIP enrichment in polyclonal and monoclonal datasets around the IL13 and IL4 locus



		H3k4me1 134,589	H3k27ac 56,069	DNAse 117,684
1F4_E10p	4870	0.97	0.91	0.94
5F7_C9p	2868	0.97	0.92	0.96
7H5_F2p	1524	0.98	0.94	0.96
1F4_E10p-Rep2	6374	0.95	0.88	0.89
sc584-Rep1-SYDH	12924	0.94	0.80	0.91
sc584-Rep2-SYDH	4510	0.97	0.90	0.96
sc585-Rep1-SYDH	8267	0.95	0.86	0.93
sc585-Rep2-SYDH	2610	0.96	0.89	0.95
sc585_lotE3113	34333	0.88	0.76	0.78
sc585_lotF2711	36868	0.88	0.76	0.78

Figure 11.5: Overlap of called p300 peaks with regions of histone mark enrichment in ENCODE data from GM12878 cells. The overlap score (O_{XY}) shown in each box indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{XY} = |X \cap Y|/|Y|$. The ENCODE histone mark region calls were downloaded from the UCSC Genome Browser.

from the prior ENCODE data is the fixation condition. For p300, we fixed cells at 37 °C versus room temperature for the historic ENCODE data. This condition was suggested to us, specifically for p300, by Dr. Bing Ren, and is based on the idea that a longer time and elevated temperature would increase p300 cross-linking via indirect links to DNA-bound transcription factors or histones. This condition significantly improves p300 ChIP in our hands - we generated four datasets using the 1F4-E10P antibody on chromatin fixed under standard conditions and they were all unsuccessful (Figure 11.6). Of the 30,000 p300 peaks called the majority (between 76% and 88%) overlap with one or more chromatin marks associated with enhancer and promoter activity in ENCODE data (H3K27ac; H3K4me1) or with regions of DNAse Hypersensitivity (Figure 11.5), consistent with them being active enhancers and promoters. For multiple cell types, the numbers of DNAse hypersensitive regions (Neph et al. 2012; Thurman et al. 2012), H3K27ac and H3K4me1 positive regions, reported previously are typically in the tens of thousands (ENCODE Project Consortium 2012), and the number of expressed genes per cell type is between 5 and 10,000. Thus the expected number of enhancers (and p300-positive regions) is larger than the single-digit thousands of p300 peaks called in most previously available data. Therefore while reagent-specific background, including possible polyclonal cross-reactivity, could explain the thousands of p300 peaks that lack additional enhancer or promoter marks, the most parsimonious explanation for the overall very large number of p300 peaks is that prior ChIP measurements have under-estimated p300 occupancy. Our best-performing monoclonal antibody did not produce comparably high peak numbers using the same chromatin substrate, but 99% of its peaks overlap those called in the polyclonal datasets. We tested additional factors with the 37 °C fixation condition. Results were very similar to the standard condition for NRSF, H3K27ac and GABP (Figure 11.7, 11.3A-F) suggesting the more aggressive fixation condition does not result in general non-specific background. Surprisingly, the individual 37 °C ZBTB33 and PU.1 were worse than the

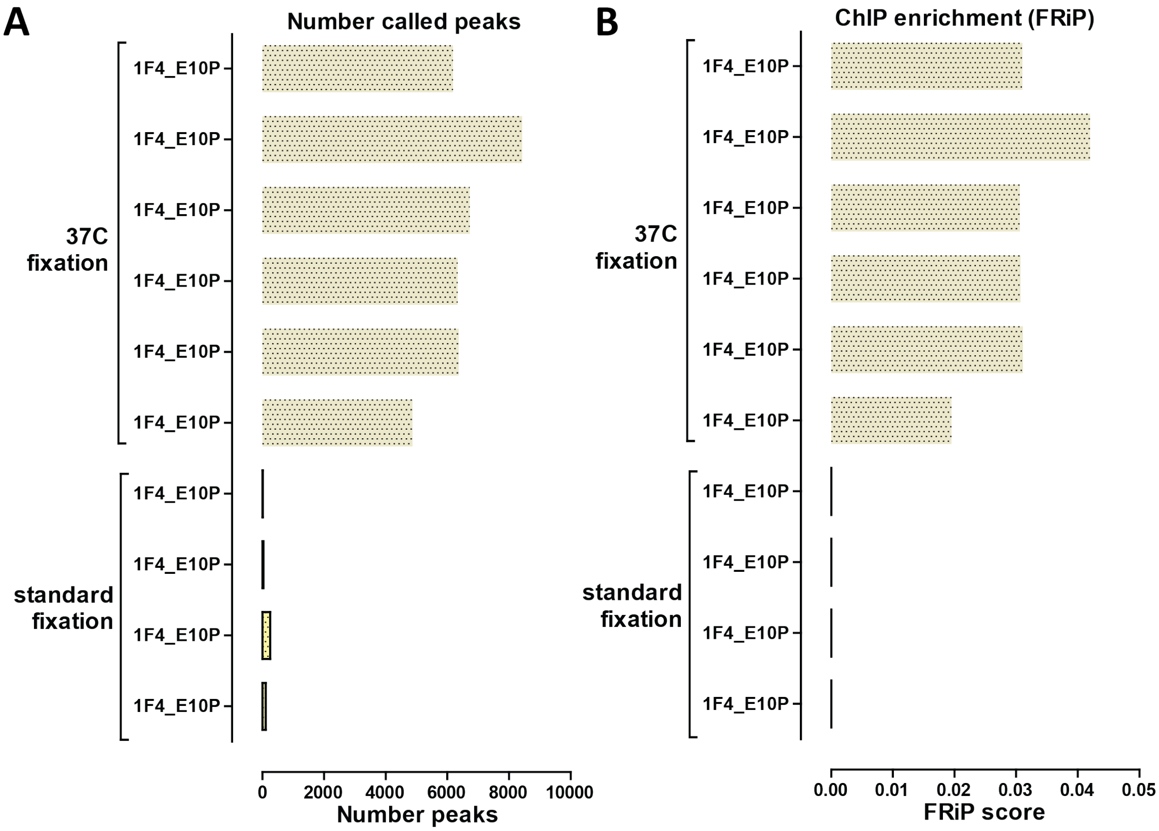
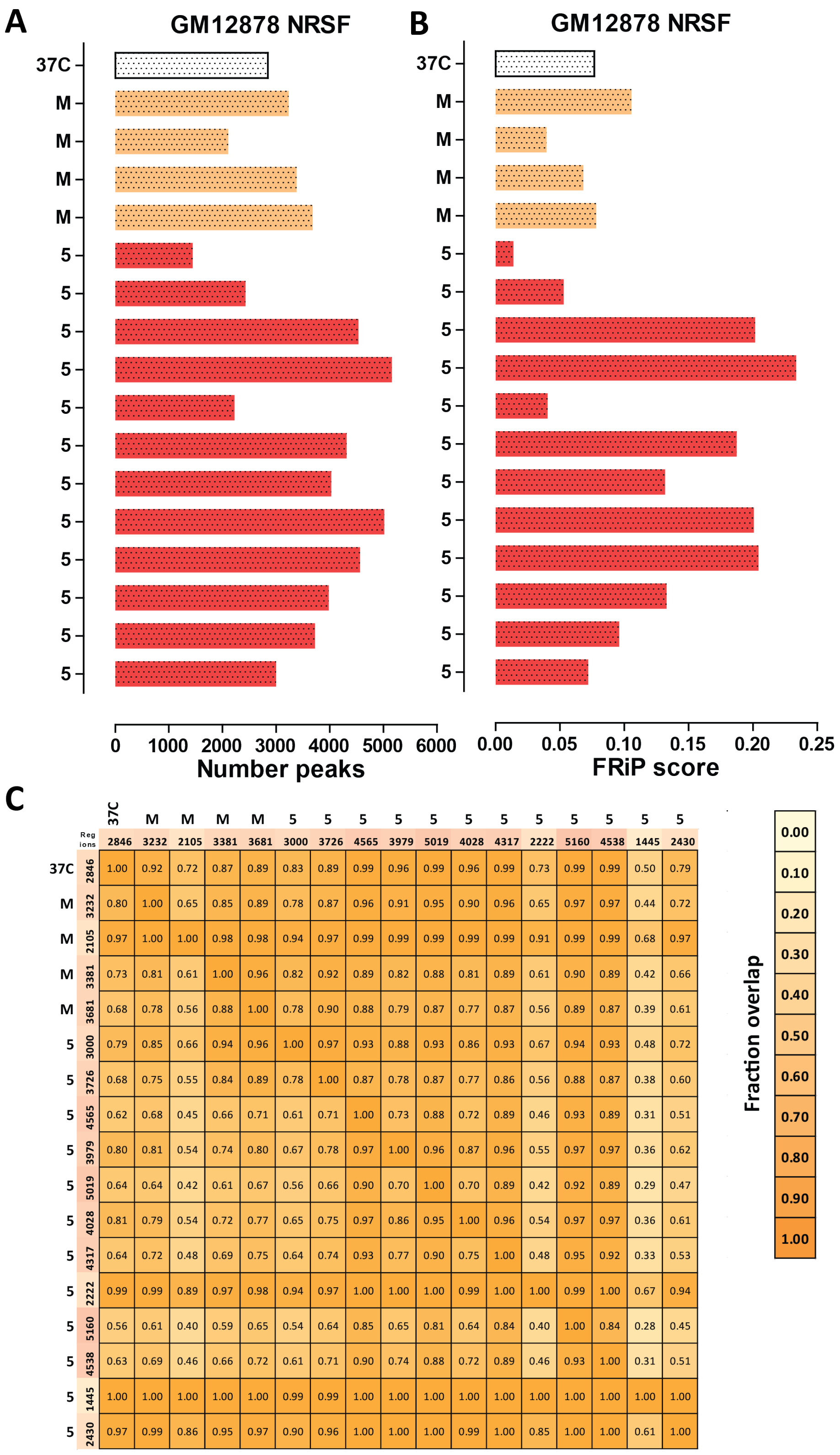


Figure 11.6: Comparison between p300 ChIP-seq resullts on GM12878 chromatin fixed under standard fixation conditions and chromatin fixed at 37 °C. The 1F4-E10P monoclonal antibody was used for all datasets. (A) Number of called peaks; (B) ChIP-enrichment measured by FRiP.



		37C	M	M	M	M	5	5	5	5	5	5	5	5	5	5	5	5
	Regions	2846	3232	2105	3381	3681	3000	3726	4565	3979	5019	4028	4317	2222	5160	4538	1445	2430
37C	2846	1.00	0.92	0.72	0.87	0.89	0.83	0.89	0.99	0.96	0.99	0.96	0.99	0.73	0.99	0.99	0.50	0.79
M	3232	0.80	1.00	0.65	0.85	0.89	0.78	0.87	0.96	0.91	0.95	0.90	0.96	0.65	0.97	0.97	0.44	0.72
M	2105	0.97	1.00	1.00	0.98	0.98	0.94	0.97	0.99	0.99	0.99	0.99	0.99	0.91	0.99	0.99	0.68	0.97
M	3381	0.73	0.81	0.61	1.00	0.96	0.82	0.92	0.89	0.82	0.88	0.81	0.89	0.61	0.90	0.89	0.42	0.66
M	3681	0.68	0.78	0.56	0.88	1.00	0.78	0.90	0.88	0.79	0.87	0.77	0.87	0.56	0.89	0.87	0.39	0.61
5	3000	0.79	0.85	0.66	0.94	0.96	1.00	0.97	0.93	0.88	0.93	0.86	0.93	0.67	0.94	0.93	0.48	0.72
5	3726	0.68	0.75	0.55	0.84	0.89	0.78	1.00	0.87	0.78	0.87	0.77	0.86	0.56	0.88	0.87	0.38	0.60
5	4565	0.62	0.68	0.45	0.66	0.71	0.61	0.71	1.00	0.73	0.88	0.72	0.89	0.46	0.93	0.89	0.31	0.51
5	3979	0.80	0.81	0.54	0.74	0.80	0.67	0.78	0.97	1.00	0.96	0.87	0.96	0.55	0.97	0.97	0.36	0.62
5	5019	0.64	0.64	0.42	0.61	0.67	0.56	0.66	0.90	0.70	1.00	0.70	0.89	0.42	0.92	0.89	0.29	0.47
5	4028	0.81	0.79	0.54	0.72	0.77	0.65	0.75	0.97	0.86	0.95	1.00	0.96	0.54	0.97	0.97	0.36	0.61
5	4317	0.64	0.72	0.48	0.69	0.75	0.64	0.74	0.93	0.77	0.90	0.75	1.00	0.48	0.95	0.92	0.33	0.53
5	2222	0.99	0.99	0.89	0.97	0.98	0.94	0.97	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	0.67	0.94
5	5160	0.56	0.61	0.40	0.59	0.65	0.54	0.64	0.85	0.65	0.81	0.64	0.84	0.40	1.00	0.84	0.28	0.45
5	4538	0.63	0.69	0.46	0.66	0.72	0.61	0.71	0.90	0.74	0.88	0.72	0.89	0.46	0.93	1.00	0.31	0.51
5	1445	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	2430	0.97	0.99	0.86	0.95	0.97	0.90	0.96	1.00	1.00	1.00	0.99	1.00	0.85	1.00	1.00	0.61	1.00

Figure 11.7: Comparison between NRSF ChIP-seq resullts on GM12878 chromatin fixed under standard fixation conditions and chromatin fixed at 37 °C. (A) Number of called regions; (B) ChIP enrichment as measured by FRiP; (C) Overlap between called peaks. The overlap score (O_{XY}) shown in each box indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{XY} = |X \cap Y|/|Y|$;

standard-fixation ones (Figure 11.3G-L), however, at present these datasets constitute a very small sample size and it is still not possible to draw comprehensive conclusions about the 37 °C fixation condition. A much wider survey of different factors will be necessary for that purpose.

11.3 Discussion

The robotic ChIP (R-ChIP) reported here was developed on a widely used commercial liquid-handling platform (Tecan) whose configuration and running program for ChIP are provided. Our initial goal was to increase ChIP-seq throughput, uniformity, and quality, while reducing investigator tedium and error in the context of a large consortium project, but this platform could also be put to work for widespread small projects through core or contract facilities. Our R-ChIP results were comparable in quality to those from the manual pipeline history by all metrics. However, this is a new protocol, and the platform's performance is not perfect. We expect that we and others will continue to make improvements. Specifically, we have observed some sporadic single reaction failures for duplicate samples on a single plate. It is our standard practice to include on each R-ChIP plate a minimum of triplicate control samples deployed across the plate geometry. We use a monoclonal reagent and a large batch of control chromatin to allow comparisons over plates through time. This allows us to evaluate each plate run and to compare it with other runs. This evaluation can be done as a QC step before committing to building the other libraries and sequencing them, which has clear economic implications.

Troubleshooting is aided by R-ChIP compared to standard manual practices. If a group of failed samples are embedded in a large R-ChIP run where the controls and other samples are successful, it becomes unlikely that the ChIP process is the source of failure, and a user can turn attention to the input sample and immune reagent (or any post-ChIP variation) as more likely problem. Of course, the overall success of ChIP-seq includes the local DNA sequencing process, which can be differentially sensitive to the mass of sample, fragment size, and other characteristics of a ChIP output.

We used R-ChIP to screen for monoclonal p300 ChIP-seq antibodies as a further test of R-ChIP. The p300 protein is a "high value" target for ChIP-seq because a map of active transcriptional enhancer candidates is often wanted. Many antibodies made against transcription factors fail in ChIP reactions, even though they work well in one or more conventional uses (e.g. standard immunoprecipitation, western blots or immunocytochemical stains). Moreover, polyclonal reagents that are ChIP-seq compatible typically vary, sometimes greatly, in their specificity and performance from lot to lot (Egelhofer et al. 2011). The upshot has been that the only way to identify a ChIP-quality antibody is to test it directly for ChIP, and the most general way to ensure reliability and unlimited supply is with a monoclonal antibody. Whether the final readout for ChIP competence is DNA sequencing or qPCR (the latter requiring known targets for the factor), the capacity to test many ChIP reactions is critical for screening. Here we used R-ChIP to identify a ChIP-grade monoclonal for p300.. The interactomes identified with this hybridoma clone, ENCITp300-1, overlap highly with prior measurements from ENCODE for the same cell line and with concurrent polyclonal determinations, confirming its specificity for p300 and the utility of R-ChIP for screening new immune reagents for ChIP. We note, however, that the data obtained with it are not as inclusive as the best ones produced using polyclonal rabbit reagents.

For p300 R-ChIP, we used chromatin from cells fixed with a modified condition (37 C, 1% formaldehyde, 30 min) which was recommended to us by Dr. Bing Ren (UC San Diego), whose laboratory has extensive experience with p300 ChIP. This improved p300 ChIP significantly in our hands compared with our standard fixation condition (see Results and Methods), while it had no detectable effect on NRSF, GABP and H3K27ac ChIP. However, we emphasize that we do not know if, or how frequently, this more stringent fixation condition will affect other factor-antibody pairs. Epitope destruction or occlusion, or elevated signals from lower affinity interactions, are among the plausible negative effects. The most positive impact is expected for proteins that interact indirectly with DNA, as p300 does.

By increasing the reliability and throughput of ChIP-seq and by liberating investigator time from a tedious and nontrivial experimental protocol, we anticipate that R-ChIP, and variations on it, will break a current bottleneck and help to advance a wide range of transcription investigations.

11.4 Methods

11.4.1 Cell growth and harvesting

Cells were grown and harvested following established ENCODE protocols (available at `http://genome.ucsc.edu/ENCODE/cellTypes.html`) with the exception of GM12878 p300 experiments for which chromatin was fixed at 37°C.

11.4.2 Chromatin Preparation

Chromatin was cross-linked by adding formaldehyde directly to the cell culture media at a final concentration of 1% and gently mixed for 10 minutes. The exception was (where indicated) fixation at 37 °C for 30 minutes, which was used for p300 experiments. In all cases, the formaldehyde reaction was quenched by adding glycine to a final concentration of 0.125M for 10 minutes. Cells were then pelleted, rinsed once in cold phosphate-buffered saline (PBS) with 1mM PMSF and once in cold MC lysis buffer (10mM Tris pH 7.5, 10mM NaCl, 3mM MgCl2, 0.5% NP-40, and Roche Complete Protease Inhibitor Cocktail) to obtain nuclear pellets. Nuclei were sonicated in RIPA buffer (PBS, 1% NP-40 Substitute, 0.5% Sodium Deoxycholate, 0.1% SDS, and Roche Complete Protease Inhibitor Cocktail) at a concentration of at least 5×10^7 nuclei/mL using a probe sonic dismembrator from Fisher Scientific (Model 550). To check for fragment size distribution after sonication, a small fraction of the sample was reverse cross-linked for two hours at 65 °C, purified using DNA purification columns from Qiagen, then loaded onto a 2% agarose EtBr E-Gel from Invitrogen.

11.4.3 Antibodies used

The following antibodies were used: an α-NRSF mouse monoclonal (12C11) from the Anderson Lab at Caltech (Mortazavi et al. 2006; Johnson et al. 2007), an α-p300 rabbit polyclonal (sc-585) from Santa Cruz Biotechnology, a mouse monoclonal α-GABP (sc-28312) from Santa Cruz Biotechnology, a mouse monoclonal α-Kaiso/ZBTB33 (sc-23871) from Santa Cruz Biotechnology, a rabbit polyclonal α-SPI1/PU.1 (sc-22805) from Santa Cruz Biotechnology, and a mouse monoclonal α-H3K27ac (306-34849) from Wako. In addition 11 α-p300 mouse monoclonals were generated in the Caltech Mouse Monoclonal Facility. Four of the α-p300 mouse monoclonals were raised against a bacterially expressed GST fusion protein containing N-terminal residues 152-213. The remaining seven antibodies were raised against a synthetic peptide from GenScript containing C-terminal residues 1526-1545.

11.4.4 Robotic-ChIP (R-ChIP) Workflow

ChIP experiments were adapted from methods previously described and optimized for performance in a 96-well plate format using a Tecan Freedom EVO 200 liquid handling robot. Reagents and labware are placed on deck of the robot.

After setup the R-ChIP workflow is completely hands-off and consists of a series of modules with a run time of ~24 hours to run, including the 12-hour reverse cross-linking step. All aspects of the setup are checked thoroughly to ensure a smooth run.

1. **Blocking and Washing of Magnetic Beads**. The Tecan begins by resuspending magnetic beads (Invitrogen M-280 Dynabeads) from the source tube with the liquid-handling arm (LiHa) and dispenses the magnetic beads into a Fisher 96-Well DeepWell ^{TM} Polypropylene known as the ChIP plate. 100 μL of beads is used for a monoclonal IP antibody and 200 μL for a polyclonal. The LiHa tips are evacuated and rinsed with ddH$_2$O between subsequent dispenses to prevent cross-contamination. 500 μL of PBS containing 5% bovine serum albumin (BSA) is then dispensed by the LiHa from a buffer reservoir (Te-Fill) to block and wash the magnetic beads. The plate containing the beads is transferred to an orbital mixer (Te-Shake) with the robotic manipulator arm (RoMa) and mixed several times for 20 seconds with a 20 second pause between each mix. The RoMa moves the bead plate to a magnetic plate for seven minutes where the beads are then pelleted in a ring allowing the multi-channel arm (MCA96) fitted with natural 200 μL tips from TipOne to aspirate liquid. These steps are repeated three more times and include an ethanol rinse of the MCA96 tips as needed to prevent cross-contamination.

2. **Binding of Antibody to Magnetic Beads**. The LiHa adds 400 μL of PBS-BSA to the antibody plate bringing the final volume to 500 μL. The antibody is then added to the beads using the MCA 96 which transfers the diluted antibody from the a 2.0-mL 96-well PlateOne V-bottom plate to the ChIP plate. For monoclonal antibodies, 5 μg of antibody were diluted in 500 μL (10 μg for polyclonals). The beads and antibody are incubated together for one hour with mixing using the Te-Shake. Any unbound antibody is then aspirated with the MCA96 and deposited into a fresh 2.0-mL 96-well PlateOne V-bottom plate for further analysis if needed.

3. **Incubation of Chromatin and Antibody-Bead Complex** The MCA96 transfers 500 μL of chromatin containing 2.5×10^7 cells from the Matrix tube rack to the ChIP plate. The chromatin and antibody bead complex are then incubated together for 2 hours during which the ChIP plate alternates between the Te-Shake and a 4 °C cool plate using the RoMa. The chromatin is stored in 1.2 mL screw-top Matrix tubes that can be arrayed on the chromatin plate as needed. Any unbound chromatin is then aspirated with the MCA96 and deposited into a fresh 2.0-mL 96-well PlateOne V-bottom plate for further analysis if needed.

4. **Washing of IP Complex**. The LiHa dispenses 500 μL of LiCl wash buffer (100mM Tris pH 7.5, 500mM Lithium Chloride, 1% NP-40, 1% sodium deoxycholate) from the Te-Fill onto the beads, which are then mixed for 20 seconds with 20 second pauses between each mix. The beads are then pelleted and the wash with LiCl buffer is repeated four more times. The LiHa then adds 500 μL of TE buffer (10mM Tris pH 7.5 and 1mM EDTA) and resuspends the beads with the Te-Shake for 20 seconds. Beads are then pelleted with the magnetic plate and any remaining buffer is aspirated and discarded by the MCA96.

5. **Elution from Beads**. The LiHa dispenses 100 μL of IP elution buffer (1% SDS and 0.1M $NaHCO_3$) from the Te-Fill and the beads are resuspended by mixing for 20 seconds with the Te-Shake. The MCA96 then aspirates the beads and transfers them from the ChIP plate to a Hard-Shell Semi-Skirted PCR Plate from Bio-Rad. The RoMa transfers a PCR lid from the storage hotel and places it on top of the PCR plate then transfers the lidded PCR plate to a DNA Engine Peltier Thermal Cycler with Remote Alpha Dock System from Bio-Rad. The top of the thermal cycler closes and places force on the PCR plate lid creating a seal. The beads are then heated for one hour at 65 °C to disassociate the IP complex from the magnetic beads.

6. **Reversal of Cross-links**. The RoMa takes the PCR plate from the thermal cycler and transfers it to the Te-Shake to resuspend the beads. The PCR plate is then transferred to the magnetic plate for pelleting of the beads. The MCA96 mounts 50 μL Tecan Pure Disposable tips from Tecan, slowly aspirates the supernatant, and transfers it to a fresh PCR plate. 10μL of proteinase K from Epicentre diluted 1:5 in proteinase K buffer (50% glycerol, 50 mM Tris-HCl pH 7.5, 0.1 M NaCl, 0.1 mM EDTA, 1 mM DTT, 10 mM $CaCl_2$, 0.1% Triton® X-100) is then added to the supernatants with the LiHa. The RoMa places a lid on the fresh PCR plate and transfers both back to the thermal cycler for a 12 hour incubation at 65 °C to reverse the cross-links. Once the incubation is finished the plate is transferred to the deck with the RoMa and the R-ChIP is complete.

11.4.5 DNA Cleanup

Samples from the R-ChIP ChIP experiments presented here were then cleaned up manually using the protocol described by Qiagen in their Qiaquick PCR purification kit with the addition that the EB buffer is heated to 55 °C prior to elution and eluted in a 50 μL volume using DNA lo-bind 1.5 mL tubes from Eppendorf. We anticipate automating this step.

11.4.6 Library building and sequencing

Library building for sequencing on the Illumina HiSeq platform was performed conventionally with barcoding to allow multiple ChIP libraries to be sequenced in a single flow cell lane, according to the Hudson Alpha ENCODE ChIP protocol.

Standard methods were used for end repair and dA addition of DNA fragments recovered from ChIPs or chromatin controls. The fragments were then ligated to Illumina Paired-End adaptor sequences and PCR-amplified to complete the adaptor sequences and introduce a 7-base DNA barcode in the i7 position. The barcode allowed mixing of multiple samples per flowcell lane. Control libraries were prepared from 500 ng of DNA from reverse crosslinked sonicated chromatin. ChIP library starting amounts varied by ChIP, with a median of 7.5 ng. Fragment size selection was achieved at the lower threshold with solid phase reversible immobilization (SPRI) technology to recover dsDNA greater than 100 bp after adaptor ligation (thereby excluding unligated adaptors) and at the upper threshold with an extension time of 30 seconds during PCR amplification. This size selection method consistently produced final DNA library fragments that ranged from ~100 to 400 bp, as determined by BioAnalysis. Final library amounts varied by ChIP, with a median of 546 ng.

Libraries were pooled in equimolar amounts and sequenced on the Illumina HiSeq2000 or HiSeq2500 with 50 bp single-end reads following the manufacturer's recommendations. Raw sequencing reads are available from GEO accession number GSE53366.

11.4.7 Data processing and analysis

Reads were aligned using Bowtie (Langmead et al. 2009), version 0.12.7, with the following settings: `‘‘-v 2 -t -k 2 -m 1 --best --strata’’`, which allow for two mismatches relative to the reference and only retain unique alignments, against the `hg19` version of the human genome (assembly downloaded from the UCSC genome browser) with the Y chromosome retained or removed depending on the sex of the cell line. Peak calling was carried out using ERANGE (Johnson et al. 2007), version 4.0, with the following settings: `--minimum 2 --ratio 3 --listPeak --shift learn --revbackground`, against matching control samples. Library complexity was estimated as described in Landt et al., 2012. Cross-correlation analysis was carried out using version 1.10.1 of `SPP` (Kharchenko et al. 2008; A. Kundaje et al., submitted) and the following parameters: `‘-s=0:2:400’`.

All additional analysis was carried out using custom-written Python scripts.

Read mapping statistics for all datasets are provided in Tables 11.1, 11.2 and 11.3.

Table 11.1: Read mapping and dataset quality statistics for robotic NRSF ChIP datasets. Quality control scores were determined using SPP as described in Landt et al. 2012 and Marinov et al. 2014

Cell Line	Factor	Plate	Well	Library	Complexity	NSC	RSC	QC	Number Peaks	FRiP	Uniquely Mapped reads
GM12878	NRSF	3365	A2	SL26735	0.84	6.165	1.215	1	3,821	0.127	12,246,881
GM12878	NRSF	3365	E6	SL26736	0.86	5.675	1.235	1	4,030	0.144	12,166,760
GM12878	NRSF	3365	F12	SL26737	0.85	3.747	1.143	1	3,349	0.103	16,531,315
GM12878	NRSF	3405	B1	SL28743	0.91	3.808	1.044	1	2,143	0.037	16,286,924
GM12878	NRSF	3405	E6	SL28744	0.89	4.343	1.107	1	2,413	0.048	19,731,464
GM12878	NRSF	3405	F12	SL28745	0.9	3.67	1.092	1	2,295	0.044	19,898,394
GM12878	NRSF	3646	B1	SL34357	0.97	2.541	0.824	0	1,569	0.015	7,567,674
GM12878	NRSF	3646	E6	SL34381	0.97	2.816	0.906	0	1,642	0.017	8,035,350
GM12878	NRSF	3646	F5	SL34380	0.88	3.915	0.811	0	2,087	0.029	4,774,294
GM12878	NRSF	3646	F6	SL34382	0.93	6.315	1.198	1	3,502	0.095	6,378,277
GM12878	NRSF	3646	F7	SL34383	0.93	8.092	1.197	1	3,547	0.095	6,404,751
GM12878	NRSF	3646	G12	SL34384	0.97	1.235	0.815	0	1,656	0.016	6,231,497
GM18278	NRSF	4028	D3	SL46179	0.87	2.729	1.411	1	3,000	0.072	8,672,020
GM18278	NRSF	4028	E3	SL46180	0.66	2.709	1.718	2	3,726	0.096	17,661,548
GM18278	NRSF	4028	A12	SL46217	0.61	4.030	1.262	1	3,979	0.132	15,999,430
GM18278	NRSF	4028	A12	SL46171	0.67	1.444	1.410	1	4,565	0.204	24,180,628

Continued on next page

Table 11.1 – *Continued from previous page*

Cell Line	Factor	Plate	Well	Library	Complexity	NSC	RSC	QC	Number Peaks	FRiP	Uniquely Mapped reads
GM18278	NRSF	4028	B11	SL46211	0.51	2.510	1.356	1	5,019	0.200	20,187,765
GM18278	NRSF	4028	B2	SL46173	0.87	2.813	1.128	1	4,028	0.131	10,454,082
GM18278	NRSF	4028	F3	SL46181	0.74	1.461	1.623	2	4,317	0.187	25,430,927
GM18278	NRSF	4028	G11	SL46216	0.89	3.999	1.170	1	2,222	0.040	13,250,634
GM18278	NRSF	4028	G2	SL46176	0.67	1.455	1.907	2	5,160	0.233	21,980,053
GM18278	NRSF	4028	G3	SL46182	0.71	1.479	1.443	1	4,538	0.201	23,259,881
GM18278	NRSF	4028	H12	SL46218	0.85	3.855	1.226	1	2,430	0.052	20,550,794
GM18278	NRSF	4028	H1	SL46172	0.92	2.018	0.975	0	1,445	0.013	24,630,268
GM12878	NRSF	Manual	Manual	ENCODE	0.78	2.121	1.089	1	3,085	0.019	11,945,180
GM12878	NRSF	Manual	Manual	ENCODE	0.89	6.594	2.081	2	3,363	0.070	16,286,742
GM18278	NRSF	Manual	Manual	SL45074	0.89	3.012	1.266	1	3,232	0.105	30,326,354
GM18278	NRSF	Manual	Manual	SL45075	0.94	3.237	1.162	1	2,105	0.039	24,260,996
GM18278	NRSF	Manual	Manual	SL45072	0.90	2.409	1.629	2	3,381	0.068	28,464,754
GM18278	NRSF	Manual	Manual	SL45073	0.87	2.286	1.843	2	3,681	0.078	37,706,507
GM18278 37 °C	NRSF	4028	B10	SL46206	0.88	1.759	1.033	1	2,846	0.076	24,978,799
Jurkat	NRSF	3365	A1	SL26729	0.94	4.948	1.017	1	1,792	0.025	14,447,433
Jurkat	NRSF	3365	B1	SL26732	0.93	4.99	1.061	1	1,899	0.029	17,281,495
Jurkat	NRSF	3365	C1	SL26733	0.78	8.33	1.225	1	3,833	0.144	12,742,979
Jurkat	NRSF	3365	D1	SL26734	0.73	12.468	1.218	1	3,995	0.144	7,551,313
Jurkat	NRSF	3365	D6	SL26730	0.92	8.756	1.141	1	2,430	0.057	13,738,597
Jurkat	NRSF	3365	H12	SL26731	0.92	5.374	1.083	1	1,906	0.029	16,829,422
Jurkat	NRSF	3405	A1	SL28740	0.97	1.075	0.516	0	277	0.001	18,013,669
Jurkat	NRSF	3405	D6	SL28741	0.85	8.854	1.144	1	2,785	0.072	13,132,743
Jurkat	NRSF	3405	H12	SL28742	0.73	8.507	1.141	1	2,937	0.078	14,966,401
Jurkat	NRSF	3435	A1	SL29213	0.96	1.125	0.739	0	1,020	0.006	17,997,043
Jurkat	NRSF	3435	D6	SL29214	0.84	6.077	1.183	1	2,852	0.084	19,801,560
Jurkat	NRSF	3435	H12	SL29215	0.84	6.727	1.189	1	3,062	0.098	18,127,638
Jurkat	NRSF	3549	A1	SL31830	0.7	6.928	1.1	1	2,233	0.045	18,458,174
Jurkat	NRSF	3549	D6	SL31851	0.91	3.993	0.957	0	1,702	0.020	16,754,683
Jurkat	NRSF	3549	H12	SL31882	0.89	1.182	0.852	0	1,331	0.010	19,956,019

Table 11.2: Read mapping and dataset quality statistics for p300 datasets. Quality control scores were determined using SPP as described in Landt et al. 2012 and Marinov et al. 2014

Cell Line	Antibody	Library	Complexity	NSC	RSC	QC	Number Peaks	FRiP	Uniquely Mapped reads
GM12878 37 °C	1A3-F8p	SL31840	0.85	1.244	0.293	-1	3	0	18,894,243
GM12878 37 °C	1F4-E10	SL31838	0.90	1.645	0.422	-1	4,870	0.019	13,446,233
GM12878 37 °C	2E10-D10	SL31839	0.87	1.364	0.269	-1	6	0	13,916,005
GM12878 37 °C	2F4-A8	SL31832	0.85	1.217	0.292	-1	40	0	17,928,468
GM12878 37 °C	2F6-F7	SL31831	0.72	1.541	0.366	-1	9	0	15,450,180
GM12878 37 °C	3B4-G6	SL31835	0.95	1.463	0.266	-1	45	0.001	13,546,541
GM12878 37 °C	3H6-B6	SL31834	0.92	1.242	0.242	-2	5	0	15,393,775
GM12878 37 °C	4C5-A1	SL31833	0.83	1.439	0.287	-1	15	0	12,550,517
GM12878 37 °C	5D2-A1	SL31836	0.89	1.422	0.270	-1	2	0	11,297,381
GM12878 37 °C	5F7-C9	SL31841	0.86	1.365	0.392	-1	2,868	0.010	14,992,412
GM12878 37 °C	7H5-F2	SL31842	0.83	1.447	0.421	-1	1,524	0.051	14,582,010
GM12878 37 °C	1F4-E10	SL34359	0.96	1.836	0.509	0	6,374	0.031	4,802,800

Continued on next page

Table 11.2 – *Continued from previous page*

Cell Line	Antibody	Library	Complexity	NSC	RSC	QC	Number Peaks	FRiP	Uniquely Mapped reads
GM12878 37 °C	sc585 lot# E3113	SL34362	0.95	3.467	1.281	1	34,333	0.317	5,811,746
GM12878 37 °C	sc585 lot# F2711	SL34358	0.94	3.659	1.239	1	36,868	0.342	6,688,765
GM12878 37 °C	1F4-E10	SL46203	0.88	1.673	0.852	0	6,333	0.030	20,062,190
GM12878 37 °C	1F4-E10	SL46209	0.87	2.027	0.874	0	6,725	0.030	16,221,052
GM12878 37 °C	sc585 lot# E3113	SL46202	0.88	2.867	1.628	2	28,447	0.257	26,343,063
GM12878 37 °C	sc585 lot# E3113	SL46205	0.92	2.585	1.431	1	11,369	0.066	27,085,241
GM12878 37 °C	sc585 lot# F2711	SL46204	0.93	2.907	1.381	1	15,093	0.091	15,854,599
GM12878 37 °C	1F4-E10	SL45094	0.96	1.798	0.993	0	8,430	0.042	31,243,370
GM12878 37 °C	1F4-E10	SL45095	0.96	1.663	0.978	0	6,181	0.031	33,548,275
GM12878	1F4-E10	SL45092	0.97	1.224	0.580	0	108	0.0004	26,071,005
GM12878	1F4-E10	SL45093	0.97	1.237	0.588	0	252	0.0009	26,022,658
GM12878	1F4-E10	SL46207	0.91	1.146	0.217	-2	37	0.0001	18,794,116
GM12878	1F4-E10	SL46208	0.91	1.152	0.236	-2	17	0.0000	20,417,157
GM12878	sc584	ENCODE	0.92	1.549	0.759	0	12,924	0.063	15,906,721
GM12878	sc584	ENCODE	0.91	1.639	0.765	0	4,510	0.018	16,950,416
GM12878	sc585	ENCODE	0.86	2.088	1.258	1	8,267	0.043	23,366,821
GM12878	sc585	ENCODE	0.88	1.292	0.698	0	2,610	0.011	20,403,419

Table 11.3: Read mapping and dataset quality statistics for H3K27ac, GABP, ZBTB33, PU.1 and input datasets. Quality control scores were determined using SPP as described in Landt et al. 2012 and Marinov et al. 2014

Cell Line	Factor	Rep	R/M	Library	Complexity	NSC	RSC	QC	Number Peaks	FRiP	Uniquely Mapped reads
GM12878 37 °C	Input	Rep1		SL45100	0.96	1.308	0.748	0			25,588,571
GM12878 37 °C	Input	Rep2		SL45101	0.96	1.242	0.613	0			28,422,425
GM12878	Input	Rep3		SL45098	0.96	1.475	0.763	0			21,218,915
GM12878	Input	Rep4		SL45099	0.96	1.381	0.820	0			26,213,457
GM12878	GABP	Rep1	M	SL45068	0.86	2.456	2.343	2	5,694	0.144	28,778,500
GM12878	GABP	Rep2	M	SL45069	0.90	2.615	2.296	2	4,675	0.096	29,626,523
GM12878	H3K27ac	Rep3	M	SL45090	0.89	1.629	1.945	2	35,570	0.502	31,263,444
GM12878	H3K27ac	Rep4	M	SL45090	0.87	1.699	1.811	2	34,619	0.497	36,587,615
GM12878	H3K27ac	Rep1	M	SL45090	0.92	1.463	1.593	2	28,580	0.384	30,476,218
GM12878	H3K27ac	Rep2	M	SL45090	0.93	1.444	1.541	2	32,345	0.386	25,852,868
GM12878	Input	Rep1		SL45096	0.98	1.210	0.683	0			33,254,931
GM12878	Input	Rep2		SL45097	0.98	1.090	0.272	-1			32,985,584
GM12878	PU.1	Rep1	M	SL45076	0.88	6.960	2.143	2	19,779	0.182	27,166,940
GM12878	PU.1	Rep2	M	SL45077	0.89	6.197	2.166	2	17,346	0.143	31,549,754
GM12878	ZBTB33	Rep1	M	SL45080	0.94	1.992	1.250	1	2,066	0.030	22,662,929
GM12878	ZBTB33	Rep2	M	SL45081	0.95	1.614	1.159	1	1,242	0.016	27,577,513
GM12878 37 °C	GABP	Rep1	R	SL46201	0.85	4.153	1.538	2	4,630	0.111	18,416,079
GM12878 37 °C	H3K27ac	Rep1	R	SL46215	0.87	1.542	3.067	2	36,216	0.509	21,994,873
GM12878 37 °C	PU.1	Rep1	R	SL46212	0.89	1.513	0.656		1,961	0.007	22,811,938
GM12878 37 °C	ZBTB33	Rep1	R	SL46213	0.95	1.373	0.376		475	0.004	10,385,335
GM12878	GABP	Rep1	R	SL46174	0.73	3.779	2.651		5,158	0.151	10,708,570
GM12878	GABP	Rep2	R	SL46175	0.63	3.524	3.101		6,170	0.198	13,707,692
GM12878	H3K27ac	Rep3	R	SL46199	0.88	1.590	2.402		40,102	0.511	21,229,446
GM12878	H3K27ac	Rep4	R	SL46200	0.88	1.553	2.801		37,038	0.520	24,739,874
GM12878	H3K27ac	Rep1	R	SL46197	0.92	1.379	1.974		29,449	0.337	20,680,756

Continued on next page

Table 11.3 – *Continued from previous page*

Cell Line	Antibody	Rep	R/M	Library	Complexity	NSC	RSC	QC	Number Peaks	FRiP	Uniquely Mapped reads
GM12878	H3K27ac	Rep2	R	SL46198	0.93	1.358	1.861		32,267	0.343	21,320,237
GM12878	PU.1	Rep1	R	SL46183	0.79	9.692	2.561	2	22,376	0.246	21,431,025
GM12878	PU.1	Rep2	R	SL46184	0.91	2.242	1.561	2	7,383	0.040	23,919,950
GM12878	ZBTB33	Rep1	R	SL46187	0.92	1.891	1.203	1	1,343	0.016	17,536,255
GM12878	ZBTB33	Rep2	R	SL46188	0.92	1.796	1.032	1	1,157	0.013	19,265,589

Part IV

Other Projects

12

The role of Piwi in piRNA-guided transcriptional silencing and establishment of repressive chromatin

The data in this chapter was generated by the Fejes-Tóth and Aravin labs; my contribution was in carrying out the analysis for it. Most of the material in it consists of what was previously published as:

Le Thomas A*, Rogers AK*, Webster A*, Marinov GK*, Liao SE, Perkins EM, Hur JK, Aravin AA, and Fejes-Tóth K. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev* **27**:390–399. doi: 10.1101/gad.209841.112.

The paper is reprinted in Appendix F.

I have also added some further analysis that I did that refutes certain claims about the way Piwi functions in the nucleus that appeared in the literature after our paper was published.

Abstract

In the metazoan germline, piwi proteins and associated piwi-interacting RNAs (piRNAs) provide a defense system against the expression of transposable elements. In the cytoplasm, piRNA sequences guide piwi complexes to destroy complementary transposon transcripts by endonucleolytic cleavage. However, some piwi family members are nuclear, raising the possibility of alternative pathways for piRNA-mediated regulation of gene expression. We found that *Drosophila* Piwi is recruited to chromatin, colocalizing with RNA polymerase II (Pol II) on polytene chromosomes. Knockdown of Piwi in the germline increases expression of transposable elements that are targeted by piRNAs, whereas protein-coding genes remain largely unaffected. Derepression of transposons upon Piwi depletion correlates with increased occupancy of Pol II on their promoters. Expression of piRNAs that target a reporter construct results in a decrease in Pol II occupancy and an increase in repressive H3K9me3 marks and heterochromatin protein 1 (HP1) on the reporter locus. Our results indicate that Piwi identifies targets complementary to the associated piRNA and induces transcriptional repression by establishing a repressive chromatin state when correct targets are found. More recently, a different model for Piwi's action has been proposed, which features Piwi binding strongly and very specifically to repetitive elements in the genome (even those that are not expressed). I show why that model is wrong and based on flawed data.

12.1 Introduction

Diverse small RNA pathways function in all kingdoms of life, from bacteria to higher eukaryotes. In eukaryotes, several classes of small RNA associate with members of the Argonaute protein family, forming effector complexes in which the RNA provides target recognition by sequence complementarity, and the Argonaute provides the repressive function. Argonautes-small RNA complexes have been shown to regulate gene expression both transcriptionally and

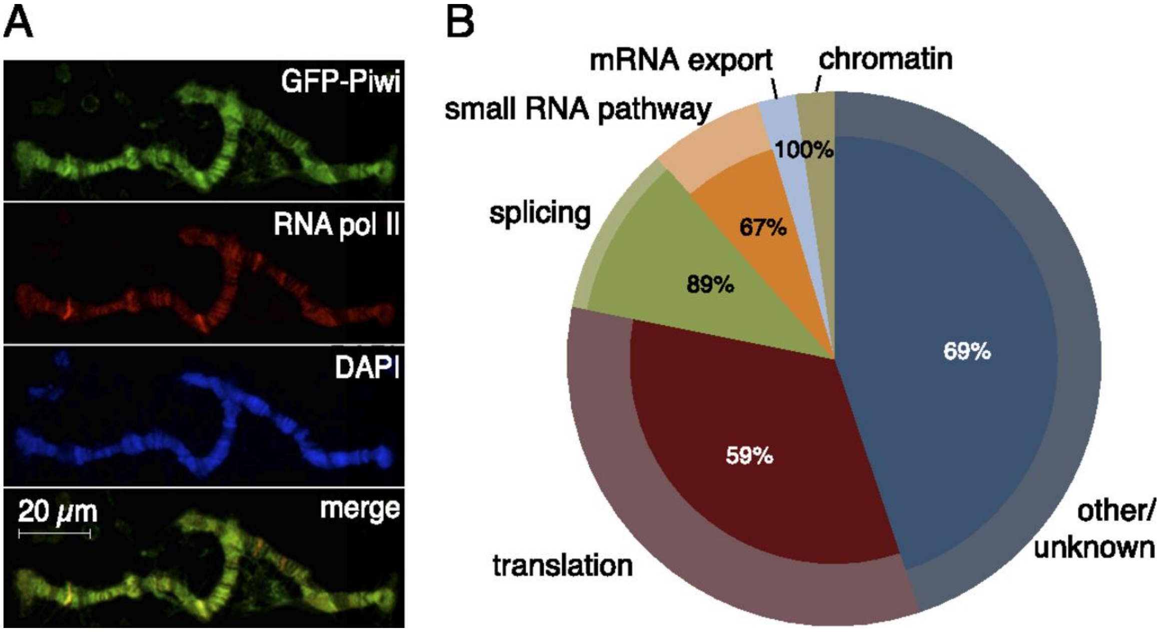
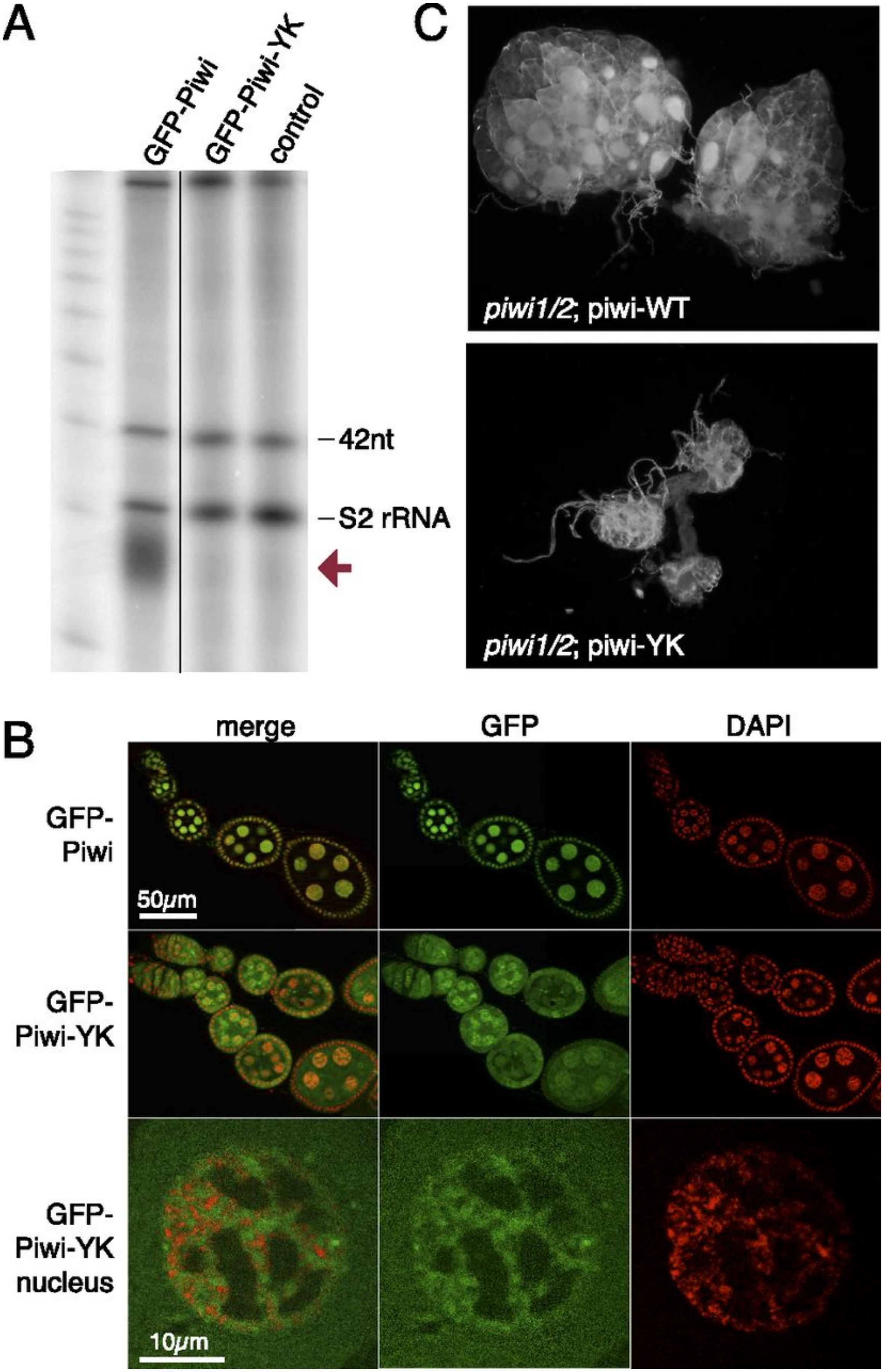


Figure 12.1: Piwi associates with chromatin and nuclear transcripts. (A) Polytene chromosomes from *Drosophila* nurse cells expressing GFP-Piwi on the *otu*[7]/*otu*[11] background. Piwi pattern on chromosomes correlates with Pol II staining. (B) Mass spectrometry analysis of Piwi interaction partners. Piwi complexes were precipitated in the presence and absence of RNase A. The outer circle represents classification of Piwi-associated proteins based on GO term analysis. The inner pies represent the fraction of each group whose association with Piwi depends on RNA (percentage indicated). Note that chromatin, splice, and mRNA export factors are virtually absent after RNase A treatment.

post-transcriptionally. Post-transcriptional repression involves cleavage of target RNA through either the endonucleolytic activity of Argonautes or sequestering targets into cytoplasmic ribonucleoprotein (RNP) granules (Hutvagner & Simard 2008).

The mechanism of transcriptional repression by small RNAs has been extensively studied in fission yeast and plants. Several studies showed that Argonautesmall RNA complexes induce transcriptional repression by tethering chromatin modifiers to target loci. In fission yeast, the effector complex containing the Argonaute and the bound siRNA associates with the histone H3 Lys 9 (H3K9) methyltransferase Clr4 to install repressive H3K9-dimethyl marks at target sites (Nakayama et al. 2001; Maison & Almouzni 2004; Sugiyama et al. 2005; Grewal & Jia 2007). Methylation of histone H3K9 leads to recruitment of the heterochromatin protein 1 (HP1) homolog Swi6, enhancing silencing and further promoting interaction with the Argonaute complex. The initial association of Ago with chromatin, however, requires active transcription (Ameyar-Zazoua et al. 2012; Keller et al. 2012). Plants also use siRNAs to establish repressive chromatin at repetitive regions. Contrary to yeast, heterochromatin in plants is marked by DNA methylation, although repression also depends on histone methylation by a Clr4 homolog (Soppe et al. 2002; Onodera et al. 2005). Although siRNA-mediated gene silencing is predominant on repetitive sequences, it is not limited to these sites. Constitutive expression of dsRNA mapping to promoter regions results in production of corresponding siRNAs, de novo DNA methylation, and gene silencing (Mette et al. 2000; Matzke et al. 2004).

In metazoans, small RNA pathways are predominantly associated with post-transcriptional silencing. One class of small RNA, microRNA, regulates expression of a large fraction of protein-coding genes (Friedman et al. 2009). In *Drosophila*, siRNAs silence expression of transposable elements (TEs) in somatic cells (Chung et al. 2008; Ghildiyal et al. 2008) and target vi-


A
GFP-Piwi
GFP-Piwi-YK
control
42nt
S2 rRNA
C
piwi1/2; piwi-WT
piwi1/2; piwi-YK
B
merge
GFP
DAPI
GFP-Piwi
50μm
GFP-Piwi-YK
GFP-Piwi-YK nucleus
10μm

ral genes upon infection (Galiana-Arnoux et al. 2006; Wang et al. 2006; Zambon et al. 2006). Another class of small RNAs, Piwi-interacting RNAs (piRNAs), associates with the Piwi clade of Argonautes and acts to repress mobile genetic elements in the germline of both *Drosophila* and mammals (Siomi et al. 2011). Analysis of piRNA sequences in *Drosophila* revealed a very diverse population of small RNAs that primarily maps to transposon sequences and is derived from a number of heterochromatic loci called piRNA clusters, which serve as master regulators of transposon repression (Brennecke et al. 2007). Additionally, a small fraction of piRNAs seems to be processed from the mRNA of several host protein-coding genes (Robine et al. 2009; Saito et al. 2009). The *Drosophila* genome encodes three piwi proteins: Piwi, Aubergine (AUB), and Argonaute3 (AGO3). In the cytoplasm, AUB and AGO3 work together to repress transposons through cleavage of transposon transcripts, which are recognized through sequence complementarity by the associated piRNAs (Vagin et al. 2006; Agger et al. 2007; Brennecke et al. 2007; Gunawardane et al. 2007).

In both *Drosophila* and mammals, one member of the Piwi clade proteins localizes to the nucleus. Analogously to small RNA pathways in plants, the mouse piRNA pathway is required for de novo DNA methylation and silencing of TEs (Carmell et al. 2007; Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008); however, the exact mechanism of this process is unknown. In *Drosophila*, DNA methylation is absent; however, several studies indicate that elimination of Piwi from the nucleus causes changes in histone marks on TEs (Klenov et al. 2011; Pöyhönen et al. 2012), yet a genome-wide analysis of Piwi's effect on chromatin marks and transcription is lacking.

We showed that Piwi interacts with chromatin on polytene chromosomes in nurse cell nuclei. We found that Piwi exclusively represses loci that are targeted by piRNAs. We showed that Piwi-mediated silencing occurs through repression of transcription and correlates with installment of repressive chromatin marks at targeted loci.

12.2 Results

To analyze the role of Piwi in the nucleus, we generated transgenic flies expressing a GFP-tagged Piwi protein (GFP-Piwi) under the control of its native regulatory region. GFP-Piwi was expressed in the ovary and testis in a pattern indistinguishable from the localization of native Piwi and was able to rescue the piwi-null phenotype as indicated by ovarian morphology, fertility, transposon expression, and piRNA levels. GFP-Piwi was deposited into the mature egg and localized to the pole plasm; however, contrary to a previous observation (Brower-Toland et al. 2007), we did not detect Piwi expression outside of the ovary and testis in third instar larvae or adult flies. We also did not observe the association of Piwi with polytene chromosomes in salivary gland cells of third instar larvae. In both follicular and germline cells of the *Drosophila* ovary, GFP-Piwi localized exclusively in the nucleus, with slightly higher concentrations apparent in regions enriched for DAPI, indicating a possible interaction with chromatin. To gain further insight into Piwi localization in the nucleus, we took advantage of the fact that nurse cell chromosomes are polytenized and can be visualized on the otu mutant background (Mal'ceva et al. 1997). Analysis of polytene chromosomes from nurse cells demonstrated that GFP-Piwi associates with chromatin in a specific banding pattern. Interestingly,

Figure 12.2 *(preceding page)*: Piwi function, but not its nuclear localization, requires piRNA association. (A) The Piwi-YK mutant does not associate with piRNA. Immunoprecipitation of PiwipiRNA complexes was performed with GFP antibody on ovaries from GFP-Piwi and GFP-Piwi-YK transgenic flies and a control strain. Small RNAs were isolated, 5'-labeled, and resolved on a denaturing gel. The same amount of 42-nucleotide RNA oligonucleotides was spiked into all samples prior to RNA isolation to control for loss of RNA during isolation and labeling. piRNAs (red arrow) are absent in the Piwi-YK complex. (B) GFP-Piwi-YK is present in the nuclei of nurse cells and colocalizes with chromatin (DAPI-stained areas). (C) The Piwi-YK mutant does not rescue the morphological changes caused by the piwi-null mutation. Dark-field images of ovaries where either the wild-type *piwi* or the *piwi-YK* transgene has been backcrossed onto the piwi-null background.

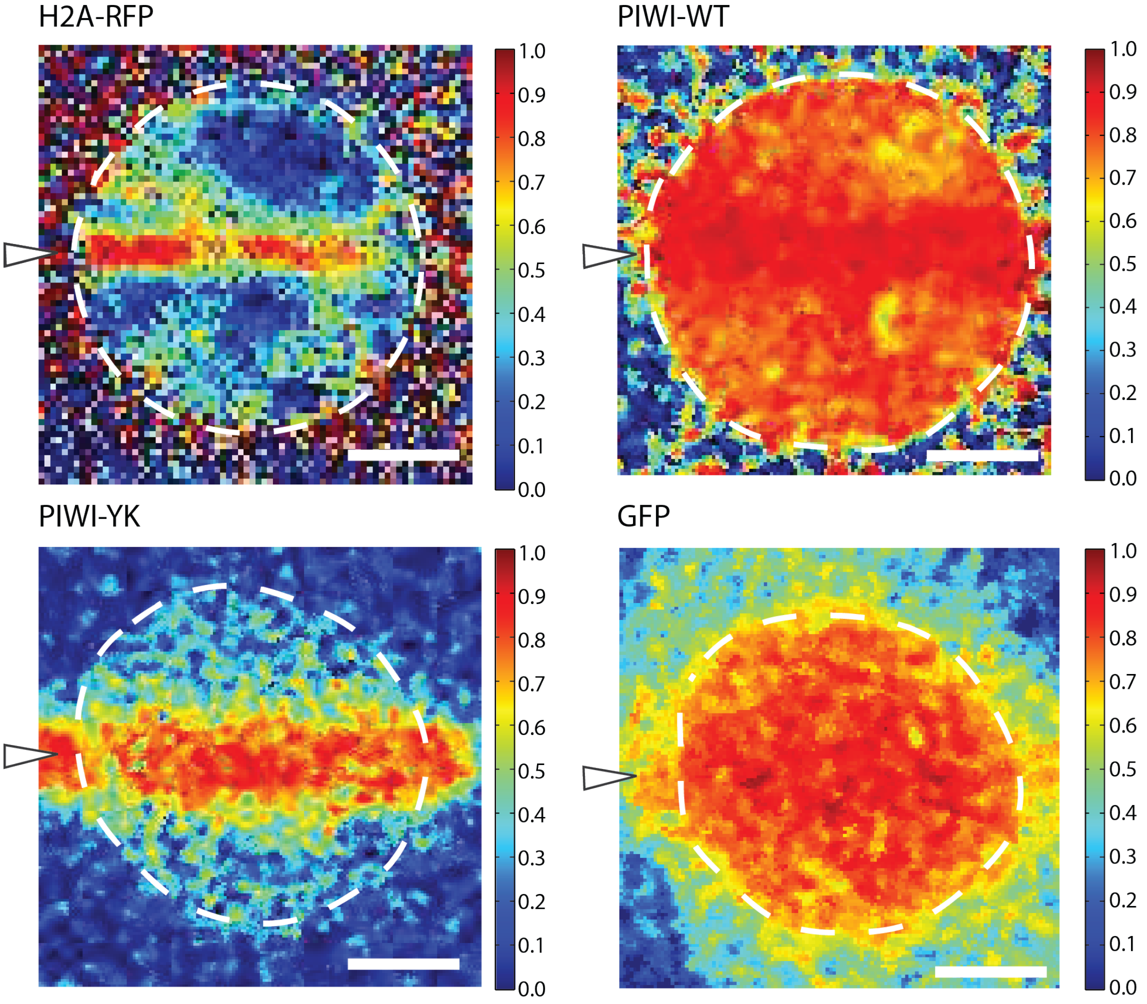
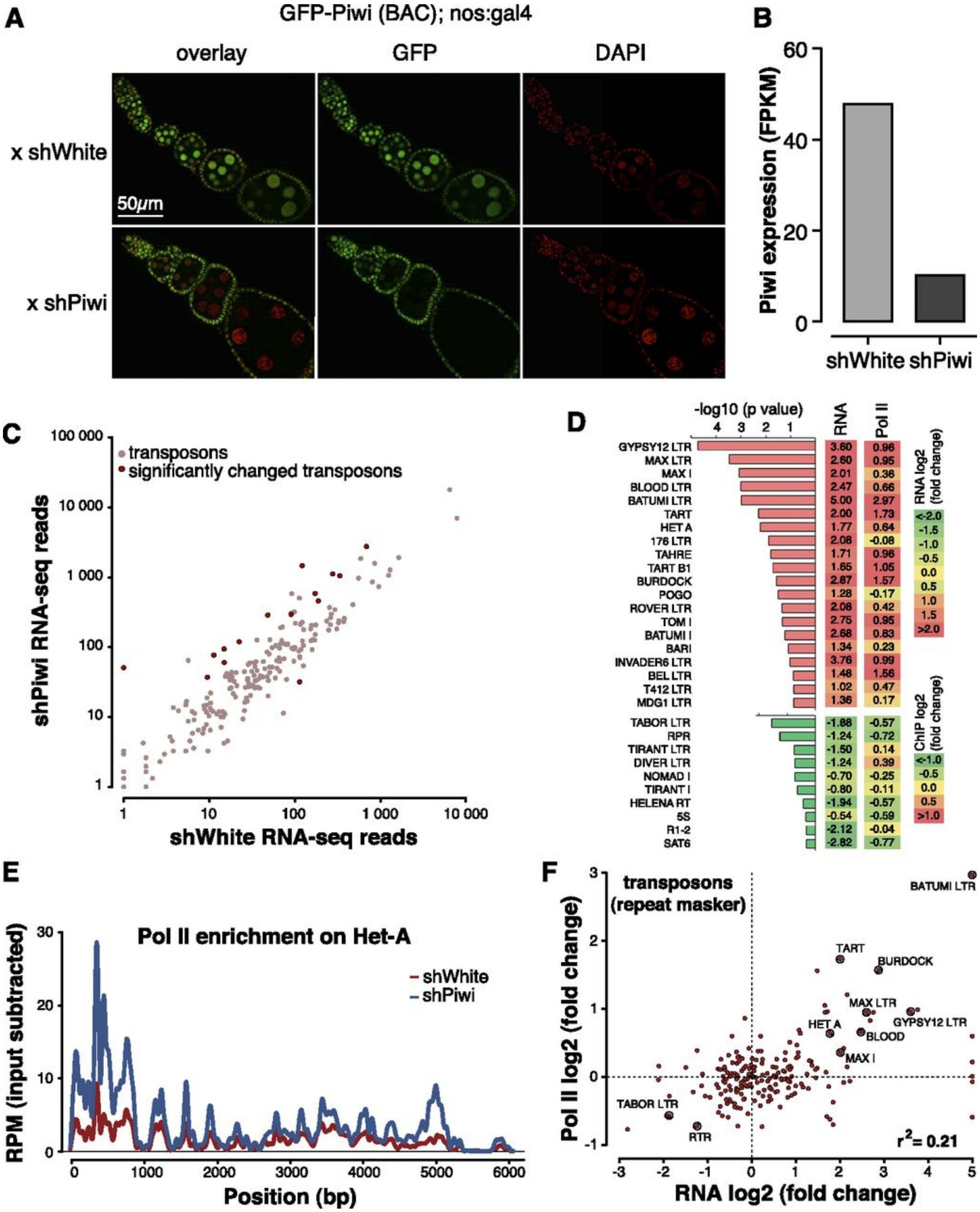


Figure 12.3: Fluorescence Loss in Photobleaching (FLIP) experiments indicate fast redistribution of most of nuclear Piwi and slower movement of the Piwi-YK mutant. Amount of fluorescence decrease after 110 bleaching iterations for H2A-RFP, GFP-Piwi and GFP-Piwi-YK mutant and GFP in a nurse cell nucleus is shown. In each case significant fluorescence loss (red pixels) is observed along the bleach axis. Both GFP and WT GFP-Piwi has extensive loss of fluorescence ($\geq 75\%$) across much of the nucleus, except for specific loci. GFP-Piwi-YK mutant exhibits far less change ($\leq 40\%$) in regions far from the site of bleaching. H2A-RFP control undergoes very little change in intensity away from the bleach region. Note that the apparent slower redistribution of free GFP is likely due to simultaneous nuclear import from the unbleached cytoplasmic pool. Bars = 5μm. Arrowheads indicate position of bleach stripe across the nucleus.

coimmunostaining showed that a GFP-Piwi signal on polytene chromosomes generally overlaps with the RNA polymerase II (Pol II) signal, which marks sites of active transcription (Figure 12.1A).

In order to identify factors that might be responsible for targeting Piwi to chromatin, we immunoprecipitated Piwi complexes from the *Drosophila* ovary and analyzed Piwi interaction partners by mass spectrometry. We purified Piwi complexes from ovaries of three different transgenic lines expressing GFP-Piwi, myc-Piwi, or Flag-Piwi using antibodies against each respective tag. As a control, we used flies expressing free GFP in the ovary. We identified ¿50 factors that showed significant enrichment in all three Piwi purifications but were absent in the control. We were unable to identify chromatin-associated


A
GFP-Piwi (BAC); nos:gal4
overlay
GFP
DAPI
x shWhite
x shPiwi
50μm
B
Piwi expression (FPKM)
60
40
20
0
shWhite
shPiwi
C
transposons
significantly changed transposons
shPiwi RNA-seq reads
shWhite RNA-seq reads
100 000
10 000
1 000
100
10
1
1
10
100
1 000
10 000
D
-log10 (p value)
4 3 2 1
RNA
Pol II
GYPSY12 LTR 3.60 0.96
MAX LTR 2.60 0.95
MAX I 2.01 0.36
BLOOD LTR 2.47 0.66
BATUMI LTR 5.00 2.97
TART 2.00 1.73
HET A 1.77 0.64
176 LTR 2.08 -0.08
TAHRE 1.71 0.96
TART B1 1.65 1.05
BURDOCK 2.87 1.57
POGO 1.28 -0.17
ROVER LTR 2.08 0.42
TOM I 2.75 0.95
BATUMI I 2.68 0.83
BARI 1.34 0.23
INVADER6 LTR 3.76 0.99
BEL LTR 1.48 1.56
T412 LTR 1.02 0.47
MDG1 LTR 1.36 0.17
TABOR LTR -1.88 -0.57
RPR -1.24 -0.72
TIRANT LTR -1.50 0.14
DIVER LTR -1.24 0.39
NOMAD I -0.70 -0.25
TIRANT I -0.80 -0.11
HELENA RT -1.94 -0.57
5S -0.54 -0.59
R1-2 -2.12 -0.04
SAT6 -2.82 -0.77
RNA log2 (fold change)
<-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
>2.0
ChIP log2 (fold change)
<-1.0
-0.5
0.0
0.5
>1.0
E
Pol II enrichment on Het-A
RPM (input subtracted)
30
20
10
0
shWhite
shPiwi
Position (bp)
0
1000
2000
3000
4000
5000
6000
F
transposons (repeat masker)
Pol II log2 (fold change)
RNA log2 (fold change)
3
2
1
0
-1
-3
-2
-1
0
1
2
3
4
5
BATUMI LTR
TART
BURDOCK
MAX LTR
GYPSY12 LTR
HET A
BLOOD
MAX I
TABOR LTR
RTR
r^2= 0.21

factors that directly associate with Piwi but identified several RNA-binding proteins that associate with nascent transcripts, such as splicing (Rm62, Pep, Ref1, Yps, CG9684, CG31368, CG5728, and Mago) and nuclear export (Tho2 and Hpr1) factors (Figure 12.1B). Upon RNase A treatment prior to immunoprecipitation, the presence of most of these RNA-binding proteins in purified Piwi complexes was eliminated.

Piwi proteins are believed to find their targets through sequence complementarity of the associated piRNA. In fact, it has been proposed that lack of the associated piRNA leads to destabilization of piwi proteins and to Piwi's inability to localize to the nucleus (Saito et al. 2009; Haase et al. 2010; Olivieri et al. 2010; Handler et al. 2011; Ishizu et al. 2011). On the other hand, Piwi has been proposed to have functions that are independent of its role in transposon control by regulating stem cell niche development (Cox et al. 1998; Klenov et al. 2011). To address the role of piRNA in translocation of Piwi into the nucleus and its function, we generated transgenic flies expressing a point mutant Piwireferenced as Piwi-YKthat is deficient in piRNA binding due to a substitution of two conserved amino acid residues (Y551L and K555E) in the 5 phosphate-binding pocket (Kiriakidou et al. 2007; Djuranovic et al. 2010). The Piwi-YK mutant was expressed in *Drosophila* follicular and germ cells at levels similar to that of wild-type Piwi but was completely devoid of associated piRNA (Figure 12.2A). In contrast to wild-type Piwi, Piwi-YK could be found in the cytoplasm, supporting the existence of a quality control mechanism that prevents entrance of unloaded Piwi into the nucleus (Ishizu et al. 2011). Nevertheless, a significant amount of piRNA-deficient Piwi localized to the nucleus (Figure 12.2B). Similar to wild-type Piwi, Piwi-YK seemed to associate with chromatin, as indicated by its localization in DAPI-stained regions of the nuclei, and this is consistent with fluorescence loss in photobleaching (FLIP) experiments that demonstrated reduced nuclear mobility compared with free diffusion (Figure 12.3). Based on sterility and ovarian morphology, the *piwi-YK* transgene was unable to rescue the piwi-null phenotype despite its nuclear localization (Figure 12.2C), indicating that while piRNA binding is not absolutely essential for stability and nuclear localization of Piwi, it is required for Piwi function.

To directly test the function of Piwi in the nucleus, we analyzed the effect of Piwi deficiency on gene expression and chromatin state on a genome-wide scale. Piwi mutant females have atrophic ovaries caused by Piwi deficiency in somatic follicular cells (Lin and Spradling 1997; Cox et al. 1998), which precludes analysis of Piwi function in null mutants. Instead, we used RNAi knockdown to deplete Piwi in germ cells while leaving it functionally intact in somatic follicular cells. The Piwi knockdown flies did not exhibit gross morphological defects in the ovary; however, they showed drastic reduction in GFP-Piwi expression in germ cells and were sterile (Figure 12.4A and B). To analyze the effect of Piwi deficiency on the steady-state transcrip-

Figure 12.4 *(preceding page)*: Piwi transcriptionally represses TEs. (A) Piwi knockdown is efficient and specific to ovarian germ cells as indicated by GFP-Piwi localization. GFP-Piwi; Nanos-Gal4-VP16 flies were crossed to control shRNA (shWhite) or shPiwi lines. Piwi is specifically depleted in germ cells and not in follicular cells, consistent with expression of the Nanos-Gal4-VP16 driver. (B) Piwi expression as measured by RNA-seq in the Piwi knockdown and control lines. Note that Piwi expression is unaffected in follicular cells, leading to relatively weak apparent knockdown in RNA-seq libraries from whole ovaries. (C) Effect of Piwi knockdown on the expression of TEs. Two biological replicate RNA-seq experiments were carried out, and differential expression was assessed using DESeq. Transposons that show significant change ($p < 0.05$) are indicated by dark-red circles. Out of 217 individual RepeatMasker-annotated TEs, 15 show a significant increase in expression upon Piwi knockdown. (D) The change in the levels of TE transcripts and Pol II occupancy on their promoters upon Piwi knockdown. Twenty up-regulated and 10 down-regulated transposons with the most significant changes in expression level are shown. Note the low statistical significance for down-regulated transposons. For a complete list of transposons, see Supplemental Figure S2. (E) Pol II signal over the Het-A retrotransposon in control flies (shWhite; red) and upon Piwi knockdown (shPiwi; blue). (F) Increased abundance of transposon transcripts upon Piwi depletion correlates with increased Pol II occupancy over their promoters ($r^2 = 0.21$). Note that the majority of elements do not show significant change in either RNA abundance or Pol II occupancy.

RNA log2(Fold Change)

<-2.00	-1.50	-1.00	-0.50	0.00	0.50	1.00	1.50	>2.00

ChIP log2(Fold Change)

<-1.00	-0.50	0.00	0.50	>1.00

	-log10 (p)	RNA	Pol II
Gypsy12_LTR	4.75	3.6	0.96
MAX_LTR	3.48	2.6	0.95
MAX_I	3.067	2.01	0.36
BLOOD_LTR	3.002	2.47	0.66
BATUMI_LTR	2.989	5	2.97
TART	2.307	2	1.73
HETA	2.235	1.77	0.64
DM176_LTR	1.904	2.08	-0.08
TAHRE	1.813	1.71	0.96
TART_B1	1.714	1.65	1.05
BURDOCK_LTR	1.562	2.87	1.57
POGO	1.505	1.28	-0.17
ROVER-LTR_DM	1.351	2.08	0.42
TOM_I	1.342	2.75	0.95
BATUMI_I	1.23	2.68	0.83
BARI_DM	1.097	1.34	0.23
Invader6_LTR	1.042	3.76	0.99
BEL_LTR	0.909	1.48	1.56
T412LTR	0.884	1.02	0.47
MDG1_LTR	0.879	1.36	0.17
GTWIN_LTR	0.819	1.01	0.28
BLOOD_I	0.756	0.89	0.14
S_DM	0.73	0.9	0.01
G_DM	0.699	0.86	0.11
Gypsy12A_LTR	0.663	2.16	0.15
DOC5_DM	0.641	1.3	0.24
DM412B_LTR	0.636	1.05	0.3
Copia_LTR	0.585	1.66	0.95
ACCORD_LTR	0.578	2.16	1.21
TC1-2_DM	0.564	0.86	-0.07
DM412	0.556	0.54	0.01
Gypsy6A_LTR	0.543	1.07	0.69
DMLTR5	0.519	0.86	0.16
Gypsy4_LTR	0.512	1.68	0.07
TRANSIB1	0.44	1.84	-0.37
Copia_I	0.434	1.48	0.13
POGON1	0.43	0.81	-0.03
S2_DM	0.419	0.84	-0.04
IDEFIX_LTR	0.398	1.44	0.11
Gypsy_LTR	0.38	0.82	0.14
NINJA_LTR	0.377	1.74	-0.14
Invader4_LTR	0.376	1.05	0.36
G6_DM	0.367	0.56	0.26
ACCORD_I	0.358	1.1	0.64
QUASIMODO_LTR	0.356	0.74	0.08
BURDOCK_I	0.343	1.45	-0.01
ROVER-I_DM	0.338	1.12	-0.1
Copia2_LTR_DM	0.333	1.84	-0.7
DMTOM1_LTR	0.304	0.55	0
Gypsy8_I	0.303	0.41	-0.15
DOC6_DM	0.301	1.25	0.04
Invader1_LTR	0.3	1.77	-0.62
G4_DM	0.296	0.99	-0.13
NTS_DM	0.296	0.81	-0.3
I_DM	0.293	0.54	-0.11
Gypsy10_LTR	0.283	0.43	-0.33
BLASTOPIA_LTR	0.281	1.16	-0.15
Copia2_I	0.277	1.7	-0.58
Stalker3_LTR	0.266	0.84	-0.03
Stalker2_LTR	0.256	0.64	-0.44
XDMR_DM	0.249	0.37	-0.36
Gypsy8_LTR	0.24	0.54	-0.22
TRANSIB3	0.238	0.58	-0.21
Mariner2_DM	0.236	0.36	-0.06
MICROPIA_LTR	0.232	0.75	0.25
Invader2_I	0.22	0.82	-0.06
Stalker2_I	0.209	0.83	-0.16
TRANSPAC_I	0.201	0.9	-0.27
MDG3_I	0.194	0.71	0.09
TC1_DM	0.192	0.68	-0.04
FROGGER_LTR	0.192	5	0.22
BARI1	0.179	0.6	0.06
BEL_I	0.172	0.86	0.51
Jockey2	0.166	0.28	-0.19
QUASIMODO_I	0.159	0.55	0.14
MINOS	0.155	0.72	-0.03
HOBO	0.154	0.25	-0.28
DM1731_LTR	0.146	0.76	0.36
DM176_I	0.14	0.44	-0.18
DIVER_I	0.14	0.5	0.02
DM297_I	0.139	0.4	-0.02
Gypsy7_LTR	0.132	5	0.6
ACCORD2_I	0.123	0.24	0.03
DNAREP1_DM	0.123	0.22	-0.05
HETRP_DM	0.122	0.14	0.38
NOMAD_LTR	0.12	0.19	-0.42
IDEFIX_I	0.11	0.47	0.15
PROTOP_B	0.104	0.24	0.26
Gypsy9_I	0.104	0.52	0
PROTOP_A	0.096	0.23	0.2
Invader2_LTR	0.086	0.4	-0.16
TRANSIB4	0.084	0.66	-0.03
DOC3_DM	0.08	0.06	0.15
MARINA	0.08	0.12	-0.08
LOOPER1_DM	0.072	0.06	-0.35
TV1I	0.072	0.21	0.38
Gypsy3_LTR	0.068	0.35	0.05
Gypsy3_I	0.059	0.06	0
FB4_DM	0.059	0.15	-0.23
DMCR1A	0.05	0.01	0
PROTOP	0.049	0.01	0.27
Invader4_I	0.043	0.05	0.09
Invader5_LTR	0.038	0.97	-0.24
ZAM_LTR	0.034	0.4	-0.28
STALKER4_I	0.025	0	0.02
HMSBEAGLE_I	0.017	0.09	-0.01
TRANSPAC_LTR	0.013	0.1	0.25
Gypsy6_LTR	0.013	0.03	0.14
OSVALDO_LTR	0	5	-0.59
TLD2	0	5	-0.07
UHU	0	5	NA
G7_DM	0	0.23	-0.73
MDG3_LTR	0	0.2	0.6
TABOR_LTR	1.546	-1.88	-0.57
DMRPR	1.264	-1.24	-0.72
TIRANT_LTR	0.77	1.5	0.14
DIVER_LTR	0.734	-1.24	0.39
NOMAD_I	0.731	0.7	-0.25
TIRANT_I	0.629	0.8	-0.11
HELENA_RT	0.443	-1.94	-0.57
R1-2_DM	0.341	-2.12	-0.04
DMSAT6	0.339	-2.82	-0.77
LINEJ1_DM	0.223	-0.51	-0.33
Gypsy11_I	0.211	-0.22	0.27
Gypsy5_LTR	0.201	-1.12	0.04
GTWIN_I	0.196	-1.33	-0.22
XDMR	0.194	-1.06	-0.28
DIVER2_LTR	0.194	-2.11	0.16
ROO_I	0.181	-0.17	-0.32
Gypsy7_I	0.165	-1.14	-0.13
BS3_DM	0.152	-0.77	0.12
SAR_DM	0.143	-0.76	-0.52
TRANSIB2	0.142	-0.48	-0.19
MDG1_I	0.14	-0.23	-0.16
ROOA_LTR	0.137	-0.73	0.11
FW_DM	0.129	-0.55	0.06
G2_DM	0.124	-0.11	-0.14
ARS406_DM	0.112	0.72	-0.43
DMRT1C	0.111	-0.58	-0.33
ROO_LTR	0.11	-0.33	-0.45
TABOR_I	0.107	-0.19	-0.33
BS2	0.093	-0.65	0.01
MICROPIA_I	0.089	-0.58	-0.03
BS	0.088	-0.26	-0.46
DM297_LTR	0.088	-0.02	0.04
DMRT1A	0.084	-0.64	-0.06
R2B_DM	0.079	-0.45	0.6
Gypsy5_I	0.073	-0.53	0.32
ZAM_I	0.072	-0.56	-0.39
PENELOPE	0.067	-0.95	-0.3
Ulysses_I	0.065	-0.76	-0.16
R2_DM	0.064	-0.39	-0.44
Gypsy4_I	0.063	-0.46	0.05
Baggins1	0.058	-0.71	0.11
IVK_DM	0.054	-0.35	-0.22
Gypsy6_I	0.053	-0.34	0.2
DOC	0.052	-0.23	0.09
Invader6_I	0.048	-0.39	0.05
Gypsy2_LTR	0.048	-0.7	0.26
Transib5	0.046	-0.6	-0.2
ROOA_I	0.045	-0.35	-0.17
Gypsy9_LTR	0.045	-0.18	-0.05
Gypsy2_I	0.043	-0.01	0.01
Gypsy_I	0.04	-0.33	0.03
G5A_DM	0.038	-0.35	-0.18
CIRCE	0.036	-0.61	-0.08
DMRP1	0.035	-0.39	-0.39
G3_DM	0.035	-0.29	0.3
Gypsy12_I	0.035	-0.07	0.42
M4DM	0.027	-0.1	-0.14
OSVALDO_I	0.027	-0.49	0.01
DMRT1B	0.027	-0.32	-0.19
Invader1_I	0.026	-0.33	-0.01
DIVER2_I	0.026	-0.14	-0.12
STALKER4_LTR	0.019	-0.61	0.11
R1_DM	0.017	-0.02	0.18
NINJA_I	0.016	-0.06	0.01
BLASTOPIA_I	0.015	-0.09	-0.54
DOC4_DM	0.015	-0.07	0.04
ACCORD2_LTR	0.013	-0.44	0.21
BS4_DM	0.013	-Inf	-0.8
HELENA	0.013	-Inf	-0.26
HSATII	0.013	-Inf	-1.54
Invader3_I	0.01	-0.06	0.12
FW3_DM	0.005	-0.51	-0.13
Invader3_LTR	0.003	-0.29	0.23
DOC2_DM	0.002	-0.24	-0.12
G5_DM	0.002	-0.18	0.13
Gypsy10_I	0.001	-0.24	-0.08
PLACW_DM	3E-04	-0.14	0.16
Transib-N1_DM	0	-0.05	-0.11
FROGGER_I	0	-0.1	-0.03
DM1731_I	0	-0.15	-0.03
FW2_DM	0	-0.24	-0.25
Gypsy11_LTR	0	-0.32	0.29
Invader5_I	0	-0.48	-0.36
RSP	0	-0.62	0.15
ALA_DM	NA	NA	-0.44
Bilbo	NA	NA	NA
DMHMR2	NA	NA	0.09
FTZ_DM	NA	NA	0.12
FUSHI_DM	NA	NA	NA
Helitron1_DM	NA	NA	0.33
TLD1	NA	NA	NA
TLD3	NA	NA	0.12
TRAM_I	NA	NA	-0.97
TRAM_LTR	NA	NA	NA

tome as well as the transcription machinery, we performed RNA sequencing (RNA-seq) and Pol II chromatin immunoprecipitation (ChIP) combined with deep sequencing (ChIP-seq) experiments from Piwi knockdown and control flies.

In agreement with previous observations that implicated Piwi in transposon repression (Saito et al. 2006; Aravin et al. 2007; Brennecke et al. 2007), we found that steady-state transcript levels of several TEs were increased upon Piwi knockdown in germ cells (Figure 12.4C and D; Figure 12.5). We found little to no change of RNA levels for transposons whose activity is restricted to follicular cells of the ovary, indicating that the observed changes are indeed due to loss of Piwi in the germline (Figure 12.5). The analysis of Pol II ChIP-seq showed that Pol II occupancy increased over promoters of multiple TEs (Figure 12.4DF; Figure 12.6). Indeed, the change in steady-state levels of transposon transcripts upon Piwi depletion correlated with changes of Pol II occupancy (Figure 12.4F). This result demonstrates that Piwi ensures low levels of transposon transcripts through a repressive effect on the transcription machinery.

To test whether Piwi-mediated transcriptional repression is accompanied by a corresponding change in chromatin state, we used ChIP-seq to analyze the genome-wide distribution of the repressive H3K9me3 mark in the ovary upon Piwi knockdown. We identified 705 genomic loci at which the level of H3K9me3 significantly decreased. More than 90% of the regions that show a decrease in the H3K9me3 mark upon Piwi depletion overlapped TE sequences, compared with the 33% that is expected from random genome sampling (Figure 12.7A). Furthermore, these regions tend to be located in the heterochromatic portions of the genome that are not assembled on the main chromosomes (Figure 12.7B). Only 20 of the identified regions localized to the euchromatic parts of the genome. Of these, 15 (75%) contained potentially active annotated copies of transposons. Taken together, our results indicate that Piwi is required for installment of repressive H3K9me3 chromatin marks on TE sequences of the genome.

While the vast majority of protein-coding host genes did not show significant changes in transcript level or Pol II occupancy upon Piwi knockdown, the expression of a small set of protein-coding genes (150 genes with a p-value < 0.05) was significantly increased (Figure 12.8A; Table Figure 12). There are several possible explanations for Piwi's effect on host gene expression. First, failure in the piRNA pathway might cause up-regulation of several genes that generate piRNAs in wild-type ovaries (Robine et al. 2009; Saito et al. 2009). However, the genes up-regulated in Piwi-deficient ovaries were not enriched in piRNAs compared with other genes. Second, H3K9me3 marks installed on TE sequences in a Piwi-dependent manner might spread into neighboring host genes and repress their transcription, as was recently demonstrated in a follicular cell culture model (Sienski et al. 2012). To address this possibility, we analyzed genomic positions of the genes whose expression was increased upon Piwi knockdown relative to genomic regions that showed a decrease in H3K9me3 marks. We found that up-regulated genes did not show a significant change in the H3K9me3 mark (Figure 12.8B; Figure 12.9). Furthermore, the few genes located close to the regions that show a decrease in H3K9me3 signal had unaltered expression levels upon Piwi knockdown. Next, we analyzed the functions of up-regulated genes using gene ontology (GO) term classifications and found significant enrichment for proteins involved in protein turnover and stress and DNA damage response pathways (Figure 12.8C). Particularly, we found that 31 subunits of the proteasome complex were overexpressed. Therefore, our analysis indicates that up-regulation of specific host genes is likely a secondary response to elevated transposon levels and genomic damage.

In contrast to host genes, transcripts of TEs

Figure 12.5 *(preceding page)*: Piwi regulates transposon levels through transcriptional repression. The change in the levels of transposable element transcripts and RNA Polymerase II occupancy upon Piwi knockdown is shown. RNA-seq and ChIP-seq experiments were carried out in shWhite and shPiwi ovaries in two replicates. Differential expression was assessed using DESeq (see methods). The first column shows the statistical significance of the observed expression change (in log_{10}(p-value)); upregulated and downregulated genes are sorted separately in order of decreasing significance. The second column shows the average change in RNA levels as defined by DESeq. The third column shows the average change in Pol II occupancy between the two replicate experiments.

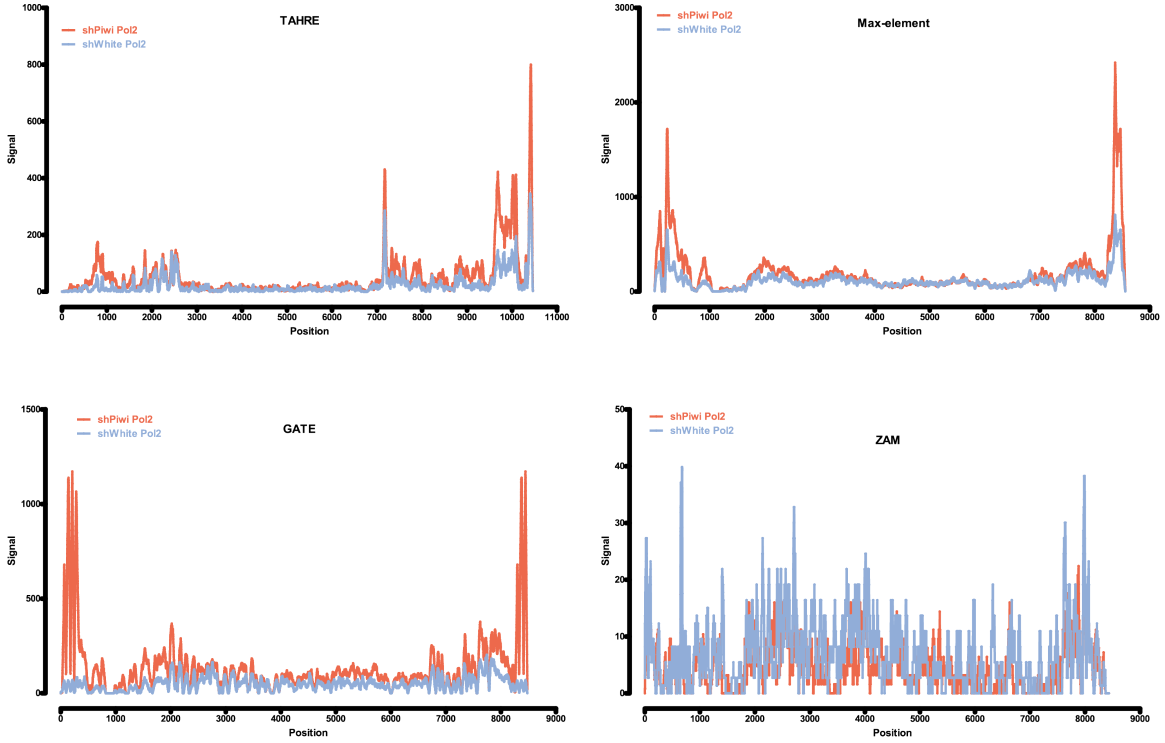


Figure 12.6: Piwi depletion increases RNA Pol II association with promoters of transposable elements. RNA Polymerase II ChIP-seq signal over the consensus sequences of selected transposable elements in the control (shWhite) and Piwi-depleted (shPiwi) ovaries. Pol II occupancy increases in the promoter regions (LTRs) of transposons upon germline knockdown of Piwi. Transposons expressed in somatic follicular cells such as ZAM are not affected.

are targeted by piRNA. To directly address the role of piRNA in Piwi-mediated transcriptional silencing, we took advantage of a fly strain that expresses artificial piRNAs against the *lacZ* gene, which are loaded into Piwi complexes and are able to repress *lacZ* reporter expression in germ cells (Figure 12.10A; Josse et al. 2007; Muerdter et al. 2012). Expression of piRNAs that are antisense to the reporter gene caused transcriptional silencing of the *lacZ* gene as measured by Pol II occupancy (Figure 12.10B). Furthermore, we found that piRNA-induced silencing of the reporter gene was associated with an increase in the repressive H3K9me3 mark and HP1 occupancy and a decrease in the abundance of the active H3K4me2/3 marks at the reporter locus (Figure 12.10C). This result is in good agreement with the genome-wide effect of Piwi depletion on distribution of the H3K9me3 mark and suggests that transcriptional silencing correlates with the establishment of a repressive chromatin structure and is mediated by piRNAs that match the target locus.

12.3 Discussion

Little is known about the function of nuclear piwi proteins. The nuclear piwi in mice (Miwi2) affects DNA methylation of TEs (Carmell et al. 2007; Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008). Several recent reports implicate *Drosophila* Piwi in regulation of chromatin marks on transposon sequences (Lin and Yin 2008; Klenov et al. 2011; Wang and Elgin 2011; Sienski et al. 2012). The mechanism of these processes is unknown in both organisms. Previously, Piwi was shown to associate with polytene chromosomes in salivary gland cells and colocalize with HP1, a chromodomain protein that binds to heterochromatin and a few loci in euchromatin, suggesting that HP1 mediates Piwi's interaction with chromatin (Brower-Toland et al. 2007). However, recent results showed that the putative HP1-binding site on Piwi is dispensable for Piwi-mediated transposon silencing (Wang and Elgin 2011).

We did not detect Piwi expression outside of the ovary and testis, including in salivary gland cells, using a GFP-Piwi transgene expressed under native regulatory elements. We detected GFP-Piwi on polytene chromosomes in ovarian nurse cells that have a germline origin; however, it localizes in a pattern that largely does not overlap with HP1. FLIP experiments with GFP-Piwi indicated a relatively fast rate of fluorescence redistribution as compared with histone H2A (Figure 12.3), implying a transient interaction of Piwi with chromatin. Our proteomic analysis of Piwi complexes isolated from *Drosophila* ovaries did not identify chromatin-associated factors but revealed several RNA-binding proteins, such as splicing and nuclear export factors that bind nascent RNA transcripts (Fig. 1B). Importantly, the interaction of most of these RNA-binding proteins with Piwi was dependent on RNA, indicating that Piwi associates with nascent transcripts. As Piwi itself lacks DNA- and RNA-binding domains (beyond the piRNA-binding domain), it is likely that the recruitment of Piwi to chromatin is through interactions with other RNA-binding proteins or sequence-specific interactions between Piwi-bound piRNA and nascent transcripts.

Using specific Piwi knockdown in germ cells of the *Drosophila* ovary, we analyzed the effect of Piwi depletion on gene expression, the transcription machinery, and H3K9me3 chromatin marks genome-wide. In agreement with previous results (Klenov et al. 2011), we found up-regulation of several TEs upon Piwi knock-

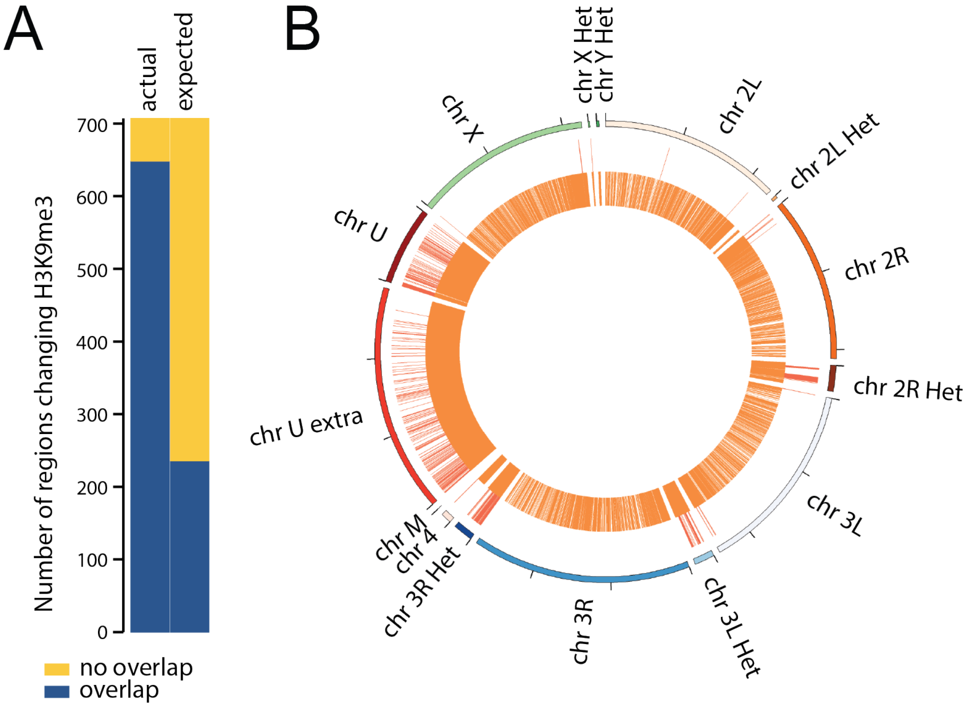


Figure 12.7: Piwi-induced transcriptional repression correlates with establishment of a repressive chromatin state. (A) Overlap between genomic regions of H3K9me3 depletion upon Piwi knockdown and TEs. Two replicates of H3K9me3 ChIP-seq experiments were carried out on control and Piwi-depleted ovaries, and enriched regions were identified using DESeq (see the Materials and Methods for details). A total of 705 regions show significant ($p < 0.05$) decrease in H3K9me3 occupancy upon Piwi knockdown, while only 30 regions showed a similarly significant increase. Out of the 705 regions that show a decrease in H3K9me3 marks upon Piwi knockdown, 91% (646) overlap with TE sequences compared with the 33% expected from random genome sampling. (B) Genomic positions of H3K9me3-depleted regions upon Piwi depletion (outer circle) and RepeatMasker-annotated transposons (inner circle). Note that almost all regions are localized in heterochromatic and repeat-rich portions of the genome (Het, chrU, and chrUExtra chromosomes).

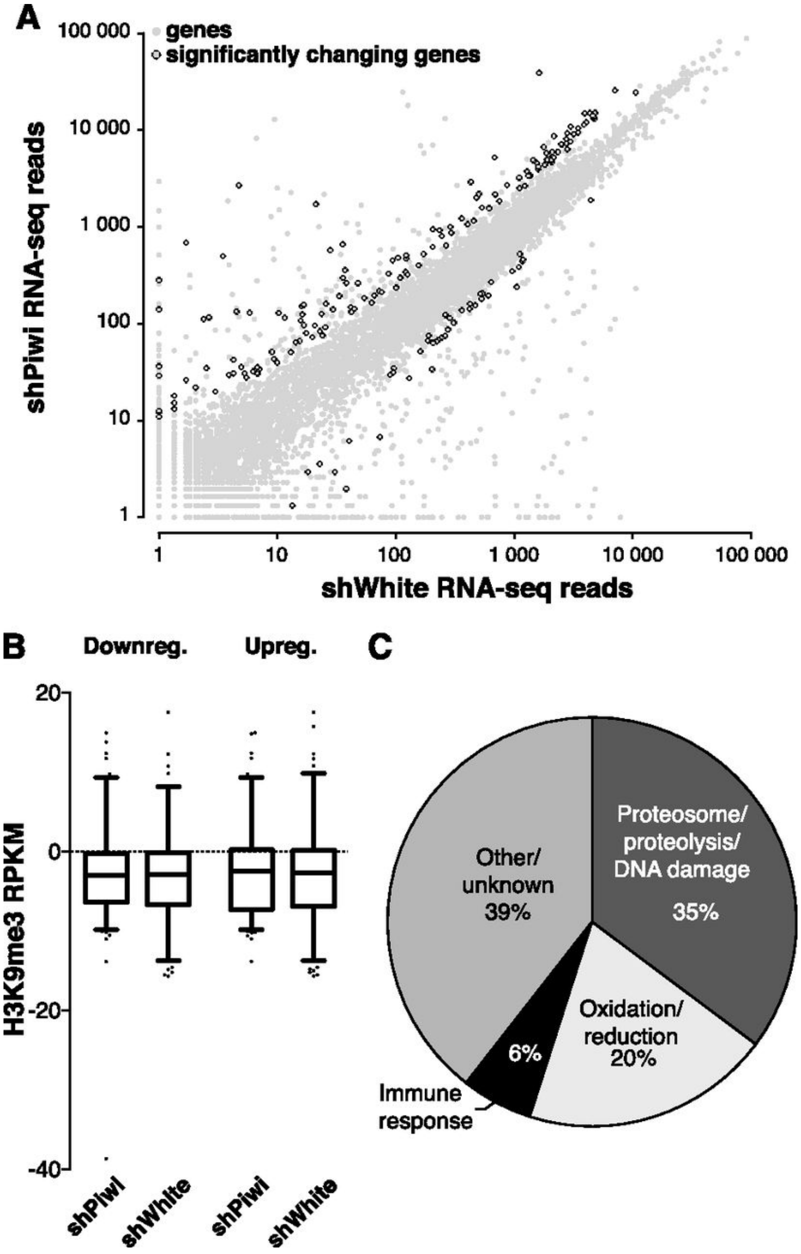


Figure 12.8: Piwi does not directly repress protein-coding genes (A) Effect of Piwi knockdown on the expression of genes. Two replicate RNA-seq experiments were carried out, and differential expression was assessed using DESeq. Genes that show significant change ($p < 0.05$) are indicated by black circles. The vast majority of genes does not change significantly upon germline Piwi knockdown (shPiwi) compared with control (shWhite). (B) H3K9me3 mark density does not change over genes that show a significant change in expression upon Piwi knockdown (see Figure 12.4C). Up-regulated and down-regulated genes are plotted separately. Signal indicated is after background subtraction. (C) Functional analysis of up-regulated genes by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) reveals activation of the protein degradation and DNA damage response pathways. Percentages of all up-regulated genes are indicated.

down (Figure 12.4C). The TEs that did not change their expression upon germline knockdown of Piwi might be expressed exclusively in somatic follicular cells of the ovary, such as the *gypsy* retrotransposon. Alternatively, some elements present in the genome might not have transcriptionally active copies, or the cytoplasmic AUB/AGO3 proteins may efficiently silence them at the post-transcriptional level.

The increase in steady-state levels of RNA upon Piwi depletion strongly correlates with an increase in Pol II occupancy on the promoters of transposons (Figure 12.4D,F; Figure 12.5). This result suggests that Piwi represses transposon expression at the transcriptional level, although we cannot completely exclude the possibility of an additional post-transcriptional effect. It was shown previously that depletion or mutation of Piwi leads to depletion of the repressive H3K9me3 mark and an increase in the active H3K4me2/3 marks on several transposon sequences (Klenov et al. 2011; Wang and Elgin 2011). Our ChIP-seq data extend these results to a genome-wide scale, proving that trans-

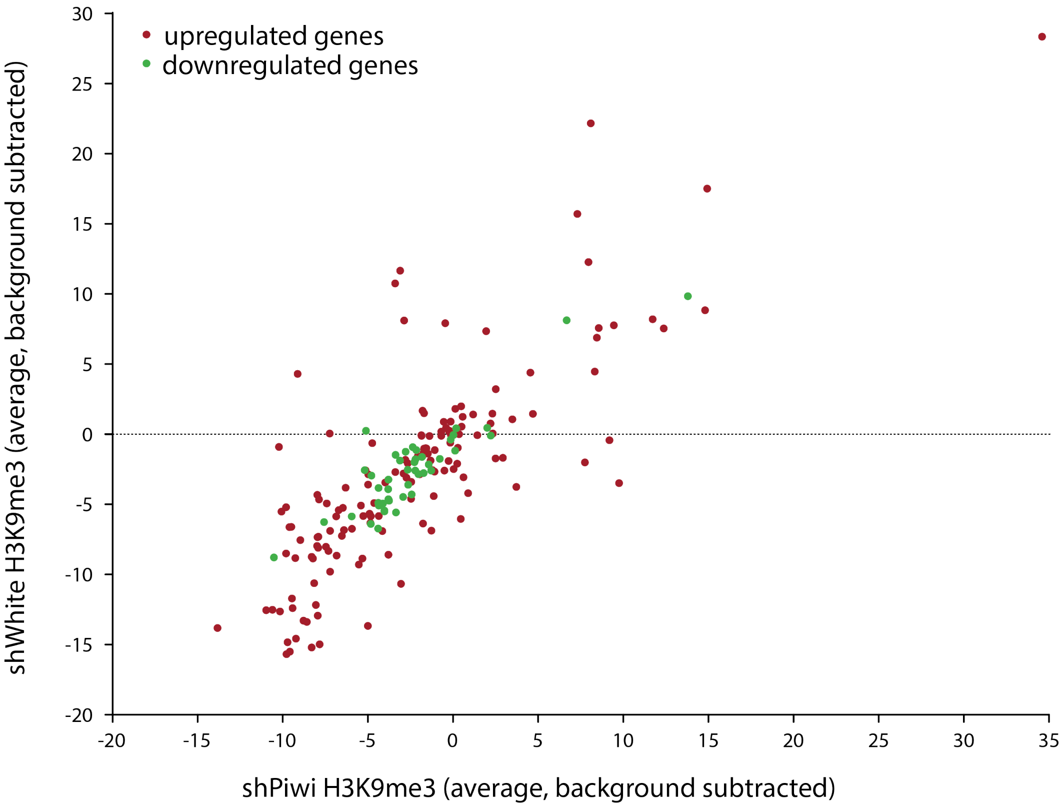


Figure 12.9: Piwi depletion does not alter H3K9me3 occupancy over differentially expressed genes. Scatter plot indicating average H3K9me3 mark levels upon Piwi depletion (shPiwi) and control (shWhite) over genes that were previously identified in the RNA-seq experiments to be differentially expressed upon Piwi knockdown. (red: upregulated genes, green: downregulated genes). The average signal of two biological replicates was taken after subtraction of the corresponding input signals.

posons are indeed the sole targets of Piwi, and demonstrate that changes in histone marks directly correlate with transcriptional repression.

Piwi depletion in the germline does not affect expression of the majority of host genes, although a small fraction of genes changes expression (Figure 12.8A). One possible mechanism of the effect Piwi has on host genes is the spreading of repressive chromatin structure from transposon sequences to adjacent host genes. Indeed, such a spreading and the resulting repression of host gene transcription were observed in an ovarian somatic cell (OSC) culture model (Sienski et al. 2012). However, we did not find significant changes in the H3K9me3 mark for genes that are up-regulated upon germline depletion of Piwi, arguing against this mechanism playing a major role in host gene regulation. Instead, we found that the majority of host genes whose expression is increased as a result of Piwi depletion participate in protein turnover (e.g., proteasome subunits) and stress and DNA damage response pathways, indicating that they might be activated as a secondary response to cellular damage induced by transposon activation. The different effect of Piwi depletion on host gene expression in ovary and cultured cells might be explained by the fact that silencing of host genes due to transposon insertion would likely have a strong negative effect on the fitness of the organism but could be tolerated in cultured cells. Accordingly, new transposon insertions that cause repression of adjacent host genes should be eliminated from the fly population but can be detected in cultured cells. In agreement with this explanation, the majority of cases of repressive chromatin spreading in OSCs were observed for new transposon insertions that are absent in the sequenced *Drosophila* genome. Indeed, it was shown that the vast majority of new transposon insertions is present at a low frequency in the *Drosophila* population, likely due to strong negative selection (Petrov et al. 2003). Such selection was primarily attributed to the ability of TE sequences to cause recombination and genomic rearrangements. We proposed that in addition to the effects on recombination, the selection against transposons can be driven by their negative impact on host gene expression in the germline linked to Piwi-mediated chromatin silencing.

How does Piwi discriminate its proper targetstransposonsfrom host genes? In the case of cytoplasmic Piwi proteins AUB and AGO3,

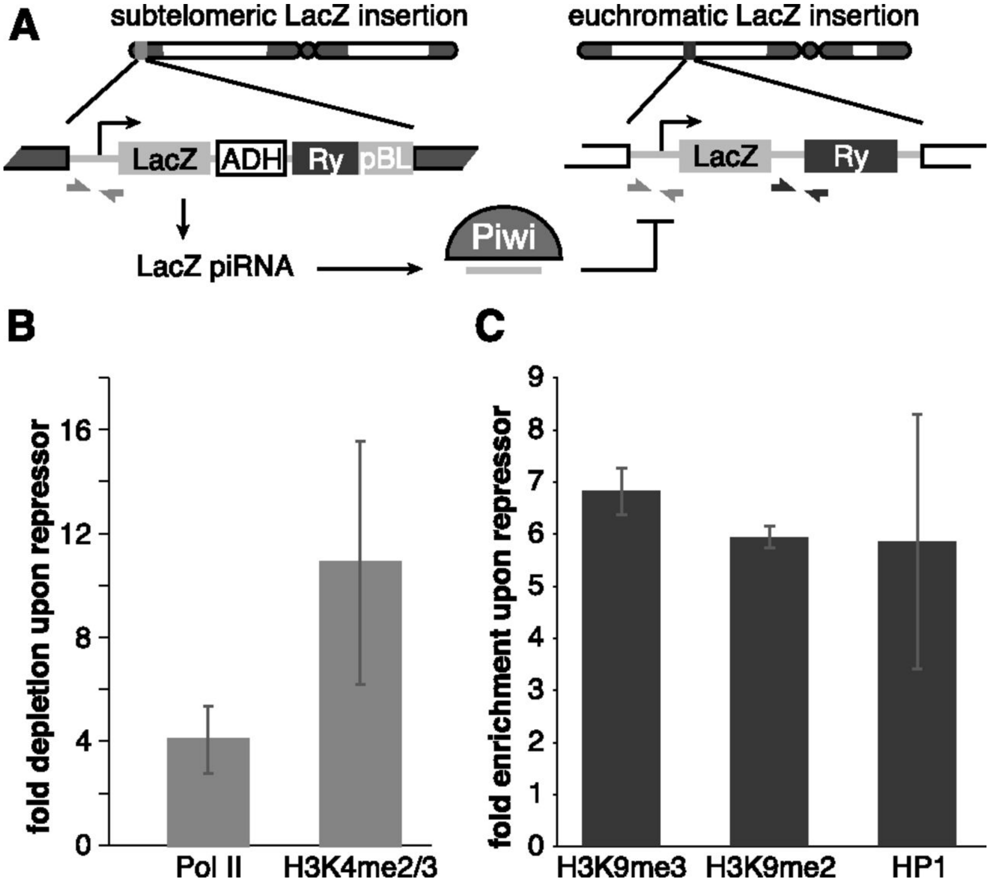


Figure 12.10: piRNA-dependent targeting of Piwi to a reporter locus leads to establishment of a repressive chromatin state and transcriptional silencing. (A) The mechanism of *trans*-silencing mediated by artificial piRNA and a schematic representation of the repressor and reporter *lacZ* constructs. The repressor construct is inserted in a subtelomeric piRNA cluster, leading to generation of piRNA from its sequence. Primers mapping to both constructs used for the Pol II and H3K4me2/3 ChIP-quantitative PCR (qPCR) are shown by light-gray arrows; primers specific to the reporter locus used for the H3K9me3, H3K9me2, and HP1 ChIP-qPCR are indicated by dark-gray arrows. (B) piRNAs induce transcriptional repression of the *lacZ* reporter. Pol II and H3K4me2/3 signals decreased on the *lacZ* promoter in the presence of artificial piRNAs as measured by ChIP-qPCR. Shown is the fold depletion of signal in flies that carry both repressor and reporter constructs compared with control flies that have only the reporter construct. The signal was normalized to RP49. (C) piRNAs induce an increase in H3K9me3 and H3K9me2 marks and HP1 binding as measured by ChIP-qPCR. Shown is the fold increase of corresponding ChIP signals downstream from the *lacZ* reporter in flies that carry both repressor and reporter constructs compared with control flies that have only reporter construct. The signal was normalized to RP49.

recognition and post-transcriptional destruction of TE transcripts is guided by associated piRNAs. Our results indicated that piRNAs provide guidance for transcriptional silencing by the nuclear Piwi protein as well. First, in contrast to host genes that are not targeted by piRNAs, TE transcripts, which are regulated by Piwi, are recognized by antisense Piwi-bound piRNA (Brennecke et al. 2007). Second, a Piwi mutant that is unable to bind piRNA failed to rescue the piwi-null mutation despite its ability to enter the nucleus. Finally, expression of artificial piRNAs that target a reporter locus induced transcriptional silencing associated with an increase in repressive H3K9me3 and HP1 chromatin marks and a decrease in the active H3K4me2/3 marks (Figure 12.10B and C). In contrast, the tethering of Piwi to chromatin in a piRNA-independent fashion by fusing Piwi with the lacI DNA-binding domain that recognizes lacO sequences inserted upstream of a reporter gene did not lead to silencing of the reporter (data not shown). Together, our results demonstrated that piRNAs are the essential guides of Piwi to recognize its targets for transcriptional repression.

It is tempting to propose that, similar to Argonautes in fission yeast, *Drosophila* Piwi directly recruits the enzymatic machinery that establishes the repressive H3K9me3 mark on its targets. Establishment of repressive marks can lead to stable chromatin-based transcriptional silencing that does not require further association of Piwi with target loci. This model explains why we found that Piwi is relatively mobile in the nucleus, indicative of only a transient interaction with chromatin. The Piwi-mediated transcriptional silencing has an interesting parallel in *Caenorhabditis elegans*, where the Piwi protein PRG-1 and associated 21U RNAs are able to induce stable transgenerational repression that correlates with formation of silencing chromatin marks on target loci. Interestingly, PRG-1 and 21U RNAs are necessary only for initial establishment of silencing, while continuing repression depends on siRNA and the WAGO group of Argonautes (Ashe et al. 2012; Bagijn et al. 2012; Buckley et al. 2012; Shirayama et al. 2012). Future studies should reveal the pathway that leads to transcriptional repression downstream from Piwi in *Drosophila* and the differences from and similarities to other species.

12.4 Materials and methods

12.4.1 Drosophila stocks

Nanos-Gal4-VP16 (BL4937), *UASp-shWhite* (BL33623), *UASp-shPiwi* (BL 33724), and Chr. I and II Balancer (BL7197) were purchased from the Bloomington Stock Center. GFP-Piwi-expressing flies (see below) were backcrossed onto the *piwi1/piwi2* (available from Bloomington Stock Center) background or the *otu7/otu11* (available from Bloomington Stock Center) background, respectively. *LacZ* reporter lines were a generous gift from S. Ronsseray.

12.4.2 Generation of transgenic fly lines

The GFP-Piwi, 3xFlag-HA-Piwi, and myc-Piwi constructs were generated using bacterial recombineering (Gene Bridges Counter Selection kit) to insert the respective tag after the start codon of the Piwi genomic region cloned in BAC clone BACN04M10. The KpnIXbaI genomic fragment that contains the Piwi gene and flanking sequences was transferred to corresponding sites of the pCasper4 vector to create pCasper4/tagged Piwi.

The pCasper4/GFP-Piwi construct was used to generate pCasper4/GFP-Piwi-YK with two point mutations, Y551I and K555E. Mutations were introduced by PCR, amplifying products corresponding to a 3.1-kb upstream fragment and a 2.58-kb downstream fragment. The upstream fragment included a unique XbaI site at the 5 end of the amplicon and overlapped 39 base pairs (bp) with the downstream fragment, which included a unique BamHI site at its 3 end. The single XbaIBamHI fragment was generated by overlap PCR with outside primers and cloned into corresponding sites of pCasper4/GFP-Piwi to replace the wild-type fragment. Transgenic flies were generated by P-element-mediated transformation (BestGene).

12.4.3 Immunoprecipitation of Piwi proteins and RNA gel of piRNA

Dissected ovaries were lysed in lysis buffer (20 mM HEPES at pH 7.0, 150 mM KCl, 2.5 mM MgCl, 0.5% Triton X-100, 0.5% Igepal, 100 U/mL RNasin [Promega], EDTA-free Complete

Protease Inhibitor Cocktail [Roche]) and supernatant clarified by centrifugation. Supernatant was incubated with anti-eGFP polyclonal antibody (Covance) conjugated to Protein-G Dynabeads at 4 °C. Beads were spiked with 5 pmol of synthesized 42-nucleotide RNA oligomer to assess purification efficiency, proteinase K-digested, and phenol-extracted. Isolated RNA was CIP-treated, radiolabeled using PNK and γ-^{32}P-labeled ATP, and run on a 15% urea-PAGE gel. Western blots of ovary lysate and anti-eGFP immunoprecipitates were obtained from 8% SDS-PAGE gels and probed with polyclonal rabbit anti-eGFP antibody to confirm expression of the full-length transgene.

12.4.4 Mass spectrometric analysis of Piwi interaction partners

Lysis and clarification of ovary samples were performed as described above using lysis buffer with reduced detergent (0.1% Triton X-100, 0.1% Igepal). Piwi proteins with Flag, Myc, or GFP tag were purified from *Drosophila* ovaries using corresponding antibodies covalently coupled to M-270 epoxy Dynabeads (Invitrogen) (Cristea et al. 2005). Immunoprecipitation of free GFP from GFP-expressing ovaries was used as a negative control. Immunoprecipitations were performed in the presence or absence of RNase A (100 μg/mL; 30 min at 25 °C). Piwi and copurified interacting proteins were resolved on NuPAGE Novex 4%12% Bis-Tris gels and stained with colloidal Coomassie blue. Gel fragments that contained protein bands were excised and in-gel-trypsinized, and the peptides were extracted following the standard protocol of the Proteome Exploration Laboratory at California Institute of Technology. Peptide analyses were performed on an LTQ-FT Ultra (Thermo Fisher Scientific) equipped with a nanoelectrospray ion source (Thermo Fisher Scientific) connected to an EASY-nLC. Fractionation of peptides was performed on a 15-cm reversed-phase analytical column (75-μm internal diameter) in-house-packed with 3-μm C18 beads (ReproSil-Pur C18-AQ medium; Dr. Maisch GmbH). Acquired spectra were searched against the *Drosophila melanogaster* proteome using the search engine Mascot (Matrix Science, version 2.2.06), and protein inferences were performed using Scaffold (Proteome Software, version 3).

12.4.5 Antibodies

eGFP antibody (rabbit polyclonal serum; Covance) was affinity-purified in the Aravin/Tóth laboratories. Anti-myc (Millipore), anti-Flag (Sigma), Pol II (ab5408), and Pol II pSer5 (ab5131) are commercially available.

12.4.6 Imaging of ovaries

Ovaries were fixed in 4% PFA in PBS for 20 min, permeabilized in 1% Triton X-100 in PBS, DAPI-stained (Sigma-Aldrich), washed, and mounted in 50% glycerol/PBS. Images were captured using an AxioImager microscope; an Apotome structured illumination system was used for optical sections (Carl Zeiss).

12.4.7 FLIP

FLIP time series were captured on an LSM510 confocal microscope equipped with a 40×/0.9 NA Imm Corr multi-immersion objective. Ovaries were dissected into halocarbon 700 oil (Sigma) and mounted under a 0.17-mm coverslip (Carl Zeiss) immediately before imaging. Two initial baseline images were captured, followed by 80100 iterations consisting of two bleach iterations at 100% laser power (488 nm or 543 nm for GFP- and RFP-tagged proteins, respectively), followed by two images with reduced illumination intensity. FLIP series were cropped and median-filtered with a 2-pixel radius to reduce noise using FIJI (Schindelin et al. 2012) and the "Rigid Body" function of the StackReg plugin (Thévenaz et al. 1998) to correct drift when needed. Using Matlab software (The Mathworks), images were background-subtracted and corrected for acquisition bleaching. A value representing the true loss of intensity relative to the initial prebleach images, where 0 indicates no change in intensity and 1 represents complete photobleaching, was calculated for each pixel and each bleach/capture cycle and plotted with a color lookup table and calibration bar. Scale bars and annotations were made in Inkscape (`http://inkscape.org`).

12.4.8 Preparation of polytene squashes for immunofluorescence

Flies carrying the GFP-Piwi BAC construct were backcrossed onto the *otu*[7] and *otu*[11] background. Progeny from the cross of the two lines were grown at 18 °C. Stage 712 egg chambers were separated and transferred to a polylysine-coated microscopic slide into PBST. From here, the "smush" protocol was followed (Johansen et al. 2009), but PFA cross-linking was reduced to 10 min. Slides were imaged using an AxioImager microscope and a 63× oil immersion objective (Carl Zeiss).

12.4.9 ChIP, ChIP-seq, and RNA-seq

ChIP was carried out using standard protocols (Moshkovich and Lei 2010). ChIP-seq and RNA-seq library construction and sequencing were carried out using standard protocols following the general principles described by Johnson et al. (2007) and Mortazavi et al. (2008), respectively. For quantitative PCR (qPCR) primers, see 12.2. GO term analysis of genes upregulated upon Piwi knockdown was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang et al. 2009a,b) and FlyBase for additional assignment of GO terms.

12.4.10 High throughput data analysis

Except for where specifically specified otherwise, all data processing was carried out using custom-written python scripts. The dm3/BDGP assembly, release 5 version of the *Drosophila melanogaster* genome was used.

12.4.11 ChIP-seq and ChIP-seq data processing

Sequencing libraries were sequenced on the Illumina HiSeq 2000 (50bp reads). The resulting sequencing reads were trimmed down to 36bp and mapped against the genome using Bowtie 0.12.7 (Langemad et al. 2009) with the following settings: `''-v 2 --best --strata''` retaining only uniquely mappable reads with up to two mismatches. Read mapping statistics for ChIP-seq datasets processed this way are presented in 12.3.

12.4.12 Gene expression quantification using RNA-seq

RNA-seq libraries were built from polyA-selected RNA from fly ovaries following standard protocols (Mortazavi & Williams et al. 2008) and sequenced on the HiSeq 2000 (50bp reads). For the purposes of expression quantification, reads were mapped as 50mers, using TopHat 1.4.1 (Trapnell et al. 2009) and splice junctions from the ENSEMBL62 dm3 annotation with otherwise default settings. Gene expression was quantified in RPKMs/FPKMs (**R**eads/**F**ragments **P**er **K**ilobase per **M**illion mapped reads/fragments) for the refSeq annotation (downloaded from the UCSC browser) with Cufflinks 2.0.2 (Trapnell et al. 2010; Trapnell et al. 2012). Read mapping statistics for these libraries are presented in 12.4.

12.4.13 Repeat analysis

The usual practice when mapping ChIP-seq data is to retain only unique alignments as the ambiguity of the allocation of multimapper seriously confounds most analyses. In this study it was necessary to examine repeats but not absolutely necessary to properly allocate multimappers to each individual repeat. I therefore adopted the following two strategies for processing our ChIP-seq and RNA-seq data and examining ChIP enrichment over the expression of repeat elements:

12.4.13.1 Analysis on RepeatMasker-annotated repeat elements

Both ChIP-seq and RNA-seq reads were trimmed down to the same length (36bp) and again aligned with Bowtie 0.12.7 against the dm3 genome but this time with the following options: "`-v 0 -a --best --strata -q`", i.e. no mismatches and an unlimited number of locations to which a read can map to. Read mapping statistics for these alignments are presented in 12.5. For each read r, an integer multiplicity score NH_r was defined (corresponding to the number of positions in the genome the read maps to) and for each individual instance of a repeated element RE (as defined in the RepeatMasker repeat element annotation downloaded from UCSC) an RPM score was calculated as follows:

$$RPM_{RE} = \sum_{r \in RE} \frac{1}{NH_r} \qquad (12.1)$$

A combined repeat RPM score was calculated as the sum of the RPMs for each individual instance of that repeat:

$$TotalRPM_{RE} = \sum_{RE} RPM_{RE} \qquad (12.2)$$

For RNA-seq data, repeat expression change was assessed as the RPM ratio between the shPiwi and shWhite libraries. For Pol II ChIP-seq data, an additional confounding factor exists as the differences in signal between two regions is the result of the combination of the actual change in occupancy and the difference in ChIP strength between the two experiments. I therefore used the total Pol II RPMs over transcription start sites in order to assess the difference in ChIP strength and derive a normalization factor to be used for rescaling of the repeat RPMs of one libraries so that they are comparable to those in the other (here, this factor turned out to be very close to 1).

12.4.13.2 Analysis on consensus repeat sequences

An orthogonal strategy for the analysis of repeat occupancy and expression change that we employed was to map reads against consensus repeat sequences (obtained from FlyBase version FB2012_05 (McQuilton et al. 2012)). Reads were mapped with the following settings: "`-v 3 -a --best --strata -q`" (allowing for up to 3 mismatches and unlimited number of multimappers). Read mapping statistics for these alignments can be found in 12.6. Read counts for each repeat were calculated (normalizing for multimapper multiplicity as described above) and normalized for sequencing depth against the total number of reads mappable to the genome (derived from the alignment without limits to read multiplicity discussed in the previous section) and finally, normalized for the length of the consensus sequences (RPKMs).

Results from both analyses were very similar and so only plots for RepeatMasker repetitive elements are shown.

12.4.14 Differential expression and occupancy analysis

In order to identify differentially expressed genes and transposons we used a combination of eXpress quantification (Roberts & Pachter 2013) and DESeq (Anders & Huber 2010) differential read count analysis. For each replicate, RNA-seq reads were aligned against the transcriptome and the quantification values for all transcripts belonging to the same gene were summed to derive gene-level quantifications. The "effective counts" values were used for downstream analysis. As only a minority of reads align to transposons, differential expression analysis only on transposons is not reliable. For this reason, I combined raw read counts for transposons (derived for the RepeatMakser annotation as described above or for the consensus sequences) with the eXpress quantifications on genes and ran DESeq to evaluate the statistical significance of the observed expression changes over the two shWhite and shPiwi replicates.

Differential occupancy of H3K9me3 was estimated as follows. First, the genome was divided into 1000bp bins and the H3K9me3 read count was estimated for each using the alignments generated with unlimited number of locations a read can map (dividing each alignment by the read multiplicity as discussed above). Next, DESeq was run on the H3K9me3 replicates to identify regions enriched or depleted upon Piwi knock down (p-value of 0.05 threshold was applied). Neighboring depleted regions were merged into contiguous clusters.

Pol II occupancy change over transposons was estimated from the combined RPM values for RepeatMasker transposons and from RPKM values for consensus transposons after taking into account that the difference in ChIP signal between two regions is the result of the combination of the actual change in occupancy and the difference in ChIP strength between the two experiments. I therefore used the total Pol II RPMs over TSSs in order to assess the difference in ChIP strength and derive a normalization factor to be used for rescaling of the repeat RPMs of libraries so that they are comparable to those in the other (this factor turned out to be close to 1 for both sets of replicates).

Table 12.1: List of genes significantly upregulated upon Piwi knockdown. Shown are the . DESeq $log_2(FoldChange)$ and p-values as calculated from two biological replica

Gene	$log_2(foldchange)$	p-value
CG14628	Inf	6.77E-10
CG15056	Inf	8.05E-03
CG18823	Inf	2.77E-02
CG31054	Inf	3.28E-12
CG4984	Inf	2.59E-02
Sdic1	Inf	3.80E-02
yellow-c	Inf	9.49E-03
blanks	9.97	7.11E-04
Rpt6R	9.5	3.17E-05
CG32259	7.65	3.91E-02
Rpt3R	6.42	1.73E-07
Oseg5	6.32	8.24E-07
Shawl	6.1	2.02E-09
CG18193	5.63	1.56E-02
CG15201	5.37	4.01E-02
CG12493	5.24	1.87E-02
TrxT	5.19	5.90E-03
Salt	5.14	4.61E-02
CG4650	4.74	2.73E-02
CR18854	4.59	3.91E-11
Rbp4	4.48	1.69E-03
PebIII	4.42	5.82E-03
CG5791	4.33	1.52E-02
CG13321	4.26	2.81E-09
CG3884	3.79	1.47E-02
CG12655	3.68	3.73E-03
CG10151	3.45	6.51E-05
CG5281	3.32	9.62E-05
GstD2	3.32	3.00E-02
CG30108	3.3	3.83E-06
IM1	3.3	1.59E-02
CG10440	3.23	2.18E-02
CG34291	3.2	3.13E-02
CG16758	3.14	1.34E-02
CG6776	3.1	3.14E-05
Cyp12d1-p	3.03	1.37E-03
CG18186	2.94	1.18E-05
Obp99b	2.86	5.56E-04
CG1600	2.82	2.48E-04
CG13936	2.79	4.55E-02
Hsp70Ab	2.77	7.62E-03
CG7470	2.7	2.51E-04
Gfat1	2.65	4.23E-03
CG9960	2.6	2.87E-03
Ptp52F	2.58	1.58E-03
GstD10	2.58	4.32E-02
GstD5	2.57	2.29E-02
Mdr49	2.57	1.13E-02
Lsd-1	2.48	7.31E-04

Continued on next page

Table 12.1 – *Continued from previous page*

Gene	$log_2(FoldChange)$	p-value
scpr-A	2.47	3.65E-03
GstE5	2.45	3.60E-02
Cyp28d1	2.34	1.05E-02
CG7408	2.34	4.42E-02
CG9380	2.3	1.04E-02
CG15347	2.28	2.26E-02
CG14629	2.27	1.03E-02
CG32572	2.26	7.74E-03
CG5399	2.24	4.98E-03
Jheh3	2.2	8.99E-03
CG5171	2.19	3.17E-02
CG9743	2.17	4.68E-02
Hsp23	2.13	8.02E-04
RpS19b	2.1	4.61E-02
Lip4	2.07	6.69E-03
Hsp70Aa	2.06	8.74E-05
IM2	2.05	3.62E-02
Pomp	2	8.31E-04
pncr008	1.99	4.42E-03
CG5853	1.96	1.08E-02
CG9360	1.93	2.94E-02
CG30104	1.93	5.42E-03
CG12290	1.92	2.58E-02
ref(2)P	1.92	1.26E-03
Prosalpha5	1.92	1.56E-03
CR42871	1.91	3.78E-02
Pros28.1	1.86	1.72E-03
Pros35	1.86	5.95E-03
CG6299	1.85	5.75E-03
Prosbeta7	1.8	3.75E-03
CG15445	1.79	5.28E-03
qsm	1.78	1.13E-02
CG11378	1.78	2.50E-02
DnaJ-H	1.76	2.53E-03
CG17331	1.74	4.46E-03
Jheh1	1.73	8.66E-03
dgo	1.7	2.67E-02
IM3	1.69	3.05E-02
CG3348	1.69	4.28E-02
Prosbeta5	1.68	8.07E-03
CG5958	1.67	1.50E-02
Prosbeta1	1.65	6.22E-03
Hmu	1.65	1.08E-02
msd1	1.64	7.74E-03
CG4199	1.64	1.08E-02
cathD	1.63	9.09E-03
CG10208	1.62	1.45E-02
Gel	1.61	1.41E-02
GstE3	1.61	1.75E-02
Prosbeta2	1.6	6.70E-03
sev	1.58	2.74E-02

Continued on next page

Table 12.1 – *Continued from previous page*

Gene	$log_2(FoldChange)$	p-value
Prosalpha7	1.58	7.14E-03
CG5167	1.57	2.87E-02
Lsm10	1.57	1.72E-02
Rpn9	1.57	9.83E-03
Rpn6	1.56	1.13E-02
Rpt1	1.55	8.58E-03
CG2046	1.55	6.56E-03
CG5384	1.55	1.59E-02
CG12795	1.54	7.79E-03
Pros29	1.53	1.19E-02
Roc1a	1.53	1.11E-02
Rpn12	1.52	2.12E-02
CG13779	1.51	8.89E-03
Cyp9f2	1.51	7.47E-03
Pros54	1.51	3.31E-02
Pros26	1.49	1.46E-02
Tsf1	1.49	3.31E-03
Pros25	1.47	1.99E-02
CG33099	1.46	3.51E-02
Pros45	1.46	1.90E-02
Cyp12d1-d	1.41	3.26E-02
CG11885	1.41	3.85E-02
p47	1.4	1.86E-02
Rpt4	1.39	4.25E-02
Uch-L3	1.39	2.20E-02
CG6218	1.36	2.05E-02
Sirt4	1.36	3.52E-02
PHGPx	1.36	1.86E-02
Rpn11	1.36	2.56E-02
Mov34	1.36	2.08E-02
CG12398	1.36	3.46E-02
CalpB	1.35	3.57E-02
Jheh2	1.32	3.59E-02
Clc	1.31	2.97E-02
Ube3a	1.31	3.51E-02
borr	1.28	4.07E-02
Irc	1.28	3.78E-02
Txl	1.27	2.78E-02
Rpn3	1.27	2.72E-02
CG42488	1.23	2.32E-02
TER94	1.21	3.78E-02
Ice	1.19	4.30E-02
CG4572	1.18	3.84E-02
Cyt-b5	1.17	3.81E-02
Prosbeta3	1.16	4.38E-02
CG4673	1.16	4.35E-02
CG13349	1.15	4.32E-02
CG9436	1.12	4.70E-02
SelG	1.11	4.04E-02

Table 12.2: PCR primers

Name	sequence
RP49-f(14)	CCGCTTCAAGGGACAGTATCTG
RP49-r(14)	ATCTCGCCGCAGTAAACGC
lacZpromoter-f	ATCGCCCTTCCCAACAGTTGC
lacZpromoter-r	TTCTGGTGCCGGAAACCAGG
lacZreporter-f	TGCACATTTTGCAGGAGTACGGC
lacZreporter-r	GATTTCGGCGCGACTGCTACC

Table 12.3: ChIP-seq datasets read mapping statistics

Library	Read Length	Uniquely mapped reads
Ovary shPiwi Rep1 H3K9me3	36	11,093,401
Ovary shPiwi Rep1 Input	36	23,783,156
Ovary shPiwi Rep1 Pol II	36	21,233,655
Ovary shWhite Rep1 H3K9me3	36	17,745,203
Ovary shWhite Rep1 Input	36	22,091,234
Ovary shWhite Rep1 Pol II	36	18,377,757
Ovary shPiwi Rep2 H3K9me3	36	22,467,219
Ovary shPiwi Rep2 H3K9me3 Input	36	14,843,946
Ovary shPiwi Rep2 Pol II	36	9,627,221
Ovary shPiwi Rep2 Pol II Input	36	2,985,999
Ovary shWhite Rep2 H3K9me3	36	21,135,950
Ovary shWhite Rep2 H3K9me3 Input	36	16,619,035
Ovary shWhite Rep2 Pol II	36	5,731,448
Ovary shWhite Rep2 Pol II Input	36	1,629,660

Table 12.4: RNA-seq datasets read mapping statistics (TopHat 1.4.1 mappings)

Library	Read Length	Unique	Multi	Unique splices	Multi splices
Ovary	50	19,868,793	3,249,894	2,021,378	31,552
Ovary shWhite Rep1	50	4,266,297	868,256	389,035	5,895
Ovary shPiwi Rep1	50	5,886,236	906,534	606,030	8,962
Ovary shWhite Rep2	50	10,345,357	1,186,659	607,786	18,881
Ovary shPiwi Rep2	50	12,764,829	1,393,823	1,177,886	25,302

Table 12.5: Repeat analysis mapping statistics (whole genome with unlimited multimappers, zero mismatches)

Library	Read Length	Unique	Multi
Ovary shPiwi Rep1 H3K9me3	36	9,469,110	4,511,259
Ovary shPiwi Rep1 Input	36	20,029,978	2,042,023
Ovary shPiwi Rep1 Pol II	36	17,994,285	1,994,455
Ovary shWhite Rep1 H3K9me3	36	15,101,194	5,076,952
Ovary shWhite Rep1 Input	36	18,568,175	1,435,948
Ovary shWhite Rep1 Pol II	36	15,589,380	1,675,468
Ovary shWhite Rep1 RNA-seq	36	3,682,085	6,376,989
Ovary shPiwi Rep1 RNA-seq	36	5,119,512	5,808,312
Ovary shWhite Rep2 RNA-seq	36	8,658,005	4,005,709
Ovary shPiwi Rep2 RNA-seq	36	10,573,906	3,641,282
Ovary shPiwi Rep2 H3K9me3	36	13,315,195	3,808,164
Ovary shPiwi Rep2 H3K9me3 Input	36	13,489,170	3,501,374
Ovary shPiwi Rep2 Pol2	36	8,137,867	1,183,428
Ovary shPiwi Rep2 Pol2 Input	36	2,424,728	698,521
Ovary shWhite Rep2 H3K9me3	36	19,021,830	9,010,645
Ovary shWhite Rep2 H3K9me3 Input	36	12,018,516	5,698,668
Ovary shWhite Rep2 Pol2	36	4,858,338	824,157
Ovary shWhite Rep2 Pol2 Input	36	1,303,208	873,869

Table 12.6: Repeat analysis mapping statistics (consensus repeats)

Library	Read Length	Unique	Multi
Ovary shWhite Rep1 RNA-seq	36	14,016	4,615
Ovary shPiwi Rep1 RNA-seq	36	39,413	9,692
Ovary shWhite Rep2 RNA-seq	36	15,309	7,910
Ovary shPiwi Rep2 RNA-seq	36	27,691	10,559
Ovary shPiwi Rep1 H3K9me3	36	2,720,971	283,437
Ovary shPiwi Rep1 Input	36	1,123,614	133,470
Ovary shPiwi Rep1 Pol II	36	515,368	109,711
Ovary shWhite Rep1 H3K9me3	36	3,208,049	318,559
Ovary shWhite Rep1 Input	36	739,854	83,425
Ovary shWhite Rep1 Pol II	36	346,044	74,633
Ovary shPiwi Rep2 H3K9me3	36	5,487,961	469,778
Ovary shPiwi Rep2 H3K9me3 Input	36	2,819,017	340,768
Ovary shPiwi Rep2 Pol II	36	380,988	79,937
Ovary shPiwi Rep2 Pol II Input	36	318,557	38,475
Ovary shWhite Rep2 H3K9me3	36	5,556,191	463,857
Ovary shWhite Rep2 H3K9me3 Input	36	1,718,729	205,205
Ovary shWhite Rep2 Pol II	36	220,634	52,925
Ovary shWhite Rep2 Pol II Input	36	174,554	25,171

12.5 No evidence that Piwi binds to the majority of transposons in the *Drosophila* genome

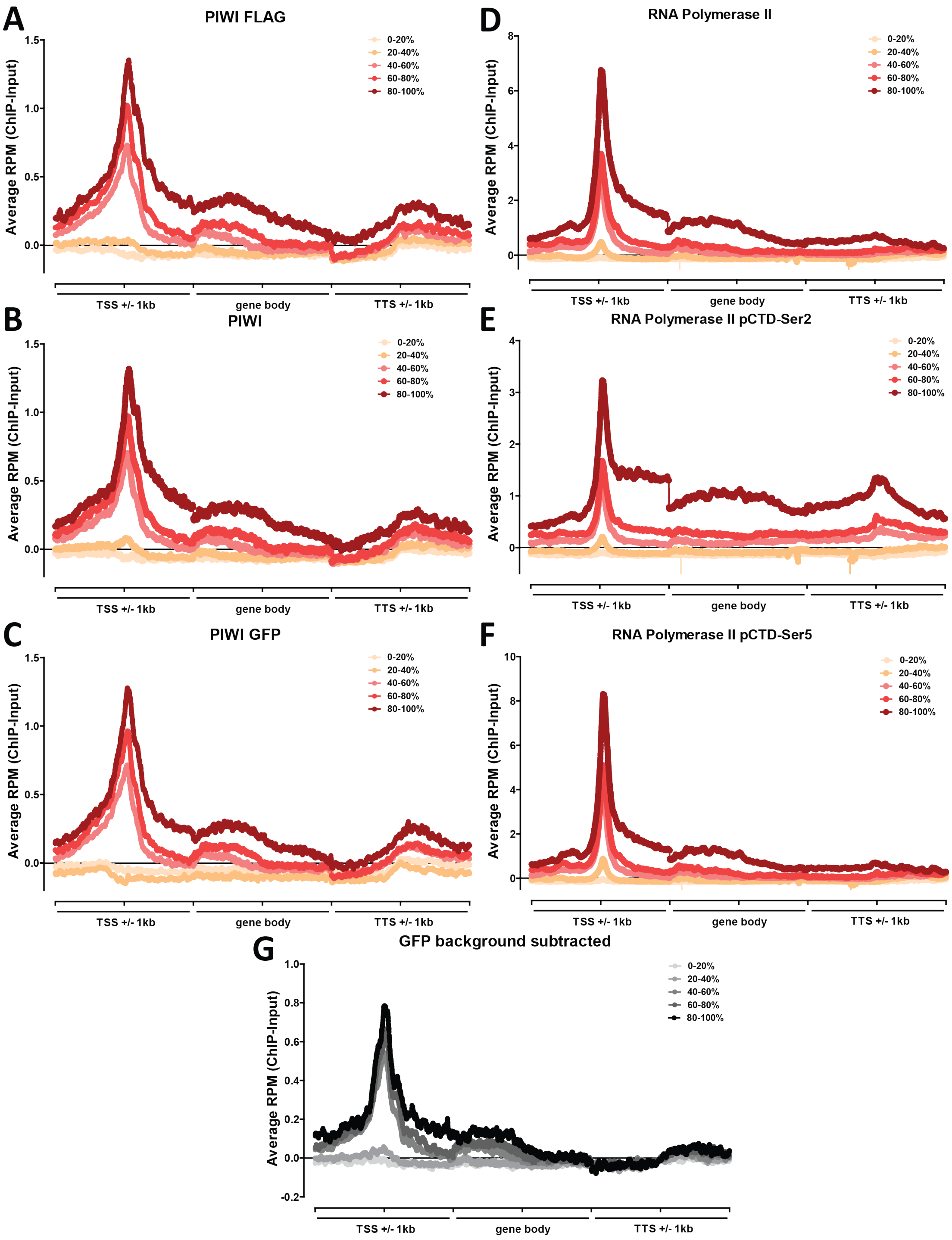
The model suggested by the findings described above, as well as in other recent studies (Sienski et al. 2012; Rozhkov et al. 2013; Ge & Zamore 2013), is one of Piwi scanning the transcriptome for piRNA-matching sequences and initiating transcriptional silencing when such matches are found. An expectation based on this model is that Piwi would be found to physically associate with transcribed genes and with more highly expressed transposable elements but not necessarily with most transposons, which are silenced and expressed only at very low levels.

We tried to test this prediction using Piwi ChIP-seq. Initial experiments using traditional fixation conditions were unsuccessful (data not shown) likely due to the transient and indirect nature of association of Piwi with chromatin (Piwi is likely associating with transcribed RNAs and maybe in some way with the RNA Polymerase machinery but is not necessarily directly interacting with DNA). We reasoned that fixation with a long-arm crosslinking agent such as ethylene glycolbis(succinimidylsuccinate) (EGS) (Abdella et al. 1979; Zeng et al. 2006), which we had previously employed successfully to stabilize protein-DNA and protein-protein interactions (see Li et al. 2012), could result in a successful Piwi ChIP.

We obtained a pattern seemingly consistent with Piwi binding to active genes, as the Piwi ChIP-seq signal was concentrated around transcription start sites and its strength correlated strongly with gene expression levels in datasets generated using a Piwi antibody, a FLAG-tagged version of Piwi, and a GFP-tagged version of Piwi (Figure 12.11A-C). This pattern was very similar to the one observed for RNA Polymerase II (Figure 12.11D) although Piwi enrichment over background was considerably lower. However, an unsettling feature of this pattern was the fact that Piwi was greatly concentrated to TSSs, more similar to the Ser5-phosphorylated form of the RNA Polymerase II CTD Figure 12.11F), which is associated with transcriptional initiation, than the profiles seen in ChIP-seq against RNA Polymerase II CTD pSer2 (Figure 12.11E), which is associated with transcriptional elongation (Buratowski, 2009). Such an observation is not consistent with the scanning model as piRNAs are not concentrated close to the TSS and Piwi would presumably need to scan the whole transcript to find regions complementary to them. This, the findings described in the second chapter of Part III (in particular, the observation that strong read clustering is more often seen in IgG controls than in sonicated inputs, which we used initially for normalization in our analysis), and the suspicion that EGS cross-linking might exacerbate the known sonication biases towards open chromatin (as it may tightly crosslink nucleosomes to each other, making them refractory to sonication in the way heterochromatin is; Auerbach et al. 2009; Teytelman et al. 2009; Gaulton et al. 2010) suggested that the apparent Piwi enrichment might be an artifact of fixation. This was confirmed when we carried out ChIP-seq against GFP in EGS-fixed cells not expressing any GFP or GFP-fusion protein and observed the same pattern as what we saw in Piwi datasets.

Soon after the publication of our work (which excluded all Piwi ChIP-seq data), a study appeared claiming to present the first genome-wide analysis of Piwi binding to the fly genome (Hwang et al. 2013). Its results were very surprising as the authors found that Piwi localizes extensively and highly specifically to transposon sequences, with 87% of reads originating from transposable elements. Transposable elements comprise only about a quarter of the *D. melanogaster* genome based on the repeatMasker repeat element annotation, which would make this Piwi dataset one of the most highly enriched ChIP-seq datasets in existence. Enrichment levels approaching a FRiP value of 0.5 are only seen with proteins associating constitutively and/or very tightly with DNA such as histones, CTCF, RNA Polymerase II (Landt et al. 2012; Marinov et al. 2014) and TFAM (Wang et al. 2013). That a protein that has been so notoriously difficult to ChIP due to the transient nature of its association with chromatin could exhibit even higher ChIP enrichment than what is only sometimes observed with the best performing in ChIP factors seemed inconceivable.

This prompted us to carry out a close examination of the Piwi ChIP-seq data from Hwang et al., which revealed that their Piwi ChIP failed completely and the claimed enrichment over repetitive elements is entirely the result of improper handling of the data. Hwang et


A
PIWI FLAG
Average RPM (ChIP-Input)
1.5
1.0
0.5
0.0
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb
B
PIWI
Average RPM (ChIP-Input)
1.5
1.0
0.5
0.0
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb
C
PIWI GFP
Average RPM (ChIP-Input)
1.5
1.0
0.5
0.0
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb
D
RNA Polymerase II
Average RPM (ChIP-Input)
8
6
4
2
0
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb
E
RNA Polymerase II pCTD-Ser2
Average RPM (ChIP-Input)
4
3
2
1
0
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb
F
RNA Polymerase II pCTD-Ser5
Average RPM (ChIP-Input)
10
8
6
4
2
0
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb
G
GFP background subtracted
Average RPM (ChIP-Input)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0-20%
20-40%
40-60%
60-80%
80-100%
TSS +/- 1kb
gene body
TTS +/- 1kb

al. employed a highly unusual data processing pipeline that involved non-standard read mapping settings and a normalization procedure that amplifies small differences between ChIP and input, but most crucially, they included both unique alignments and multiread alignments (reads mapping to multiple locations in the genome) without normalizing in any way for the number of locations a read can map to, effectively treating all such alignments as separate unique reads. As transposable elements are the primary cause for the presence of repetitive regions in genomes, it is no surprise that such a large fraction of "reads" originated from them.

I illustrate this in several ways here, using an H3K9me3 dataset (a classic heterochromatin histone modification) from Muerdter et al. 2013 and modENCODE transcription factor ChIP-seq data for comparison. Figures 12.12, 12.13 and 12.14 show the effect of data processing (see the following Methods section for details) on the appearance of the Piwi profile over transposable elements (using the three genomic region featured in genome browser snapshots in Hwang et al. 2013). When only unique reads are examined and when multireads are normalized for their multiplicity, no Piwi enrichment is apparent over transposons. The highly localized to repeats distribution of Piwi becomes apparent only when multiread alignments are treated as individual uniquely aligned reads and even then it is also present to a very similar extent in the input dataset.

The same conclusions were drawn from an analysis of the global distribution of Piwi and input signal over transposable elements (Figures 12.15, 12.16A and 12.17). Piwi ChIP-seq was indistinguishable from background and also from modENCODE transcription factor ChIP-seq datasets, for which there is no expectation of high levels of localization to transposons (Figure 12.17). In contrast, H3K9me3 exhibited strong and significant enrichment over background (Figures 12.15 and 12.16B) over transposable elements. Thus the published by Hwang et al. Piwi ChIP-seq is of extremely poor quality and does not demonstrate high levels of localization of Piwi to transposons.

In conclusion, the question what exactly Piwi's distribution over the genome is remains to be directly answered experimentally.

12.6 Reanalysis of Hwang et al. 2013; Methods

Except for where specifically specified otherwise, all data processing was carried out using custom-written python scripts. The dm3/BDGP assembly, release 5 version of the *Drosophila melanogaster* genome was used.

12.6.1 ChIP-seq data processing

Sequencing reads (36bp in data from Huang et al. 2013, paired 75bp reads in data from Muerdter et al. 2013; mixed read lengths trimmed down to 36bp in modENCODE data) were mapped against the genome using Bowtie 0.12.7 (Langmead et al., 2009) with the following settings: `''-v 2 -k 2 -m 1 --best --strata''` for unique 36bp alignments, `''-v 3 -k 2 -m 1 --best --strata''` for unique 2x75bp alignments, and `''-v 0 -a --best --strata''` for alignments in which multireads were retained. The `-X 1000` option was applied and only concordant read pairs were retained for 2x75bp H3K9me3 data. Read mapping statistics for ChIP-seq datasets processed this way are presented in Supplementary Table 3. Read mapping statistics for all alignments are presented in Supplementary Tables 1 and 2.

Three different types of signal tracks were then generated.

Figure 12.11 ***(preceding page)*****: Relationship between ChIP-signal and gene expression in Piwi, RNA Polymerase Two and GFP IgG control datasets**. Shown are metagenes profiles of the ChIP signal (in RPM) over genes (with the background subtracted), with the 2kb (±1kb) regions around the transcription start site (TSS) and transcription termination site (TTS) shown to scale and the rest of the gene body rescaled to 2kb length (genes shorter than 4kb were excluded). Gene were additionally split into 5 quantiles according to their expression levels as measured by RNA-seq. (A) ChIP-seq on FLAG-tagged Piwi ; (B) ChIP-seq on Piwi using a Piwi antibody; (C) ChIP-seq on GFP-tagged Piwi; (D) ChIP-seq pon RNA Polymerase II; (E) ChIP-seq against RNA Polymerase II pSer2; (F) ChIP-seq against RNA Polymerase II pSer5; (G) ChIP-seq against GFP.

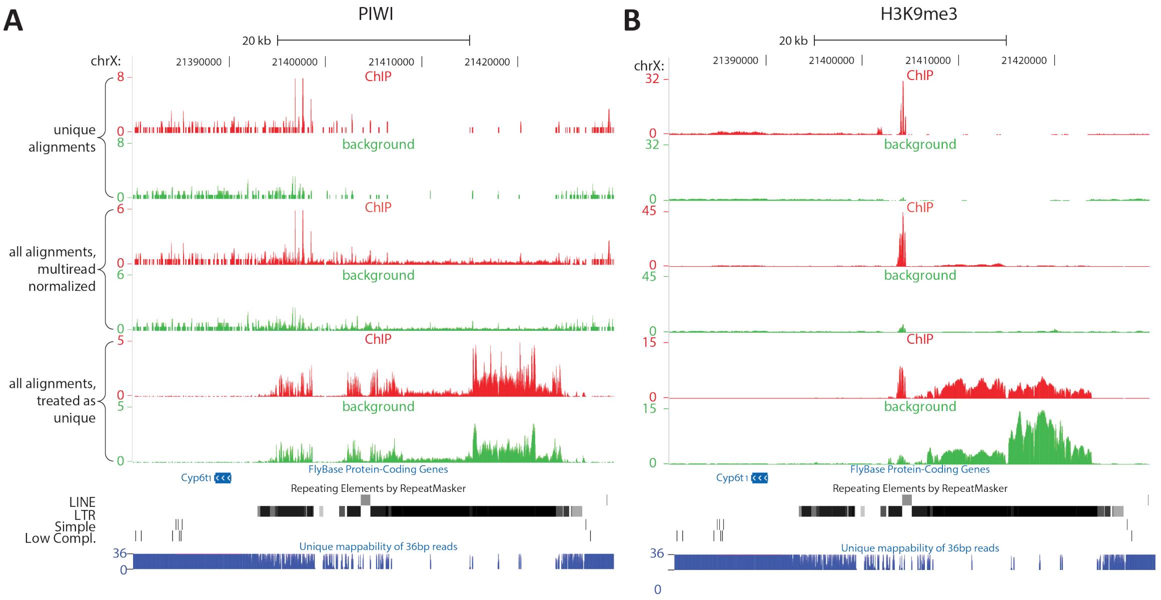


Figure 12.12: Effect of data processing on apparent Piwi occupancy over repetitive elements. Shown is the region from Fig.2B of Huang et al. 2013. (A) Piwi ChIP-seq and background (input) data from Huang et al. 2013 (B) H3K9me3 ChIP-seq and background data from Muerdter et al. 2013. For each dataset, four tracks are shown: 1) unique alignments; 2) all alignments, with multireads normalized for read multiplicity (as described in Methods); 3) all alignments, with all reads treated as unique (analogous but not identical to the processing procedure of Huang et al.); 4) data processed as in Huang et al. 2013. The striking enrichment of Piwi over repetitive elements is only observed when no multiread normalization is applied. Note than in this case a similar enrichment is observed in the background as well. Strong H3K9me3 enrichment is observed only over a short stretch corresponding to a LINE element if multiplicity is taken into consideration. If all reads are treated as unique then H3K9me3 shows a similar profile as Piwi.

1. Unique tracks retaining uniquely mapping reads only, normalized to RPMs (**R**eads **P**er **M**illion mapped reads) according to the following formula:

$$S_{c,i} = \frac{|R_{c,i}|}{\frac{|R|}{10^6}} \tag{12.3}$$

   Where $S_{c,i}$ is the signal score for position i on chromosome c, $|R|$ is the total number of mapped reads, and $|R_{c,i}|$ is the number of reads covering position i on chromosome c.

2. Tracks normalized for read multiplicity based on all alignable reads, where the normalization to RPMs is carried out as follows:

$$S_{c,i} = \frac{\sum\limits_{R \in R_{c,i}} \frac{1}{NH_R}}{\frac{|R|}{10^6}} \tag{12.4}$$

   Where NH_R is the number of locations in the genome a read maps to.

3. Tracks generated using all alignments without normalization for multiplicity, i.e. treating each individual alignment A as if it is a uniquely mappable read:

$$S_{c,i} = \frac{|A_{c,i}|}{\frac{|A|}{10^6}} \tag{12.5}$$

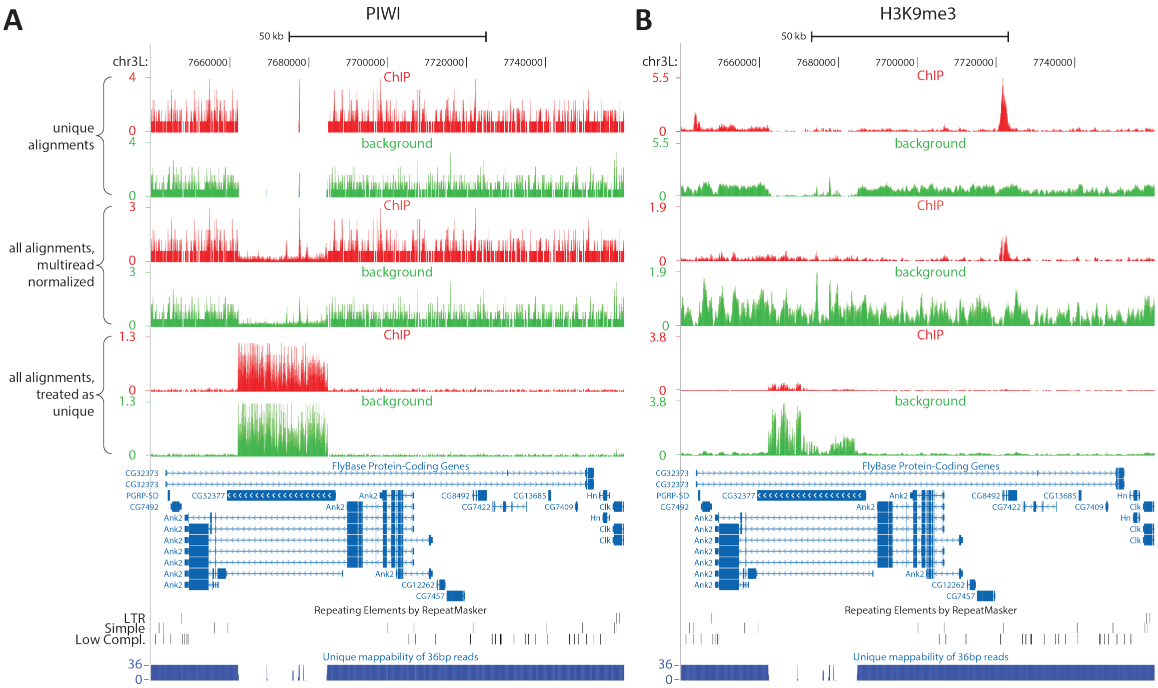


Figure 12.13: Effect of data processing on analysis of Piwi occupancy of repetitive elements. As described in Figure 1 for the genomic region shown in Fig. 2D of Huang et al. 2013. Piwi enrichment is only observed if multiplicity is not taken into consideration. Note that the enrichment over the repetitive ank sequences is stronger in the H3K9me3 background than in the ChIP, indicating the lack of enrichment even if multiplicity is not taken into consideration.

12.6.2 Analysis of RepeatMasker-annotated repeat element coverage

The RepeatMasker repeat element annotation downloaded from UCSC (Kent et al. 2002) was used for all repeat analysis. An RPM score was calculated for each repeat using the following formula:

$$RPM_{RE} = \frac{\sum_{R \in RE} \frac{1}{NH_R}}{\frac{|R|}{10^6}} \qquad (12.6)$$

12.6.3 Analysis of consensus-sequence repeat element coverage

Consensus repetitive elements for *Drosophila melanogaster* were downloaded from FlyBase (Marygold et al. 2013). Reads were trimmed down to 36bp as this was the read length of the Piwi ChIP-seq data from Huang et al. 2013. Reads were then aligned against the Flybase repetitive element consensus sequences using Bowtie 0.12.7 (Langmead et al., 2009) with the following settings: `''-v 3 -a --best --strata''`, i.e. allowing for up to 3 mismatches, and unlimited number of locations a read can map to. Read counts were calculated for each repetitive element and normalized to RPM against the total number of reads aligning to the whole genome (with unlimited number of locations a read can map to) as follows:

$$RPM_{RE_c} = \frac{|R \in RE_c|}{\frac{|R|}{10^6}} \qquad (12.7)$$

where RE_c refers to the consensus repetitive element.

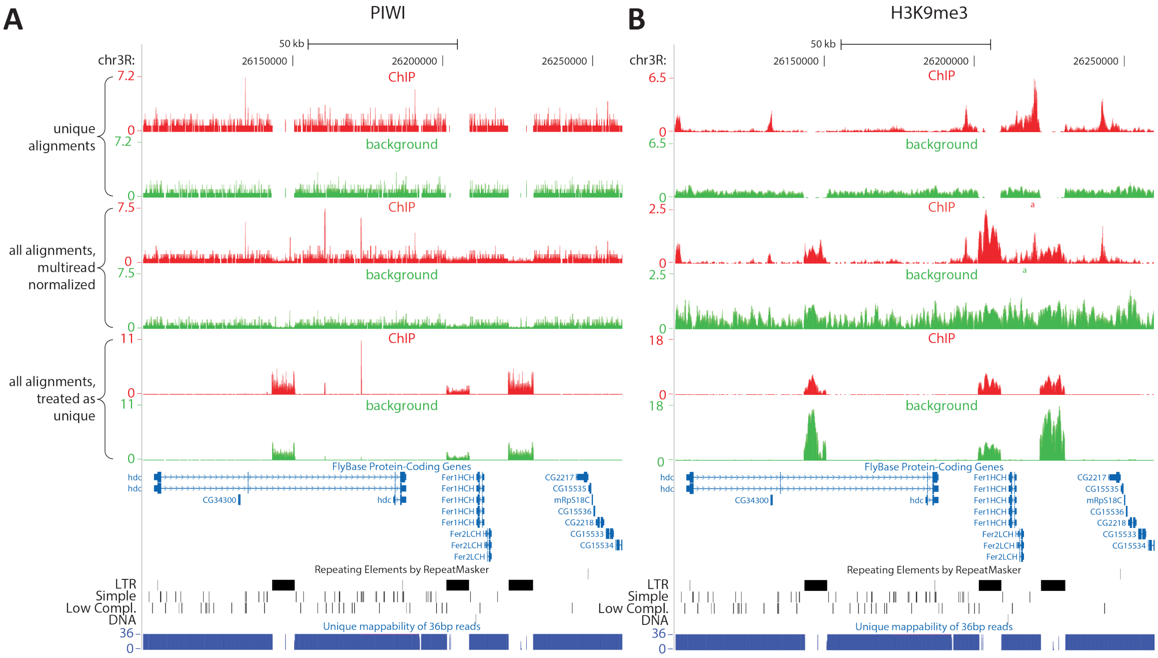


Figure 12.14: Effect of data processing on analysis of Piwi occupancy of repetitive elements. As described in Figure 1 for the genomic region shown in Fig. 2C of Huang et al. 2013. Piwi enrichment is only observed if multiplicity is not taken into consideration. In Contrast to Fig. 1 and Fig S3 in this snapshot at least some of the repeats identified by Huang et al to show Piwi enrichment do show H3K9me3 enrichment even if multiplicity is taken into consideration indicating that these regions are targeted for heterochromatinization.

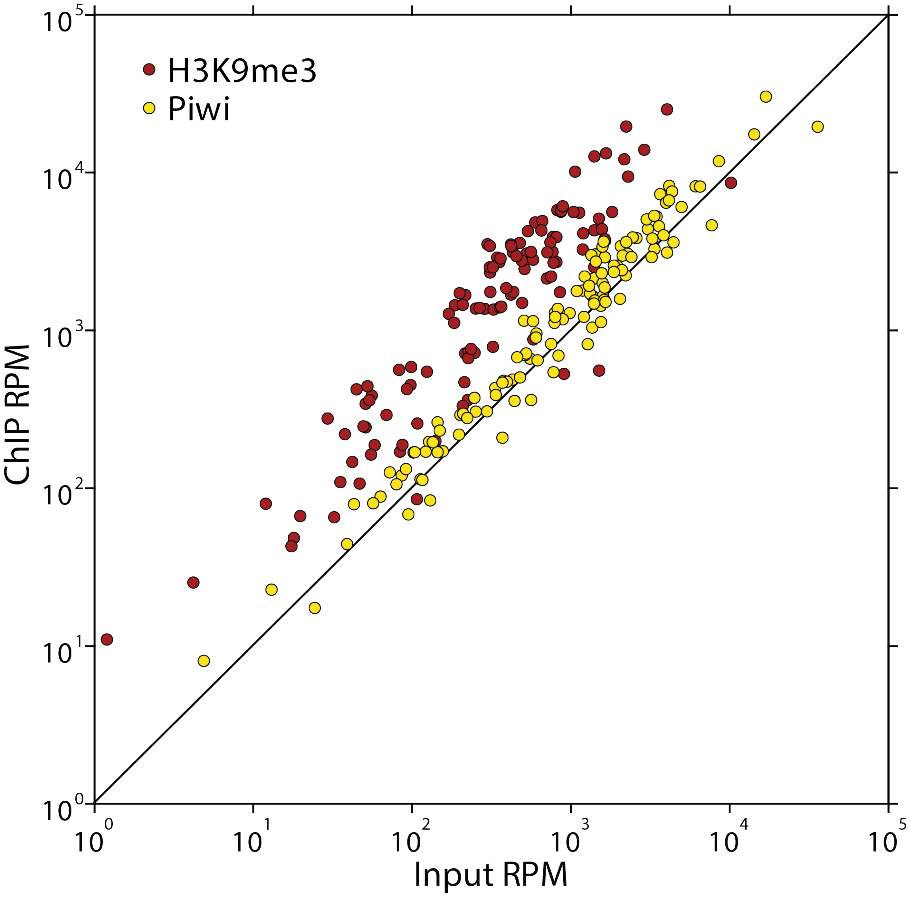


Figure 12.15: Enrichment of Piwi and H3K9me3 over consensus repetitive elements. Shown are the Input and ChIP RPMs for H3K9me3 (red, from Muerdter et al. 2013) and for Piwi (yellow, from Huang et al) over transposon consensus sequences (flybase (Marygold et al. 2013)). All reads were trimmed down to 36bp (the read length of the Piwi ChIP-seq data from Huang et al. 2013) and aligned against the consensus sequences allowing up to 3 mismatches. Read counts were calculated for each repetitive element and normalized to RPM against the total number of reads aligning to the whole genome (with unlimited number of locations a read can map to). A clear overall enrichment over repeats is observed for H3K9me3. In contrast, the Piwi-ChIP dataset from Huang et al. is very similar to the background.

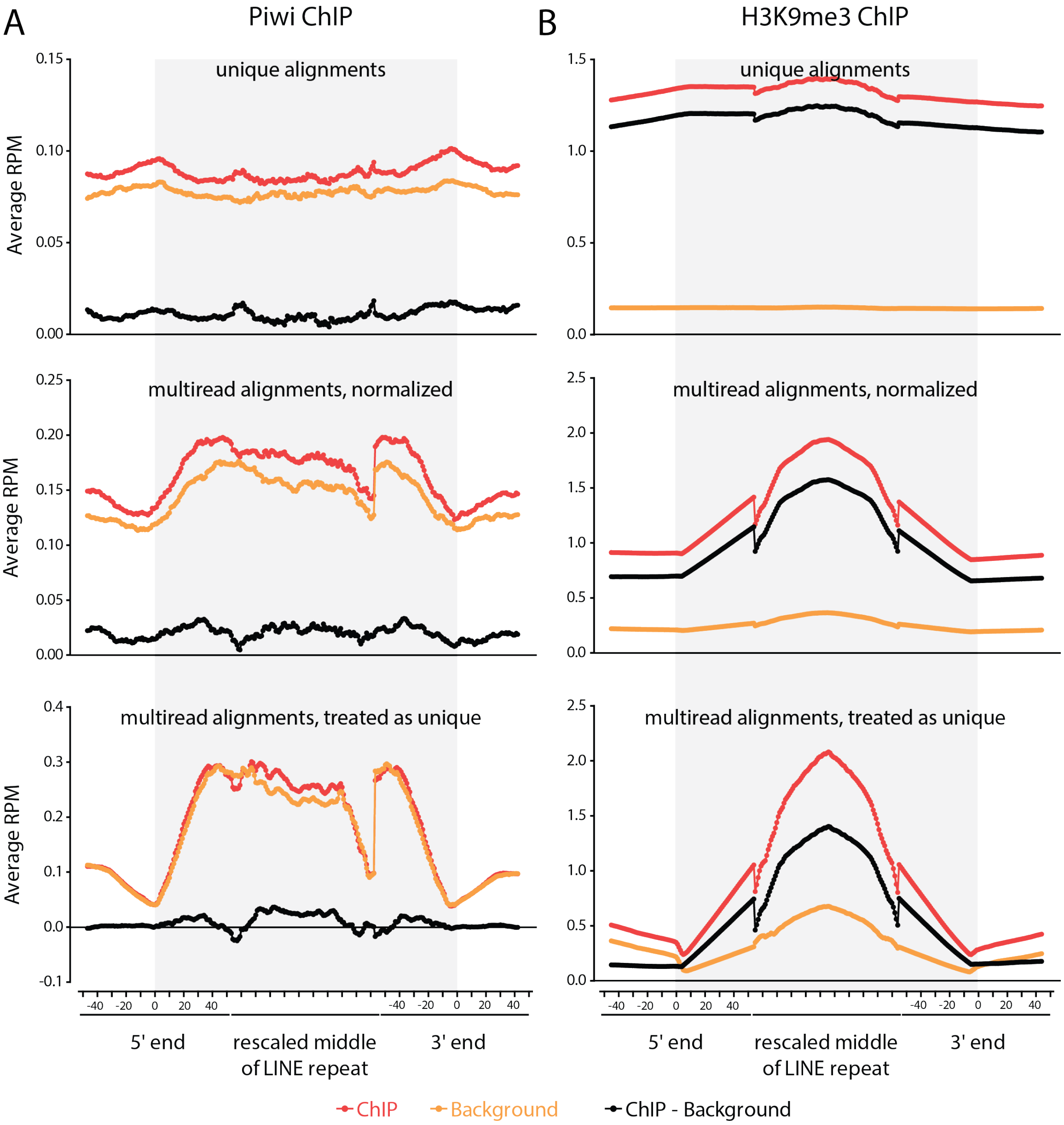


Figure 12.16: Genome-wide enrichment of Piwi and H3K9me3 over repetitive elements. Shown is the average signal distribution over LINE repetitive elements for ChIP (red) and background (yellow) datasets for Piwi from Huang et al. 2013 (A) and for H3K9me3 from Muerdter et al. 2013 (B). The background-normalized enrichment is indicated in black. The 100bp around the beginning and the end of individual elements are shown to scale, the rest of each LINE elements is rescaled to 100 units. The repeatMasker repetitive element annotation available from the UCSC Genome Browser was used. A clear enrichment over background is observed in H3K9me3 datasets, even when only uniquely aligning reads are considered. In contrast, the Piwi dataset from Huang et al. 2013 is indistinguishable from background.

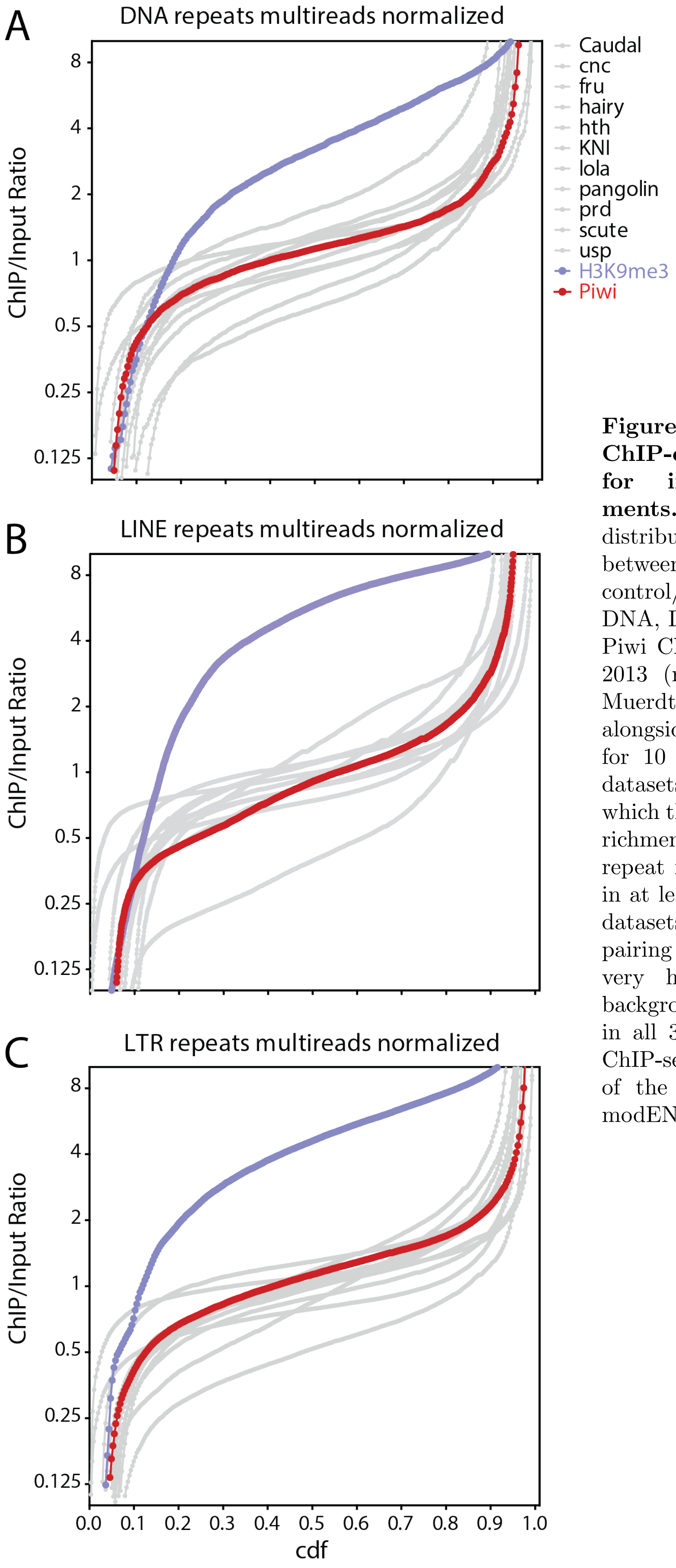


Figure 12.17: Distribution of ChIP-over-control enrichment for individual repetitive elements. Shown is the cumulative distribution function (cdf) of the ratio between the total ChIP RPM and control/background RPM for each DNA, LINE or LTR repetitive element. Piwi ChIP-seq data from Huang et al. 2013 (red) and H3K9me3 data from Muerdter et al. 2013 (blue) are plotted alongside the cumulative distribution for 10 transcription factor ChIP- seq datasets from modENCODE (gray), for which there is no expectation of high enrichment over repetitive elements. Only repeat instances with at least 10 RPM in at least one of the ChIP and control datasets for each ChIP/background pairing were included. H3K9me3 shows very high average enrichment over background over most of the elements in all 3 classes. In contrast the Piwi ChIP-seq data falls in the middle of the distribution of cdf curves for modENCODE transcription factors.

Table 12.7: ChIP-seq datasets read mapping statistics; Huang et al. 2013 Piwi and Muerdter et al. 2013 H3K9me3

Library	Read Length	Alignment policy	Unique reads	Multi-reads
Piwi ChIP (Huang et al)	36	`-v 2 -k 2 -m 1`	1,252,047	
Piwi ChIP (Huang et al)	36	`-v 0 -a`	571,778	173,696
background (Huang et al)	36	`-v 2 -k 2 -m 1`	1,803,636	0
background (Huang et al)	36	`-v 0 -a`	960,165	268,324
H3K9me3 ChIP	2x75	`-v 0 -a`	47,243,150	50,690,870
H3K9me3 ChIP	2x75	`-v 3 -k 2 -m 1`	53,692,762	0
input	2x75	`-v 0 -a`	75,933,354	13,978,550
input	2x75	`-v 3 -k 2 -m 1`	121,920,616	0

Table 12.8: ChIP-seq datasets read mapping statistics; modENCODE

Library	Read Length	Alignment policy	Unique reads	Multi-reads
Caudal-Embryos-0-4h-ChIP-Rep1	36	`-v 0 -a`	6,634,927	1,449,839
Caudal-Embryos-0-4h-Input-Rep1	36	`-v 0 -a`	7,758,011	2,263,029
KNI-Embryos-8-16h-ChIP-Rep1	36	`-v 0 -a`	1,739,675	527,196
KNI-Embryos-8-16h-Input-Rep1	36	`-v 0 -a`	1,424,498	471,903
cnc-Adult-Female-ChIP-Rep1	36	`-v 0 -a`	13,427,663	1,438,195
cnc-Adult-Female-Input-Rep1	36	`-v 0 -a`	16,303,018	1,770,286
fru-Embryos-0-8h-ChIP-Rep1	36	`-v 0 -a`	1,165,616	352,686
fru-Embryos-0-8h-Input-Rep1	36	`-v 0 -a`	1,371,922	381,865
hairy-Embryos-0-8h-ChIP-Rep1	36	`-v 0 -a`	1,073,716	245,579
hairy-Embryos-0-8h-Input-Rep1	36	`-v 0 -a`	1,368,786	271,948
hth-Embryos-0-8h-ChIP-Rep1	36	`-v 0 -a`	1,147,405	292,100
hth-Embryos-0-8h-Input-Rep1	36	`-v 0 -a`	1,391,304	376,075
lola-Embryos-0-12h-ChIP-Rep1	36	`-v 0 -a`	666,154	322,079
lola-Embryos-0-12h-Input-Rep1	36	`-v 0 -a`	1,179,350	312,687
pangolin-Embryos-0-8h-ChIP-Rep1	36	`-v 0 -a`	1,354,532	240,999
pangolin-Embryos-0-8h-Input-Rep1	36	`-v 0 -a`	1,252,479	547,165
prd-Embryos-0-12h-ChIP-Rep1	36	`-v 0 -a`	1,349,189	464,935
prd-Embryos-0-12h-Input-Rep1	36	`-v 0 -a`	1,391,042	447273
scute-Embryos-0-12h-ChIP-Rep1	36	`-v 0 -a`	12,440,506	2,471,660
scute-Embryos-0-12h-Input-Rep1	36	`-v 0 -a`	1,391,042	447,273
usp-Embryos-0-12h-ChIP-Rep1	36	`-v 0 -a`	830,111	252,276
usp-Embryos-0-12h-Input-Rep1	36	`-v 0 -a`	1,270,822	246,580

13

Single-cell heterogeneity in the noncoding transcriptome during iPS cell reprogramming

This chapter contains a study on transcriptomic changes on the single-cell level during iPS repogramming that was intended to be published (but at the time of writing this thesis has not yet been accepted for publication) as:

Kim DH, Marinov GK, Singer ZS, Pepke S, Williams BA, Schroth GP, Elowitz MB, Wold BJ. Single-cell heterogeneity in the noncoding transcriptome during iPS cell reprogramming.

My role was in carrying out most of the computational analysis for it (except the Self-Organizing Map part). I note that it features single-cell RNA-seq that was generated before we established our approaches for correcting for technical noise with a pool/split design and before we made it our standard to include spike-in quantification standards and work with absolute copy-per-cell estimates of gene expression. This is the reason why the manuscript ignored the question of technical noise and the data was analyzed as if there is no noise. Nevertheless we were able to derive useful biological insights from the data.

Abstract

Somatic cell reprogramming into induced pluripotent stem (iPS) cells (Takahashi & Yamanaka 2006; Takahashi et al. 2007; Wernig et al. 2007) involves widespread changes in the protein-coding transcriptome, which have been extensively characterized at the population level (Buganim et al. 2013; Loh et al. 2011). Recent studies have shown that acquisition of pluripotency occurs in a stepwise manner, where functionally related protein-coding genes are activated in distinct waves (Buganim et al. 2012; O'Malley et al. 2013; Polo et al. 2012; Hansson et al. 2012). However, the dynamic changes in the noncoding transcriptome during reprogramming are poorly understood. Here we characterize the transcriptomes of individual reprogramming iPS cells and show that numerous long noncoding RNAs (lncRNAs) are heterogeneously expressed using single-cell RNA sequencing (RNA-seq) and single-molecule RNA FISH (smFISH). At a systems level, activation of the endogenous pluripotency network led to an unexpected global decrease in protein-coding transcriptome variation. Notably though, reprogramming iPS cells failed to fully recapitulate a lncRNA expression repertoire that is more prominent and stably expressed in the pluripotent state. Resetting of the noncoding transcriptome therefore appears incomplete in most iPS cells, even at late stages of reprogramming. Loss-of-function experiments showed that lncRNAs activated during reprogramming (LADR), many of which associate with chromatin regulatory proteins (Guttman et al. 2011; Zhao et al. 2010), are required for generating iPS cells and silencing lineage-specific genes. Transcriptome analysis of iPS

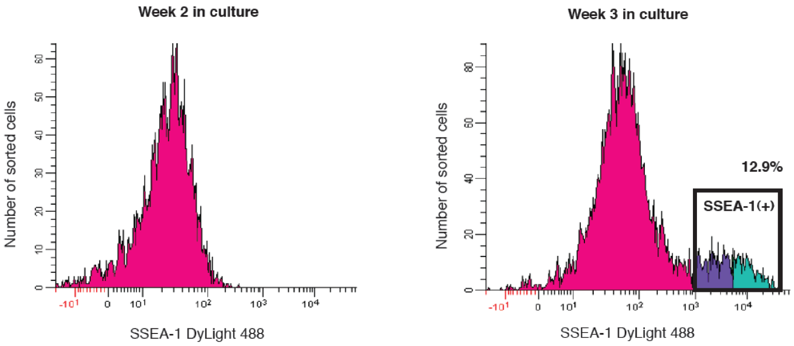


Figure 13.1: Flow cytometry analysis of SSEA-1 in reprogramming TTFs. SSEA-1 expression on reprogramming TTFs after doxycycline exposure for 2-3 weeks in culture, as determined by flow cytometry.

LADR knockdowns showed that two specific lncRNAs, LADR1 and LADR2, corepress a common gene set, indicating combinatorial control of lineage-specific genes by these lncRNAs. Taken together, our findings reveal that functionally important lncRNAs are stochastically active and rate-limiting, with the capacity to directly affect downstream differentiation genes during reprogramming

13.1 Introduction, Results and Discussion

Epigenetic reprogramming is understood to be clonal in nature (Tchieu et al. 2010), wherein individual cells ultimately convert to the pluripotent state. An ectopic pulse of Oct4, Sox2 Klf4, and Myc (OSKM) expression can initiate a lengthy reprogramming process that requires weeks in culture to produce iPS cells (Yamanaka 2009). This process has both stochastic and deterministic elements (Yamanaka 2009; Buganim

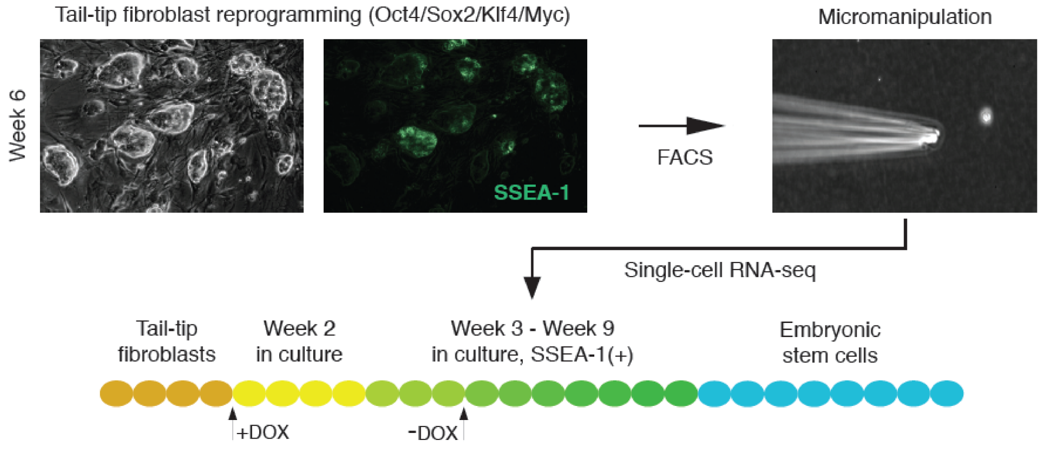


Figure 13.2: Experimental outline. Live-cell imaging, FACS isolation, micromanipulation, and single-cell RNA-seq library generation from tail-tip fibroblasts, SSEA-1(-) and SSEA-1(+) reprogramming iPS cells, and embryonic stem cells. DOX, doxycyline. FACS, fluorescence-activated cell sorting.

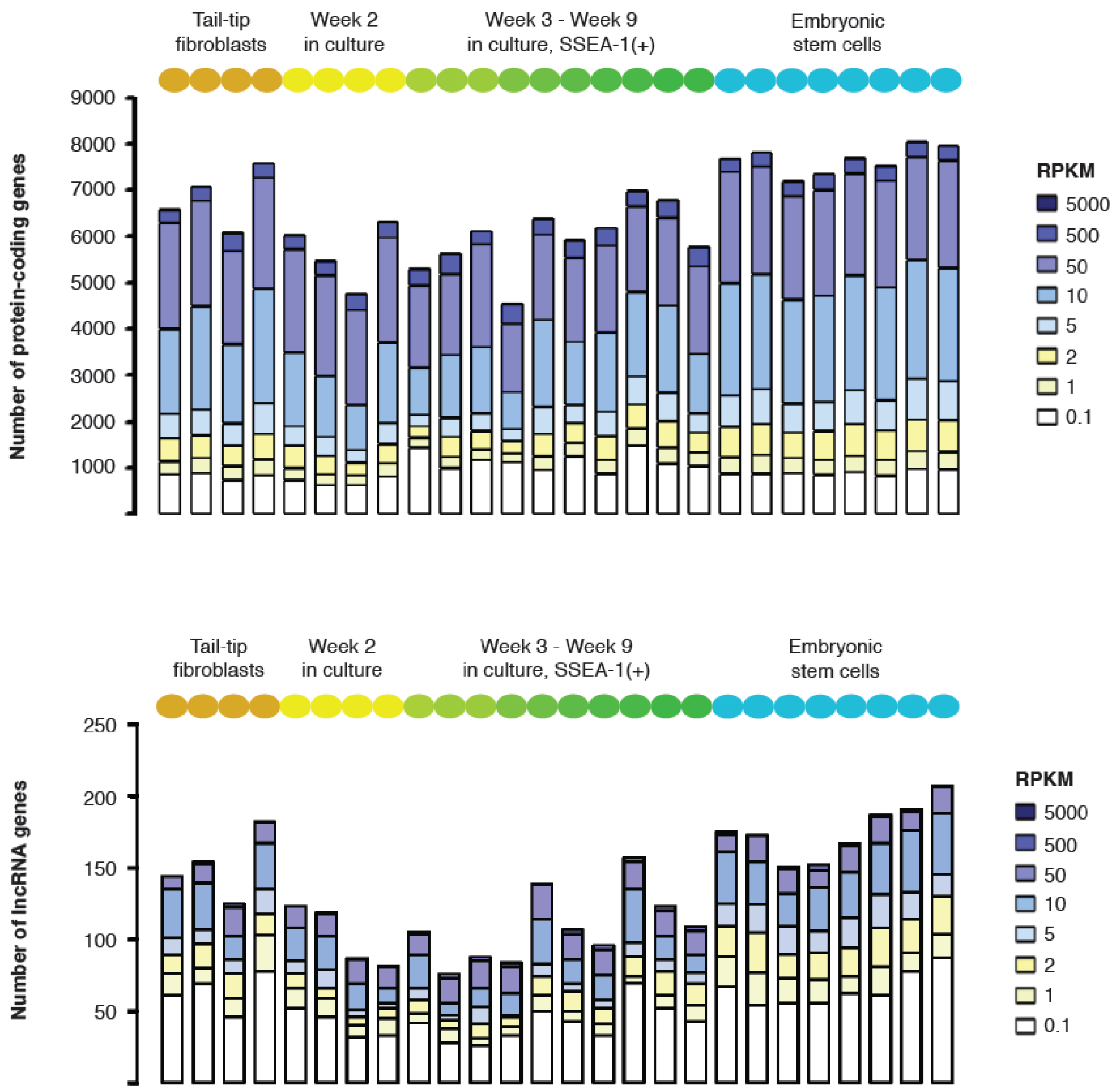


Figure 13.3: Protein-coding and lncRNA genes in single-cell libraries. Number of genes detected in single-cell RNA-seq libraries, according to abundance class. RPKM, Reads Per Kilobase per Million mapped reads.

et al. 2012), and only a small fraction of cells become pluripotent. Conventional protein-coding transcriptome studies of reprogramming, performed at the population level, have identified key transcriptional regulators and chromatin remodeling proteins (Buganim et al. 2013; Loh et al. 2011). Some of those remodeling factors have been shown to associate with lncRNAs (Guttman et al. 2011; Zhao et al. 2010; Lee 2012; Rinn & Chang 2012), but previous studies have not examined the entire coding and noncoding transcriptomes during reprogramming. For both coding and noncoding RNAs, population level measurements obscure individual cell differences by mixing and mutual dilution, blurring both known and new RNA signatures of cell states and phenotypes. Recent studies have begun to address the limitation of population-based approaches by using single-cell techniques to examine small and specific subsets of known, protein-coding genes (Polo et al. 2012; Buganim et al. 2012), but noncoding genes have yet to be characterized systematically during reprogramming, and full transcriptomes have not been

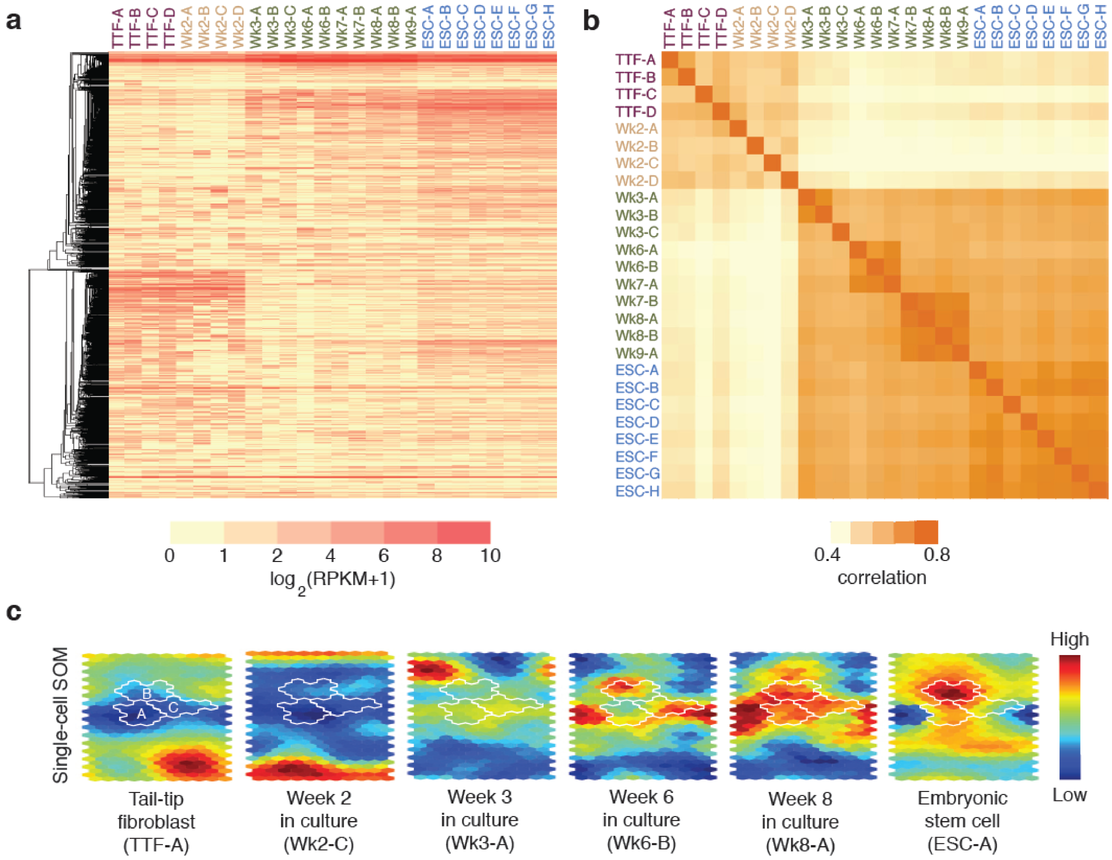


Figure 13.4: Global decrease in transcriptome variation during reprogramming. a, Hierarchical clustering of protein-coding genes detected in single-cell RNA-seq libraries. RPKM, Reads Per Kilobase per Million mapped reads. b, Correlation matrix for single-cell RNA-seq libraries using protein-coding genes. c, Visualization of individual cell transcriptomes using the self-organizing map (SOM). Colorbar indicates normalized expression values of clustered genes, as determined by single-cell RNA-seq.

measured in single cells. Here we performed RNA-seq and smFISH on individual cells drawn from a reprogramming stimulus timecourse, extending the single-cell view to the entire coding and noncoding transcriptomes.

We characterized the single-cell transcriptomes of reprogramming cells by capturing full-length poly(A)$^+$ RNA from individual cells (Ramsköld et al. 2012). We isolated tail-tip fibroblasts (TTFs) from the "reprogrammable mouse" (Carey et al. 2010), which express OSKM in a doxycycline (dox)-dependent manner. TTFs exposed to dox for 2 weeks remained negative for the SSEA-1 reprogramming marker (Buganim et al. 2013), and SSEA-1 positive (+) cells first appeared after 3 weeks of dox induction (Figure 13.1). After 4 weeks of culturing in dox, we obtained SSEA-1(+) iPS colonies that proliferated in the absence of OSKM (Figure 13.2). We sorted SSEA-1(+) cells at 3-9 weeks from the time of OSKM initiation, isolated cells using micromanipulation, and generated single-cell RNA-seq libraries (Figure 13.2 and Table 13.1). Additionally, we constructed RNA-seq libraries from single embryonic stem cells (ESCs) to characterize the transcriptomes of the pluripotent state.

We detected ∼5,000-8,000 protein-coding genes in each single-cell library out of 12,482 protein-coding genes detected at >1 RPKM (Mortazavi & Williams et al. 2008) in the union set of all libraries (Figure 13.3). Additionally, we found that ∼100-200 lncRNA genes were expressed in individual cells, out of the set of 525

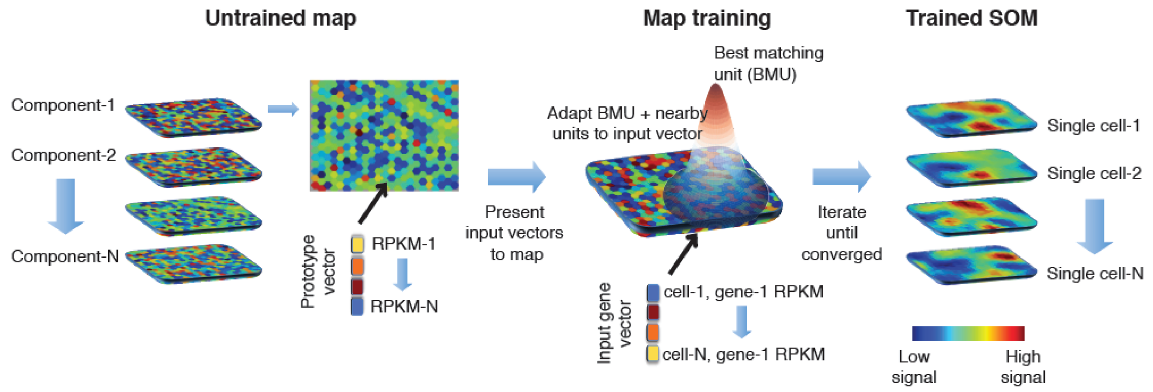


Figure 13.5: The self-organizing map (SOM). A gene is clustered according to the minimum distance of its expression vector from prototype vectors assigned to units in a 2D grid. Initial vectors can be chosen in a variety of ways. In this work, they are initialized by mapping the first two principal components of the data onto the grid. Training proceeds by incrementally moving each prototype toward input vectors that map near it, using a weighting that decreases with map distance from the best matching unit (BMU). The trained SOM consists of prototypes adapted to input data and exhibits spatial organization of units in larger-scale clusters across the grid. Colorbar represents log transformation of normalized data vectors, where normalization is performed on a gene-by-gene basis by subtracting the vector mean and dividing by its standard deviation.

lncRNAs detected at >1 RPKM (Figure 13.3). To examine global differences between single cell transcriptomes, we performed hierarchical clustering of all protein-coding genes (>1 RPKM) (Figure 13.4A). The single-cell transcriptomes of week 2 (Wk2) cells remained most similar to TTFs, but starting at week 3 (Wk3), the transcriptomes of SSEA-1(+) cells began to more closely resemble the ESC transcriptome (Figure 13.4A). Unexpectedly, SSEA-1(+) cells also exhibited a large global decrease in transcriptome variation during reprogramming (Figure 13.4B). The overall systems level picture thus suggests that the reprogramming process entrains participating cells and quashes a level of cell-tocell variability that typifies the TTFs.

In order to cluster, visualize, and search for functional relationships in the single-cell transcriptome data, we generated a self-organizing map (SOM, Kohonen 2013) (Figure 13.4C, Figure 13.5 and Figure 13.6). The SOM integrated data from all cells and projected the resulting clustering onto a two-dimensional topological map, in which proximity on the map reflected similarity of gene expression vectors to each other across all cells. As expected, pluripotency factors (e.g. Nanog, Rex1, Esrrb, Sall4, Oct4) and chromatin remodeling proteins (e.g. Suz12, Jarid1b, Tet1, Tet2, Dpy30) clustered together (cluster A) (Figure 13.6C), and this cluster showed enrichment for the gene ontology (GO) terms “stem cell development” (Bonferroni-corrected $p = 4.88 \times 10^{-3}$) and “chromatin organization” (Bonferroni-corrected $p = 8.61 \times 10^{-6}$). Cluster B was expressed most highly in ESCs and included several key regulators of germ cell development (e.g. Prdm14, Stella) involved in the GO term “nucleic acid metabolic process” (Bonferroni-corrected $p = 5.41 \times 10^{-8}$). Notably, the adjacent cluster (cluster C) (Figure 13.4C) contained several lncRNAs that associate with the chromatin regulatory proteins from cluster A (Guttman et al. 2011; Zhao et al. 2010), including the Polycomb protein Suz12 and Jarid1b. While these lncRNA genes had similar activation kinetics to the pluripotency and germ cell factors, they were coordinately regulated with a different module of genes, including those involved in the GO term “RNA binding” (Bonferroni-corrected $p = 1.33 \times 10^{-6}$). These initial observations from the SOM clustering focused attention on germ cell genes and lncRNAs, together with RNA-associated proteins.

Coinciding with the global reduction in cell-to-cell variation, numerous pluripotency factors were activated by Wk3, including Esrrb, Dppa2, Utf1, and Lin28 (Figure 13.7A), which are predictor genes for successful reprogramming (Buganim et al. 2012). Other pluripotency genes activated by Wk3 included Tcfcp2l1, Fbxo15, Klf2, Fgf4, Dppa4, and Nr0b1, as well as the epigenetic regulators Wdr5, Dnmt3b, and Dnmt3l (Figure 13.7A). Many of these genes are thought to be activated in a deterministic manner, based on single-cell measurements from a pluripotency gene panel (Buganim et al. 2012). While our results are generally consistent with these observations (Figure 13.8), we found that a group of germ cell genes were expressed more heterogeneously during reprogramming (Figure 13.2A and Figure 13.9). Three key germ cell genes in particular, Blimp1, Stella, and Prdm1418, were coordinately expressed at week

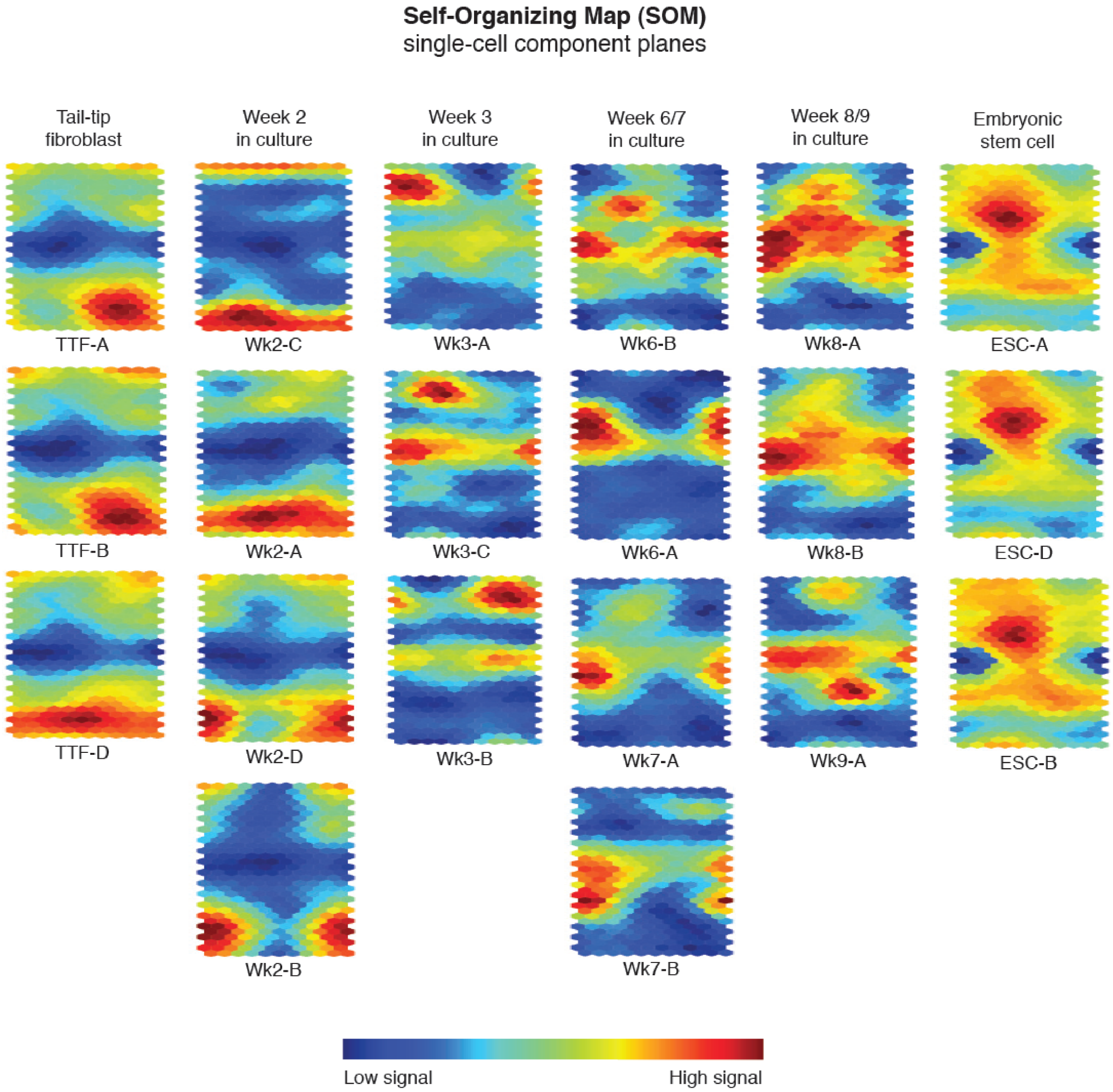


Figure 13.6: Single-cell components of the self-organizing map (SOM). Each single-cell SOM component represents one single-cell RNA-seq library at a defined time-point during OKSM-induced reprogramming, as indicated. Colorbar represents *log* transformation of normalized data vectors, where normalization is performed on a gene-bygene basis by subtracting the vector mean and dividing by its standard deviation.

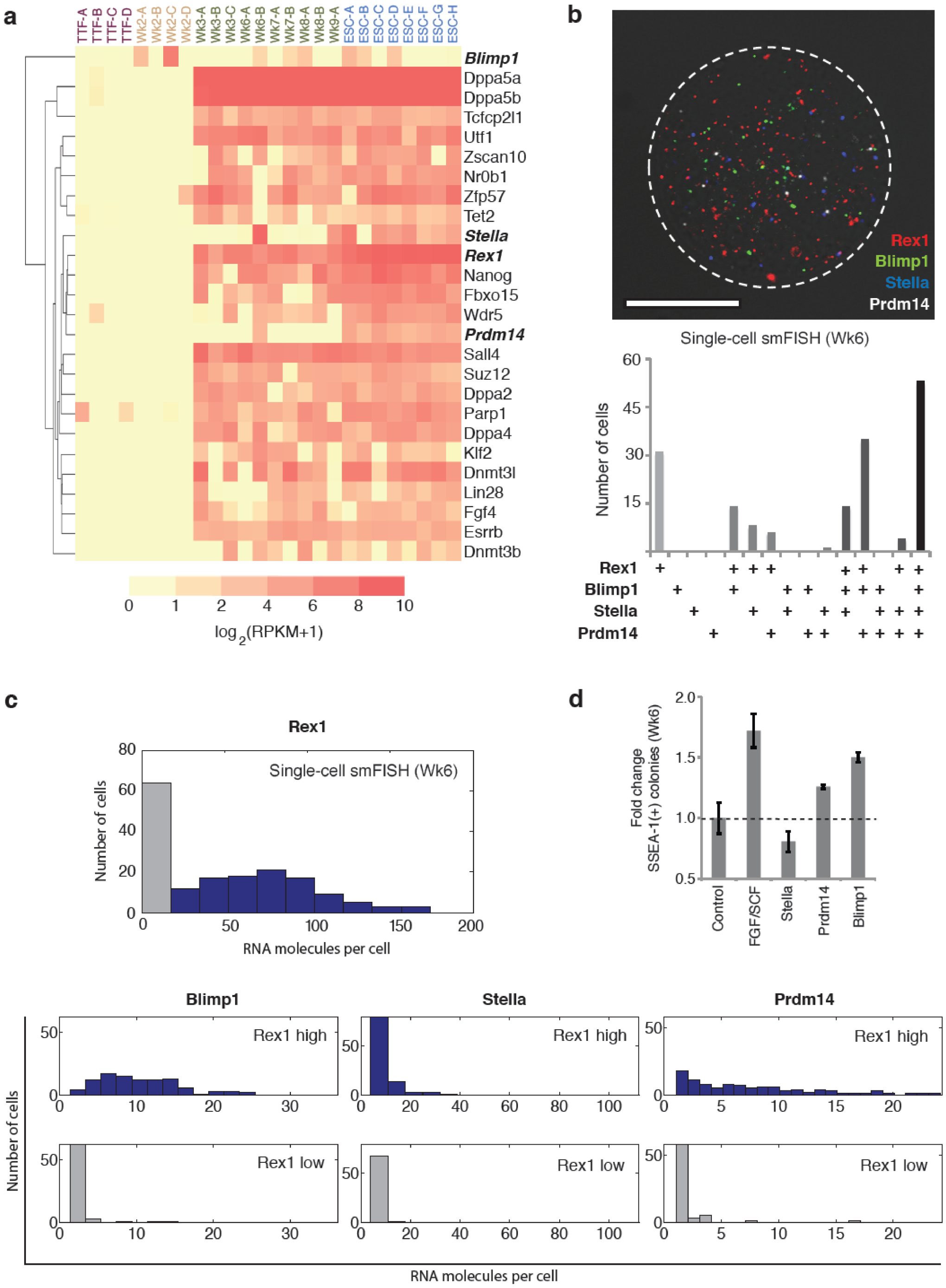


Figure 13.7: Late activation kinetics of germ cell-related genes during reprogramming. a, Hierarchical clustering of a subset of pluripotency- and germ cell-related genes in single-cell RNAseq libraries. Italicized genes in bold indicate genes examined using smFISH. RPKM, Reads Per Kilobase per Million mapped reads. b, Single-cell smFISH of reprogramming iPS cells at week 6 (Wk6) in culture. Scale bar, 10 um. c, Histograms showing the distributions of mRNA molecules per cell as determined by smFISH. d, Reprogramming efficiencies of tail-tip fibroblasts treated with indicated cytokines or expression vectors, as determined by the number of SSEA-1(+) colonies using live-cell imaging. Error bars indicate S.D. ($n = 3$).

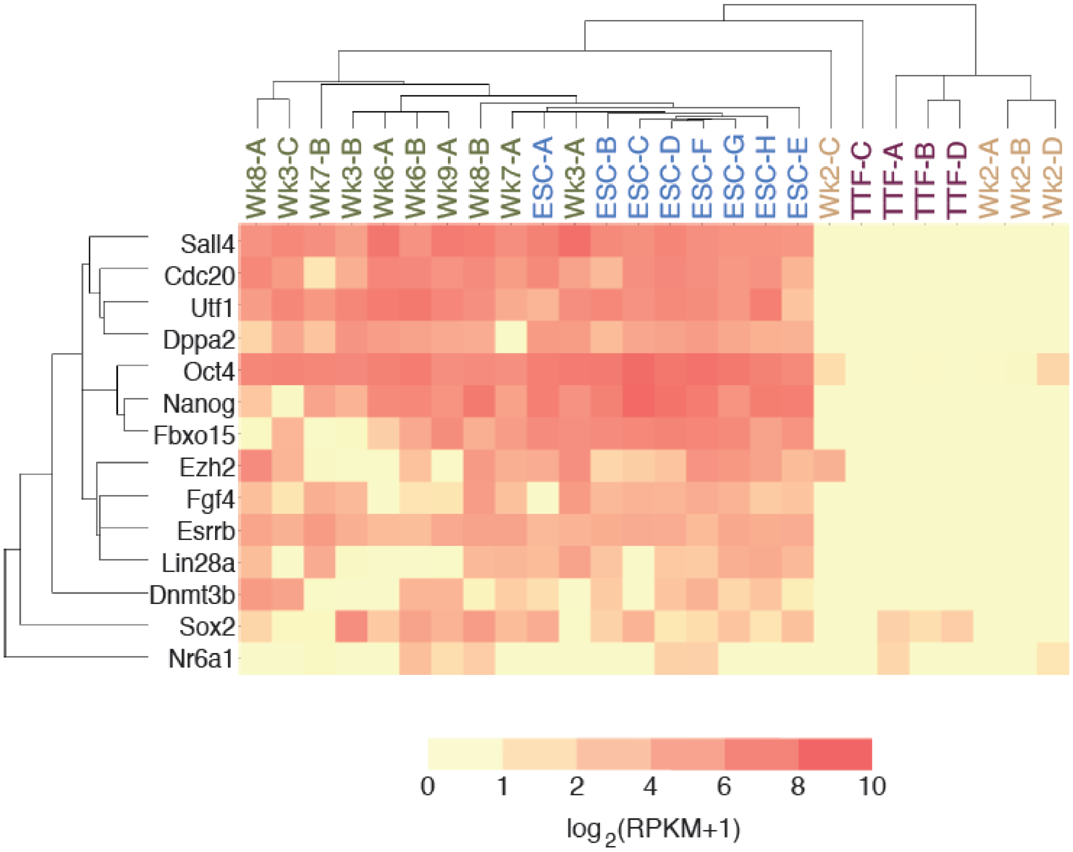


Figure 13.8: Genes involved in Buganim et al. reprogramming hierarchy. Hierarchical clustering of genes involved in a previously reported hierarchical phase of reprogramming. RPKM, Reads Per Kilobase per Million mapped reads.

6 (Wk6) (Figure 13.7A), following pluripotency factor activation at Wk3 (e.g. Rex1, Nanog).

To validate our single-cell RNA-seq results, we used 4-channel smFISH (Raj et al. 2008) as an orthogonal, amplification-independent method to count Blimp1, Stella, Prdm14, and Rex1 transcripts (Figure 13.7B and Figure 13.10), as well as Oct4 and Sox2 (Figure 13.10), in hundreds of cells at Wk6 ($n = 303$). Consistent with the single-cell RNA-seq data, Blimp1, Stella, and Prdm14 were almost always detected only in cells that expressed Rex1 (Figure 13.7B,C). Blimp1 and Prdm14 were mainly expressed in cells with high levels of Rex1, while Stella was expressed in cells with low Rex1 (Figure 13.7C). These results suggest that activation of key germ cell genes may be part of a later and hitherto unappreciated set of limiting molecular events in the reprogramming progression, which predicts that early gain-of-function experiments would increase efficiency. Overexpression of Blimp1 or Prdm14 enhanced reprogramming efficiency by 50% and 26%, respectively (Figure 13.7D), though Stella alone had negligible effect. Seeking independent evidence for a germ-cell network role in reprogramming, we also found that culture conditions (bFGF/SCF/LIF) that induce dedifferentiation of primordial germ cells into pluripotent embryonic germ cells also enhanced reprogramming efficiency by 72% (Figure 13.7D). These results suggest that a set of regulators of epigenetic reprogramming in the germline (Magnúsdóttir et al. 2012) are also engaged during somatic cell reprogramming.

We next tested the functional significance of lncRNAs during reprogramming. Of the 525 lncRNAs expressed at >1RPKM in our single-cell RNA-seq libraries (Figure 13.11A), 240 lncRNAs have previously been reported to physically interact with Polycomb repressive complex 2 in ESCs (Zhao et al. 2010), suggesting that they could be needed to silence lineage-specific genes during reprogramming. We also identified 27 lncRNAs within our single-cell data that associate with additional chromatin-modifying enzymes in ESCs (Guttman et al. 2011), many of which were previously reported to act as inhibitors (e.g. Suv39h1, Yy1) or enhancers (e.g. Ring1b, Eset, Suz12, Jarid1b, Jarid1c) of reprogramming (Onder et al. 2012). A group of robustly expressed lncRNAs in ESCs (asterisk, Figure 13.11A) was notably more variable during the reprogramming time-series at Wk3-Wk9 when compared to ESCs ($p < 0.05$, Kolmogorov-Smirnov test), with no individual cell attaining the high fractional activation observed consistently in ESCs (Figure 13.11B). Interestingly, the majority of these lncRNAs associate with chromatin-modifying proteins in ESCs (Guttman et al. 2011; Zhao et al. 2010) (Figure 13.11B) and are also heterogeneously expressed during epigenetic reprogramming in individual primordial germ cells (PGC) (Magnúsdóttir et al. 2013) (13.13).

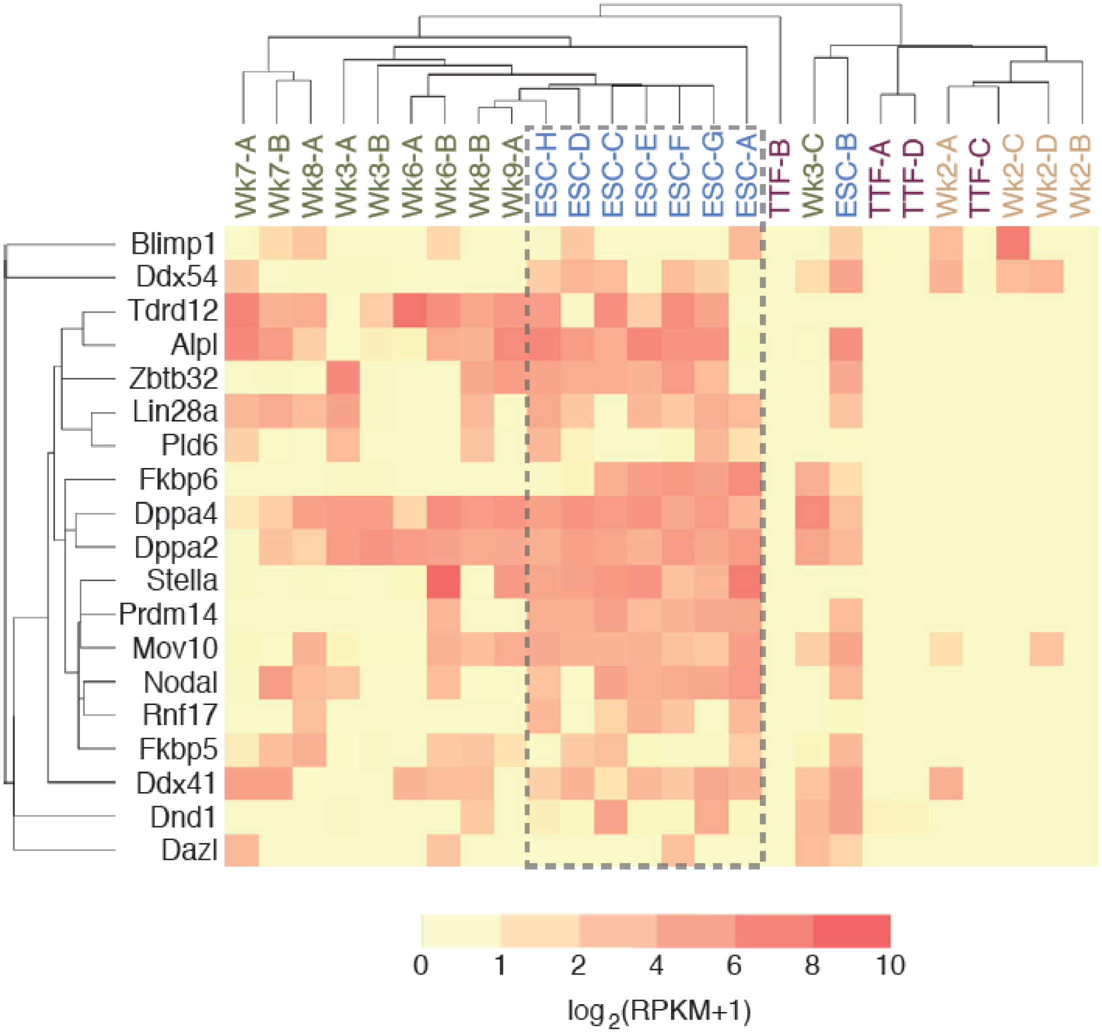


Figure 13.9: Heterogeneity in germ cell-related gene expression. Hierarchical clustering of a subset of germ cell-related genes in single-cell RNA-seq libraries. Dotted line box highlights germ cell-related gene expression signatures prominent in pluripotent ESCs. RPKM, Reads Per Kilobase per Million mapped reads.

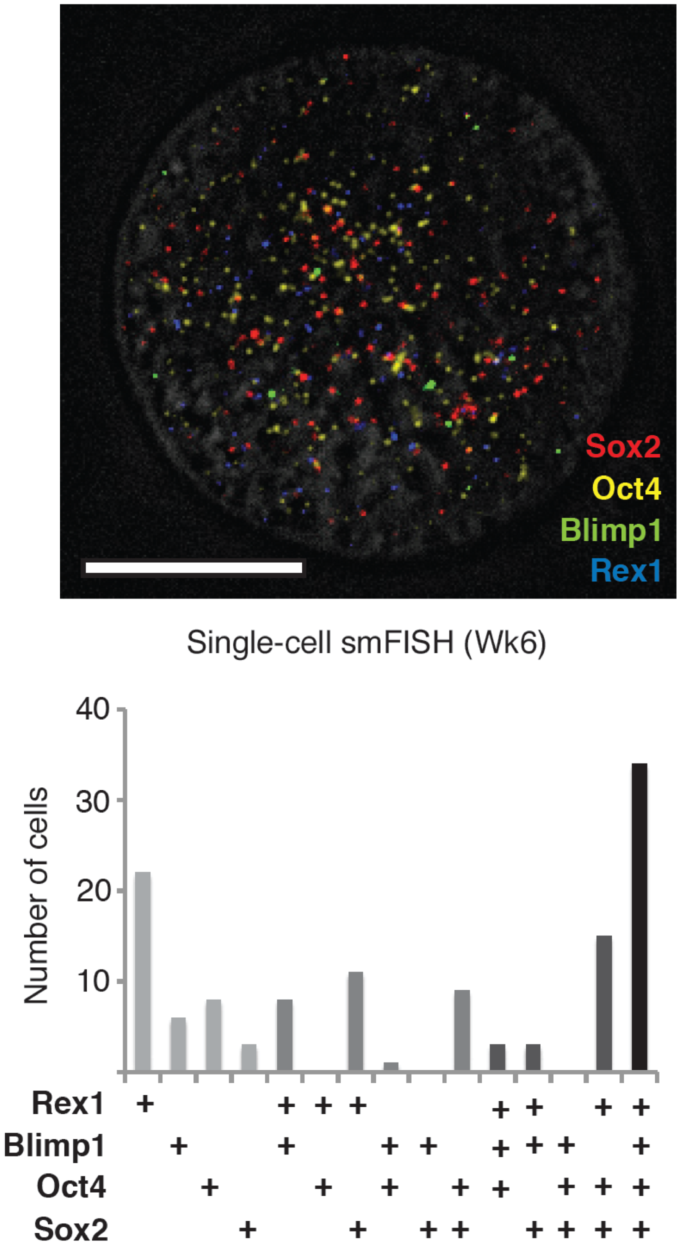


Figure 13.10: smFISH of reprogramming iPS cells. Single-cell 4-channel smFISH of reprogramming iPS cells at week 6 (Wk6) in culture. Scale bar, 10μm.

To further explore lncRNA heterogeneity during reprogramming, we used smFISH to determine the expression of three LADRs in hundreds of cells (n = 351) (13.11C). These Polycomb-associated lncRNAs were expressed at low/undetectable levels in TTFs and were first detected by Wk2 (LADR1, LADR3) or Wk3 (LADR2), as determined by single-cell RNA-seq (13.11B). In single-molecule measurements, LADR3 expression was aberrantly low at Wk6 when compared to ESCs, which might explain their stochastic detectability using single-cell RNA-seq. By Wk9, LADR3 levels became comparable to ESCs (13.11D). In contrast, the LADR2 expression profile showed substantial stochastic variation, with a subset of cells resembling ES, and another group expressing aberrantly high levels at Wk6 that were even more prominent at Wk9, when compared to the more uniform distribution in ESCs (13.11E). Lastly,

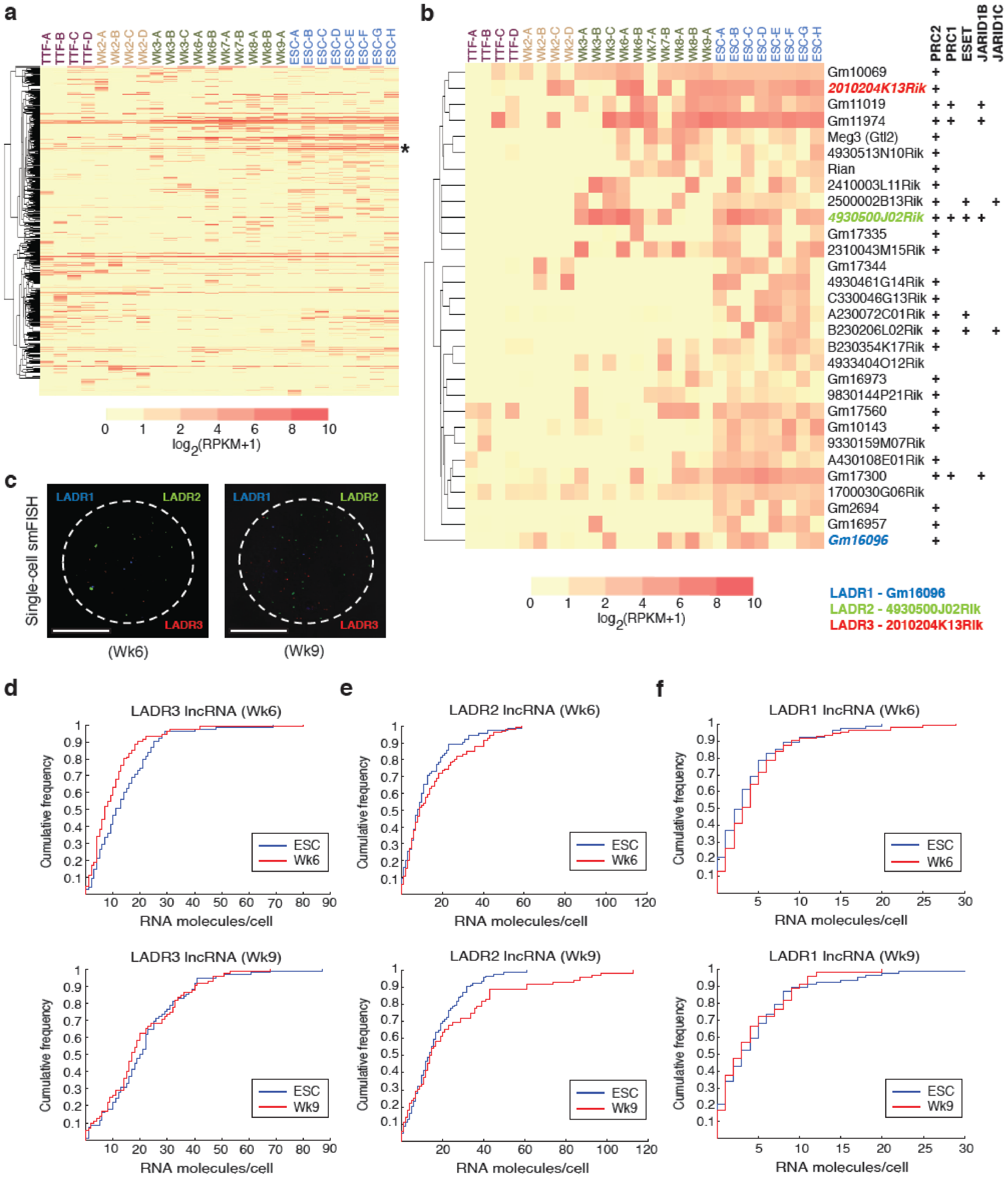


Figure 13.11: Single-cell heterogeneity in lncRNA expression during reprogramming. a, b, Hierarchical clustering of lncRNA genes detected in single-cell RNA-seq libraries (a) and a subset of ESC-enriched lncRNAs (asterisk, b) and their known associations with chromatin regulators (plus). RPKM, Reads Per Kilobase per Million mapped reads. c, Single-cell smFISH of reprogramming iPS cells at week 6 (Wk6) and week 9 (Wk9) in culture. Scale bar, 10 um. d, e, f, Cumulative distribution function plots of lncRNA molecules per cell, as determined by smFISH.

the distributions of LADR1 expression at both Wk6 and Wk9 were relatively uniform and indis-

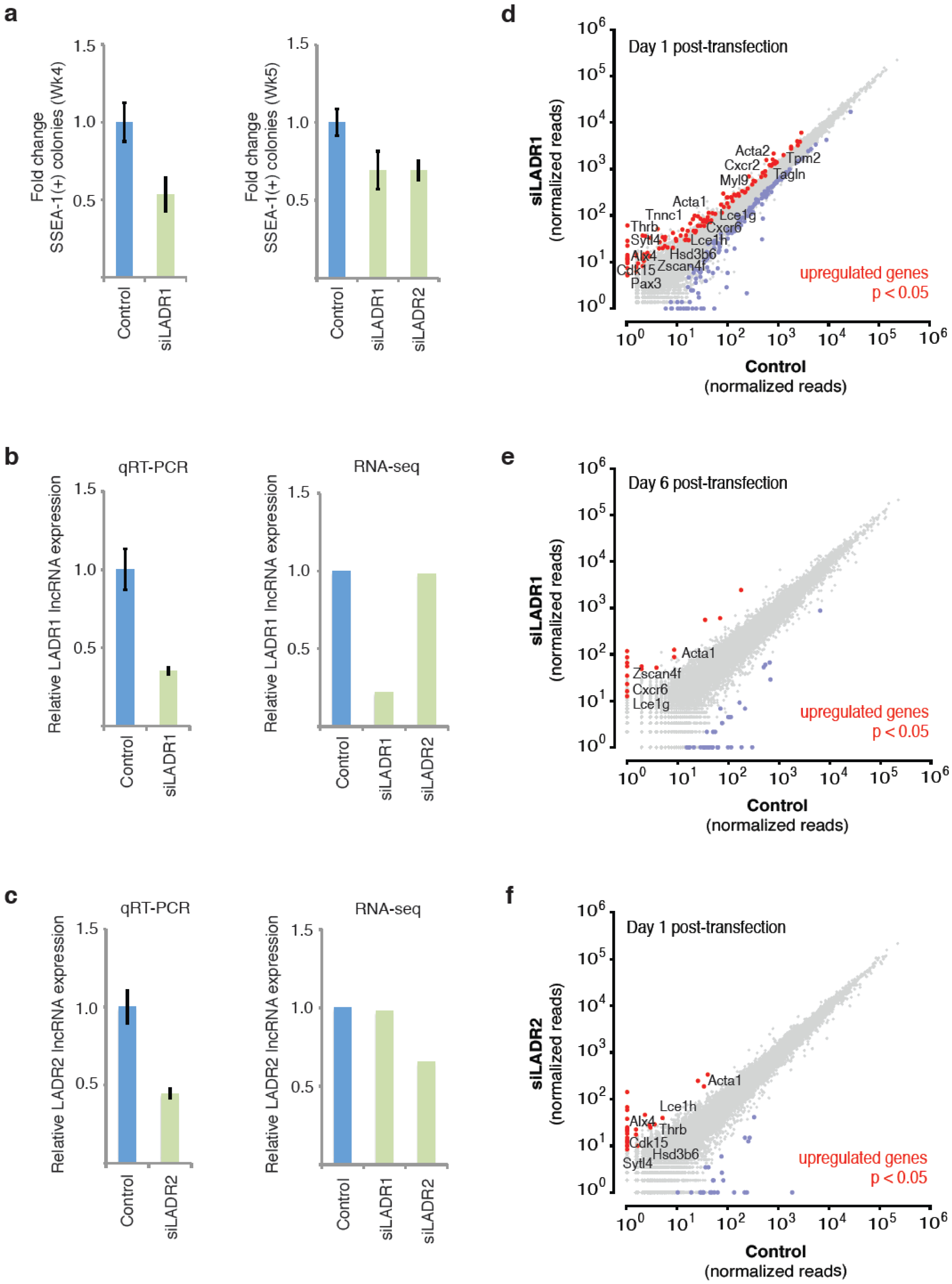


Figure 13.12: Silencing of lineage-specific genes by lncRNAs during reprogramming. a, Reprogramming efficiencies of tail-tip fibroblasts treated with indicated siRNAs, as determined by the number of SSEA-1(+) colonies using live-cell imaging. Error bars indicate S.D. ($n = 3$). b, c, qRT-PCR and RNA-seq quantification of lncRNA expression levels upon transfection of siRNAs targeting LADR1 or LADR2. d, e, f, Differential expression analysis of significantly upregulated (red dots) or downregulated genes (blue dots) in iPS cells deficient for LADR1 or LADR2, as determined by RNA-seq.

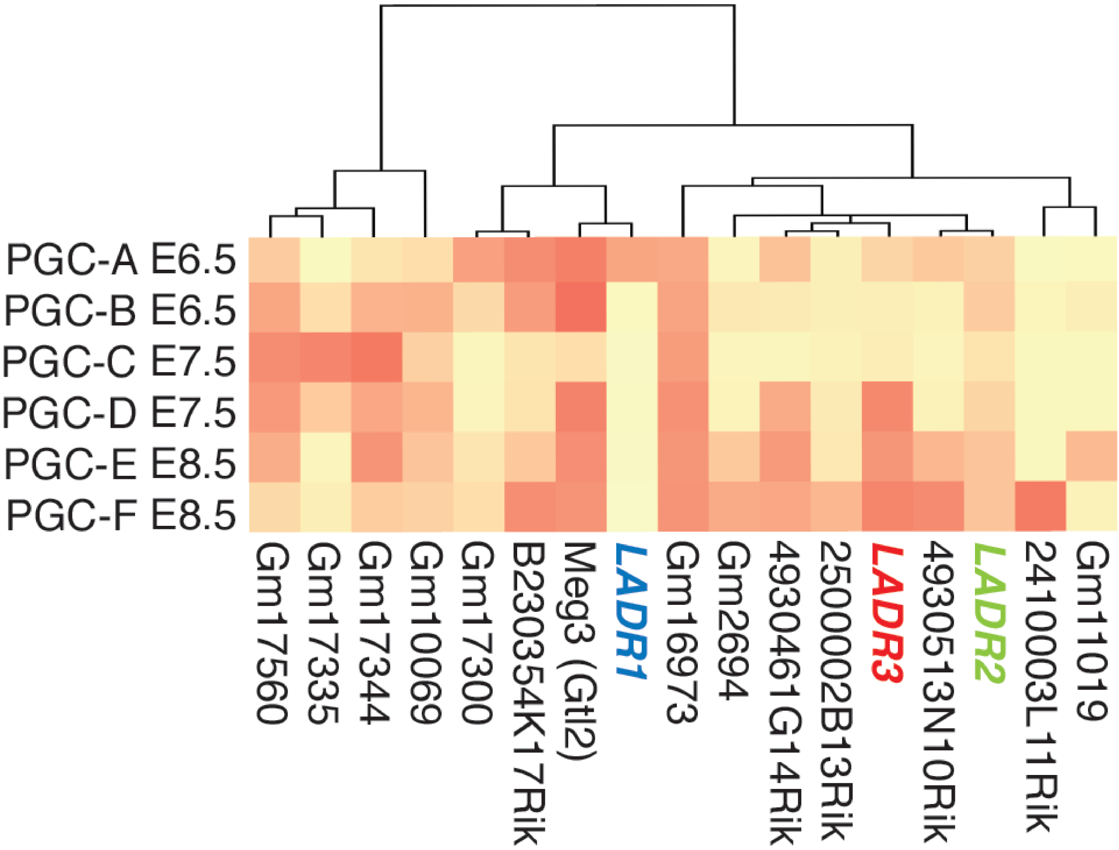


Figure 13.13: lncRNA expression in individual primodial germ cells. Hierarchical clustering of ESC-enriched lncRNAs expressed in at least one primordial germ cell from previously published single-cell RNA-seq . RPKM, Reads Per Kilobase per Million mapped reads.

tinguishable from ESCs, with a subset of cells lacking LADR1 expression (13.11F). Taken together with LADR2 and LADR3, these results highlight a spectrum of cell-to-cell variability for individual lncRNA activation during reprogramming.

Given that individual lncRNAs can modulate the expression of hundreds of protein-coding genes (Guttman et al. 2011), heterogeneity in the noncoding transcriptome may exert broad effects on the protein-coding transcriptome during reprogramming. To test whether Polycomb-associated lncRNAs were functionally important for reprogramming, we performed loss-of-function studies using small interfering RNAs (siRNAs) to attenuate the levels of LADR1 and LADR2, at the time when they were first detected by single-cell RNA-seq at Wk2 and Wk3, respectively. LADR1 knockdown at Wk2 led to a ~50% reduction in the number of SSEA-1(+) colonies by Wk4, and LADR1 or LADR2 knockdown at Wk3 led to a ~30% reduction in SSEA-1(+) colony formation by Wk5 (13.12A). Given the known functions of Polycomb in silencing lineage-specific genes, we used RNA-seq to examine iPS cells deficient for LADR1 or LADR2, to determine whether they were required for gene silencing. Both qRT-PCR and RNA-seq confirmed that siRNAs against LADR1 (siLADR1) and LADR2 (siLADR2) reduced the levels of their respective target lncRNAs, while RNA-seq also showed that siLADR1 and siLADR2 were sequence-specific and did not affect the levels of LADR2 and LADR1, respectively (13.12B,C). LADR1 knockdown led to up-regulation of numerous muscle-related genes, including Pax3, Acta1, Acta2, Tpm2, Tagln, Myl9, and Tnnc1 (13.12D), indicating that LADR1 normally plays a role in silencing these lineage-specific genes during reprogramming. When all differentially expressed genes ($p < 0.05$) were examined, the most enriched annotated GO term was "locomotion" (Bonferroni-corrected $p = 6.34 \times 10^{-7}$), consistent with a functional role for LADR1 in silencing muscle lineage genes.

To examine whether any genes were persistently up-regulated upon loss of LADR1, we performed RNA-seq on iPS cells at day 6 post-transfection of siLADR1. Only 4 genes that were upregulated at day 1 post-transfection remained up-regulated at day 6: Acta1, a skeletal muscle actin, Cxcr6, Lce1g, and Zscan4f (13.12E), which is heterogeneously expressed in a small fraction of ESCs that transit through a two-cell (2C) embryo-like state (Zalzman et al. 2010). For all differentially expressed genes ($p < 0.05$), the most enriched annotated GO term was "MRF (myogenic regulatory factor) binding" (Bonferroni-corrected $p = 3.33 \times 10^{-3}$). Unexpectedly, when we also examined iPS cells deficient for LADR2 by RNA-seq, we found that 7 genes were up-regulated in both the LADR1- and LADR2-deficient iPS cells, including Acta1 and the homeodomain transcription factor Alx4 (13.12F). These findings suggest combinatorial control of a common set of genes by LADR1 and LADR2, indicating that lncRNAs can act together to silencing lineage-specific genes during reprogramming.

This initial study of transcriptome-wide single-cell expression patterns focused attention on lncRNA heterogeneity at both early and late stages of reprogramming, by comparison with fully pluripotent cells. Experimentally perturbing the levels of some of these lncRNAs affected the efficiency of iPS cell derivation. Additionally, numerous lncRNAs that appear stochastic during reprogramming associate with one or more chromatin regulatory proteins (Guttman et al. 2011; Zhao et al. 2010), and our results demonstrated that perturbing these lncRNAs can alter the normal course of expression for lineage-specific genes. Notably, even some late-stage iPS cells exhibited lncRNA heterogeneity and quantitative dysregulation (e.g. LADR2) relative to pluripotent ES cells. We suggest that incomplete and incorrect expression of such lncRNAs could explain the intriguing and therapeutically relevant phenomenon of epigenetic memory in iPS cells (Kim et al. 2010; Polo et al. 2010).

13.2 Methods

13.2.1 iPS cell reprogramming.

Tail-tip fibroblast (TTF) cultures were established from 3-8 day old reprogrammable mice homozygous for both the tet-inducible OSKM polycistronic cassette and the ROSA26-M2rtTA allele (Carey et al. 2010). TTFs were cultured in ES medium (DMEM, 15% FBS, sodium bicarbonate, HEPES, nonessential amino acids, penicillin-streptomycin, L-glutamine, β-mercaptoethanol, 1000 U/mL LIF) with doxycycline and grown on 6-well plates coated with 0.1% gelatin and irradiated MEF feeder cells. For gain-of-function, reprogramming cells were transiently transfected 1 or more times with Blimp1, Prdm14, or Stella TrueORF cDNA plasmids (Origene) using Lipofectamine LTX with Plus Reagent (Life) between weeks 3-4 after OSKM induction. For loss-of-function, reprogramming cells were transiently transfected 1 or more times with lncRNA-targeting siRNAs (IDT) using Lipofectamine RNAiMAX (Life) at early (between weeks 1-4 after OSKM induction) and late (week 6+ after OSKM induction) stages of iPS cell reprogramming. Reprogramming efficiencies were determined by plating equal numbers of cells in triplicate and counting the number of SSEA-1 positive iPS cell colonies using StainAlive SSEA-1 DyLight 488 antibody (Stemgent) and live-cell imaging, where cells were incubated with antibody (1:100) for 2 hours and washed 3 times with PBS. SSEA-1 DyLight 488 positive cells at specified time-points during reprogramming were isolated using flow cytometry on an iCyt Mission Technology Reflection Cell Sorter inside a Baker Bioguard III biosafety cabinet. Single-cell and bulk sample cDNA synthesis and amplification. cDNA synthesis was performed using the Smart-Seq protocol as previously described (Ramsköld et al. 2012). Briefly, the SMARTer Ultra Low RNA kit for Illumina sequencing (Clontech) was used to generate and amplify cDNA from single cells isolated using a micromanipulator or from bulk samples. Intact single cells were deposited directly into hypotonic lysis buffer. Poly(A)+ RNA was reverse transcribed through oligo dT priming to generate full-length cDNA, which was then amplified using 20-22 cycles. cDNA length distribution was assessed using High Sensitivity DNA kits on a Bioanalyzer (Agilent).

13.2.2 Single-cell and bulk sample RNA-seq library generation and sequencing

Single-cell and bulk sample RNA-seq libraries were constructed using the Nextera DNA Sample Prep kit (Illumina). Briefly, cDNA was "tagmentated" at 55 °C with Nextera transposase, and tagmented DNA was purified using Agencourt AMPure XP beads (Beckman Coulter). Purified DNA was amplified using 5 cycles of Nextera PCR, and library quality was assessed using High Sensitivity DNA kits on a Bioanalyzer (Agilent). Libraries were sequenced on the Illumina HiSeq2000. Single-end reads of 50bp or 100bp length were obtained.

13.2.3 Read mapping and expression quantification

All reads were trimmed down to 50bp (if necessary) and mapped to the mouse genome (version mm9) with TopHat (Trapnell et al. 2009) (version 1.2.1) while supplying splice junctions annotated in the ENSEMBL63 set of transcript models. RPKMs for the ENSEMBL63 annotation were obtained using Cufflinks (Trapnell et al. 2010, version 1.0.3) with otherwise default settings. Single-cell libraries ($n = 3$) displaying very low numbers of detected genes were excluded from analysis, as while it is possible that they represent accurate measurements of so far unappreciated biological variability, technical failure of library building is at present the more likely explanation for such observations. For downstream analysis, the biotype classification of genes and transcripts in the ENSEMBL annotation was used to identify noncoding genes. Hierarchical clustering (Spearman rank correlation, unless otherwise indicated) was carried out using Cluster 3.02 (de Hoon et al. 2004) and visualized using Java Treeview (Saldanha 2004). For differential expression analysis, we aligned reads against the refSeq mouse transcriptome using Bowtie 0.12.72 (Langmead et al. 2004). Expression levels were then estimated using eXpress version 1.3.0 (Roberts & Pachter 2013), with gene-level effective counts and RPKM values derived from the sum of the corresponding values for all isoforms of a gene. The effective count values were then used as input to DESeq (Anders & Huber 2010) to assess differential expression.

13.2.4 qRT-PCR

Total RNA was isolated using Direct-zol (Zymo Research) and reverse transcribed using random hexamers or lncRNA-specific primers (IDT, sequences available upon request) and Superscript III reverse transcriptase (Invitrogen) per manufacturers instructions. Real-time PCR was performed on a LightCycler (Roche) using SYBR Green Supermix (Bio-Rad) and normalized to Actin.

13.2.5 Single-molecule fluorescence in situ hybridization

smFISH was performed as previously described (Raj et al. 2008). Up to 48 DNA probes per target mRNA or lncRNA were synthesized and conjugated to Alexa fluorophore 488, 555, 594, or 647 (Life Technologies) and then purified by HPLC. Cells were trypsinized, fixed in 4% Formaldehyde, and permeabilized in 70% ethanol overnight. Cells were then hybridized with probe overnight at 30 °C, in 20% Formamide, 2X SSC, 0.1g/mL Dextran Sulfate, 1mg/mL *E. coli* tRNA, 2mM Vanadyl ribonucleoside complex, 0.1% Tween 20 in nuclease free water. Samples were washed twice in 20% Formamide, 2X SSC, and Tween 20 at 30 °C, and then twice in 2X SSC + 0.1% Tween at RT. 1μL of hybridized cells was placed between #1 coverslips and flattened. Automated grid-based acquisition was performed on a Nikon Ti-E with Perfect Focus System, Semrock FISH filtersets, Lambda LS Xenona Arc Lamp, 60× 1.4NA oil objective, and Coolsnap HQ2 camera. Semi-automated dot detection and segmentation was performed using custom-built MATLAB software with a Laplacian-of-Gaussian Kernel, using Otsu's method to determine "dotness" threshold across all cells in the dataset.

13.2.6 Self-organizing maps

The 5000 genes with the greatest variance among the libraries were used for training a self-organizing map (SOM) (Kohonen 1982; Kohonen 2013). Prior to SOM training, the data vectors were normalized on a gene-by-gene basis by subtracting each vector mean and dividing by its standard deviation. The SOM was constructed using the R package `kohonen`. The total number of map units was set to the heuristic value $5\sqrt{N}$, where N is the number of data vectors. The map grid was initialized with the first two principal components of the data multiplied by a sinusoidal function to yield smooth toroidal boundary conditions. Training lasted 200 epochs (presentations of the data) during which the radius within which units were adapted toward the winning unit decreased linearly from $h/8$ to 2 units, where h is the map height (always chosen as the direction of largest length). Further analysis, including clustering and visualization, was performed with custom python code. Clusters were seeded by the local minima of the U-matrix, with a value for each unit defined as the average of the vector difference between that unit's prototype and its six neighbors on the hexagonal grid. All other unit prototypes were then assigned to clusters according to the minimum vector distance to a seed unit. The lists of clustered genes were submitted to the Princeton GO TermFinder (Boyle et al. 2004) server (`http://go.princeton.edu`) in order to determine enriched terms.

Table 13.1: Read mapping statistics for single-cell RNA-seq libraries.

Single cells	Unique reads	Unique splices	Multi reads	Multi splices
TTF-A	4,754,379	1,793,789	1,016,116	17,993
TTF-B	3,186,598	1,330,084	705,038	25,690
TTF-C	5,882,879	1,802,289	1,364,458	14,894
TTF-D	4,150,724	1,651,656	799,307	12,882
Wk2-A	6,796,932	2,388,019	1,320,707	28,850
Wk2-B	6,695,477	2,128,141	1,687,672	29,942
Wk2-C	7,321,838	2,460,231	1,319,304	23,164
Wk2-D	3,766,443	1,565,189	1,006,975	27,793
Wk3-A	10,817,581	1,468,929	7,879,702	13,590
Wk3-B	10,544,532	1,005,115	4,389,066	9,234
Wk3-C	15,297,126	1,725,119	5,606,955	17,313
Wk6-A	6,649,903	812,609	3,487,925	7,031
Wk6-B	16,445,945	1,629,904	8,522,770	19,809
Wk7-A	20,598,921	2,543,587	11,733,247	27,259
Wk7-B	13,242,715	1,516,497	7,271,470	15,986
Wk8-A	12,817,740	1,535,044	6,579,353	17,672
Wk8-B	14,308,754	1,453,584	7,336,354	13,362
Wk9-A	13,643,846	1,753,756	7,938,751	16,795
ESC-A	8,280,645	2,934,602	2,123,766	26,275
ESC-B	7,072,853	2,610,021	2,225,725	24,739
ESC-C	6,227,982	2,182,842	1,853,443	17,492
ESC-D	5,048,404	1,767,981	1,550,137	15,637
ESC-E	9,095,244	3,412,551	2,766,101	32,967
ESC-F	3,061,161	1,177,880	889,317	10,666
ESC-G	6,711,997	2,475,645	2,177,537	20,003
ESC-H	4,115,001	1,578,915	1,268,654	13,544

Part V

Conclusions and Towards the Future

14

Third-Generation Sequencing Technologies and Functional Genomics Studies

High-throughput sequencing technologies, in particular the Illumina platform, form the basis of most of the work described in this thesis. The short nature of the reads they generate, however, has also presented numerous challenges to data analysis and as repeatedly mentioned so far. Platforms that produce long reads have now emerged, and here I present my perspective on the implications of these technologies, their strength and their expected limitations, on the future of functional genomics research.

Abstract

In recent years, "second generation" sequencing technologies have revolutionized multiple aspects of biomedical research, in particular genome sequencing and functional genomic studies. However, the short-read nature of the data produced by second generation sequencing instruments has presented numerous challenges to data analysis and interpretation in both areas due to the specifics of library generation, read alignment, assembly and a number of other issues. So called "Third-generation" sequencing technologies promise to alleviate a lot of these difficulties by providing a combination of single-molecule sequencing and/or much longer read lengths. This is expected to greatly benefit *de novo* genome sequencing and genome resequencing efforts, but it also has the potential to transform functional genomics studies by resolving existing issues that second-generation technologies have so far not been able to conclusively address, and by opening completely novel research directions. In the same time, certain functional genomic applications are very well suited to the short-read format and have at this point reached maturity, and are therefore less likely to change significantly in the future. Here, the anticipated impact of further developments in sequencing technology is reviewed, together with the still unmet challenges to data quality that will have to be resolved in order to answer the major unresolved questions in the field.

14.1 Introduction

The completion of the sequence of the human genome in the early 2000s (Lander et al. 2001; Venter et al. 2001; International Human Genome Sequencing Consortium 2004) was the culmination of many years development of genomic science and provided the foundation for an explosion in the further advancement of our understanding of the structure and function of genomes during the next decade. While a lot can be learned from the sequence of the genome and the annotation of the genes in it, full understanding of the relationship between the genomic sequence on one side, and cellular and organismic phenotypes on the other, requires deep and comprehensive understanding of the mechanisms of regulation of gene expression, the genomic regulatory elements through which it is carried out, and their dynamics (Hood & Galas 2003; ENCODE Project Consortium 2004). For this reason, a key component of the advances in genomics following the completion the human genome sequences has been the development of functional genomic tools for measuring gene expression, interactions between proteins and DNA, the activity of regulatory elements, and many others.

For about a decade, between the late 1990s and the late 2000s, functional genomics was dominated by DNA microarray technology, which is based on the hybridization of DNA molecules in a sample against a known set of complementary sequences situated on an array. Initially, the availability of genome (or transcriptome) sequences allowed the development of microarrays designed to measure gene expression levels (Schena et al. 1995; Lashkari et al. 1997). Later, the combination of chromatin immunoprecipitation and microarrays (ChIP-on-Chip) enabled the mapping of the occupancy of transcription factors in promoter regions or over the whole genome (Iyer et al. 2001; Ren et al. 2000). Microarray-based techniques delivered numerous insights into genome biology (ENCODE Project Consortium 2007); however, they were still a less-than-ideal solution to the major challenges in the field, as they suffered from issues with hybridization artifacts, the lack of single base pair resolution, and the limitation of measurements to only sequences included on the array. The latter, especially, made difficult not only the assaying of the whole human genomes, but imposed a major limitation in terms of which organisms were available to be studied: a new microarray had to be manufactured for each species, and the process of designing and producing arrays was slow, cumbersome and expensive.

The sequencing of the human genome relied entirely on assembly of the genome from reads of several hundred base pairs (bp) length generated using the Sanger sequencing method (Sanger et al. 1977), which requires extensive sample preparation and has low throughput. As a result, it cost several billion dollars. Later sequencing projects for organisms with similarly sized genomes were less costly, but still carried a price tag in the millions of dollars. This stimulated the development of so called "second-" or "next-generation" (NGS) high-throughput sequencing technologies in the mid-2000s, which promised to make genome sequencing much cheaper and faster. The first such technology was 454 pyrosequencing (Margulies et al. 2005), followed shortly by Polonator sequencing (Shendure et al. 2005), Solexa (later Illumina) (Bentley etal. 2008), ABI SOLiD (McKernan et al. 2009), Helicos (Harris et al. 2008), and more recently, Ion Torrent (Rothberg et al. 2011). Initially, these technologies delivered much shorter reads than Sanger sequencing did: a few tens to hundreds of thousands reads, with a read distribution in the low hundreds of bp (454), or a few hundreds of thousands to a few million reads that were just 20-25 bp long (Solexa/Illumina). Very short read lengths pose severe challenges to *de novo* genome assembly (Whiteford et al. 2005; Alkan et al. 2011), but they are much better suited for functional genomic applications, and this is where they were first applied and made their mark, helping them become well-established (Wold & Myers 2008). Small RNA species such as miRNAs and piRNAs (Bartel 2004; Aravin et al. 2007) are mostly less than 30bp long which enabled the direct sequencing of the whole cellular repertoire of small RNAs very early in the development of NGS technologies and greatly stimulated the development of the field (Ruby et al. 2006; Brennecke et al. 2007). The coupling of the ChIP assay with high-throughput sequencing (ChIP-seq) allowed the truly genome-wide identification of protein-DNA interactions (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007; Mikkelsen et al. 2007), while the direct sequencing of reverse-transcribed RNA fragments provided single base pair-resolution view of the transcriptome (Nagalakshmi et al. 2008; Mortazavi et al. 2008; Cloonan et al. 2008; Sultan et al. 2008; Wilhelm et al. 2008; Wang et al. 2008). By obviating the need for the design and manufacturing of arrays for each genome, sequencing-based assays allowed the application of functional genomics approaches to any species with a sequenced genome (discussed in depth in the last chapter of the thesis). A wide array of "seq-assays" has been developed in the last few years (Table 14.1) targeting almost every imaginable aspect of chromatin, transcriptional and RNA biology, and as a result sequencing has gradually replaced arrays as the method of choice for assaying of nucleic acids in functional genomics (ENCODE Project Consortium 2011).

As technology has improved, the number of reads and their length have increased significantly and the cost of sequencing has dropped; in

Table 14.1: Seq-based functional genomic assays.

Group of assays	Assay	Detection of / Description	References
Genomic Occupancy	ChIP-seq	Protein-DNA interactions	Johnson et al. 2007; Barski et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007
	ChIP-exo-seq	High-resolution protein-DNA interactions	Rhee & Pugh 2011; Rhee & Pugh 2012
	ChIRP-seq	RNA-DNA interactions	Chu et al. 2011
	CHART-seq	RNA-DNA interactions	Simon et al. 2011
	Chem-seq	Genome-wide localization of small-molecules	Anders et al. 2014
Chromatin interactions	4C	Targeted physical interactions between distant genomic regions	Dostie et al. 2007
	5C	Targeted physical interactions between distant genomic regions	Bau et al. 2011; Umbarger et al. 2011;
	Hi-C	Physical interactions between distant genomic regions	Lieberman-Aiden et al. 2009; Umbarger et al. 2011
	ChIA-PET	Protein-mediated interactions between distant genomic regions	Fullwood et al. 2009; Li et al. 2010; Handoko et al. 2011; Li et al. 2012
Open chromatin	DNAse-seq	DNAse accessible regions	Hesselberth et al. 2009; Song et al. 2011; Boyle et al. 2011
	FAIRE-seq	Shearing-susceptible open chromatin	Gaulton et al. 2010; Song et al. 2011
	Sono-seq	Shearing-susceptible open chromatin	Auerbach et al. 2010
	ATAC-seq	Transposition-mediated mapping of accessible chromatin	Buenrostro et al. 2013
	DGF	Digital Genomic Footprinting	Neph et al. 2012
	DNAse-FLASH	Fragment-length analysis of DNAse hypersensitivity	Vierstra et al. 2014
Replication timing	Repli-seq	Newly replicated DNA	Hansen et al. 2010
DNA methylation	RRBS	Reduced representation bisulfite sequencing	Meissner et al. 2008
	BS-seq	Whole-genome bisulfite sequencing	Lister et al. 2008; Lister et al. 2009
	PBAT	Whole-genome bisulfite sequencing	Miura et al. 2013
	MeDIP-seq	Methylation-enriched regions	Down et al. 2008
	MethylCap-Seq	Methylation-enriched regions	Brinkman et al. 2010
	oxBS-seq	Mapping of sites of 5-hydroxymethylcytosine methylation	Booth et al. 2012
	TAB-seq	Mapping of sites of 5-hydroxymethylcytosine methylation	Yu et al. 2012
Transcriptomics	RNA-seq	Various long transcripts	Mortazavi et al. 2008; Nagalakshmi et al. 2008; Sultan et al. 2008; Wilhelm et al. 2008; Marioni et al. 2008
	Small RNA sequencing	Small RNA species	Ruby et al. 2006; Brennecke et al. 2007
	CAGE	Capped 5’ ends of transcripts	Kodzius et al. 2006; Balwierz etal. 2009; Plessy et al. 2010

Continued on next page

Table 14.1 – *Continued from previous page*

Group of assays	Assay	Detection of / Description	References
Transcriptomics	3P-seq, PAS-seq, MAPS, PolyA-seq	Polyadenylation sites	Jan et al. 2011; Yoon et al. 2010; Derti et al. 2012; Fox-Walsh et al. 2011; Shepard et al. 2011
	RNA-PET	Paired 5' and 3' transcript ends	Fullwood et al. 2009; Ruan & Ruan 2012
	PARE	Endonuclease Degradation products	German et al. 2009
	GRO-seq	Global Nuclear Run-On products	Core et al. 2008
	NET-seq	Nascent RNA molecules	Churchman et al. 2011
	RAMPAGE	Promoter mapping	Batut et al. 2013
	PARE-seq	Mapping of RNA ends	German et al. 2008
	TIF-seq	Mapping of transcript ends	Pelechano et al. 2013
	PEAT	Transcription initiation	Ni et al. 2012
Single–cell transcriptomics	CEL-seq	Single-cell RNA-seq	Hashimshony et al. 2012
	SMART-seq	Single-cell RNA-seq	Ramsköld et al. 2012
	SMART-seq2	Single-cell RNA-seq	Picelli et al. 2013
	STRT	Single-cell RNA-seq	Islam et al. 2011
	Quartz-seq	Single-cell RNA-seq	Sasagawa et al. 2013
RNA-protein interactions	HITS-CLIP	UV cross-linked protein-RNA interactions	Licatalosi et al. 2008; Chi et al. 2009
	PAR-CLIP	UV cross-linked protein-RNA interactions	Hafner et al. 2010
	iCLIP	UV cross-linked protein-RNA interactions	König et al. 2010
	RIP-seq	RNA coimmunoprecipitated with proteins	Zhao et al. 2010
RNA-RNA interactions	CLASH	Mapping RNA-RNA interactions	Kudla et al. 2011
RNA modifications	MeRIP-seq	Mapping of RNA methylation sites	Meyer et al. 2012
	ICE	Mapping A-to-I RNA editing sites	Sakurai et al. 2010
RNA structure	PARS	Genome-wide RNA structure determination	Kertesz et al. 2010
	FRAG-seq	Genome-wide RNA structure determination	Underwood et al. 2010
	SHAPE-seq	Targeted RNA structure determinations	Lucks et al. 2011
	HRF-seq	Determination of RNA accessibility	Kielpinski & Vinther 2014
Ribosome profiling	Ribo-seq	Genome-wide mapping of ribosome occupancy	Ingolia et al. 2009
High-throughput functional assays	Massively parallel functional assays	Simultaneous measurements of the enhancer activity of very large number of constructs	Patwardhan et al. 2012; Melnikov et al. 2012
	STARR-seq	Genome-wide measurement of enhancer activity	Arnold et al. 2013

the same time improved analytical tools tailored to the now well-understood specifics of the data coming from the major platforms have been developed. As a result, human genome resequencing and the study of human genetic variation, cancer genomics, the *de novo* assembly of newly sequenced genomes and metagenomics are now also thriving fields currently mostly based on NGS technologies (1000 Genomes Project Consortium 2010; 1000 Genomes Project Consortium 2012; Mardis et al. 2010; Gnerre et al. 2011; Human Microbiome Project Consortium 2012; Garraway & Lander 2013; Bradnam et al. 2013; Gilbert & Dupont 2011; Lappalainen et al. 2013; Khurana et al. 2013; i5K Consortium 2013). However, the reads generated by these platforms are still short enough to present considerable difficulties in the analysis and interpretation of data, genomes assembled *de novo* from short-read data are still highly fragmented and incomplete (Alkan et al. 2011; Koboldt et al. 2010; Earl et al. 2011; Bradnam et al. 2013), and the cost of sample preparation and the computational infrastructure investments needed to generate and handle the data are still substantial. This has provided an incentive for further sequencing technology development that is both even cheaper and in the same time improves on the current inadequacies of NGS data. As a result "third-generation" sequencing (TGS) technologies are currently emerging. There is some debate whether the term "TGS" should even be used at this point given that there is much less of a sharp divide between these technologies and the NGS technologies compared to the paradigm shift relative to Sanger sequencing that NGS platforms triggered. I will nevertheless still use it here for simplicity, but I will define what exactly I mean by it first. TGS sequencing delivers much longer reads that, in contrast with most NGS technologies, originate from single founder nucleic acid molecules and not from amplified clones (i.e. single-molecule sequencing). The long read lengths promise to greatly simplify and improve *de novo* genome assembly, the study of genomic structural variation and metagenomics, but they also have the potential to once again transform the practices of some areas of functional genomic research. In the same time their single-molecule nature comes at the cost of lowered accuracy. Here, I discuss the functional genomic areas in which TGS technologies are expected to have the greatest impact, as well as the areas, which are at this point mature and for which TGS will not provide much benefit over NGS. NGS can therefore be expected to remain dominant for the foreseeable future in these applications. In particular, transcriptomics and the study of DNA methylation are highlighted, and the anticipated requirements towards the characteristics and quality of data necessary for the promised impact to materialize are examined.

14.2 Overview of second generation sequencing technologies

A common feature of most NGS sequencing technologies is the use of clonally amplified clusters of DNA sequences, the sequence of which is read one or several bases at a time using a variety of sequencing-by-synthesis readout strategies that rely on the signal boost due to the presence of large numbers of identical source molecules. This enables the generation of high-quality sequence reads but it has also limited the read lengths that can be achieved as errors accumulate during each synthesis step in different pieces of DNA in a cluster and eventually proper phasing between individual sequences in a cluster is lost. The most successful NGS technology has been the Solexa/Illumina reverse terminator chemistry, and it and 454 will be used to illustrate the common characteristics of NGS platforms. Illumina sequencing is based on the attachment of DNA sequences to complementary primers immobilized on a glass surface, followed by clonal bridge amplification of each sequence in order to form a cluster of identical sequences. Then, these sequences are read one base at a time using sequencing-by-synthesis relying on reversible fluorescent dye-terminator nucleotides differently colored for each base that can be scanned by a high-resolution microscope after addition, then cleaved off and another based added. This provides high-quality sequencing reads with very low error rates, with errors mostly consisting of base-pair substitutions. The HiSeq incarnation of the technology was initially capable of generating more than three billion individual reads of lengths longer than 100bp, but with subsequent improvements this has now increased to up to 2x250, and 2x500 reads have been generated on the MiSeq platform. The 454 technology was[1] based on the clonal amplification of individual DNA sequences within emulsion droplets containing beads with primers attached to them. Single beads are placed inside the wells of an optic chip and sequencing relies on adding one of the four nucleotides, one at a time; when a nucleotide is incorporated by DNA polymerase to a complementary position in the template, an inorganic pyrophosphate is released which is used to determine the identity

[1] As of the time the last edits of this text were put in place, 454 was scheduled to be phased out within a year

of the base. Polymerase will proceed adding nucleotides over stretches of multiple instances of the same base pair; the resulting signal scales linearly with the length of such homopolymers only up to a point, and as a result indels are the major source of errors with this technology (and other technologies where multiple bases are read at a time). The reads generated by 454 are longer than those generated by most other NGS platforms, even approaching the length of Sanger reads, however their number was always limited (to around a million at most), which limited its applications. In addition indel errors are more difficult to deal with during read alignment and assembly than base substitutions.

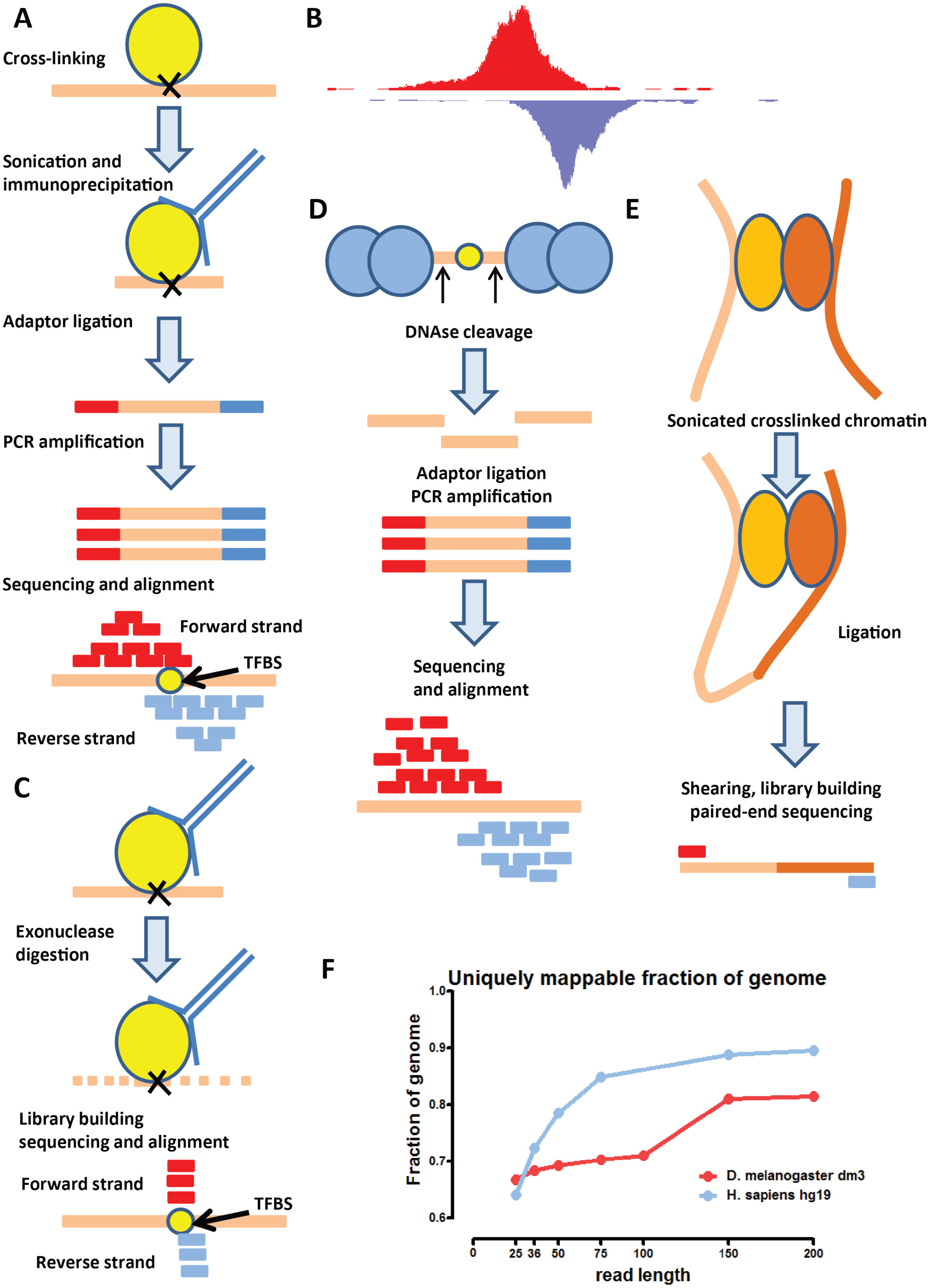
14.3 Third-generation sequencing technologies

Newer sequencing technologies continue to be constantly developed, and some of them are very similar to the NGS strategies outlined above in their characteristics (such as the most recent newcomer to become established on the market, Ion Torrent). The defining features of TGS technologies can be summarized as much longer read lengths combined with the ability to sequence single DNA molecules rather than multiple clones in clusters (Schadt et al. 2010). The first single-molecule sequencing platform was Helicos; however reads generated by Helicos were very short (Harris et al. 2008; Ozsolak et al. 2010; Orlando et al. 2011) and had high error rates; as a result (together with some other undesirable properties of the instrument) Helicos is, at the time of the writing of this text, largely a footnote in the history of sequencing. Single-molecule sequencing has the benefit of much simplified library preparation, which eliminates a lot of the representation biases and artifacts introduced into current sequencing libraries; however it comes at the cost of much increased error rates as reading the sequence of single molecules accurately is considerably more challenging than reading out massively amplified clonal populations.

Two companies have so far presented commercially available TGS technologies: Pacific Biosciences (PacBio) and Oxford Nanopores (although data from the latter has yet to be published[2]). PacBio's SMRT (Single-Molecule Real Time) sequencing technology is based on anchoring individual DNA polymerase molecules and DNA templates into Zero-Mode Waveguide (ZMW) nanowells, and then observing the incorporation of fluorescently labeled nucleotides in real time (Eid et al. 2009). It is possible to do so with single molecules because ZMW wells are smaller than the wavelength of visible light which cannot enter the bottom of the well; by illuminating the well from the bottom, only the bottom volume of the well is visible. Fluorescently labeled nucleotides diffuse very fast in and out of the well but when incorporated by DNA polymerase (an orders-of-magnitude slower process than diffusion), they are held in the well for much longer, which enables the identification of the DNA sequence. The technology allows for the generation of read lengths greater than Sanger sequencing (up to several and even tens of kilobases; read length is limited by the lifetime of the polymerase molecule which is degraded by the laser light used to read fluorescent nucleotides). Its limitations include the high error rate (up to 15%) and the at present low number of sequencing reads generated (a single SMRT cell only generates several tens of thousands of reads). Error rates can be improved by generating circularized single-stranded templates which can be sequenced several times to derive a consensus (Travers et al. 2010); this, however, comes at the cost of decreasing the effective read lengths.

Nanopore sequencing is a very promising approach towards sequencing nucleic acids (and, potentially, other biological heteropolymers too), based on the characteristic changes in electric current through a nanopore situated in an impermeable membrane that are induced by different nucleotides passing through it; as each base passes through the pore, the current changes in a way that is unique to that base allowing its identification (Branton et al. 2008; Manrao et al. 2010; Cherf et al. 2012; Manrao et al. 2012). The method was first proposed about two decades ago (Kasianowicz et al. 1996; Deamer et al. 2000); however, building a working sequencer has been a major challenge as simple electrophoresis of DNA through a nanopore occurs too fast for the sequence to be read, which has necessitated the development of methods to slow down the rate of translocation through the pore. The commercial launch of such an instrument has finally been announced

[2] As of late April 2014


A
Cross-linking
Sonication and
immunoprecipitation
Adaptor ligation
PCR amplification
Sequencing and alignment
Forward strand
TFBS
Reverse strand
B
C
Exonuclease
digestion
Library building
sequencing and alignment
Forward strand
TFBS
Reverse strand
D
DNAse cleavage
Adaptor ligation
PCR amplification
Sequencing
and alignment
E
Sonicated crosslinked chromatin
Ligation
Shearing, library building
paired-end sequencing
F
Uniquely mappable fraction of genome
Fraction of genome
1.0
0.9
0.8
0.7
0.6
read length
0
25
36
50
75
100
150
200
D. melanogaster dm3
H. sapiens hg19

by Oxford Nanopores Technologies in the last two years. However, actual sequence data generated by nanopores is still not publicly available, and therefore key details regarding error rates (which, due to the single-molecule nature of the method, are certain to be significantly higher than Illumina sequencing, but possibly lower than those of SMRT sequencing) and the cost of generating a given number of reads remain unknown. Still, nanopore sequencing holds the long-term promise of delivering reads that are tens or even hundreds of kilobases long, with minimal to no sample prep, little to no sensitivity to the fragment length distribution of the input library (a limitation of both Illumina and PacBio instruments, which do not sequence short and long fragments with the same efficiency), and further in its development, the direct identification of modified nucleotides and eventually direct RNA and even protein sequencing. Its characteristics make it a particularly attractive candidate for being the next transformative technology in the sequencing world, if it delivers on its promises.

Below, the expected impact of these TGS technologies on the different subfields of functional genomic research is reviewed in the context of the experimental and analytical chal-

Figure 14.1 *(preceding page)*: Functional genomic assays for measuring chromatin occupancy, openness and interactions. (A) In ChIP-seq, proteins are crosslinked to DNA, chromatin is sonicated down to fragments of at most 300-400bp in size, and immunoprecipitated with an antibody against the protein of interest. The resulting set of DNA sequence fragments is then converted into a sequencing library and sequenced. (B) A characteristic asymmetric distribution of reads on the forward and reverse strand around the occupancy site is observed, with the distance between the peaks on each strand corresponding to the average fragment length. (C) In ChIP-exo-seq, the high-resolution modification of ChIP-seq, crosslinked fragments are subjected to 5'-to'-3' λ exonuclease treatment; the exonuclease is processive but is blocked by the site of crosslinking. As a result, the 5' ends of sequencing fragments in the final library are very highly enriched immediately around the site where the protein of interest is crosslinked to DNA. (D) DNAse-seq and its variations are based on the high sensitivity of DNA that is not protected by nucleosomes to DNAse cleavage. The resulting DNA fragments are then converted into libraries and sequenced. (E) Assays measuring chromatin interactions rely on the fact that such interactions are mediated by proteins; crosslinking of DNA to proteins and of proteins to proteins leads to the formation of complexes in which the ends of DNA fragments originating from distant genomic locations are brought in close physically proximity and can be ligated to each other (of course, so can be the ends of each fragment on its own, and this is a major source of noise in the final libraries). The ligation products are then subjected to further processing (with the details varying on the protocol) with the end result being the generation of chimeric DNA fragments each end of which originates from one of the interacting genomic loci. These fragments are then sequenced in a paired-end format. Note that in all these assays the size of the fragments being sequenced is small (a few hundred base pairs at most), and their short length is actually important to the resolution of the assay. (F) The fraction of the human and *Drosophila melanogaster* genomes that is uniquely mappable at different read lengths. Mappability was evaluated as follows: for each read length r, a set of "sequencing reads" was generated by creating one such read starting at each position in the genome. The reads were then mapped to the genome using Bowtie (version 0.12.7; Langmead et al. 2009) while retaining only unique alignments, and read coverage C (in raw read counts) was calculated for each position in the genome. The mappable fraction of the genome MF_G was then calculated as follows:

$$MF_G = \frac{\sum\limits_{c \in G} |c|}{\sum\limits_{c \in G} \sum\limits_{p=1}^{|c|} I(C_{c,p} \geq r)} \tag{14.1}$$

where I is the indicator function, c is a chromosomes in the genome, $|c|$ is the length of a chromosome, and p are the individual positions in each chromosome.

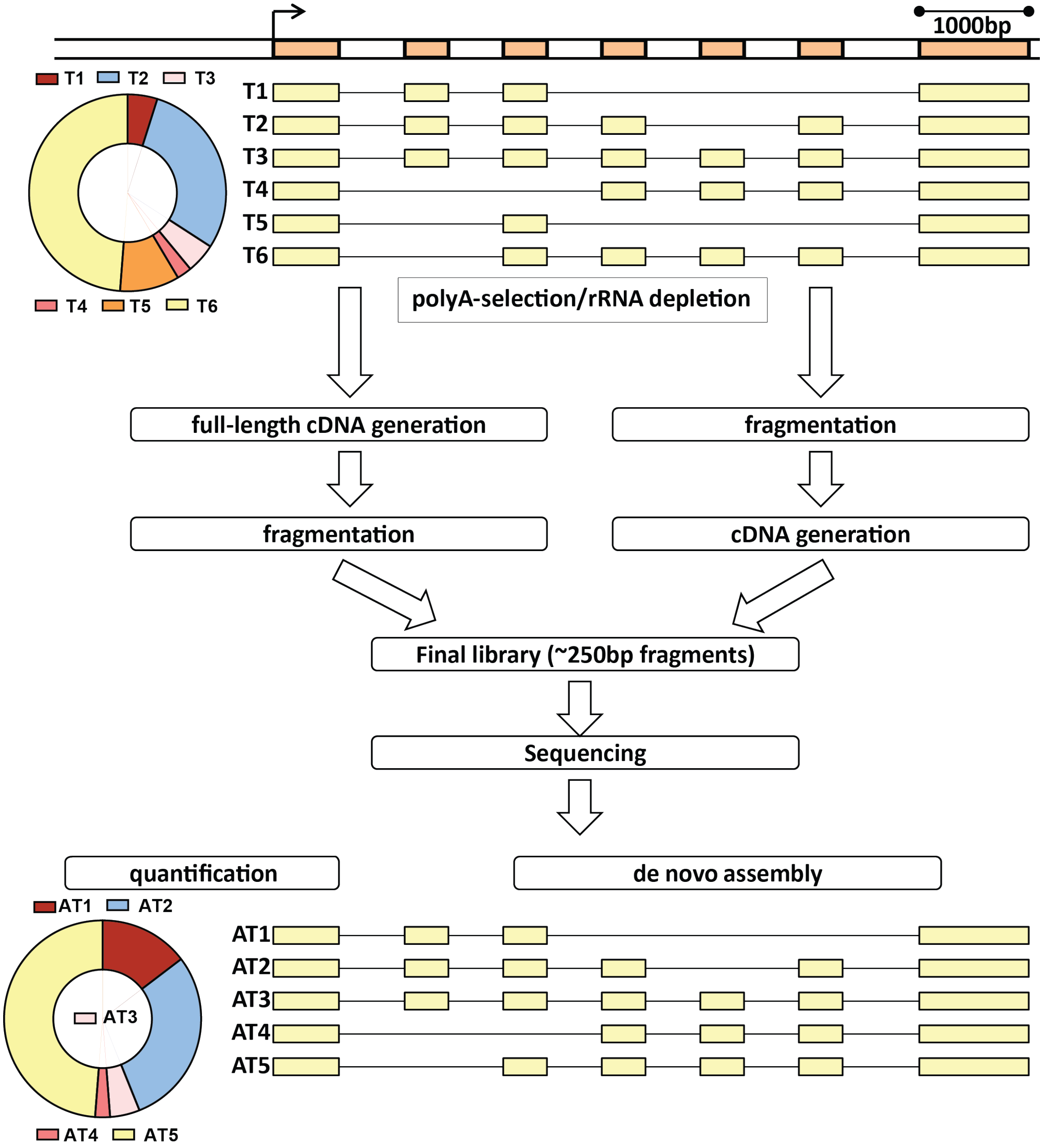


Figure 14.2: General strategies for contemporary RNA-seq measurements of the transcriptome. A hypothetical gene expressing six different alternative transcripts (T1 to T6) in the relative ratios indicated in the pie chart is shown. The input RNA may first be polyA-selected or rRNA-depleted; the transcribed mRNAs are then either subjected to random fragmentation (as in the original protocol described in Mortazavi et al. 2008), and then converted to cDNA (using, for example, random priming). Alternatively, the mRNAs can be converted to full-length (to the extent the input mRNA is full-length and the reverse transcription reaction proceeds to completion) cDNA molecules (such as in the SMART-seq protocol; Ramsköld et al. 2012) and then fragmented. In either case, a final library of fragment size usually in the 150–350bp range is generated, much shorter than the length of the original transcripts. The transcripts have to be then assembled and/or quantified using probabilistic methods, which does not always return results true to the original biological reality. In this case, this is illustrated by following the approach adopted by Cufflinks (Trapnell et al. 2010) and assembling the minimum number of transcripts that can explain the data (assembled transcripts AT1 to AT5), which, however, results in the loss of one transcript and not fully accurate isoform-level quantification.

lenges that the field at present faces

14.4 Functional genomics assays and third-generation sequencing

14.4.1 ChIP-seq and derivatives for the measurement of genomic occupancy

Our current understanding of gene regulatory mechanisms revolves around the extremely complex interplay between the binding of sequence specific transcription factor to regulatory elements in the genome (in the immediate vicinity of promoters of genes or to enhancer sequences located very far upstream or downstream of promoters), which affects transcription by the recruitment or inhibition of the transcriptional machinery and the induction of changes in the chromatin state, mainly covalent modifications on histone tails nearby (Kouzarides 2007). In the same time, chromatin state also influences transcription factor binding, with, for example, many transcription factors being unable to bind to chromatin in closed inactive conformation (Zaret & Carroll 2011). Thus measuring the genomic location of binding events of transcription factors and chromatin modifying enzymes, and the distribution of histone modifications, in diverse cell types and conditions, is of critical importance for full understanding of the process of gene regulation.

ChIP-seq is at present the standard tool for accomplishing this task. As shown in Figure 14.1, a ChIP-seq experiment begins with the chemical cross-linking of proteins bound to DNA, shearing the cross-linked chromatin to size of a few hundred bp at most (typically below 200), immunoprecipitating the DNA fragments bound to the protein of interest, reversing cross-links, and building a sequencing library by ligating sequence adapters and PCR amplification; a parallel library is built from crosslinked chromatin without immunoprecipitation for comparison and normalization purposes when calling binding sites. Usually, a short tag (initially 36bp, later 50bp, with longer read lengths of 1x100 or even 2x100 increasingly common now) from only one end of the DNA fragment is sequenced and aligned to the genome. Crucially, because adapters are ligated only in one direction relative to the original genomic strands of the fragment and the length of fragments is variable, reads mapping to the forward and reverse strands distribute in a characteristic asymmetric way around the position where the target protein binds to DNA (if the protein binds to specific locations in the genome in a sequence specific manner; elongating RNA polymerase and histone marks spread along large genomic domains do not exhibit that behavior), and this information is used to more precisely define transcription factor binding sites Figure 14.1A and B) (Kharchenko et al. 2008). Derivatives of the ChIP assay have been developed that aim at identifying the binding sites of RNA molecules – ChIRP-seq (Chu et al. 2011) and CHART-seq (Simon et al. 2011) – as well as the chromatin occupancy of small molecules (Chem-seq; Anders et al. 2014).

So far, a significant limitation in the practice of ChIP-seq has been the bottleneck created by the process of performing the ChIP reaction, which has traditionally been slow, tedious and low-throughput. Automated robotic protocols for carrying it out have now been developed (Aldridge et al. 2013; Gasper et al., in press), and coupled with the automation of library generation promise to enable a major increase in throughput, allowing up to 96 samples to be efficiently processed in the same time (although it should be noted that even then there will still be a bottleneck in the workflow, one that will remain for the foreseeable future: the crosslinking and sonication steps; unless very large amounts of chromatin from the same source are analyzed, large numbers of samples will still have to be crosslinked and fragmented manually).

The other area where improvements are needed in the ChIP-seq assay is achieving truly single base-pair resolution. Recently, the ChIP-exo-seq variation of ChIP-seq has been developed, which addresses this issue by combining ChIP with 5'-to-3 λ exonuclease digestion of the crosslinked fragments. Exonuclease processivity is blocked by the site of the crosslink thus providing precisely phased sequencing ends right around the protein-DNA interaction site, with higher resolution than traditional ChIP-seq, especially in regions where closely spaced binding of multiple transcription factor molecules occurs (Rhee & Pugh 2011; Rhee & Pugh 2012; Figure 14.1C). Because improving resolution is the key area of further development of the assay, the longer reads generated by TGS platforms will not be of much advantage in ChIP-seq. Sequencing longer reads can improve the ability to detect binding events over a larger fraction of the genome as it will make a more of the genome uniquely mappable (Figure 14.1F shows the uniquely mappable fraction of the human and fly genomes as a function of read length). However, first, the length of reads can only be as long as that of the input fragments, and shorter fragments usually lead to better resolution (Figure 14.1B), and second, Illumina read lengths, especially in the paired end format, are already covering the range of fragment sizes observed in a typical ChIP-seq library. In addition, large numbers of reads are necessary for the comprehensive identification of transcription factor binding sites (in the tens of millions of reads for mammalian-sized genomes; Landt et al. 2012), and even larger numbers are optimal for broad-source histone marks (Jung et al. 2012); at present, second-generation sequencing technologies are comfortably delivering that many reads, and of very high quality too. However, if true single molecule sequencing with no library preparation that can generate a very large number of reads of comparable quality becomes available at acceptable cost, it would eliminate the need for PCR amplification together with the various biases introduced by it and it would enable working with very small amounts of samples. This is currently challenging (Shankaranarayanan et al. 2011; Adli et al. 2011), but has great potential importance to provide insight into the working of rare cell types in the body.

14.4.2 DNAse hypersensitivity and other open chromatin assays

Active regulatory elements in the genome (enhancers and promoters) are characterized by increased chromatin accessibility. This property can be used in order to identify them: increased chromatin accessibility manifests itself as elevated susceptibility to DNAse cleavage. DNAse I hypersensitivity mapping has been used for decades to study individual loci (Maniatis & Ptashne 1973), and paired with NGS technologies has allowed the genome-wide detection of DNAse hypersensitive sites by sequencing the resulting DNA ends (Hesselberth et al. 2009; Song et al. 2011; Boyle et al. 2011; Figure 14.1D). Other methods for identifying open chromatin regions rely on the preferential segregation of open chromatin regions into the aqueous phase when cross-linked chromatin is phenol-chloroform extracted (FAIRE-seq and Sono-seq; Gaulton et al. 2010; Song et al. 2011; Auerbach et al. 2010). For all of these methods, resolution and depth of sequencing is a key consideration. Indeed, sequencing DNAse I digested chromatin to a depth of nearly half a billion reads yields high-resolution maps of individual transcription factor footprints (and is even labeled separately as Digital Genomic Footprinting, or DGF; Neph et al. 2012a; Neph et al. 2012b), and more recently, analysis of the different fragment lengths produced by DNAse digestion (DNAse-FLASH; Vierstra et al. 2014) has proved very useful for understanding nucleosome architecture and transcription factor binding in detail. For these reasons, similar reasoning to the one outlined above for ChIP-seq applies regarding the utility of third generation sequencing technologies.

14.4.3 Mapping long-range chromatin interaction

Key components of eukaryotic gene regulatory networks are enhancer elements, regulatory sequences located far away from the promoters of the genes they regulate. ChIP-seq can identify potential enhancers but as these elements can be located very far away from their target genes, even “skipping” over one or multiple genes (Lettice et al. 2003), it is usually not possible to assign an enhancer to its corresponding promoter (or promoters) with absolute certainty. The relationship between enhancers and promoters is far from the only known type of long-range physical interactions between genomic elements; in recent years, appreciation for the dynamic 3D structure of the nucleus has been steadily growing, and structures as transcriptional factories that bring multiple genes in close genomic proximity have been proposed. Identification of these long-range interactions is of major importance for understanding the biology of the nucleus and the logic of gene regulation.

The chromosome conformation capture (3C) technique was the first one developed to tackle this issue and to test the interaction between two candidate genomic loci (Dekker et al. 2002). The advent of NGS technology has allowed to develop derivatives of 3C that measure interactions between large sets of candidate loci (4C and 5C; Bau et al. 2011; Umbarger et al. 2011), or on a fully genome-wide scale (Hi-C; Lieberman-Aiden et al. 2009; Umbarger et al. 2011), while the ChIA-PET assay measures long-range interactions mediated by a particular protein (Fullwood et al. 2009; Li et al. 2010; Handoko et al. 2011; Li et al. 2012). These assays rely on the chemical crosslinking of protein-mediated interactions between distant genomic loci, the subsequent shearing of chromatin and the ligation of the ends held together by the proteins under dilute conditions so that ligation between DNA ends in different complexes floating in solution is prevented (Figure 14.1E). After library building, short reads are generated from both ends of the resulting chimeric fragments and aligned to the genome. These fragments are once again short, sometimes extremely short (in the case of the original ChIA-PET protocol, only very short stretches of sequences on each end are informative due to the use of Type IIS restriction enzymes during library building), it is not expected that TGS technologies will initially have a great impact in this field.

14.4.4 Mapping DNA methylation genome-wide

Numerous modifications of DNA bases playing a biological role have been described, especially in prokaryotes and single-cell eukaryotes (Mruk & Kobayashi 2014; Gommers-Ampt et al. 1993; van Luenen et al. 2012).The one that has attracted the most attention in multicellular eukaryotes, due to its role in epigenetic regulation, is the methylation of the 5 position of cythosine (5mC), particularly in the context of CpG dinucleotides (Bird 1986; Fuks 2005; Miranda & Jones 2007). While it was first identified many decades ago (Wyatt & Cohen 1952), in the last few years 5-hydroxymethylcytosine (5hmC) has also begun to emerge as a biologically important modification (Kriaucionis et al. 2009; Tahiliani et al. 2009; Guo et al. 2011). The classical role of 5mC in mammalian systems is in the CpG context in promoter-associated CpG islands. The methylation of a CpG island is associated with the repression of gene expression from the associated promoter (Fuks 2005; Miranda & Jones 2007)), which is of vital importance during embryonic development, for the establishment of imprinted loci and cancer progression, among many other processes. In addition to this classical view, genome-wide profiling of the modification in both mammalian systems and in other clades of the tree of life has revealed a much more complex picture involving methylation over gene bodies and in non-CpG contexts (Lister et al. 2008; Lister et al. 2009; Zemach et al. 2010; Huff & Zilberman 2014; see also an extensive discussion on the topic in the final chapter).

Two general strategies exist and have been in wide use for profiling DNA methylation: enrichment for methylated DNA and bisulfite sequencing. Enrichment methods rely on immunoprecipitation with antibodies specific for 5mC (MeDIP; Weber et al. 2005) or on enrichment using the methyl binding domains (MBD) of naturally occurring proteins (MethylCap; Cross et al. 1994). Both methods can be coupled with NGS sequencing and the nature of the data generated is similar to that of ChIP-seq (Down et al. 2008; Brinkman et al. 2010). The drawback is that they do not provide single base-pair resolution of methylation events but only enrich for regions with elevated methylation levels. Base-pair resolution is provided by bisulfite sequencing (BS; Frommer et al. 1992; Clark et al. 1994). Treatment of DNA with bisulfite converts unmethylated cytosine to uracil but leaves 5mC unaffected; as a result 5mC is sequenced as cytosine while unmethylated cytosine as thymine. The resulting libraries can be sequenced and aligned against the genome and methylation levels assessed at the level of individual base pairs. As the cost of sequencing whole genomes has been until recently prohibitively high for routine whole-genome BS sequencing, reduced-representation bisulfite sequencing approaches (RRBS) have been developed; in RRBS, restriction enzymes are used to cleave specific positions in the genome and the methylation status of the surrounding nucleotides is assessed after bisulfite conversion and sequencing (Meissner et al. 2008). With decreasing costs, whole-genome BS-seq is becoming more widely used even in mammalian systems (Lister et al. 2009; Lister et al. 2011).

BS-seq assays present considerable analytical challenges due to the nature of methylation events and bisulfite conversion. Alignment of BS-seq sequencing reads is a non-trivial informatics problem with numerous trade-offs between sensitivity and specificity that have to be made as a result of the conversion of cytosines to thymine, the potential for heterogeneity of methylation events between CpGs in close proximity to each other, and a number of other issues (Krueger et al. 2012). In addition, bisulfite treatment does not differentiate between 5mC

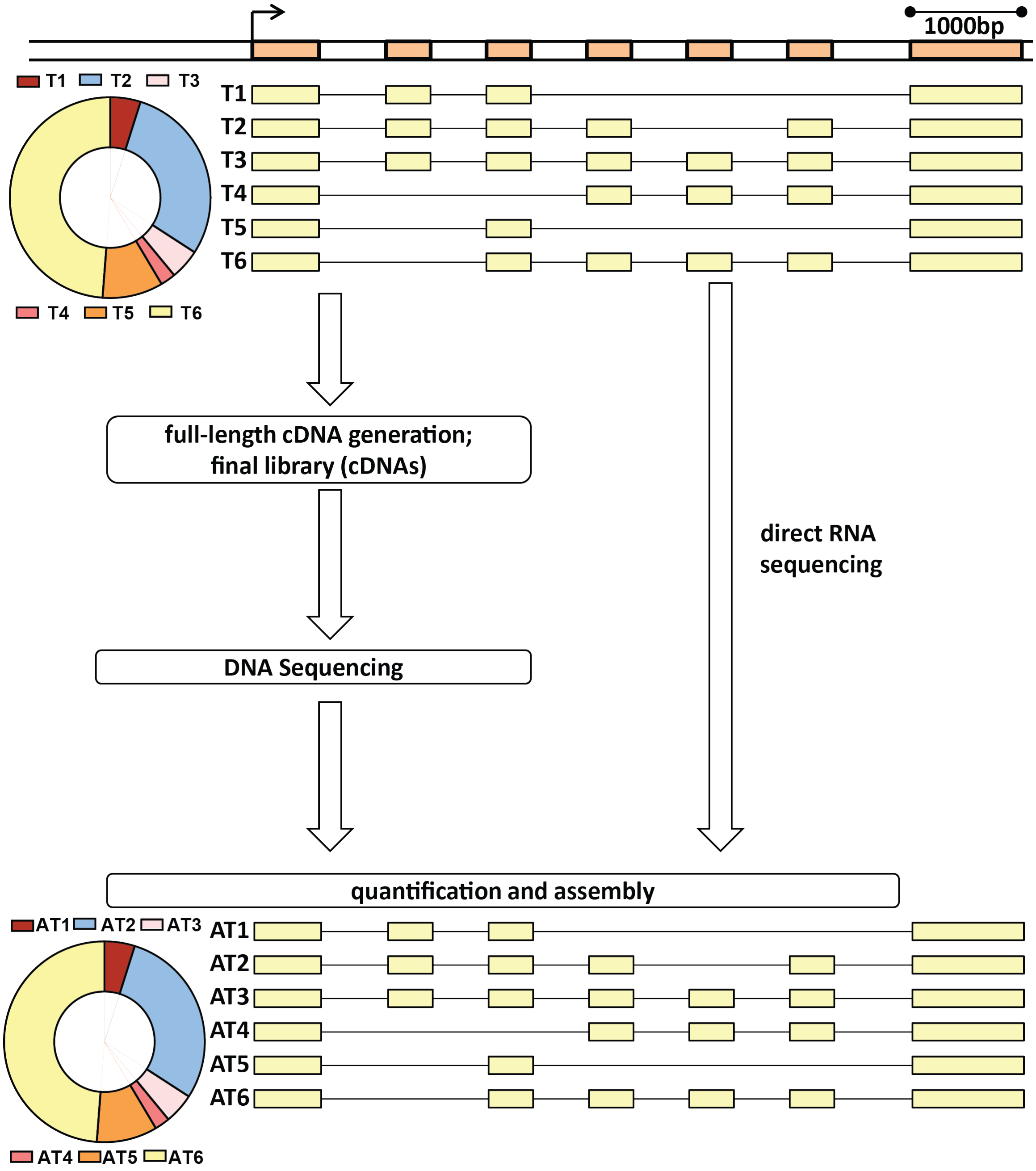


Figure 14.3: Future long-read RNA-seq format. The same gene shown in Figure 14.2 is depicted here too. If the appropriate sequencing technology is available, RNA can be converted into cDNA and the cDNA directly sequenced (preferably without amplification, if possible). An even better option would be to sequence RNA directly, which is in principle possible with nanopore sequencing but still some way from becoming a commercially available reality. Note that the sequencing has to be carried out to a sufficiently high depth for results to be representative for all genes in the dynamic range of the transcriptome (meaning tens of millions of reads should be generated). The problems of transcript assembly and transcript-level quantification become greatly simplified and likely actually solvable in the great majority of cases with data of this kind, unlike the insurmountable computational and epistemological challenges presented by current datasets.

and 5hmC and as a result additional assays have had to be developed to measure its levels. Finally, it is at present difficult to examine the phasing of methylation events between maternal and paternal chromosomes due to the short nature of NGS sequencing reads.

TGS technologies promise to deliver a solution to many of these issues by avoiding bisulfite conversion and reading methylation events directly over long stretches of DNA. The ability of the PacBio platform to directly detect 5mC has been demonstrated based on the characteristic delay in nucleotide incorporation by the polymerase at 5mC positions (Flusberg et al. 2010; Clark et al. 2012). Both 5mC and 5hmC have also been shown to induce characteristic changes in the current through nanopores (Clarke et al. 2009; Wallace et al. 2010), thus potentially providing a way to directly read methylation events over very long DNA sequences with very minimal sample preparation (and correspondingly lower cost), potentially even from single cells. The long reads are advantageous because they will allow the reliable allelic phasing of methylation status, which is at present very difficult with short reads. If error rates can be sufficiently minimized, these technologies could enable us to dive much deeper into the detailed workings of the epigenome than currently possible.

14.4.5 Transcriptomics

The area of functional genomics where TGS technologies can be expected to have the greatest impact is transcriptomics. The interaction between the immense complexity of the transcriptome, the short length of current sequencing reads and the limitations of library building protocols and sequencing platforms has presented some very difficult analytical challenges to the field, which longer reads should be able to address if generated in sufficient numbers.

14.4.5.1 Long RNA molecules and RNA-seq

The primary tool for characterizing transcriptomes today is RNA-seq. A typical RNA-seq experiment aims at measuring mRNA molecules

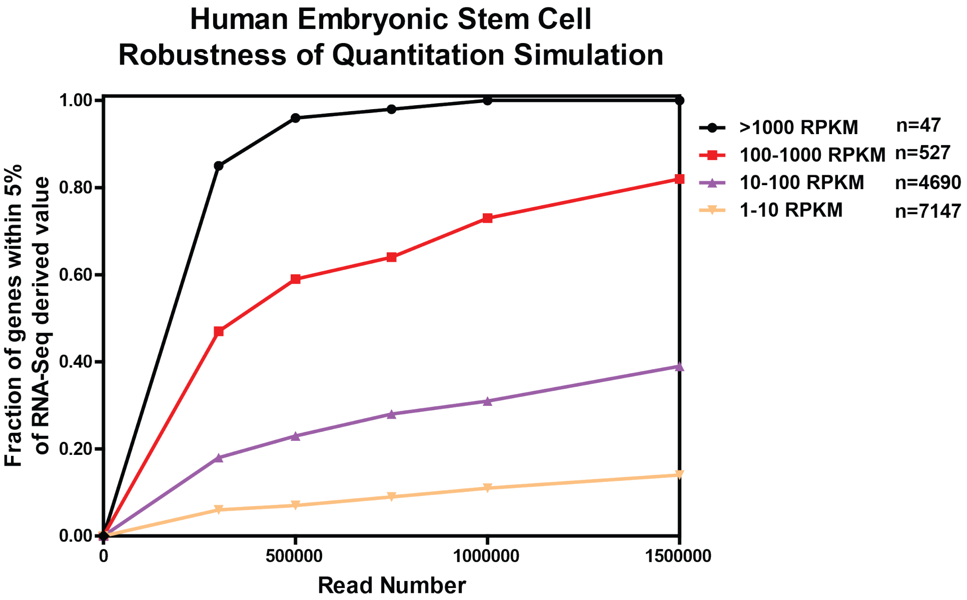


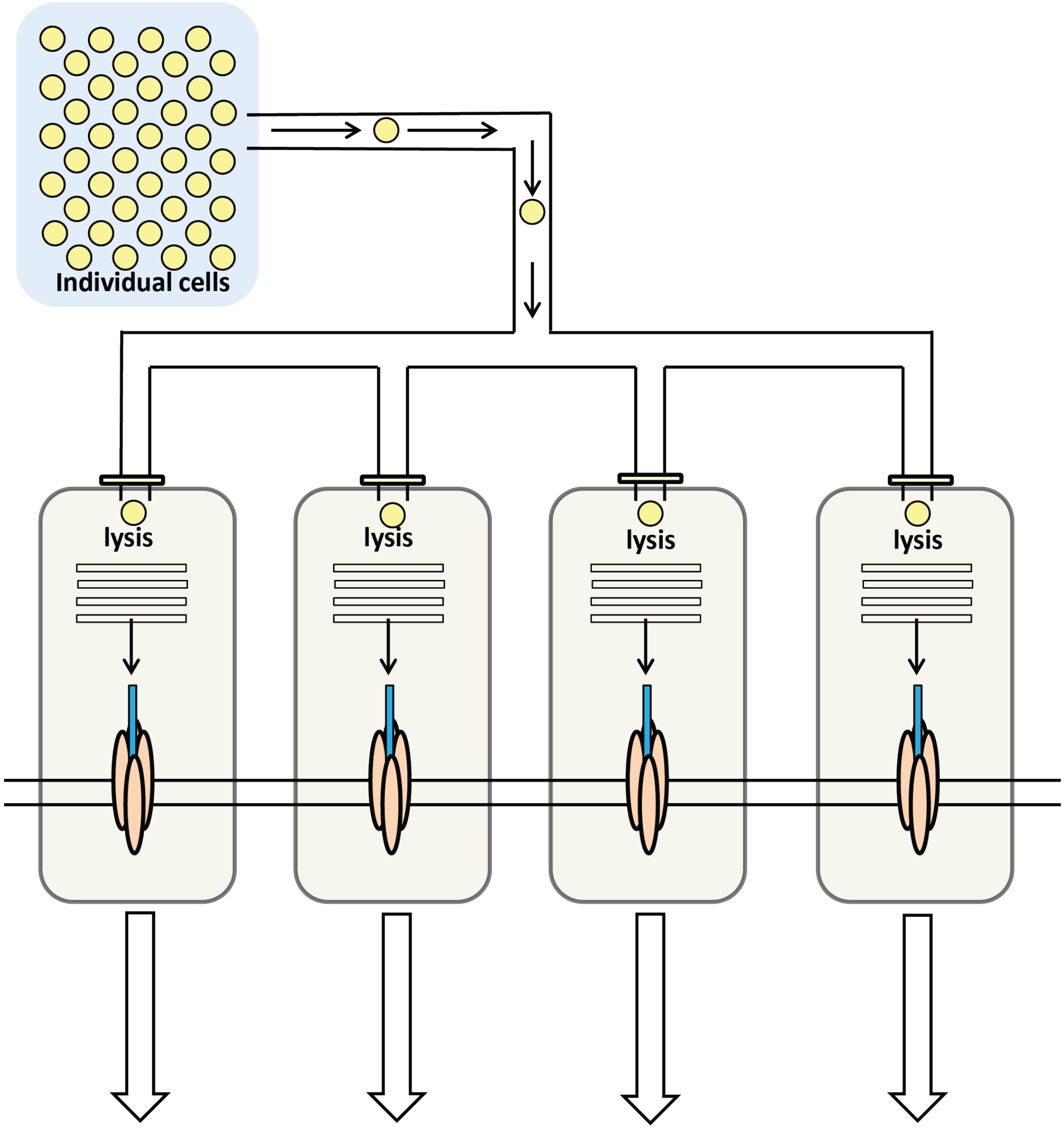
Figure 14.4: Robustness of long-read RNA-seq to sequencing depth. Gene-level FPKMs for the H1-hESC cell line (2x75bp ENCODE data from the Wold lab) were used a starting point. Assuming the relative FPKM abundances correspond to real relative abundances, a long-read transcriptome was simulated as follows. First, the FPKM for each gene was multiplied by 10×10^4. Then the resulting transcriptome was sampled at different sequencing depths, assuming that 1 long read corresponds to 1 transcript. Gene-level expression values were calculated in TPM (Transcripts Per Million transcripts sequenced), and the fraction of genes expressed at different FPKM levels (upper right) that were quantified within 5% of their original relative abundance was calculated.

and involves the selection of polyadenylated RNAs, their fragmentation to a size usually below 200-300bp, conversion of the fragments to cDNA, PCR amplification and sequencing of the resulting fragments, either from one end or from both ends as paired-end reads. Other protocols may feature alternative sequence of steps (Figure 14.3) but the general principle remains the same: long RNA molecules are converted into much shorter DNA fragments in the final library and then sequenced. Several varieties of the library-building protocol that preserve information about which strand reads originate from (”stranded” RNA-seq; Levin et al. 2010) exist. In addition, while what is most often measured is polyA-selected mRNAs, non-polyadenylated transcripts can also be specifically targeted using various strategies for depleting ribosomal RNA; Chen et al. 2011). Finally, specific very rare transcripts can be specifically captured and subjected to RNA-seq (Mercer et al. 2011).

The resulting datasets contain an enormous amount of information about the transcriptome at a single base-pair resolution (Djebali et al. 2012; Chapter 2 of this thesis). Splicing events can be directly quantified using sequencing reads that cross splice junctions and new splice isoforms can be identified (Katz et al. 2010). Chimeric transcripts resulting from chromosome translocations playing a role in cancer biology can be identified (Levin et al. 2009; Zhang et al. 2010; Kinsella et al. 2011; Kim et al. 2011; Sakarya et al. 20102; Li et al. 2011; McPherson et al. 2011; Levin et al. 2009). RNA-editing events can be cataloged (Li et al. 2011; Peng et al. 2012; Bahn et al. 2012; Park et al. 2012) and expression bias towards the maternal or paternal chromosome can be measured (Rozowsky et al. 2011; Reddy et al. 2012; Chapter 3 of this thesis). New classes of transcripts, such as long intergenic non-coding RNAs (lincRNAs), can be discovered, annotated and quantified (Cabili et al. 2011; Guttman et al. 2010). Finally, newly-sequenced genomes can in principle be annotated *de novo* using RNA-seq data.

While RNA-seq datasets have already provided highly useful insights into all of the above areas, as extensively discussed in Chapters 2 and 3, two very important classes of problems have remained unsolved at a satisfactory level for all biological applications: isoform-level quantification and *de novo* transcript assembly. The ability to faithfully carry out these tasks is of critical importance for the study of alternative splicing, alternative initiation and termination (Lenhard et al. 2012; Sandberg et al. 2008; Jan et al. 2011), and for the accurate annotation of genomes. These are unsolved problems not because of lack of sufficient computational sophistication, but simply because the information needed to solve them in all cases is often simply not present in the data. The median length of annotated mRNAs in the human genome is in the 2-3kb range while the length of sequence reads has only recently approached 150-250b, still far shorter than a full-length mRNA, necessitating the use of probabilistic methods to parse them between all available isoforms, a non-trivial computational problem for which a unique solution not always exists. Not only that, but the situation is posed to worsen as annotations get more and more comprehensive by including more and more alternative isoforms for each gene – the ability of isoform-level quantification algorithms to accurately parse reads between the transcripts of a gene is inversely proportional to the number of isoforms annotated for it. Even if longer (but still shorter than the longest transcripts in the genome) reads were available, it would not be advisable to use them for purposes other than assembly because this would introduce a number of undesirable biases in datasets (see discussion in Chapter 3 for details). Long RNA fragments present more opportunities for the formation of secondary structures, which affect reverse transcription in unpredictable ways and increase coverage non-uniformity, and even if this was not the case, long fragments create representation biases against shorter transcripts.

It has now become abundantly clear that the only viable solution to these problems is to sequence full-length RNAs using long-read TGS technologies, as this will provide the long-range connectivity information that is missing in current RNA-seq datasets and which will allow the accurate assembly and transcript-level quantification of even the most complex loci. This will be possible even with relatively high error rates, as, first, in many cases a reference genome of high quality will be available, and second, hybrid strategies, which combine TGS reads with high-quality Illumina reads, using the latter for error correction (Au et al. 2012; Au et al. 2014) can be used. However, it is still not clear whether all requirements that need to be met for short-read RNA-seq to be displaced by long-read RNA-seq will in fact be satisfied by TGS technologies. The primary ones are sequencing depth and library



	Cell 1	Cell 2	Cell 3	Cell K
G_1T_1	$C_{1,1,1}$	$C_{1,1,2}$	$C_{1,1,3}$	$C_{1,1,k}$
G_1T_2	$C_{1,2,1}$	$C_{1,2,2}$	$C_{1,2,3}$	$C_{1,2,k}$
G_1T_3	$C_{1,3,1}$	$C_{1,3,2}$	$C_{1,3,3}$	$C_{1,3,k}$
……	……	……	……	……
G_nT_m	$C_{n,m,1}$	$C_{n,m,2}$	$C_{n,m,3}$	$C_{n,m,k}$

preparation. First, tens of millions of reads are still going to be needed for accurate quantification even with long reads (Figure 14.4). This is far beyond what is economically feasible with current PacBio output. It is not clear what the throughput of nanopore sequencers is going to be, but it has the theoretical potential to be much higher as the speed of translocation through pores is very fast (indeed the main challenge impeding their development has been how to slow it down). Second, sequencing RNA on the PacBio platform has so far required the partitioning of samples into different length classes, preparing separate libraries for each and then sequencing them separately. Such approach makes quantification of the whole sample pretty much impossible. Therefore, practical RNA long-read sequencing will have to be done on a platform that is not biased towards or against fragment of certain sizes. Once again, this is in theory a characteristic of nanopore sequencing, but it remains to be seen how real-life instruments will operate. Nanopores have one more potential feature, perhaps the most desirable of all, and it is the ability to sequence RNA directly (Ayub & Bayley 2012; Ayub et al. 2013; Cracknell et al. 2013). Working instruments capable of direct RNA sequencing are still some time from being commercially available. However, they are the most promising candidate for delivering what would be perhaps the end point of development of RNA-seq technology: protocols based on direct RNA sequencing would remove all of the enzymatic steps that are sources of various biases in current protocols (such as reverse transcription and PCR), in addition to providing long reads (Figure 14.3).

14.4.5.2 Small RNA sequencing

The sequencing of small RNAs was one of the very first applications of NGS sequencing but ever since most NGS-based small RNA sequencing studies have primarily aimed at identification and annotation of small RNAs rather than quantification and comparison between samples. The reason is that while multiple protocols for building small RNA libraries exist, relying either on ligation or oligonucleotide-tailing, they all introduce very serious representation biases into the final libraries, making it difficult to compare the levels of individual small RNA species (Linsen et al, 2009; Hafner et al. 2011; Toedling et al. 2012). A technology allowing for direct RNA-sequencing would be ideally suited for this problem, and again, nanopore-based platforms could in principle accomplish this, though whether the quality of the data will be sufficiently high to displace current sequencing platforms remains to be seen. Initial steps in this direction have already been reported (Wang et al. 2011; Gu & Wang

Figure 14.5 ***(preceding page)*****: The single-cell RNA-seq of the future**. In addition to the less-then-ideal aspects common to all current RNA-seq protocols, single-cell RNA-seq faces the challenge of maximizing capture efficiency (the probability that each original RNA molecule will be captured and represented in the final library, i.e. single molecule capture probability or p_{smc}). Single cells contain a finite and limited number of founder RNA molecules, and it is vitally important that each and every one of them is counted, and counted just once, if one is to obtain accurate measurements of the transcriptome of each individual cell. Unfortunately, p_{smc} is at present nowhere near 1, and is a source of significant technical stochasticity between individual cells. Sequencing very large numbers of cells (Shalek et al. 2013; Jaitin et al. 2014) can recover common patterns in cell-to-cell diversity in large populations of cells, but it is still highly desirable to overcome the technical stochasticity by generating truly accurate measurements. The best way to achieve that is to eliminate as many of the enzymatic steps between founder RNA molecules and sequencing as possible. Ultimately, this means direct RNA sequencing, which is in principle possible with nanopore sequencing though it still lies quite some time into future. Ideally, it would be incorporated into a microfluidic system, which channels single cells into individual microfluidic chambers, in which they are first lysed, and then their RNA content is passed through nanopores embedded into the wall of the chamber. This would provide digital counts of isoform-level transcript abundances in absolute copies per cell numbers (depicted as $C_{G,T,K}$, where C stands for absolute transcript copy numbers, G for the gene a transcript belongs to, T for the transcript itself, and K for the index of each individual cell), potentially also including the non-polyadenylated portion of the transcriptome (which has received virtually no attention in single-cell transcriptomics so far as all currently available protocols feature a polyA-selection step).

2013; Gu et al. 2013).

14.4.5.3 Single-cell transcriptomics and epigenomics

The vast majority of functional genomic measurements are performed on large populations of cells. This is largely due to the limitations of many of the experimental protocols (for example, it is most likely not possible to perform ChIP-seq on single cells due to the inefficiency of cross-linking) but it has the end result of masking cell-to-cell variation and noise, the presence of distinct subpopulations within the larger population and other very interesting biological phenomena that only manifest themselves when examined on the level of single cells. Ideally, single cell functional genomic measurements would be routinely available, and this has prompted the development of protocols for sequencing both the genomes (Xu et al. 2012; Hou et al. 2012) and the transcriptomes of single cells (Tang et al. 2009; Tang et al. 2010; Tang et al. 2011; Islam et al. 2011; Ramsköld et al. 2012; Marinov et al. 2014). Here is also the place to note that the RPKM normalization widely used for RNA-seq quantification is only a relative measurement of gene expression levels (and so is RNA-seq itself in general in its current form) as it measures the allocation of a given number of reads between genes/transcripts but not the absolute levels of transcript copies per cell. It is in principle possible to obtain estimates of the average number of copies per cell by precisely tracking down the number of cells and amounts of purified RNA that went into a library (Mortazavi et al. 2008) but this information is typically not available and even when it exists, it is no match for actually knowing the number of copies for individual single cells. However, because enormously larger amounts of DNA are needed for sequencing on NGS platforms relative to the amount of RNA present in a single cell, current single-cell transcriptomics protocols all involve massive amplification of fragmented RNA, typically in two rounds of amplification separated by a fragmentation step. As a result, information about the number of transcript copies in a cell is lost, in addition to biases introduced by PCR amplification, reverse transcription, and the stochastically variable capture rate of the original transcripts (p_{smc}, single-molecule capture probability), most of which are present at low copy number per cell to begin with. There are two partial solutions to this problem: first, the use of spike-in standards of know abundance and the subsequent recalibration of FPKMs to estimated copies per cell (Marinov et al. 2014; Islam et al. 2011), and second, the use of unique molecular identifiers that track the number of founder molecules (Shalek et al. 2013; Islam et al. 2014). Both approaches are far from ideal. The first one is not entirely quantitative, as spike-ins are not necessarily present in exactly the fixed number of copies that are on average inputted in each reaction, while all current variations of the second involve the tagging of one end of transcripts and result in the loss of the rest of it, with the corresponding consequences for the ability to analyze alternative splicing and allelic biases on the single cell level. Neither of them resolves the technical stochasticity problem either, which can only be eliminated by eliminating the enzymatic steps in protocols that are its source.

The most viable way, in which such an advance can be achieved, is the direct sequencing of RNA from single cells, using nanopores. Whether and when this will be practically possible remains an open question, but it is the only technology that has the potential to be the basis of a single-cell RNA-seq assay of the kind shown in Figure 14.5. Microfluidics-based single-cell genomics devices are already in widespread use (such as the Fluidigm C_1 system; Shalek et al. 2013; Wu et al. 2014) and proven their usefulness. It is in principle possible to design a microfluidics platform that integrates the sorting of individual cells into microfluidic chambers with nanopores in each of them, capable of direct RNA sequencing of the RNA content of each cell after lysis. This would provide a direct readout of the absolute abundances of all RNA species in a cell, hopefully with minimal bias, resolving most of the technical issues with current protocols and platforms.

Similarly promising are the prospects for the application of TGS platforms to single-cell epigenomics. While assays measuring genomic occupancy and chromatin structure are ill-suited for single-cell measurements, DNA methylation could be well measured on the level of single cells with a very long-read single-molecule sequencing platform. The examination of large numbers of allelically phased single-cell DNA methylation profiles should reveal a great deal about the dynamics and regulation of epigenetic DNA modifications.

There are numerous formidable technical

challenges to be overcome for these prospects to become reality. The concept of nanopore sequencing has been applied in practice in ways that are not directly compatible with the vision outlined above, due to the numerous difficult technical issues that have had to be tackled to even get this far. These include: the need to slow down the rate of DNA translocation through the pore, the need to ensure that nanopores are loaded with DNA, (these two have meant that some library prep has always had to be applied depending on the specifics of the approach adopted, in order to bring DNA in contact with the pore and the surface it is embedded in), the use of designs that feature protein nanopores embedded in a lipid membrane (which means that at any given moment multiple bases are present in the pore and the sequence is reconstructed from the reading of several bases at a time instead of just one; this basically precludes the application of the method to methylated DNA), and others. Thus further advances in miniaturization and manufacturing will have to be made to enable true direct single-base pair readout, to ensure that all nucleic acid molecules in a single cell are read efficiently by the pore, and in the case of direct RNA sequencing on single cells, to further minimize library preparation. Nevertheless, the fundamental features of nanopore sequencing make it the only candidate technology that has the potential to deliver all the information that is at present impossible to obtain using existing tools and protocols, provided that, of course, solutions are found to the challenges that still remain unresolved.

14.5 Concluding Remarks

Advances in sequencing technology have been the primary driver of development in genomics for most of the existence of the field. New sequencing technologies have repeatedly enabled us to ask questions that were not accessible prior to that. The refinement of sequencing technology is not complete – we still do not have the sequencing capabilities we would like to have, both in the area of genome sequencing (where large and highly repetitive genomes are at present practically impossible to assemble) and in several areas of functional genomics, in particular transcriptomics (where we need to be able to sequence full-length transcripts, at very high sequencing depth, preferably without having to use PCR and reverse transcription). However, we can be reasonably certain that these problems will be eventually resolved and we will be able to carry out the measurements we want, whether it is thanks to currently emerging third generation sequencing technologies, or through further developments beyond that.

In the same time, many functional genomic assays, in particular those centered on chromatin biology, are either not expected to derive large benefits from these anticipated developments, or the changes in the nature of the data generated will not be radical. We can therefore consider methods such as ChIP-seq to be at this point mature. This means that, first, the analytical tools developed for working with them, such as the ones described in Part 3 of this thesis, will remain relevant for quite some time into the future, and second, that it is time to shift the emphasize of functional genomic research efforts from developing and improving assays and methods to using the data we can generate and analyze with them to address biological questions (a common criticism of the field, which I personally cannot disagree with, has been its overt focus on technological development). Once the needed advances in the area of transcriptomics outlined above are achieved, the same reasoning will apply for it too. The biological questions, the exploration of which is of particular interest to me, are discussed in the last chapter of the thesis.

15

The extent of functionality in the human genome

In this chapter, I present my personal view on the question of how much of human genome is functional. This is a subject that generated much attention after the publication of the main ENCODE papers in late 2012. As a result, a perspective on the issue was put together by members of the ENCODE Consortium, which I had a hand in putting together, and which has been published as:

Kellis M*, Wold BJ*, Snyder MP*, Bernstein BE*, Kundaje A‡, Marinov GK‡, Ward LD‡, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski L, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard TJP, Kent WJ, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos J, Weng Z, White KP, Hardison RC. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**:6131-6138. doi: 10.1073/pnas.1318948111.

The original text of the paper is reprinted in Appendix M. I contributed to it the coverage analysis, using the set of publicly available ENCODE element files as input as well as newly generated histone mark ChIP-seq region calls provided by Anshul Kundaje (Stanford University). The conservation analysis was added by Luke Ward (Broad Institute and CSAIL, Massachusetts Institute of Technology). I will not reiterate all of the points made in that manuscript here, but will instead emphasize the ones that were either not made or did not feature prominently in it. Also, it should be noted that here I focus almost exclusively on ENCODE results; a more extensive treatment of all the issues associated with this topic can be found in the next chapter.

Abstract

The completion of the sequence of the human genome gave us a rich source of information about certain features of it such as genes and repetitive elements. However, a complete understanding of how the human genome functions requires also the comprehensive identification and characterization of all other functional elements in the genome, especially those involved in the regulation of gene expression. In addition, the annotation of the gene content of the human genome carried out as part of the initial sequencing effort was far from complete, in particular with respect to noncoding RNA species. To fill these gaps in our knowledge, the ENCODE Consortium was set up with the goal of generating an exhaustive genome-wide map of functional elements in the human genome. The main approach that the ENCODE project adopted towards achieving that goal was the use of functional genomic assays to produce maps of biochemical activity across the genome. Its efforts resulted in a collection of such maps that covered the majority (≥80%) of the genome with reproducibly detectable biochemical activity, in stark contrast with prior studies using evolutionary conservation, which

usually estimate no more than 10% of the genome to be functional. Here, a discussion of the sources of this discrepancy, the various lines of evidence for function, the nature of "function", and what the most likely true value is (much lower than 80%), is presented.

15.1 Introduction

The sequencing of the human genome (Lander et al. 2011; Venter et al. 2001; International Human Genome Sequencing Consortium 2004) was a monumental achievement in our quest to understand the genetic basis of human biology. However, on its own it was not sufficient, as, while it provided a list of genes, the identity of the regulatory elements that control them were largely unknown. These elements reside mostly in the noncoding portions of the genome, and their importance is illustrated by the fact that the majority of known trait-associated sequence variants reside outside of protein coding exons (Kleinjan & Lettice 2008; Hindorff et al. 2009; Nicolae et al. 2010; Zhong et al. 2010). In addition, the complexity of the transcriptome had by no means been exhaustively explored, with the corresponding absence of understanding of what roles the RNA species not immediately apparent in the human genome sequence and its initial annotation might be playing. It is with the goal of addressing these gaps in knowledge that the ENCODE Project was set up in the early 2000s (ENCODE Project Consortium 2004). Initially, in its pilot phase, it used tiling microarrays to comprehensively assay transcript abundance and diversity, regions of open chromatin, transcription factor occupancy and histone modifications over 1% of the human genome (ENCODE Project Consortium 2007). It revealed the significant complexity of biochemical activity over the genome (for example, in the form of the pervasive transcription of the targeted regions). However, microarrays are not the ideal technology for the detailed functional genomic characterization of the genome, especially with respect to the transcriptome, as they do not provide a truly single base-pair resolution, and suffer from limitations to their dynamic range, relatively high noise levels and a number of other issues (Royce et al. 2005). In this context, it was fortunate that the beginning of the second, genome-wide production phase of ENCODE coincided with the advent of high-throughput sequencing technology, which quickly displaced microarrays as the primary platform for functional genomics research. Sequencing-based assays are characterized by greatly diminished noise levels and very high resolution (single-base pair in the case RNA-seq and whole-genome bisulphite sequencing), and confidence in the results they deliver is correspondingly higher. The second phase of ENCODE detected reproducible biochemical activity over 80% of the genome, which lead to a heated discussion over whether this is evidence that 80% of human genome is functional or not, ranging from rejection of that idea (Eddy 2012; Eddy 2013; Doolittle 2013; Graur 2013; Niu & Jiang 2012) to its acceptance (Germain et al. 2014; Tragante et al. 2014; Mattick & Dinger 2013). The reason such a position is controversial is that decades of research in population genetics and evolutionary biology have converged onto a coherent view of the human genome as one that consists of a tiny fraction of functional DNA and a majority of nonfunctional DNA, often referred to as "junk" (Ohno 1973). This "junk" DNA exists primarily because the power of natural selection in mammalian lineages is insufficient to efficiently eliminate it, while the balance of mutational biases (in particular, transposable element insertions as well as small insertion and deletions) is on average directed towards expansion, with the strength of natural selection being limited by the low effective population size (N_e) of these species (Lynch 2007a; Lynch 2007b). This subject is elaborated on in depth in the next chapter.

15.2 The three types of evidence for genomic function

A main driving factor behind this debate is the varying weight that is given to the different types of evidence for the functionality of a given region of the genome according to each view. There are three approaches (complementary to each other) for evaluating functionality: biochemical, genetic, and evolutionary, with biochemical evidence being the primary type that ENCODE collected. These are briefly reviewed below, and are also summarized in Table 15.1

15.2.1 Biochemical evidence

Functional elements in the genome exhibit certain biochemical activities when their function is expressed and exercised. These activities include the production of mRNA in the case of protein coding genes and of functional non-coding RNAs from non-coding genes, the binding of transcription factors, chromatin modifying and remodeling complexes, and other proteins to enhancers, promoters, insulators and other regulatory elements, the establishment of characteristic combinations of histone marks over regulatory regions and certain portions of genes, the open chromatin structure of enhancers and promoters, and others. Biochemical activities can be measured genome-wide using functional genomic assays: RNA-seq for transcripts (Mortazavi et al. 2008), ChIP-seq for the occupancy of DNA by proteins and the distribution of histone marks along the genome (Johnson et al. 2007; Barski et al. 2007; Mikkelsen et al. 2007), DNAseq-seq (and its high-resolution version DGF, or Digital Genomic Footprinting; Hesselberth et al. 2009; Neph et al. 2012a), and FAIRE-seq (Formaldehyde-Assisted Identification of Regulatory Elements; Song et al. 2010). This is the main approach that the ENCODE Consortium adopted in order to identify candidate functional elements in the genome, and has also been successfully used to identify them in other systems, in particular to find candidate enhancers regions in mammalian genomes (Visel et al. 2009; Rada-Iglesias et al. 2011; Creyghton et al. 2010).

The main advantage of the biochemical approach is that it provides a direct readout of the biochemical activities in which a functional element is involved. However, the presence of biochemical activity it on its own not sufficient evidence that a given region of DNA is functional – functional elements are biochemically active, but the opposite is not necessarily true (Table 15.1), and biochemical activity can occur over regions of DNA with little to no functional significance.

15.2.2 Genetic evidence

The genetic dissection of the biological role that candidate functional elements may play, in loss- and gain-of-function settings, is the classic, gold-standard approach for defining functionality. This can be accomplished through the study of naturally occurring mutants, the targeted generation of loss-of-function mutants or of RNAi knock-downs (Berns et al. 2004), through the use of transfection assays to measure the activity of candidate enhancer regions, and others. Genetic evidence provides very strong indication for functionality, however, traditionally most genetic approaches have been low-throughput and labor-intensive, in particular in human systems. This has begun to change recently, with the appearance of high-throughput functional assays (Patwardhan et al. 2012; Melnikov et al. 2012; Kheradpour et al. 2013), and of simplified and widely accessible genome editing protocols (Jinek et al. 2012). One problem still remains, however, and it is that not all functional elements display a phenotype upon genetic manipulation. Famously, deletion of ultraconserved elements in mice is known to sometimes lead to viable animals (Ahituv et al. 2007); this might be the result of redundancy with other functional elements or, alternatively, phenotypes may not be visible in laboratory conditions with fitness costs being incurred only in the diverse environmental conditions encountered in the wild. Thus while positive results using genetic tests for functionality can be straightforwardly interpreted as strong evidence for it, negative results do not constitute correspondingly strong evidence for its absence (Table 15.1).

15.2.3 Evolutionary evidence

While experimental tests for functionality suffer from the issue of phenotypes not always being visible under laboratory conditions, throughout the process of evolution functionality is constantly being tested all the time without this constraint. Regions of the genome, the sequence of which is of major functional significance, are subjected to strong purifying selection, and can be detected in multiple genome alignments as conserved, in contrast to the rest of the genome, which tends to evolve largely neutrally and as a result its sequence diverges to a much greater extent between different lineages. This method has been widely used to find conserved noncoding elements in mammalian and other genomes (Nobrega et al. 2003; Cliften et al. 2003; Boffelli et al. 2003; Siepel et al. 2003). It is also the source of the most conservative minimal estimates for the fraction of the genome that is functional – between 5 and 10% (Lindblad-Toh et al. 2011). However, while strong sequence conservation implies functionality, the opposite is not always true – there is extensive evidence that regulatory elements can and often do turnover

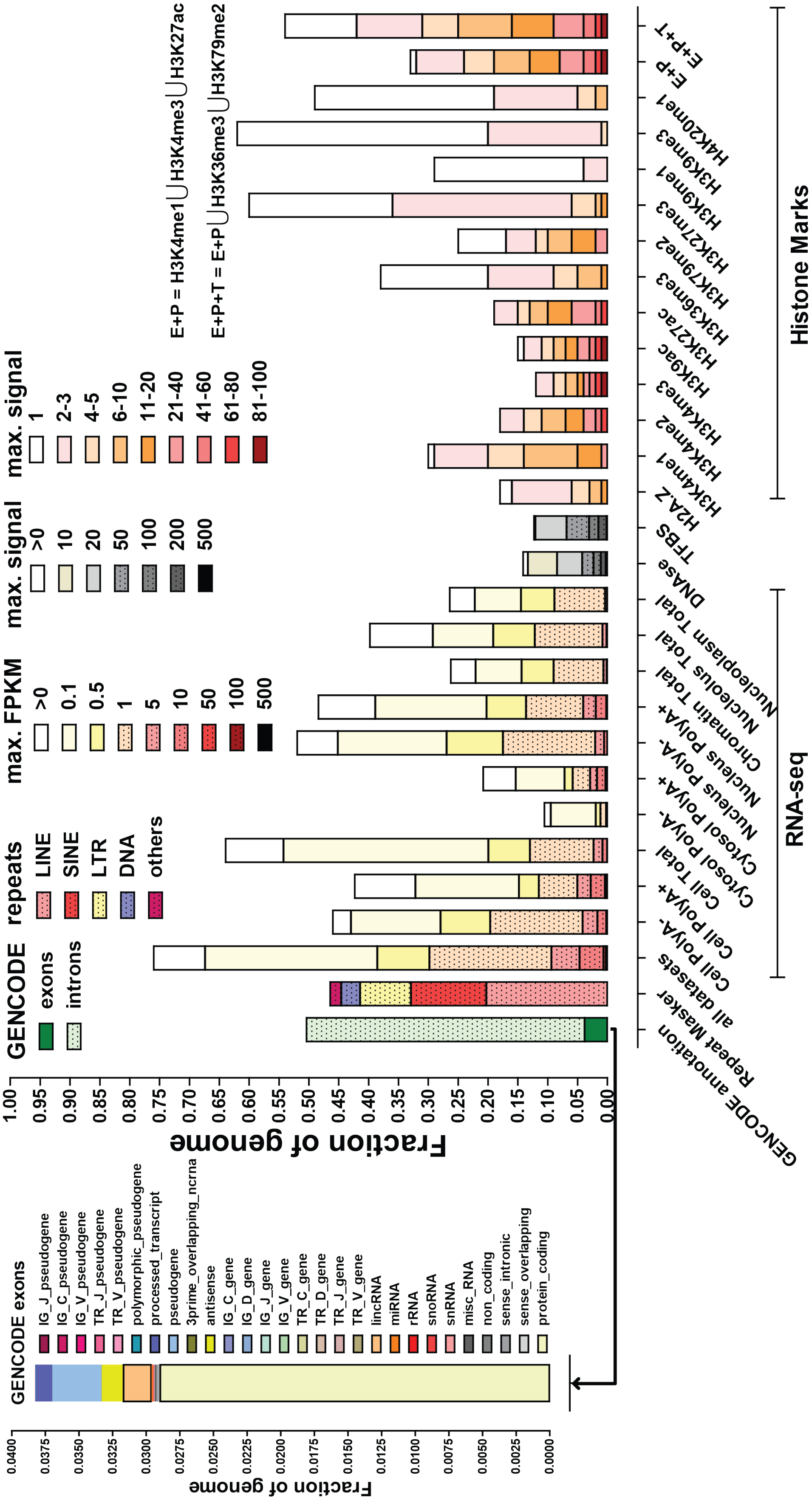

GENCODE exons
IG_J_pseudogene
IG_C_pseudogene
IG_V_pseudogene
TR_J_pseudogene
TR_V_pseudogene
polymorphic_pseudogene
processed_transcript
pseudogene
3prime_overlapping_ncrna
antisense
IG_C_gene
IG_D_gene
IG_J_gene
IG_V_gene
TR_C_gene
TR_D_gene
TR_J_gene
TR_V_gene
lincRNA
miRNA
rRNA
snoRNA
snRNA
misc_RNA
non_coding
sense_intronic
sense_overlapping
protein_coding
Fraction of genome
0.0400
0.0375
0.0350
0.0325
0.0300
0.0275
0.0250
0.0225
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Fraction of genome
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
GENCODE
exons
introns
repeats
LINE
SINE
LTR
DNA
others
max. FPKM
>0
0.1
0.5
1
5
10
50
100
500
max. signal
>0
10
20
50
100
200
500
max. signal
1
2-3
4-5
6-10
11-20
21-40
41-60
61-80
81-100
E+P = H3K4me1⋃H3K4me3⋃H3K27ac
E+P+T = E+P⋃H3K36me3⋃H3K79me2
GENCODE annotation
Repeat Masker
all datasets
Cell PolyA-
Cell PolyA+
Cell Total
Cytosol PolyA-
Cytosol PolyA+
Nucleus PolyA-
Nucleus PolyA+
Chromatin Total
Nucleolus Total
Nucleoplasm Total
DNase
TFBS
H2A.Z
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K36me3
H3K79me2
H3K27me3
H3K9me1
H3K9me3
H4K20me1
E+P
E+P+T
RNA-seq
Histone Marks

Table 15.1: Approaches for evaluating the functionality (F) of segments of the genome. Note that the table is presented as if each criterion produces binary results and functionality is also a binary characteristic, however, the biological reality is, of course, different and all of these are in fact continuously distributed. A means a positive score according to each criterion.

Approach (A)	$F \overset{?}{\Rightarrow} A$	$\neg F \overset{?}{\Rightarrow} \neg A$	$A \overset{?}{\Rightarrow} F$	$\neg A \overset{?}{\Rightarrow} \neg F$
Biochemical	yes	no	not always	yes
Genetic	yes	yes	yes	not always
Evolutionary	not always	mostly yes	yes	not always

relatively rapidly on an evolutionary timescale, and as a consequence are not always detectable in multiple genome alignments (McGaughey et al. 2008; Meader et al. 2010; Lohmueller et al. 2011). Thus absence of conservation does not necessarily imply lack of function (Table 15.1).

15.3 What fraction of the human genome is functional?

The question of how much of the human genome is functional has fascinated researchers for a long time, ever since it was recognized that genome size does not correlate with perceived organismal complexity (the so called C-value paradox, Thomas et al. 1971), and even more so after more recently it was realized that the number of genes in a genome also does not correlate with it, with the human genome containing barely more genes (20,000) than the genome of the nematode *Caenorhabditis elegans* (19,000) and only half the number of genes that many plants have (≥40,000 in some cases). A well-established within the field of molecular evolution explanation for the C-value paradox has been that, aside from cases of polyploidy, the observed differences in genome size are largely the result of the different amount of "junk" DNA that different lineages have accumulated in their genomes. This idea has been consistently supported by the results of comparative genomics, which have re-

Figure 15.1 *(preceding page)*: Summary of coverage of the human genome by ENCODE data. Shown is the fraction of the human genome covered by ENCODE elements in at least one cell line/tissue for each assay as well as genomic coverage by annotated genes and repetitive elements. Version 16 of the GENCODE annotation (Harrow et al. 2012) was used to calculate coverage by annotated genes. Detailed breakdown of the coverage of the genome by the exons of protein coding genes and various non-coding transcripts and pseudogenes is shown separately. The Repeat Masker annotation downloaded from the USCS Genome Browser was used to calculate coverage of the genome by repetitive elements. For transcripts, coverage was calculated from RNA-seq derived contigs (Djebali et al. 2012) separated into abundance classes by FPKM values. Note that FPKMs are not directly comparable between different subcellular fractions as they reflect relative abundances within a fraction rather than average absolute transcript copy numbers per cell. Depending on the total amount of RNA in a cell, 1 transcript copy per cell corresponds to between 0.5 and 5 FPKM in PolyA+ whole cell samples according to current estimates (with the upper end of that range corresponding to small cells with little RNA and vice versa). "All RNA" refers to all RNA-seq experiments, including all subcellular fractions. DNAse hypersensitivity and transcription factor (TFBS) and histone mark ChIP-seq coverage was calculated similarly but divided according to signal strength. "Motifs+footprints" refers to the union of occupied sequence recognition motifs for transcription factors as determined by ChIP-seq and as measured by digital genomic footprinting, with the purple portion of the bar representing the genomic space covered by bound motifs in ChIP-seq. Signal strength for ChIP-seq data for histone marks was determined based on the p-value of each enriched region (the $-log_{10}$ of the p-value is shown), using peak calling procedures tailored to the broadness of occupancy of each modification (Supplementary Methods). "E+P and "E+P+T" refer to the union of coverage by histone marks associated with enhancers and promoters ("E+P") or enhancers, promoters and transcriptional activity ("E+P+T").

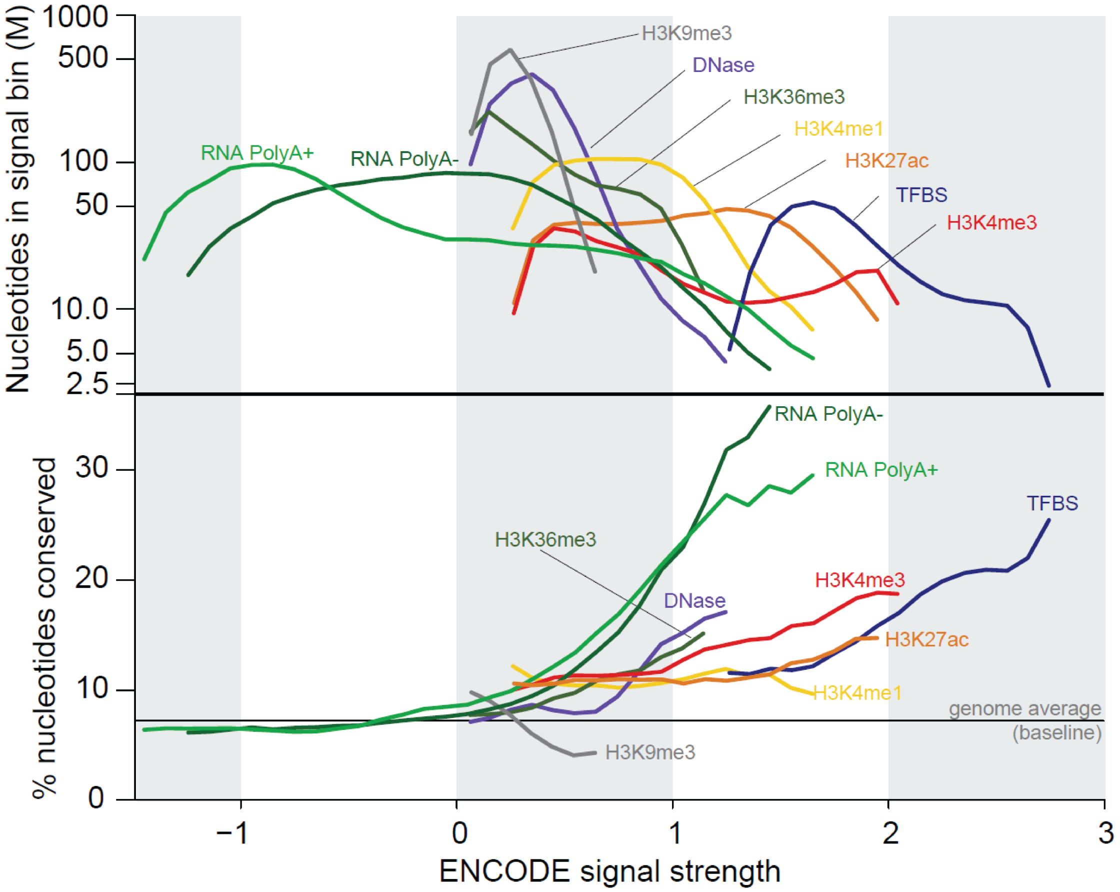


Figure 15.2: Relationship between ENCODE signal and conservation.. Signal strength of ENCODE functional annotations were defined as follows: log_{10} of signal intensity for DNase and TFBS, log_10 of RPKM for RNA, and log_{10} of $-log_{10}$ P value for histone modifications. Annotated regions were binned by 0.1 units of signal strength. (A) The number of nucleotides in each signal bin was plotted. (B) The fraction of the genome in each signal bin covered by conserved elements (by genomic evolutionary rate profiling) was plotted.

peatedly estimated the fraction of the human genome that is conserved within mammals to be below 10% (Mouse Genome Sequencing Consortium 2002; Lindblad-Toh et al. 2011), and by the fact that half of the human genome consists of decayed copies of transposable elements. However, proposals that most of the human genome is in fact functionally important, even though it is not conserved on the sequence level, and that the regulatory complexity hidden in the noncoding and nonconserved portions of the genome is what is responsible for the organizational complexity of the human body and even our cognitive abilities have been repeatedly made (see discussion in the following chapter).

15.3.1 The “biochemically active” 80% fraction

It is in this context that ENCODE’s result that ≥80% of the human genome is biochemically active appeared in the scientific literature. However, while the number became widely popular, its origin was not explained properly, so where exactly does the 80% figure come from?

It should first be noted that given the nature of the functional genomic assays used to generate the data, which cover 80% of the genome with significant and reproducible signal, 80% is really largely equivalent to 100%. The high-throughput sequencing platforms used during this phase of the ENCODE Project generate reads of between 25 and 100bp in length, however, the human genome contains many repetitive and highly similar to each other sequences, meaning that a fraction of it is not uniquely mappable with reads of such length. As only unique reads were considered during analysis, the effectively "visible" portion of the human genome was only slightly larger than 80% of it (Figure 14.1), i.e. ENCODE elements in fact cover nearly 100% of the accessible part of the genome.

Second, ENCODE invested great effort into ensuring the quality of the data produced and the reproducibility of the candidate elements detected (Landt et al. 2012; Li et al. 2011); thus the detected coverage is generally unlikely to be the result of experimental artifacts. However, detailed investigation of where the coverage of the whole genome originates from is needed before conclusions about its significance are made. To this end I generated the summary shown in Figure 15.1, where the coverage of the human genome by different types of data is shown, as a function of signal strength, together with the coverage of the genome by exons and introns of annotated genes, and by repetitive elements. Several types of relevant data were generated by ENCODE, and their properties need to also be understood:

1. **Transcription factor ChIP-seq**. Maps of the genomic occupancy of over 120 human transcription factors were generated using ChIP-seq. However, transcription factors bind to short stretches of DNA sequence, usually 6-8bp long, more rarely up to ∼20bp, while ChIP-seq libraries consist of fragments of average length ∼200bp. As a result, binding regions called from ChIP-seq data are several hundred bases long even though the causative sequence is typically only less than 10bp in length.

2. **Maps of DNAse hypersensitivity regions**. Similarly to ChIP-seq, the identified regions of DNAse hypersensitivity can be several hundred bases long but are caused by the binding of transcription factors and other proteins to DNA sequences of considerably shorter length.

3. **DNAse footprinting.** Very deeply sequenced DNAse libraries provide digital genomic footprints of DBA occupancy and while they are still somewhat longer than actual transcription factor binding sites, they provide a more refined mapping of the contacts between non-nucleosomal proteins and DNA in the genome.

4. **Histone mark ChIP-seq**. The following histone modifications were mapped across a wide variety of cell types: H3K4me3 (a mark associated with active promoters), H3K4me2 (promoters), H3K4me1 (enhancers), H3K9ac (promoters), H3K27ac (ehnancers and promoters), H3K36me3 (transcriptional elongation), H3K79me2 (transcriptional elongation), H3K27me3 (transcriptional repression, in particular when mediated by Polycomb complexes), H3K9me3 (repressed heterochromatin), H3K9me1 and H4K20me1 (of less clear function), and the histone variant H2A.Z (associated with promoter regions). However, a histone state can be induced by sequence elements much shorter than the genomic space occupied by the nucleosome carrying the corresponding marks. For example, an enhancer region might induce histone modifications over several nucleosomes on each side (or just one side; Kundaje et al. 2012).

5. **RNA-seq**. RNA-seq was carried out on polyadenylated RNA from whole cells (the most commonly targeted portion of the transcriptome as these are the characteristics of messenger RNAs and most lincRNAs), but also separately on polyadenylated and nonpolyadenylated RNA from whole cells and from subcellular fractions (primarily nucleus and cytosol, and in a few cell lines, nucleolus, nucleoplasm and chromatin). While the resolution of RNA-seq is single base-pair and its dynamic range is vast, it still has an imperfection and it is that it only measures the relative abundance of transcripts but not their absolute abundances (Löven et al. 2012). This is best illustrated by the following thought experiment. Imagine two different cellular conditions A and B; all genes are upregulated by a factor of 2 in B relative to A, however, when RNA-seq libraries from each are sequenced, the FPKM metric used to calculate gene expression will be the same for each gene. Naturally, one would want to know how many copies of each transcript are present in each cell but this information is generally not avail-

able in RNA-seq datasets without significant modifications to experimental design. It is generally understood that 1 FPKM unit corresponds to between 0.5 and 5 FPKMs (depending on how much total RNA there is in each cell on average: small cells with little RNA per cell would be expected to have higher FPKM-per-copy values) in polyadenylated samples from whole cells, but not even a rough such estimate is available for subcellular fractions. It is likely, however, that the abundance of transcripts in them is significantly lower than that in whole-cell polyA+ samples.

Coverage by annotated elements. According to version 13 of the GENCODE annotation, 50% of the human genome is covered by the exons and introns of annotated genes. Of this, <4% consists of exons, of which ∼2.9% is exons of protein coding genes (including the open reading frames and the untranslated regions), ∼0.2% is exons of lincRNAs, and ∼0.7% is pseudogenes; the rest is introns (Figure 15.1). Also, according to the RepeatMasker annotation, 45% of the genome is covered by repetitive elements. These numbers are relevant, because introns are, of course, transcribed and can be expected to be detected, in particular in the form of not yet spliced pre-mRNAs, in the nuclear and nucleoplasmic subcellular fractions.

Coverage by RNA-seq data. A total of 75% of the genome is covered by RNA-seq elements across all datasets. It is worth first going over the generation of these elements. The ENCODE transcriptome analysis effort (Djebali et al. 2012) eventually settled on generating RNA contigs from the data instead of relying on *de novo* transcript reconstruction (as it is a difficult and still not fully solved problem). RNA contigs were generated based on the overlap of mapped reads on the same strand, FPKMs were calculated for each of them, and they were then subjected to a non-parametric irreproducible discovery rate (npIDR) filtering to narrow down the final list of elements to those that are reproducible. Thus regions for which functionality is not *a priori* expected such as intronic fragments within subcellular fractions were definitely included in the final list of elements.

Even more importantly, much of the observed coverage was derived from elements with very low FPKM values and from subcellular fractions. The fraction of the genome covered by ≥1 FPKM elements is 30% across all fractions, slightly above 10% if only whole cell PolyA+ samples are considered, and ∼5% in Cytosol PolyA+ samples. It is nearly 20% and 15% in nuclear PolyA– and PolyA+ samples, respectively (Figure 15.1).

Coverage by transcription factors occupancy sites and DNAse. Around 15% of the genome is covered by DNAse hypersensitivity and transcription factor occupancy regions (Figure 15.1). In each case, there is a smaller fraction of the genome covered by high signal levels, and a larger fraction covered by signals of lower intensity. This relationship has been noted numerous times in the past (Landt et al. 2012) and it is an open research question to what extent there is a correlation between signal strength and functionality – there are certainly good reasons to think that low-level occupancy may not on average be as functionally important as strongly occupied sites are; some empirical evidence in support of this view has been published in *Drosophila melanogaster* (Fisher et al. 2012).

In addition, up to 10% of the genome is occupied by DGF footprints and by motifs of known transcription factors located within called transcription factor ChIP-seq occupancy regions (Figure 2 in Appendix M).

Coverage by histone marks. As mentioned above, coverage by histone marks is not necessarily a good measure of the extent of the functionality of the genome, as its resolution is not high enough due to the fact that the sequences inducing a specific chromatin state are often much smaller than the region of the genome occupied by that space. It should also be noted that the numbers shown in Figure 15.1 were generated using a broader set of histone marks in terms of the number of cell lines included, further extending total coverage, but also that a more conservative region calling pipeline was used that resulted in regions of significantly shorter length than those used in the main ENCODE publication (ENCODE Project Consortium 2012). Thus they are not necessarily directly relevant to the origin of the 80% number but they are informative on their own. Using this more conservative set, a total of ≥35% of the human genome is covered by regions of enrichment for marks associated with active enhancers and promoters, and ≥55% by marks associated with active enhancers, promoters and transcribed regions, and the same trend of a small fraction of the genome with high signal and a larger portion with lower signal is observed.

15.3.2 The level of biochemical signal correlates with evolutionary conservation

It is clear from the considerations presented above that assuming that biochemical activity measured in these ways is equivalent to functionality is not a viable strategy, as first, the resolution of many of the assays is not sufficiently high, and second, it is not possible to distinguish biochemical noise from functionally significant activity. As discussed at length in the next chapter, it is not reasonable to expect that all transcription in the human genome is functional and that all trascription factor binding events are of major regulatory importance, even when they are reproducibly detected. This means that integrative use of all three criteria for functionality should be used to assess the functional significance of the candidate functional elements identified using functional genomic tools.

We took a first step in this direction by examining the relationship between the strength of the biochemical signal measured by the various assays and level of sequence conservation (Figure 15.2). Strikingly, positive correlation between signal intensity and evolutionary conservation was observed for all types of data, with two exceptions: the H3K4me1 curve was mostly flat and dipped slightly downwards at the high end of the signal distribution, and H3K9me3 exhibited a strong negative correlation with conservation. The former is not straightforward to interpret at present, but the latter is very clearly related to the fact that H3K9me3-modified nucleosomes are a core component of repressed heterochromatin, and repressed heterochromatin is a classic location for “junk” DNA as heterochromatinization is used by cells to silence transposable elements. It is therefore no surprise that regions with more H3K9me3 are less conserved than the genomic average.

That there is strong correlation between conservation and signal levels in RNA-seq, DNAse and transcription factor ChIP-seq datasets also makes sense, and this has clear implications for the interpretation of the majority low-signal coverage of the genome – it is more likely that much (though by no means all) of it is nonfunctional and represents biochemical noise.

15.3.3 The most likely estimate for the fraction of the human genome that is functional

At present we cannot provide a definitive answer to the question how much of the human genome is functional. This is in part because we do not have all the data we would like to have and which we need to tackle it, but also, in even larger part, because it is a question the answer to which is highly dependent on definitions. There is as of now no universally agreed upon definition of what function is, what should be called a functional element (Doolittle 2013; Graur et al. 2013), and at what resolution.

We are naturally inclined to think of function as a binary characteristic that a region of DNA either does or does not have. However, this is not how biology works – the functional significance of individual base pairs, regulatory elements, portions of genes, and even whole genes is distributed on a continuum. The lower end of this continuum, where what is pure biochemical noise meets the currently marginally functional elements, is perhaps where a lot of evolutionary innovation takes place, and it will never be straightforward to place a dividing line and declare that all nucleotides on one side of it are functional while everything else is not. And it might have to be individual nucleotides and not larger regions of DNA as even within well-established functional elements not all nucleotides are of equal significance. One of the most fundamental components of our biology, the genetic code provides a very good example for that, with its often degenerate third positions in codons, and this is even more so for other functional elements. Table 15.2 shows a fictional but typical in its structure position weight matrix (PWM) representing the binding preference of a hypothetical eukaryotic factor. Some positions are very strongly constrained, some can tolerate more than one nucleotide, yet others are not constrained at all and a binding site could be fully functional with each of the four nucleotides in that position. Individual transcription factor binding sites can therefore exhibit significant tolerance to substitutions (though significantly less to deletions and insertions). An example of even less constraint is provided by the 3'UTR of genes, which contain sequences, recognized by miRNAs and by RNA binding proteins, but these are often embedded within sequence that is largely unconstrained. Similarly

Table 15.2: An example of a position weight matrix (PWM) describing the binding preferences of a transcription factor. The PWM score is defined as the fraction of binding sites for the factor in which each base is found in the indicated positions.

Position\Nucleotide	A	T	C	G
1	1	0	0	0
2	0	1	0	0
3	0.5	0	0	0.5
4	0.25	0.25	0.25	0.25
5	0.25	0.25	0.25	0.25
6	0	0.5	1.5	0
7	1	0	0	0
8	0	1	0	0

low constrains are likely operating with respect to the sequence of long non-coding RNAs. Finally, there are transcriptional phenomena where the act of transcription is important but the sequence of the transcripts produced is not, for example transcription interference, in which the production of noncoding RNAs upstream of and through the promoter of a gene inhibits its expression, and numerous other variations of the same theme (Martens 2004; Petruk et al. 2006; Shearwin et al. 2005; Hirota et al. 2008; Uhler et al. 2007; Kuehner & Brow 2008; Thiebaut et al. 2008; Palmer et al. 2009). It is far from clear how to classify the regions that produce such transcripts with respect to whether their sequence is functional or not.

Perhaps the most natural measure of functionality would be the selective coefficient s associated with the presence of the fixed allele in the genome with respect to its hypothetical alternatives and especially compared to the total absence of the element. This is, of course, continuously distributed over many (technically an infinite number) orders of magnitude, thus it cannot provide a hard estimate for how much of the genome is functional. It is also very difficult to measure accurately (near impossible in humans) and is to a great extent dependent on environmental conditions therefore not constant in time. But even if we had a way to measure it with absolute accuracy, we would still be left with the problem of defining what an allele is and what its alternatives are. For example, the fitness effect of substitutions of individual positions in a transcription factor binding site will differ according to their importance for occupancy by that factor, which will relate in some way to the factor's PWM, but even the most severe of these effects will be much smaller than deletion of the whole site. The same reasoning applies on multiple progressively higher levels. Within an individual enhancer there might be several binding sites for the same transcription factor and inactivation of one of them can be compensated but deletion of the whole enhancer will have more serious fitness consequences. A gene might be regulated by multiple enhancers, which are completely or partially redundant, and deletion of some of them can be tolerated, but not of multiple ones, and so on. It matters greatly what resolution we use when we define functional elements and the most appropriate choices of resolution may not be the same for different kinds of elements.

These are all difficult issues, and the prospects of ever achieving conclusive resolution of all of them are slim. But we do not necessarily have to solve them, as obtaining an accurate estimate of how much of the genome is functional is really of very little practical significance and is in the opinion of the author primarily driven by our collective inability to have an objective understanding of our genome and of ourselves as a biological species. The important questions are first, whether most of the genome consists of "junk DNA" and second, what regions of the genome (as opposed to how much of it) are functional and in what ways. Even though a lot of it is biochemically active, the conclusion reached decades ago that most of the genome is "junk" is in no way overturned by ENCODE data, as shown here and in the next chapter, and the answers to the second question will be derived from detailed functional analysis of individual candidate elements, an endeavor that will be greatly facilitated by the advent of genome editing tools and massively parallel functional assays, but will nevertheless still require an immense amount of effort.

Still, if we are to place a rough estimate of how much of the human genome, in my opinion we should use a definition of function that includes the presence of selected effect on the sequence of a given region of DNA (Doolittle 2013; Graur et al. 2013). The most comprehensive comparative genomics effort across mammals (Lindblad-Toh et al. 2011) estimated that at least 5.5% of the human genome is under purifying selection. This is certainly an underestimate as mammalian lineages seem to be subject to particularly massive turnover of distal regulatory elements (Villar et al. 2014). Lineage-specific constraint might be therefore the more relevant metric, one that will certainly produce higher estimates (Lohmueller et al. 2011). Although the large numbers of sequenced human genomes needed to conclusively answer this question are not yet available, initial studies have produced estimates in the neighborhood of 10% of the genome (Ward & Kellis 2012). This is consistent with the limits on the fraction of the human genome that might be under sequence constraint imposed by the mutation rate (as was understood back in the 1960s, if all of our genome was functional and constrained on the sequence level, given the empirically measured mutation rate, each new generation would suffer from debilitating mutations, which is not the case, therefore the fraction of the genome that is sequence-constrained has to be low), and is the most useful in terms of how we think about the genome rough estimate.

16

The origins of genomic complexity and the Tree-of-Life ENCODE

In this chapter, I summarize my view of the results of the ENCODE Project from an evolutionary perspective, especially in the context of the controversy it generated regarding the extent of functionality of the human genome. It consists in large part my vision for the future, but it was also written because a proper response to the controversy coming from within the consortium was, at the time of writing not available. The opinions presented here are, of course, solely mine and not those of the Consortium as a whole.

Abstract

The publications of the results of the first genome-wide phase of the ENCyclopedia Of DNA Elements (ENCODE) project as well as its sister modENCODE projects in *Drosophila melanogaster* and *Caenorhabditis elegans* were landmark moments in our progress towards understanding the biology of eukaryotic genomes as functional genomic characterization of eukaryote species was carried out for the first time at such depth. However, discussion of the actual results of the human ENCODE project was overshadowed by the portrayal of its conclusions as debunking the well-established concept of "junk DNA", and while questioning this interpretation is fully justified, some of it extended into questioning the utility of the field of functional genomics as a whole. I have two goals here. First, I discuss how ENCODE results are entirely consistent with existing nonadaptive frameworks for understanding the origins of genome complexity. Second, I describe the usefulness of and highlight the need for ENCODE-style characterization of a wide diversity of genomes across the tree of life, in particular in the protozoan groups that account for most of the diversity of eukaryotes. Such projects are becoming feasible with recent technical advances and can be expected to resolve a number of important open questions. They would help more rigorously test the different hypotheses about the origins of genome architecture as wide variations of genome sizes and structures exist and intersect with similarly wide variations in organismal complexity. They would also clarify what the truly fundamental principles of eukaryotic gene regulation are, as radical departures from the familiar from opisthokonts and flowering plants genome organization and mechanisms of gene regulation have been found in other eukaryotic lineages, but in general very little is known in detail about these groups. The comprehensive functional genomic characterization approaches pioneered by ENCODE are ideally suited for addressing these gaps in our knowledge.

16.1 Introduction, or a historic overview of what is in our genome

Perhaps the most fundamental question in all of biology concerns the relationship between genotype and phenotype. Understanding that relationship is the ultimate goal of genome biology, both for purely intellectual reasons and for very practical ones as it is what figuring out the genetic basis of diseases reduces to in the end. Knowing how genomes function is crucial for accomplishing this, and involves understanding both the set of molecules that the genome encodes and the mechanisms of gene expression regulation during development and in response to changing environmental conditions.

There are two approaches, complementary to each other, towards achieving these goals. The first one is the very detailed functional genomic characterization of the genomes of certain species, in particular that of humans, and involves the exhaustive identification of functional (i.e. relevant to the organism's phenotype) genomic elements (genes, transcript, regions with regulatory and structural roles). The second one recalls the old saying that nothing in biology makes sense except in the light of evolution (Dobzhansky 1973) and aims at identifying the general principles driving the evolution of genomes, the establishment of certain features in them, and ultimately, understanding the human genome as a product of these principles in action. The former is the approach taken first by the Human Genome Project (Lander et al. 2001; Venter at al. 2001; International Human Genome Sequencing Consortium 2004) and later by the ENCODE consortium, the latter has been pursued by researchers in the fields of molecular evolution and evolutionary genomics. However, a narrative that the main result of the ENCODE Project has been the debunking of the existence of "junk DNA" (DNA, the sequence of which is of little, or even negative, consequence for organismal fitness) emerged. This prompted an at times quite vigorous debate for and against this proposition (Mattick & Dinger 2013; Graur et al. 2013; Eddy 2012; Eddy 2013; Doolittle 2013), and took attention away from the real scientific results of the ENCODE Consortium and other large-scale functional genomic initiatives.

The intellectual roots of the "controversy" go back deep in history, and perhaps can even ultimately be traced back all the way to the 19th century and Charles Darwin's works that laid the foundation of evolutionary theory (Darwin 1859). The history of evolutionary biology since that time is long and complex, but if there is a major discontinuity in it, that is the period when a quantitative explanatory framework for understanding how the frequency of genotypes changes in population was developed in the form of population genetics in the first half of the 20th century (Fisher 1930; Haldane 1932), a framework that incorporated the Mendelian principles of inheritance and is still the foundational basis for all work in the field. Mendel's work was published in Darwin's time (Mendel 1866), however, it was not widely noticed and the proper integration of evolutionary theory with genetics did not happen until the principles of the latter were rediscovered at the turn of the century (de Vries 1900; Correns 1900) and the discipline was further developed. As a result, while the modern theory of evolution recognizes multiple evolutionary forces - mutation, genetic drift, migration and natural selection - only the last one featured prominently in Darwin's writings and to this day, due to the cultural importance of Darwin and the attention his work has deservedly received, thinking about evolution has been excessively skewed towards viewing all of its outcomes as the result of adaptation (Gould & Lewontin 1979; Brenner 1998; Lynch 2007b; Lynch 2007c). The "hardening" of the Modern Synthesis (Huxley 1942) in the mid-20th century also greatly contributed to this state of affairs, which later developments in the opposite direction have only partially altered.

How we think about evolution greatly affects how we think about genome biology, and vice versa. Advances in our knowledge of genome function have been critical for the development of evolutionary theory. Throughout the 20th century, new discoveries of the molecular features of genomes, the biology of RNA and the mechanisms for regulating gene expression have gone hand in hand with placing them in an evolutionary context, with improved understanding in both areas being the end result. The concept of the "gene" as an individual unit of inheritance was developed around the turn of the 20th century (de Vries 1989; Johannsen 1909), and around the same time it was understood that genetic material is physically organized into chromosomes (Sutton 1902; Sutton 1903; Boveri 1904; Morgan et al. 1915). However, even though DNA was discovered long before that (Miescher 1871), it was not until the 1940s that it was confirmed that it is the carrier of genetic information (Avery et al. 1944; Hershey & Chase 1952). In retrospect it is remarkable that much of the mathematical foundations of population genetics, still standing strong today, was worked out in the absence of understanding of the molecular biology of heredity. The subsequent discovery of the structure of DNA (Watson & Crick 1953a; Watson & Crick 1953b), the elucidation of the genetic code and the basic mechanisms of gene expression (Crick 1958; Crick et al. 1961; Lengyel et al. 1961; Nirenberg & Matthaei 1961; Yanofsky et al. 1964; Sarabhai et al. 1964; Nirenberg & Leder 1964; Marcker & Sanger et al. 1964; Holley et al. 1965a; Holley et al. 1965; Weigert & Garen 1965; Brenner et al. 1965; Crick 1966; Khorana et al. 1966) and the formulation of the Central Dogma of molecular biology (that genetic information cannot flow back from protein to nucleic acids or between proteins; Crick 1970) filled that gap in knowledge, and facilitated the development of the neutral and nearly neutral theories of molecular evolution in the late 1960s and the 1970s (Kimura 1968; King & Jukes 1969; Ohta 1973; discussed in more detail below).

The main question in the study of genome biology since then has been how the expression of genes is regulated, as differential gene regulation is the process that is ultimately behind the establishment of different cell states during development and in response to environmental stimuli. Much of our progress has consisted of a growing appreciation of the complexity of the roles that noncoding DNA (ncDNA) and noncoding RNAs (ncRNAs) play in these processes. The foundations of the study of gene regulation were laid by studies of the bacterial *lac* operon (Jacob & Monod 1961) and the λ phage (Ptashne 1967), but it took quite a bit longer for a rudimentary understanding of it in eukaryotes to emerge. An early and important observations from studies of DNA reassociation kinetics (C_0t curves) was that the amount of repetitive DNA in multicellular organisms is much higher in the more organizationally "complex" species than it is in the "lower" ones, explaining most of the large variation in genome size seen between them (Britten & Kohne 1968). As a consequence, an early theory of gene regulation featured a prominent role for repetitive DNA in the regulation of gene expression in multicellular organisms (Britten & Davidson 1968; Britten & Davidson 1969).

However, repetitive DNA was later understood to be the product of transposable element insertions, which were discovered through genetic means years earlier (McClintock 1950; McClintock 1953) and even suggested to control genes (McClintock 1951; McClintock 1956). When placed in the context of the nearly neutral theory of molecular evolution, developed in the 1970s, it eventually came to be seen as parasitic "junk" (Orgel & Crick 1980; Doolittle & Sapienza 1980), the result of transposable elements reproducing themselves within the genome with the sole purposes of making more copies of themselves.

Around the same time, the first pseudogenes were identified (Jacq et al. 1997; Hardison et al. 1979; Fritsch et al. 1980; Vanin et al. 1980; Nishioka et al. 1980; Lauer et al. 1980). Pseudogenes are portions of the genome that are clearly derived from inactivated copies of formerly functional protein coding genes. They have also been long understood to constitute "junk" DNA in their majority.

In the early 1980s, the first transcriptional enhancers were found (Banerji et al. 1981; Banerji et al. 1983; Gillies et al. 1983), sequence elements capable of stimulating the expression of genes from a long distance and irrespective of their orientation relative to genes and their promoters. Later, insulator (elements blocking the action of an enhancer when situated between it and its target promoter) and other regulatory elements were also identified, primarily from studies of β-globin and a limited number of other loci (Emerson et al. 1985; Forrester et al. 1986; Grosveld et al. 1987; Udvardy et al. 1985; Chung et al. 1993). Eventually it became clear that gene expression in multicellular eukaryotes is in large part controlled not just by transcription factor binding sites in their promoter proximal region but also by regulatory elements residing away from genes in noncoding space that can act at large distance (with extreme examples of enhancers residing nearly 1Mb away from their target known; Lettice et al. 2003).

While histone proteins have been known since the late 19th century, it was in the early 1970s that it was found that unlike prokaryotes eukaryotic chromatin is organized into nucleosomes (Kornberg 1974; Olins & Olins 1974). It was later shown that such a chromatin organization has a repressive effect on transcription

(Grunstein 1990), a consequence of which is that overcoming this barrier is key component of both the regulation and execution of gene expression. That histones carry various chemical modification, in particular in their N-terminal tails, was also known for a long time, but only in the mid- and late 1990s that was it understood that these marks are deposited and removed in a dynamic and regulated manner (Brownell et al. 1996). We now know that chromatin modifications and chromatin remodeling play key role in all aspects of chromatin biology, as histone modifications constitute a form of code that is specifically written and read by various proteins and is critical for the orderly execution of biochemical processes operating on chromatin (Jenuwein & Allis 2001; Kouzarides 2007; Li et al. 2007). They, together with the methylation of position 5 of cytosine residues in DNA (Johnson & Coghill 1925; Hotchkiss 1948) also play a vital role in the epigenetic specification of cell states (Holliday & Pugh 1975; Riggs 1975; Goldberg et al. 2007; Bernstein et al. 2007).

In the late 1970s and early 1980s it was realized that genes in eukaryotes are interrupted by introns (Berget et al. 1977; Chow et al. 1977), which are spliced out before the mature mRNA is translated, with the process sometimes generating alternative splicing products (King & Piatigorsky 1983; Schwarzbauer et al. 1983). The evolutionary origins of splicing have been much debated and the outcome of their study has had major implications for how we understand the evolution of life (see discussion below).

In the 1980s, another interesting phenomenon was observed, the editing of the sequence of RNAs through the modification, or even the insertion or replacement, of individual bases, initially in the mitochondria of the kinetoplastid protozoans (Benne et al. 1986; Feagin et al. 1987; Feagin et al. 1988; Shaw et al. 1988) but later also in the nuclear genomes of animals and many other eukaryotes.

Since the 1970s, an ever expanding universe of ncRNA species carrying out a wide variety of cellular functions has been identified, in eukaryotes and in other organisms. These include among others:

1. **snRNAs**, or U-RNAs, small nuclear RNAs that are core components of the spliceosomal machinery necessary for excising introns during splicing.

2. **snoRNAs**, small nucleolar RNAs that guide the chemical modifications of other RNAs, such as the ribosomal and transport RNAs.

3. The **SRP RNA** (Walter & Blobel 1982), component of the signal recognition particle used to target proteins to the endoplasmic reticulum.

4. **Antisense transcripts**, first discovered in bacteria where antisense transcription can be used to inhibit translation by base pairing with the sense transcript (Mizuno et al. 1984).

5. the phenomenon of RNA interference (Fire et al. 1998), induced by double stranded RNA, and by siRNAs (small interfering RNAs; Hamilton & Baulcombe 1999; Elbashir et al. 2001)), in which the expression of genes is inhibited post-transcriptionally through the degradation of mRNAs complementary to these small RNAs.

6. **miRNAs**, 21-23nt small RNAs (Lee et al. 1983; Reinhart et al. 2000; Pasquinelli et al. 2000) that can inhibit the translation of genes and/or target them for cleavage, in particular through binding to their 3'UTRs.

7. **lncRNAs/lincRNAs**, long (intergenic) noncoding RNAs that do not code for proteins but function as RNAs, such as the *Xist* and *Tsix* (Borsani et al. 1991; Brown et al. 1991; Lee et al. 1999), and *roX* (Meller et al. 1997) RNAs, involved in the establishment of dosage compensation of sex chromosomes in mammals and *Drosophila*, respectively, as well as numerous others (Ji et al. 2003; Wang et al. 2002; DeChiara & Brosius 1987).

8. **The telomerase RNA**, a core component of the machinery responsible for the maintenance of chromosome ends in eukaryotes (Greider & Blackburn 1987; Greider & Blackburn 1989; Shippen-Lentz & Blackburn 1990).

9. **7SK RNA** (Reddy et al. 1984), which regulates the activity of the transcription elongation factor P-TEFb (Diribarne G, Bensaude 2009), **Y RNAs** (Lerner et al. 1981; Christov et al. 2006), **Vault RNAs** (Kedersha & Rome 1986), and numerous

others, the precise functions of which are less clear.

Remarkably, all of these regulatory mechanisms and ncRNAs were discovered in the absence of complete genome sequences, in the course of biochemical and cell and molecular biology studies. One of the very relevant to the ENCODE debate discoveries, that the genome is pervasively transcribed was also made in the pregenomic era. The C_0t curve methodology used to find that much of the genomes of multicellular organisms consists of repetitive sequence, was also applied to the transcriptome in the late 1970s. A large fraction of the genome was found to be transcribed (Hough et al. 1975; Holland et al. 1980), but also to be present at very low copy number, significantly less than one RNA molecule per cell.

Successive major technical advances have enabled both the generalization of these findings and the discovery of many additional layers of complexity. These include the development of DNA sequencing technology (Sanger et al. 1997; Maxam & Gilbert 1977), which made possible the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001; International Human Genome Sequencing Consortium 2004) and the genomes of the main model organisms (Goffeau 1996; *C. elegans* Sequencing Consortium.; Adams et al. 2000; Mouse Genome Sequencing Consortium 2002); the development of microarray technology for measuring RNA expression levels (Schena et al. 1995; Lashkari et al. 1997) and the genomic occupancy of proteins (Iyer et al. 2001; Ren et al. 2000), and the more recent advent of high-throughput sequencing technologies (Shendure et al. 2005; Margulies et al. 2005; Bentley et al. 2008; McKernan et al. 2009; Harris et al. 2008; Rothberg et al. 2011) and the myriad of applications it has found in the form of various functional genomic *-seq assays for (Wold & Myers 2008): RNA-seq for the study of the transcriptome, at the level of large cell populations (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Pan et al. 2008; Sultan et al. 2008; Wang et al. 2008; Wilhelm et al. 2008) and individual cells (Tang et al. 2009; Tang et al. 2010; Islam et al. 2011; Hashimshony et al. 2012; Ramsköld et al. 2012; Picelli et al. 2013; Islam et al. 2014; Wu et al. 2014), CAGE for the mapping of the 5' ends of capped transcripts (Kodzius et al. 2006; Balwierz et al. 2009), GRO-seq (Core et al. 2008) for measuring the instantaneous rate of transcription, ribosome profiling for the measuring translational activity (Ingolia et al. 2009), ChIP-seq for the high-resolution genome-wide profiling of protein-DNA interactions (Johnson & Mortazavi et al. 2007; Barski et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007), DNAse-seq (Hesselberth et al. 2009; Song et al. 2011), FAIRE-seq (Gaulton et al. 2010; Song et al. 2011) and DGF (Neph et al. 2012a; Neph et al. 2012b) for the mapping of regions of open chromatin in the genome, BS-Seq for assessing levels of DNA methylation (Lister et al. 2008; Lister et al. 2009; Meissner et al. 2008), ChIA-PET (Fullwood et al. 2009; Handoko et al. 2010; Li et al. 2010; Li et al. 2012), and Hi-C (Lieberman-Aiden et al. 2009; Umbarger et al. 2011; Zhang et al. 2012; Dixon et al. 2012) for studying the three-dimensional physical organization of genomes, as well as numerous others.

The genomic era has delivered the, exhaustive identification of previously known functional components of the genome as well as the discovery of a number of novel RNA species and new phenomena in transcriptional and regulatory biology. These advances include:

1. The comprehensive cataloging of gene content. Initially, >30,000 protein coding genes were reported in the human genome (Lander et al. 2001). This number has gone down after subsequent refinement of annotations and has stabilized around 20,000 (Harrow et al. 2012).

2. The genome-wide identification of conserved noncoding sequences from multiple genome alignments,. Such sequences are strong candidates for functional regulatory elements (Hardison et al. 1997; Hardison 2000; Siepel et al. 2005; Bejerano et al. 2004; Woolfe et al. 2004; Margulies et al. 2003; Cooper et al. 2004). In total, while <2% of the human genome consists of protein coding sequence, the sequence-constrained fraction of it is at least 5.5% (Lindblad-Toh et al. 2011).

3. The genome-wide identification of miRNAs using both computational and experimental methods (Lagos-Quintana et al. 2001; Lau et al. 2003; Grad et al. 2003; Lai et al. 2003; Lagos-Quintana et al. 2003; Aravin et al. 2003; Houbaviy et al. 2003; Lim et al. 2003a; Lim et al. 2003b; Bartel

2004), of which several hundred are now known in vertebrate genomes.

4. The discovery of multiple additional classes of small RNAs in various organisms, with piRNAs, 24-29nt RNAs playing a crucial role in the defense of the genome against transposable element proliferation being perhaps the most significant (Aravin et al. 2006; Girard et al. 2006; Grivna et al. 2006; Lau et al. 2006; Ruby et al. 2006; Aravin et al. 2007a; Aravin et al. 2007b; Batista et al. 2008; Brennecke et al. 2007; Lin 2007; Aravin et al. 2007a; Brennecke et al. 2007; Gunawardane et al. 2007; Houwing et al., 2007; Bagijn et al. 2012; Ashe et al. 2012; Shirayama et al. 2012; Lee et al. 2012; Vazquez et al. 2004; Peragine et al. 2004).

5. The significant expansion of the list of lincRNA genes, of which several thousands are now known in vertebrate genomes (Guttman et al. 2009; Khalil et al. 2009; Guttman et al. 2011; Dinger et al. 2008; Mercer et al. 2008; Pauli et al. 2012), the functional role of a number of which have been investigated in detail (Sleutels et al. 2002; Sunwoo et al. 2009; Tian et al. 2010; Loewer et al. 2010; Gupta et al. 2010; Huarte et al. 2010; Grote et al. 2013; Hacisuleyman et al. 2013; Sun et al. 2013; Kretz et al. 2012).

6. The identification of alternative splicing events in the genome, initially through the sequencing of expressed sequence tags (EST; Adams et al. 1991; Adams et al. 1995), and later using microarrays and RNA-seq, which has shown that the vast majority of mammalian genes can be transcribed into more than one isoform.

7. The growing appreciation of the molecular and functional complexity of the transcriptome driven by the discoveries of numerous novel RNA species and transcriptional phenomena (Gingeras 2009), a functional role for some of which has been shown:

   7.1 The discovery of eRNAs (enhancer RNAs). These RNAs are transcribed bidirectionally from active enhancers and there is evidence that their transcription is necessary for the positive regulation of the genes targeted by the enhancer from which they originate (Koch et al. 2008; Kim et al. 2010; Ørom et al. 2010; Melo et al. 2013; Li et al. 2013; Lam et al. 2013; Hah et al. 2013; Mousavi et al. 2013).

   7.2 The discovery of circular RNAs. Numerous examples of circularized RNA molecules, arising from the nonlinear splicing of introns, have been reported over the decades (Hsu & Coca-Prados 1979; Cocquerelle et al. 1992; Capel et al. 1993; Cocquerelle et al. 1993; Surono et al. 1999; Zaphiropoulos 1996; Zaphiropoulos 1997; Li & Lytton 1999; Dixon et al. 2005; Burd et al. 2010). However, it was the advent of RNA-seq that allowed for the identification of a large number of them on a genome-wide scale (Salzman et al. 2012; Salzman et al. 2013; Memczak et al. 2013; Hansen et al. 2013; Wang et al. 2014). Functional characterization of individual cases has suggested that they play a regulatory role by acting as miRNA "sponges" sequestering miRNAs and making them unavailable for repression of their target genes.

   7.3 The ceRNA hypothesis, which proposes that mRNAs, and in particular transcribed pseudogenes and lincRNAs compete for the binding of miRNAs, and therefore ceRNA molecules can be used to modulate the efficiency of miRNA-mediated repression (Salmena et al. 2011; Karreth & Pandolfi 2013; Ala et al. 2013; Karreth et al. 2011; Tay et al. 2014a; Tay et al. 2014b).

   7.4 Widespread antisense transcription, in particular in the form of *cis*-NATs (Natural Antisense Transcripts), which have been proposed to play a role in the regulation of the expression of their cognate genes (Vanhée-Brossollet & Vaquero C 1998; Lehner et al. 2002; Wang et al. 2005; Yelin et al. 2003; Cheng et al. 2005; Katayama et al. 2005; Korneev & O'Shea 2005; Kiyosawa et al. 2003).

   7.5 The pervasively transcribed genome. As mentioned already, it had been known that the genome is pervasively transcribed at a low level for decades, but microarray-based studies in the early 2000s in mammals (Cheng et al. 2005; Manak et al. 2006; Johnson et al. 2005; Kapranov et al. 2002; Kapranov et al. 2005; Clark et al. 2011; Bertone et al. 2004; Kampa et al. 2004; Kapranov et al. 2007; Carninci et al. 2005), fly (Stolc et al. 2004), rice (Li et al. 2006), and yeast (David et al. 2006; Dutrow et al. 2008) presented further direct evidence for it and attracted a lot of attention to this phenomenon (Van Bakel 2010; Clark et al. 2011).

7.6 Promoters are bidirectionally transcribed. GRO-seq studies and the deep sequencing of the small RNA fraction of the transcriptome have shown that promoters are bidirectionally transcribed (Core et al. 2008; Seila et al. 2008; Xu et al. 2009), although generally only the sense transcript produces a stable mRNA (Almada et al. 2013).

7.7 CUTs (Cryptic Unstable Transcripts), SUTs (Stable Uncharacterized Transcripts) and PROMPTS (PROMoter associated Pervasive Transcripts), RNA species originating from intergenic and intragenic regions, which are normally present at low levels (higher for SUTs) and become robustly detectable upon inactivation of RNA degradation pathways such as the exosome (Wyers 2005; Thiebaut 2006; Thompson & Parker 2006; Davis & Ares 2006; Neil et al. 2009; Preker et al. 2008).

7.8 RNA species of unknown functional significance associated with transcription starts sites (TSSs) and transcription terminations sites (TTSs), such as TSS-RNAs (20-90bp bidirectionally transcribed, TSS-associated RNAs; Seila et al. 2008), tiRNAs (∼18bp RNAs bidirectionally associated with transcription initiation sites; Taft et al. 2009a; Taft et al. 2009b; Taft et al. 2010), Promoter-Associated Short RNAs (PASRs) and Promoter-Associated Long RNAs (PALRs) (Kapranov et al. 2007; Fejes-Toth et al. 2009), and Termination Associated Short RNAs (TASRs) (Kapranov et al. 2007).

It is in the context of these developments that the ENCODE Project was set up in the early 2000s (The ENCODE Project Consortium 2004) and later carried out, as a follow up to the Human Genome Project, and with the goal of comprehensively identifying the functional elements in the human genome. The first, pilot phase of the ENCODE Project concluded in 2007 (The ENCODE Project Consortium 2007); it focused on assaying a selected 1% of the genome using high-density tiling arrays. It demonstrated the utility of the large-scale functional genomic characterization of genomes, but also generated some controversy as it delivered a message of pervasive transcription and biochemical activity throughout the genome, which was portrayed as debunking of the concept of junk DNA (Weiss 2007; Sample 2007). The pilot phase of ENCODE was followed by a genome-wide production phase, which was also accompanied by companion modENCODE projects in fly and worm (Celniker et al. 2009) and later by a mouse ENCODE project (Mouse ENCODE Consortium 2012). The beginning of the second phase of ENCODE coincided with the adoption of high-throughput sequencing, which allowed a truly genome-wide coverage of the genome, at much higher resolution and with less noise than microarrays did, significantly increasing the confidence in the signals observed. The publication of the results of these projects (Gerstein et al. 2010; modENCODE Consortium 2010; Kharchenko et al. 2011; Négre et al. 2011; ENCODE Project Consortium 2011; ENCODE Project Consortium 2012; Djebali & Davis et al. 2012; Gerstein et al. 2012; Thurman et al. 2013; Neph et al. 2012; Wang et al. 2012) also emphasized the extent to which the genome is biochemically active and was strongly represented as a proof against the existence of large amounts of nonfunctional DNA in the human genome. This has resulted in even more heated arguments than the pilot phase generated, a debate which has at times moved beyond attacking the conclusions of the project and into doubting the basic premises of functional genomic studies. A major factor behind this course of events has been the tendency to view ENCODE data primarily through the prism of a panadaptationist understanding of genome evolution, while ignoring alternative theories, in which nonadaptive evolutionary forces have been a main driver of the evolution of genome organization, and which have enjoyed wide acceptance within the molecular evolution community for some time. Below I overview these competing perspectives on the subject before interpreting ENCODE results in what is in my opinion the proper context.

16.2 The adaptive view of the evolution of genome complexity

There is a long tradition in biology of providing adaptive explanations for most observations. This goes back to the fact that natural selection was the main theme of Darwin's foundational work on the subject (Darwin 1859) but is also because panadaptationist views have dominated popular presentations of evolution (Dawkins 1986; Dawkins 1996), and because of the explanatory utility of adaptation (Mayr 1983). It is in the spirit of this tradition that the growing appreciation of the complexity of metazoan transcriptional regulation and RNA biology has been interpreted, and in turn the results of the ENCODE project have been widely viewed through the lens of an ultra-adaptationist explanatory framework (for example, Mattick & Dinger 2013). Specific propositions that are often argued for include the following:

16.2.1 The absence of sequence conservation does not mean that nonconserved sequences are not functional

A traditionally widely used criterion for assessing the functional significance of genomic segments is the phylogenetic conservation of their sequence. Strong sequence conservation is the result of the action of purifying selection, which means such sequences are subject to significant evolutionarily constraint and highly likely to be functional. However, while conservation is very strong evidence for functionality, the absence of conservation does not necessarily imply lack of function, and numerous examples of both conserved and nonconserved functionalities conferred by rapidly turning over at the sequence level functional elements are known (Smith et al. 2004; Meader et al. 2010; Ponting & Hardison 2011; see also discussion below). The existence of nonconserved functional elements is often extrapolated to a proposition that most or even the whole genome has a function in the absence of sequence conservation (Pang et al. 2006; Pheasant & Mattick 2007; Oldmeadow et al. 2010; Mattick & Dinger 2013). It is also sometimes argued that the conservation criterion is based on the circular reasoning of assuming that repetitive elements evolve neutrally and then using their rate of evolution as a reference to identify the constrained portion of the genome (Pheasant & Mattick 2007; Mattick & Dinger 2013).

16.2.2 Biochemical activity implies functionality

Functional DNA elements exhibit biochemical activity in the form of transcription and occupancy by transcription factors, and other regulatory or architectural chromatin-associated and RNA-binding proteins. thus detection of such biochemical activity does suggests possible functionality for a given region of the genome. However, this is often taken further to argue that the detection of biochemical activity always means a given region of the genome is functional, and such interpretations are a primary reason why the results of the ENCODE Project and of earlier efforts reporting pervasive transcription in mammalian genomes have been perceived as debunking the concept of "junk" DNA.

16.2.3 Repetitive DNA of transposable-element origin is functional

Transposons are traditionally understood to be "selfish" DNA sequences existing solely in order to propagate themselves, and thus an archetypal example of "junk DNA" (Orgel & Crick 1980; Doolittle & Sapienza 1980). However, transposons have been a rich source of material for evolutionary innovation and have been exapted into functional roles on numerous occasions, at the level of individual transposable element insertions (Norris et al. 1995; Vansant & Reynolds 1999; Rebollo et al. 2012; Chen et al. 2009; Krull et al. 2007; Lynch et al. 2011; Medstrand et al. 2001; Naito et al. 2009; Peaston et al. 2004; Schmidt et al. 2012; Santangelo et al. 2007; Bejerano et al. 2006; Faulkner et al. 2009; Kunarso et al. 2010) and even globally (Singh et al. 1985; Espinoza et al. 2004; Allen et al.

2004; Fornace & Mitchell 1986; Li et al. 1999; Mariner et al. 2008; Liu et al. 1995). These and other observations (for example, the somatic retrotransposition observed in the human brain and in cancer cells; Peaston et al. 2004; Muotri et al. 2005; Coufal et al. 2009; Baillie et al. 2011; Lee et al. 2012; Evrony et al. 2012) have been extrapolated into interpreting transposable elements as a vital functional regulatory component of the human genome (Makalowski 2003; Shapiro 1999; Shapiro 2005; Shapiro JA & von Sternberg).

16.2.4 Pseudogenes have functions

Pseudogenes are another classic example of "junk DNA", for which examples of possible exaptation have accumulated in recent years. From such observations a function for the majority or even all of them is generalized (Balakirev & Ayala 2003; Pink et al. 2011; Muro et al. 2011; Li et al. 2013). There are several known mechanisms though which a pseudogene could play a functional role. First, antisense pseudogene transcripts could regulate the expression of sense transcripts from the parental gene (McCarrey & Riggs 1986), some possible examples of which have been reported (Korneev et al. 1999; Hawkins & Morris 2010). Second, pseudogene-derived small RNAs can have a regulatory effect on the parent genes (Tam et al. 2008; Watanabe 2008). A role for pseudogenes in regulating mRNA stability has also been proposed (Hirotsune et al. 2003; Piehler et al. 2008). Finally, pseudogene-derived ceRNAs can act as miRNA sponges (Tay et al. 2014b), affecting the expression of the parent gene.

16.2.5 Functionally important alternative splicing is widespread

Numerous examples of alternative splicing generating different protein products with distinct functions have accumulated since the discovery of splicing (for example, Lynch & Maniatis 1996; Kornblihtt et al. 1996; Graveley 2002; Liao et al. 2005; Izquierdo 2005; Venables 2012), and the number of human genes for which multiple splice products have been detected has been constantly increasing as technology moved from EST sequencing to splicing microarrays and eventually to RNA-seq (Mironov et al. 1999; Croft et al. 2000; Xu et al. 2002; Johnson et al. 2003; Kwan et al. 2008; Wang et al. 2008; Harrow et al. 2012). It now includes the great majority of multiexonic genes in the human genome. Much of this splicing has been reported to be tissue-specific (Pan et al. 2004; Xing & Lee 2005; Wang et al. 2008), and these observations have been interpreted as evidence for the widespread prevalence of adaptively important functional alternative splicing in complex multicellular animals (Kim et al. 2007; Romero et al. 2007; Stamm et al. 2005). The vast universe of alternative splicing products could play a crucial role in expanding the protein coding repertoire of the genome and are proposed to explain the perceived contradiction between the high level of organismal complexity of humans and the fact that we do not have a larger number of genes than other species. Early theories for the evolution of splicing also viewed it from an adaptive angle, by suggesting that genes existed in pieces containing separate functional domains very early in evolution and splicing allows for the shuffling of these domains and the generation of increased protein diversity, which was selectively beneficial (Gilbert et al. 1997; de Souza et al. 1996; Kriventseva 2003). While these theories are now largely rejected (see a more detailed discussion of this subject in the next section), the idea that the presence of introns and splicing is a major causal factor driving increased organismal complexity (Mattick 1994) is still very much alive (Chen et al. 2014).

Additional functions of alternative splicing products have also been proposed. For examples, it is commonly observed that a significant fraction of alternative splicing products contain truncated ORFs and are expected to be subject to nonsense mediated decay (NMD). It has been suggested that the regulated production of such isoforms may serve as an additional mechanism for the regulation of protein expression (Lewis et al. 2003; McGlincy & Smith 2008; Cuccurese et al. 2008).

16.2.6 The central importance of ncRNA and of "exotic" transcripts for the emergence of organismal complexity

Examples of previously unknown functional role of ncRNAs and the complexity of metazoan RNA biology have repeatedly been interpreted as providing an explanation for the high sophistication of organismal organization in complex multicellular animals, and even for the evolution of human intelligence (Mattick & Gagen 2001; Frith et al. 2005; Mattick 2004; Amaral et al. 2008; Mattick et al. 2010; Mattick 2007; Taft et al. 2007; Mattick 2009; Mattick 2011; Liu et al. 2013; Slack 2006). The latter is often based on examples of the expression and activity of such ncRNA species in brain tissue (Mercer et al. 2008; Mehler & Mattick 2006).

These and other proposals of similar nature paint a picture of the genome and organismal evolution in which practically every detail of genome and organismal biology is the product of selective evolutionary forces and is of major adaptive importance for organismal fitness. Within this framework, a high level of complexity of transcriptional and RNA biology is needed in order for organismal complexity to emerge, which in turn is understood to be vastly higher in humans than in other animals, with a correspondingly intricate, largely RNA-mediated regulatory mechanisms.

16.3 The nonadaptive view of genome evolution

In contrast to the understanding of the human genome as lacking "junk" DNA and consisting almost entirely of functional sequence, a diverse set of empirical observations and theoretical considerations, starting in the middle of the 20th century, and significantly enhanced more recently with the advent of comparative genomics, strongly suggest that a large portion of it is indeed junk. The key concept here is the idea of the selection-drift barrier. A foundational result in population genetics states that the power of natural selection to influence allele frequencies is constrained by the magnitude of the selective coefficient s of a given mutation and the effective population size N_e. Specifically, when:

$$|s| < \frac{1}{4N_e} \qquad (16.1)$$

in a diploid sexually reproducing species, mutations evolve effectively neutrally and are "invisible" to natural selection. This has the important consequences that first, mutations with negative effects on fitness will not be weeded out by selection, and second, beneficial mutations are not guaranteed fixation, provided the magnitude of the selective disadvantage they confer is sufficiently low. The value of s for which this is true is increasingly higher the lower the effective population size N_e is.

The general nature of this relationship has been known since early on in the development of population genetics (Wright 1931) but never featured prominently in the Modern Synthesis, and especially in the "hardened" panselectionist version of it that eventually became widely popular. The development of the neutral and nearly neutral theories of molecular evolution in the 1960s and 1970s (Kimura 1968; Kimura 1983; King & Jukes, 1983; Ohta 1973) posed a challenge to panadaptationism, and combined with early data on knowable at the time, even if imprecisely, parameters such as mutation rates and genome sizes, to the proposal that large portions of the human genome are nonfunctional, "junk DNA" (Ohno 1972). An enormous variety in genome sizes, spanning orders of magnitude, was observed between organisms with similar level of organismal complexity and even between closely related species (Mirsky & Ris 1951; Rothfels et al. 1966; Ohno & Atkin 1966), a discrepancy eventually termed the "C-value paradox" (Thomas 1971). It was best explained by proposing that only a small fraction of the genome consists of genes and other functional sequences. In mammals, it was estimated that the rate of deleterious mutations is $\sim 10^{-5}$ per locus, and that the size of the human genome is $\sim 3 \times 10^9$ base pairs. Given these numbers the maximum number of human genes was evaluated to be $\sim 30,000$ and the fraction of the human genome occupied by genes and their regulatory elements to be ~6% (Ohno 1972). Notably, these numbers are remarkably close to what was found when the whole human and mouse genomes were sequenced, annotated and compared (Lander et al. 2001; Venter et al. 2001; Mouse Genome Sequencing Consortium 2002; Harrow et al. 2012), and by more recent efforts to identify the sequence-constrained elements in a much wider collection of sequenced mammalian genomes (Lindblad-Toh et al. 2011). Also, contemporary studies on the mutation rate in the human genome using more sophisticated measurement tools have largely corroborated the old estimates for the values of the key population genetic parameters of our lineage (Lynch 2010b; Keightley 2012).

The concept of "junk" DNA was further strengthened by the improved understanding of the nature of selfish transposable DNA elements (Doolittle & Sapienza 1980; Orgel & Crick 1980), introns and pseudogenes and, eventually, by the fraction of the fully sequenced genomes they occupy. A total of at minimum 45% of the human genome consists of transposable elements (mostly decayed copies), and close to half of it is introns (according to the most comprehensive currently available annotation, GENCODE, Harrow et al. 2012; note that introns, of course, contain many transposons so these are overlapping sets).

As genome sequencing costs went down with continuous improvements in technology and an ever increasing number of sequenced genomes became available, it has in recent years become possible to place our knowledge about the genomes of humans and the few key model organisms in the context of a much wider phylogenetic sampling. This has enabled the comprehensive assessment of the driving forces of genome evolution across the tree of life. A pluralistic view of evolution, in which the nonadaptive evolutionary forces play a major role, has emerged from this research program (Lynch 2007c; Koonin 2011). Nonadaptive explanations for the evolution of a large number features of genomic organization and gene expression regulation that are fundamental to eukaryotic biology have been proposed based on the integrative analysis of the selective and mutational pressures influencing their evolution and the population genetic environments of different lineages (Lynch 2002; Lynch & Conery 2003; Lynch 2005; Lynch 2006a; Lynch 2006b; Lynch 2007a; Lynch 2007b; Lynch 2007c; Koonin 2011). The following are most relevant to the debate about the evolutionary forces that have shaped mammalian genomes.

16.3.1 Transposable element content

Transposable elements have had a major influence over the evolution of eukaryotic genomes. Their role has sometimes been "constructive", in cases when individual transposable element insertions have been later exapted into novel regulatory and other functional elements, of which a number of examples have been documented in various species (Rebollo et al. 2012; Chen et al. 2009; Krull et al. 2007; Lynch et al. 2011; Medstrand et al. 2001; Naito et al. 2009; Peaston et al. 2004; Schmidt et al. 2012; Santangelo et al. 2007; Bejerano et al. 2006; Kunarso et al. 2010; and many others). Global roles of transposable elements in cellular processes have also been described, for example the upregulation of *B2* SINE repeats in mouse and of *Alu* elements in humans upon cellular stress and their role in the subsequent global repression of transcription (Singh et al. 1985; Espinoza et al. 2004; Allen et al. 2004; Fornace & Mitchell 1986; Li et al. 1999; Mariner et al. 2008; Liu et al. 1995). Nevertheless, the overall effects of transposable elements on organismal fitness are negative, which is evident by the existence of vitally important, dedicated to their silencing and the prevention of their expansion systems, such as piRNAs (Aravin et al. 2006; Aravin et al. 2007a; Aravin et al. 2007b; Guzzardo et al. 2013). The detrimental effects of actively transposing repetitive elements are obvious, as they can insert into and disrupt the function of genes, but even decayed copies confer a slight selective disadvantage as they increase the size of the mutational target in the genome (Lynch 2007c). The same mechanisms that lead to the exaptation of transposons into novel regulatory elements can also lead to the misregulation of the expression of important genes.

From the point of view that all of the content of genomes is adaptive, it would therefore be expected that either genomes should contain no transposons (as they would be weeded out by natural selection) or that all transposon insertions would have functional roles. However, genomes display a wide variation in their transposable element content, which is not straightforward to explain under that model. As a rule, very few transposable elements are found in prokaryote genomes. In contrast, on average a much larger fraction of eukaryote genomes is occupied by transposons, and a clear trend is observed from unicellular to large multicellular eukaryotes, with transposable elements comprising a small portion of the genomes of the former (and in some rare cases being completely absent; Gardner et al. 2002) and a significant part of the genomes of the latter, sometimes even the majority. The maize genome, for example, consists of 85% transposons (Schnable et al. 2009), and the extremely large genomes in the tens and hundreds of Gbs range, which have until very recently been almost impossible to completely sequence, likely contain even more transposable elements (for example, Nystedt et al. 2013). These variations in transposable element content are readily explainable by taking into account the population genetic environment of different lineages. The long-term effective population size N_e is typically $\geq 10^9$ for prokaryotes, $\sim 10^7$–10^8 for most single-celled eukaryotes, in the neighborhood of 10^6 for small invertebrates and annual plants, and in the 10^4–10^5 range for large multicellular organisms such as mammals and trees (Lynch 2006b; Lynch 2007c). Across the tree of life, an inverse correlation is observed between the abundance of transposons and N_e, which makes sense considering that the selective coefficient for each individual insertion is negative but small in absolute value, thus they are visible to selection only in lineages with large N_e, in which natural selection is highly efficient, such as prokaryotes, while they are free to proliferate in lineages with a low N_e, such as humans and other mammals (Lynch 2007c).

16.3.2 The number and length of introns

The presence of introns is one of the most remarkable features of eukaryotic gene expression but their existence also presents us with the puzzle of why eukaryotic genes have them in the first place. The presence of introns poses numerous challenges to the proper expression of genes as they have to be properly spliced out, which, as is the case with all biochemical processes, cannot be relied on to occur with absolute efficiency, and in addition, depends on the presence of additional functional sequence elements to direct it. These elements can be and often are a subject to mutations that disrupt proper splicing with detrimental effects to fitness, as demonstrated by the large number of human genetic diseases that are due to mutations affecting splicing (Cooper & Mattox 1997; Douglas & Wood 2011; Lynch 2006b). In this context, it is not clear why introns exist at all, as gene expression would be carried out with significantly less trouble and more faithfully without them. A commonly cited reason for this is the expansion of the protein repertoire afforded by alternative splicing, however, while there certainly is a lot of complexity in the splicing products generated in mammals, it is far from clear how much of it represents actual functionally important alternative splicing events (to be discussed in more detail later). This explanation also fails to account for the observed differences in the distribution of the number of introns and their length across the tree of life. Spliceosomal introns are restricted to eukaryotes and absent from prokaryotes. The latter instead contain self-splicing introns but those are few in number in each prokaryotic genome, and even they seem to be absent from archaea with a few exceptions that might be the result of horizontal gene transfer (Dai & Zimmerly 2003). Within eukaryotes, there are extremely large differences in intron content, from the nearly intron-free genomes of single-celled organisms such as *Encephalitozoon cuniculi* (Katinka et al. 2001) to the long and numerous introns of mammals and many green plants. A popular in the past explanation for the existence of introns was that they appeared very early in the evolution of life and that genes were pieced together from individual exons, each of which might have carried a separate protein domain or some other functional unit. This has been known as the "introns-early" hypothesis (Gilbert 1978; Gilbert 1987; Doolittle 1978; Darnell 1978; Blake 1979; Gilbert et al. 1997; de Souza et al. 1996) and is somewhat corroborated by the observation that protein domains are indeed sometimes encoded by separate exons (Roy et al. 1999; Fedorov et al. 2003) but this is far from true for all exons, and it is, once again, difficult to reconcile with the complete absence of spliceosomal introns in prokaryotes. A more likely and consistent with data scenario for their evolution has finally emerged in recent years (Koonin 2006; Martin & Koonin 2006). Numerous studies have shown that the last common ancestor of eukaryotes (LECA) was very intron-rich as many intron positions are shared between deeply diverging branches of the eukaryote tree such as plants and animals (Figure 16.1), suggesting a common origin. Subsequent intron gains have been largely limited to individual lineages while many other clades have primarily experienced intron losses (Carmel et al. 2007a; Carmel et al. 2007b; Collins & Penny 2005; Csuros et al. 2011; Fedorov et al. 2002; Rogozin et al. 2003; Rogozin et al. 2005; Roy 2006; Roy & Gilbert 2005). In the same time, it has long been known that structural similarities exist between the self-splicing Group II introns found in prokaryotes and sometimes in eukaryotic organelles (Cech 1986; Lambowitz & Zimmerly 2004) one one side, and spliceosomal RNAs on the other, strongly suggesting that the current spliceosomal splicing system of eukaryotic evolved from ancestrally self-splicing introns, which eventually lost the ability to self-splice leading to the evolution of mechanisms to ensure their proper splicing in *trans*. The intron-rich nature of the LECA might have been due to a wave of Group II intron insertions associated with the ancient endosymbiosis event between an archaeal or archaea-like prokaryote with the α-proteobacterial ancestor of mitochondria (Koonin 2006; Martin & Koonin 2006).

Whatever the mechanisms of their initial establishment in eukaryotes, the subsequent evolution and current distribution of introns within them, and their minimal presence in prokaryotes are well explained by the interplay between mutations, selection and the population genetic environment of different lineages (Lynch 2002; Lynch 2006b; Lynch 2007c). The negative effect on fitness of introns is dependent on the number of base pairs n that are critical for their proper splicing and on the mutation rate μ, which accounts for the probability of their inactivation (Lynch 2002). This implies that introns are only going to be "visible" to and removed by natural selection when $1/n < N_e\mu$ (Lynch 2006b). Organisms with low values of N_e often have elevated mutation rates (Lynch 2010a; Sung et al. 2012) but the values of $N_e\mu$ in these lineages are still comfortably below this threshold, while prokaryotes and some eukaryotes with enormous effective population sizes are well above it (Lynch 2006b; Lynch 2007c).

16.3.3 Variation in genome size

.

Another parameter that varies widely across the tree of life is the total size of the genome. Prokaryote genomes are very compact, with the largest ones known barely exceeding 10Mb (Dagan et al. 2013; Chang et al. 2011). This is still smaller than even the smallest genomes of free-living eukaryotes (Derelle et al. 2006), with typical genome sizes for unicellular eukaryotes in the range of tens of Mbs. Invertebrate genomes are on average hundreds of Mbs while vertebrate genomes are typically a few Gb in size with extreme examples of tens or even hundreds of Gb also known (Gregory et al. 2007). Similar extent of variation in genome size is observed in land plants.

There are several different ways in which genomes can expand. This can happen through the expansion of transposable elements, through the proliferation and lengthening of introns, through the expansion of other non-coding DNA, and through the duplication of genes (these are, of course, not mutually exclusive - introns, for example, often contain numerous transposon insertions). Of these, the duplication of genes seems to have been a relatively minor component as the number of genes only varies over one to two orders of magnitude and the average mRNA length does not vary much between species. Most of the variation in genome size across the tree of life is accounted for by differences in transposon content, intron numbers and length and the amount of other non-coding DNA, with transposons being the most significant contributor. How mutation and genetic drift have shaped the distribution of transposons and introns in eukaryotes was discussed above but it should be noted that expansion of other non-coding DNA is also thought to carry a slight negative fitness cost due to the increase in the size of the mutational target it represents (Lynch 2006b; Lynch 2007c). The increase in genome size in some eukaryote lineages can then be thought of as a direct consequence of their low effective population size (indeed, as with transposons and introns, a negative correlation between N_e and genome size is observed). In the absence of strong selection acting on mutations with small selective effects, genomes are free to expand provided the balance of mutational forces (the rate of small insertions and transposable elements insertions versus the rate of deletions) is in that direction (Petrov 2002). It has to be noted that it is possible that the story is more complex - eukaryotes have not been able to rid themselves of transposons through natural selection on the level of individual transposable element insertions, but they have developed systems for repressing their expression and proliferation (Aravin et al. 2007b), leading to a decrease in the selective disadvantage of individual insertions, greater tolerance to their presence, and, somewhat paradoxically, likely opening the door for their further proliferation (Fedoroff 2012).

16.3.4 The expansion in regulatory and organismal complexity

.

Gene regulation in multicellular eukaryotes is very complex on multiple levels, in contrast to the situation in prokaryotes and the yeast species studied so far (the only unicellular eukaryotes for which detailed understanding of gene regulation has been worked out so far). In the latter organisms, the expression of genes is typically controlled by short regulatory regions proximal to the transcription start site. In marked contrast, in addition to promoters, multiple other regulatory elements control the expression of genes in eukaryotes, many of them situated at great genomic distance from promoters. These regulatory elements serve as binding sites for and integrate the input of multiple transcription factors. Transcription factors in turn, form highly complex gene regulatory networks (GRNs), especially during development (Davidson 2006; Carroll 2008). Notably, different regulatory elements can be responsible for the expression of the same gene in different cell types/tissues and rewiring of GRNs has been a major mechanism behind the diversification of metazoans in the past (that being a major result of many years of evo-devo research).

This is the basic picture that has been known for some time; how our understanding of it has been altered by genomics data will be discussed later. Here, it should be noted that a major question from an evolutionary perspective is how this level of complexity came to be. It has often been by default assumed that it is adaptive but solid arguments have been proposed for why this might not be the case (Lynch 2007a; Lynch 2007b). It is indisputable that increases in regulatory complexity have led to adaptations of organisms to their environment. However, first, it is far from clear that the same result could not be achieved with significantly less convoluted in their workings systems, and second, the genomic changes that lead to this complexification are not themselves adaptive (Lynch 2007a; Lynch 2007b). There are several mechanisms through which regulatory complexity can increase. First, as previously mentioned, transposable element insertions can lead to the generation of new regulatory elements. Second, novel such elements can arise *de novo*. Third, duplication of existing regulatory elements followed by functional divergence can lead to the evolution of new regulatory functions. The latter is in a way similar to what happens during the evolution of paralogous genes, one possible fate of which is described by the divergence and subfunctionalization model (Force et al. 1999; Force et al. 2005), in which following duplication of a gene carrying out multiple functions in the cell, each duplicate copy is free to lose some of them as long as the other retains that functionality, leading to the system being locked in a state in which both copies are essential (it is, of course, also possible for paralogs to become neofunctionalized, acquiring new functionalities not present in the ancestral gene). Something similar might be happening with regulatory elements: following duplication of an initial enhancer responsible for the expression of a gene in multiple tissues, the new copies diverge or acquire new functions, by losing and/or accumulating new transcription factor binding sites, with the end result being that the expression of the gene is driven by different enhancers in different cell types, or that the gene is expressed in cell types, in which it previously was not. This might in the end be adaptive, but importantly, the series of genomic changes in all three types of events that lead to increase of regulatory complexity are not - they in fact have slightly negative effects on fitness due to the increase in the size of the mutational target they represent (Lynch 2007a) and would therefore be expected to be weeded out if the power of natural selection is sufficiently strong. Indeed, this seems to be the case in lineages with large N_e, in which this condition is met - prokaryotes and many unicellular eukaryotes have streamlined genomes in which genes are regulated by promoter-proximal elements occupying limited amounts of genomic real estate, in stark contrast to the situation in mammals.

This insight fundamentally changes the way we view the evolution of complexity in biological systems given the close relationship between increases in regulatory complexity and corresponding increases in organismal complexity. Traditionally, complexity is seen as adaptive, but it seems that in fact the main reason it has evolved is that because it could, in conditions of sufficiently low effective population sizes to allow it, through constructive neutral evolution mechanisms (Stoltzfus 1999; Stoltzfus 2012; Speijer 2011; Lukes et al. 2011; Gray et al. 2010), rather than as the direct result of adaption. Of course, there is a positive feedback loop operating here – the population genetic environment most conductive to this kind of evolution is typical for large-bodied multicellular lineages, for basic ecological reasons having to do with their physical size and the resource requirements it imposes. But large-bodied multicellular lineages are also the ones that would be expected to be most "complex" in their organization, and in turn complex body plans are often conductive to increases in body size and lowering of the effective population size. This likely also explains why no prokaryotes ever evolved multicellularity – their large effective population and the resulting very strong purifying selection to which they are subjected made impossible the complexification of gene regulation (and possibly gene content too) necessary for it. In contrast, the lowered compared to prokaryotes N_e of unicellular eukaryotes allowed in some lineages an evolutionary ratchet of paired increases in size and complexity and further lowering of N_e to take place, leading to the eventual evolution of complex multicellular organisms such as humans.

From these lines of observations and thinking, a very different view of genome evolution and complexity has emerged (Lynch 2007c; Koonin 2004; Koonin 2009; Koonin 2011), in which the interplay between selection, mutation and drift is central, and the major role that nonadaptive processes seem to have played in the evolution of complexity is prominently featured. Lineages with large effective population sizes tend to be small in size and with streamlined genome and this is the dominant mode of evolution (Wolf & Koonin 2013) as it is these lineages that comprise the majority of the diversity of life on the planet (Figure 16.1). In contrast, complex large-bodied organisms have large genomes, with lots of non-coding DNA, large transposable element content, and complex gene regulation, i.e. the have genomes existing in what has been referred to as a "highly entropic" state (Koonin 2011), in which the informational content per unit of DNA is low. The appearance of these traits is linked to the emergence of organismal complexity, but is not a primary causative agent for it. In fact, it would of course, be quite remarkable that were thus proven to be otherwise, natural selection has been unable to drive the accumulation of such embellishments in the lineages, in which it is strongest, as is well known from firmly established population genetics principles. Many other adaptive explanations for the expansion of ncDNA within multicellular (such as buffering against mutations, role in chromosome structure, selection for nuclear and cell size, and numerous others; Vinogradov 1998; Yunis & Yasmineh 1971; Zuckerkandl 1976; Zuckerkandl 1977; Comings 1972; Cavalier-Smith 1978; Cavalier-Smith 2005; Patrushev & Minkevich 2006; Beaton & Cavalier-Smith 1999; Gall 1981) are also usually similarly inconsistent with this reasoning. The nonadaptive understanding of how the human genome evolved to its present state places the results of the ENCODE project and functional genomic data in general in a dramatically different perspective.

16.4 The cultural context of the debate

As an important side note, the larger cultural context of the debate has to also be mentioned, as the panadaptationist point of view of the human genome (as well as any claim that a radically new theory of evolution overturning the old "dogma" has been developed, whether it is because of the impact of epigenetics, ncRNAs, evo-devo, lateral gene transfer, mechanisms for directed adaptive mutations of an almost Lamarckian kind (Koonin & Wolf 2009), or something else (examples in Shapiro 2002; Shapiro 2009; Shapiro 2013) has unfortunately been coopted by various creationist groups, especially Intelligent Design proponents (see Dembski 1998; Behe 2003; Wells 2011 for examples). The idea that large portions of the human genome consist of nonfunctional and even selfishly propagated and slightly detrimental to an organism's fitness DNA does not sit well with the belief that it was the product of a benevolent intelligent designer, both because it implies and provides more evidence for evolution, and because of the theological implications of such a genome if it was in fact designed in that form. For such reasons, creationists have vehemently attacked the concept of "junk DNA" (Walkup 2000; Wieland 1994; Woodmorappe 2000; Bergman 2001; Jerlström 2000) and thus any portrayal of all of the genome as being functional (see for example von Sternberg 2002 and von Sternberg & Shapiro 2005; Grossmann 2013), and more recently the public portrayal of ENCODE results, both of the pilot phase (ENCODE Project Consortium 2007) and especially the first genome-wide production phase (ENCODE Project Consortium 2012), have been warmly welcomed by them (Wells 2013). This outcome serves to emphasize the importance of the precise and clear communication to the public of the most rigorous scientific understanding of genome function, otherwise there is a real danger that great harm may be done to science education and the public understanding of science, areas the current state of which is already constantly decried (for example, it has remained the case for many decades that nearly half of the population of the United States completely rejects both the theory and the fact of evolution; Miller et al. 2006), with the corresponding long-term consequences for society.

16.5 ENCODE results and their interpretation

It is not easy to summarize the results of the ENCODE project in a few sentences as its greatest contribution to date is probably the large number of individual interesting stories rather than the emergence of overarching previously unrecognized themes. Still, the sheer scale and comprehensiveness of the data helped shed light on a number of issues previously debated but not fully resolved (which does not mean all of them have in fact been conclusively resolved). I list some of the most contentious issues below and discuss the proper (and improper) interpretations of ENCODE data with respect to them.

Before I do this, I have to point out that the controversy around ENCODE and junk DNA seems to have arisen mainly due to the large amount of writings and commentaries about ENCODE and a few misinterpreted passages within the main integration paper most of them have been based on, not on the actual content of the numerous ENCODE papers. The integrative paper (ENCODE Project Consortium 2012) states (emphasize mine):

> These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions
>
> . . .
>
> **Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure)**
>
> . . .
>
> The vast majority (80.4%) of the human genome participates in at least one biochemical RNA– and/or chromatin-associated event in at least one cell type

Keeping this definition in mind, there can be little controversy about the claim that >80% of the genome has been assigned a "function", as clearly the word "function" was used in a way different from how biologists have traditionally understood it (Doolittle 2013). Regrettably, the definition was separated from the 80% figure in the writings, press releases and commentaries about the project and the story communicated to the public and the scientific community was that ENCODE has debunked the concept of "junk DNA" (Pennisi 2012; Hurtley 2012; Kolata 2012; Jha 2012; Hall 2012a; Hall 2012b; Harmon 2012; Brown & Boytchev 2012). However, the findings of ENCODE are in no way in contradiction with the concept of "junk DNA", they can be comfortably interpreted within the non-adaptive framework of understanding genome complexity described above, and in certain ways, they actually corroborate many of its components.

16.5.1 Pervasive transcription

The reports of pervasive transcription in mammalian genomes in the early 2000s were all based on microarray technology, which is well known to suffer from a number of issues regarding its resolution, dynamic range and noise levels (Royce et al. 2005; see discussion in Van Bakel et al. 2010 and Clark et al. 2011). The advent of high-throughput sequencing and the development of RNA-seq eliminated a lot of these issues, providing base pair-resolution digital readout of transcriptional products spanning nearly as many orders of magnitude of expression levels as the sequencing depth of the dataset.

The pilot phase of the ENCODE project was also microarray-based, as this was the only technology available at the time, and focused on only 1% of the human genome. (ENCODE Project Consortium 2004; ENCODE Project Consortium 2007). It delivered a message of pervasive transcription, however, because of the aforementioned issues with arrays and also because only 1% of the genome was visible to it, what the results of an in-depth sequencing-based transcriptomic study of the whole genome would be was of great interest. The genome-wide production-phase of the project involved both the sequencing of polyA-selected RNAs (which has traditionally been the most widely used approach for studying the transcriptome as the presence of a polyA tail is a common feature of mRNAs and many lincRNAs), as well as the sequencing of the polyadenylated and non-polyadenylated RNAs from various subcellular fractions (primarily the nucleus and the cytosol, plus the chromatin, the nucleolus and the nucleoplasm in some cells) representing the less studied portion of the transcriptome. This was done across a wide variety of cell types and it confirmed beyond any doubt that the genome is indeed pervasively transcribed (Djebali et al. 2012; ENCODE Project Consortium 2012), with around 75% of the genome being covered by reproducibly detectable transcripts (Kellis et al. 2014).

The reality and functionality of pervasive transcription has generated a lot of controversy over the years (Struhl 2007; van Bakel et al. 2010; Mercer et al. 2011; Dinger et al. 2009; Clark et al. 2011) but there is a clean resolution of it, one that is well supported by the more recent RNA-seq data, and recognized going back to the time of the early C_0t curve studies of the transcriptome: pervasive transcription is indeed real but it happens infrequently, the transcripts it produces are present at quite low levels and in all likelihood the vast majority of transcripts in this expression range have no functional significance. In addition to great sensitivity and base-pair resolution, among the many advantages of RNA-seq over microarrays is also its superior dynamic range (Mortazavi et al. 2008). When the abundance of non-exonic transcript coverage in ENCODE data is examined, it turns out that the majority of it is indeed due to transcripts present in very low amounts, often likely to be significantly less than one transcript copy per cell, and in subcellular fractions other than total cellular polyadenylated RNA. The fraction of the human genome covered by substantially abundant transcripts is between 10 and 30%. This is well above the $\sim 3\%$ of the genome occupied by exons (according to GENCODE), but a good portion of it is due to instances of coverage in intronic regions (which were already known to be transcribed) and of transcription extending beyond the known 3' boundaries of annotated genes, i.e. not examples of dramatically new phenomena. In addition, strong positive correlation between read coverage in RNA-seq data and sequence conservation is observed (Kellis et al. 2014), further corroborating this interpretation.

It is not at all surprising that large portions of a large, "entropic" genome, such as ours, are pervasively transcribed at some point in the life of cells (Struhl 2007). Eukaryotic genomes have to solve the complicated task of properly identifying and regulating promoter regions and transcription start sites within a vast genomic space. To assume that only the annotated, highly expressed, protein coding and non-coding genes, would ever be transcribed, is equivalent to assuming that these organisms have achieved perfection in the area of gene regulation. This goes directly against what we know about the population genetic environment of these lineages. One of the deep insights derived from the non-adaptive, population genetics-centric view of genome evolution and complexity described above has been that organisms can only increase the precision and specificity of biochemical processes to the extent that the power of natural selection allows it. The power of natural selection is inversely correlated with the effective population size N_e and vertebrates have existed in a state of very low N_e for hundreds of millions of years. They are therefore among the organisms for which the least amount of "perfection" in the workings of their biochemical systems can be expected. This has been best studied with respect to the per-generation mutation rate, which is indeed highest in lineages with low N_e (Lynch 2010a; Sung et al. 2012), and there have also been initial studies on the rate of misincorporation of bases during transcription though general conclusions cannot yet be drawn (Gout et al. 2013). Still, the theoretical expectation is that large-bodied eukaryotes with large genomes will turn out to have the lowest transcriptional fidelity per unit of transcribed sequence, including with respect to the specification of sites of transcription initiation. The sequence elements specifying eukaryotic promoters (the TATA box, *Inr*, DPE, BRE, etc.; Lifton et al. 1978; Buratowski et al. 1989; Deng & Roberts 2005; Lagrange et al. 1998; Burke & Kadonaga) are short and degenerate and are far from restricted to annotated promoters. Given this fact, pervasive low-level transcriptional initiation from cryptic promoters, and possibly even the existence of relatively stronger ones producing transcripts with little functional consequence, is something to be expected. This seems to be corroborated by a recent study (Venters & Pugh 2013) which used ChIP-exo-seq (Rhee & Pugh 2011; Rhee & Pugh 2012) to generate high resolution genome-wide binding maps of the TATA-binding protein (TBP), a core general transcription factor involved in transcription initiation, in multiple cell lines also studied by the ENCODE project. It found tens of thousands of TBP binding sites in non-coding regions, many of them containing the promoter-associated sequence elements, producing RNAs, and with chromatin structure similar to that of protein coding promoters. This was interpreted as evidence for widespread functionality of these promoters; however, the TBP occupancy of these sites was markedly lower than that over the promoters of protein coding genes. This explains why they are not readily identifiable with the lower resolution provided by conventional ChIP-seq, and means that such observations are entirely consistent with the majority of them being non-functional. Such understanding is not contradicted by the observation that the low-abundance pervasively transcribed RNAs often exhibit cell-type specificity, as transcription is in general repressed by the presence of nucleosomes. Some chromatin states are more conductive to cryptic transcription than others, and chromatin states do differ between different cell types.

16.5.2 “Exotic” transcription

The pervasiveness of “Exotic” transcripts is also by no means an argument against most of the genome being junk. Some of these (such as the ones normally degraded by the exosome) likely fall in the category discussed in the preceding section. But various RNAs associated with the promoters and termination sites of genes (tiRNAs, TSS-RNAs, PASRs, PALRs, TASRs, etc.) could very well be well-defined RNA species and this still does not serve as a valid argument against “junk” DNA. First, they are associated with the exons and promoters of genes and are therefore mostly conserved at the sequence level. Second they could well be, and in fact likely are, a normal part of the transcriptional cycle, where they may play functional roles or may be an inevitable side product of it (future research will have to establish what, if any, these roles may be).

In addition, transcription can have functions on its own, without the RNA molecules produced being sequence-constrained. Numerous examples of phenomena such as transcriptional interference, where the transcription of other genes, or of noncoding intergenic RNAs, either interferes or aids the expression of downstream genes (through the prevention of initiation or by opening chromatin and enabling it) have been presented in the past (Martens 2004; Petruk et al. 2006; Shearwin et al. 2005; Hirota et al. 2008; Uhler et al. 2007; Kuehner & Brow 2008; Thiebaut et al. 2008; Palmer et al. 2009). The act of transcription in such cases is functional, but the sequence of the transcripts produced may be of little significance.

It is in similar light that RNA species like eRNAs can be interpreted. There is indeed evidence that the transcription of enhancers is important for their function (Ørom et al. 2010; Melo et al. 2013; Li et al. 2013; Lam et al. 2013), but in the absence of detailed mechanistic understanding why (something, which future research will hopefully elucidate) and of sequence conservation beyond the transcription factor binding sites within the enhancer, such observations are entirely consistent with the production of such RNAs being the functionally important component in the process rather than the RNAs themselves.

16.5.3 lincRNAs

In recent years, long non-coding RNAs have received a great deal of attention, both in the scientific community and outside of it. In popular communications, they have often been portrayed as overturning the foundations of our understanding of how RNA functions in cells. Even though this was most definitely not a main message of it, the subject has often been lumped together with the debate about the ENCODE project thus it is proper to discuss it here too. We are at the beginning of exploring the diversity and functional importance of these molecules, with new examples of the vital biological roles they play in various systems being described constantly and many more are certain to come in the future. However, their existence does not represent such a radical paradigm shift as is often claimed.

First, the novel discovery has been how many of them there are out there, not that they exist. Long noncoding RNAs such as *Xist* and *Tsix* (Borsani et al. 1991; Brown et al. 1991; Lee et al. 1999, *roX* in *Drosophila* (Meller et al. 1997) and multiple others have been known for nearly more than two decades.

Second, at this point there have been multiple studies identifying lincRNAs from RNA-seq data in several mammals (Cabili et al. 2011; Guttman et al. 2010; Derrien et al. 2012; Pauli et al. 2012; Washietl et al. 2014; Necsulea et al. 2014) and they all identify at most around 10,000 putative lincRNA genes. Further sampling of rare cell types will likely yield some more, especially given the generally higher tissue specificity of lincRNAs compared to protein coding genes (Cabili et al. 2011). However, the total of all lincRNA exons still occupies a minor fraction of the human genome (many of them are shorter than mRNAs; the average lincRNA transcript in GENCODE V16 is ∼950bp long, while the average GENCODE V16 transcript of a protein coding gene is ∼1.7kb long), and it is far from clear that all of them will turn out to be functionally important. The expression levels of the typical lincRNA are much lower than those of protein coding genes (with the well-characterized in the past ones being among the most highly expressed). Of course, low expression levels do not necessarily imply absence of function on their own – RNA molecules can certainly play vital functions even at low abundance levels. This is especially true if they act in *cis*. In many such cases only a few copies would be expected to be present at any time, and indeed *cis* mechanisms for their action have been proposed (Koziol & Rinn 2010).

Third, the first comparative studies of lincRNAs within vertebrates have been recently published (Washietl et al. 2014; Necsulea et al. 2014). They found significant evolutionary malleability of the precise splicing patterns of lincRNAs, which together with their generally low levels of sequence conservations suggests that sequence constraint may exist for only some portions of these transcripts, and they also observed that significant fraction of human lincRNAs that are specific to our species, with these human-specific lincRNAs exhibiting significantly higher repeat content. There is evidence that on average these lincRNAs are subject to positive selection in the human lineage (Washietl et al. 2014), and some of them undoubtedly are, but overall these patterns are also consistent with an explanation for the existence of many of them as the result of a normal process of birth and death of noncoding genes within intergenic space (especially mediated by transposable element insertions), or from previously protein coding genes (famously, the *Xist* lincRNA seems to have evolved as a result of the pseudogenization of a protein-coding gene, Duret et al. 2006), with some being exapted and conserved throughout evolution and many others eventually decaying.

The detailed functional analysis of each individual lincRNA using classical genetic tools will be need to adequately answer the question of how many of them have function, and what it is. Recently, studies taking the first steps in that direction have appeared; for example Sauvageau et al. knocked out 18 lincRNAs in mouse and found detectable phenotypes for 5 of them (Sauvageau et al. 2013). These numbers, however, cannot be extrapolated for all lincRNAs, first, because the sample size is still small, and second, because lack of phenotype upon knockout in laboratory conditions does not necessarily imply lack of functions (for example, deletion of ultraconserved sequence elements sometimes still results in viable mice; Ahituv et al. 2007). Nevertheless, as whole the reports portraying lincRNAs as completely overturning our understanding of RNA and genome biology are definitely exaggerated.

16.5.4 Alternative splicing and initiation

That the vast majority of human genes has the capacity to and do sometimes produce more than one isoform is at this point beyond dispute. The GENCODE annotation contains nearly 150,000 isoforms of the 20,000 protein coding genes, and more will likely be discovered when a deeper sampling of rare cell types becomes available. However, the functional significance of all this splicing complexity is still unclear as the compendium of actively regulated alternative splicing events of validated functionally is still tiny in comparison with the total number of isoforms. The available data is entirely consistent with the vast majority of isoforms detected in RNA-seq being the result of errors of the splicing machinery, the fidelity of which cannot be expected to be perfect for the same reasons outlined above with respect to pervasive transcription and the transcriptional initiation machinery. One line of possible evidence that functional alternative splicing is indeed a highly prevalent phenomenon would be the detection of widespread regulated switching of isoforms between cell lines, and some evidence in that direction has been presented (Wang et al. 2008).

The ENCODE transcriptome characterization effort (Djebali et al. 2012) found multiple transcripts to be expressed for each gene, with the complexity of expressed splicing products increasing with the complexity of its set of annotated transcript models (i.e. how many isoforms for the gene are present in the annotation), but the question of isoform switching between cell lines was not prominently addressed. Further analysis of some of the same data (Gonzàlez-Porta et al. 2013) concluded that for most genes, one major isoform is dominant and it is consistently most highly expressed across many cell types and tissues. Such results are fully compatible with the interpretation of most of the minor isoforms as noise (Melamud & Moult 2009; Sorek et al. 2004). It should be noted that even if the minor isoforms are highly variable between cell types, this is not strong evidence for their functionality, as this could be the result of the different sets of splicing regulators that these cells express, which could have influence on the splicing of genes other than their primary functional targets.

Thus at present, the question of how many functionally important alternative splicing events there are in the human genome is not fully resolved, and based on all the information we have there are no grounds for claiming that it is so widespread and functionally important that it makes the difference between the human species and "lower" life forms. For it to be adequately addressed, both large-scale experimental advances and detailed study of individual cases will be needed. It should be heavily stressed that all results from transcript-level quantification and assembly efforts based on short-read RNA-seq data are highly contingent upon the ability of the software used in such studies to accurately carry out these tasks. Unfortunately, this is an extremely difficult computational problem and still a major challenge (Steijger et al. 2013; Engström et al. 2013), the only satisfying solution to which will be the advent of long-read sequencing technologies capable of delivering the needed for the analysis of the transcriptome sequencing depths. This will eliminate the need to computationally assemble transcripts from reads much shorter than the length of mRNAs and parse reads between isoforms using statistical methods based on incomplete and sometimes even misleading data due to various read coverage biases in the data. Pioneering efforts in that direction have recently been published (Sharon et al. 2013; Au et al. 2013), but much further progress is needed to fully resolve the issue. Even when this happens though, only a list of candidate events to be further studied will be available, which will then have to be subjected to detailed functional testing to assess their functional importance.

16.5.5 The very large number of putative regulatory elements

Regulatory elements in eukaryote genomes are marked by occupancy by transcription factors or insulator proteins, and are typically exhibit increased DNAse hypersensitivity due to the occlusion of nucleosomes caused by the binding of these proteins to DNA. Global ChIP-seq and DNAse-seq maps of transcription factor occupancy and of DNAse hypersensitive sites in the genome are a highly informative way of mapping putative regulatory elements. The ENCODE Project produced many such maps across a wide variety of cell types (Gerstein et al. 2012; Wang et al. 2012; Thurman et al. 2012; Neph et al. 2012a; Neph et al. 2012b), and they suggest the existence of a very large number of potential distal regulatory elements. The reproducible sites of enriched signal in these assays occupy up to 20% of the genome; however, the resolution of ChIP-seq and DNAse-seq is lower than the footprints of transcription factor binding sites, inflating this number somewhat. Still, ∼5.7% of the genome was occupied by footprints as directly measured by digital genomic footprinting (DGF), the high-resolution version of the DNAse assay, which allows more precise identification of DNAse-protected DNA (ENCODE Project Consortium 2012). As with other ENCODE measurements, these results by no mean invalidate the concept that most of the genome is nonfunctional.

First, such observations in fact corroborate the idea that extensive regulatory complexification is facilitated by population genetic environments characterized by very low N_e (Lynch 2007a). As discussed above, the genomic changes leading to expansion of regulatory complexity are not directly adaptive but they can be tolerated if they are not too deleterious relative to the power of drift in a population. This allows regulatory elements to be duplicated or arise *de novo*, then subfunctionalize and/or be coopted in the regulation of nearby genes. Subfunctionalization eventually leads to the gene needing an increased number of regulatory elements for its proper expression in different cell types. Something very similar was observed by a recent study (Kieffer-Kwon et al. 2013), in which distal regulatory elements in mouse ES and B cells were identified. Functional dissection of individual such elements was then carried out by knocking them out using genome editing, and different enhancers were found to be responsible for the expression of the same gene in the two different cell types.

Second, it is far from clear whether all transcription factor occupancy sites identified by high-throughput studies are in fact functional. There are multiple lines of evidence casting doubt on such an interpretation. A broad observation of all ChIP-seq studies has been that transcription factors bind to many sites near genes that are of apparently little relevance to the previously well-established functional roles of the factor (Cao et al. 2010). It is entirely possible that transcription factors bind to regions of chromatin that are in state conductive to their binding, but without having specified selectively important effect on all of them; in fact, this is quite likely given the degenerate nature of the sequence recognition motifs of eukaryotic transcription factors. In addition, a very wide dynamic range of occupancy strength is seen in ChIP-seq assays (Landt et al. 2012), which follows a power-law like distribution with a small number of sites showing very strong occupancy and a very long tail of low-occupancy sites. There is no simple relationship between the strength of occupancy signal and functionality, as both high- and low-signal functional sites are observed, but ChIP-seq and DNAse-seq signal is generally correlated with sequence conservation (Kellis et al. 2014), and studies suggesting that low-occupancy sites in *D. melanogaster* mostly lack enhancer activity have been published (Fisher et al. 2012). It is thus premature to conclude that each and every ChIP-seq or DNAse-seq peak is functionally important without subjecting it to tests for enhancer activity and other functional assays.

That the evolution of regulatory elements in vertebrates is driven in large part by nonadaptive processes seems to be corroborated by recent studies assessing the conservation and divergence of transcription factor binding between species (Villar et al. 2014). Such studies have so far been carried out only in flies (Bradley et al. 2010; He et al. 2011; Ni et al. 2012) and in vertebrates (Loh et al. 2006; Odom et al. 2007; Conboy et al. 2007; Kunarso et al. 2010; Schmidt et al. 2010; Stefflova et al. 2013; Schmidt et al. 2012; Martin et al. 2011; The mouse ENCODE Consortium 2014), which is admittedly a limited sample. Some patterns have nevertheless already emerged: a significantly higher conservation of transcription factor occupancy sites is observed in flies than in vertebrates, with very high rates of turnover found in the latter (Villar et al. 2014), although it should be stressed that in many cases the turnover of binding sites does not translate into turnover of the regulation of their target genes, i.e. often regulatory elements controlling a given gene are lost and replaced by different regulatory elements putatively playing the same role. These changes in occupancy can be mediated by sequence alterations in the recognition sequences targeted by each factor but even more often they are the result in changes in the recognition sequence of other factors occupying the same loci in a combinatorial fashion. The differences between flies and vertebrates can be interpreted as the result of the differences in the population genetic environment of the two groups (Villar et al. 2014). Flies have two orders of magnitude higher N_e than most vertebrate species, the result of which is an order of magnitude smaller genome, a much higher fraction of which is under selective constraint (potentially more than 50% (Siepel et al. 2005; Andolfatto 2005; Drosophila 12 Genomes Consortium 2007), compared to <10% in mammals (Lindblad-Toh et al. 2011). The lowered strength of selective constraint as a result of the low N_e of vertebrates allows for a more rapid evolution of regulatory elements in these lineages. Of note, similar evolutionary factors might be behind the rapid evolution of the lincRNA repertoire in our lineage, as discussed above (Nesculea et al. 2014; Washietl et al. 2014).

It should be explicitly pointed out that the reasoning outlined above concerns the origin of regulatory complexity, not necessarily the current functions of its individual components. States of irreducible complexity, in which all parts of the system are indeed vital for organism fitness, can be achieved via the mechanisms of constructive neutral evolution. It is also true that increases in complexity through nonadaptive means likely facilitate the emergence of organismal complexity as increases in the number of regulatory elements regulating genes allow for their expression in new cells/tissues or the emergence of new cell types. However, it remains true that regulatory complexity itself may not be strictly necessary for increases in organismal complexity – it could very well be the case that a fully functional human organism could be "built" with a much more streamlined and efficient system of regulatory relationships between transcription factors and their targets than the one we observe in our genome.

16.5.6 The fraction of the human genome that is functional

Much scientific and rhetorical effort has been invested into trying to pin down a specific number for the fraction of the human genome that is functional. There are good reasons to think that the obsessive fixation on obtaining a precise number is misguided:

1. Any such number will ultimately depend on the definition of what a "functional element" is. However, it has been in practice impossible to reach universal agreement on such a definition.

2. Among others, one reason for this state of affairs is that "functionality" is not a binary characteristic that a given DNA base pair in the genome either does or does not have. Changes in DNA sequence in different regions of the genome can differ vastly in the magnitude of their effect on phenotypes and fitness. Thus "functionality" is best understood as being continuously distributed, and consequently any estimate for the total amount of "functional" DNA in the genome will be highly contingent upon an arbitrary threshold-dependent definition of what function is.

3. On a most fundamental level, the important question with major implications for how we think about our genome is whether most of it consists of "junk" DNA or not. Most of the human genome indeed does seem to be "junk" DNA and this is true irrespective of whether we estimate the amount of functional DNA to be 5%, 15% or some other number constituting a minority fraction of it.

4. The unfinished (and monumental in its magnitude) task, which does have real importance, is to understand the role of all candidate functional elements in the genome in shaping phenotypes, largely through the classical (though greatly aided and sped up by technological advances such as genome editing and high-throughput functional assays) approaches that have produced the extensive amount of knowledge we have about a handful of loci in humans and some of the major model systems. Agreeing on a precise number for how much of the human genome is functional has little relevance to these efforts as it does not necessarily change the null hypotheses and the priors with which the study of individual regulatory elements and ncRNAs will be approached.

16.5.7 The contributions of ENCODE

The excessive focus on the "junk" DNA debate has overshadowed the real scientific advances that the ENCODE Project has contributed to and it is therefore useful to summarize the major ones here. While doing this, it should be remembered that the ENCODE Consortium was set up with the goal of identifying the functional elements in the human genomes, and not with the goal of finding radically new principles of gene regulation as there was no *a priori* reason to think such mechanisms would be discovered. That no such discovery was made was therefore no surprise and no reason for disappointment; significant progress towards the main goal of the project was made though it has become apparent that reaching it is going to be significantly more complicated and laborious than perhaps hoped for in the beginning. Specifically, ENCODE delivered:

1. **Lists of candidate functional elements**. The complexity of the transcriptome and of the transcription factor binding landscape in the genome, especially when interpreted in the light of the nonadaptive view of genome evolution, means that no candidate functional element identified through a high-throughput functional genomic assays, whether it is an enhancer, a noncoding RNA or an alternatively spliced isoform of a gene, can be considered functional without subsequent confirmation of its significance and dissection of its functional components. This is a necessary activity, without the completion of which a complete understanding of gene regulation in the human genome will be difficult to achieve. Fortunately, while such testing has previously been very labor-intensive, highly parallel reporter assays (Melnikov et al. 2012; Patwardhan et

al. 2012; Smith et al. 2013; Kheradpour et al. 2013; Arnold et al. 2013) and readily applicable genome editing tools (Jinek et al. 2012) have recently become available, promising to greatly speed up the process of validation.

2. **Annotation of noncoding variants associated with human phenotypic variation**. Genome-wide association studies (GWAS) of phenotypic variation in the human population have revealed that the majority of trait-associated sequence variants reside in the noncoding portions of the genome and are preferentially associated with regulatory regions (Hindorff et al. 2009; Nicolae et al. 2010; Zhong et al. 2010). The intersection between the GWAS annotations of such variants and ENCODE maps of candidate functional elements, especially those of transcription factor occupancy, has been (Maurano et al. 2012; Ward & Kellis 2012a; Ward & Kellis 2012b; Boyle et al. 2012; Vernot et al. 2012; Schaub et al. 2012; Hardison 2012) and will be a highly informative source of understanding of the mechanisms through which sequence variation impacts phenotype (examples in Bauer et al. 2013; Hardison & Blobel 2013).

3. **The development of the functional genomics toolkit**. The development of many of the functional genomics assays that have become the workhorses of research in the field was driven by researchers within the ENCODE Consortium (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Kodzius et al. 2006; Fullwood et al. 2009; Johnson & Mortazavi et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007; Hesselberth et al. 2009; Song et al. 2011; Gaulton et al. 2010; Neph et al. 2012a). The experience it has had working with a large number of such datasets and the need to analyze them jointly have led to the development of standardized best practices for their execution (Kharchenko et al. 2008; Landt et al. 2012; Marinov et al. 2014; Jung et al. 2014; Ernst & Kellis 2010; Ernst et al. 2011; Ernst & Kellis 2012; Mortazavi et al. 2013; Hoffman et al. 2013). The Consortium has also pioneered many of the existing tools for integrative analysis of functional genomic datasets (Ernst & Kellis 2010; Buske et al. 2011; Ernst et al. 2011; Ernst & Kellis 2012; Hoffman et al. 2012; Mortazavi et al. 2013; Hoffman et al. 2013; Xie et al. 2013). These methods will serve as a foundation for large scale functional genomics study of many systems in the future as I extensively discuss below.

16.6 The Tree-of-Life ENCODE

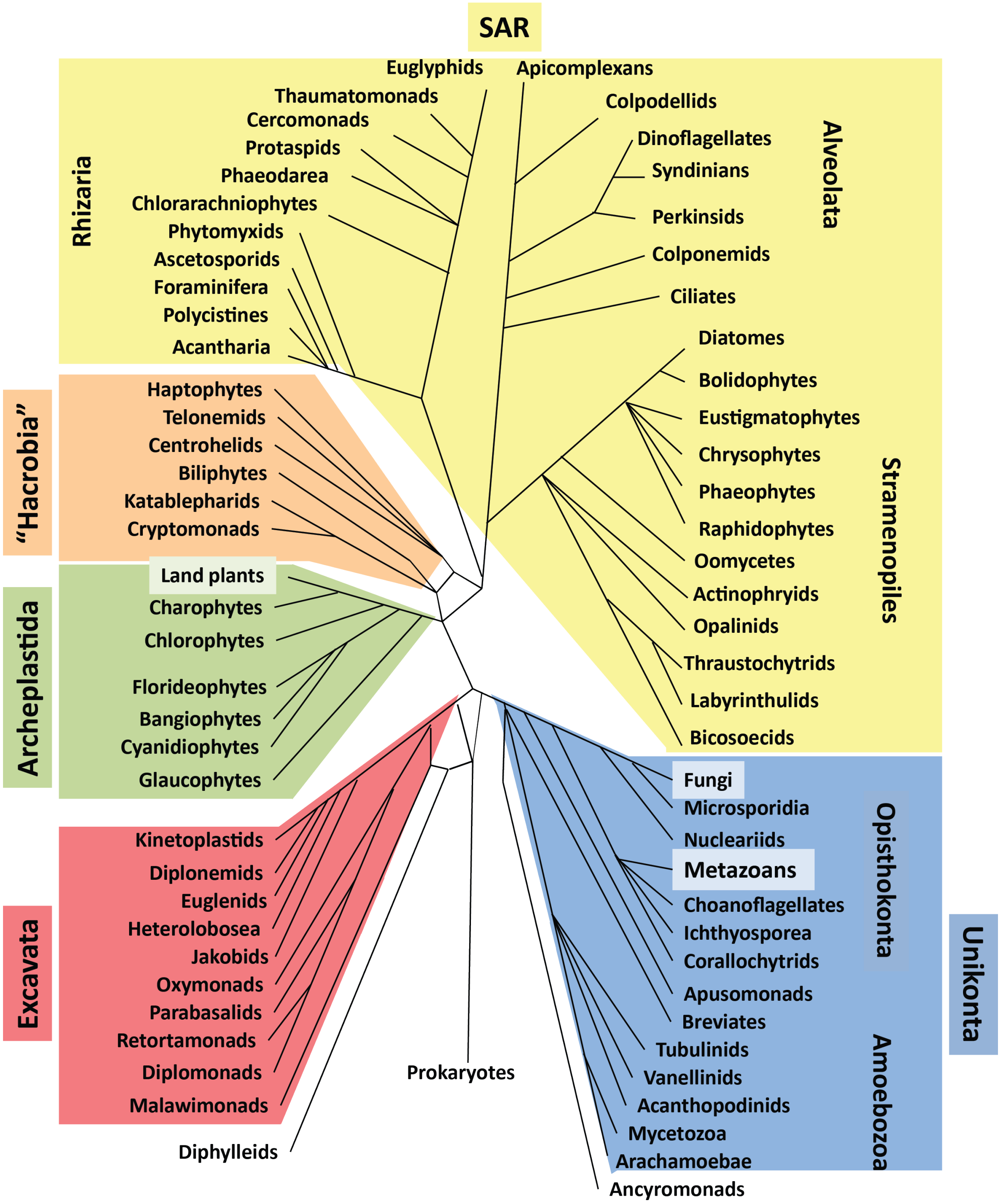
The question in the heart of the debate surrounding the results of the ENCODE Project is the relationship between the complexity of genome architecture and the complexity of organismal organization. Through the lens of panadaptationism, the experimentally demonstrated biochemical complexity of transcriptional regulation, the products of transcription, and of RNA biology, is viewed as an integral causative component agent behind the organismal complexity of humans. This is especially true if the common view of the human species as the highest achievement of evolution is adopted. As already discussed, one way of looking at the relationship of genomic and organismal complexity sees the two as forming a positive feedback loop, in which increased organismal complexity leads to larger organismal size, lowered N_e, and increased tolerance towards further increases in genomic complexity. This in turn may facilitate more regulatory innovations leading to further complexification of organismal organization. However, this is at present only a general trend observed largely based on the comparison of the very general features of sequenced genomes. Even at this level, it remains to be generalized across the whole tree of life – the sampling of completely sequenced genomes is nowhere near complete in terms of coverage of the major eukaryote lineages and the multicellular groups that independently evolved within them – but more importantly, it has not yet been tested by direct biochemical measurements of functional genomic complexity. The integration of the results of the ENCODE and modENCODE and mouse ENCODE projects will provide many insights into these questions. However, all of these species are metazoans and animals are only one of a very large number of deeply diverging lineages of eukaryote (Figure 16.1). In addition to these four major model organisms, the yeast *Saccharomyces cerevisiae* has been subject of extensive functional genomic characterization (Lee et al. 2002), and significant amount of work has been done on the plant model organism *Arabidopsis thaliana*, but even in the latter case a large scale dissection of regulatory complexity has not been embarked on. Thus we have a significant (yet still far from complete) functional genomic knowledge of only a handful of species belonging to only three major lineages (out of many dozens) within two of the five to eight major subdivisions of the eukaryotes (Parfrey et al. 2005; Adl et al. 2012). A major expansion of this list is highly desirable for a number of overlapping reasons discussed below. Fortunately, the work done by the ENCODE Project combined with current technological developments has now enabled such studies. Based on the history of biology in recent decades, there are reasons to believe that they will provide deep insights into these questions, and potentially open up many new research directions.

16.6.1 The any-organism-ENCODE

One of the less appreciated consequences of the advent of next-generation sequencing and the phasing out of microarray technology has been that now any organism is in principle accessible for functional genomic dissection as all that is needed is a sequenced genome, without the need to go through the slow, complex and expensive procedure of generating microarrays for each species. The availability of sequenced genomes is not exactly a solved problems for eukaryotes, especially for those with larger and repeat-rich genomes, where the nature of short-read sequencing has made obtaining anything significantly better than highly fragmented assemblies extremely difficult (Alkan et al. 2011). However, this situation is set to improve considerably with increased throughput from existing long-read sequencing technologies (Eid et al. 2009; Schadt et al. 2010; Kuleshov et al. 2014) and the long-awaited arrival of functioning nanopore sequencing (Kasianowicz et al. 1996; Deamer & Akeson 2000; Branton et al. 2008; Clarke et al. 2009; Cherf et al. 2012; Manrao et al. 2012). It is reasonable to assume that in the coming years it will become possible to assemble at high quality and contiguity all genomes, even the very large ones that are now outside of the realm of the possible and the current gaps in our sampling of the phylogenetic diversity will be filled.

Once a genome is sequenced, the various *seq assays (Wold & Myers 2008) can be carried out on it, and most of them are at this point mature. By their very nature most techniques assaying the occupancy of proteins on nucleic acids are tailored to short-read technologies as the DNA or RNA fragments subjected to sequencing are at most a few hundred base pairs in size. For these reasons long-read sequencing is of little utility to ChIP-seq, DNAse-seq, CLIP-seq and other such assays, and of even less utility to high-resolution versions of them such as ChIP-exo-seq (Rhee & Pugh 2011; Rhee & Pugh 2012). The approaches and methodologies developed so far for processing, quality evaluation, analysis and integration of these kinds of data developed as part of the ENCODE Project will therefore continue to be relevant long into the future.

While DNAse, Hi-C and RNA-seq assays are generic in nature in the sense that no special reagents are needed, ChIP, ChIA-PET and CLIP assays require antibodies specific to the targeted protein. Many histone modifications are highly phylogenetically conserved and the same antibodies can be used in deeply divergent species, but working ChIP-validated antibodies are generally only available for a small fraction of human transcription factors and other chromatin-associated proteins and for even fewer such targets in the major model systems. The advent of genome editing will hopefully alleviate this problem. CRISPR-mediated genome editing (Jinek et al. 2012) has recently emerged as a powerful tool for manipulating genomes and has been successfully used in a very wide variety of systems (Dickinson et al. 2013; Chen et al. 2013; Auer et al. 2014; Jiang et al. 2013a; Jiang et al. 2013b; Mali et al. 2013; DiCarlo et al. 2013; Friedland et al. 2013; Gratz et al. 2013; Hwang et al. 2013; Chang et al. 2013; Jao et al. 2013; Cong et al. 2013; Li et al. 2013a; Li et al. 2013b; Li et al. 2013c; Nekrasov et al. 2013; Shan et al. 2013; Tzur et al. 2013; Waaijers et al. 2013; Wang et al. 2013; Chiu et al. 2013; Lo et al. 2013; Katic & Großlhans 2013; Kondo & Ueda 2013), including for the knock-in of tags such as GFP into endogenous loci (Dickinson et al. 2013; Chen et al. 2013; Auer et al. 2014). Such approaches, when combined with recently developed high-throughput chromatin immunoprecipitation methods (Aldridge et al. 2014; the R-ChIP protocol described ear-


SAR
Rhizaria
Euglyphids
Thaumatomonads
Cercomonads
Protaspids
Phaeodarea
Chlorarachniophytes
Phytomyxids
Ascetosporids
Foraminifera
Polycistines
Acantharia
Apicomplexans
Colpodellids
Dinoflagellates
Syndinians
Perkinsids
Colponemids
Ciliates
Alveolata
Diatomes
Bolidophytes
Eustigmatophytes
Chrysophytes
Phaeophytes
Raphidophytes
Oomycetes
Actinophryids
Opalinids
Thraustochytrids
Labyrinthulids
Bicosoecids
Stramenopiles
"Hacrobia"
Haptophytes
Telonemids
Centrohelids
Biliphytes
Katablepharids
Cryptomonads
Archeplastida
Land plants
Charophytes
Chlorophytes
Florideophytes
Bangiophytes
Cyanidiophytes
Glaucophytes
Excavata
Kinetoplastids
Diplonemids
Euglenids
Heterolobosea
Jakobids
Oxymonads
Parabasalids
Retortamonads
Diplomonads
Malawimonads
Diphylleids
Prokaryotes
Fungi
Microsporidia
Nucleariids
Metazoans
Choanoflagellates
Ichthyosporea
Corallochytrids
Opisthokonta
Apusomonads
Breviates
Unikonta
Tubulinids
Vanellinids
Acanthopodinids
Mycetozoa
Arachamoebae
Amoebozoa
Ancyromonads

lier here) open the door to potentially assaying the whole set of transcription factors of a species, especially in unicellular eukaryotes.

The major area in which significant changes are both expected and needed is transcriptomics. Most RNA molecules are far too long to be sequenced from end to end with current short-read sequencers (with the various small RNA species being the major exception). This has posed immense difficulties for the assembly of full-length transcripts from RNA-seq data, a problem the accurate solution of which is of crucial importance for annotating genomes and for the study of alternative transcription initiation, splicing and other RNA processing events. The ability to generate large numbers of very long reads covering full-length transcripts should solve many of these problems.

Still, many of the tools are already in place to enable the generation of ENCODE-level in their size and scope functional genomic compendiums for pretty much any species of interest, and this can be quite rapid and inexpensive compared to the scale of the effort and investment it took to carry out the first genome-wide phase of the ENCODE and modENCODE projects. Thanks to the continued advances in technology and automation in the near future it will be feasible (both in terms of manpower and in terms of cost) for the comprehensive large-scale functional genomic characterization of a whole organism to be carried out by individual laboratories, especially in the cases of unicellular eukaryotes. It should be noted though, that for the integration of such efforts between laboratories to be possible, standardized protocols and stringent control of data quality will be needed, of the kind that large consortia such as ENCODE have invested significant effort in developing (Landt et al. 2012; Marinov et al. 2014).

The major promise this holds is the ability to learn a lot about the genome biology of previously non-model organisms orders of magnitude faster and cheaper than the decades of effort that had to be invested into accumulating the knowledge we have about the human genome and the genomes of the major model systems. The ENCODE Project did not reveal fundamentally new paradigms of gene regulation and the functional organization of the human genome, however, first, given the depth in which our genome has been studied in the past, it would have been a major, and not entirely pleasant surprise if it had in fact done so; and second, it did recover a lot of what was previously known about it. For example, the integrative analysis of multiple histone marks did return the known, in some cases from detailed mechanistic studies, correlations between these marks and between the marks and various genic and intergenic features in the genome, in addition to finding some new chromatin states that were not recognized before. Of course, not everything about how a genome functions can be learned from high-throughput functional genomics, as evident by the above-mentioned interpretative difficulties we face when trying to assess the functional importance of all observed biochemical activity. Still, it remains true that first, obtaining the list of potential parts is always a major step forward, and second, that the general principles of functional organization can be inferred without the

Figure 16.1 ***(preceding page)*****: Major eukaryotic clades, their established and putative relationships with each other, and the place of the human lineage within them**. The tree is derived from Keeling 2013. A variety of other topologies differing in both minor and major ways have been proposed by other authors and it is likely that the true phylogenetic relationships differ from the ones shown here; continued revisions are therefore to be expected for the foreseeable future, especially with the continued discovery of previously unknown deeply diverging lineages and whole-genome sequencing of representatives of lineages for which genome sequences are not available at present. The clades to which the major model organisms that have been the workhorses of functional genomic research belong are highlighted: Metazoans (*Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*; Fungi (*Saccharomyces cerevisiae*), and land plants (*Arabidopsis thaliana*). These lineages comprise a fairly small portion of the known eukaryotic diversity. Note that not all major clades are shown and that almost certainly not all major clades are even known as new lineages continue to be identified, the discovery of *Chromera velia* as a sister lineage of apicomplexans and its importance to understanding the evolution of parasitism in the latter being a very good example (Moore et al. 2008; Oborník et al. 2009; Dorrell et al. 2014; Weatherby & Carter 2013).

detailed annotation of each and every candidate functional element.

16.6.2 Understanding the biology of crops and pathogens

While the focus of this text is on the origins, evolution and significance of genome complexity, questions the answers to which may necessitate the study of obscure organisms the main claim to fame of which lies in the amazing evolutionary innovations their lineages have come up with, the genomic approaches described in the previous section will be of even greater practical relevance for figuring out the biology of plant crops, eukaryotic pathogens (protists such as *Plasmodium*, *Toxoplasma*, *Leishmania*, *Trichomonas*, *Trypanosoma*, *Entamoeba*, and *Giardia*, various parasitic worms, and numerous others) and any species of importance to humans, the genome of which contains significant amounts of at present poorly annotated noncoding DNA (although it should be noted that the genome biology of many such species is deeply intriguing on its own).

One of the major contribution to understanding human biology that the ENCODE Project has made has been the annotation of some noncoding GWAS variants, as discussed above. This work is by no means finished – the next round of the project should bring us closer to the final goal – but it does represent a pioneering effort in this direction.

To the extent that the same genomic architecture is shared between mammals and plants, it is quite likely that the same problem of a lot of explanatory variants residing in noncoding regions of the genome will be faced by large-scale sequencing studies aiming at understanding the genetic basis of variation between different plant cultivars and between different strains of other economically important species, in proportion with the amount of functional noncoding DNA they possess. Based on ENCODE's experience with the human genome, it can be expected that comprehensive mapping of transcription factor binding sites and other regulatory elements will be needed to understand the trait-associated variants in these genomes, with the approaches developed for tackling these questions in humans providing invaluable help.

Similarly, functional genomic approaches will be of tremendous benefit for dissecting the regulatory biology of pathogen species. This is not only of practical, but also of fundamental biological importance, given that as a rule, pathogens have the most complicated life cycles of all organisms, and the same genome is capable of encoding the development of morphologically very different life forms. While multiple such genomes have already been sequenced, at present knowledge of how their gene regulation intersects with developmental mechanisms remains very limited.

16.6.3 Mapping the rewiring of gene regulatory networks in evolution

A major results of extensive studies in the field of evolution of development (evo-devo) over the last few decades has been that the evolution of body plans seems to be in larger extent the result of changes in the regulation of genes, especially developmental regulators, rather than being primarily due to changes in the gene repertoire of different lineages. Often the same molecules are repeatedly utilized in the development of very different structures, both across the metazoan phylogeny and within the same organism. The rewiring of gene regulatory networks has been at the core of these changes (Davidson 2006; Peter & Davidson 2011). While the detailed functional characterization of individual loci using classical genetic approaches (Davidson et al. 2002a; Davidson et al. 2002b) will remain indispensable, the path towards a complete understanding of the evolution of development will be significantly more easily traveled if the targets of the major developmental regulators are comprehensively mapped and their conservation and divergence during the evolution of different groups studied in detail. Given that regulatory elements are often not conserved on the sequence level (Romano & Wray 2003; Balhoff & Wray 2005; Ludwig et al. 2005; Hare et al. 2008; see also discussion above on transcription factor binding site turnover), functional genomics methods for mapping transcription factor binding sites and other regulatory elements of the kind that the ENCODE Project Consortium has extensively used will be required to accomplish this task. At present such studies face major hurdles due to the lack of suitable immune reagents and the difficulty of obtaining material of sufficient quantities and purity from specific developmental stages and tissues/cell types in many lineages of key interest for understanding metazoan evolution. In some systems, these challenges may remain unsolved for a very long time, yet technological advances in genome editing and in the isolation of specific subsets of cells/nuclei from embryos/tissues (Deal & Henikoff 2010; Steiner et al. 2012; Henryet al. 2012; Southall et al. 2013; Schauer et al. 2013) should make such studies feasible in many others.

16.6.4 Understanding the evolutionary origins and the diversity of eukaryotic regulatory biology

As their answers are what is needed to enable the manipulation of biological systems, the questions asked in biochemical and molecular biology research tend to be of the "what" and "how" kind, i.e. we pick apart the individual components of these systems and identify the relationships between them. From an evolutionary perspective the "why" questions are just as important. Behind a lot of the arguments about the ENCODE Project and what its results mean about our view of our genome stands the question "Why is mammalian regulatory biology the way it is?". A perfectly valid possible answer to this question might be that "this is the only way it could be" and if we did not have any examples of significant deviations from the regulatory principles we observe in our genome, there would be no way to reject that explanation. On the other hand, if such deviations do in fact exist, then we know that there are other ways the system might operate and we are forced to find an explanation for why it has diverged between different lineages. The classic model systems already give us plenty of examples of such deviations from the organization of the human genome. As mentioned above, all known prokaryotes have compact streamlined genomes with little intergenic DNA, no spliceosomal introns, few repetitive elements, genes organized in operons, and in general, very little that could be potentially classified as "junk" DNA and far less of the baroque regulatory complexity of vertebrates. Within eukaryotes, the model yeast species also have small compact genomes, with little intergenic DNA, their introns are fewer in number and short in length, and gene regulation seems to be operating mostly through promoter-proximal regulatory elements. The genomes of *D. melanogaster* and of *C. elegans* are more similar in organization to ours but are still and order of magnitude more compact, and in *C. elegans* many genes are transcribed as polycistronic units and then *trans*-spliced to splicing leader sequences to generate mRNAs.

We do have a general theory that explains many of these differences as a result of the interplay between natural selection, mutational biases and genetic drift (Lynch 2007c). But not all aspects of regulatory biology have been examined through an evolutionary lens, and far from all of eukaryotic diversity has been studied from such perspective. It could well be, and is in fact, highly likely that novel insights into the origins and functions of the core features of eukaryotic transcription regulation and RNA biology will be derived from the comparative study of regulatory mechanisms across the tree-of-life, including all deeply diverging eukaryote lineages that have received little attention so far.

This approach has already proven invaluable in understanding the deep evolutionary origins and functional significance of some core features of mammalian genome biology. A prime example is DNA methylation. The primary role of 5-methylcytosine DNA methylation has been traditionally understood to be repression (Fuks 2005; Miranda & Jones 2007), based on extensive research on the scale of individual genes and the whole genome in mammals and in flowering plants (Lister et al. 2008; Lister et al. 2009). Significant differences in the patterns of methylation have been uncovered between the two lineages (Law & Jacobsen 2010). In mammals, cytosines are methylated in the context of CG dinucleotides, by the *de novo* DNA methylases DNMT3a and DNMT3b (Okano et al. 1998; Okano et al. 1999), and by the maintenance DNA methylase DNMT1 (Bestor et al. 1988). All CG dinucleotides in mammalian genomes are methylated, including gene bodies, except for the so called CpG islands, clusters of elevated density of CG nucleotides associated with the promoters of genes (Bird 1986; Gardiner-Garden & Frommer 1987), in contrast with the rest of genome where CG nucleotides tend to be eliminated as methylated cytosines can undergo spontaneous deamination an turn into thymines. CpG islands are differentially methylated in the context of the developmental repression of lineage-specific genes and methylation is also important for the silencing of transposons but the methylation of gene bodies of less understood significance (Kulis et al. 2013). In addition, whole-genome profiling of 5mC in embryonic stem cells has also revealed that cytosines in the CHG and CHH sequence contexts (where H stands for A, T or G) can also be methylated (Lister et al. 2009). In contrast, in flowering plants, methylation is restricted to transposons and other repetitive elements, where it serves repressive function and occurs in all three sequences contexts. It is deposited de novo by the DRM2 enzyme and maintained by DMT1 in the CG context, CMT3 in the CHG context, and by persistent de novo methylation in the CHH context, with all these proteins belonging to the same family of enzymes (Law & Jacobsen 2010).

Even though 5mC methylation has been lost on more than one occasion (for example, yeast such as *Saccharomyces cerevisiae* and the nematode *C. elegans*), the presence of methylation and DNA methylation enzymes of the same family in very deeply diverging lineages suggests deep evolutionary conservation of the methylation pathway going back to the LECA, where it likely played a role in silencing transposons. However, whether different methylation patterns and functions exist in other organisms had not been clear until several studies in the last few years used genome-wide bisulphite sequencing to profile the genome-wide distribution of 5mC in multiple species, both in the major eukaryotic groups that model organisms belong to and in more deeply diverging lineages (Zemach et al. 2010; Feng et al. 2010; Huff & Zilberman 2014). These studies found some examples of unusual methylation patterns (for example, in the green alga *Chlamydomonas*, non-CpG methylation is highly enriched within the exons of genes) and concluded that both gene-body and non-CpG methylation were ancestral to eukaryotes, with this pattern then undergoing modification in diverging lineages, likely due to differential constraints on TE proliferation experience by them (Zemach et al. 2010).

A more recent study of DNA methylation (Huff & Zilberman 2014) extended the taxon sampling to diatoms (*Phaeodactylum tricornutum*, *Fragilariopsis cylindrus*, *Thalassiosira pseudonana*), the pelagophyte stramenopile *Aureococcus anophagefferens*, the haptophyte *Emiliania huxleyi*, and the prasinophyte chlorophytes *Bathycoccus prasinos*, *Ostreococcus lucimarinus*, and *Micromonas pusilla*. Remarkably, it found a completely novel methylation pattern in some of these species (*A. anophagefferens*, *E. huxleyi*, *B. prasinos*, *O. lucimarinus*, and *M. pusilla*) characterized by DNA methylation of CpG dinucleotides situated in linker histone regions, with a periodicity corresponding to the length of nucleosome spacing in each species. CpG methylation was coupled to and directly influenced a correspondingly tight nucleosome positioning pattern, as the methylated cytosines disfavor the formation of nucleosomes. Even more remarkably, CG dinucleotides were actually enriched in nucleosome linker regions and overall in the genomes of these organisms, contrary to what is observed in the genomes of most other lineages (where CG nucleotides are typically depleted due to the spontaneous deamination of methylcytosine) indicating that they are subject of active maintenance by selective forces. Another important surprise was that these novel DNA methylation patterns were generated by a different DNA methylase, DNMT5, which is of the same family as the DNMT3 and DNMT1 enzymes (Ponger & Li 2005), while DNMT1, and often DNMT3 too, is not present in their genomes. Given that DNMT5 is found in very deeply diverging groups of eukaryotes, it is likely that the last common ancestor eukaryotes contained both DNMT1 and DNMT5. Apparently, DNMT5's enzymatic activity is highly biased against methylating nucleosomal DNA, which explains its preference for nucleosome linker DNA. It has been known for quite some time that nucleosomes disfavor methylation (Robertson et al. 2004; Gowher et al. 2005; Takeshima et al. 2006; Takeshima et al. 2008; Felle et al. 2011; Jiang et al. 2011; Kelly et al. 2012), and specific amino acids changes in Dnmt3b have been identified that confer enhanced nucleosome methylation ability in mammals (Shen et al. 2010). In contrast, Dnmt5 seems to have evolved in the opposite direction, disfavoring nucleosomes to an much greater extent, and candidate amino acid residues responsible for this shift in preferences were identified (Huff & Zilberman 2014).

This linker histone methylation pattern observed in these organisms has been interpreted as arising due to selective pressure towards compactness of the nucleus. All of these species are marine algae, the lifestyle of which favors small cell sizes and compact nuclei (as this might confer enhanced light absorption and quicker growth). Using DNA methylation to position nucleosomes was suggested to eliminate the need for bulky chromatin remodeling complexes (Huff & Zilberman 2014). Whether this is the case remains to be confirmed by future studies (and is, of course, a hypothesis that is well suited for testing by identifying the components of the chromatin remodeling complexes in the genomes of these organisms and globally mapping their genomic occupancy). Nevertheless, these results are highly intriguing for a few reasons:

1. A completely novel DNA methylation pattern was found, one that has apparently evolved independently and convergently in deeply diverging eukaryote lineages

2. This pattern is governed by a previously poorly characterized member of the DNMT family.

3. It provided further insight into the relationship between DNA methylation and nucleosome positioning, including in mammalian systems .

4. A potential connection between evolutionary constraints on cell and nuclear size and genome architecture was identified.

5. A plausible explanation for the high genomic GC content in these species was found, in contrast to most eukaryotic genomes, which tend to be AT-rich.

We can be reasonably certain that many other such surprises await discovery in the genomes of unicellular eukaryotes, and they will have significant impact on our thinking about the core features of mammalian regulatory biology. Several instances of the evolution of radical departures from the standard model of eukaryotic genomic organization and gene regulation are already known, although they have rarely been studied in detail.

The genome biology of ciliates provides one such example. Ciliates are unicellular eukaryotes belonging to the alveolates clade, together with two other major groups, the dinoflagellates and the apicomplexans (Figure 16.1). The ciliate *Tetrahymena thermophila* has been a model system for many decades, the study of which has resulted in a number of fundamental biological discoveries, such as the discovery of self-splicing RNAs (Kruger et al. 1982), the relationship between histone acetylation and gene activation (Brownell et al. 1996), telomerase (Greider & Blackburn 1985), and others. One of the defining features of ciliates is the presence of a macronucleus and a micronucleus. The micronucleus is diploid and transcriptionally inert while the macronucleus is highly polyploid and is where gene expression takes places. The micronucleus can divide mitotically and meiotically and in effect constitutes the “germline”, while the “somatic” macronucleus divides amitotically (Wolfe 1967; Ammermann 1971), with no known mechanisms of guaranteeing equal separation of genetic material; instead its high polyploidy is what ensures that each daughter macronucleus receives the full set of genes. The most striking feature of this system is that the macronucleus is derived from the micronucleus through a complex process involving the excision of large portions of the micronuclear genome, from 20-30% in some ciliate groups, such as *Tetrahymena* and *Paramecium*, to more than 95% in others such as *Euplotes*, *Stylonychia* and *Oxytricha* (Jahn & Klobutcher 2002). The excised fragments (called internal eliminated sequences or IESs) are both often similar to transposons in structure and generally enriched for transposable elements (Baird et al. 1989; Wuitschick et al. 2002; Fillingham et al. 2004). The process of IES elimination is dependent on transposase enzymes (Baudry et al. 2009; Cheng et al. 2010; Nowacki et al. 2009) and is carried out through complex small RNA- (Mochizuki et al. 20012; Mochizuki & Gorowsky 2004; Mochizuki & Gorowsky 2005; Aronica et al. 2008; Lepere et al. 2008; Lepere et al. 2009; Schoeberl et al. 2012; Fang et al. 2012; Zahler et al. 2012) and long RNA-mediated (Prescott et al. 2003; Nowacki et al. 2008) epigenetic mechanisms that guide their excision.

The macronuclear genomes of several ciliates have been sequenced: *Tetrahymena thermophila* (Eisen et al. 2006), *Paramecium tetraurelia* (Aury et al. 2006), *Ichthyophthirius multifiliis* (Coyne et al. 2011), and most recently, *Oxytricha trifallax* (Swart et al. 2013). From these and prior studies, multiple differences in the genome reduction patterns and mechanisms between the different ciliates lineages have emerged. In all species, the elimination of IESs results in the fragmentation of the micronuclear chromosomes into smaller micronuclear chromosomes, a process that involves the addition of new telomeric sequences to the ends of the new chromosomes (Nowacki et al. 2009). However, while macronuclear chromosomes are still a relatively small number in *Tetrahymena* (225), *Ichthyophthirius* (71) and *Paramecium* (∼200), each of them contains many genes and is generally organized like a typical eukaryotic chromosomes.

In contrast, in *Oxytricha* and in related ciliates such as *Stylonychia* and *Euplotes*, the macronuclear genome exists in the form of thousands of nanochromosomes (Lawn et al. 1978; Swanton et al. 1980; Swart et al. 2013), which in the *Oxytricha* macronuclear genome contain just a single gene, rarely 2 or more (Swart et al. 2013). Not only that, but in *Oxytricha* IESs are not just excised, but the genes exist in a scrambled nonlinear form in the micronucleus and have to be "unscrambled" and put back together in the correct order when the macronucleus is formed (Prescott 1999; Fuhrmann et al. 2013). This processes can lead to the formation of alternative nanochromosome isoforms for the same gene, some of which are incomplete and likely nonfunctional (Herrick et al. 1987a; Herrick et al. 1987b; Klobutcher et al. 1988), and even allows for the possibility of "alternative DNA splicing" (Fass et al. 2011). The organization of single-gene nanochromosomes is most curious, as they consist of a 20bp-long telomere sequence on each end, a 5' untranscribed sequence (UTS) that is on average 73bp long, the gene, and a 3'UTS that is on average 25bp long. UTRs are also very short (on average 34bp for the 5'UTR and 25bp for the 3'UTR). Very little research has been done on the mechanisms of gene regulation in these organisms, however such extremely small noncoding regions pose obvious questions regarding the way transcriptional regulation is mediated by transcription factors and histone modification patterns in these systems. The regulatory noncoding regions surrounding transcription start sites in other eukaryotes are usually significantly longer, and transcription initiation stats and regions of transcriptional elongation are marked by nucleosomes (each of which occupies between 150 and 200bp of DNA) containing specific histone marks (these modifications are present in the macronuclei of *Tetrahymena* and *Paramecium* but little work has been done on *Oxytricha* and not much is known about its macronuclear chromatin). One possibility is that some part of gene expression regulation may be accomplished at the level of the control of DNA copy number. Substantial variation in copy number is observed between different nanochromosomes, nanochromosome copy number is somewhat correlated with the expression of their genes (Xu et al. 2012) and mechanisms for the RNA-mediated epigenetic regulation of DNA copy number have been proposed (Nowacki et al. 2010; Heyse et al. 2010). However, regulation of copy number would be expected to be slower than the direct regulation of transcription, and to be somatically heritable, thus it would not be well suited to situations in which fast response to quickly changing environmental conditions is needed, and transcriptional and/or post-transcriptional regulation has to be playing a significant role in the biology of these organisms. In the future, it will of great interest to explore it in more detail as well as the evolution of the nanochromosome format and all the associated changes in nuclear and regulatory biology it necessitates within the ciliate clade. Perhaps even more intriguing are the possibilities such systems offer for understanding gene regulation in general – the holy grail of the field has always been the ability to build detailed computational models of the regulation of gene expression that are fully predictive of its outcome, but this has turned out to be very difficult in practice, one of the main reasons for which has been our absence of good understanding of the roles of transcription factor binding cooperativity, the integration of long-range interactions between distal enhancers and promoters, and the effect of preexisting chromatin states on transcription factor binding. It is certainly possible that all these phenomena are also important in the biology of nanochromosomes but if this is not the case and gene regulation in these organisms is governed by the binding of just a few TFs to few and mostly promoter-proximal sites, as suggested by the limited sequence space around nanochromosome promoters, many of the currently confounding variables would be absent, providing us with a simplified system allowing us to better understand the interplay between the remaining ones (in particular, the interactions between transcription factors and between histone states), knowledge that can later be used in more conventional eukaryotic systems.

While ciliates engage in complex rearrangements of their genomes, on more than one occasion the even more unusual direction of dispensing with most gene regulation at the transcriptional level has been followed. One relatively well-studied lineage, in which this has happened, is the trypanosomatid kinetoplastids. Kinetoplastids as a whole are a group of excavates (Figure 16.1) that contains both free-living and parasitic lineages (with parasitism apparently having evolved multiple times in their evolutionary history; Simpson et al. 2006), but only the trypanosomatids have been extensively studied as they include several major human pathogens (*Trypanosoma* and *Leishmania*). Several *Trypanosoma* and *Leishmania* genomes have been sequenced Ivens et al. 2005; Downing et al. 2011; Berriman et al. 2005; El-Sayed et al. 2005; Peacock et al. 2007), as well as the genome of the *Phytomonas* spp. trypanosomatid, which parasites on plants (Porcel et al. 2013). These genomes display an unusual genome organization – they are compact, containing very few introns, and most strikingly, genes are grouped in long units of several dozens to more than a hundred, which are transcribed as single polycistronic transcripts. These transcripts are then subject to *trans*-splicing through the addition of splicing leader (SL) sequences to the 5' ends of the individual genes. *Trans*-splicing on its own is found in a few other eukaryotic groups (most notably, in nematodes; Krause & Hirsh 1987; Huang & Hirsh 1989), but in trypanosomes the whole genome is transcribed as polycistronic units, the units contain exceptionally large number of genes, the genes within individual units have no discernible functional relationship with each other, and most importantly, there seems to be no regulation of gene expression at the level of transcription (reviewed in Campbell et al. 2003; Martínez-Calvillo et al. 2010; Kramer 2012). This is in marked contrast with the complex life cycles of these organisms, which certainly requires a lot of regulation of gene expression, meaning that it has to happen at the post-transcriptional level. One mechanism might be differential *trans*-splicing (Gupta et al. 2013), in which the addition of SL sequences to different positions leads to functionally distinct proteins, and indeed differential *trans*-splicing seems to be widespread (Siegel et al. 2010; Nilsson et al. 2010; Kolev et al. 2010) though whether it has functional significance in all cases is not known. Another mechanism is the regulation of mRNA stability, evidenced by the fact that changes in mRNA levels are observed for a portion of trypanosomatid genes in different stages of the life cycle (Leifso et al. 2007; Saxena et al. 2007; Minning et al. 2009; Queiroz et al. 2009; Kabani et al. 2009; Jensen et al. 2009; Veitch et al. 2010; Lahav et al. 2011; Depledge et al. 2009; Rochette et al. 2009; Srividya et al. 2007; Alcolea et al. 2010;). Such regulation might be mediated by RNA binding proteins (Estévez 2008; Dallagiovanna et al. 2008), riboswitches, or other mechanisms. Finally, regulation at the translational and possibly the post-translational level seems to be widespread (Bente et al. 2003; Nugent et al. 2004; McNicoll et al. 2006; Leifso et al. 2007; Rosenzweig et al. 2008; Vasquez et al. 2014).

While we know enough to conclude that the regulatory biology in trypanosomatids seems to be happening almost entirely at the RNA level, overall we know very little about the inner workings of these systems, the nature of the *cis*-acting regulatory elements and the logic of regulatory circuits that operate in them, and the evolutionary pressures the drove/allowed their evolution. The independent evolution of *trans*-splicing in nematodes and in other groups is often explained as a consequence of their compact genomes and fast generation times, and the parasitic lifestyle of trypanosomatids might have something to do with their unusual genomic organization. The study of free-living kinetoplastids and the related diplonemid and euglenid lineages should shed light on some of these issues. Limited published genomic data on the free-living kinetoplastid *Bodo saltans* (Santana et al. 2001; Jackson et al. 2008), belonging to the eubodonids, the closest to trypanosomatids lineage (Deschamps et al. 2011), suggests that *trans*-splicing is an ancestral feature of all kinetoplastids, however, whether the lack of transcriptional regulation is also ancestral remains to be seen. The tools for studying the functional genomics of RNA-protein interactions are approaching maturity (Rinn & Ule 2014; Mittal & Zavolan 2014; McHugh et al. 2014), and although they remain more difficult to carry out than ChIP and other chromatin assays, they will be crucial for the untangling of the regulatory networks in these organisms. Of note, the next round of the ENCODE Project features as one of its goals the large-scale identification of the binding sites of a large number of human proteins; the insights from this effort will be informative for the study of kinetoplastid biology and vice versa.

The dinoflagellates are another group that has evolved in a similar direction (in fact, there are many convergent features common to kinetoplastids and dinoflagellates; Lukes et al. 2009); however while they share certain features with other groups, dinoflagellates go beyond anything observed elsewhere and reach wholly new levels of "oddness", exhibiting the most radical known departures from our conventional view of the way an eukaryotic cells operates in numerous aspects of their biology (Hackett et al. 2004; see discussion of organellar genome biology below for more examples), to an extent that they used to be thought as intermediates between prokaryotes and eukaryotes (Dodge 1965). The dinoflagellates are a highly successful and diverse lineage of alveolates, containing both heterotrophic and autotrophic groups, with photosynthetic capacity being the result of secondary and even serial secondary endosymbiosis (Keeling 2009; Keeling 2010). They are unique among all eukaryotes in that their nuclei seem to contain little or no histones (Rizzo & Nooden 1972; Rizzo 2003), chromatin is permanently condensed (Dodge & Greuet 1987), chromosomes exist in a liquid crystalline state (Rill et al. 1989), and up to 70% of thymine bases in DNA are replaced by 5-hydroxymethylcytosine (Rizzo et al. 1987). The negative charge of DNA has been suggested to be neutralized by divalent cations instead of histone proteins (Levi-Setti et al. 2008). For a long time it was thought the histones are completely absent from their genomes, but EST and transcriptome sequencing efforts have conclusively shown that dinoflagellates in fact do possess histone genes (Hackett 2005; Jaeckisch et al. 2012; Roy & Morse 2012; Bayer et al. 2012). These results, however, by no means resolve the mystery of dinoflagellates genome biology as the reason histones were believed to be absent is that they were not detectable biochemically and that the protein-to-DNA ration in dinoflagellate chromatin is about 1:10, compared to the typical 1:1 ratio in all other eukaryotes, i.e. even though histones are present, they are either expressed only at certain stages of the life cycle or they are only bound to a tiny fraction of the genome. More recently, an abundant nuclear protein that might be playing a histone-like role of apparent Phycodnaviridae viral origin was found (Gornik et al. 2012).

Unfortunately, knowledge of dinoflagellate genome organization and gene regulation is very limited owing to their extremely large genomes, which have so far precluded whole genome sequencing. The smallest genomes in the group are ∼1.5Gb (for example, *Symbiodinium*; LaJeunesse et al. 2005), with most other species possessing larger genomes, up to more than 100Gb (for example, *Prorocentrum micans*; Veldhuis et al. 1997). What little is known is derived from transcriptome sequencing (Hackett 2005; Jaeckisch et al. 2012; Roy & Morse 2012; Bayer et al. 2012) and the sequencing of small portions of the genome (McEwan et al. 2008). The available information suggests that these genomes contain large numbers of genes (larger than the 20,000 protein coding genes in the human genomes, potentially up to 40,000 or more); however, due to its very large total size, the genome has low gene density. Genes are often organized in tandem arrays (Bachvaroff & Place 2008); however, unlike those found in kinetoplastids, dinoflagellate genes arrays usually consist of the same gene repeated many times. *Trans*-splicing of SL sequences is widespread (Lidie & van Dolah 2007; Zhang et al. 2007; McEwan et al. 2008; Lin et al. 2010) and it seems that transcriptional regulation is limited, similarly to kinetoplastids, though much further work will be needed to understand to what extent.

These peculiarities pose numerous questions regarding the nature of gene regulation, whether and what role histones, other chromatin-associated proteins, and transcription factors (which seem to be limited in number and diversity in dinoflagellates) play in it, the three-dimensional organization of dinoflagellate genomes, how it compares to that of other eukaryotes and what influence it has on gene expression, and most importantly, what evolutionary forces shaped these genomes in such a strange from our perspective way. The transcriptome of *Perkinsus marinus*, the representative species of the closest to the dinoflagellates lineage, the perkinsids, has been sequenced and it too uses splice leader *trans*-spicing. Perkinsids, however, have a full set of histones and use them as all other eukaryotes do (Gornik et al. 2012). Fortunately, the whole-genome sequencing of dinoflagellates genomes is expected to become feasible in the near future thanks to the advent of long-read sequencing technologies. This will in turn enable the application of functional genomics tools to the study of these fascinating organisms and their closest relatives, which should shed light on the evolution of this outstanding section of the eukaryote tree.

16.6.5 The evolution of the histone code

In the last nearly two decades, much progress has been made in deciphering the histone code (Kouzarides 2007). The association of a number of histone modifications with certain transcriptional states and chromatin processes is now well known. For examples, methylation of lysine 4 on histone 3 (H3K4me3) is a signature mark of active promoters (Bernstein et al. 2002; Santos-Rosa et al. 2002; Guenther et al. 2007), enhancers are marked by H3K4me1 and H3K27ac (Heintzman et al. 2007; Heintzman et la. 2009; Creyghton et al. 2010; Rada-Iglesias et al. 2009), H3K27me3 is associated with repression mediated by Polycomb proteins (Simon & Kingston 2013; Zheng & Chen 2013), H3K9me3 is found in repressed heterochromatin and has a positive feedback loop relationship with DNA methylation (Hashimoto et al. 2010), etc. In some cases, we have a quite detailed mechanistic understanding of the role histone modifications play in these processes; a classic example is H3K36me3, which is associated with transcribed genes, where it is deposited in the process of transcriptional elongation and functions to recruit histone deacetylases. The deacetylases in turn remove the acetylation marks also deposited during elongation in order to prevent intragenic transcription from cryptic promoters, as acetylated histones exist in a more open and conductive to transcription conformation (Lee & Shilatifard 2007). However, in addition to the few well studied examples, a large number of poorly understood histone modifications have been detected through mass spectrometry (Freitas et al. 2004), thus the deciphering of the code is very far from complete, especially at the level of understanding the mechanistic biochemical roles individual modifications play. Notably, histone marks are consistently found in particular combinations (though not necessarily physically on the same histone tails and at the same time) constituting specific chromatin "states" associated with certain parts of genes and with intergenic features such as regulatory elements (Ernst & Kellis 2010; Ernst et al. 2011; Ernst & Kellis 2012; Mortazavi et al. 2013).

The sequence of histone proteins is very deeply conserved and most of the well-known histone modifications are accordingly shared by deeply diverging lineages, suggesting they were ancestral to all eukaryotes. However, it is far less clear whether the less known modifications are also similarly conserved, and more importantly, whether the chromatin states observed in metazoans are also ancestral to all eukaryotes. Based on what we know from the available data, the answer seems to be that they can be evolutionarily malleable, even within animals. For example, comparison of modENCODE ChIP data on a number of histone modifications revealed that while H3K27me3 and H3K9me3 colocalize in the fly genome, they are found in distinct domains in *C. elegans*. Studies in other organisms have discovered a number of other deviations from the well-known patterns of mammalian chromatin structure.

H2A.Z is a variant of H2A well-known for its association with promoter regions in animals, yeast and plants (Henikoff 2008; Jin et al. 2009). But in *Plasmodium falciparum* H2A.Z instead demarcates all intergenic regions (Bártfai et al. 2010; Hoeijmakers et al. 2012), and in *Trypanosoma brucei* H2A.Z together with the unique to kinetoplastids variants H2BV, H3V, and H4V marks the boundaries of the polycistronic units (Janzen et al. 2006; Mandava et al. 2008). As is the case with H3K9me3, H3K9me2 is generally a repressive mark, but in diatom genomes, it has been found to be associated with actively transcribed genes (Lin et al. 2012). A region either free of nucleosomes or containing labile nucleosomes (Jin et al. 2009) is found around promoters in most model systems, but in *Dictyostelium discoideum* it has been reported that extended such regions are found both around the 5' and the 3' end of genes, where they are precisely demarcated by Poly-A tracts and Poly-T tracts, respectively (Chang et al. 2012).

From the perspective of understanding the origins of the histone code and the reasons for its current form in mammalian genomes, it will be of great interest to carry out a systematic epigenomic analysis of chromatin modifications and chromatin states across the tree of life, identify the major deviations from the familiar patterns, and, if possible, the evolutionary forces behind their appearance. Integrative methods for analyzing histone mark ChIP-seq data will be of invaluable help in this endeavor (Ernst & Kellis 2010; Ernst et al. 2011; Ernst & Kellis 2012; Mortazavi et al. 2013).

16.6.6 Testing the competing theories for the origin of genomic complexity

At this point in time we have a very general, well supported by multiple lines of evidence theory of the evolution of genome complexity, in which the concept of "junk DNA" and the role of non-adaptive processes in shaping genome architecture have a prominent place (Lynch M. 2007c; Koonin 2011), organismal complexity is not understood to be the direct result of adaptive increases in genomic complexity, and humans are not perceived to be the pinnacle of evolutionary progress (Koonin 2004). These lines of evidence include:

1. Population genetics theory and what is known about the population genetic environment of different lineages. Population genetics predicts that lineages with lower long-term effective population size N_e will accumulate larger numbers of neutral and slightly disadvantageous genomic changes as the efficiency of natural selection is reduced when N_e is low.

2. The C-value paradox, the observation that there is no relationship between organismal complexity and genome size (Thomas 1971), that orders of magnitude of differences in genome size between species of comparable morphological complexity are observed, and that many species with genomes vastly larger than the human genome, including unicellular ones, exist.

3. The g-value paradox, the observation that the number of protein coding genes that organisms have does not correlate with organismal complexity (Hahn & Wray 2002). Many plants and even unicellular eukaryotes have more genes that mammals do.

4. The general inverse correlation between genome size and N_e in different lineages across the tree of life.

5. The closely related general inverse correlation between N_e and genomic features such as transposable element content, intron length and size.

This view is not shared by all researchers, with vocal opinions in support of the position that “junk DNA” does not exist having been repeatedly raised (for example, Mattick & Dinger 2013). However, there are numerous reasons why such a position is untenable. First, the null hypothesis should always be neutral evolution and neutral adaptive significance for any trait examined, as is the standard practice in molecular evolution research. Functionality has to be demonstrated in a positive way by rejecting this null hypothesis. This is despite suggestions that the null hypothesis should be reversed and lack of function is what should be demonstrated explicitly (see discussion in Bhattacharjee 2014). Second, all adaptive explanations for the evolution of certain genomic features will have to be supported by population genetics arguments (because “nothing in evolution makes sense except in the light of population genetics”; Lynch M. 2007b). This has so far typically not been done – all such arguments have been verbal rather than quantitative and that is when evolution was even considered, which has not always been the case. In contrast, population genetics-oriented analysis has mostly returned results pointing in the completely opposite direction. Third, to argue that there is no junk DNA in the human genome is equivalent to arguing that there is no junk DNA anywhere in the tree of life, otherwise one would have to elevate the human genome to a very special position compared to other organisms, directly contradicting one of the most fundamental insights of evolutionary biology, that humans (or, in a more relaxed version of the same statement, vertebrates in general) are part of a continuum with all other life forms (as a curious side note, it should be noted that the “junk” DNA debate pops up primarily when the human genome is discussed but significantly less often when the genomes of other organisms are concerned – for example, the publication of the modENCODE papers generated no such controversy – suggesting that our view of ourselves as a species has a lot to do with the persistent resurfacing of this discussion every time we probe deeper into our genome). Thus in order to reject the existence of junk DNA, it will have to be shown to not exist not only in humans but in all other lineages, and there are numerous cases in which it is much more difficult to even suggest possible functions for certain DNA sequences than it is for the vast noncoding portions of mammalian genomes. A good example of such an objection is the “onion test” (Gregory 2007) requiring that proposals that all eukaryotic DNA is the result of adaptive evolution should be able to explain why onion species in the *Allium* genus need more DNA than humans, and why the DNA content of different *Allium* species varies more than five-fold; many other such examples are also known, from the giant genomes of lingfish and salamanders (40 times bigger than the human genome) to the equally gigantic genomes of the unicellular dinoflagellates mentioned previously. It is not clear why these species would need orders of magnitude more noncoding DNA, with all the proposed adaptive roles it might be playing than mammals. And then there is what can be considered the ultimate example of junk DNA – DNA that is deleted in somatic genomes, an example of which are the transposon-rich IESs of ciliates (as well as what happens in some other species, reviewed in Kloc & Zagrodzinska 2001), which are present in the inert and transcriptionally silent "germline" micronuclear genome of these organisms and are excised and absent from the "somatic" macronuclear genome. While genes are scrambled and alternative DNA splicing might be happening in some ciliates, in others the excision of IESs results in a reliably colinear splicing of the functional DNA segments, thus arguments that larger protein diversity is produced in this way (which have also been used to support the adaptive significance of RNA splicing) do not apply. Even if we adopt the biochemical criterion for assessing the functionality of DNA sequences (i.e. that if a DNA sequenced is transcribed or bound by transcription factors) as our sole guide, IESs largely fail to qualify as being functional as the micronucleus is generally transcriptionally silent (though it should be noted that high-resolution functional genomic analysis of micronuclear gene expression in species like *Oxytricha* has not been carried out). Numerous other examples can be presented (see discussion below on organellar genomes).

Yet while the general nonadaptive theory is sound, there are still numerous incompletely resolved issues both regarding the details of genome evolution in particular groups and the general forces behind the evolution of genomic and organismal complexity across the tree of life. Also, it has not been comprehensively tested with respect to many aspects of functional genomic complexity and other features of genome biology. Despite the controversy surrounding the project, thanks in no small part to the efforts of the ENCODE Project Consortium, we are now in a position, in which the experimental and analytical tools to obtain conclusive answers to many of these questions are in existence, by unleashing this functional genomics machinery on the large known eukaryotic diversity. Below I list and discuss some of them.

16.6.6.1 The relationship between genome size, genome complexity and organismal complexity

From the limited genomic sampling of organisms we have at our disposal, we know that there is a general pattern of correlation between organismal size and organismal complexity on one side, and genomic size and functional genomic complexity, on the other. This is understood to be the result of the lowered N_e associated with larger physical size, as already outlined. However, setting aside all other considerations, the adaptive view of genome complexity is also consistent with this general pattern, and indeed the vast amount of non-coding RNA and the generally increased complexity of genomic architecture in large multicellular organisms has often been interpreted as evidence for the existence of a relationship between the two that is both causal and adaptive (Liu et al. 2013).

It is possible to conclusively distinguish between these two competing explanations thanks to the existence of natural control groups. One of them are the metazoans with extremely large (much larger than ours) genomes. Few would consider these species to be more "complex" than humans, and so if all the hallmarks of genomic complexity listed below scale up in correlation with genome size, then this would be solid evidence for the absence of absolute functional and causal relationship between the complexity of organismal organization and that of gene regulation and gene expression (Doolittle 2013). So far it has been technically and economically infeasible to generate high-quality assemblies for any genome much larger than the mammalian genome size, but steps forwards toward making this a real possibility have already been made (Nystedt et al. 2013; Neale et al. 2014; Zimin et al. 2014; Wegrzyn et al. 2014), and advances in sequencing technology promise to eventually solve the problem. Of note, for such a comparison to be valid, the large genome size should not be the result of extensive polyploidy, which might be the case with some of the known species with huge genomes, but this is unlikely to be the case for all of them.

Even better, we have multiple natural controls for testing our theories explaining the association of genomic complexity with the complexity of multicellular organisms, as multicellularity arose in more than one lineage, and the extent of variation of parameters of the population genetic environment such as N_e is comparable between some of them. Within these large clades, we observe large variations in genome size between different species of lineages at comparable levels of organismal complexity (Gregory et al. 2007).

Multicellularity evolved on at least six occasions, in metazoans, the land plants lineage, in red algae, in brown algae and on multiple separate occasions in fungi (Knoll 2011). In two of these groups, the land plants and the animals, a remarkable convergence of the genome characteristics correlated in very similar ways with values of N_e is seen. The N_e of invertebrates is on the order of 10^6 while it goes down to 10^4 in large-bodied vertebrates (Lynch 2007c), and the genomes of the former are on average an order of magnitude smaller (few hundred megabases) and contain much less noncoding DNA that those of the latter (which are gigabase-sized). Similar differences in the average N_e are observed between trees and annual plants. Groups like the gymnosperms have some of the biggest on average genomes of all eukaryotes (between 2 and 36Gb; Murray et al. 2012), while many annual plants have small genomes. However, a number of small annual flowering plants also have giant genomes, complicating the picture (even after accounting for the rampant polyploidy in plants) suggesting much remains to be learned about the balance between their effective populations size, the direction of mutational biases, and the tolerance of the biology of these organisms towards genomic expansion. Still, it remains true that many similarities between the extremes of genomic size in plants and animals exist and it will be very informative to study and compare the functional complexity of these lineages. The smallest known land plant genomes are just 63Mb in size (*Genlisea margaretae*; Greilhuber et al. 2006) and the size of the model organism *Arabidopsis thaliana* genome is also quite small, at 120−150Mb (Arabidopsis Genome Initiative 2000). More recently, an even more highly compressed plant genome was sequenced, that of *Utricularia gibba* at 82Mb (Ibarra-Laclette et al. 2013) and it revealed marked reduction of noncoding DNA and little transposable element activity. At the other extreme, plant genomes as big as 152Gb are known (*Paris japonica*; Pellicer et al. 2010) and many species with genomes in the tens of Gb range are known (Zonneveld 2009; Zonneveld 2010). In animals, genomes as small as 20Mb are known (in some parasitic nematodes; Leroy et al. 2007) and as big as 130Gb (the lungfish *Protopterus aethiopicus*; Pedersen 1971), and large variation is seen even within the major animal phyla. For example, the smallest arthropod genomes are on the order of 100Mb (the *Strepsiptera* insects, Johnston et al. 2007; the mite *Tetranychus urticae* at 90Mb; Grbic et al. 2011), but species with truly massive genomes are also known (the amphipod *Ampelisca macrocephala* at 63.2Gb; Rees et al. 2007). Within vertebrates, amphibians range from 0.95Gb for some frogs (*Limnodynastes ornatus*; Olmo & Morescalchi 1978) to more than 120Gb in some salamanders (*Necturus lewisi*; Olmo 1973). The smallest fish genomes are on the order of 360Mb, such as the sequenced genomes of *Tetraodon nigroviridis* (Jaillon et al. 2004) and *Fugu rubripes* (Aparicio et al. 2002) and the even smaller genome of *Tetraodon fluviatilis* (Brainerd et al. 2001), while the >130Gb genomes of lungfish were already mentioned. Significant variation is observed even within mammals: bat genomes tend to be smaller that the typical for mammal $\sim$3Gb (the smallest known bat genome is that of *Miniopterus schreibersi* at 1.7Gb; Capanna & Manfredi Romanini 1971), while the genome of vizcacha rat *Tympanoctomys barrerae* is 8.4Gb, although it, unusually for a mammal, seems to be tetraploid (Gallardo et al. 1999), and thus the title of largest mammalian genome belongs to the 6.3Gb genome of the cape golden mole *Chrysochloris asiatica* (Redi et al. 2007). Remarkably, the genome size of birds is typically significantly smaller than that of mammals, in convergence with what is seen in bats, with the largest bird genome being that of the flightless and large-sized ostrich *Struthio camelus* at 2.16Gb (Eden et al. 1978) suggesting that flying may be a common reason for this reduction (Tiersch & Wachtel 1991; Hughes & Hughes 1995; Smith & Gregory 2009). It will be highly intriguing to know what the functional complexity of genomes across these extremes of genome size occurring at otherwise similar levels of organismal organization is. Of note, the other lineages that have evolved multicellularity also display wide variation in genome size but not to the same extent: the known genome sizes of red algae are between 100Mb and 2.8Gb (Kapraun & Freshwater 2012), those of brown algae are between 200Mb and 3.6Gb (though at the high end, polyploidy may play a role; Phillips et al. 2011), and the largest fungal genomes reach 800Mb (Kullman et al. 2005).

Perhaps even more intriguing though are the giant genomes of unicellular organisms. The N_e of unicellular eukaryotes is larger than that of invertebrates and correspondingly on average they have smaller genomes (usually <100Mb), fewer and shorter introns and fewer transposons

(Lynch 2007c). However, examples of very big genomes are known among them. An estimate of more than 700Gb is often quoted for some free-living amoebas (Friz 1968) but it is quite likely that this is due to high levels of polyploidy. Yet plenty of other, more reliable examples of giant genomes exist. The enormous dinoflagellates genomes were already discussed, but large genomes have been reported for other algae too: for example, while the two sequenced diatom genomes, those of *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* stand at 27Mb and 34Mb respectively, ((Bowler et al. 2008; Armbrust et al. 2004), a value of >25Gb has been reported for *Coscinodiscus asteromphalus* (Shuter et al. 1983). The existence of such large genomes in organisms with algal lifestyle is a bit puzzling; dinoflagellates are a highly successful and diverse group presumably subjected to some of the same evolutionary pressures as other algal lineages, which have generally lead to genome streamlining and reduction (for example, the smallest genome of a free-living eukaryotes is the 12.6Mb genome of *Ostreococcus tauri*, Derelle et al. 2006; the smallest known eukaryotic genome in general belongs to the parasite *Encephalitozoon intestinalis* at 2.3Mb (Vivarès & Méténier 2000). Thus it is not far from clear what drove the evolution of such large genomes, although it has been suggested that dinoflagellate effective population size may in fact be very low (Watts et al. 2013). Even more intriguing is the question of what the content of these genomes is. As mentioned above, because organismal complexity typically correlates both with low N_e and with increased genome size, it is still possible to claim adaptive significance of all the extra noncoding DNA. However, this argument is much more difficult to make in the case of a unicellular species with a very large genome, as while such organisms may have far from simple lifestyles, they do not have to execute an incredibly complicated developmental program and specify hundreds of different cell types, each with its own gene expression program, the reasons usually cited as a reason for the functionality of all noncoding DNA and RNAs in humans. The functional genomic study of dinoflagellates and especially of other more "normal" in their biology protists, should be highly illuminative with respect to these questions.

16.6.6.2 The number and complexity of regulatory elements

The large number of putative regulatory elements has been a major result of the ENCODE Project (Neph et al. 2012; Thurman et al. 2012). As discussed above, whether all of them are in fact functional is an open question, but their existence can be interpreted both as the result of the need for them for the specification of the complex vertebrate body plan with all its numerous cell types (Levine & Tjian 2003), and as a nonadaptive consequence of the greater tolerance towards genomic changes conferred by the low N_e of vertebrates. The two explanations are far from mutually exclusive, and in fact both are likely to be correct to an extent, but in order to parse their relative contribution, natural control groups within vertebrates with larger and smaller genomes will have to be studied using ENCODE-like approaches for mapping putative regulatory regions. Examples of such groups include the salamander and lungfish species with huge genomes, the larger mammalian genomes, the fish species with small genomes such as fugu, and birds and bats with their on average two-fold reduced compared to the mammalian mean genomes, as well as the giant and compressed genomes within the different invertebrate groups. If the number of putative regulatory elements scales up with genome size then this would be convincing evidence in support of the nonadaptive origin of regulatory complexity in large multicellular species.

The other key control group would be plants. Historically, the study of enhancers has been a metazoan-centric enterprise, with little work having been done in plants, and it would be of great biological interest and practical relevance to learn more about them, especially if any major differences in the general architecture of long-range gene regulation exist in the larger plant genomes compared to what is seen in mammals. With respect to functional complexity, as plants present a similarly wide range of genome sizes to animals, if a similar scaling of the number of regulatory elements with genome size and N_e is observed in them (as is seen for the number of introns and transposons), this would further corroborate their nonadaptive origin. The nonadaptive hypothesis would be strengthened even further if the same is seen in unicellular species.

16.6.6.3 The genomic changes associated with the evolution of multicellularity

A defining feature of multicellularity is the differentiation of cells into distinct cell types, which is traditionally understood to require increases in regulatory complexity. At present we do not have a detailed genomic understanding of how multicellularity evolved but from the genomes of representatives of the closest to metazoans lineages, the choanoflagellates *Monosiga brevicollis* (King et al. 2008) and *Salpingoeca rosetta* (Fairclough et al. 2013), and the filasterean *Capsaspora owczarzaki*, we do know that these lineages already possessed members of a number of common metazoan transcription factor families. Understanding what regulatory networks look like in these organisms and how they were rewired later in metazoan evolution, as well as similar studies in the other lineages where multicellularity arose should be highly informative regarding its origins.

16.6.6.4 The prevalence of functional regulated alternative splicing and the general level of splicing complexity

The inverse-correlation relationship between N_e and intron number and size is well understood (Lynch 2002; Lynch 2006b; Lynch 2007c). However, how many introns there are in a gene is a different question from how many alternative splice products are generated and what their functional significance is. Whether a similar inverse-correlation relationship between the complexity of splicing and population genetic parameters exists is highly relevant to the debate about the functional significance of most of the observed alternative splicing events in mammalian genomes. Under the nonadaptive model of the evolution of splicing complexity, improvements in the fidelity of the splicing machinery would be limited by the minimum value of the level of negative effect on fitness such errors have that is visible to selection (which in turn is constrained by N_e). Comparison of RNA-seq datasets across the eukaryotic tree of life should shed light on these questions. It should be noted that because of issues with the variable complexity of cell type composition in the samples that can be practically obtained from different species, such questions are best answered by single-cell sequencing. For example, a population of unicellular organisms can exist in a reasonably uniform cell state but for a multicellular organism, only samples from whole organs containing many cell types with presumably different splicing patterns could be available, making the direct comparison of splicing complexity problematic as it would be artificially elevated in the latter case. At present, single-cell RNA-seq is not yet up to this task due to the large stochastic noise levels of current protocols (Marinov et al. 2014) but future improvements in experimental techniques should resolve these issues and enable such studies.

Notable curious cases are already known. The genome of the chlorarachniophyte *Bigelowiella natans* was recently sequenced and its transcriptome analyzed using RNA-seq (Curtis et al. 2012). Remarkably, while the genome as a whole is highly streamlined, it contains numerous introns and exhibits very high levels of alternative splicing, similar to what is seen in human brains (it also contains more protein-coding genes than the human genome, more than $21,000$). Why a single-celled organism would need so much alternative splicing is not clear, and indeed most of it has been interpreted as the result of errors in the splicing machinery (Curtis et al. 2012), but why this is tolerated in *Bigelowiella natans* and not in other algae is not clear. It is also curious that as a chlorarachniophyte, *Bigelowiella natans* possesses another eukaryotic genome, that of the nucleomorph remnant of the nucleus of its photosynthetic secondary endosymbiont (see discussion below). While nucleomorph genomes provide the most extreme example of the reduction of an eukaryotic genome (Archibald & Lane 2009), the *Bigelowiella natans* nucleomorph genome is intron-rich even though the introns are extremely short, 18-21bp on average (Gilson et al. 2006).

16.6.6.5 The number of lncRNAs

As discussed above, one explanation for the number and fast evolution of lncRNAs in vertebrate genomes is that many of them are the product of an evolutionary treadmill of gene birth and death, which is allowed to generate much larger numbers of lncRNAs in big genomes due to looser constraints on this process. Therefore whether the number of lncRNAs scales up with genome size and whether this is true across all eukaryotic groups is of major importance for the way we think of these RNAs. The work on lncRNAs outside of vertebrates has been limited so far. An RNA-seq study revealed 164 novel lncRNAs in the streamlined genome of *Plasmodium falciparum* (Liao et al. 2014), and 60 lncRNA candidates were identified in a prior tiling array-based work (Broadbent et al. 2011), numbers that vastly smaller than the up to 10,000 lincRNAs found in humans. Another recent study (Li et al. 2014) found 1,704 high-confidence lncRNAs in the 2.3Gb maize genome (*Zea mays*; Schnable et al. 2009) but much more data points are needed to draw general conclusions.

16.6.6.6 The prevalence, nature and conservation of "exotic" RNA species

Work on "exotic" RNAs such as the various promoter and transcription termination-associated RNAs and on eRNAs has been limited to mammals and other traditional model systems. and even in these organisms, the functional significance of these molecules has not been investigated, with some exceptions (Melo et al. 2013; Li et al. 2013; Lam et al. 2013; Mousavi et al. 2013; Memczak et al. 2013; Hansen et al. 2013). Major unresolved questions remain regarding whether these RNAs are universal features of eukaryotic gene expression and if not, what the patterns of their evolution across different lineages are. For example, large genomes with low gene density, in which genes are regulated by distantly located enhancers, seem to have evolved multiple times from an ancestral state characterized by a much more compact genomic architecture. Whether features such as eRNAs are present in all such lineages, and if yes, whether they were ancestral or evolved convergently and why, is of significant interest. As with other aspects of eukaryote biology, some notable deviations from what is observed in traditional model systems are already known, one of them being promoter-associated bidirectional transcription. It is a common features of mammalian promoters (Core et al. 2008; Seila et al. 2008; Xu et al. 2009), however, antisense transcripts are unstable and preferentially degraded, perhaps due to the differences in the frequency of polyadenylation and splice sites in the sense and antisense direction, which help the transcriptional machinery determine the proper orientation (Almada et al. 2013). But in the diplomonad *Giardia lamblia* sterile, noncoding bidirectional antisense transcripts have been reported to be abundant and polyadenylated, representing up to 20% of total cDNA (Elmendorf et al. 2001; Teodorovic et al. 2007). Diplomonads are a very deeply diverging lineage (Figure 16.1) and the *Giardia* genome is highly compressed (Morrison et al. 2007), with its transcriptional apparatus being simplified compared to other eukaryotic and missing a number of components. The details of this unusual transcriptional biology and the reasons for its evolution are at present unknown.

16.7 Organellar genomes and the principles of genome evolution

Finally, a few words need to be said about organellar genomes and what they tell us about the relative contribution of the different evolutionary forces shaping genomes. While organelles were not a focus of the ENCODE and modENCODE projects, there is little reason to think the answers to the questions whether the existence of "junk DNA" in large amounts is permissible by evolution, and whether the picture of genome organization derived from metazoans and other opisthokonts is representative of all eukaryotes, would be different with respect to their genomes. Just as it is true that if there is no junk DNA is the human genome then no junk DNA should be expected in the genomes of all other organisms, it is also true that if there is no junk DNA in the human genome then there should be no junk DNA in organellar genomes. It so happens that both very difficult to refute examples of "junk DNA" and a large diversity of genome organizations and complicated embellishments in gene expression and RNA processing are found in the organellar genomes of various organisms (Barbrook et al. 2010). Organelles are also one of the systems the evolution of which provides a textbook example supporting the mutation burden hypothesis for the evolution of genomic organization and complexity (Lynch 2006). The study of the diversity of organellar genomes should therefore provide some very helpful insights into these issues.

We should first briefly review our knowledge of organellar genomes and their structure across the tree of life. Organelles evolved as a result of endosymbiosis (Altmann 1890; Mereschkowsky 1905; Sagan 1967), and their genomes are a remnant of their free-living past. There were two primary endosymbiotic events in the history of eukaryotes. First, mitochondria evolved as a result of the endosymbiosis of a α-proteobacterial prokaryote (John & Whatley 1975; Yang et al. 1985) and the ancestor of modern eukaryotes, which was most likely either an archaeon or an archaeon-like lineage (Lake et al. 1984; Ribeiro & Golding 1998; Cox et el. 2008; Williams et al. 2013; Embley & Martin 2006; Gribaldo et al. 2010). This event may have in fact even provided the primary evolutionary driving force behind the origin of modern eukaryotes and their features (Martin & Koonin 2006; Koonin 2006). All modern eukaryotes possess either mitochondria, which with very few exceptions (Abrahamsen et al. 2004; Henriquez et al. 2005) contain their own genome (Nass et al. 1965), or the remnants of mitochondria in the form of hydrogenosomes (Lindmark & Müller 1973) and mitosomes (Tovar et al. 1999; Tovar et al. 2003; Williams et al. 2002) that, with few exceptions (Boxma et al. 2005), have lost it (Embley & Martin 2006; Embley et al. 2003; van der Giezen 2009). Later in eukaryote evolution, one lineage acquired a second endosymbiont of cyanobacterial origin (Keeling 2004), which evolved into the modern chloroplast. Chloroplasts were later lost in multiple lineages and acquired again several times through secondary endosymbiosis (the engulfment of a plastid-containing eukaryotes by another eukaryote) and even tertiary endosymbiosis (Cavalier-Smith 2002; Stoebe & Maier 2002; Archibald & Keeling 2002; Keeling 2009; Keeling 2010).

The main theme in the evolution of organellar genomes has been the transfer of genes from them to the nucleus. There seems to be a constant, ongoing process of integration of parts of the mitochondrial genome into the nuclear genome, evidenced by the presence of multiple partial copies of the mitochondrial DNA in mammalian genomes (NUMTs, Hazkani-Covo et al. 2010). A similar process acts on the plastid genome (Ayliffe et al. 1998; Huang et al. 2003). Some of these nuclear insertions of organellar genes subsequently evolve regulatory elements enabling their expression in the nucleus and sequences ensuring their targeting and importing into organelles. The organellar copy of the gene can then be lost without fitness consequences. It has to be noted that this process is not unidirectional - organellar proteomes also contain many genes that did not originate from the genome of the original endosymbiont but are instead either of nuclear origin or come from other organelles (for example, a significant fraction of the plastid proteome is of non-cyanobacterial ancestry; Suzuki & Miyagishima 2010). The end result has been the great reduction of gene content in organellar genomes in all lineages studied. The most gene-rich mitochondrial genomes are those of *Reclinomonas americana*, which contains 62 protein coding genes (including its own apparently ancestral bacterial-type RNA polymerase), and of other jacobid species(Lang et al. 1997; Burger et al. 2103). At the other extreme, the mitochondrial genomes of *Plasmodium* and other apicomplexans, and those of dinoflagellates are extremely reduced in terms of gene content, containing as little as 3 genes (Vaidya & Mather 2009; Nash et al. 2007). Plastid genomes have generally retained a larger number of genes, between 90 and 250 genes (Green 2011), with a few exception associated with the loss of photosynthetic capacity (for example, the apicomplexans; Sato 2011).

These differences in gene content are confined within a relatively narrow band of variability compared to the extreme differences in organellar genome size and organization observed within the known eukaryotic diversity. The best studied and most familiar organellar genomes are those of mammalian mitochondria. The human mitochondrial genome is 16,571bp long and contains 13 protein coding genes, 22 tRNAs and 2 rRNAs (Anderson et al. 1981; Bibb et al. 1981). It is circular mapping (meaning that it can be represented as a circle but it does not necessarily adopt a single-circle conformation in vivo; Bendich 1993) and extremely densely packed with genes, with no introns and only one non-coding region, referred to as the D-loop, which plays a central role in the initiation of transcription and replication. Transcription is carried out by a dedicated polymerase (POLRMT), which is encoded in the nucleus and is of apparent phage origin (Masters et al. 1987), and three polycistronic transcripts are produced from both strands, which are subsequently posttranscriptionally processed to produce the mature message molecules. This economy of DNA content is a common feature of all animal mitochondrial genomes: the size of the smallest ones is ~11kbp (for example, the chaetognath *Sagitta decipens*; Miyamoto et al. 2010), while the largest ones are ∼43kbp long (in *Trichoplax*; Dellaporta et al. 2006). More significant variations in structure exist within animals (see discussion below) but they are dwarfed by the extraordinary diversity in the size and organization of mitochondrial genomes within other eukaryotes. The best known contrast is that between the mitochondrial genomes of animals and green plants. The latter have large genome (typically several hundred kb) but comparable gene content, with the difference being largely due to the presence of large amounts of repeats and intronic sequences. Plant mitochondrial genomes can in some cases reach truly extreme sizes, sometimes considerably larger than the genomes of free-living bacteria. For example, *Cucurbitaceae* mitochondrial genomes, of which that of *Cucumis sativus* was recently fully sequenced (Alverson et al. 2011), can reach up to 3Mbp (Ward et al. 1981). The *Cucumis sativus* mitochondrial genome was found to be 1685kb long, yet it still has only 37 genes, with the rest of the genome consisting of repeats, expanded introns and apparently inactive sequences of nuclear, plastid and viral origin. An even more extreme example is provided by the mitochondrial genomes of angiosperms in the *Silene* genus, two of which, *Silene conica* and *Silene noctiflora* were recently sequenced and found to be 6.7 and 11.3Mb long, respectively, which is again due primarily to extreme expansion of repetitive sequences (Sloan et al. 2012).

The differences in mitochondrial genome size between plants and animals are theoretically explained as a non-adaptive consequence of the differences in the mutation rate in mitochondria between the two lineages (Lynch et al. 2006). The mitochondrial mutation rate μ in plants and mammals have evolved in opposite directions, and is orders of magnitude lower in the former than it is in the latter. Recall from the discussion of introns above that the removal of noncoding DNA such as introns by selection is facilitated by large values of $N_e\mu$; the values of N_e are similar between multicellular animals and multicellular plants but the large differences in μ explain well the observed disparities in noncoding DNA content. Of note, there are animal lineages in which μ is low compared to other metazoans (for example, cnidarians) and they happen to also be an exception of the general rule that animal mitochondria do not contain introns (Shearer et al. 2002; van Oppen et al. 2002). The true picture is likely to be more complex: for example, the aforementioned giant *Silene* mitochondrial genomes seem to have very high mutation rates combined with extreme bloating with noncoding DNA but it is not clear whether these measurements represent the long-term population-genetic environment of the lineage, and whether the repeat expansion in them is not driven by other factors.

Nevertheless, organellar genomes provide some of the at present most difficult to explain from an adaptive perspective cases of “junk” DNA. The smallest mitochondrial genomes are only 6-7kb long (*Plasmodium yoelii*, Vaidya et al. 1989; *Theileria parva*, Kairo et al. 1994), the largest are as big as the largest known genomes of free-living prokaryote, yet they never possess more than 60 protein coding genes and around 100 genes in total after taking tRNAs and rRNAs into account. Very large differences in non-coding DNA content are observed between closely related species. For example, the sizes of the mitochondrial genomes of different *Schizosaccharomyces* yeast species differ by as much as 4-fold, with little difference in gene content (Bullerwell et al. 2003). Even more strikingly, the two *Silene* species mentioned above differ by 4.5Mb in terms of mitochondrial genome size, while other species in the same genus (*Silene vulgaris* and *Silene latifolia*) have an order of magnitude smaller mitochondrial genomes than either of them (427kb and 253kb respectively). It is hard to imagine what functional role all this additional, mostly repetitive-element sequence, might be playing. Most adaptive hypotheses for the role of transposable repetitive elements, introns, and pseudogenes have been specifically tailored to large nuclear genomes packaged by histones. But organellar genomes evolved from prokaryote genomes, in which long-range regulatory interactions are not the norm, and subsequently underwent drastic reduction, as a result of which today they only contain a few dozen genes. It is far from clear how hundreds of kilobases and even megabases of non-coding DNA could all conceivably function, let alone be necessary, for their regulation, expression and processing. They also do not score highly according to the biochemical criterion for functionality. A great illustration of this was recently provided by the sequencing of the genome of *Amborella trichopoda*, the sister lineage of all other flowering plants (Amborella Genome Project 2013). The *Amborella* mitochondrial genome (Rice et al. 2013) consists of five circular-mapping chromosomes, is 3.9Mb in size, and apparently got so big by acquiring large regions (even full-length copies) of mtDNA from other plant species (totaling about six genome-equivalents), as well as plastid DNA. Many of these foreign mitochondrial genes have been pseudogenized. Notably, the expression of the endogenous and foreign genes was assayed (by targeted RT-PCR on total RNA) and only the endogenous ones were found to be expressed. These foreign mtDNA insertions therefore exhibit all the classic features of "junk DNA".

There is, however, a lot to be learned about organelle genomes, their organization and mechanisms of gene expression. This is in many ways even more so regarding mitochondria and plastids than it is about the diversity of nuclear genomes discussed above, because of the stunning diversity of organelle genome structure and organization observed within eukaryotes. As stark as the differences between eukaryotes and prokaryotes are, the nuclear eukaryotic genomes are still most likely derivatives of an ancestral genome similar in its general features to the typical blueprint a modern prokaryotic genome. Organelle genomes represent another diversification of such an ancestral state, albeit one that developed under very different evolutionary pressures. Understanding the many ways in which these genomes have been dramatically reorganized in different lineages would greatly improve our knowledge of the driving forces, and the possibilities and limitations of genome evolution.

As is the case with nuclear genomes, the best studied systems are confined to the classical model systems representing just a few of the major eukaryotic lineages. However, the textbook-example, highly compacted metazoan mitochondrial genome consisting of a single circular-mapping molecule of mtDNA generating a few long polycistronic messages is far from representative. A single circular-mapping molecule is, of course, a very common configuration, as this was the most likely ancestral state, but within it there are large variations in terms of non-coding DNA content, as discussed above. It is highly unlikely that polycistronic messages are generated in genomes in which individual genes are separated by large stretches on non-coding DNA; instead they are likely transcribed and regulated individually, but much less is known about the detailed workings of such systems than it is for mammalian mitochondria. And even within metazoans, very different topologies are observed. For example, the mitochondrial genome in *Hydra* consist of two linear pieces of mtDNA (Warrior & Gall 1985; Voigt et al. 2008), in the human louse *Pediculus humanus* it is composed of 18 individual circular molecules (Shao et al. 2009) and in the mesozoan *Dicyema shimantoense*, the initially circular mtDNA is fragmented in somatic cells into minicircles containing single genes (Watanabe et al. 1999). Other examples of fragmented and/or linear mitochondrial genomes in metazoans are also known (Smith et al. 2012; Shao et al. 2012;)

The most famous example of "weird" mitochondrial genomes are the kinetoplasts of kinetoplastids, which are composed of multiple identical maxicircles, on which the genes reside, plus hundreds of minicircles containing guide RNAs (gRNAs), with these circles being physically intertwined (Simpson 1997; Simpson et al. 1989; Lukes et al. 1998). The gRNAs are necessary for the massive amounts of RNA editing that the genes need to undergo in order to be properly expressed (Benne et al. 1986; Simpson & Thiemann 1995; Blum et al. 1990). Kinetoplastid RNA editing has been usually interpreted as a classic example of constructive neutral evolution. The most widely accepted model for its origins involves the initial acquiring of the capacity for RNA editing, which allows the mutation of coding nucleotides, which in turn leads to the eventual locking of the system into a state in which editing is essential (Covello & Gray 1993), although adaptive explanations for its origin have also been proposed (Speijer 2006).

Even more unusual is the mtDNA organization of diplonemids, a closely related to kinetoplastids lineage (see Figure 16.1). The mitochondrial genome of *Diplonema papillatum* is very large (>600kbp), consisting of about 100 circular 6-7kb molecules (Marande et al. 2005), and the organization is similar in other diplonemids (Roy et al. 2007; Kiethega et al. 2011). Most remarkably, the protein coding genes in these mitochondria are broken into small individual pieces (up to a dozen per gene) each residing on a separate minichromosome (Vlcek et al. 2011), with the rest of the minichromosome containing a highly regular and similar between different minicircles pattern of repeat motifs. The pieces are subsequently joined to produce the full-length message (Kiethega et al. 2011; Marande & Burger 2007). Another interesting case are the mitochondrial genomes of ichthyosporeans, of which

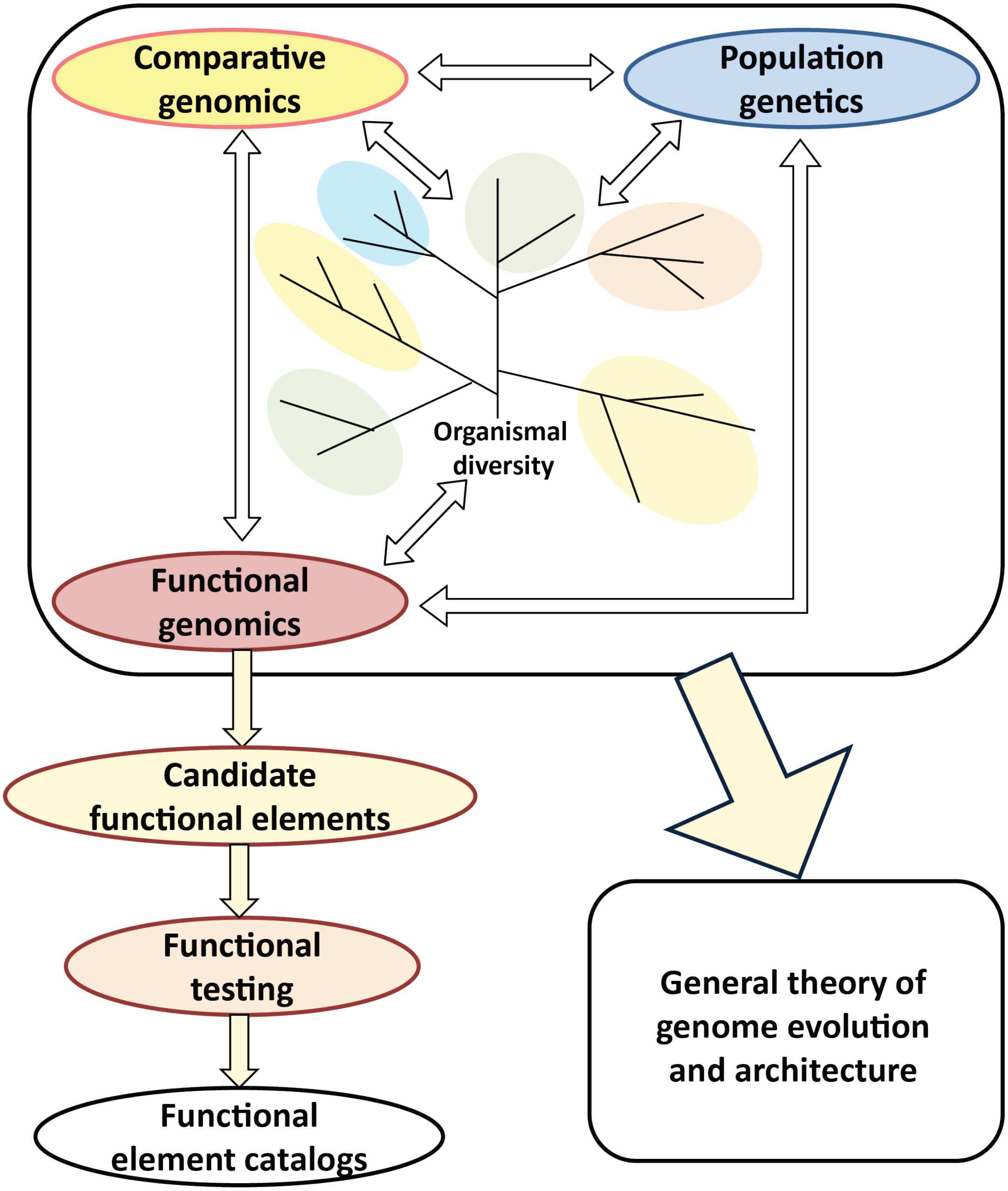

Comparative genomics
Population genetics
Organismal diversity
Functional genomics
Candidate functional elements
Functional testing
Functional element catalogs
General theory of genome evolution and architecture

Figure 16.2: Towards a general understanding of genome function.

the *Amoebidium parasiticum* one has been sequenced. It is >200kb long consisting of hundreds of numerous linear minichromosomes and contains both full-length and fragmented copies of genes (Burger et al. 2003).

As is often the case, dinoflagellates present the farthest deviation from the norm. Their mitochondrial genomes have presented a huge challenge to sequencing because they apparently exist as multiple circles of varying size and composition, with the same genes occurring in different sequence contexts, likely due to very high levels of recombination (Norman and Gray 2001; Jackson et al. 2007; Nash et al. 2008; Waller and

Jackson 2009; Slamovits et al. 2007; Kamikawa et al. 2009). The three protein-coding genes they contain occur in both full-length copies as well as in multiple varying in length and composition fragments. Remarkably, this results in the production of partial transcript fragments that are not only transcribed and polyadenylated separately, but are likely not *trans*-spliced as in diplonemids, with the exception of the *cox3* gene, as well as polycistronic transcripts and transcripts without stop codons (Jackson et al. 2007).

Dinoflagellates also display an amazing configuration of their plastid genome. The typical plastid genome is circular-mapping, between 100 and 200kb long, containing between 100 and 200 genes (depending on the lineage) and is organized into a large and small single-copy regions separated by two inverted repeats (Kolodner & Tewari 1979). Polycistronic transcripts are generated from multiple promoters, by two different RNA polymerases, one plastid-encoded and of cyanobacterial origin (PEP), and another one of phage origin and nucleus-encoded (NEP) (Yagi & Shiina 2014). Plastid genomes display less variation in their structure compared to mitochondria, yet in dinoflagellates the chloroplast genome is of most unusual nature: it is split into multiple minicircles, which are transcribed into polycistronic messages by a rolling circle mechanism (Zhang et al. 1999; Barbrook & Howe 2000; Nisbet et al. 2008; Barbrook et al. 2012).

In general, little is known about the detailed mechanisms of transcriptional and post-transcriptional regulation of organellar genomes and, in turn, what evolutionary forces have shaped them, except for some of the well-studied model organisms (though even in those latter cases much still remains to be learned). Functional genomic tools for characterizing the transcriptome and the protein-DNA and protein-RNA interaction landscapes in organelles should greatly facilitate advances in that area (Smith 2013), and initial studies have already shown the power of these approaches (Sanchez et al. 2011; Mercer et al. 2011; Liu et al. 2013; Wang et al. 2013; Marinov & Wang et al. 2014; Tanifuji et al. 2014; Hotto et al. 2013).

This is even more so the case for nucleomorph genomes. Nucleomorphs are the result of secondary endosymbiosis between two eukaryotes (Gibbs 1978; Hibberd & Norris 1984; Cavalier-Smith 2002). In most lineages with a history of such events, the nucleus and the nuclear genome of the endosymbiont have been lost following massive gene transfer to the host nucleus (Martin et al. 1998), but in two modern groups, the chlorarachniophytes and the cryptophytes, the endosymbiotic nucleus persists in the form of a nucleomorph. The chlorarachniophyte nucleomorph is of green algal origin, while the cryptophyte one is of independent red algal origin, yet these genomes display remarkable convergence in their characteristics (Moore & Archibald 2009). They are highly reduced, <1Mb in size, (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Tanifuji et al. 2011; Curtis et al. 2012), and contain only a few hundred genes. All sequenced nucleomorph genomes consist of three chromosomes, each of which contains (typically subtelomeric) rRNA genes (Silver et al. 2007), with very short intergenic spaces and often overlapping genes (Williams et al. 2005) that are themselves compacted (Lane et al. 2007), though introns are also present in most cases, sometimes in substantial numbers such as in *Bigelowiella natans*. These features represent the most extreme known case of eukaryotic genome reduction and pose very intriguing questions about the transcriptional and regulatory biology of nucleomorphs. The transcriptomes of several nucleomorphs have been recently characterized on a global scale (Tanifuji et al. 2014; Hirakawa et al. 2014), but given the extremely limited intergenic and noncoding space in these genomes, it will be also of great interest to know their chromatin structure, the histone code operating in them and its states (Müller et al. 1994; Löffelhardt 2011; Hirakawa et al. 2011), their transcription factor binding patterns, and as mentioned above, the biology of introns and splicing in them. This should shed light on the evolutionary limits on the process of gene regulation imposed by extreme genome compaction.

16.8 Conclusions

While we are still a long way from having a complete understanding of genome function and evolution, we do at this point have a quite robust explanatory framework accounting for the driving forces behind the appearance of many of the major features of eukaryote biology, and for many of the differences in genome organization, content and structure observed within the known organismal diversity. The concept of "junk" DNA features prominently in this framework, and it is by no means debunked by the results of the ENCODE Project. ENCODE data is entirely consistent with our previous understanding of mammalian genome biology, which it does not overturn but instead adds to and enriches. Indeed, it would have been quite distressing if major rethinking was necessary, given the immense amount of research on and knowledge about the subject that has accumulated over many decades. The real contribution of the ENCODE Project to understanding human biology has been the identification of the candidate functional elements in the genome, each of which will, however, have to be subsequently functionally dissected to fully understand its role in the normal functioning of the cell, development and disease.

Another major, and so far overlooked, contribution of the ENCODE has been the role it has played in the development of the tools and techniques to carry out large-scale functional genomic characterization of genomes. This is critical for achieving the goal of complete understanding of the fundamental principles driving the evolution of eukaryote genomes and the mechanisms of carrying out and regulating gene expression that they use. From this perspective, the study of the comparatively well-characterized human genome served as a proof of principle: the ENCODE Project did not necessarily find completely new principles of gene regulation and RNA biology, but it recovered a lot of what was previously known about the functional organization of our genome, and added further pieces to the puzzle opening new research directions to be pursued in the future. While the detailed study of the individual components of gene expression and gene regulatory systems using classical genetic and biochemical tools will always remain essential, it is now possible to use functional genomic tools to advance the understanding of how a genome works by decades compared to the trajectory this process followed for the traditional model systems.

This would be of little significance if all genomes were the same in their organization, but fortunately, this is very far from being the case. The last two decades have revealed a breathtaking diversity of approaches that different lineages have adopted to solving the problem of being an eukaryote cell, some of them representing truly drastic deviations from the classical textbook depiction of these processes. This diversity provides a great opportunity to understand the evolutionary forces and limitations shaping genome architecture. Thanks to the rapid advances in sequencing and functional genomic methods (to which researchers working on ENCODE Project have made major contributions), all genomes are now either accessible to study (or poised to soon become so) using the powerful tools available to us, allowing the tackling of these major questions, at the deepest level. Such an endeavor should feature the close integration of the disciplines of comparative genomics, population genetics and functional genomics (Lawrie & Petrov 2014), and a fully fleshed, universally agreed upon, theory of genome evolution should eventually emerge from it, together with the detailed understanding of the function of individual genomes (Figure 16.2).

Part VI

Appendices

A

Effects of sequence variation on differential allelic transcription factor occupancy and gene expression

Originally published as:

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, Crawford GE, Wold B, Willard HF, Myers RM. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research.* **22**(5):860–869. doi: 10.1101/gr.131201.111.

Effects of sequence variation on differential allelic transcription factor occupancy and gene expression

Timothy E. Reddy,[1,2] Jason Gertz,[1] Florencia Pauli,[1] Katerina S. Kucera,[2] Katherine E. Varley,[1] Kimberly M. Newberry,[1] Georgi K. Marinov,[3] Ali Mortazavi,[3,4] Brian A. Williams,[3] Lingyun Song,[2] Gregory E. Crawford,[2] Barbara Wold,[3] Huntington F. Willard,[2] and Richard M. Myers[1,5]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [2]Duke Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA; [3]Department of Biology, California Institute of Technology, Pasadena, California 91125, USA

A complex interplay between transcription factors (TFs) and the genome regulates transcription. However, connecting variation in genome sequence with variation in TF binding and gene expression is challenging due to environmental differences between individuals and cell types. To address this problem, we measured genome-wide differential allelic occupancy of 24 TFs and *EP300* in a human lymphoblastoid cell line GM12878. Overall, 5% of human TF binding sites have an allelic imbalance in occupancy. At many sites, TFs clustered in TF-binding hubs on the same homolog in especially open chromatin. While genetic variation in core TF binding motifs generally resulted in large allelic differences in TF occupancy, most allelic differences in occupancy were subtle and associated with disruption of weak or noncanonical motifs. We also measured genome-wide differential allelic expression of genes with and without heterozygous exonic variants in the same cells. We found that genes with differential allelic expression were overall less expressed both in GM12878 cells and in unrelated human cell lines. Comparing TF occupancy with expression, we found strong association between allelic occupancy and expression within 100 bp of transcription start sites (TSSs), and weak association up to 100 kb from TSSs. Sites of differential allelic occupancy were significantly enriched for variants associated with disease, particularly autoimmune disease, suggesting that allelic differences in TF occupancy give functional insights into intergenic variants associated with disease. Our results have the potential to increase the power and interpretability of association studies by targeting functional intergenic variants in addition to protein coding sequences.

[Supplemental material is available for this article.]

Variation in protein coding sequence is interpretable, owing to our knowledge of gene models and the triplet code. Recent studies that utilize exome sequencing take advantage of this knowledge to predict loss-of-function and nonsense mutations (Meyerson et al. 2010; Teer and Mullikin 2010). However, predicting the effects of DNA sequence variation in the large noncoding parts of the genome remains a largely unsolved problem. While transcription factors (TFs) preferentially bind DNA at definable sequence motifs, the motifs are often degenerate and are rarely predictive of binding (Tompa et al. 2005). Recent advances in DNA sequencing technologies allow genome-wide empirical measures of TF occupancy (i.e., chromatin immunoprecipitation followed by sequencing, or ChIP-seq; Johnson et al. 2007; Robertson et al. 2007), revealing that differences in TF occupancy between individuals are common (Kasowski et al. 2010; McDaniell et al. 2010). Furthermore, combining ChIP-seq with personal human genome sequencing has identified instances in which a TF preferentially binds one allele over the other in the same cell type (McDaniell et al. 2010), which we call differential allelic occupancy. Because differential allelic occupancy compares TF binding between alleles in the same nucleus, it is controlled for environmental differences between individuals and cell types and therefore provides a more direct connection between genome sequence and regulatory function than do population-based studies.

To understand the functional consequences of allelic differences in TF occupancy, it is important to measure allelic differences in expression in the same cells. Numerous approaches have been developed to measure differential allelic expression in select genes (e.g., Yan et al. 2002; Gimelbrant et al. 2007; Serre et al. 2008; Main et al. 2009; Zhang and Borevitz 2009; Zhang et al. 2009), with current estimates that 10% of human genes have allele-specific expression (Gimelbrant et al. 2007; Zhang et al. 2009). High-throughput sequencing can identify allelic imbalances in expression when complete genome sequences for both the parents are available, for example in F1 fly hybrids (McManus et al. 2010). When a complete genome sequence is available for a trio of related humans, RNA-seq (Mortazavi et al. 2008) can be used to measure genome-wide allelic imbalances in human gene expression (Degner et al. 2009; Pickrell et al. 2010). However, measurement of differential allelic expression with RNA-seq is limited to genes with heterozygous exonic sequences, which represents less than half of human transcripts.

In this work, we sought to better understand the functional consequences of genomic variation, both on TF occupancy and on gene expression. To do so, we first characterized differential allelic

[4]Present address: Developmental & Cell Biology, University of California, Irvine CA 92697, USA.
[5]Corresponding author.
E-mail rmyers@hudsonalpha.org.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.131201.111.

22:860–869 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

occupancy for 24 TFs and the cofactor *EP300*, as well as heritability of TF occupancy for a subset of those factors. In addition, we measured differential allelic expression using both RNA-seq as well as ChIP-seq of RNA polymerase II (RNA Pol2). The latter enabled prediction of differential allelic expression of genes with homozygous exons but heterozygous introns (Knight et al. 2003), revealing many additional otherwise undetectable instances of differential allelic expression. Together, the results provide many insights into how genome sequence impacts TF occupancy, and the extent to which that occupancy impacts downstream gene expression. The results may also have the potential to improve our understanding of disease, as we found numerous intergenic variants associated with autoimmune diseases to also be differentially bound by TFs. Ultimately, targeting intergenic regions shown to have functional consequence may improve future microarray- and sequencing-based association studies by increasing coverage with only a modest effect on statistical power.

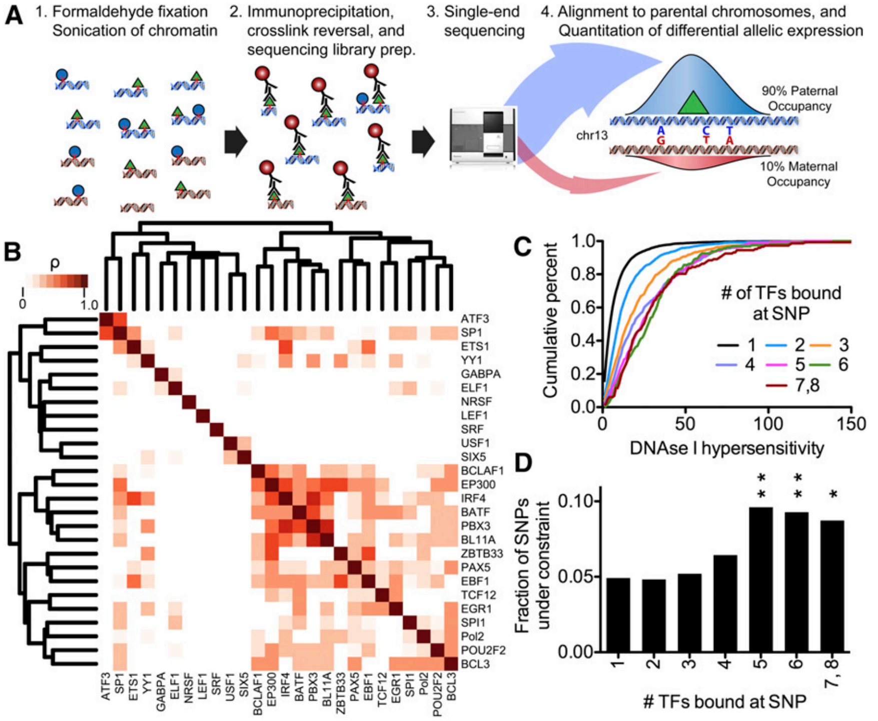


Figure 1. (*A*) Diagram of method used to measure differential allelic TF occupancy. First, chromatin was formaldehyde-fixed and sonicated. Cross-linked TF-binding complexes were then immunoprecipitated with an antibody specific for the TF of interest. The co-precipitated DNA was recovered and subjected to high-throughput single-end sequencing. Reads were aligned to maternal and paternal versions of the GM12878 genome according to data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). For each binding site, differential allelic occupancy was called when reads aligned to a single allele significantly more often than would be expected by random. (*B*) Spearman correlation of allelic imbalance at sites of TF co-occupancy throughout the genome. The color of the boxes indicates the correlation coefficient, with white indicating nonsignificant correlation ($P >$ 0.05). The tree shows complete linkage hierarchical clustering. (*C*) We classified heterozygous variants by the number of TFs binding at that variant. Shown is the cumulative distribution of DNase I hypersensitivity signal at all occupied heterozygous variants in each class, as indicated in the legend. (*D*) For each class of heterozygous variants (as defined in *C*), the fractions of variants with phastCons score >0.5. Asterisks ([**] $P < 0.01$; [*] $P < 0.05$) indicate statistical significance compared to the uniquely bound variants as described in Methods.

Results

Transcription factors often cluster together on the same alleles in regions of open chromatin

To survey the allelic *cis*-regulatory landscape, we performed ChIP-seq on 24 sequence-specific human TFs and the transcriptional co-activator *EP300* in a lymphoblastoid cell line (LCL), GM12878, generated by EBV immortalization of cells from a female (Supplemental Table 1). Whole genome sequencing has been performed on this cell line and on LCLs derived from both of her parents (The 1000 Genomes Project Consortium 2010), and we aligned sequence reads to both the maternal and paternal versions of the genome (see Methods; Figure 1A). We identified 157,586 high-confidence TF occupied regions, of which 20,013 (13%) overlap at least one heterozygous single nucleotide polymorphism (SNP). We found 1094 (5.5%) of heterozygous sites with a significant difference in occupancy between parental chromosomes for at least one TF (false discovery rate, or FDR, <5%) (Supplemental Table 2). When a single binding site covered multiple variants, we observed a consistent allelic imbalance across all variants in the binding site (Supplemental Fig. 1). Differential allelic occupancy was also reproducible between biological replicates (Supplemental Fig. 2), evenly distributed across autosomes (Supplemental Fig. 3), and not substantially biased in favor of the reference allele (Supplemental Table 3). On the X chromosome, TFs predominantly bound the maternal homolog (Supplemental Fig. 4), consistent with reports of a strong bias toward paternal X inactivation in the GM12878 cell line (McDaniell et al. 2010).

We found evidence that TFs commonly interact with each other on the genome, especially at regions with differential allelic occupancy. Overall, 30% of autosomal TF binding sites with significant differential allelic occupancy overlapped another such site (Supplemental Table 4), and the overlaps appeared to follow a power-law distribution (Supplemental Fig. 5). In comparison, we found on average 15% of binding sites overlapping one another among an equal number of sites for which we did not detect significant differential allelic occupancy. The greater overlap in sites of differential allelic occupancy was unlikely to occur by random ($P = 8 \times 10^{-6}$) according to permutation tests that take into account potential biases resulting from antibody-specific variation in ChIP-seq signal strength and the average size of binding sites between different factors and between binding sites with and without differential allelic occupancy (see Supplemental Methods). When multiple TFs bound the same heterozygous SNP, they frequently resided on the same allele, as indicated by positive correlations between allelic occupancy at co-bound SNPs (Fig. 1B; Supplemental Figs. 6, 7). On the contrary, in no case did we observe pairs of TFs that regularly bound the same position on alternate autosomes. In some cases, the factors may bind together in heteromeric complexes. For example, occupancy of the transcriptional co-activator *EP300* correlated with that of many TFs. However, overall, we did not find evidence of known protein–protein interactions supporting our observed correlated occupancy (Persico et al. 2005). Instead, the TF hubs may either include novel TF–TF associations or may be a more general feature of the genomic landscape (MacArthur et al. 2009). Chromatin state may also play a role either in increasing TF occupancy at variants bound by multiple TFs, or in maintaining a state established by pioneer factors. In support of this hypothesis, the DNA near TF hubs had increased sensitivity to DNase I when compared with regions bound by a single factor (Fig. 1C). The result indicates either that these regions of open chromatin were more accessible to TFs before binding, or that the recruitment of many TFs to these regions resulted in more extensive and stable chromatin remodeling. The co-occupied variants may also have particular functional significance, as they were more likely to be evolutionarily conserved than variants bound by a single factor (Fig. 1D). Together, the results reveal the existence of hubs of coordinated differential allelic gene regulation involving multiple TFs throughout the human genome.

Most differential allelic occupancy results from variation outside the DNA binding motif

To better understand the mechanisms underlying differential allelic occupancy, we investigated the genetic contributions to allelic occupancy. Kasowski and colleagues previously found that variation of NFKB binding between different individuals significantly associated with disruption of the NFKB binding motif (Kasowski et al. 2010), and others have suggested a similar relationship may be found for differential allelic occupancy (McDaniell et al. 2010). We therefore sought to determine generally across many TFs how often differences in the primary TF binding site correspond to differential allelic occupancy. We first evaluated the location of heterozygous SNPs in autosomal TF binding sites. We found that, after controlling for biases in read coverage and variant density, differentially occupied sites were strongly enriched for heterozygous SNPs within 50 bp of the position of maximal ChIP signal (Fig. 2A), indicating that they may be the most functionally important nucleotides. We then compared the rate at which heterozygous SNPs occurred at motif versus non-motif intergenic positions (Supplemental Table 5), a ratio we designate dM/dI. Generally across all factors and limited to autosomes, we found that heterozygous variants in motifs were nearly three times more likely to occur in differentially bound sites ($\overline{dM/d1}$ = 2.47) than in equally bound sites ($\overline{dM/d1}$ = 0.80) (Fig. 2B). Compared with an estimated background rate calculated from randomly chosen 5-kb promoter regions ($\overline{dM/d1}$ = 0.98), we found motif-disrupting mutations were significantly enriched in differentially bound regions and significantly depleted in equally bound regions ($P < 1 \times 10^{-100}$ for both cases; see Methods). As expected and consistent with reports of inter-individual variation of NFKB binding (Kasowski et al. 2010), the bound alleles were overall more similar to the consensus motif than the unbound alleles (Supplemental Fig. 8). Differential allelic occupancy ranged from subtle to absolute. Binding sites with the greatest allelic difference in occupancy corresponded to the presence of a canonical binding motif and to mutation of that motif (Fig. 2C,D). However, variants in known binding motifs explained only ~12% of instances of differential allelic occupancy. While the exact percentage is dependent on many factors, it appears that the minority of differential allelic occupancy can be attributed to mutation of a canonical TF binding motif. Instead, our results suggest that there are different regimes of variation in TF binding. At the minority of differentially occupied binding sites, mutation of a canonical binding motif drives strong allelic differences in TF occupancy. Meanwhile, at the majority of differentially occupied sites, TFs bind DNA at weak or noncanonical binding motifs. In such cases, smaller differences in occupancy occur, perhaps via genetic disruption of a cofactor binding site or differences in chromatin structure (McDaniell et al. 2010; Gertz et al. 2011)

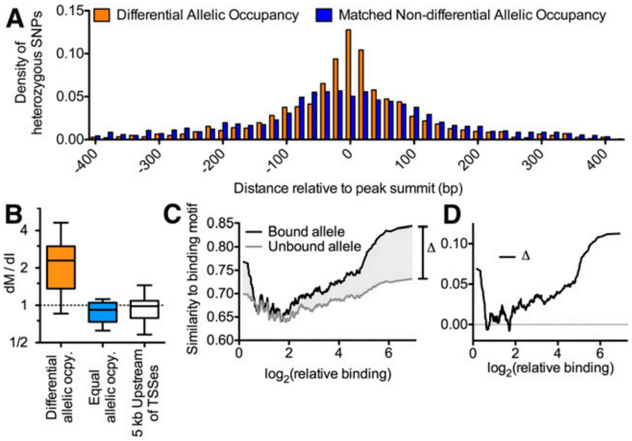


Figure 2. (*A*) Histogram of the distance of heterozygous SNPs from the location of maximal ChIP-seq signal for sites with (orange) and without (blue) differential allelic TF occupancy. To control for potential observation biases resulting from high read coverage at variants near the center of binding sites, the sites of equal allelic occupancy were chosen to match the differential allelic occupancy in two ways. First, for each site of differential allelic occupancy, we required the total number of aligned reads covering heterozygous variants in the matched site to be equivalent. Second, we required that the total number of variants in each binding site was also equivalent. If a suitably matched site did not exist, the site was excluded from the sites of differential allelic occupancy for this analysis. Using this strategy, the distribution of aligned reads at heterozygous variants was not significantly different between the sites of differential allelic occupancy and the matched set of equal allelic occupancy (P = 0.15, two-sided Wilcoxon rank-sum test). (*B*) The ratio of the rate of motif-disrupting to non-motif-disrupting intergenic mutations (dM/dI) across all sited of differential allelic TF occupancy (orange), and at TF binding sites that lack significant differential allelic occupancy (blue). To allow comparison with *cis*-regulatory DNA, the distribution of dM/dI is also shown for regions 5 kbp upstream of 10,000 randomly chosen TSSs (white). Whiskers show 95% confidence intervals. We excluded TFs for which we only observed a single motif-disrupting variant across all binding sites. (*C*) For the bound (black) or unbound (gray) allele at all sites of differential allelic occupancy, the similarity to TF binding motif (as a fraction of the optimal match) at sites of heterozygosity (*y*-axis) plotted against relative binding (the ratio of reads aligning to the bound vs. unbound allele; *x*-axis). Data were smoothed over a 32-data-point sliding window. The shaded region labeled Δ indicates the amount of difference in motif similarity between bound and unbound alleles, and is plotted in panel *D*.

RNA Pol2 occupancy predicts differential allelic expression of genes with homozygous exons

To evaluate the effects of differential allelic occupancy on expression, we used ultrahigh-throughput mRNA sequencing (RNA-seq) (Mortazavi et al. 2008) to measure differential allelic gene expression across the human genome (Pant et al. 2006; Gimelbrant et al. 2007; Zhang et al. 2009). To avoid biases from mapping to the reference genome (Degner et al. 2009; Pickrell et al. 2010), we assembled complete paternal and maternal GM12878-specific versions of all RefSeq transcripts. We then sequenced the transcriptome and aligned the reads to the parental transcripts (Fig. 3A; Supplemental Table 6). We identified significant (FDR < 5%) differential allelic expression for 381 (9%) of the 4194 expressed RefSeq transcripts with heterozygous variants in exonic regions (Fig. 3B). The results were reproducible between biological replicates ($r^2 = 0.88$, $P < 2 \times 10^{-27}$) (Supplemental Fig. 9), and validation with Sanger sequencing reproduced results from six of six tested genes (Supplemental Fig. 10; Gertz et al. 2011). Differences in allelic expression were often subtle: 166 (52%) of the 322 autosomal genes identified had less than a twofold difference in expression between alleles. Known imprinted genes (Morison et al. 2005; Pollard et al. 2008) and X-linked genes were the exception, nearly all of which had a greater than twofold allelic expression difference. Most X-chromosomal genes were transcribed from the maternal copy (Supplemental Figs. 11, 12), as expected, given the paternal X inactivation bias in GM12878 cells (McDaniell et al. 2010; Kucera et al. 2011). We also identified differential allelic expression of seven long noncoding RNAs (Supplemental Fig. 13). Monoallelic expression of *XIST* (Brown et al. 1991) and *KCNQ1OT1* (Weksberg et al. 2003; Nagano and Fraser 2009) is necessary for silencing gene expression on the opposite alleles, and it remains to be seen if any of the additional five that we identified have a similar function (Mohammad et al. 2009; Malecova and Morris 2010).

Allelically imbalanced gene regulation likely results from regulatory sequences that are not in exons, and therefore both heterozygous and homozygous genes may have differential allelic expression. However, measurement of differential allelic expression with RNA-seq is limited to genes with heterozygous exonic sequences, which represents only 39% of the transcripts in GM12878. Chromatin immunoprecipitation of RNA Pol2 isolates DNA from both exons and introns, enabling genome-wide prediction of differential allelic expression of genes with homozygous exons but heterozygous introns (Knight et al. 2003). Aggregating allelic RNA Pol2 ChIP-seq signals across gene bodies, we predicted differential allelic expression for 654 (6.3%) of the 10,353 genes with sufficient coverage of RNA Pol2 at heterozygous variants. The genes included 456 autosomal that lacked exonic heterozygous variants and could not be evaluated with RNA-seq. When we found significant differential allelic expression of X-linked genes, we predicted expression from the expected allele giving us perfect specificity (Fig. 3C). However, not all X-linked genes reached our significance threshold, some of which may escape inactivation. Comparing to a chromosome-wide study of genes subject to or escaping from X inactivation (Carrel and Willard 2005), we estimated that our analysis of RNA Pol2 occupancy achieves 66% sensitivity in predicting X inactivation or escape. Given the perfect specificity, relaxing our significance criteria combined with deeper sequencing may improve the sensitivity. However, for the purposes of this study, we were more concerned with ensuring a high true positive rate. As a further positive control, we measured differential allelic expression and RNA Pol2 occupancy in complementary clonal isolates of GM12878 with paternal or maternal X chromosomes inactivated. For both RNA-seq and RNA Pol2 occupancy, we predicted that >80% of genes with differential allelic expression were transcribed from the expected X chromosome in these clonal cell populations (Supplemental Figs. 14–16). On the autosomes, however, we see strong concordance in allelic expression among clonal isolates as well as with the original cell population (Supplemental Fig. 17). Searching for evidence of random monoallelic expression that could explain the observed differential allelic expression (Gimelbrant et al. 2007), we found that 13.5% of genes with differential allelic expression in one clone were either bi-allelic or expressed from the homologous chromosome in a different clone (Supplemental Table 7). While only a limited number of clones were studied, the result suggests that the minority of differential allelic expression results from random monoallelic expression. Across the autosomes, allelic differences in RNA Pol2 across the gene body positively predicted allelic differences in expression for 135 (92%) of the 146 genes that were also detected in RNA-seq ($P = 1 \times 10^{-27}$, Fisher's exact test). That variation in differential allelic RNA Pol2 occupancy significantly but imperfectly explains variation in gene expression ($r^2 = 0.48$, $P < 1 \times 10^{-16}$) (Supplemental Fig. 18) may result both from technical noise in genome-wide measurements of allelic RNA Pol2 occupancy as well as from biological sources such as differential rates

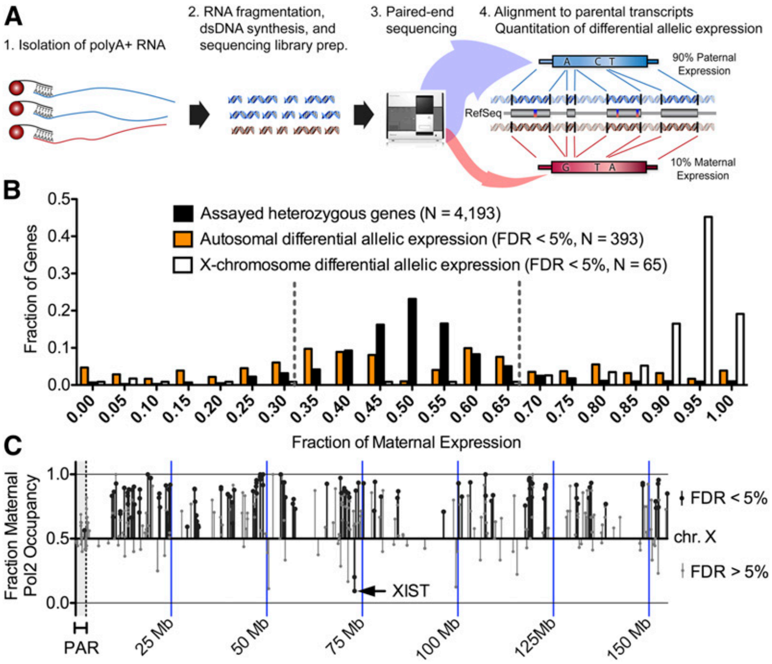


Figure 3. (*A*) Diagram of our method for using RNA-seq to measure differential allelic expression. First, poly(A)$^+$ RNA was isolated using magnetic beads conjugated to oligo(dT) nucleotides. After RNA fragmentation, dsDNA was synthesized and subjected to paired-end sequencing on an Illumina Genome Analyzer. Reads were then aligned to GM12878-specific maternal and paternal versions of all RefSeq transcripts. Differential allelic expression was called when significantly more reads aligned to a single allele than would be expected by random. (*B*) Distribution of the fraction of maternal expression for all heterozygous genes (black), autosomal genes with differential allelic expression (orange), and X-chromosomal genes with differential allelic expression (white). (*C*) Prediction of differential allelic expression (*y*-axis) along the X chromosome (*x*-axis) using allelic occupancy of RNA Pol2. (Black lines) Significant differential allelic RNA Pol2 occupancy; (gray lines) nonsignificant binding. The shaded region on the *left* indicates the pseudoautosomal region that is not inactivated. All significant differential allelic occupancy predicted expression as expected. Genes that do not achieve statistical significance in the inactivated region of the X were a mix of genes that are known to escape inactivation as well as false negatives.

of transcriptional initiation or elongation, or by allelic differences in RNA stability. Combining evidence of differential allelic expression from RNA-seq and from RNA Pol2 ChIP-seq, we thus identified 910 genes with differential allelic expression in GM12878. The list of all genes with differential allelic expression is provided in Supplemental Materials.

Transcription factor occupancy is more directly inherited than gene expression

While differential allelic occupancy and expression are prevalent in an individual, understanding the extent to which these traits are inherited is critical to understanding how they contribute to heritable disease risk. To investigate, we measured genome-wide both the occupancy of five TFs (*GABPA*, *POU2F2* a.k.a. *OCT2*, *PAX5*, *SPI.1* a.k.a. *PU.1*, and *YY1*) and also gene expression in LCLs derived from both the mother and the father of the GM12878 donor. When a TF had differential allelic occupancy at a heterozygous autosomal variant in GM12878, and each parent was homozygous for one of the alleles, the allele with stronger binding in GM12878 had greater ChIP-seq signal in the corresponding parent in 81% of cases, significantly more often than previously reported for *CTCF* (McDaniell et al. 2010) (P = 1.5 × 10^{-5}, binomial test). We also found that the extent of differential allelic occupancy in GM12878 strongly correlated with differential occupancy between the parental LCLs (Spearman's ρ = 0.75) (Fig. 4A). On the contrary, differential allelic expression of autosomal genes was less directly heritable than differential allelic occupancy (ρ = 0.24, P = 2.1 × 10^{-6}) (Fig. 4B), with the more highly expressed allele in GM12878 having greater expression in the corresponding parental cell line for 60% of genes (P = 3 × 10^{-4}; Fisher's exact test). The reduced heritability of expression likely reflects the integration of a complex mixture of regulatory contributions from both parents, acting both in *cis* and in *trans*, as well as epigenetic contributions. In comparison, individual TF binding sites appear to be more strongly determined by local sequence signals and less affected by the surrounding genomic milieu.

Genes with differential allelic expression are expressed at lower levels in many human cell lines

To investigate the comparatively weak inheritance of gene expression, we first looked for evidence of mechanisms that compensate for allelic differences in the expression of autosomal genes. To do so, we used RNA Pol2 occupancy to identified genes with and without evidence of differential allelic expression, and used RNA-seq to compare expression between the two sets of genes. To control for potential biases due to sample size and RNA Pol2 coverage, for each gene with differential allelic expression we selected a matched gene with a similar amount of RNA Pol2 coverage at heterozygous positions (see Supplemental Methods). If allelic imbalances in autosomal gene expression were compensated, we would expect an overall similar level of expression between the two sets of genes. Contrary to this hypothesis, we found that genes with differential allelic expression have substantially and significantly lower expression than genes expressed equally from both alleles (Fig. 4C). The result is independent of the read coverage threshold, as we have reproduced the result at the RNA Pol2 ChIP-seq coverage threshold ranging from 25× to 120× (Supplemental Table 8). To see if the increased allelic variability of lowly expressed genes was specific to GM12878 cells, we measured gene expression of eight additional cell lines and found that the same genes were significantly less expressed in those cell lines as well (Supplemental Fig. 19). Therefore, it appears that genes with differential allelic occupancy generally have lower expression, perhaps due to fundamental differences in the *cis*-regulatory landscape surrounding these genes. With the exception of immunoglobulin genes and the proto-

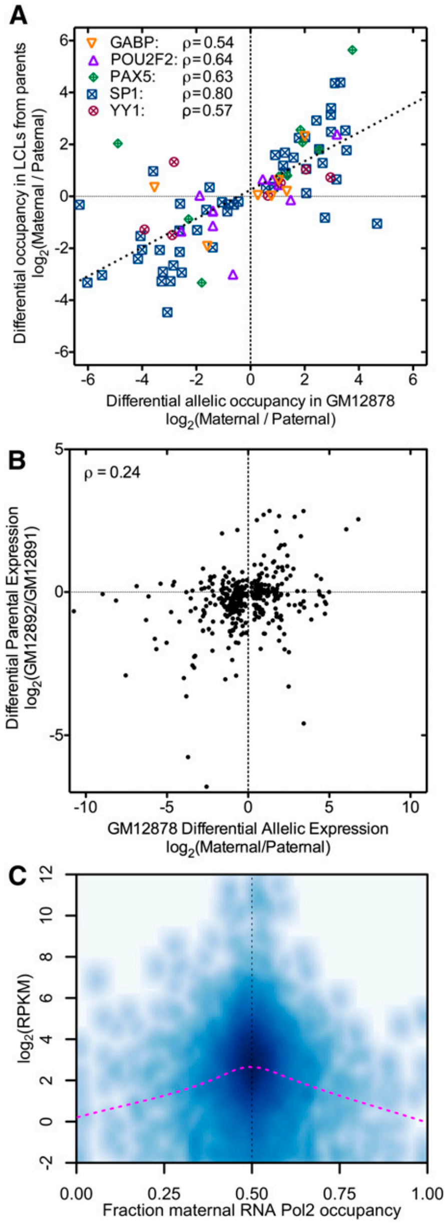


Figure 4. (*A*) Inheritance of allelic TF occupancy. The log-ratio of occupancy of the indicated TFs in the maternally versus paternally derived LCLs (*y*-axis) is plotted against the allelic occupancy of the same factors in GM12878 (*x*-axis). For each site plotted (*N* = 85), we required that both parents were homozygous for alternate alleles. Combining all points together, the overall correlation is ρ = 0.75, and for 88% of sites, the more bound allele in GM12878 was also more bound in the corresponding parent. (*B*) Similar to *A*, the log-ratio of expression from the parental LCLs plotted as a function of the allelic expression in GM12878. (*C*) Genes with differential allelic expression have overall lower expression in GM12878. For each gene with expression >0.25 RPKM, the gene expression (*y*-axis) is shown as a function of differential allelic RNA Pol2 occupancy (*x*-axis). (Darker shading) Greater density of values; (magenta line) less smoothing over the data.

cadherin-gamma cluster, both known to exhibit monoallelic expression patterns (Kaneko et al. 2006), we did not find evidence that genes with differential allelic expression were enriched for particular classes or functions of proteins.

Transcription factor occupancy explains expression up to 100 kb from transcription start sites

One of the major advantages of studying differential allelic occupancy and expression is the potential to link intergenic variants implicated in diseases with functional changes in TF occupancy and gene expression. It is therefore important to know the extent to which allelic TF occupancy correlates with allelic gene expression, especially considering our finding that gene expression was weakly heritable. Overall, we found more TF and cofactor occupancy at variants associated with regulation of gene expression (Montgomery et al. 2010) than would be expected by random (see Supplemental Methods), strongly suggesting that the occupancy we measured does indeed impact gene expression. To investigate further, we evaluated the local *cis*-regulatory landscape of autosomal genes to determine if differential allelic TF occupancy occurred near genes with differential allelic expression. We found that differential allelic occupancy was significantly closer to genes with differential allelic expression than without ($P = 5.0 \times 10^{-15}$, Wilcoxon test comparing the distance to the nearest TSS of a gene with differential vs. equal allelic expression) (Fig. 5A). In contrast, binding sites with equal allelic occupancy were on average no closer to genes with imbalanced or balanced allelic expression ($P = 0.21$, two-sided Wilcoxon test) (Fig. 5B). The fact that differential allelic occupancy occurred closer to genes with differential allelic expression did not result from differences in the total number of observed binding sites, but instead from a greater fraction of the TF binding sites around genes with differential allelic expression having differential allelic occupancy. Specifically, 6.8% of sites within 100 kb of a TSS with differential allelic expression had differential allelic occupancy, compared to 3.9% of sites within 100 kb of a TSS without differential allelic expression ($P < 1 \times 10^{-20}$, Fisher's exact test). Finally, we did not observe a significant difference in the total number of binding sites in the same regions. The association between differential allelic occupancy and expression suggests we may be able to observe a functional relationship between the two.

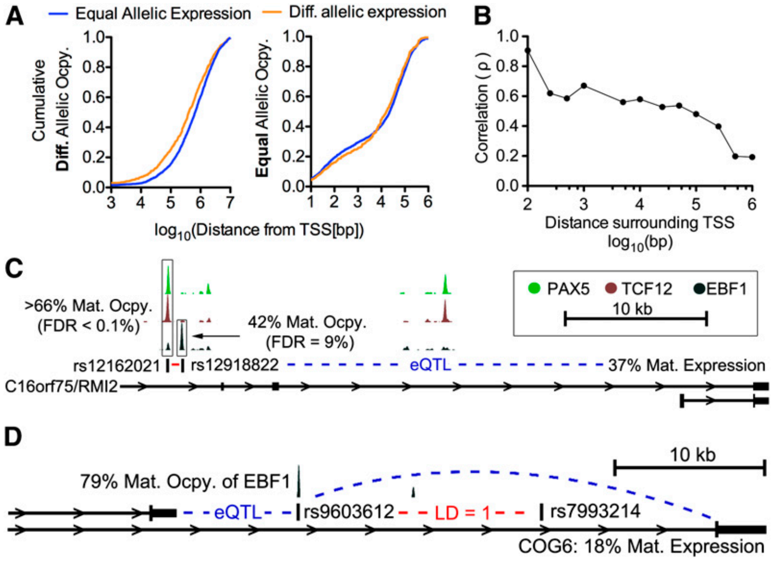


Figure 5. (*A*) Cumulative distribution of the distance from the TSS (*x*-axis) to the nearest site of differential allelic occupancy for all autosomal genes with differential allelic (orange) or equal allelic (blue) expression. (*Left*) All genes with differential allelic expression, where the difference between the two distributions is highly significant. (*Right*) Genes with equal allelic expression, and there is no significant difference between the two distributions. (*B*) Spearman's correlation (*y*-axis) of allelic occupancy with allelic expression within the distance from autosomal TSSs indicated on the *x*-axis. For each point, we aggregated all allelic occupancy (both for sites with and without a significant allelic imbalance) at the indicated distance around all genes with significant differential allelic expression. Then, for every gene with at least a single site with a significant differential allelic occupancy, we calculate Spearman's correlation coefficient and plot. Detailed scatter plots are included in Supplemental Figure 20. (*C*) Differential allelic occupancy of multiple factors at variants either directly or through perfect linkage disequilibrium ($R^2 = 1$; red dash) with celiac disease. Nearby, *RMI2* (also known as *C16orf75*) is predominantly expressed from the maternal allele, and the regulatory interaction is supported by expression quantitative trait loci (eQTL) mapping. (*D*) Similar to *C*, allelic occupancy of *EBF1* at a variant associated (via linkage disequilibrium) with psoriasis corresponds with differential allelic expression of *COG6*. Again, the regulatory interaction is supported by eQTL analysis.

Limited to autosomal cases in which we found allelic imbalance both in occupancy and in expression, the ability of allelic occupancy to explain allelic expression depended on the proximity of binding to the transcription start site (TSS). In the few cases where we observed allelic occupancy within 100 bp of the TSS, we found strong positive correlation between allelic occupancy and expression from the same allele ($\rho = 0.91$, $N = 13$). Meanwhile, allelic occupancy at intervals between 1 and 100 kb from the TSS weakly explained expression ($\rho = 0.45$, $N = 290$). More than 100 kb from the TSS, differential allelic occupancy did not significantly explain expression ($\rho = 0.06$, $N = 760$) (Fig. 5B). The results show that differential allelic occupancy does indeed correspond to differential allelic expression, and may therefore give functional hypotheses to intergenic disease-associated variants. Notably, while the analysis included binding from all TFs and did not attempt to distinguish activating from repressive binding sites or factors, we observed an overall positive correlation. The result suggests either that the TFs chosen in the study are more commonly activating than repressing, or alternatively that activating sites are more amenable to detection by ChIP-seq.

Allelic variation in TF occupancy in GM12878 provides insights into autoimmune disease

The majority of genomic variants associated with disease using genome-wide association studies (GWAS) are intergenic and have unclear regulatory consequences. TF binding sites may give functional insights into the variants identified. Using our observations of TF binding and differential allelic occupancy, we investigated a compilation of disease-associated variants (Hindorff et al. 2009) for potential overlaps that suggest function. Overlap with differential allelic occupancy is particularly interesting because the variant may also explain the difference in TF occupancy between the two alleles. We found 155 unique autosomal variants that were either directly associated with disease, or that were in perfect linkage disequilibrium ($R^2 = 1$) with a disease-associated variant, that also occurred in a heterozygous TF binding site. The overlap was unlikely to occur by random when compared to a set of variants matched on distance relative to a TSS and on minor allele frequency (Supplemental Table 9). Of those variants, we found 21 instances of disease-associated variants that occurred in a site of differential allelic occupancy. More than 75% of the disease-associated variants are associated with autoimmune diseases, including variants associated with multiple sclerosis, celiac disease, Type 1 diabetes, systemic lupus erythematosus, and psoriasis (Supplemental Table 10). The result is especially compelling considering that the functional differences are identified in a cell type relevant for immune modulation (B-cells), and is in agreement with recent findings of a study evaluating genome-wide chromatin states in the same cells (Ernst et al. 2011). As an example, we found a cluster of TFs including *EBF1* and *PAX5*—two key factors in B-cell development—binding with a more than twofold preference to the maternal (protective) allele at variants in complete linkage disequilibrium with the celiac disease-associated variant rs12928822 (Dubois et al. 2010). The variants are found near isoforms of *RMI2*, a gene important for genomic stability. In our study, *RMI2* also shows differential allelic expression, but from the opposite homolog. Furthermore, evidence from expression quantitative trait loci (eQTL) mapping (Dubois et al. 2010) substantiates the presence of a regulatory interaction between the variant and the *RMI2* (Fig. 5C). In another example, we found differential allelic occupancy of *EBF1* at the psoriasis-associated variant rs9603612 and expression of the nearby gene *COG6*, a gene involved in the structure of the Golgi apparatus, again from the opposite homolog (Fig. 5D; Liu et al. 2008). Again, eQTL linkage between the variant and COG6 supports the presence of a regulatory interaction (Zeller et al. 2010).

Discussion

Understanding the impact of genetic variation on gene regulation remains a major challenge in deciphering the human transcriptional regulatory code. To uncover functional noncoding variants we used ultra-high throughput sequencing to measure genome-wide gene expression and occupancy of RNA Pol2, of the transcriptional co-activator *EP300*, and of 24 sequence-specific TFs in the female LCL GM12878. By aligning sequence reads to versions of the reference human genome modified to include homozygous and heterozygous variants identified by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010), we measured allelic differences both in gene expression and in TF occupancy. In doing so, we have produced an extensive and detailed map of transcripts that show allelic bias in expression and alleles that impact TF binding.

Comparing genomic occupancy between multiple TFs, we found that hubs of TF occupancy occur frequently in the human genome: ~15% of the TF binding sites in our study overlapping a binding site for another factor. An abundance of TF-binding hubs have also been found in fly (MacArthur et al. 2009) and may be a common feature of the *cis*-regulatory landscape in complex genomes. The hubs often exhibited a coordinated reaction to functional variants. In such cases, the co-occupying factors bound similarly to the same allele, suggesting a cooperative behavior at such sites. The overabundance of allelically imbalanced hubs also suggests that TF hubs are particularly sensitive to genetic variation, and that genetic polymorphism can destabilize occupancy across the entire hub as opposed to that of a single factor. We also found that the DNA in the most populated hubs had greater evolutionary conservation, suggesting they may play an important role as enhancers of distal gene regulation.

To link allelic TF occupancy to gene expression outcomes, we also characterized differential allelic gene expression across the genome. We used a combination of techniques to measure allelic gene expression. While RNA sequencing gives a direct measurement of allelic gene expression, we found that the majority of protein-coding genes have no heterozygous variants in their exons. Leveraging the ability of ChIP-seq to detect elongating RNA Pol2 at heterozygous variants in introns and to serve as a proxy for gene expression, we developed a complementary approach to measure genome-wide allelic expression of exonically homozygous genes. Our findings suggest that differential allelic expression is as common in genes with genetically identical transcripts as in genes with genetically different transcripts, and that the majority of differential allelic expression is therefore not detectable by comparing mRNA abundance. Comprehensively characterizing such cases of cryptic differential allelic expression may be important in better understanding haploinsufficiency-based disease by revealing many more instances of monoallelic gene expression than are currently known.

Looking across all genes with differential allelic expression, we found that such genes are more likely to be lowly expressed, even in unrelated cell lines. The finding may indicate a closer link between gene expression and evolutionary conservation than has previously been shown. The protein-coding sequences of highly expressed genes are in general more conserved than that of lowly expressed genes (Pal et al. 2001; Subramanian and Kumar 2004; Wall et al. 2005), and our findings suggest that the transcriptional regulation of highly expressed genes is also more conserved. Similarly, it has also been shown that genes with expression limited to specific tissues have less constrained protein coding sequence (Duret and Mouchiroud 2000), and we found evidence that genes with differential allelic expression are expressed in fewer tissues (Supplemental Fig. 21). It may be that the evolutionary pressures or other mechanisms of constraint introduced by increased and organism-wide expression act more broadly than protein coding sequence and also limit allelic variation in the regulation of the same genes.

With a more complete characterization of differential allelic expression, we were able to link allelic TF occupancy to these genes, showing that differential allelic occupancy is more prevalent near differential allelic expression. Ultimately, we found allelic occupancy within 100 bp of the TSS to be highly predictive of expression. However, while we detected significant associations between occupancy and expression up to 100 kb away from a TSS, the associations were comparatively very weak. The finding highlights the ongoing challenge of understanding the extent to which distal *cis*-regulatory elements contribute to expression, and may underlie the weak penetrance that genetic variation at many intergenic variants has in genome-wide association studies. It is also important to note that, while many factors are known to act both as an activator and a repressor, we did not observe any systematic inverse relationships between allelic TF occupancy and expression. The result may be explained by studies in inducible systems that have found the repressive activity of TFs to be predominantly associated with occupancy distal to the TSS (e.g., Cheng et al. 2009; Reddy et al. 2009).

Targeted exon sequencing is becoming a common tool for identifying rare coding variants that may be associated with disease. From genome wide association studies it is clear that many regulatory variants are also associated with disease, but due to their predominantly intronic or intergenic location (Hindorff et al. 2009) as well as the complex nature of *cis*-regulation, such variants are more difficult to functionally interpret. The compendium of functional noncoding variants we have identified provide a resource for identifying noncoding polymorphisms that are likely to have an effect on genomic function, suggesting a compromise between GWAS and exon sequencing. By using a capture approach that includes functional intergenic regions in addition to exons, targeted sequencing can explore a greater fraction of the potentially functional genome while limiting the number of hypotheses being tested. By expanding exon sequencing to include targeted regulatory regions, it may therefore be possible to identify rare intergenic variants that are significantly associated with disease. Meanwhile, the prior knowledge of particular TFs bound in each region provides a mechanistic hypothesis to investigate in more detail, overcoming another of the major challenges in existing association studies (Freedman et al. 2011). That many of the functional variants identified in this study overlap with previously identified disease associated SNPs provides hope that augmenting disease studies with targeted sequencing of functional regulatory variation will ultimately be a successful strategy.

Methods

Cell growth

Biological replicates of GM12878, GM12891, and GM12892 cells were grown in RPMI 1640 media with 2 mM L-glutamine, 15% fetal bovine serum, and 1% penicillin-streptomycin at 37°C under 5% carbon dioxide.

ChIP-seq

We performed ChIP experiments and prepared the immunoprecipitated DNA for sequencing on an Illumina Genome Analyzer as described (Johnson et al. 2007). We selected factors to include both ubiquitous TFs and cofactors (e.g., *SP1* and *EP300*), and factors specific to the development of B-cells (e.g., *POU2F2*, *SPI1*, *PBX3*, *BCL3*, and *EBF1*). Antibodies used are listed in Supplemental Table 1. For each factor, we produced ≥12 million 36 nucleotide reads per biological replicate. We aligned reads to the GM12878-specific reference genome using Bowtie (Langmead et al. 2009) with options "-n 2 -l 36 -k 1–best", and removed alignments mismatching at any heterozygous SNP. To avoid potential biases resulting from amplification artifacts, we collapse all sequences identified multiple times to a single instance. To define binding regions, we used QuEST (Valouev et al. 2008) with "stringent peak calling parameters". For each binding region, we estimated the fraction of maternal (paternal) occupancy as the fraction of mini-contig alignments that mapped to the maternal (paternal) chromosome.

For RNA Pol2, we produced 64 million additional paired-end 100-bp reads by using a similar protocol and the Illumina HiSeq 2000 sequencer. We aligned each end independently against the GM12878-specific reference genome using Bowtie (Langmead et al. 2009) with options "–best–strata -n 2 -m 10 -k 1", and excluded alignments that mismatched at any heterozygous SNP. We predicted the fraction of maternal expression as the fraction of mini-contig alignments across each RefSeq gene that mapped to the maternal allele. To ensure stringency, we only considered genes with reads aligning to at least three heterozygous SNPs.

RNA-seq

Paired-end RNA-seq experiments were performed in biological replicate as described previously (Trapnell et al. 2010). Replicate one and two were sequenced to a depth of 44 and 25 million paired-end 75-bp reads, respectively. We aligned reads to the reference transcriptome using Bowtie (Langmead et al. 2009) with parameters "-a–best–strata" and default paired-end settings. The parameters were chosen to allow alignment to multiple isoforms. We then removed any alignments that resulted in mismatches at heterozygous SNPs. Finally, we aligned RNA-seq reads to the reference transcriptome, and estimated the fraction of expression from the maternal (paternal) chromosome as the fraction of reads mapping to a heterozygous SNP that contain the maternal (paternal) allele.

Sequence alignment and determination of differential allelic occupancy and expression

To measure differential allelic occupancy, we constructed a GM12878-specific reference genome that allowed concurrent alignment to both the maternal and paternal genome as suggested by Degner et al. (2009). Maternal and paternal genome sequences were determined using variants in the March 2010 data release by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). To construct the maternal and paternal genomes, we first altered homozygous SNPs in the hg18 reference genome to match the GM12878 genotype. Then, for each heterozygous SNP with discernable parent-of-origin, we replaced the SNP and the flanking 35 bp (for a 36-bp read length) with a paternal and a maternal version of the sequence. We then combined overlapping sequences such that any read aligning to a parental sequence will overlap a heterozygous SNP and vice versa. For RNA Pol2, we used RefSeq genes instead of peak calls, and only considered genes with reads aligning to at least three heterozygous SNPs.

To measure differential allelic expression, we aligned RNA-seq reads to a GM12878-specific reference transcriptome that included both maternal and paternal versions of all transcripts with a heterozygous variant in an exon. To do so, we first assembled sequences for all RefSeq transcripts from the hg18 reference human genome. We then corrected all homozygous SNPs to match the sequence of GM12878. Then, we created a paternal and maternal version of each transcript with a heterozygous exon by changing heterozygous nucleotides to match the parental chromosome, if known.

We performed a number of additional filtering steps to remove false positives. First, to remove artifacts due to incorrect genome sequence and copy number variation, we removed from analysis variants with a substantial allelic bias in sequencing of input control DNA (i.e., DNA from chromatin that was cross-linked and sonicated, but not immunoprecipitated). We also removed variant calls that were discordant with sequencing of the GM12878 genome as performed by Complete Genomics (Drmanac et al. 2010). Next, we filtered reads that aligned to positions in the genome for which either the maternal or paternal sequence were not unique and could have therefore arisen from a different location, as sequences aligning to such positions are inherently biased to a single allele (Degner et al. 2009). To do so, we simulated every possible 36-bp read that would overlap a heterozygous variant. We then aligned all such reads to the maternal and paternal genomes, and noted every genomic position that did not have a unique 36-bp alignment for either the maternal or paternal version (i.e., reads for which the maternal or paternal variant could also align elsewhere in the genome, or could originate from elsewhere in the genome). The additional screening step reduced the number of sites of differential allelic occupancy by 1.5%. Lastly, we removed 10 (<0.05% of total) SNPs that overlapped regions of aneuploidy as measured by microarray experiments (Supplemental Table 11).

To determine statistical significance of differential allelic expression or occupancy, we used a binomial test against the null hypothesis that an equal number of reads maps to each chromosome. For all statistical testing, we require a 7× coverage threshold because it is the minimum number of reads required to achieve significance with a binomial test. We corrected for multiple hypotheses using the method of Benjamini and Hochberg (Benjamini and Hochberg 1995) implemented in the R statistical package.

Identification of differential allelic occupancy at disease-associated variants

Disease-associated variants were obtained from the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies on April 19, 2011. We then expanded the list to include all variants known to be in perfect linkage disequilibrium ($R^2 = 1$) in individuals of central European ancestry according to the HapMap project. Comparing the list with resequencing of the GM12878 genome, we identified all disease-associated variants that are heterozygous in GM12878. Finally, we identified all such variants that also had significant differential allelic occupancy by one or more TFs at the same SNP.

To determine if the overlap with TF occupancy was greater than expected by random, we used a permutation approach. To do so, we randomly assigned disease association among the phased (i.e., where the inheritance of each allele is unambiguous) heterozygous variants in GM12878, controlling for observation biases in GWAS studies in three ways: (i) maintaining a matched distribution of minor allele frequencies (with 5% absolute value difference), (ii) maintaining a matched distance to the TSS of the nearest RefSeq gene (with 1 kb), and (iii) maintaining both similar minor allele frequency (within 10% absolute value difference) and similar distance to the nearest RefSeq TSSs (within 2 kb). For the third group, we used relaxed stringency in order to assure that we could find enough matched sets. For (i) and (ii), we performed 1000 random sets and for (iii) we used 150 random sets. We then count the number of unique variants that overlap TF binding from our study, and describe the resulting distribution in Supplemental Table 9.

Data access

All ChIP-seq and RNA-seq data are publicly available from the ENCODE repository on the UCSC Genome Browser. Details of accession numbers can be found in Supplemental Tables 12 and 13. In addition, processed data specific to our study including allele-specific alignments, aggregation over variants, binding site calls, and aggregation of allelic alignments over those called binding sites are available online at http://hudsonalpha.org/sites/default/files/DataSets/Myerslab/Differential_allelic_occupancy_and_expression.

Acknowledgments

We thank Chris Gunter, Greg Cooper, and the members of the Myers lab for contributions and suggestions. This work was funded by NHGRI ENCODE Grant 5U54HG004576 to R.M.M. and B.W. Support for T.E.R. was from NIH/NIAMS fellowship 5T32AR007450.

Authors' contributions: T.E.R., J.G., K.E.V., H.F.W., and R.M.M. conceived and designed the study, T.E.R. performed and interpreted the analysis, and wrote the manuscript. J.G. carried out the cloning-based validation of the RNA-seq experiments. F.P. and K.M.N. carried out the ChIP-seq experiments and RNA-seq experiments for the clonal isolates of GM12878. L.S. and G.E.C. performed and contributed to the interpretation of the DNase I hypersensitivity experiments. K.S.K. and H.F.W. designed and created the clonal GM12878 isolates, including determining the X inactivation state. G.K.M., A.M., B.A.W., and B.W. designed and performed the RNA-seq experiments. All authors contributed to the editing of the manuscript.

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349:** 38–44.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434:** 400–404.

Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D, et al. 2009. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19:** 2172–2184.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25:** 3207–3212.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327:** 78–81.

Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A, et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42:** 295–302.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17:** 68–74.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, et al. 2011. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43:** 513–518.

Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, Kucera KS, Willard HF, Myers RM. 2011. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet* **7:** e1002228. doi: 10.1371/journal.pgen.1002228.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318:** 1136–1140.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kaneko R, Kato H, Kawamura Y, Esumi S, Hirayama T, Hirabayashi T, Yagi T. 2006. Allelic gene regulation of *Pcdh-α* and *Pcdh-γ* clusters involving both monoallelic and biallelic expression in single Purkinje cells. *J Biol Chem* **281:** 30551–30560.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in transcription factor binding among humans. *Science* **328:** 232–235.

Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. 2003. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet* **33:** 469–475.

Kucera KS, Reddy TE, Pauli F, Gertz J, Logan JE, Myers RM, Willard HF. 2011. Allele-specific distribution of RNA polymerase II on female X chromosomes. *Hum Mol Genet* **20:** 3964–3973.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, Wise C, Miner A, Malloy MJ, Pullinger CR, et al. 2008. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet* **4:** e1000041. doi: 10.1371/journal.pgen.1000041.

MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV, et al. 2009. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10:** R80. doi: 10.1186/gb-2009-10-7-r80.

Main BJ, Bickel RD, McIntyre LM, Graze RM, Calabrese PP, Nuzhdin SV. 2009. Allele-specific expression assays using Solexa. *BMC Genomics* **10:** 422. doi: 10.1186/1471-2164-10-422.

Malecova B, Morris KV. 2010. Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. *Curr Opin Mol Ther* **12:** 214–222.

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328:** 235–239.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res* **20:** 816–825.

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11:** 685–696.

Mohammad F, Mondal T, Kanduri C. 2009. Epigenetics of imprinted long noncoding RNAs. *Epigenetics* **4:** 277–286.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464:** 773–777.

Morison IM, Ramsay JP, Spencer HG. 2005. A census of mammalian imprinting. *Trends Genet* **21:** 457–465.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Nagano T, Fraser P. 2009. Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mamm Genome* **20:** 557–562.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158:** 927–931.

Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. 2006. Analysis of allelic differential expression in human white blood cells. *Genome Res* **16:** 331–339.

Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G. 2005. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* (Suppl 4) **6:** S21. doi: 10.1186/1471-2105-6-S4-S21.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.

Pollard KS, Serre D, Wang X, Tao H, Grundberg E, Hudson TJ, Clark AG, Frazer K. 2008. A genome-wide approach to identifying novel-imprinted genes. *Hum Genet* **122:** 625–634.

Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. 2009. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res* **19:** 2163–2171.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4:** 651–657.

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, et al. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. *PLoS Genet* **4:** e1000006. doi: 10.1371/journal.pgen.1000006.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168:** 373–381.

Teer JK, Mullikin JC. 2010. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* **19:** R145–R151.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137–144.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5:** 829–834.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci* **102:** 5483–5488.

Weksberg R, Smith AC, Squire J, Sadowski P. 2003. Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Hum Mol Genet* **12:** R61–R68.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* **297:** 1143.

Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, et al. 2010. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5:** e10693. doi: 10.1371/journal.pone.0010693.

Zhang X, Borevitz JO. 2009. Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* **182:** 943–954.

Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM, et al. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6:** 613–618.

Received August 26, 2011; accepted in revised form February 1, 2012.

B

The ENCODE Project Consortium - An integrated encyclopedia of DNA elements in the human genome

Originally published as:

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012. **489**(7414):57–74. doi: 10.1038/nature11247.

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with regard to non-coding RNAs, alternatively spliced transcripts and regulatory sequences. Systematic analyses of transcripts and regulatory information are essential for the identification of genes and regulatory regions, and are an important resource for the study of human biology and disease. Such analyses can also provide comprehensive views of the organization and variability of genes and regulatory information across cellular contexts, species and individuals.

The Encyclopedia of DNA Elements (ENCODE) project aims to delineate all functional elements encoded in the human genome[1–3]. Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure). Comparative genomic studies suggest that 3–8% of bases are under purifying (negative) selection[4–8] and therefore may be functional, although other analyses have suggested much higher estimates[9–11]. In a pilot phase covering 1% of the genome, the ENCODE project annotated 60% of mammalian evolutionarily constrained bases, but also identified many additional putative functional elements without evidence of constraint[2]. The advent of more powerful DNA sequencing technologies now enables whole-genome and more precise analyses with a broad repertoire of functional assays.

Here we describe the production and initial analysis of 1,640 data sets designed to annotate functional elements in the entire human genome. We integrate results from diverse experiments within cell types, related experiments involving 147 different cell types, and all ENCODE data with other resources, such as candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained regions. Together, these efforts reveal important features about the organization and function of the human genome, summarized below.

- The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.
- Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.
- Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.
- It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.
- Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.
- Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.

ENCODE data production and initial analyses

Since 2007, ENCODE has developed methods and performed a large number of sequence-based studies to map functional elements across the human genome[3]. The elements mapped (and approaches used) include RNA transcribed regions (RNA-seq, CAGE, RNA-PET and manual annotation), protein-coding regions (mass spectrometry), transcription-factor-binding sites (ChIP-seq and DNase-seq), chromatin structure (DNase-seq, FAIRE-seq, histone ChIP-seq and MNase-seq), and DNA methylation sites (RRBS assay) (Box 1 lists methods and abbreviations; Supplementary Table 1, section P, details production statistics)[3]. To compare and integrate results across the different laboratories, data production efforts focused on two selected

*Lists of participants and their affiliations appear at the end of the paper.

BOX 1

ENCODE abbreviations

RNA-seq. Isolation of RNA sequences, often with different purification techniques to isolate different fractions of RNA followed by high-throughput sequencing.
CAGE. Capture of the methylated cap at the 5′ end of RNA, followed by high-throughput sequencing of a small tag adjacent to the 5′ methylated caps. 5′ methylated caps are formed at the initiation of transcription, although other mechanisms also methylate 5′ ends of RNA.
RNA-PET. Simultaneous capture of RNAs with both a 5′ methyl cap and a poly(A) tail, which is indicative of a full-length RNA. This is then followed by sequencing a short tag from each end by high-throughput sequencing.
ChIP-seq. Chromatin immunoprecipitation followed by sequencing. Specific regions of crosslinked chromatin, which is genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then subjected to high-throughput sequencing to determine the regions in the genome most often bound by the protein to which the antibody was directed. Most often used are antibodies to any chromatin-associated epitope, including transcription factors, chromatin binding proteins and specific chemical modifications on histone proteins.
DNase-seq. Adaption of established regulatory sequence assay to modern techniques. The DNase I enzyme will preferentially cut live chromatin preparations at sites where nearby there are specific (non-histone) proteins. The resulting cut points are then sequenced using high-throughput sequencing to determine those sites 'hypersensitive' to DNase I, corresponding to open chromatin.
FAIRE-seq. Formaldehyde assisted isolation of regulatory elements. FAIRE isolates nucleosome-depleted genomic regions by exploiting the difference in crosslinking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low). FAIRE consists of crosslinking, phenol extraction, and sequencing the DNA fragments in the aqueous phase.
RRBS. Reduced representation bisulphite sequencing. Bisulphite treatment of DNA sequence converts unmethylated cytosines to uracil. To focus the assay and save costs, specific restriction enzymes that cut around CpG dinucleotides can reduce the genome to a portion specifically enriched in CpGs. This enriched sample is then sequenced to determine the methylation status of individual cytosines quantitatively.
Tier 1. Tier 1 cell types were the highest-priority set and comprised three widely studied cell lines: K562 erythroleukaemia cells; GM12878, a B-lymphoblastoid cell line that is also part of the 1000 Genomes project (http://1000genomes.org)[55]; and the H1 embryonic stem cell (H1 hESC) line.
Tier 2. The second-priority set of cell types in the ENCODE project which included HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells and primary (non-transformed) human umbilical vein endothelial cells (HUVECs).
Tier 3. Any other ENCODE cell types not in tier 1 or tier 2.

sets of cell lines, designated 'tier 1' and 'tier 2' (Box 1). To capture a broader spectrum of biological diversity, selected assays were also executed on a third tier comprising more than 100 cell types including primary cells. All data and protocol descriptions are available at http://www.encodeproject.org/, and a User's Guide including details of cell-type choice and limitations was published recently[3].

Integration methodology

For consistency, data were generated and processed using standardized guidelines, and for some assays, new quality-control measures were designed (see refs 3, 12 and http://encodeproject.org/ENCODE/dataStandards.html; A. Kundaje, personal communication). Uniform data-processing methods were developed for each assay (see Supplementary Information; A. Kundaje, personal communication), and most assay results can be represented both as signal information (a per-base estimate across the genome) and as discrete elements (regions computationally identified as enriched for signal). Extensive processing pipelines were developed to generate each representation (M. M. Hoffman *et al.*, manuscript in preparation and A. Kundaje, personal communication). In addition, we developed the irreproducible discovery rate (IDR)[13] measure to provide a robust and conservative estimate of the threshold where two ranked lists of results from biological replicates no longer agree (that is, are irreproducible), and we applied this to defining sets of discrete elements. We identified, and excluded from most analyses, regions yielding untrustworthy signals likely to be artefactual (for example, multicopy regions). Together, these regions comprise 0.39% of the genome (see Supplementary Information). The poster accompanying this issue represents different ENCODE-identified elements and their genome coverage.

Transcribed and protein-coding regions

We used manual and automated annotation to produce a comprehensive catalogue of human protein-coding and non-coding RNAs as well as pseudogenes, referred to as the GENCODE reference gene set[14,15] (Supplementary Table 1, section U). This includes 20,687 protein-coding genes (GENCODE annotation, v7) with, on average, 6.3 alternatively spliced transcripts (3.9 different protein-coding transcripts) per locus. In total, GENCODE-annotated exons of protein-coding genes cover 2.94% of the genome or 1.22% for protein-coding exons. Protein-coding genes span 33.45% from the outermost start to stop codons, or 39.54% from promoter to poly(A) site. Analysis of mass spectrometry data from K562 and GM12878 cell lines yielded 57 confidently identified unique peptide sequences in intergenic regions relative to GENCODE annotation. Taken together with evidence of pervasive genome transcription[16], these data indicate that additional protein-coding genes remain to be found.

In addition, we annotated 8,801 automatically derived small RNAs and 9,640 manually curated long non-coding RNA (lncRNA) loci[17]. Comparing lncRNAs to other ENCODE data indicates that lncRNAs are generated through a pathway similar to that for protein-coding genes[17]. The GENCODE project also annotated 11,224 pseudogenes, of which 863 were transcribed and associated with active chromatin[18].

RNA

We sequenced RNA[16] from different cell lines and multiple subcellular fractions to develop an extensive RNA expression catalogue. Using a conservative threshold to identify regions of RNA activity, 62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons. Of these bases, only 5.5% are explained by GENCODE exons. Most transcribed bases are within or overlapping annotated gene boundaries (that is, intronic), and only 31% of bases in sequenced transcripts were intergenic[16].

We used CAGE-seq (5′ cap-targeted RNA isolation and sequencing) to identify 62,403 transcription start sites (TSSs) at high confidence (IDR of 0.01) in tier 1 and 2 cell types. Of these, 27,362 (44%) are within 100 base pairs (bp) of the 5′ end of a GENCODE-annotated transcript or previously reported full-length messenger RNA. The remaining regions predominantly lie across exons and 3′ untranslated regions (UTRs), and some exhibit cell-type-restricted expression; these may represent the start sites of novel, cell-type-specific transcripts.

Finally, we saw a significant proportion of coding and non-coding transcripts processed into steady-state stable RNAs shorter than 200 nucleotides. These precursors include transfer RNA, microRNA, small nuclear RNA and small nucleolar RNA (tRNA, miRNA, snRNA and snoRNA, respectively) and the 5′ termini of these processed products align with the capped 5′ end tags[16].

Table 1 | Summary of transcription factor classes analysed in ENCODE

Acronym	Description	Factors analysed
ChromRem	ATP-dependent chromatin complexes	5
DNARep	DNA repair	3
HISase	Histone acetylation, deacetylation or methylation complexes	8
Other	Cyclin kinase associated with transcription	1
Pol2	Pol II subunit	1 (2 forms)
Pol3	Pol III-associated	6
TFNS	General Pol II-associated factor, not site-specific	8
TFSS	Pol II transcription factor with sequence-specific DNA binding	87

Protein bound regions

To identify regulatory regions directly, we mapped the binding locations of 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types using ChIP-seq (Table 1, Supplementary Table 1, section N, and ref. 19); 87 (73%) were sequence-specific transcription factors. Overall, 636,336 binding regions covering 231 megabases (Mb; 8.1%) of the genome are enriched for regions bound by DNA-binding proteins across all cell types. We assessed each protein-binding site for enrichment of known DNA-binding motifs and the presence of novel motifs. Overall, 86% of the DNA segments occupied by sequence-specific transcription factors contained a strong DNA-binding motif, and in most (55%) cases the known motif was most enriched (P. Kheradpour and M. Kellis, manuscript in preparation).

Protein-binding regions lacking high or moderate affinity cognate recognition sites have 21% lower median scores by rank than regions with recognition sequences (Wilcoxon rank sum P value $<10^{-16}$). Eighty-two per cent of the low-signal regions have high-affinity recognition sequences for other factors. In addition, when ChIP-seq peaks are ranked by their concordance with their known recognition sequence, the median DNase I accessibility is twofold higher in the bottom 20% of peaks than in the upper 80% (genome structure correction (GSC)[20] P value $<10^{-16}$), consistent with previous observations[21–24]. We speculate that low signal regions are either lower-affinity sites[21] or indirect transcription-factor target regions associated through interactions with other factors (see also refs 25, 26).

We organized all the information associated with each transcription factor—including the ChIP-seq peaks, discovered motifs and associated histone modification patterns—in FactorBook (http://www.factorbook.org; ref. 26), a public resource that will be updated as the project proceeds.

DNase I hypersensitive sites and footprints

Chromatin accessibility characterized by DNase I hypersensitivity is the hallmark of regulatory DNA regions[27,28]. We mapped 2.89 million unique, non-overlapping DNase I hypersensitive sites (DHSs) by DNase-seq in 125 cell types, the overwhelming majority of which lie distal to TSSs[29]. We also mapped 4.8 million sites across 25 cell types that displayed reduced nucleosomal crosslinking by FAIRE, many of which coincide with DHSs. In addition, we used micrococcal nuclease to map nucleosome occupancy in GM12878 and K562 cells[30].

In tier 1 and tier 2 cell types, we identified a mean of 205,109 DHSs per cell type (at false discovery rate (FDR) 1%), encompassing an average of 1.0% of the genomic sequence in each cell type, and 3.9% in aggregate. On average, 98.5% of the occupancy sites of transcription factors mapped by ENCODE ChIP-seq (and, collectively, 94.4% of all 1.1 million transcription factor ChIP-seq peaks in K562 cells) lie within accessible chromatin defined by DNase I hotspots[29]. However, a small number of factors, most prominently heterochromatin-bound repressive complexes (for example, the TRIM28–SETDB1–ZNF274 complex[31,32] encoded by the *TRIM28*, *SETDB1* and *ZNF274* genes), seem to occupy a significant fraction of nucleosomal sites.

Using genomic DNase I footprinting[33,34] on 41 cell types we identified 8.4 million distinct DNase I footprints (FDR 1%)[25]. Our *de novo* motif discovery on DNase I footprints recovered ~90% of known transcription factor motifs, together with hundreds of novel evolutionarily conserved motifs, many displaying highly cell-selective occupancy patterns similar to major developmental and tissue-specific regulators.

Regions of histone modification

We assayed chromosomal locations for up to 12 histone modifications and variants in 46 cell types, including a complete matrix of eight modifications across tier 1 and tier 2. Because modification states may span multiple nucleosomes, which themselves can vary in position across cell populations, we used a continuous signal measure of histone modifications in downstream analysis, rather than calling regions (M. M. Hoffman *et al.*, manuscript in preparation; see http://code.google.com/p/align2rawsignal/). For the strongest, 'peak-like' histone modifications, we used MACS[35] to characterize enriched sites. Table 2 describes the different histone modifications, their peak characteristics, and a summary of their known roles (reviewed in refs 36–39).

Our data show that global patterns of modification are highly variable across cell types, in accordance with changes in transcriptional activity. Consistent with previous studies[40,41], we find that integration of the different histone modification information can be used systematically to assign functional attributes to genomic regions (see below).

DNA methylation

Methylation of cytosine, usually at CpG dinucleotides, is involved in epigenetic regulation of gene expression. Promoter methylation is typically associated with repression, whereas genic methylation correlates with transcriptional activity[42]. We used reduced representation bisulphite sequencing (RRBS) to profile DNA methylation quantitatively for an average of 1.2 million CpGs in each of 82 cell lines and tissues (8.6% of non-repetitive genomic CpGs), including CpGs in intergenic regions, proximal promoters and intragenic regions (gene bodies)[43], although it should be noted that the RRBS method preferentially targets CpG-rich islands. We found that 96% of CpGs exhibited differential methylation in at least one cell type or tissue

Table 2 | Summary of ENCODE histone modifications and variants

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

assayed (K. Varley *et al.*, personal communication), and levels of DNA methylation correlated with chromatin accessibility. The most variably methylated CpGs are found more often in gene bodies and intergenic regions, rather than in promoters and upstream regulatory regions. In addition, we identified an unexpected correspondence between unmethylated genic CpG islands and binding by P300, a histone acetyltransferase linked to enhancer activity[44].

Because RRBS is a sequence-based assay with single-base resolution, we were able to identify CpGs with allele-specific methylation consistent with genomic imprinting, and determined that these loci exhibit aberrant methylation in cancer cell lines (K. Varley *et al.*, personal communication). Furthermore, we detected reproducible cytosine methylation outside CpG dinucleotides in adult tissues[45], providing further support that this non-canonical methylation event may have important roles in human biology (K. Varley *et al.*, personal communication).

Chromosome-interacting regions

Physical interaction between distinct chromosome regions that can be separated by hundreds of kilobases is thought to be important in the regulation of gene expression[46]. We used two complementary chromosome conformation capture (3C)-based technologies to probe these long-range physical interactions.

A 3C-carbon copy (5C) approach[47,48] provided unbiased detection of long-range interactions with TSSs in a targeted 1% of the genome (the 44 ENCODE pilot regions) in four cell types (GM12878, K562, HeLa-S3 and H1 hESC)[49]. We discovered hundreds of statistically significant long-range interactions in each cell type after accounting for chromatin polymer behaviour and experimental variation. Pairs of interacting loci showed strong correlation between the gene expression level of the TSS and the presence of specific functional element classes such as enhancers. The average number of distal elements interacting with a TSS was 3.9, and the average number of TSSs interacting with a distal element was 2.5, indicating a complex network of interconnected chromatin. Such interwoven long-range architecture was also uncovered genome-wide using chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)[50] applied to identify interactions in chromatin enriched by RNA polymerase II (Pol II) ChIP from five cell types[51]. In K562 cells, we identified 127,417 promoter-centred chromatin interactions using ChIA-PET, 98% of which were intra-chromosomal. Whereas promoter regions of 2,324 genes were involved in 'single-gene' enhancer–promoter interactions, those of 19,813 genes were involved in 'multi-gene' interaction complexes spanning up to several megabases, including promoter–promoter and enhancer–promoter interactions[51].

These analyses portray a complex landscape of long-range gene–element connectivity across ranges of hundreds of kilobases to several megabases, including interactions among unrelated genes (Supplementary Fig. 1, section Y). Furthermore, in the 5C results, 50–60% of long-range interactions occurred in only one of the four cell lines, indicative of a high degree of tissue specificity for gene–element connectivity[49].

Summary of ENCODE-identified elements

Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element (detailed in Supplementary Table 1, section Q). The broadest element class represents the different RNA types, covering 62% of the genome (although the majority is inside of introns or near genes). Regions highly enriched for histone modifications form the next largest class (56.1%). Excluding RNA elements and broad histone elements, 44.2% of the genome is covered. Smaller proportions of the genome are occupied by regions of open chromatin (15.2%) or sites of transcription factor binding (8.1%), with 19.4% covered by at least one DHS or transcription factor ChIP-seq peak across all cell lines. Using our most conservative assessment, 8.5% of bases are covered by either a transcription-factor-binding-site motif (4.6%) or a DHS footprint (5.7%). This, however, is still about 4.5-fold higher than the amount of protein-coding exons, and about twofold higher than the estimated amount of pan-mammalian constraint.

Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages, these proportions must be underestimates of the total amount of functional bases. However, many assays were performed on more than one cell type, allowing assessment of the rate of discovery of new elements. For both DHSs and CTCF-bound sites, the number of new elements initially increases rapidly with a steep gradient for the saturation curve and then slows with increasing number of cell types (Supplementary Figs 1 and 2, section R). With the current data, at the flattest part of the saturation curve each new cell type adds, on average, 9,500 DHS elements (across 106 cell types) and 500 CTCF-binding elements (across 49 cell types), representing 0.45% of the total element number. We modelled saturation for the DHSs and CTCF-binding sites using a Weibull distribution ($r^2 > 0.999$) and predict saturation at approximately 4.1 million (standard error (s.e.) = 108,000) and 185,100 (s.e. = 18,020) sites, respectively, indicating that we have discovered around half of the estimated total DHSs. These estimates represent a lower bound, but reinforce the observation that there is more non-coding functional DNA than either coding sequence or mammalian evolutionarily constrained bases.

The impact of selection on functional elements

From comparative genomic studies, at least 3–8% of bases are under purifying (negative) selection[4–11], indicating that these bases may potentially be functional. We previously found that 60% of mammalian evolutionarily constrained bases were annotated in the ENCODE pilot project, but also observed that many functional elements lacked evidence of constraint[2], a conclusion substantiated by others[52–54]. The diversity and genome-wide occurrence of functional elements now identified provides an unprecedented opportunity to examine further the forces of negative selection on human functional sequences.

We examined negative selection using two measures that highlight different periods of selection in the human genome. The first measure, inter-species, pan-mammalian constraint (GERP-based scores; 24 mammals[8]), addresses selection during mammalian evolution. The second measure is intra-species constraint estimated from the numbers of variants discovered in human populations using data from the 1000 Genomes project[55], and covers selection over human evolution. In Fig. 1, we plot both these measures of constraint for different classes of identified functional elements, excluding features overlapping exons and promoters that are known to be constrained. Each graph also shows genomic background levels and measures of coding-gene constraint for comparison. Because we plot human population diversity on an inverted scale, elements that are more constrained by negative selection will tend to lie in the upper and right-hand regions of the plot.

For DNase I elements (Fig. 1b) and bound motifs (Fig. 1c), most sets of elements show enrichment in pan-mammalian constraint and decreased human population diversity, although for some cell types the DNase I sites do not seem overall to be subject to pan-mammalian constraint. Bound transcription factor motifs have a natural control from the set of transcription factor motifs with equal sequence potential for binding but without binding evidence from ChIP-seq experiments—in all cases, the bound motifs show both more mammalian constraint and higher suppression of human diversity.

Consistent with previous findings, we do not observe genome-wide evidence for pan-mammalian selection of novel RNA sequences (Fig. 1d). There are also a large number of elements without mammalian constraint, between 17% and 90% for transcription-factor-binding regions as well as DHSs and FAIRE regions. Previous studies could not determine whether these sequences are either biochemically active, but with little overall impact on the organism, or under lineage-specific selection. By isolating sequences preferentially inserted into

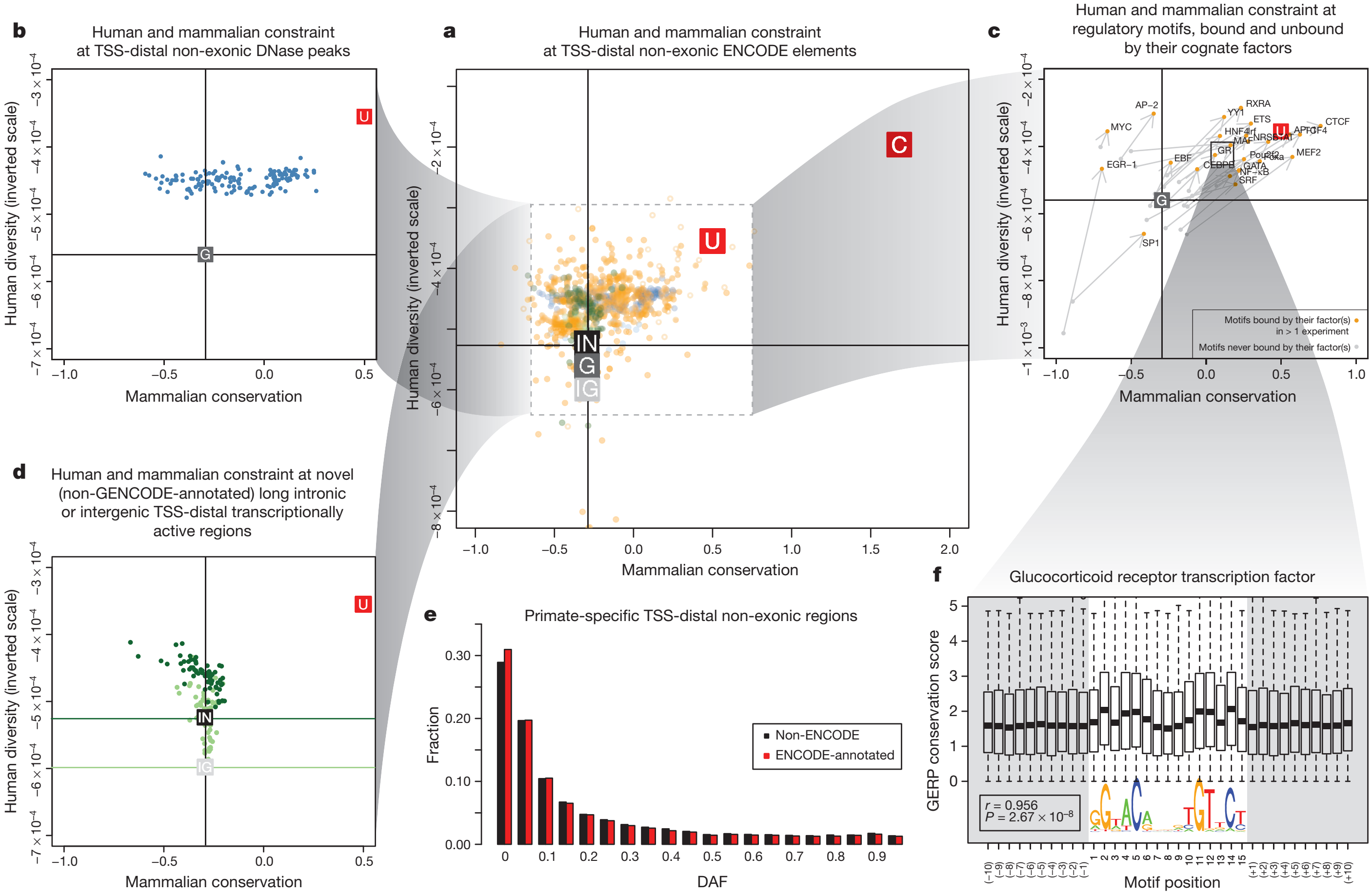


Figure 1 | Impact of selection on ENCODE functional elements in mammals and human populations. **a**, Levels of pan-mammalian constraint (mean GERP score; 24 mammals[8], x axis) compared to diversity, a measure of negative selection in the human population (mean expected heterozygosity, inverted scale, y axis) for ENCODE data sets. Each point is an average for a single data set. The top-right corners have the strongest evolutionary constraint and lowest diversity. Coding (C), UTR (U), genomic (G), intergenic (IG) and intronic (IN) averages are shown as filled squares. In each case the vertical and horizontal cross hairs show representative levels for the neutral expectation for mammalian conservation and human population diversity, respectively. The spread over all non-exonic ENCODE elements greater than 2.5 kb from TSSs is shown. The inner dashed box indicates that parts of the plot have been magnified for the surrounding outer panels, although the scales in the outer plots provide the exact regions and dimensions magnified. The spread for DHS sites (**b**) and RNA elements (**d**) is shown in the plots on the left. RNA elements are either long novel intronic (dark green) or long intergenic (light green) RNAs. The horizontal cross hairs are colour-coded to the relevant data set in **d**. **c**, Spread of transcription factor motif instances either in regions bound by the transcription factor (orange points) or in the corresponding unbound motif matches in grey, with bound and unbound points connected with an arrow in each case showing that bound sites are generally more constrained and less diverse. **e**, Derived allele frequency spectrum for primate-specific elements, with variations outside ENCODE elements in black and variations covered by ENCODE elements in red. The increase in low-frequency alleles compared to background is indicative of negative selection occurring in the set of variants annotated by the ENCODE data. **f**, Aggregation of mammalian constraint scores over the glucocorticoid receptor (GR) transcription factor motif in bound sites, showing the expected correlation with the information content of bases in the motif. An interactive version of this figure is available in the online version of the paper.

the primate lineage, which is only feasible given the genome-wide scale of this data, we are able to examine this issue specifically. Most primate-specific sequence is due to retrotransposon activity, but an appreciable proportion is non-repetitive primate-specific sequence. Of 104,343,413 primate-specific bases (excluding repetitive elements), 67,769,372 (65%) are found within ENCODE-identified elements. Examination of 227,688 variants segregating in these primate-specific regions revealed that all classes of elements (RNA and regulatory) show depressed derived allele frequencies, consistent with recent negative selection occurring in at least some of these regions (Fig. 1e). An alternative approach examining sequences that are not clearly under pan-mammalian constraint showed a similar result (L. Ward and M. Kellis, manuscript submitted). This indicates that an appreciable proportion of the unconstrained elements are lineage-specific elements required for organismal function, consistent with long-standing views of recent evolution[56], and the remainder are probably 'neutral' elements[2] that are not currently under selection but may still affect cellular or larger scale phenotypes without an effect on fitness.

The binding patterns of transcription factors are not uniform, and we can correlate both inter- and intra-species measures of negative selection with the overall information content of motif positions. The selection on some motif positions is as high as protein-coding exons (Fig. 1f; L. Ward and M. Kellis, manuscript submitted). These aggregate measures across motifs show that the binding preferences found in the population of sites are also relevant to the per-site behaviour. By developing a per-site metric of population effect on bound motifs, we found that highly constrained bound instances across mammals are able to buffer the impact of individual variation[57].

ENCODE data integration with known genomic features

Promoter-anchored integration

Many of the ENCODE assays directly or indirectly provide information about the action of promoters. Focusing on the TSSs of protein-coding transcripts, we investigated the relationships between different ENCODE assays, in particular testing the hypothesis that RNA expression (output) can be effectively predicted from patterns of chromatin modification or transcription factor binding (input). Consistent with previous reports[58], we observe two relatively distinct types of promoter: (1) broad, mainly (C+G)-rich, TATA-less promoters; and (2) narrow, TATA-box-containing promoters. These promoters have distinct patterns of histone modifications, and transcription-factor-binding sites are selectively enriched in each class (Supplementary Fig. 1, section Z).

We developed predictive models to explore the interaction between histone modifications and measures of transcription at promoters, distinguishing between modifications known to be added as a consequence of transcription (such as H3K36me3 and H3K79me2) and other categories of histone marks[59]. In our analyses, the best models had two components: an initial classification component (on/off) and a second quantitative model component. Our models showed that activating acetylation marks (H3K27ac and H3K9ac) are roughly as informative as activating methylation marks (H3K4me3 and H3K4me2) (Fig. 2a). Although repressive marks, such as H3K27me3 or H3K9me3, show negative correlation both individually and in the model, removing these marks produces only a small reduction in model performance. However, for a subset of promoters in each cell line, repressive histone marks (H3K27me3 or H3K9me3) must be used to predict their expression accurately. We also examined the interplay between the H3K79me2 and H3K36me3 marks, both of which mark gene bodies, probably reflecting recruitment of modification enzymes by polymerase isoforms. As described previously, H3K79me2 occurs preferentially at the 5′ ends of gene bodies and H3K36me3 occurs more 3′, and our analyses support the previous model in which the H3K79me2 to H3K36me3 transition occurs at the first 3′ splice site[60].

Few previous studies have attempted to build qualitative or quantitative models of transcription genome-wide from transcription factor levels because of the paucity of documented transcription-factor-binding regions and the lack of coordination around a single cell line. We thus examined the predictive capacity of transcription-factor-binding signals for the expression levels of promoters (Fig. 2b).

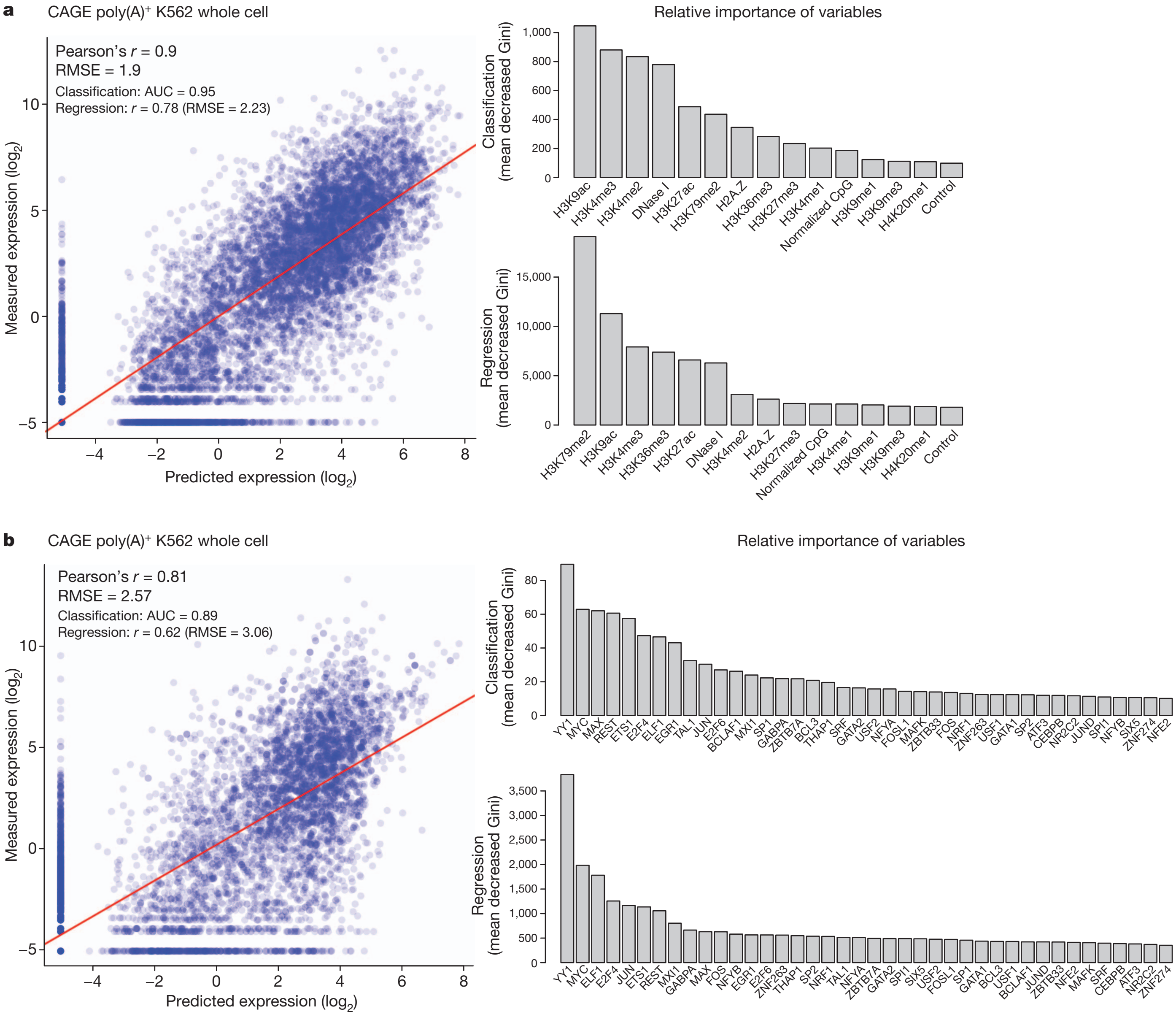


Figure 2 | Modelling transcription levels from histone modification and transcription-factor-binding patterns. a, **b**, Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each case the scatter plot shows the output of the correlation models (*x* axis) compared to observed values (*y* axis). The bar graphs show the most important histone modifications (**a**) or transcription factors (**b**) in both the initial classification phase (top bar graph) or the quantitative regression phase (bottom bar graph), with larger values indicating increasing importance of the variable in the model. Further analysis of other cell lines and RNA measurement types is reported elsewhere[59,79]. AUC, area under curve; Gini, Gini coefficient; RMSE, root mean square error.

In contrast to the profiles of histone modifications, most transcription factors show enriched binding signals in a narrow DNA region near the TSS, with relatively higher binding signals in promoters with higher CpG content. Most of this correlation could be recapitulated by looking at the aggregate binding of transcription factors without specific transcription factor terms. Together, these correlation models indicate both that a limited set of chromatin marks are sufficient to 'explain' transcription and that a variety of transcription factors might have broad roles in general transcription levels across many genes. It is important to note that this is an inherently observational study of correlation patterns, and is consistent with a variety of mechanistic models with different causal links between the chromatin, transcription factor and RNA assays. However, it does indicate that there is enough information present at the promoter regions of genes to explain most of the variation in RNA expression.

We developed predictive models similar to those used to model transcriptional activity to explore the relationship between levels of histone modification and inclusion of exons in alternately spliced transcripts. Even accounting for expression level, H3K36me3 has a positive contribution to exon inclusion, whereas H3K79me2 has a negative contribution (H. Tilgner *et al.*, manuscript in preparation). By monitoring the RNA populations in the subcellular fractions of K562 cells, we found that essentially all splicing is co-transcriptional[61], further supporting a link between chromatin structure and splicing.

Transcription-factor-binding site-anchored integration

Transcription-factor-binding sites provide a natural focus around which to explore chromatin properties. Transcription factors are often multifunctional and can bind a variety of genomic loci with different combinations and patterns of chromatin marks and nucleosome organization. Hence, rather than averaging chromatin mark profiles across all binding sites of a transcription factor, we developed a clustering procedure, termed the Clustered Aggregation Tool (CAGT), to identify subsets of binding sites sharing similar but distinct patterns of chromatin mark signal magnitude, shape and hidden directionality[30]. For example, the average profile of the repressive histone mark H3K27me3 over all 55,782 CTCF-binding sites in H1 hESCs shows poor signal enrichment (Fig. 3a). However, after grouping profiles by signal magnitude we found a subset of 9,840 (17.6%) CTCF-binding sites that exhibit significant flanking H3K27me3 signal. Shape and orientation analysis further revealed that the predominant signal profile for H3K27me3 around CTCF peak summits is asymmetric, consistent with a boundary role for some CTCF sites between active and polycomb-silenced domains. Further examples are provided in Supplementary Figs 5 and 6 of section E. For TAF1, predominantly found near TSSs, the asymmetric sites are orientated with the direction of transcription. However, for distal sites, such as those bound by GATA1 and CTCF, we also observed a high proportion of asymmetric histone patterns, although independent of motif directionality. In fact, all transcription-factor-binding data sets in all cell lines show predominantly asymmetric patterns (asymmetry ratio >0.6) for all chromatin marks but not for DNase I signal (Fig. 3b). This indicates that most transcription-factor-bound chromatin events correlate with structured, directional patterns of histone modifications, and that promoter directionality is not the only source of orientation at these sites.

We also examined nucleosome occupancy relative to the symmetry properties of chromatin marks around transcription-factor-binding sites. Around TSSs, there is usually strong asymmetric nucleosome occupancy, often accounting for most of the histone modification signal (for instance, see Supplementary Fig. 4, section E). However, away from TSSs, there is far less concordance. For example, CTCF-binding sites typically show arrays of well-positioned nucleosomes on either side of the peak summit (Supplementary Fig. 1, section E)[62]. Where the flanking chromatin mark signal is high, the signals are often asymmetric, indicating differential marking with histone modifications (Supplementary Figs 2 and 3, section E). Thus, we confirm on a genome-wide scale that transcription factors can form barriers around which nucleosomes and histone modifications are arranged in a variety of configurations[62–65]. This is explored in further detail in refs 25, 26 and 30.

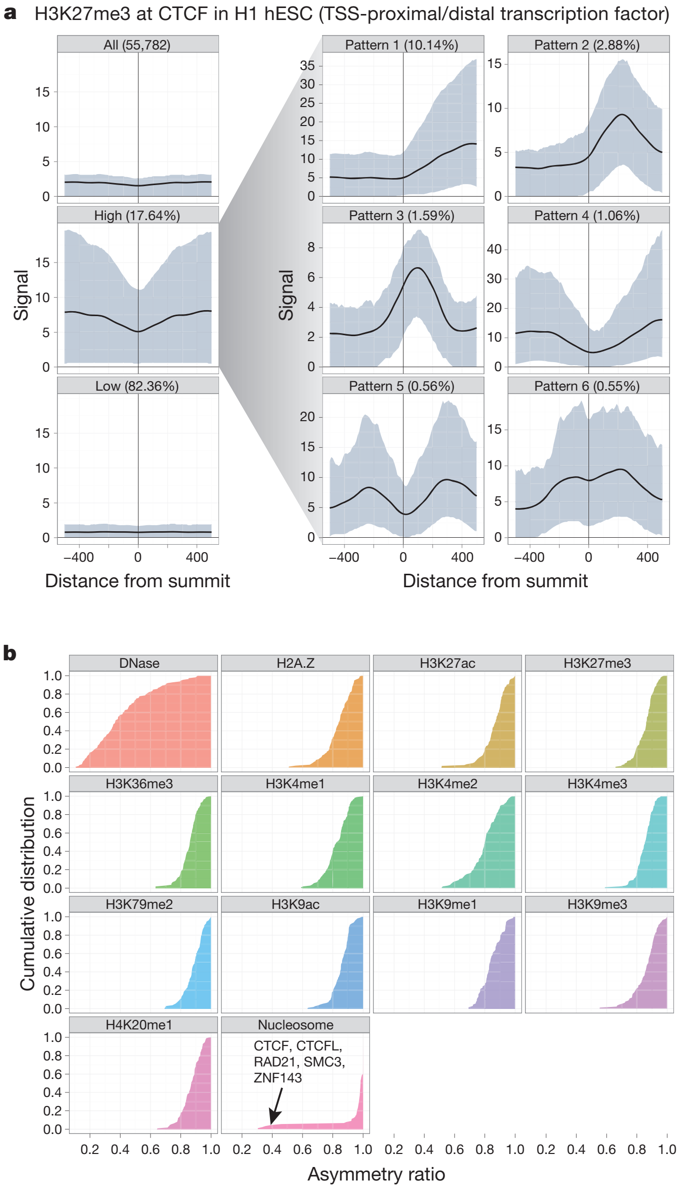


Figure 3 | Patterns and asymmetry of chromatin modification at transcription-factor-binding sites. a, Results of clustered aggregation of H3K27me3 modification signal around CTCF-binding sites (a multifunctional protein involved with chromatin structure). The first three plots (left column) show the signal behaviour of the histone modification over all sites (top) and then split into the high and low signal components. The solid lines show the mean signal distribution by relative position with the blue shaded area delimiting the tenth and ninetieth percentile range. The high signal component is then decomposed further into six different shape classes on the right (see ref. 30 for details). The shape decomposition process is strand aware. **b**, Summary of shape asymmetry for DNase I, nucleosome and histone modification signals by plotting an asymmetry ratio for each signal over all transcription-factor-binding sites. All histone modifications measured in this study show predominantly asymmetric patterns at transcription-factor-binding sites. An interactive version of this figure is available in the online version of the paper.

Transcription factor co-associations

Transcription-factor-binding regions are nonrandomly distributed across the genome, with respect to both other features (for example, promoters) and other transcription-factor-binding regions. Within the

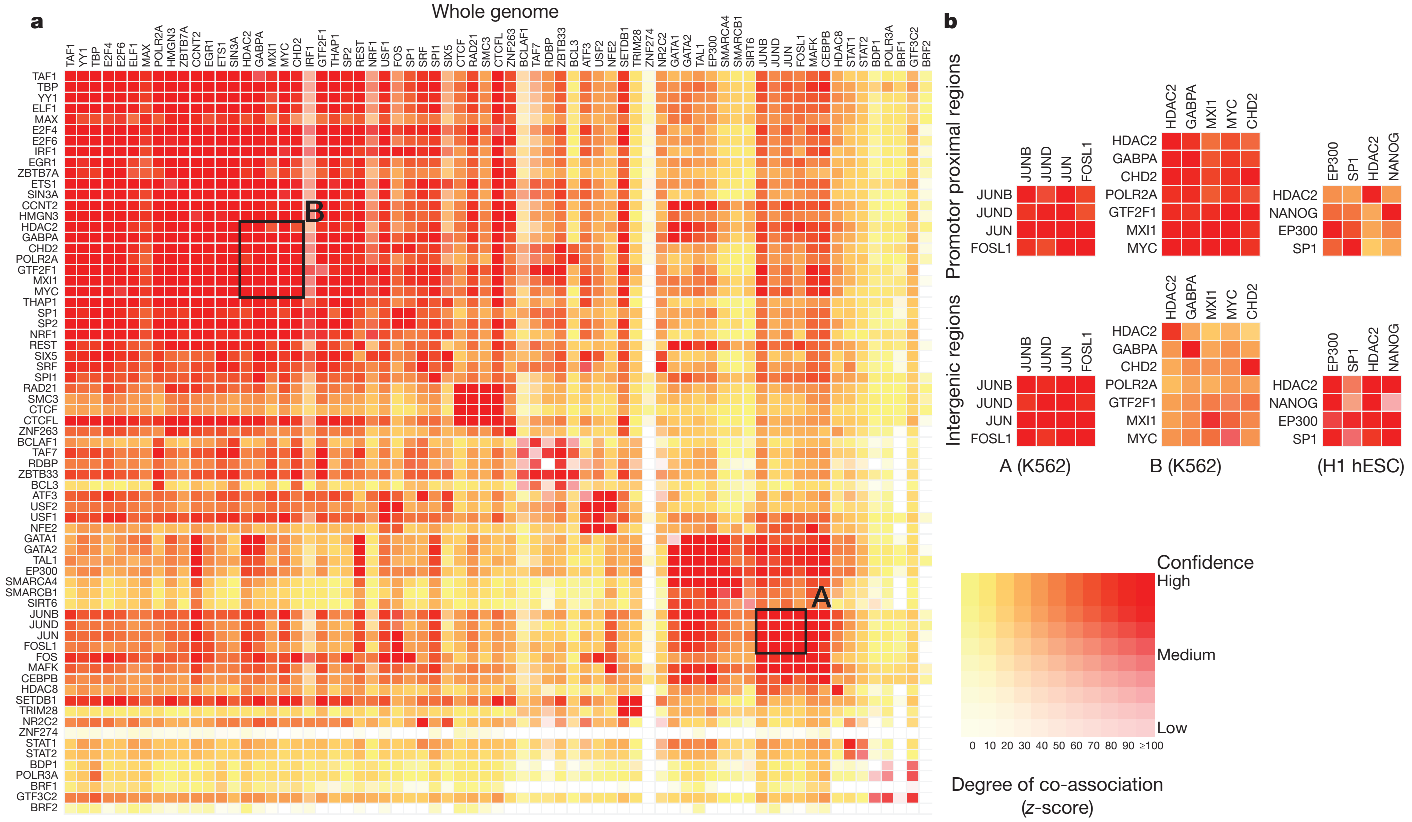


Figure 4 | Co-association between transcription factors. a, Significant co-associations of transcription factor pairs using the GSC statistic across the entire genome in K562 cells. The colour strength represents the extent of association (from red (strongest), orange, to yellow (weakest)), whereas the depth of colour represents the fit to the GSC[20] model (where white indicates that the statistical model is not appropriate) as indicated by the key. Most transcription factors have a nonrandom association to other transcription factors, and these associations are dependent on the genomic context, meaning that once the genome is separated into promoter proximal and distal regions, the overall levels of co-association decrease, but more specific relationships are uncovered. **b**, Three classes of behaviour are shown. The first column shows a set of associations for which strength is independent of location in promoter and distal regions, whereas the second column shows a set of transcription factors that have stronger associations in promoter-proximal regions. Both of these examples are from data in K562 cells and are highlighted on the genome-wide co-association matrix (**a**) by the labelled boxes A and B, respectively. The third column shows a set of transcription factors that show stronger association in distal regions (in the H1 hESC line). An interactive version of this figure is available in the online version of the paper.

tier 1 and 2 cell lines, we found 3,307 pairs of statistically co-associated factors ($P < 1 \times 10^{-16}$, GSC) involving 114 out of a possible 117 factors (97%) (Fig. 4a). These include expected associations, such as Jun and Fos, and some less expected novel associations, such as TCF7L2 with HNF4-α and FOXA2 (ref. 66; a full listing is given in Supplementary Table 1, section F). When one considers promoter and intergenic

Table 3 | Summary of the combined state types

Label	Description	Details*	Colour
CTCF	CTCF-enriched element	Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3; CTCF is known to recruit the cohesin complex.	Turquoise
E	Predicted enhancer	Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act at enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be *cis*-regulatory regions. Enriched for sites for the proteins encoded by *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A)− fraction.	Orange
PF	Predicted promoter flanking region	Regions that generally surround TSS segments (see below).	Light red
R	Predicted repressed or low-activity region	This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by *REST* and some other factors (for example, proteins encoded by *BRF2*, *CEBPB*, *MAFK*, *TRIM28*, *ZNF274* and *SETDB1* genes in K562 cells).	Grey
TSS	Predicted promoter region including TSS	Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments.	Bright red
T	Predicted transcribed region	Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A)$^{+}$ RNA, especially cytoplasmic.	Dark green
WE	Predicted weak enhancer or open chromatin *cis*-regulatory element	Similar to the E state, but weaker signals and weaker enrichments.	Yellow

* Where specific enrichments or overlaps are identified, these are derived from analysis in GM12878 and/or K562 cells where the data for comparison is richest. The colours indicated are used in Figs 5 and 7 and in display of these tracks from the ENCODE data hub.

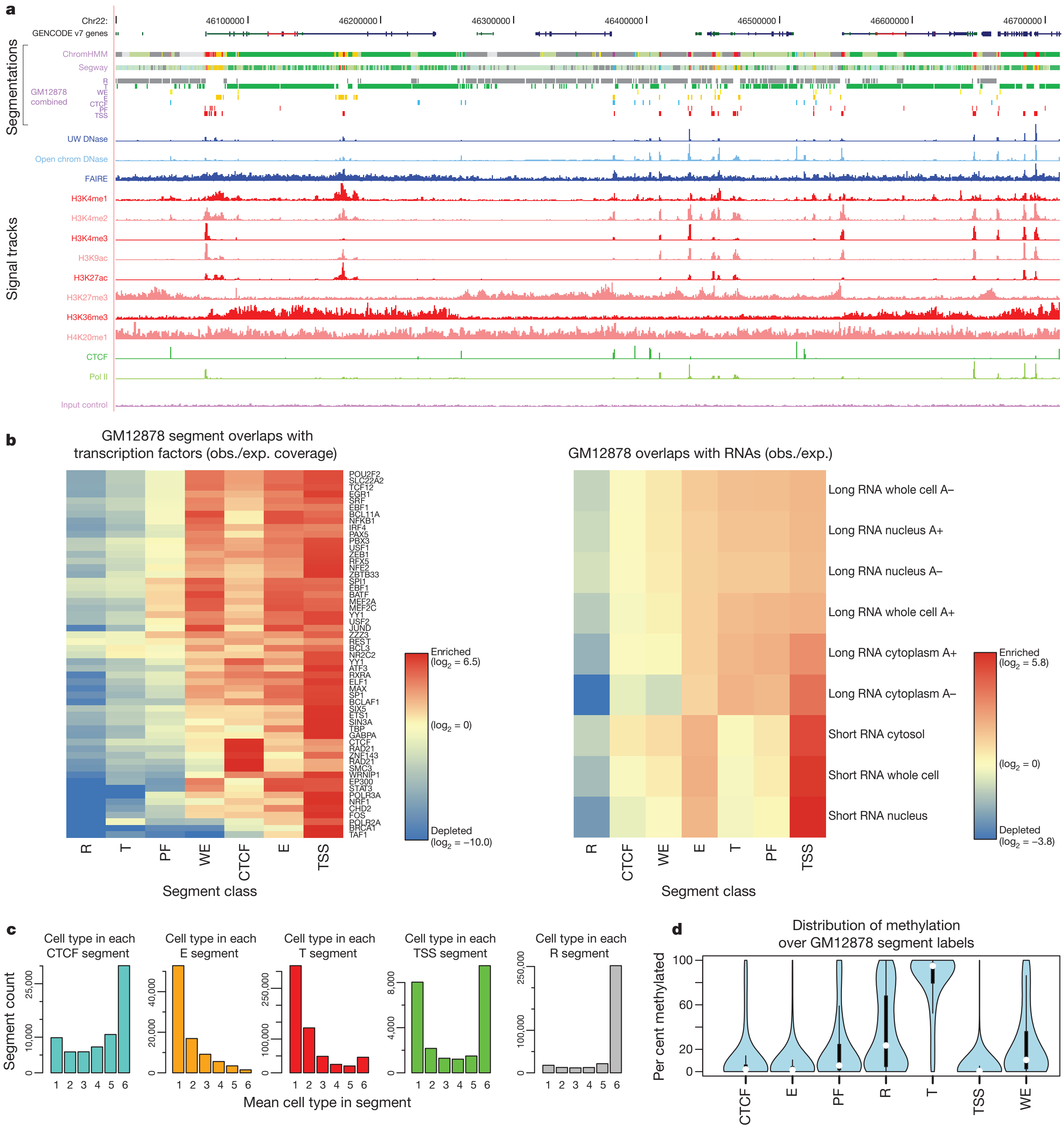


Figure 5 | Integration of ENCODE data by genome-wide segmentation. **a**, Illustrative region with the two segmentation methods (ChromHMM and Segway) in a dense view and the combined segmentation expanded to show each state in GM12878 cells, beneath a compressed view of the GENCODE gene annotations. Note that at this level of zoom and genome browser resolution, some segments appear to overlap although they do not. Segmentation classes are named and coloured according to the scheme in Table 3. Beneath the segmentations are shown each of the normalized signals that were used as the input data for the segmentations. Open chromatin signals from DNase-seq from the University of Washington group (UW DNase) or the ENCODE open chromatin group (Openchrom DNase) and FAIRE assays are shown in blue; signal from histone modification ChIP-seq in red; and transcription factor ChIP-seq signal for Pol II and CTCF in green. The mauve ChIP-seq control signal (input control) at the bottom was also included as an input to the segmentation. **b**, Association of selected transcription factor (left) and RNA (right) elements in the combined segmentation states (*x* axis) expressed as an observed/expected ratio (obs./exp.) for each combination of transcription factor or RNA element and segmentation class using the heat-map scale shown in the key besides each heat map. **c**, Variability of states between cell lines, showing the distribution of occurrences of the state in the six cell lines at specific genome locations: from unique to one cell line to ubiquitous in all six cell lines for five states (CTCF, E, T, TSS and R). **d**, Distribution of methylation level at individual sites from RRBS analysis in GM12878 cells across the different states, showing the expected hypomethylation at TSSs and hypermethylation of genes bodies (T state) and repressed (R) regions.

regions separately, this changes to 3,201 pairs (116 factors, 99%) for promoters and 1,564 pairs (108 factors, 92%) for intergenic regions, with some associations more specific to these genomic contexts (for example, the cluster of HDAC2, GABPA, CHD2, GTF2F1, MXI1 and MYC in promoter regions and SP1, EP300, HDAC2 and NANOG in intergenic regions (Fig. 4b)). These general and context-dependent associations lead to a network representation of the co-binding with many interesting properties, explored in refs 19, 25 and 26. In addition, we also identified a set of regions bound by multiple factors representing high occupancy of transcription factor (HOT) regions[67].

Genome-wide integration

To identify functional regions genome-wide, we next integrated elements independent of genomic landmarks using either discriminative training methods, where a subset of known elements of a particular class were used to train a model that was then used to discover more instances of this class, or using methods in which only data from ENCODE assays were used without explicit knowledge of any annotation.

For discriminative training, we used a three-step process to predict potential enhancers, described in Supplementary Information and ref. 67. Two alternative discriminative models converged on a set of ~13,000 putative enhancers in K562 cells[67]. In the second approach, two methodologically distinct unbiased approaches (see refs 40, 68 and M. M. Hoffman *et al.*, manuscript in preparation) converged on a concordant set of histone modification and chromatin-accessibility patterns that can be used to segment the genome in each of the tier 1 and tier 2 cell lines, although the individual loci in each state in each cell line are different. With the exception of RNA polymerase II and CTCF, the addition of transcription factor data did not substantially alter these patterns. At this stage, we deliberately excluded RNA and methylation assays, reserving these data as a means to validate the segmentations.

Our integration of the two segmentation methods (M. M. Hoffman *et al.*, manuscript in preparation) established a consensus set of seven major classes of genome states, described in Table 3. The standard view of active promoters, with a distinct core promoter region (TSS and PF states), leading to active gene bodies (T, transcribed state), is rediscovered in this model (Fig. 5a, b). There are three 'active' distal states. We tentatively labelled two as enhancers (predicted enhancers, E, and predicted weak enhancers, WE) due to their occurrence in regions of open chromatin with high H3K4me1, although they differ in the levels of marks such as H3K27ac, currently thought to distinguish active from inactive enhancers. The other active state (CTCF) has high CTCF binding and includes sequences that function as insulators in a transfection assay. The remaining repressed state (R) summarizes sequences split between different classes of actively repressed or inactive, quiescent chromatin. We found that the CTCF-binding-associated state is relatively invariant across cell types, with individual regions frequently occupying the CTCF state across all six cell types (Fig. 5c). Conversely, the E and T states have substantial cell-specific behaviour, whereas the TSS state has a bimodal behaviour with similar numbers of cell-invariant and cell-specific occurrences. It is important to note that the consensus summary classes do not capture all the detail discovered in the individual segmentations containing more states.

The distribution of RNA species across segments is quite distinct, indicating that underlying biological activities are captured in the segmentation. Polyadenylated RNA is heavily enriched in gene bodies. Around promoters, there are short RNA species previously identified as promoter-associated short RNAs (Fig. 5b)[16,69]. Similarly, DNA methylation shows marked distinctions between segments, recapitulating the known biology of predominantly unmethylated active promoters (TSS states) followed by methylated gene bodies[42] (T state, Fig. 5d). The two enhancer-enriched states show distinct patterns of DNA methylation, with the less active enhancer state (by H3K27ac/H3K4me1 levels) showing higher methylation. These states also have an excess of RNA elements without poly(A) tails and methyl-cap RNA, as assayed by CAGE sequences, compared to matched intergenic controls, indicating a specific transcriptional mode associated with active enhancers[70]. Transcription factors also showed distinct distributions across the segments (Fig. 5b). A striking pattern is the concentration of transcription factors in the TSS-associated state. The enhancers contain a different set of transcription factors. For example, in K562 cells, the E state is enriched for binding by the proteins encoded by the *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes. We tested a subset of these predicted enhancers in both mouse and fish transgenic models (examples in Fig. 6), with over half of the elements showing activity, often in the corresponding tissue type.

The segmentation provides a linear determination of functional state across the genome, but not an association of particular distal regions with genes. By using the variation of DNase I signal across cell lines, 39% of E (enhancer associated) states could be linked to a proposed regulated gene[29] concordant with physical proximity patterns determined by 5C[49] or ChIA-PET.

To provide a fine-grained regional classification, we turned to a self organizing map (SOM) to cluster genome segmentation regions based on their assay signal characteristics (Fig. 7). The segmentation regions were initially randomly assigned to a 1,350-state map in a two-dimensional toroidal space (Fig. 7a). This map can be visualized as a two-dimensional rectangular plane onto which the various signal distributions can be plotted. For instance, the rectangle at the bottom left of Fig. 7a shows the distribution of the genome in the initial randomized map. The SOM was then trained using the twelve different ChIP-seq and DNase-seq assays in the six cell types previously analysed in the large-scale segmentations (that is, over 72-dimensional space). After training, the SOM clustering was again visualized in two dimensions, now showing the organized distribution of genome segments (lower right of panel, Fig. 7a). Individual data sets associated with the genome segments in each SOM map unit (hexagonal cells) can then be visualized in the same framework to learn how each additional kind of data is distributed on the chromatin state map. Figure 7b shows CAGE/TSS expression data overlaid on the randomly initialized (left) and trained map (right) panels. In this way the trained SOM highlighted cell-type-specific TSS clusters (bottom panels of Fig. 7b), indicating that there are sets of tissue-specific TSSs that are distinguished from each other by subtle combinations of ENCODE

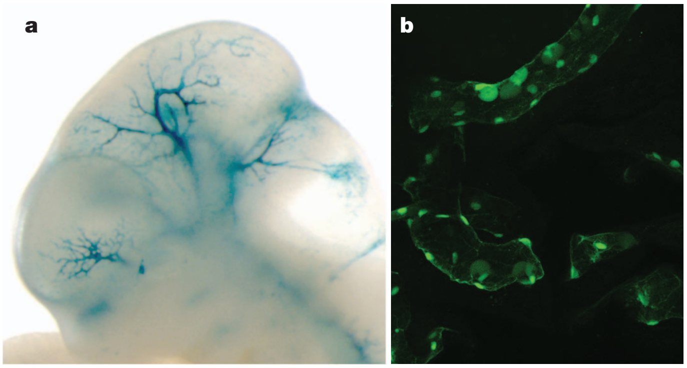


Figure 6 | Experimental characterization of segmentations. Randomly sampled E state segments (see Table 3) from the K562 segmentation were cloned for mouse- and fish-based transgenic enhancer assays. **a**, Representative LacZ-stained transgenic embryonic day (E)11.5 mouse embryo obtained with construct hs2065 (EN167, chr10: 46052882–46055670, GRCh37). Highly reproducible staining in the blood vessels was observed in 9 out of 9 embryos resulting from independent transgenic integration events. **b**, Representative green fluorescent protein reporter transgenic medaka fish obtained from a construct with a basal *hsp70* promoter on meganuclease-based transfection. Reproducible transgenic expression in the circulating nucleated blood cells and the endothelial cell walls was seen in 81 out of 100 transgenic tests of this construct.

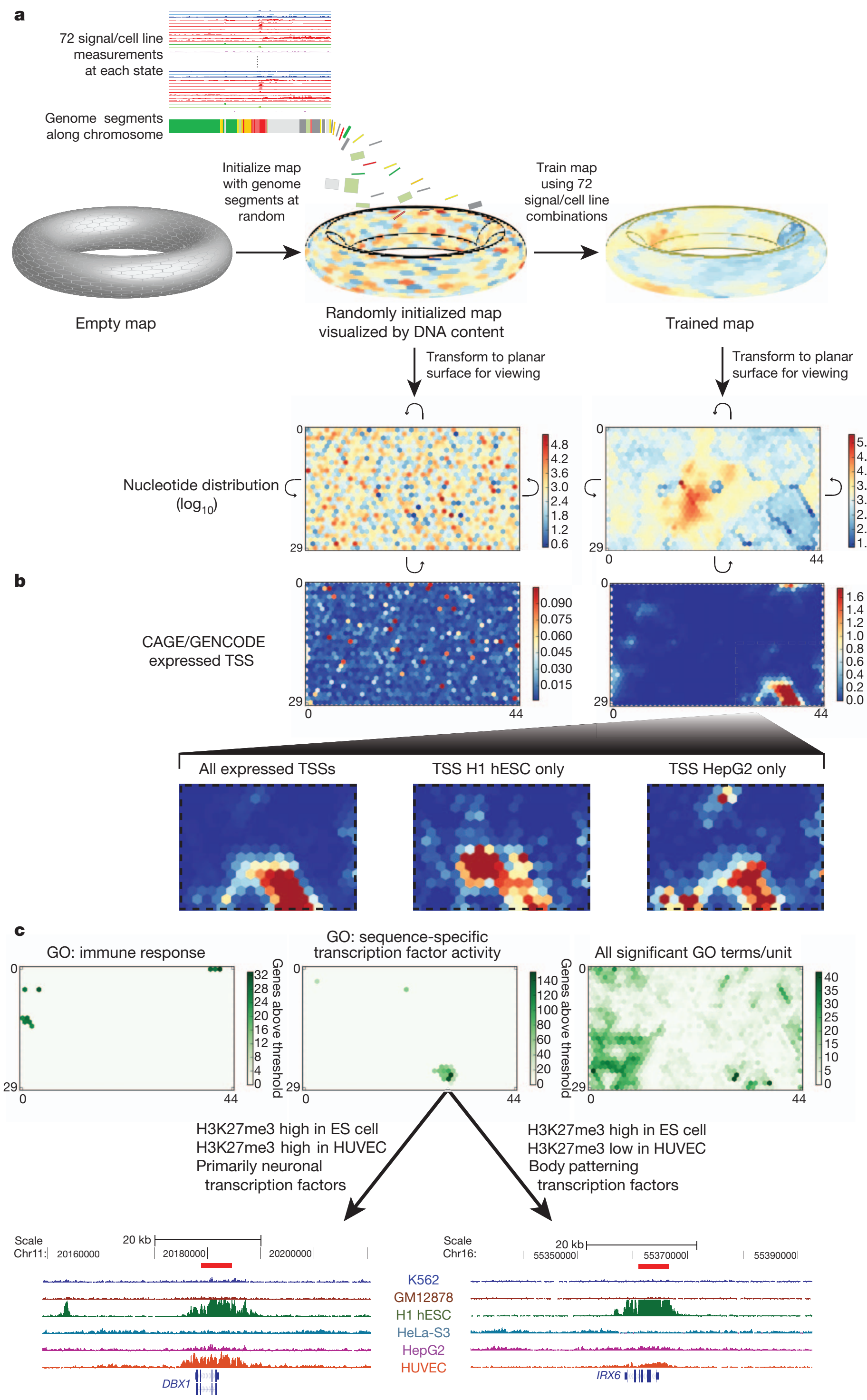


Figure 7 | High-resolution segmentation of ENCODE data by self-organizing maps (SOM). **a**–**c**, The training of the SOM (**a**) and analysis of the results (**b**, **c**) are shown. Initially we arbitrarily placed genomic segments from the ChromHMM segmentation on to the toroidal map surface, although the SOM does not use the ChromHMM state assignments (**a**). We then trained the map using the signal of the 12 different ChIP-seq and DNase-seq assays in the six cell types analysed. Each unit of the SOM is represented here by a hexagonal cell in a planar two-dimensional view of the toroidal map. Curved arrows indicate that traversing the edges of two dimensional view leads back to the opposite edge. The resulting map can be overlaid with any class of ENCODE or other data to view the distribution of that data within this high-resolution segmentation. In panel **a** the distributions of genome bases across the untrained and trained map (left and right, respectively) are shown using heat-map colours for $\log_{10}$ values. **b**, The distribution of TSSs from CAGE experiments of GENCODE annotation on the planar representations of either the initial random organization (left) or the final trained SOM (right) using heat maps coloured according to the accompanying scales. The bottom half of **b** expands the different distributions in the SOM for all expressed TSSs (left) or TSSs specifically expressed in two example cell lines, H1 hESC (centre) and HepG2 (right). **c**, The association of Gene Ontology (GO) terms on the same representation of the same trained SOM. We assigned genes that are within 20 kb of a genomic segment in a SOM unit to that unit, and then associated this set of genes with GO terms using a hypergeometric distribution after correcting for multiple testing. Map units that are significantly associated to GO terms are coloured green, with increasing strength of colour reflecting increasing numbers of genes significantly associated with the GO terms for either immune response (left) or sequence-specific transcription factor activity (centre). In each case, specific SOM units show association with these terms. The right-hand panel shows the distribution on the same SOM of all significantly associated GO terms, now colouring by GO term count per SOM unit. For sequence-specific transcription factor activity, two example genomic regions are extracted at the bottom of panel **c** from neighbouring SOM units. These are regions around the *DBX1* (from SOM unit 26,31, left panel) and *IRX6* (SOM unit 27,30, right panel) genes, respectively, along with their H3K27me3 ChIP-seq signal for each of the tier 1 and 2 cell types. For *DBX1*, representative of a set of primarily neuronal transcription factors associated with unit 26,31, there is a repressive H3K27me3 signal in both H1 hESCs and HUVECs; for *IRX6*, representative of a set of body patterning transcription factors associated with SOM unit 27,30, the repressive mark is restricted largely to the embryonic stem (ES) cell. An interactive version of this figure is available in the online version of the paper.

chromatin data. Many of the ultra-fine-grained state classifications revealed in the SOM are associated with specific gene ontology (GO) terms (right panel of Fig. 7c). For instance, the left panel of Fig. 7c identifies ten SOM map units enriched with genomic regions associated with genes associated with the GO term 'immune response'. The central panel identifies a different set of map units enriched for the GO term 'sequence-specific transcription factor activity'. The two map units most enriched for this GO term, indicated by the darkest green colouring, contain genes with segments that are high in H3K27me3 in H1 hESCs, but that differ in H3K27me3 levels in HUVECs. Gene function analysis with the GO ontology tool (GREAT[71]) reveals that the map unit with high H3K27me3 levels in both cell types is enriched in transcription factor genes with known neuronal functions, whereas the neighbouring map unit is enriched in genes involved in body patterning. The genome browser shots at the bottom of Fig. 7c pick out an example region for each of the two SOM map units illustrating the difference in H3K27me3 signal. Overall, we have 228 distinct GO terms associated with specific segments across one or more states (A. Mortazavi, personal communication), and can assign over one-third of genes to a GO annotation solely on the basis of its multicellular histone patterns. Thus, the SOM analysis provides a fine-grained map of chromatin data across multiple cell types, which can then be used to relate chromatin structure to other data types at differing levels of resolution (for instance, the large cluster of units containing any active TSS, its subclusters composed of units enriched in TSSs active in only one cell type, or individual map units significantly enriched for specific GO terms).

The classifications presented here are necessarily limited by the assays and cell lines studied, and probably contain a number of heterogeneous classes of elements. Nonetheless, robust classifications can be made, allowing a systematic view of the human genome.

Insights into human genomic variation

We next explored the potential impact of sequence variation on ENCODE functional elements. We examined allele-specific variation using results from the GM12878 cells that are derived from an individual (NA12878) sequenced in the 1000 Genomes project, along with her parents. Because ENCODE assays are predominantly sequence-based, the trio design allows each GM12878 data set to be divided by the specific parental contributions at heterozygous sites, producing aggregate haplotypic signals from multiple genomic sites. We examined 193 ENCODE assays for allele-specific biases using 1,409,992 phased, heterozygous SNPs and 167,096 insertions/deletions (indels) (Fig. 8). Alignment biases towards alleles present in the reference genome sequence were avoided using a sequence specifically tailored to the variants and haplotypes present in NA12878 (a 'personalized genome')[72]. We found instances of preferential binding towards each parental allele. For example, comparison of the results from the POLR2A, H3K79me2 and H3K27me3 assays in the region of *NACC2* (Fig. 8a) shows a strong paternal bias for H3K79me2 and POL2RA and a strong maternal bias for H3K27me3, indicating differential activity for the maternal and paternal alleles.

Figure 8b shows the correlation of selected allele-specific signals across the whole genome. For instance, we found a strong allelic correlation between POL2RA and BCLAF1 binding, as well as negative correlation between H3K79me2 and H3K27me3, both at genes (Fig. 8b, below the diagonal, bottom left) and chromosomal segments (top right). Overall, we found that positive allelic correlations among the 193 ENCODE assays are stronger and more frequent than negative correlations. This may be due to preferential capture of accessible alleles and/or the specific histone modification and transcription factor, assays used in the project.

Rare variants, individual genomes and somatic variants

We further investigated the potential functional effects of individual variation in the context of ENCODE annotations. We divided NA12878 variants into common and rare classes, and partitioned these into those overlapping ENCODE annotation (Fig. 9a and Supplementary Tables 1 and 2, section K). We also predicted potential functional effects: for protein-coding genes, these are either non-synonymous SNPs or variants likely to induce loss of function by frame-shift, premature stop, or splice-site disruption; for other regions, these are variants that overlap a transcription-factor-binding site. We found similar numbers of potentially functional variants affecting protein-coding genes or affecting other ENCODE annotations, indicating that many functional variants within individual genomes lie outside exons of protein-coding genes. A more detailed analysis of regulatory variant annotation is described in ref. 73.

To study further the potential effects of NA12878 genome variants on transcription-factor-binding regions, we performed peak calling using a constructed personal diploid genome sequence for NA12878 (ref. 72). We aligned ChIP-seq sequences from GM12878 separately against the maternal and paternal haplotypes. As expected, a greater fraction of reads were aligned than to the reference genome (see Supplementary Information, Supplementary Fig. 1, section K). On average, approximately 1% of transcription-factor-binding sites in GM12878 cells are detected in a haplotype-specific fashion. For instance, Fig. 9b shows a CTCF-binding site not detected using the

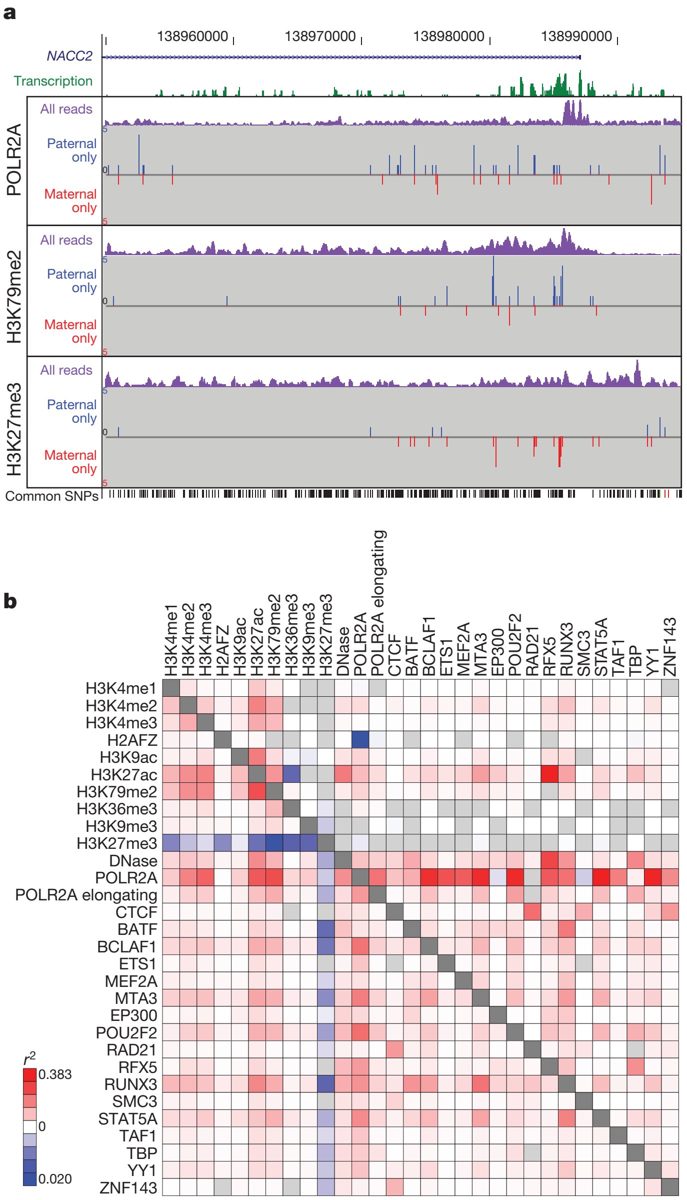


Figure 8 | Allele-specific ENCODE elements. a, Representative allele-specific information from GM12878 cells for selected assays around the first exon of the *NACC2* gene (genomic region Chr9: 138950000–138995000, GRCh37). Transcription signal is shown in green, and the three sections show allele-specific data for three data sets (POLR2A, H3K79me2 and H3K27me3 ChIP-seq). In each case the purple signal is the processed signal for all sequence reads for the assay, whereas the blue and red signals show sequence reads specifically assigned to either the paternal or maternal copies of the genome, respectively. The set of common SNPs from dbSNP, including the phased, heterozygous SNPs used to provide the assignment, are shown at the bottom of the panel. *NACC2* has a statistically significant paternal bias for POLR2A and the transcription-associated mark H3K79me2, and has a significant maternal bias for the repressive mark H3K27me3. **b**, Pair-wise correlations of allele-specific signal within single genes (below the diagonal) or within individual ChromHMM segments across the whole genome for selected DNase-seq and histone modification and transcription factor ChIP-seq assays. The extent of correlation is coloured according to the heat-map scale indicated from positive correlation (red) through to anti-correlation (blue). An interactive version of this figure is available in the online version of the paper.

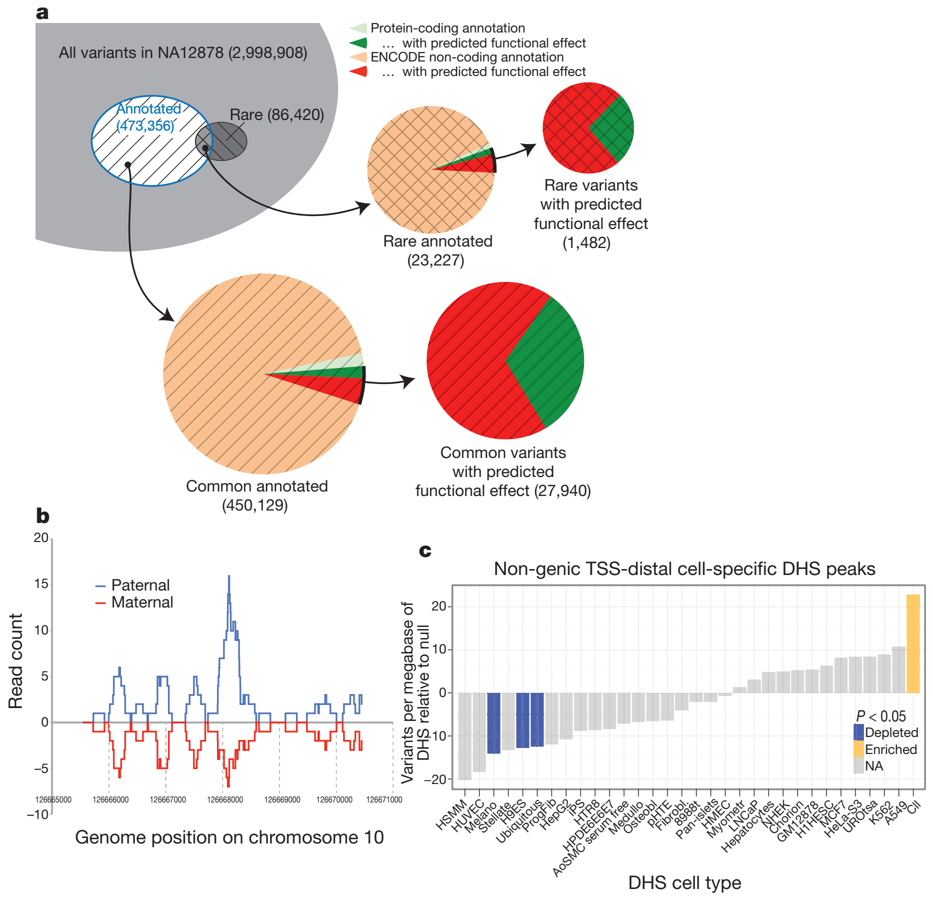


Figure 9 | Examining ENCODE elements on a per individual basis in the normal and cancer genome. **a**, Breakdown of variants in a single genome (NA12878) by both frequency (common or rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project[55])) and by ENCODE annotation, including protein-coding gene and non-coding elements (GENCODE annotations for protein-coding genes, pseudogenes and other ncRNAs, as well as transcription-factor-binding sites from ChIP-seq data sets, excluding broad annotations such as histone modifications, segmentations and RNA-seq). Annotation status is further subdivided by predicted functional effect, being non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations. A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category. **b**, One of several relatively rare occurrences, where alignment to an individual genome sequence (paternal and maternal panels) shows a different readout from the reference genome. In this case, a paternal-haplotype-specific CTCF peak is identified. **c**, Relative level of somatic variants from a whole-genome melanoma sample that occur in DHSs unique to different cell lines. The coloured bars show cases that are significantly enriched or suppressed in somatic mutations. Details of ENCODE cell types can be found at http://encodeproject.org/ENCODE/cellTypes.html. An interactive version of this figure is available in the online version of the paper.

reference sequence that is only present on the paternal haplotype due to a 1-bp deletion (see also Supplementary Fig. 2, section K). As costs of DNA sequencing decrease further, optimized analysis of ENCODE-type data should use the genome sequence of the individual or cell being analysed when possible.

Most analyses of cancer genomes so far have focused on characterizing somatic variants in protein-coding regions. We intersected four available whole-genome cancer data sets with ENCODE annotations (Fig. 9c and Supplementary Fig. 2, section L). Overall, somatic variation is relatively depleted from ENCODE annotated regions, particularly for elements specific to a cell type matching the putative tumour source (for example, skin melanocytes for melanoma). Examining the mutational spectrum of elements in introns for cases where a strand-specific mutation assignment could be made reveals that there are mutational spectrum differences between DHSs and unannotated regions (0.06 Fisher's exact test, Supplementary Fig. 3, section L). The suppression of somatic mutation is consistent with important functional roles of these elements within tumour cells, highlighting a potential alternative set of targets for examination in cancer.

Common variants associated with disease

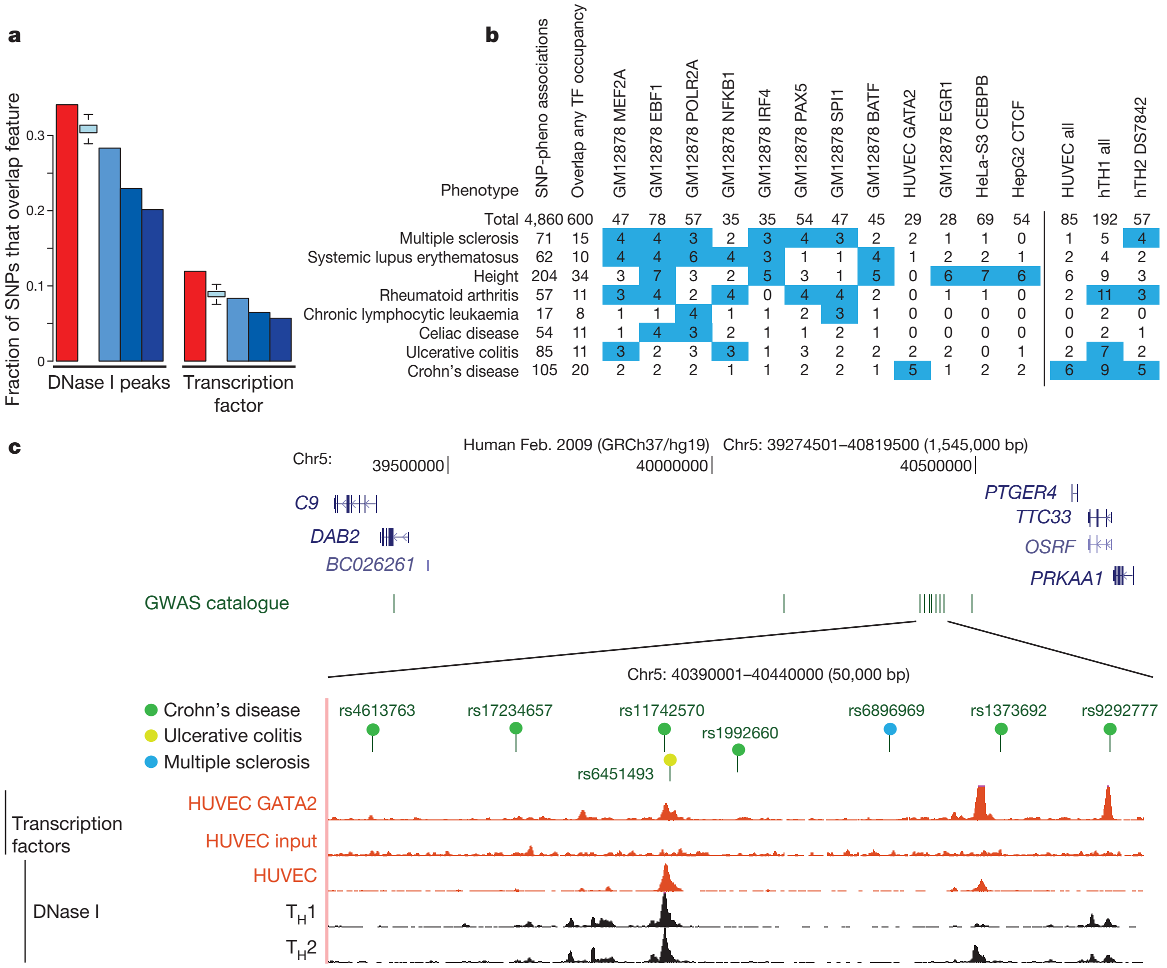
In recent years, GWAS have greatly extended our knowledge of genetic loci associated with human disease risk and other phenotypes. The output of these studies is a series of SNPs (GWAS SNPs) correlated with a phenotype, although not necessarily the functional variants. Notably, 88% of associated SNPs are either intronic or intergenic[74]. We examined 4,860 SNP–phenotype associations for 4,492 SNPs curated in the National Human Genome Research Institute (NHGRI) GWAS catalogue[74]. We found that 12% of these SNPs overlap transcription-factor-occupied regions whereas 34% overlap DHSs (Fig. 10a). Both figures reflect significant enrichments relative to the overall proportions of 1000 Genomes project SNPs (about 6% and 23%, respectively). Even after accounting for biases introduced by selection of SNPs for the standard genotyping arrays, GWAS SNPs show consistently higher overlap with ENCODE annotations (Fig. 10a, see Supplementary Information). Furthermore, after partitioning the genome by density of different classes of functional elements, GWAS SNPs were consistently enriched beyond all the genotyping SNPs in function-rich partitions, and depleted in function-poor partitions (see Supplementary Fig. 1, section M). GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types (see Supplementary Fig. 2, section M).

Examining the SOM of integrated ENCODE annotations (see above), we found 19 SOM map units showing significant enrichment for GWAS SNPs, including many SOM units previously associated with specific gene functions, such as the immune response regions.

Thus, an appreciable proportion of SNPs identified in initial GWAS scans are either functional or lie within the length of an ENCODE annotation (~500 bp on average) and represent plausible candidates for the functional variant. Expanding the set of feasible functional SNPs to those in reasonable linkage disequilibrium, up to 71% of GWAS SNPs have a potential causative SNP overlapping a DNase I site, and 31% of loci have a candidate SNP that overlaps a binding site occupied by a transcription factor (see also refs 73, 75).

The GWAS catalogue provides a rich functional categorization from the precise phenotypes being studied. These phenotypic categorizations are nonrandomly associated with ENCODE annotations and there is marked correspondence between the phenotype and the identity of the cell type or transcription factor used in the ENCODE assay (Fig. 10b). For example, five SNPs associated with Crohn's disease overlap GATA2-binding sites (*P* value 0.003 by random permutation or 0.001 by an empirical approach comparing to the GWAS-matched SNPs; see Supplementary Information), and fourteen are located in DHSs found in immunologically relevant cell types. A notable example is a gene desert on chromosome 5p13.1 containing eight SNPs associated with inflammatory diseases. Several are close to or within DHSs in T-helper type 1 (T_H1) and T_H2 cells as well as peaks of binding by transcription factors in HUVECs (Fig. 10c). The latter cell line is not immunological, but factor occupancy detected there could be a proxy for binding of a more relevant factor, such as GATA3, in T cells. Genetic variants in this region also affect expression levels of *PTGER4* (ref. 76), encoding the prostaglandin receptor EP4. Thus, the ENCODE data reinforce the hypothesis that genetic variants in 5p13.1 modulate the expression of flanking genes, and furthermore provide the specific hypothesis that the variants affect occupancy of a GATA factor in an allele-specific manner, thereby influencing susceptibility to Crohn's disease.

Nonrandom association of phenotypes with ENCODE cell types strengthens the argument that at least some of the GWAS lead SNPs are functional or extremely close to functional variants. Each of the associations between a lead SNP and an ENCODE annotation remains a credible hypothesis of a particular functional element



Phenotype	SNP-pheno associations	Overlap any TF occupancy	GM12878 MEF2A	GM12878 EBF1	GM12878 POLR2A	GM12878 NFKB1	GM12878 IRF4	GM12878 PAX5	GM12878 SPI1	GM12878 BATF	HUVEC GATA2	GM12878 EGR1	HeLa-S3 CEBPB	HepG2 CTCF	HUVEC all	hTH1 all	hTH2 DS7842
Total	4,860	600	47	78	57	35	35	54	47	45	29	28	69	54	85	192	57
Multiple sclerosis	71	15	**4**	**4**	**3**	2	**3**	**4**	**3**	2	2	1	1	0	1	5	**4**
Systemic lupus erythematosus	62	10	**4**	**4**	**6**	**4**	**3**	1	1	**4**	1	2	2	1	2	4	2
Height	204	34	3	**7**	3	2	**5**	3	1	**5**	0	**6**	**7**	**6**	6	9	3
Rheumatoid arthritis	57	11	**3**	**4**	2	**4**	0	**4**	**4**	2	0	1	1	0	2	**11**	**3**
Chronic lymphocytic leukaemia	17	8	1	1	**4**	1	1	2	**3**	1	0	0	0	0	0	2	0
Celiac disease	54	11	1	**4**	**3**	2	1	1	2	1	0	0	0	0	0	2	1
Ulcerative colitis	85	11	**3**	2	3	**3**	1	3	2	2	2	2	0	1	2	**7**	2
Crohn's disease	105	20	2	2	2	1	1	2	2	1	**5**	1	2	2	**6**	**9**	**5**

Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data. **a**, Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at http://main.genome-browser.bx.psu.edu (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. **b**, Aggregate overlap of phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical *P*-value threshold ≤0.01 (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The *P* value for the total number of phenotype–transcription factor associations is <0.001. **c**, Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper T_H1 and T_H2 cells. An interactive version of this figure is available in the online version of the paper.

class or cell type to explore with future experiments. Supplementary Tables 1–3, section M, list all 14,885 pairwise associations across the ENCODE annotations. The accompanying papers have a more detailed examination of common variants with other regulatory information[19,25,29,73,75,77].

Concluding remarks

The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome. Our analyses have revealed many novel aspects of gene expression and regulation as well as the organization of such information, as illustrated by the accompanying papers (see http://www.encodeproject.org/ENCODE/pubs.html for collected ENCODE publications). However, there are still many specific details, particularly about the mechanistic processes that generate these elements and how and where they function, that require additional experiments to elucidate.

The large spread of coverage—from our highest resolution, most conservative set of bases implicated in GENCODE protein-coding gene exons (2.9%) or specific protein DNA binding (8.5%) to the broadest, most general set of marks covering the genome (approximately 80%), with many gradations in between—presents a spectrum of elements with different functional properties discovered by ENCODE. A total of 99% of the known bases in the genome are within 1.7 kb of any ENCODE element, whereas 95% of bases are within 8 kb of a bound transcription factor motif or DNase I footprint. Interestingly, even using the most conservative estimates, the fraction of bases likely to be involved in direct gene regulation, even though incomplete, is significantly higher than that ascribed to protein-coding exons (1.2%), raising the possibility that more information in the human genome may be important for gene regulation than for biochemical function. Many of the regulatory elements are not constrained across mammalian evolution, which so far has been one of the most reliable indications of an important biochemical event for the organism. Thus, our data provide orthologous indicators for suggesting possible functional elements.

Importantly, for the first time we have sufficient statistical power to assess the impact of negative selection on primate-specific elements, and all ENCODE classes display evidence of negative selection in these unique-to-primate elements. Furthermore, even with our most conservative estimate of functional elements (8.5% of putative DNA/protein binding regions) and assuming that we have already sampled half of the elements from our transcription factor and cell-type diversity, one would estimate that at a minimum 20% (17% from protein binding and 2.9% protein coding gene exons) of the genome participates in these specific functions, with the likely figure significantly higher.

The broad coverage of ENCODE annotations enhances our understanding of common diseases with a genetic component, rare genetic diseases, and cancer, as shown by our ability to link otherwise anonymous associations to a functional element. ENCODE and similar studies provide a first step towards interpreting the rest of the genome—beyond protein-coding genes—thereby augmenting common disease genetic studies with testable hypotheses. Such information justifies performing whole-genome sequencing (rather than exome only, 1.2% of the genome) on rare diseases and investigating somatic variants in non-coding functional elements, for instance, in cancer. Furthermore, as GWAS analyses typically associate disease to SNPs in large regions, comparison to ENCODE non-coding functional elements can help pinpoint putative causal variants in addition to refinement of location by fine-mapping techniques[78]. Combining ENCODE data with allele-specific information derived from individual genome sequences provides specific insight on the impact of a genetic variant. Indeed, we believe that a significant goal would be to use functional data such as that derived from this project to assign every genomic variant to its possible impact on human phenotypes.

So far, ENCODE has sampled 119 of 1,800 known transcription factors and general components of the transcriptional machinery on a limited number of cell types, and 13 of more than 60 currently known histone or DNA modifications across 147 cell types. DNase I, FAIRE and extensive RNA assays across subcellular fractionations have been undertaken on many cell types, but overall these data reflect a minor fraction of the potential functional information encoded in the human genome. An important future goal will be to enlarge this data set to additional factors, modifications and cell types, complementing the other related projects in this area (for example, Roadmap Epigenomics Project, http://www.roadmapepigenomics.org/, and International Human Epigenome Consortium, http://www.ihec-epigenomes.org/). These projects will constitute foundational resources for human genomics, allowing a deeper interpretation of the organization of gene and regulatory information and the mechanisms of regulation, and thereby provide important insights into human health and disease. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

METHODS SUMMARY

For full details of Methods, see Supplementary Information.

Received 24 November 2011; accepted 29 May 2012.

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
2. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
3. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
5. Chiaromonte, F. *et al.* The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 245–254 (2003).
6. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
7. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324,** 389–392 (2009).
8. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
9. Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. *Genome Res.* **17,** 1245–1253 (2007).
10. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21,** 1769–1776 (2011).
11. Asthana, S. *et al.* Widely distributed noncoding purifying selection in the human genome. *Proc. Natl Acad. Sci. USA* **104,** 12410–12415 (2007).
12. Landt, S. G. *et al.* ChIP-seq guidelines and practices used by the ENCODE and modENCODE consortia. *Genome Res.* http://dx.doi.org/10.1101/gr.136184.111 (2012).
13. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).
14. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* http://dx.doi.org/10.1101/gr.135350.111 (2012).
15. Howald, C. *et al.* Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.* http://dx.doi.org/10.1101/gr.134478.111 (2012).
16. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* http://dx.doi.org/10.1038/nature11233 (this issue).
17. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* http://dx.doi.org/10.1101/gr.132159.111 (2012).
18. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).
19. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* http://dx.doi.org/10.1038/nature11245 (this issue).
20. Bickel, P. J., Boley, N., Brown, J. B., Huang, H. Y. & Zhang, N. R. Subsampling methods for genomic inference. *Ann. Appl. Stat.* **4,** 1660–1697 (2010).
21. Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genet.* **7,** e1001290 (2011).
22. Li, X. Y. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol.* **12,** R34 (2011).

23. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–455 (2011).
24. Zhang, Y. *et al.* Primary sequence and epigenetic determinants of *in vivo* occupancy of genomic DNA by GATA1. *Nucleic Acids Res.* **37,** 7024–7038 (2009).
25. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* http://dx.doi.org/10.1038/nature11212 (this issue).
26. Whitfield, T. W. *et al.* Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13,** R50 (2012).
27. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57,** 159–197 (1988).
28. Urnov, F. D. Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. *J. Cell. Biochem.* **88,** 684–694 (2003).
29. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* http://dx.doi.org/10.1038/nature11232 (this issue).
30. Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.* http://dx.doi.org/10.1101/gr.136366.111 (2012).
31. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. III. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16,** 919–932 (2002).
32. Frietze, S., O'Geen, H., Blahnik, K. R., Jin, V. X. & Farnham, P. J. ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS ONE* **5,** e15082 (2010).
33. Boyle, A. P. *et al.* High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).
34. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6,** 283–289 (2009).
35. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
36. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
37. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
38. Hon, G. C., Hawkins, R. D. & Ren, B. Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.* **18,** R195–R201 (2009).
39. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Rev. Genet.* **12,** 7–18 (2011).
40. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
41. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5,** e1000566 (2009).
42. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnol.* **27,** 361–368 (2009).
43. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
44. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87,** 953–959 (1996).
45. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).
46. Dekker, J. Gene regulation in the third dimension. *Science* **319,** 1793–1794 (2008).
47. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
48. Lajoie, B. R., van Berkum, N. L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nature Methods* **6,** 690–691 (2009).
49. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* http://dx.doi.org/10.1038/nature11279 (this issue).
50. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).
51. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
52. Borneman, A. R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317,** 815–819 (2007).
53. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.* **39,** 730–732 (2007).
54. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
55. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
56. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
57. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biol.* **13,** R49 (2012).
58. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev. Genet.* **8,** 424–436 (2007).
59. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13,** R53 (2012).
60. Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nature Struct. Mol. Biol.* **17,** 1495–1499 (2010).
61. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* http://dx.doi.org/10.1101/gr.134445.111 (2012).
62. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
63. Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16,** 6677–6690 (1988).
64. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
65. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).
66. Frietze, S. *et al.* Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* **13,** R52 (2012).
67. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
68. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9,** 473–476 (2012).
69. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
70. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Struct. Mol. Biol.* **18,** 956–963 (2011).
71. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* **28,** 495–501 (2010).
72. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
73. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* http://dx.doi.org/10.1101/gr.137323.112 (2012).
74. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
75. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* http://dx.doi.org/10.1101/gr.136127.111 (2012).
76. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3,** e58 (2007).
77. Vernot, B. *et al.* Personal and population genomics of human regulatory variation. *Genome Res.* http://dx.doi.org/10.1101/gr.134890.111 (2012).
78. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470,** 264–268 (2011).
79. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* http://dx.doi.org/10.1101/gr.136838.111 (2012).
80. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463,** 943–947 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank additional members of our laboratories and institutions who have contributed to the experimental and analytical components of this project. We thank D. Leja for assistance with production of the figures. The Consortium is funded by grants from the NHGRI as follows: production grants: U54HG004570 (B. E. Bernstein); U01HG004695 (E. Birney); U54HG004563 (G. E. Crawford); U54HG004557 (T. R. Gingeras); U54HG004555 (T. J. Hubbard); U41HG004568 (W. J. Kent); U54HG004576 (R. M. Myers); U54HG004558 (M. Snyder); U54HG004592 (J. A. Stamatoyannopoulos). Pilot grants: R01HG003143 (J. Dekker); RC2HG005591 and R01HG003700 (M. C. Giddings); R01HG004456-03 (Y. Ruan); U01HG004571 (S. A. Tenenbaum); U01HG004561 (Z. Weng); RC2HG005679 (K. P. White). This project was supported in part by American Recovery and Reinvestment Act (ARRA) funds from the NHGRI through grants U54HG004570, U54HG004563, U41HG004568, U54HG004592, R01HG003143, RC2HG005591, R01HG003541, U01HG004561, RC2HG005679 and R01HG003988 (L. Pennacchio). In addition, work from NHGRI Groups was supported by the Intramural Research Program of the NHGRI (L. Elnitski, ZIAHG200323; E. H. Margulies, ZIAHG200341). Research in the Pennachio laboratory was performed at Lawrence Berkeley National Laboratory and at the United States Department of Energy Joint Genome Institute, Department of Energy Contract DE-AC02-05CH11231, University of California.

Author Contributions See the consortium author list for details of author contributions.

Author Information The Supplementary Information is accompanied by a Virtual Machine (VM) containing the functioning analysis data and code. Further details of the VM are available from http://encodeproject.org/ENCODE/integrativeAnalysis/VM. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.B. (birney@ebi.ac.uk).

The ENCODE Project Consortium

Overall coordination (data analysis coordination) Ian Dunham[1], Anshul Kundaje[2]†; **Data production leads (data production)** Shelley F. Aldred[3], Patrick J. Collins[3], Carrie A. Davis[4], Francis Doyle[5], Charles B. Epstein[6], Seth Frietze[7], Jennifer Harrow[8], Rajinder Kaul[9], Jainab Khatun[10], Bryan R. Lajoie[11], Stephen G. Landt[12], Bum-Kyu Lee[13], Florencia Pauli[14], Kate R. Rosenbloom[15], Peter Sabo[16], Alexias Safi[17], Amartya Sanyal[11], Noam Shoresh[6], Jeremy M. Simon[18], Lingyun Song[17], Nathan D. Trinklein[3]; **Lead analysts (data analysis)** Robert C. Altshuler[19], Ewan Birney[1], James B. Brown[20], Chao Cheng[21], Sarah Djebali[22], Xianjun Dong[23], Ian Dunham[1], Jason Ernst[19]†, Terrence S. Furey[24], Mark Gerstein[21], Belinda Giardine[25], Melissa Greven[23], Ross C. Hardison[25,26], Robert S. Harris[25], Javier Herrero[1], Michael M. Hoffman[16], Sowmya Iyer[27], Manolis Kellis[19], Jainab Khatun[10], Pouya Kheradpour[19], Anshul Kundaje[2]†, Timo Lassmann[28], Qunhua Li[20]†, Xinying Lin[23], Georgi K. Marinov[29], Angelika Merkel[22], Ali Mortazavi[30], Stephen C. J. Parker[31], Timothy E. Reddy[14]†, Joel Rozowsky[21], Felix Schlesinger[4], Robert E. Thurman[16], Jie Wang[23], Lucas D. Ward[19], Troy W. Whitfield[23], Steven P. Wilder[1], Weisheng Wu[25], Hualin S. Xi[32], Kevin Y. Yip[21]†, Jiali Zhuang[23]; **Writing group** Bradley E. Bernstein[6,33], Ewan Birney[1], Ian Dunham[1], Eric D. Green[34], Chris Gunter[14], Michael Snyder[12]; **NHGRI project management (scientific management)** Michael J. Pazin[35], Rebecca F. Lowdon[35]†, Laura A. L. Dillon[35]†, Leslie B. Adams[35], Caroline J. Kelly[35], Julia Zhang[35]†, Judith R. Wexler[35]†, Eric D. Green[34], Peter J. Good[35], Elise A. Feingold[35]; **Principal investigators (steering committee)** Bradley E. Bernstein[6,33], Ewan Birney[1], Gregory E. Crawford[17,36], Job Dekker[11], Laura Elnitski[37], Peggy J. Farnham[7], Mark Gerstein[21], Morgan C. Giddings[10], Thomas R. Gingeras[4,38], Eric D. Green[34], Roderic Guigó[22,39], Ross C. Hardison[25,26], Timothy J. Hubbard[8], Manolis Kellis[19], W. James Kent[15], Jason D. Lieb[18], Elliott H. Margulies[31]†, Richard M. Myers[14], Michael Snyder[12], John A. Stamatoyannopoulos[40], Scott A. Tenenbaum[5], Zhiping Weng[23], Kevin P. White[41], Barbara Wold[29,42]; **Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis)** Jainab Khatun[10], Yanbao Yu[43], John Wrobel[10], Brian A. Risk[10], Harsha P. Gunawardena[43], Heather C. Kuiper[43], Christopher W. Maier[43], Ling Xie[43], Xian Chen[43], Morgan C. Giddings[10]; **Broad Institute Group (data production and analysis)** Bradley E. Bernstein[6,33], Charles B. Epstein[6], Noam Shoresh[6], Jason Ernst[19]†, Pouya Kheradpour[19], Tarjei S. Mikkelsen[6], Shawn Gillespie[33], Alon Goren[6,33], Oren Ram[6,33], Xiaolan Zhang[6], Li Wang[6], Robbyn Issner[6], Michael J. Coyne[6], Timothy Durham[6], Manching Ku[6,33], Thanh Truong[6], Lucas D. Ward[19], Robert C. Altshuler[19], Matthew L. Eaton[19], Manolis Kellis[19]; **Cold Spring Harbor, University of Geneva, Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis)** Sarah Djebali[22], Carrie A. Davis[4], Angelika Merkel[22], Alex Dobin[4], Timo Lassmann[28], Ali Mortazavi[30], Andrea Tanzer[22], Julien Lagarde[22], Wei Lin[4], Felix Schlesinger[4], Chenghai Xue[4], Georgi K. Marinov[29], Jainab Khatun[10], Brian A. Williams[29], Chris Zaleski[4], Joel Rozowsky[21], Maik Röder[22], Felix Kokocinski[8]†, Rehab F. Abdelhamid[28], Tyler Alioto[22,44], Igor Antoshechkin[29], Michael T. Baer[4], Philippe Batut[4], Ian Bell[45], Kimberly Bell[4], Sudipto Chakrabortty[4], Xian Chen[43], Jacqueline Chrast[46], Joao Curado[22], Thomas Derrien[22]†, Jorg Drenkow[4], Erica Dumais[45], Jackie Dumais[45], Radha Duttagupta[45], Megan Fastuca[4], Kata Fejes-Toth[4]†, Pedro Ferreira[22], Sylvain Foissac[45], Melissa J. Fullwood[47]†, Hui Gao[45], David Gonzalez[22], Assaf Gordon[4], Harsha P. Gunawardena[43], Cédric Howald[46], Sonali Jha[4], Rory Johnson[22], Philipp Kapranov[45]†, Brandon King[29], Colin Kingswood[22,44], Guoliang Li[48], Oscar J. Luo[47], Eddie Park[30], Jonathan B. Preall[4], Kimberly Presaud[4], Paolo Ribeca[22,44], Brian A. Risk[10], Daniel Robyr[49], Xiaoan Ruan[47], Michael Sammeth[22,44], Kuljeet Singh Sandhu[47], Lorain Schaeffer[29], Lei-Hoon See[4], Atif Shahab[47], Jorgen Skancke[22], Ana Maria Suzuki[28], Hazuki Takahashi[28], Hagen Tilgner[22]†, Diane Trout[29], Nathalie Walters[46], Huaien Wang[4], John Wrobel[10], Yanbao Yu[43], Yoshihide Hayashizaki[28], Jennifer Harrow[8], Mark Gerstein[21], Timothy J. Hubbard[8], Alexandre Reymond[46], Stylianos E. Antonarakis[49], Gregory J. Hannon[4], Morgan C. Giddings[10], Yijun Ruan[47], Barbara Wold[29,42], Piero Carninci[28], Roderic Guigó[22,39], Thomas R. Gingeras[4,38]; **Data coordination center at UC Santa Cruz (production data coordination)** Kate R. Rosenbloom[15], Cricket A. Sloan[15], Katrina Learned[15], Venkat S. Malladi[15], Matthew C. Wong[15], Galt P. Barber[15], Melissa S. Cline[15], Timothy R. Dreszer[15], Steven G. Heitner[15], Donna Karolchik[15], W. James Kent[15], Vanessa M. Kirkup[15], Laurence R. Meyer[15], Jeffrey C. Long[15], Morgan Maddren[15], Brian J. Raney[15]; **Duke University, EBI, University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis)** Terrence S. Furey[24], Lingyun Song[17], Linda L. Grasfeder[18], Paul G. Giresi[18], Bum-Kyu Lee[13], Anna Battenhouse[13], Nathan C. Sheffield[17], Jeremy M. Simon[18], Kimberly A. Showers[18], Alexias Safi[17], Darin London[17], Akshay A. Bhinge[13], Christopher Shestak[18], Matthew R. Schaner[18], Seul Ki Kim[18], Zhuzhu Z. Zhang[18], Piotr A. Mieczkowski[50], Joanna O. Mieczkowska[18], Zheng Liu[13], Ryan M. McDaniell[13], Yunyun Ni[13], Naim U. Rashid[51], Min Jae Kim[18], Sheera Adar[18], Zhancheng Zhang[24], Tianyuan Wang[17], Deborah Winter[17], Damian Keefe[1], Ewan Birney[1], Vishwanath R. Iyer[13], Jason D. Lieb[18], Gregory E. Crawford[17,36]; **Genome Institute of Singapore group (data production and analysis)** Guoliang Li[48], Kuljeet Singh Sandhu[47], Meizhen Zheng[47], Ping Wang[47], Oscar J. Luo[47], Atif Shahab[47], Melissa J. Fullwood[47]†, Xiaoan Ruan[47], Yijun Ruan[47]; **HudsonAlpha Institute, Caltech, UC Irvine, Stanford group (data production and analysis)** Richard M. Myers[14], Florencia Pauli[14], Brian A. Williams[29], Jason Gertz[14], Georgi K. Marinov[29], Timothy E. Reddy[14]†, Jost Vielmetter[29,42], E. Christopher Partridge[14], Diane Trout[29], Katherine E. Varley[14], Clarke Gasper[29,42], Anita Bansal[14], Shirley Pepke[29,52], Preti Jain[14], Henry Amrhein[29], Kevin M. Bowling[14], Michael Anaya[29,42], Marie K. Cross[14], Brandon King[29], Michael A. Muratet[14], Igor Antoshechkin[29], Kimberly M. Newberry[14], Kenneth McCue[29], Amy S. Nesmith[14], Katherine I. Fisher-Aylor[29,42], Barbara Pusey[14], Gilberto DeSalvo[29,42], Stephanie L. Parker[14]†, Sreeram Balasubramanian[29,42], Nicholas S. Davis[14], Sarah K. Meadows[14], Tracy Eggleston[14], Chris Gunter[14], J. Scott Newberry[14], Shawn E. Levy[14], Devin M. Absher[14], Ali Mortazavi[30], Wing H. Wong[53], Barbara Wold[29,42]; **Lawrence Berkeley National Laboratory group (targeted experimental validation)** Matthew J. Blow[54], Axel Visel[54,55], Len A. Pennachio[54,55]; **NHGRI groups (data production and analysis)** Laura Elnitski[37], Elliott H. Margulies[31]†, Stephen C. J. Parker[31], Hanna M. Petrykowska[37]; **Sanger Institute, Washington University, Yale University, Center for Genomic Regulation, Barcelona, UCSC, MIT, University of Lausanne, CNIO group (data production and analysis)** Alexej Abyzov[21], Bronwen Aken[8], Daniel Barrell[8], Gemma Barson[8], Andrew Berry[8], Alexandra Bignell[8], Veronika Boychenko[8], Giovanni Bussotti[22], Jacqueline Chrast[46], Claire Davidson[8], Thomas Derrien[22]†, Gloria Despacio-Reyes[8], Mark Diekhans[15], Iakes Ezkurdia[56], Adam Frankish[8], James Gilbert[8], Jose Manuel Gonzalez[8], Ed Griffiths[8], Rachel Harte[15], David A. Hendrix[19], Cédric Howald[46], Toby Hunt[8], Irwin Jungreis[19], Mike Kay[8], Ekta Khurana[21], Felix Kokocinski[8]†, Jing Leng[21], Michael F. Lin[19], Jane Loveland[8], Zhi Lu[57], Deepa Manthravadi[8], Marco Mariotti[22], Jonathan Mudge[8], Gaurab Mukherjee[8], Cedric Notredame[22], Baikang Pei[21], Jose Manuel Rodriguez[56], Gary Saunders[8], Andrea Sboner[58], Stephen Searle[8], Cristina Sisu[21], Catherine Snow[8], Charlie Steward[8], Andrea Tanzer[22], Electra Tapanari[8], Michael L. Tress[56], Marijke J. van Baren[59]†, Nathalie Walters[46], Stefan Washietl[19], Laurens Wilming[8], Amonida Zadissa[8], Zhengdong Zhang[60], Michael Brent[59], David Haussler[61], Manolis Kellis[19], Alfonso Valencia[56], Mark Gerstein[21], Alexandre Reymond[46], Roderic Guigó[22,39], Jennifer Harrow[8], Timothy J. Hubbard[8]; **Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UC Davis group (data production and analysis)** Stephen G. Landt[12], Seth Frietze[7], Alexej Abyzov[21], Nick Addleman[12], Roger P. Alexander[21], Raymond K. Auerbach[21], Suganthi Balasubramanian[21], Keith Bettinger[12], Nitin Bhardwaj[21], Alan P. Boyle[12], Alina R. Cao[62], Philip Cayting[12], Alexandra Charos[63], Yong Cheng[12], Chao Cheng[21], Catharine Eastman[12], Ghia Euskirchen[12], Joseph D. Fleming[64], Fabian Grubert[12], Lukas Habegger[21], Manoj Hariharan[12], Arif Harmanci[21], Sushma Iyengar[65], Victor X. Jin[66], Konrad J. Karczewski[12], Maya Kasowski[12], Phil Lacroute[12], Hugo Lam[12], Nathan Lamarre-Vincent[64], Jing Leng[21], Jin Lian[67], Marianne Lindahl-Allen[64], Renqiang Min[21]†, Benoit Miotto[64], Hannah Monahan[63], Zarmik Moqtaderi[64], Xinmeng J. Mu[21], Henriette O'Geen[62], Zhengqing Ouyang[12], Dorrelyn Patacsil[12], Baikang Pei[21], Debasish Raha[63], Lucia Ramirez[12], Brian Reed[63], Joel Rozowsky[21], Andrea Sboner[58], Minyi Shi[12], Cristina Sisu[21], Teri Slifer[12], Heather Witt[7], Linfeng Wu[12], Xiaoqin Xu[62], Koon-Kiu Yan[21], Xinqiong Yang[12], Kevin Y. Yip[21]†, Zhengdong Zhang[60], Kevin Struhl[64], Sherman M. Weissman[67], Mark Gerstein[21], Peggy J. Farnham[7], Michael Snyder[12]; **University of Albany SUNY group (data production and analysis)** Scott A. Tenenbaum[5], Luiz O. Penalva[68], Francis Doyle[5]; **University of Chicago, Stanford group (data production and analysis)** Subhradip Karmakar[41], Stephen G. Landt[12], Raj R. Bhanvadia[41], Alina Choudhury[41], Marc Domanus[41], Lijia Ma[41], Jennifer Moran[41], Dorrelyn Patacsil[12], Teri Slifer[12], Alec Victorsen[41], Xinqiong Yang[12], Michael Snyder[12], Kevin P. White[41]; **University of Heidelberg group (targeted experimental validation)** Thomas Auer[69]†, Lazaro Centanin[69], Michael Eichenlaub[69], Franziska Gruhl[69], Stephan Heermann[69], Burkhard Hoeckendorf[69], Daigo Inoue[69], Tanja Kellner[69], Stephan Kirchmaier[69], Claudia Mueller[69], Robert Reinhardt[69], Lea Schertel[69], Stephanie Schneider[69], Rebecca Sinn[69], Beate Wittbrodt[69], Jochen Wittbrodt[69]; **University of Massachusetts Medical School Bioinformatics group (data production and analysis)** Zhiping Weng[23], Troy W. Whitfield[23], Jie Wang[23], Patrick J. Collins[3], Shelley F. Aldred[3], Nathan D. Trinklein[3], E. Christopher Partridge[14], Richard M. Myers[14]; **University of Massachusetts Medical School Genome Folding group (data production and analysis)** Job Dekker[11], Gaurav Jain[11], Bryan R. Lajoie[11], Amartya Sanyal[11]; **University of Washington, University of Massachusetts Medical Center group (data production and analysis)** Gayathri Balasundaram[70], Daniel L. Bates[16], Rachel Byron[70], Theresa K. Canfield[16], Morgan J. Diegel[16], Douglas Dunn[16], Abigail K. Ebersol[71], Tristan Frum[71], Kavita Garg[72], Erica Gist[16], R. Scott Hansen[71], Lisa Boatman[71], Eric Haugen[16], Richard Humbert[16], Gaurav Jain[11], Audra K. Johnson[16], Ericka M. Johnson[71], Tattyana V. Kutyavin[16], Bryan R. Lajoie[11], Kristen Lee[16], Dimitra Lotakis[71], Matthew T. Maurano[16], Shane J. Neph[16], Fiedencio V. Neri[16], Eric D. Nguyen[71], Hongzhu Qu[16], Alex P. Reynolds[16], Vaughn Roach[16], Eric Rynes[16], Peter Sabo[16], Minerva E. Sanchez[71], Richard S. Sandstrom[16], Amartya Sanyal[11], Anthony O. Shafer[16], Andrew B. Stergachis[16], Sean Thomas[16], Robert E. Thurman[16], Benjamin Vernot[16], Jeff Vierstra[16], Shinny Vong[16], Hao Wang[16], Molly A. Weaver[16], Yongqi Yan[71], Miaohua Zhang[70], Joshua M. Akey[16], Michael Bender[70], Michael O. Dorschner[73], Mark Groudine[70], Michael J. MacCoss[16], Patrick Navas[71], George Stamatoyannopoulos[71], Rajinder Kaul[9], Job Dekker[11], John A. Stamatoyannopoulos[40]; **Data Analysis Center (data analysis)**; Ian Dunham[1], Kathryn Beal[1], Alvis Brazma[74], Paul Flicek[1], Javier Herrero[1], Nathan Johnson[1], Damian Keefe[1], Margus Lukk[74]†, Nicholas M. Luscombe[75], Daniel Sobral[1]†, Juan M. Vaquerizas[75], Steven P. Wilder[1], Serafim Batzoglou[2], Arend Sidow[76], Nadine Hussami[2], Sofia Kyriazopoulou-Panagiotopoulou[2], Max W. Libbrecht[2]†, Marc A. Schaub[2], Anshul Kundaje[2]†, Ross C. Hardison[25,26], Webb Miller[25], Belinda Giardine[25], Robert S. Harris[25], Weisheng Wu[25], Peter J. Bickel[20], Balazs Banfai[20], Nathan P. Boley[20], James B. Brown[20], Haiyan Huang[20], Qunhua Li[20]†, Jingyi Jessica Li[20], William Stafford Noble[16,77], Jeffrey A. Bilmes[78], Orion J. Buske[16], Michael M. Hoffman[16], Avinash D. Sahu[16]†, Peter V. Kharchenko[79], Peter J. Park[79], Dannon Baker[80], James Taylor[80], Zhiping Weng[23], Sowmya Iyer[27], Xianjun Dong[23], Melissa Greven[23], Xinying Lin[23], Jie Wang[23], Hualin S. Xi[32], Jiali Zhuang[23], Mark Gerstein[21], Roger P. Alexander[21], Suganthi Balasubramanian[21], Chao Cheng[21], Arif Harmanci[21], Lucas Lochovsky[21], Renqiang Min[21]†, Xinmeng J. Mu[21], Joel Rozowsky[21], Koon-Kiu Yan[21], Kevin Y. Yip[21]† & Ewan Birney[1]

[1]Vertebrate Genomics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [2]Department of Computer Science, Stanford University, 318 Campus Drive, Stanford, California 94305-5428, USA. [3]SwitchGear Genomics, 1455 Adams Drive Suite 1317, Menlo Park, California 94025, USA. [4]Functional Genomics, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [5]College of Nanoscale Sciences and Engineering, University ay Albany-SUNY, 257 Fuller Road, NFE 4405, Albany, New York 12203, USA. [6]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [7]Biochemistry and Molecular Biology, USC/Norris Comprehensive Cancer Center, 1450 Biggy Street, NRT 6503, Los Angeles, California 90089, USA. [8]Informatics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [9]Department of Medicine, Division of Medical Genetics, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195, USA. [10]College of Arts and Sciences, Boise State University, 1910 University Drive, Boise, Idaho 83725, USA. [11]Program in Systems Biology, Program in Gene Function and Expression, Department of Biochemistry and Molecular

Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [12]Department of Genetics, Stanford University, 300 Pasteur Drive, M-344, Stanford, California 94305-5120, USA. [13]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, The University of Texas at Austin, 1 University Station A4800, Austin, Texas 78712, USA. [14]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA. [15]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [16]Department of Genome Sciences, University of Washington, 3720 15th Ave NE, Seattle, Washington 98195-5065, USA. [17]Institute for Genome Sciences and Policy, Duke University, 101 Science Drive, Durham, North Carolina 27708, USA. [18]Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, 408 Fordham Hall, Chapel Hill, North Carolina 27599-3280, USA. [19]Computer Science and Artificial Intelligence Laboratory, Broad Institute of MIT and Harvard, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA. [20]Department of Statistics, University of California, Berkeley, 367 Evans Hall, University of California, Berkeley, Berkeley, California 94720, USA. [21]Computational Biology and Bioinformatics Program, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [22]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Doctor Aiguader, 88, Barcelona 08003, Catalonia, Spain. [23]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [24]Department of Genetics, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, CB 7240, Chapel Hill, North Carolina 27599, USA. [25]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, Wartik Laboratory, University Park, Pennsylvania 16802, USA. [26]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 304 Wartik Laboratory, University Park, Pennsylvania 16802, USA. [27]Program in Bioinformatics, Boston University, 24 Cummington Street, Boston, Massachusetts 02215, USA. [28]RIKEN Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [29]Division of Biology, California Institute of Technology, 156-291200 East California Boulevard, Pasadena, California 91125, USA. [30]Developmental and Cell Biology and Center for Complex Biological Systems, University of California Irvine, 2218 Biological Sciences III, Irvine, California 92697-2300, USA. [31]Genome Technology Branch, National Human Genome Research Institute, 5625 Fishers Lane, Bethesda, Maryland 20892, USA. [32]Department of Biochemistry and Molecular Pharmacology, Bioinformatics Core, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [33]Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, 185 Cambridge St CPZN 8400, Boston, Massachusetts 02114, USA. [34]National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, Building 31, Room 4B09, Bethesda, Maryland 20892-2152, USA. [35]National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane, Bethesda, Maryland 20892-9307, USA. [36]Department of Pediatrics, Division of Medical Genetics, Duke University School of Medicine, Durham, North Carolina 27710, USA. [37]National Human Genome Research Institute, National Institutes of Health, 5625 Fishers Lane, Rockville, Maryland 20892, USA. [38]Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [39]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia 08002, Spain. [40]Department of Genome Sciences, Box 355065, and Department of Medicine, Division of Oncology, Box 358081, University of Washington, Seattle, Washington 98195-5065, USA. [41]Institute for Genomics and Systems Biology, The University of Chicago, 900 East 57th Street, 10100 KCBD, Chicago, Illinois 60637, USA. [42]Beckman Institute, California Institute of Technology, 156-29 1200 E. California Boulevard, Pasadena, California 91125, USA. [43]Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Campus Box 7260, 120 Mason Farm Road, 3010 Genetic Medicine Building, Chapel Hill, North Carolina 27599, USA. [44]Centro Nacional de Análisis Genómico (CNAG), C/Baldiri Reixac 4, Torre I, Barcelona, Catalonia 08028, Spain. [45]Genomics, Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [46]Center for Integrative Genomics, University of Lausanne, Genopode Building, 1015 Lausanne, Switzerland. [47]Genome Technology and Biology, Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [48]Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [49]Department of Genetic Medicine and Development, University of Geneva Medical School, and University Hospitals of Geneva, 1 rue Michel-Servet, 1211 Geneva 4, Switzerland. [50]Department of Genetics, The University of North Carolina at Chapel Hill, 5078 GMB, Chapel Hill, North Carolina 27599-7264, USA. [51]Department of Biostatistics, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, 408 Fordham Hall, Chapel Hill, North Carolina 27599-7445, USA. [52]Center for Advanced Computing Research, California Institute of Technology, MC 158-79, 1200 East California Boulevard, Pasadena, California 91125, USA. [53]Department of Statistics, Stanford University, Sequoia Hall. 390 Serra Mall, Stanford, California 94305-4065, USA. [54]DOE Joint Genome Institute, Walnut Creek, California, USA. [55]Genomics Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, MS 84-171, Berkeley, California 94720, USA. [56]Structural Computational Biology, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, 3, 28029 Madrid, Spain. [57]School of Life Sciences, Tsinghua University, School of Life Sciences, Tsinghua University, 100084 Beijing, China. [58]Department of Pathology and Laboratory Medicine, Institute for Computational Biomedicine, Weill Cornell Medical College, 1305 York Avenue, Box 140, New York, New York 10065, USA. [59]Computer Science and Engineering, Washington University in St Louis, St Louis, Missouri 63130, USA. [60]Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Room 353A, Bronx, New York 10461, USA. [61]Center for Biomolecular Science and Engineering, Howard Hughes Medical Institute, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [62]Genome Center, University of California-Davis, 451 Health Sciences Drive, Davis, California 95616, USA. [63]Department of Molecular, Cellular, and Developmental Biology, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06511, USA. [64]Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [65]Biochemistry and Molecular Biology, University of Southern California, 1501 San Pablo Street, Los Angeles, California 90089, USA. [66]Department of Biomedical Informatics, Ohio State University, 3172C Graves Hall, 333 W Tenth Avenue, Columbus, Ohio 43210, USA. [67]Department of Genetics, Yale University, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [68]Department of Cellular and Structural Biology, Children's Cancer Research Institute–UTHSCSA, Mail code 7784- 7703 Floyd Curl Dr, San Antonio, Texas 78229, USA. [69]Centre for Organismal Studies (COS) Heidelberg, University of Heidelberg, Im Neuenheimer Feld 230, 69120 Heidelberg, Germany. [70]Basic Sciences Division, Fred Hutchinson Cancer Research Center, 825 Eastlake Avenue East, Seattle, Washington 98109, USA. [71]Department of Medicine, Division of Medical Genetics, Box 357720, University of Washington, Seattle, Washington 98195-7720, USA. [72]Division of Human Biology, Fred Hutchinson Cancer Research Center, 825 Eastlake Avenue East, Seattle, Washington 98109, USA. [73]Department of Psychiatry and Behavioral Sciences, Box 356560, University of Washington, Seattle, Washington 98195-6560, USA. [74]Microarray Informatics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [75]Genomics and Regulatory Systems Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [76]Department of Pathology, Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [77]Department of Computer Science and Engineering, 185 Stevens Way, Seattle, Washington 98195, USA. [78]Department of Electrical Engineering, University of Washington, 185 Stevens Way, Seattle, Washington 98195, USA. [79]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, Massachusetts 02115, USA. [80]Departments of Biology and Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322, USA. †Present addresses: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA (A.K.); UCLA Biological Chemistry Department, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, Jonsson Comprehensive Cancer Center, 615 Charles E Young Dr South, Los Angeles, California 90095, USA (J.E.); Department of Statistics, 514D Wartik Lab, Penn State University, State College, Pennsylvania 16802, USA (Q.L.); Department of Biostatistics and Bioinformatics and the Institute for Genome Sciences and Policy, Duke University School of Medicine, 101 Science Drive, Durham, North Carolina 27708, USA (T.E.R.); Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (K.Y.Y.); Department of Genetics, Washington University in St Louis, St Louis, Missouri 63110, USA (R.F.L.); Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA (L.A.L.D.); National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA (J.Z.); University of California, Davis Population Biology Graduate Group, Davis, California 95616, USA (J.R.W.); Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL, UK (E.H.M.); BlueGnome Ltd., CPC4, Capital Park, Fulbourn, Cambridge CB21 5XE, UK (F.K.); Institut de Génétique et Développement de Rennes, CNRS-UMR6061, Université de Rennes 1, F-35000 Rennes, Brittany, France (T.D.); Caltech, 1200 East California Boulevard, Pasadena, California 91125, USA (K.F.-T.); A*STAR-Duke-NUS Neuroscience Research Partnership, 8 College Road, Singapore 169857, Singapore (M.J.F.); St Laurent Institute, One Kendall Square, Cambridge, Massachusetts 02139, USA (P.K.); Department of Genetics, Stanford University, Stanford, California 94305, USA (H.T.); Biomedical Sciences (BMS) Graduate Program, University of California, San Francisco, 513 Parnassus Avenue, HSE-1285, San Francisco, California 94143-0505, USA (S.L.P.); Monterey Bay Aquarium Research Institute, Moss Landing, California 95039, USA (M.J.v.B.); Department of Machine Learning, NEC Laboratories America, 4 Independence Way, Princeton, New Jersey 08540, USA (R.M.); Neuronal Circuit Development Group, Unité de Génétique et Biologie du Développement, U934/UMR3215, Institut Curie–Centre de Recherche, Pole de Biologie du Développement et Cancer, 26, rue d'Ulm, 75248 Paris Cedex 05, France (T.A.); Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK (M.L.); Unidade de Bioinformatica, Rua da Quinta Grande, 6, P-2780-156 Oeiras, Portugal (D.S.); Department of Genome Sciences, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195-5065, USA (M.W.L.); Center for Bioinformatics and Computational Biology, 3115 Ag/Life Surge Building 296, University of Maryland, College Park, Maryland 20742, USA (A.D.S.).

C

Landscape of transcription in human cells

Originally published as:

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. 2012. Landscape of transcription in human cells. *Nature.* **489**(7414):101–108. doi: 10.1038/nature11233.

doi:10.1038/nature11233

Landscape of transcription in human cells

Sarah Djebali[1]*, Carrie A. Davis[2]*, Angelika Merkel[1], Alex Dobin[2], Timo Lassmann[3], Ali Mortazavi[4,5], Andrea Tanzer[1], Julien Lagarde[1], Wei Lin[2], Felix Schlesinger[2], Chenghai Xue[2], Georgi K. Marinov[4], Jainab Khatun[6], Brian A. Williams[4], Chris Zaleski[2], Joel Rozowsky[7,8], Maik Röder[1], Felix Kokocinski[9], Rehab F. Abdelhamid[3], Tyler Alioto[1,10], Igor Antoshechkin[4], Michael T. Baer[2], Nadav S. Bar[11], Philippe Batut[2], Kimberly Bell[2], Ian Bell[12], Sudipto Chakrabortty[2], Xian Chen[13], Jacqueline Chrast[14], Joao Curado[1], Thomas Derrien[1], Jorg Drenkow[2], Erica Dumais[12], Jacqueline Dumais[12], Radha Duttagupta[12], Emilie Falconnet[15], Meagan Fastuca[2], Kata Fejes-Toth[2], Pedro Ferreira[1], Sylvain Foissac[12], Melissa J. Fullwood[16], Hui Gao[12], David Gonzalez[1], Assaf Gordon[2], Harsha Gunawardena[13], Cedric Howald[14], Sonali Jha[2], Rory Johnson[1], Philipp Kapranov[12,17], Brandon King[4], Colin Kingswood[1,10], Oscar J. Luo[16], Eddie Park[5], Kimberly Persaud[2], Jonathan B. Preall[2], Paolo Ribeca[1,10], Brian Risk[6], Daniel Robyr[15], Michael Sammeth[1,10], Lorian Schaffer[4], Lei-Hoon See[2], Atif Shahab[16], Jorgen Skancke[1,11], Ana Maria Suzuki[3], Hazuki Takahashi[3], Hagen Tilgner[1]†, Diane Trout[4], Nathalie Walters[14], Huaien Wang[2], John Wrobel[6], Yanbao Yu[13], Xiaoan Ruan[16], Yoshihide Hayashizaki[3], Jennifer Harrow[9], Mark Gerstein[7,8,18], Tim Hubbard[9], Alexandre Reymond[14], Stylianos E. Antonarakis[15], Gregory Hannon[2], Morgan C. Giddings[6,13], Yijun Ruan[16], Barbara Wold[4], Piero Carninci[3], Roderic Guigó[1,19] & Thomas R. Gingeras[2,12]

Eukaryotic cells make many types of primary and processed RNAs that are found either in specific subcellular compartments or throughout the cells. A complete catalogue of these RNAs is not yet available and their characteristic subcellular localizations are also poorly understood. Because RNA represents the direct output of the genetic information encoded by genomes and a significant proportion of a cell's regulatory capabilities are focused on its synthesis, processing, transport, modification and translation, the generation of such a catalogue is crucial for understanding genome function. Here we report evidence that three-quarters of the human genome is capable of being transcribed, as well as observations about the range and levels of expression, localization, processing fates, regulatory regions and modifications of almost all currently annotated and thousands of previously unannotated RNAs. These observations, taken together, prompt a redefinition of the concept of a gene.

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

As the technologies for RNA profiling and for cell-type isolation and culture continue to improve, the catalogue of RNA types has grown and led to an increased appreciation for the numerous biological functions carried out by RNA, arguably putting them on par with the functional importance of proteins[1]. The Encyclopedia of DNA Elements (ENCODE) project has sought to catalogue the repertoire of RNAs produced by human cells as part of the intended goal of identifying and characterizing the functional elements present in the human genome sequence[2]. The five-year pilot phase of the ENCODE project[3] examined approximately 1% of the human genome and observed that the gene-rich and gene-poor regions were pervasively transcribed, confirming results of previous studies[4,5]. During the second phase of the ENCODE project, lasting 5 years, the scope of examination was broadened to interrogate the complete human genome. Thus, we have sought to both provide a genome-wide catalogue of human transcripts and to identify the subcellular localization for the RNAs produced. Here we report identification and characterization of annotated and novel RNAs that are enriched in either of the two major cellular subcompartments (nucleus and cytosol) for all 15 cell lines studied, and in three additional subnuclear compartments in one cell line. In addition, we have sought to determine whether identified transcripts are modified at their 5′ and 3′ termini by the presence of a 7-methyl guanosine cap or polyadenylation, respectively. We further studied primary transcript and processed product relationships for a large proportion of the previously annotated long and small RNAs. These results considerably extend the current genome-wide annotated catalogue of long polyadenylated and small RNAs collected by the GENCODE annotation group[6–8]. Taken together, our genome-wide compilation of subcellular localized and product-precursor-related RNAs serves as a public resource and reveals new and detailed facets of the RNA landscape.

- Cumulatively, we observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts, respectively, with no cell line showing more than 56.7% of the union of the expressed transcriptomes across all cell lines. The consequent reduction in the length of 'intergenic regions' leads to a significant

[1] Centre for Genomic Regulation and UPF, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain. [2] Cold Spring Harbor Laboratory, Functional Genomics, 1 Bungtown Road, Cold Spring Harbor, New York 11742, USA. [3] RIKEN Yokohama Institute, RIKEN Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [4] California Institute of Technology, Division of Biology, 2 Beckman Institute, Pasadena, California 91125, USA. [5] University of California Irvine, Department of Developmental and Cell Biology, 2300 Biological Sciences III, Irvine, California 92697, USA. [6] Boise State University, College of Arts & Sciences, 1910 University Drive, Boise, Idaho 83725, USA. [7] Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [8] Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [9] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [10] Centro Nacional de Análisis Genómico (CNAG), C/ Baldiri Reixac 4, Torre I, Barcelona 08028, Catalonia, Spain. [11] Department of Chemical Engineering, Norwegian University of Science and Technology, Trondheim NO-7491, Norway. [12] Affymetrix, Inc, 3380 Central Expressway, Santa Clara, California 95051, USA. [13] University of North Carolina at Chapel Hill, Department of Biochemistry & Biophysics, 120 Mason Farm Road, Chapel Hill, North Carolina 27599, USA. [14] University of Lausanne, Center for Integrative Genomics, Genopode building, Lausanne 1015, Switzerland. [15] University of Geneva Medical School, Department of Genetic Medicine and Development and iGE3 Institute of Genetics and Genomics of Geneva, 1 rue Michel-Servet, Geneva 1211, Switzerland. [16] Genome Institute of Singapore, Genome Technology and Biology, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [17] St Laurent Institute, One Kendall Square, Cambridge, Massachusetts 02141, USA. [18] Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [19] Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain. †Present address: Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.
*These authors contributed equally to this work.

overlapping of neighbouring gene regions and prompts a redefinition of a gene.

- Isoform expression by a gene does not follow a minimalistic expression strategy, resulting in a tendency for genes to express many isoforms simultaneously, with a plateau at about 10–12 expressed isoforms per gene per cell line.
- Cell-type-specific enhancers are promoters that are differentiable from other regulatory regions by the presence of novel RNA transcripts, chromatin marks and DNase I hypersensitive sites.
- Coding and non-coding transcripts are predominantly localized in the cytosol and nucleus, respectively, with a range of expression spanning six orders of magnitude for polyadenylated RNAs, and five orders of magnitude for non-polyadenylated RNAs.
- Approximately 6% of all annotated coding and non-coding transcripts overlap with small RNAs and are probably precursors to these small RNAs. The subcellular localization of both annotated and unannotated short RNAs is highly specific.

RNA data set generation

We performed subcellular compartment fractionation (whole cell, nucleus and cytosol) before RNA isolation in 15 cell lines (Supplementary Table 1) to interrogate deeply the human transcriptome. For the K562 cell line, we also performed additional nuclear subfractionation into chromatin, nucleoplasm and nucleoli. The RNAs from each of these subcompartments were prepared in replica and were separated based on length into >200 nucleotides (long) and <200 nucleotides (short). Long RNAs were further fractionated into polyadenylated and non-polyadenylated transcripts. A number of complementary technologies were used to characterize these RNA fractions as to their sequence (RNA-seq), sites of initiation of transcription (cap-analysis of gene expression (CAGE)[9]) and sites of 5′ and 3′ transcript termini (paired end tags (PET)[10]; Supplementary Fig. 1). Sequence reads were mapped and post-processed using a variety of software tools (Supplementary Table 2 and Supplementary Fig. 2). We used the mapped data to assemble and quantify *de novo* elements (exons, transcripts, genes, contigs, splice junctions and transcription start sites (TSSs)) as well as to quantify annotated GENCODE (v7) elements. Elements and quantifications were further assessed for reproducibility between replicates using a non-parametric version (npIDR, Supplementary Information) of the irreproducible detection rate (IDR) statistical test[11]. Only elements deemed to be reproducible with at least 90% likelihood were used in most analyses. The raw data, mapped data and elements were then made available by the ENCODE Data Coordination Center (DCC, http://genome.ucsc.edu/ENCODE/dataSummary.html) (Supplementary Fig. 2). These data, as well as additional data on all intermediate processing steps, are available on the RNA Dashboard (http://genome.crg.cat/encode_RNA_dashboard/).

Long RNA expression landscape

Detection of annotated and novel transcripts

The GENCODE gene (Supplementary Fig. 3a) and transcript (Supplementary Fig. 3b) reference annotation[8] captures our current understanding of the polyadenylated human transcriptome. In the samples interrogated here, we cumulatively detected 70% of annotated splice junctions, transcripts and genes (Fig. 1 and Table 1a). We also detected approximately 85% of annotated exons with an average coverage by RNA-seq contigs of 96%. The variation in the proportion of detected elements among cell lines was small (Fig. 1, width of box plots). Consistent with earlier studies, most annotated elements are present in both polyadenylated (Supplementary Table 3a) and non-polyadenylated (Supplementary Table 3b) samples[12–15]. Only a small proportion of GENCODE elements (0.4% of exons, 2.8% of splice sites, 3.3% of transcripts and 4.7% of genes) are detected exclusively in the non-polyadenylated RNA fraction.

Beyond the GENCODE annotated elements, we observed a substantial number of novel elements represented by reproducible RNA-seq contigs. These novel elements covered 78% of the intronic nucleotides and 34% of the intergenic sequences (Supplementary Fig. 4). Overall, the unique contribution of each cell line to the coverage of the genome tends to be small and similar for each cell line (Supplementary Fig. 5). We used the Cufflinks algorithm (see Supplementary Information), and predicted over all long RNA-seq samples 94,800 exons, 69,052 splice junctions, 73,325 transcripts and 41,204 genes in intergenic and antisense regions (Table 1b). These novel elements increase the GENCODE collection of exons, splice sites, transcripts and genes by 19%, 22%, 45% and 80%, respectively. The increase in the number of genes and the relatively low contribution of novel splice sites is primarily caused by the detection of both polyadenylated and non-polyadenylated mono-exonic transcripts (Supplementary Table 3). Detection of unspliced transcripts could partially be an artefact caused by low levels of DNA contamination or by incomplete determination of transcript structures.

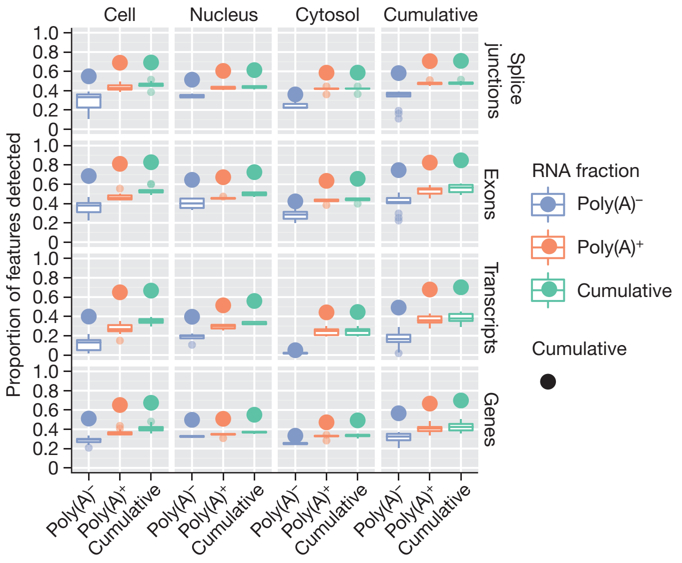


Figure 1 | A large majority of GENCODE elements are detected by RNA-seq data. Shown are GENCODE-detected elements in the polyadenylated and non-polyadenylated fractions of cellular compartments (cumulative counts for both RNA fractions and compartments refer to elements present in any of the fractions or compartments). Each box plot is generated from values across all cell lines, thus capturing the dispersion across cell lines. The largest point shows the cumulative value over all cell lines.

Independent validation of multi-exonic transcript models and the associated predicted coding products were carried out using overlapping targeted 454 Life Sciences (Roche) paired-end reads and mass spectrometry. Of approximately 3,000 intergenic and antisense transcript models tested, validation rates from 70% to 90% were observed, depending on the number of reads and IDR score. In addition, these experiments led to the identification of more than 22,000 novel splice sites not previously detected, meaning an almost eightfold increase in detection compared to the sites originally detected with RNA-seq (Supplementary Fig. 6). Using mass spectrometric analyses, we investigated what fraction of the novel Cufflinks transcript models show evidence consistent with protein expression. We produced 998,570 spectra from two cell lines (K562 and GM12878; J. Khatun *et al.*, manuscript in preparation), and mapped them to a three-frame translation of the novel Cufflinks models (Supplementary Material). At a 1% false discovery rate (FDR), we identified 419 novel models with 5 or more spectral and/or 2 or more peptide hits, of which only 56 were intergenic or antisense to GENCODE genes (Supplementary Table 4 and Supplementary Fig. 7). Thus, most novel transcripts seem to lack protein-coding capacity.

The transcriptome of nuclear subcompartments

For the K562 cell line, we also analysed RNA isolated from three subnuclear compartments (chromatin, nucleolus and nucleoplasm;

Supplementary 5). Almost half (18,330) of the GENCODE (v7) annotated genes detected for all 15 cell lines (35,494) were identified in the analysis of just these three nuclear subcompartments. In addition, there were as many novel unannotated genes found in K562 subcompartments as there were in all other data sets combined (Supplementary Table 5 and Table 1b). For all annotated (Supplementary Table 5.1) or novel (Supplementary Table 5.2) elements, only a small fraction in each subcompartment was unique to that compartment (Supplementary Table 6).

The interrogation of different subcellular RNA fractions provides snapshots of the status of the RNA population along the RNA processing pathway. Thus, by analysing short and long RNAs in the different subcellular compartments, we confirm that splicing predominantly occurs during transcription. By using RNA-seq to measure the degree of completion of splicing (Fig. 2a), we observed that around most exons, introns are already being spliced in chromatin-associated RNA—the fraction that includes RNAs in the process of being transcribed (Fig. 2b). Concomitantly, we found strong enrichment specifically of spliceosomal small nuclear RNAs (snRNAs) in this RNA fraction (see ‘Short RNA expression landscape’ later). Co-transcriptional splicing provides an explanation for the increasing evidence connecting chromatin structure to splicing regulation, and we have observed that exons in the process of being spliced are enriched in a number of chromatin marks[16,17].

Gene expression across cell lines

The analyses of RNAs isolated from different subcellular compartments also provide information concerning compartment-specific relative steady-state abundance and the post transcriptional processing state (spliced/unspliced, polyadenylated/non-polyadenylated, 5′ capped/uncapped) for each of the detected transcripts. The observed range of gene expression spans six orders of magnitude for polyadenylated RNAs (from 10^{-2} to 10^{4} reads per kilobase per million reads (r.p.k.m.)), and five orders of magnitude (from 10^{-2} to 10^{3} r.p.k.m.) for non-polyadenylated RNAs (Fig. 3 and Supplementary Fig. 8a). The distribution of gene expression is very similar across cell lines, with protein-coding genes, as a class, having on average higher expression levels than long non-coding RNAs (lncRNAs). Assuming that 1–4 r.p.k.m. approximates to 1 copy per cell[18], we find that almost one-quarter of expressed protein-coding genes and 80% of the detected lncRNAs are present in our samples in 1 or fewer copies per cell. The general lower level of gene expression measured in lncRNAs may not necessarily be the result of consistent low RNA copy number in all cells within the population interrogated, but may also result from restricted expression in only a subpopulation of cells. In some cell lines, individual lncRNAs can exhibit steady-state expression levels as high as those of protein-coding genes. This is, for example, seen in the expression of the protein-coding gene actin, gamma 1 (*ACTG1*), and the non-coding gene, *H19* (Fig. 3). *ACTG1* transcripts are part of all non-muscle cytoskeleton systems within cells and show a steady-state expression level at the population level that is at least 1–2 logs greater than *H19*, a cytosolic non-coding RNA (ncRNA). However, when measured at the individual transcript level, expression of lncRNA transcripts is comparable to that of individual protein-coding transcripts (Supplementary Fig. 8b).

Novel antisense and intergenic genes predicted in this study comprise a third clustering of RNAs with levels of expression ranging from 10^{-4} to 10^{-1} r.p.k.m. As a class, only protein-coding genes seem to be enriched in the cytosol, making the nucleus a centre for the accumulation of ncRNAs (Fig. 3). Other gene classes, such as pseudogenes and small annotated ncRNAs, also show subcellular compartmental enrichment (Supplementary Fig. 9).

Higher variability and lower pairwise correlation of expression across all cell lines is consistent with lncRNAs contributing more to cell-line specificity than protein-coding genes. Indeed, a considerable fraction (29%) of all expressed lncRNAs are detected in only one of the cell lines studied when considering the whole cell polyadenylated RNAs, whereas only 10% were expressed in all cell lines. Conversely, whereas a large fraction (53%) of expressed protein-coding genes were constitutive (expressed in all cell lines), only ~7% were cell-line specific (Supplementary Table 7 and Supplementary Fig. 10).

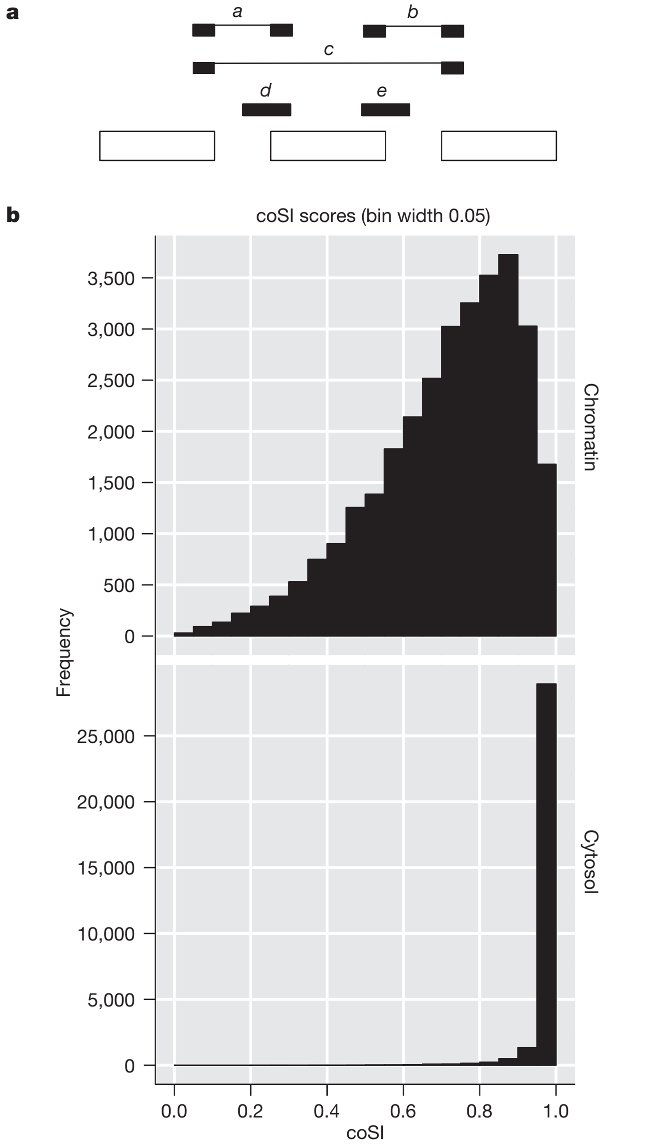


Figure 2 | Co-transcriptional splicing. a, Short read mappings for exon-based splicing completion. Read mappings that allow assessment of splicing completion around exons are shown. Reads providing evidence of splicing completion for the region containing the exon (with either exon inclusion (*a*, *b*) or exclusion (*c*)) are shown. Reads providing evidence for the splicing of the region containing the exon not being completed yet are indicated by *d* and *e*. The complete splicing index (coSI) is the ratio of (0.5(*a* + *b*) + *c*) over (0.5(*a* + *b*) + *c* + 0.5(*d* + *e*)) and can thus be broadly assumed to correspond to the fraction of RNA molecules in which the region containing the exon has already been spliced (see ref. 16). A coSI value of 1 means splicing completed, whereas a value of 0 indicates that splicing has not yet been initiated. **b,** Distribution of coSI scores computed on GENCODE internal exons. Top: distribution in the total chromatin RNA fraction. Bottom: distribution in cytosolic polyadenylated RNA fraction.

Patterns of splicing

The analysis of the expression of alternative isoforms resulted in several observations. First, isoform expression does not seem to follow a minimalistic strategy. Genes tend to express many isoforms simultaneously, and as the number of annotated isoforms per gene grows, so does the number of expressed isoforms (Fig. 4a). The increase, however, is not linear and seems to plateau at about 10–12 expressed

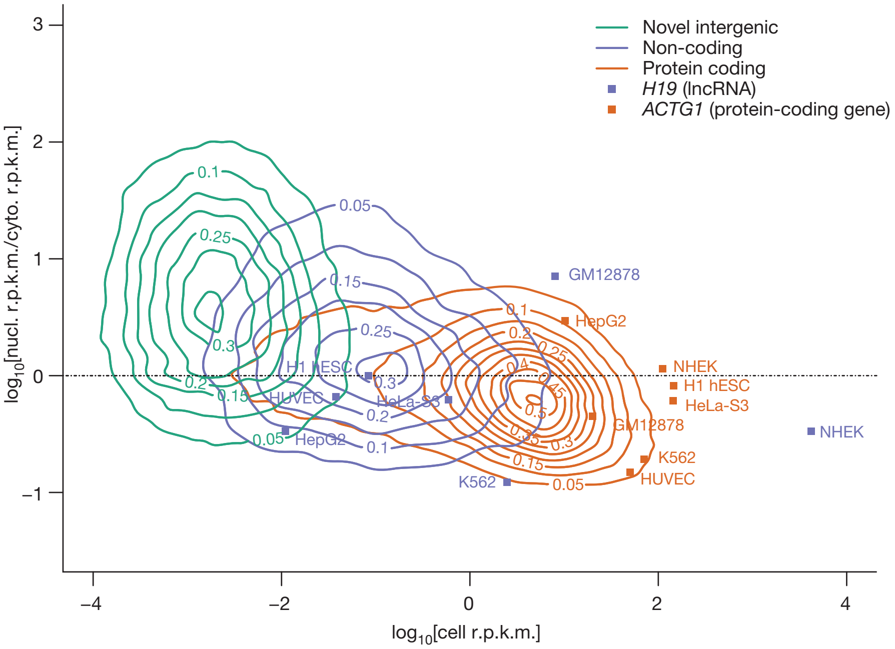


Figure 3 | Abundance of gene types in cellular compartments. Two-dimensional kernel density plots of nuclear over cytosolic enrichment (*y* axis) versus overall gene expression in the whole cell extract (*x* axis), for protein coding, long non-coding and novel genes over all cell lines. Only genes present in all three RNA extracts are displayed, as well as two representative genes (*ACTG1* in red and *H19* in blue), for which the expression in each individual cell line is shown. The actual values of the estimated kernel density are indicated by contour lines and colour shades.

isoforms per gene. However, we cannot obviously distinguish whether this is the result of multiple isoforms expressed in the same cell or of different isoforms expressed in different cells within the interrogated population. Second, alternative isoforms within a gene are not expressed at similar levels, and one isoform dominates in a given condition—usually capturing a large fraction of the total gene expression (at least 30%, even for genes with many isoforms; Fig. 4b). Third, about three-quarters of protein-coding genes have at least two different dominant/major isoforms depending on the cell line (Supplementary Fig. 11a). Fourth, the number of major isoforms per gene grows with the number of annotated isoforms; indeed, the proportion of genes with *n* isoforms that express only one major isoform is strikingly proportional to 1/*n* (Supplementary Fig. 11b). Fifth, variability of gene expression contributes more than variability of splicing ratios to the variability of transcript abundances across cell lines (Supplementary Information).

Alternative transcription initiation and termination

On the basis of RNA-seq analysis of polyadenylated RNAs, a total of 128,021 TSSs were detected across all cell lines, of which 97,778 were previously annotated and 30,243 were novel intergenic/antisense TSSs (Supplementary Table 3a). CAGE tags, filtered by a hidden Markov model (HMM)-based algorithm to differentiate between 5′ capped termini of polymerase II transcripts and recapping events[19] (Supplementary Information), identified a total of 82,783 non-redundant TSSs (Supplementary Table 8). Approximately 48% of the CAGE-identified TSSs are located within 500 base pairs (bp) of an annotated RNA-seq-detected GENCODE TSS, whereas an additional 3% are within 500 bp of a novel TSS (Supplementary Fig. 12). Notably, only ~72% of all CAGE sequencing reads map to TSSs, indicating that the remaining 30% may originate from recapping events or from a new class of TSS.

Using data collected within the ENCODE consortium[20], we carried out a comparison of the GENCODE/RNA-seq and CAGE-determined TSSs and correlated them to chromatin and DNA features characteristic of initiation of transcription, such as DNase hypersensitivity[21], chromatin modification and DNA binding elements[22,23]. All GENCODE/RNA-seq-determined TSSs were examined in each of the cell lines (Supplementary Fig. 13, column 1). Of these redundant positions, 44.7% (199,146) of the RNA-seq-supported TSSs also displayed

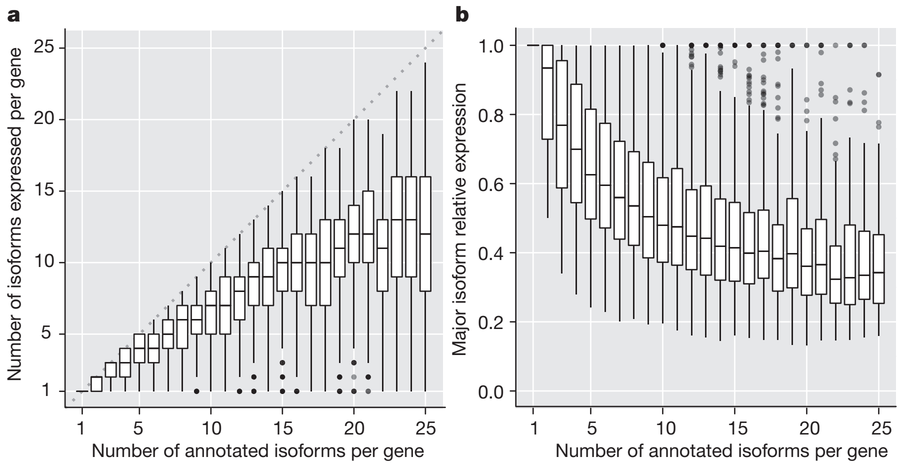


Figure 4 | Isoform expression within a gene. **a**, Number of expressed isoforms per gene per cell line. Genes tend to express many isoforms simultaneously. **b**, Relative expression of the most abundant isoform per gene per cell line. There is generally one dominant isoform in a given condition. The whiskers are defined as Q1 −1.5 × IQR to Q3 +1.5 × IQR, where IQR is the interquartile range, and Q1 and Q3 the first and third quartile, respectively. Each box plot was constructed using the number of genes with 1, 2, 3, 4, etc. up to 25 isoforms.

evidence of CAGE. Approximately half of these TSS positions are associated with at least one of the other characteristic features of transcription initiation (DNase I, H3K27ac and H3K4me3 chromatin modifications). Thus, only a small minority of the TSSs identified by either CAGE or RNA-seq/GENCODE displayed all of the characteristics of the start of transcription (presence of DNase I, H3K4me3, H3K27ac sites and either TAF1 or TBP binding). This is consistent with the possibility that regulatory regions proximal to TSSs are of more than one type.

At the 3′ end, a total of 128,824 sites mapping within annotated GENCODE transcripts were identified as potential sites of polyadenylation after trimming unmapped RNA-seq reads with long terminal polyadenine stretches[24]. About 20% of these mapped proximal to annotated polyadenylation sites (PAS) whereas the remaining 80% correspond to novel PAS of annotated genes, raising the average number of PAS per gene from 1.1 to 2.5. Generally, we observed a cell-type preference for proximal PAS (closest to the annotated stop codon) in the cytosol compared to the nucleus (Supplementary Information).

Short RNA expression landscape

Annotated small RNAs

Currently, a total of 7,053 small RNAs are annotated by GENCODE, 85% of which correspond to four major classes: small nuclear (sn)RNAs, small nucleolar (sno)RNAs, micro (mi)RNAs and transfer (t)RNAs (Table 2a). Overall we find 28% of all annotated small RNAs to be expressed in at least one cell line (Table 2a). The distribution of annotated small RNAs differs markedly between cytosolic and nuclear compartments (Supplementary Fig. 14a). We found that the small RNA classes were enriched in those compartments where they are known to perform their functions: miRNAs and tRNAs in the cytosol, and snoRNAs in the nucleus. Interestingly, snRNAs were equally abundant in both the nucleus and the cytosol. When specifically interrogating the subnuclear compartments of the K562 cell line, however, snRNAs seem to be present in very high abundance in the chromatin-associated RNA fraction (Supplementary Fig. 14b, c). This striking enrichment is consistent with splicing being predominantly co-transcriptional[16,25].

Unannotated short RNAs

We detected two types of unannotated short RNAs. The first type corresponds to subfragments of annotated small RNAs. Because we performed 36-nucleotide end-sequencing of the small RNA fraction, we expected RNA-seq reads to map to the 5′ end of the small RNAs. Supplementary Figure 15 shows the mapping profile of reads along small RNA genes. In both the nuclear and cytosolic compartments, we indeed detected accumulation of reads at the start of snoRNAs and at the guide and passenger sequences of annotated miRNAs. For snRNAs, however, we observed three prominent peaks: the expected one at the 5′ end and two smaller ones at the middle and at the 3′ end of the gene, indicating fragmentation of some snRNAs. Finally, tRNAs seem not to have any prominent sets of 5′ end fragments present at levels greater than what is seen at the annotated 5′ termini. Whereas subfragments of mature tRNAs have been reported previously, these reports were confined to distinct alleles of only a few tRNA genes[26–28].

The second and largest source of unannotated short RNAs corresponds to novel short RNAs (Table 2b) that map outside of annotated ones. Almost 90% of these are only observed in one cell line and are present at low copy numbers. Nearly 40% of these unannotated short RNAs are associated with promoter and terminator regions of annotated genes (promoter-associated short RNAs (PASRs) and termini-associated short RNAs (TASRs)), and their position relative to TSSs and transcription termination sites is similar to previous results[4].

Genealogy of short RNAs

Genome wide, 27% of annotated small RNAs reside within 8% of protein-coding and 5% within 3% of lncRNA genes (Supplementary Fig. 16). Overall, about 6% of all annotated long transcripts overlap with small RNAs and are probably precursors to these small RNAs. Although most of these small RNAs reside in introns, when controlling for relative exon/intron length, we found that exons from lncRNAs are comparatively enriched as hosts for snoRNAs (Supplementary Fig. 17a). Additionally, 8.4% of GENCODE annotated small RNAs map within novel intergenic transcripts, with most overlapping annotated tRNAs. The enrichment for tRNAs was mostly in novel intergenic transcripts derived from non-polyadenylated RNAs (Supplementary Fig. 17b). Many long RNAs, both novel and annotated, thus seem to have dual roles, as functional (protein coding) RNAs, and as precursors for many important classes of small RNAs. Using RNA-seq data from the K562 cell line, we investigated the preferential cellular localization of these RNA precursors (Supplementary Fig. 18). For mature miRNAs and tRNAs (cytosolic enrichment), the potential RNA precursors, identified as RNA-seq contigs overlapping the small RNAs, were detected to be predominantly nuclear (Supplementary Fig. 18a, d). Notably, whereas mature snRNAs were both nuclear and cytosolic, the overlapping long RNAs were observed to be primarily nuclear (Supplementary Fig. 18c). Finally, for snoRNAs (nuclear enrichment), potential long RNA precursors were decidedly observed to be both nuclear and cytosolic (Supplementary Fig. 18b). Unannotated short RNAs were found overall not to be enriched in either the nuclear or cytosolic compartment (Supplementary Fig. 18e).

RNA editing and allele-specific expression

The sequence of transcripts can differ from the underlying genomic sequence as the result of post-transcriptional editing. We developed a pipeline to filter sequencing artefacts and identify genes that are RNA edited[29]. Focusing first on GM12878, a cell line that has been deeply re-sequenced, we find a total 51,557 RNA consistent single nucleotide variants (SNVs) within genic boundaries, 65% of which are present in dbSNP. Of the remainder, 1,186 SNVs in 430 genes (Supplementary Fig. 19a) survive our most stringent filters and 88% of these are candidate adenosine to inosine A>G(I) changes. Notably, the next highest frequency of SNVs is for T>C (5%) and these occur primarily in regions with detectable antisense transcription[29]. We find similar A>G(I) frequencies of 75–84% in seven additional cell lines (Supplementary Fig. 19b). The remaining non-canonical edits amount to very few events in each cell line and are relatively evenly distributed (G>A is the third highest). These results do not support a recent report of a substantial number of non-canonical SNV edits in the RNA of human lymphoblastoid cells[30].

Using the AlleleSeq pipeline[31] on the SNPs in the GM12878 genome, we found that approximately 18% of both GENCODE annotated protein-coding and long non-coding genes exhibit allele-specific expression. The proportion of genes with allele-specific expression was similar in the three investigated RNA fractions (whole-cell, cytoplasm and nucleus; Supplementary Table 9 and Supplementary Information).

Repeat region transcription

About 18% (14,828) of CAGE-defined TSS regions overlap repetitive elements. More precisely, we find 322, 315, 507 and 1,262 intergenic CAGE clusters overlapping long interspersed element (LINE), short interspersed element (SINE), long terminal repeat (LTR) and other repeat elements, respectively (see Supplementary Information). Measuring Shannon entropy across cell lines, we found that CAGE clusters mapping to repeat regions were noticeably more narrowly expressed than CAGE clusters mapping within genic regions (Supplementary Fig. 20a). We represented the correlation of levels of expression compared to cell types as heat maps drawn separately for each of the three repeat element families (LINE, SINE and LTR) (Supplementary Fig. 20b–d). Although a large proportion of the transcripts in the human genome is thought to be initiated from repetitive elements (especially retrotransposon elements[32]), these data clearly point to cell-line specificity as the main characteristic of transcripts emanating from repeat regions.

Characterization of enhancer RNA

It has recently been reported that RNA polymerase II binds some distal enhancer regions and can produce enhancer-associated transcripts named eRNA[33–35]. We used our RNA assays to detect and characterize transcriptional activity at enhancer loci predicted genome-wide from ENCODE chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) data[20,36].

Figure 5a shows the aggregate pattern of RNA-seq and CAGE signal in a strand-specific manner around the subset of predicted gene-distal enhancers containing DNase I hypersensitive sites and centred on those sites. In these plots, as denoted by the accumulation of CAGE tags signifying TSSs, transcription initiation within the enhancer region is observed, and continues outwards for several kilobases (kb). This behaviour can be observed for the polyadenylated and non-polyadenylated RNA fractions mapping in both intronic and intergenic regions. As previously reported[33], we observe a large diversity of expression levels at each of the transcribed enhancers. Polyadenylated to non-polyadenylated RNA ratios, as well as nuclear to cytoplasmic ratios, vary at individual enhancers (Supplementary Fig. 21a, b). However, contrary to some previous reports, although most eRNAs are prevalent in the nuclear non-polyadenylated RNA fraction, some eRNAs seemed to be polyadenylated in the nucleus. This pattern was significantly different compared to transcripts from GENCODE annotated and novel predicted[20] promoters (Fig. 5b).

Transcribed enhancers on average show a significantly different pattern of chromatin modification than non-transcribed ones[37–40]. The enhancer regions displayed stronger signals for H3K4 methylation, H3K27 acetylation and H3K79 dimethylation along with higher levels of RNA polymerase II binding, all associated with

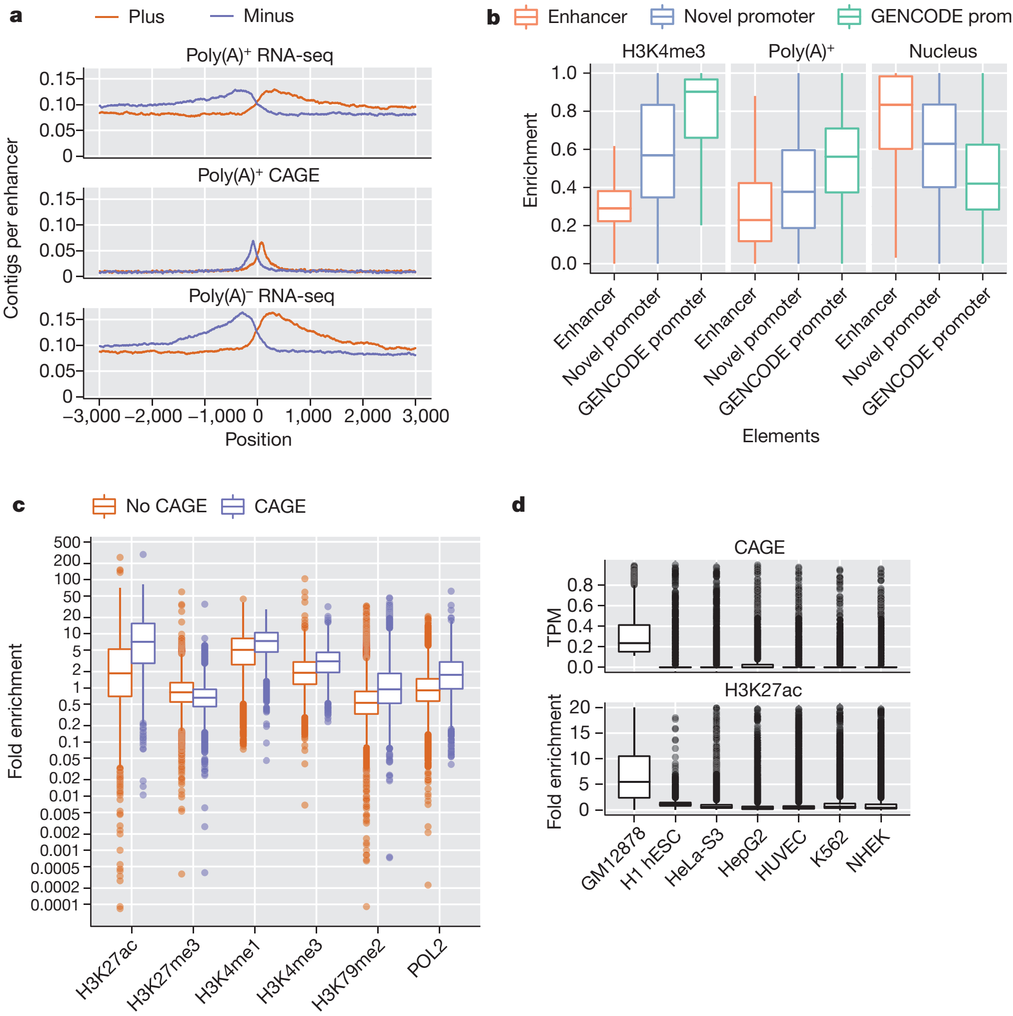


Figure 5 | Transcription at enhancers. **a**, The pattern of RNA elements around enhancer predictions[20,36] containing DNase I hypersensitive sites. The lines represent the average frequency of RNA elements (top, polyadenylated long RNA contigs; middle, CAGE tag clusters; bottom, non-polyadenylated long RNA contigs) in a genomic window around the centre of the enhancer prediction as determined by DNase I hypersensitive sites. Elements on the plus strand are shown in red, and on the minus strand in blue. **b**, Enhancer transcripts differ from promoter transcripts. The box plots compare the features of transcripts at predicted enhancer loci compared to predicted novel intergenic promoters[20] and annotated promoters[8]. H3K4me3, poly(A)$^+$ and nucleus denote the three following ratios: H3K4me3/(H3K4me3 + H3K4me1), polyadenylated/(polyadenylated + non-polyadenylated), nuclear/(nuclear + cytosolic). Enhancers are marked by higher levels of H3K4me1 compared to H3K4me3 than novel or annotated promoters (left). Enhancer transcripts show higher levels of non-polyadenylated (middle) and nuclear (right) RNA relative to promoters. **c**, Chromatin state at transcribed enhancers. Enhancer predictions with evidence of transcription (in blue; Cage tags present at predicted locus) show a different pattern of histone modification and higher levels of RNA polymerase II binding than non-transcribed predictions (red). They are enriched for H3K27 acetylation, H3K4 methylation, H3K79 dimethylation and depleted for H3K27 trimethylation. **d**, Enhancer activity and transcription is cell-type specific. Loci predicted to be active transcribed enhancers in GM12878 cells show low signal for CAGE tags (top) and for H3K27 acetylation (bottom) in other cell lines. The whiskers are defined as Q1 −1.5 × IQR to Q3 +1.5 × IQR, where IQR is the interquartile range, and Q1 and Q3 the first and third quartile, respectively.

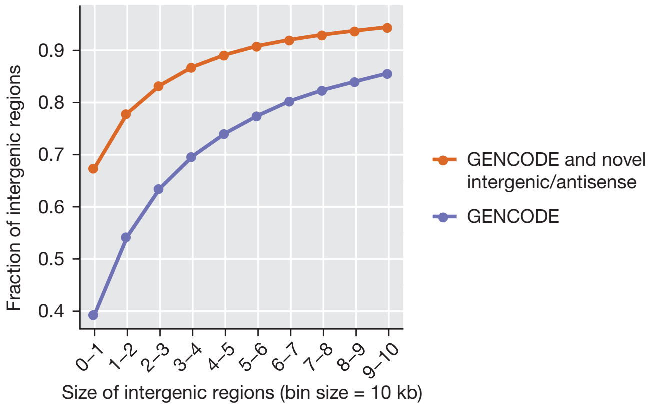


Figure 6 | Size distribution of intergenic regions. Novel genes increase the proportion of small intergenic regions.

transcriptional initiation and elongation (Fig. 5c). Both the transcripts and the chromatin states are cell-type specific (Fig. 5d). Taking the GM12878 cell line as an example, the enhancer loci producing eRNA demonstrate enrichment of CAGE tag detection (Fig. 5d, top) and the presence of H3K27ac histone modification (Fig. 5d, bottom) in this cell line compared to five other analysed cell lines. This strongly suggests that the regulatory regions governing the expression of enhancer transcripts are distinguished from regulatory regions located at the beginning of genic regions.

Concluding remarks

The cumulative coverage of transcribed regions in the 15 cell lines across the human genome is 62.1% and 74.7% for processed and primary transcripts, respectively (Supplementary Table 10 and Supplementary Fig. 22). On average, for each cell line, 39% of the genome is covered by primary transcripts and 22% by processed RNAs. No cell line showed transcription of more than 56.7% of the union of the expressed transcriptomes across all cell lines. When mapping the current RNA-seq data to the ENCODE pilot regions (Supplementary Table 10), we observed a similar, albeit higher, extent of transcriptional coverage of 73.3% for processed RNAs and 84.5% for primary transcripts. Previously reported estimates in these regions for processed and primary transcripts were 24% and 93%, respectively (Supplementary Table 2.4.3 and ref. 3). The increased genome coverage by processed RNAs stems largely from the inclusion of

Table 1 | Long polyadenylated and non-polyadenylated RNAs

Expression of GENCODE (v7) annotated elements (a)

Gene type	Detected exons† (annotation no.)	Detected splice junctions† (annotation no.)	Detected transcripts† (annotation no.)	Detected genes† (annotation no.)	Exon nucleotide coverage‡ (%)	Number of genes expressed in at least one cell line	Number of genes expressed in only one cell line	Proportion over genes expressed§ (%)	Number of genes expressed in 14 cell lines	Proportion over genes expressed‖ (%)
Long non-coding	22,381 (41,467)	8,017 (26,872)	6,521 (14,880)	5,906 (9,277)	87.5	5,906	1,386	23.5	631	10.7
Protein coding	288,322 (318,514)	194,752 (244,158)	59,822 (76,006)	18,939 (20,679)	98.1	18,939	1,082	5.7	10,571	55.8
Other*	102,000 (133,937)	19,277 (47,663)	45,410 (71,113)	10,649 (21,750)	95.2	10,649	2,453	23.0	1,896	17.8
Total annotated	412,703 (493,918)	222,046 (318,693)	111,753 (161,999)	35,494 (51,706)	96.7	35,394	4,921	13.9	13,098	37.0

Expression of GENCODE (v7) intergenic and antisense elements (b)

Category	Detected exons†	Detected splice junction†	Detected transcripts†	Detected genes†
Mono-exonic	55,683	NA	55,682	33,686
Multi-exonic	39,117	69,052	17,643	7,518
Total	94,800	69,052	73,325	41,204

NA, not applicable.
* Includes pseudogenes, miRNAs, etc.
† All elements that passed npIDR (0.1).
‡ Cumulative detected nucleotide in detected exons/total nucleotides in detected exons.
§ Proportion for genes expressed in only one cell line.
‖ Proportion for genes expressed in 14 cell lines.

Table 2 | Short RNAs

Expression of GENCODE (v7) annotated small RNA genes (a)

Gene type*	GENCODE total	Detected genes (% detected)	No. genes expressed in only one cell line (% detected)	No. genes expressed in 12 cell lines (% detected)	miRNA guide fragment‡	miRNA passenger fragment§	Internal fragments‖ of annotated small RNA (average per detected gene)
miRNA	1,756	497 (28)	59 (12)	147 (30)	454 (454)	175 (175)	18
snoRNA	1,521	458 (30)	73 (16)	223 (49)	NA	NA	60
snRNA	1,944	378 (19)	123 (33)	41 (11)	NA	NA	36
tRNA	624	465 (75)	29 (6)	197 (42)	NA	NA	52
Other†	1,209	191 (16)	69 (36)	24 (13)	NA	NA	32
Total GENCODE	7,054	1,989 (28)	353 (18)	632 (32)	NA	NA	40

Expression of unannotated short RNAs (b)

Cell compartment	Unannotated short RNAs	Exonic	Intronic	Exon–intron boundaries	Genic	Gene–intergene boundaries	Intergenic
Cell	57,393	14,116	13,773	1,818	29,707	13,048	25,906
Nucleus	82,297	19,334	40,136	5,248	64,718	7,417	16,289
Cytosol	25,455	6,183	5,605	665	12,453	6,631	12,447
Three compartments	150,165	38,969	55,061	7,552	101,582	23,185	45,081

NA, not applicable.
* Includes all other GENCODE small transcript biotypes except for pseudogenes.
† All elements that have passed npIDR (0.1).
‡ Number of detected miRNAs with an expressed annotated guide (with an annotated guide in mirbase).
§ Number of detected miRNAs with an expressed annotated passenger (with an annotated passenger in mirbase).
‖ Short RNA-seq mapping for which the 5′ end starts 5 bp after the start and ends 5 bp before the end of a detected gene.

non-polyadenylated RNAs in the current study. Other than that, given the differences in the samples studied, the selection of pilot regions with high genic content, the increase of annotated genomic regions over time, and the different technologies used to interrogate transcription, both estimates are in reasonable agreement.

As a consequence of both the expansion of genic regions by the discovery of new isoforms and the identification of novel intergenic transcripts, there has been a marked increase in the number of intergenic regions (from 32,481 to 60,250) due to their fragmentation and a decrease in their lengths (from 14,170 bp to 3,949 bp median length; Fig. 6). Concordantly, we observed an increased overlap of genic regions. As the determination of genic regions is currently defined by the cumulative lengths of the isoforms and their genetic association to phenotypic characteristics, the likely continued reduction in the lengths of intergenic regions will steadily lead to the overlap of most genes previously assumed to be distinct genetic loci. This supports and is consistent with earlier observations of a highly interleaved transcribed genome[12], but more importantly, prompts the reconsideration of the definition of a gene. As this is a consistent characteristic of annotated genomes, we would propose that the transcript be considered as the basic atomic unit of inheritance. Concomitantly, the term gene would then denote a higher-order concept intended to capture all those transcripts (eventually divorced from their genomic locations) that contribute to a given phenotypic trait. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

METHODS SUMMARY

For full details of Methods, see Supplementary Information.

Received 10 December 2011; accepted 15 May 2012.

1. Mattick, J. S. Long noncoding RNAs in cell and developmental biology. *Semin. Cell Dev. Biol.* **22,** 327 (2011).
2. The ENCODE Project Consortium.. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
3. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
4. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
5. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8,** 413–423 (2007).
6. Coffey, A. J. *et al.* The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.* **19,** 827–831 (2011).
7. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (suppl. 1), 1–9 (2006).
8. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* (in the press).
9. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature Methods* **3,** 211–222 (2006).
10. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2,** 105–111 (2005).
11. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).
12. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308,** 1149–1154 (2005).
13. Katinakis, P. K., Slater, A. & Burdon, R. H. Non-polyadenylated mRNAs from eukaryotes. *FEBS Lett.* **116,** 1–7 (1980).
14. Milcarek, C., Price, R. & Penman, S. The metabolism of a poly(A) minus mRNA fraction in HeLa cells. *Cell* **3,** 1–10 (1974).
15. Salditt-Georgieff, M., Harpold, M. M., Wilson, M. C. & Darnell, J. E. Jr. Large heterogeneous nuclear ribonucleic acid has three times as many 5′ caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Mol. Cell. Biol.* **1,** 179–187 (1981).
16. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* (in the press).
17. Tilgner, H. *et al.* Genomic analysis of ENCODE data reveals widespread links between epigenetic chromatin marks and alternative splicing. *Genome Res.* (in the press).
18. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
19. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457,** 1028–1032 (2009).
20. ENCODE Project Consortium.. An integrated encyclopaedia of DNA elements in the human genome. *Nature* http://dx.doi.org/10.1038/nature11247 (this issue).
21. Thurman, R. E. The accessible chromatin landscape of the human genome. *Nature* http://dx.doi.org/10.1038/nature11232 (this issue).
22. Gerstein, M. B. Architecture of the human regulatory network derived from ENCODE data. *Nature* http://dx.doi.org/10.1038/nature11245 (this issue).
23. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* (in the press).
24. Fu, Y. *et al.* Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* **21,** 741–747 (2011).
25. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Struct. Mol. Biol.* **18,** 1435–1440 (2011).
26. Cole, C. *et al.* Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15,** 2147–2160 (2009).
27. Kawaji, H. *et al.* Hidden layers of human small RNAs. *BMC Genom.* **9,** 157 (2008).
28. Lee, Y. S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* **23,** 2639–2649 (2009).
29. Park, E., Williams, B., Wold, B. & Mortazavi, A. A Survey of RNA Editing in the human ENCODE RNA-seq data (GRCP043). *Genome Res.* (in the press).
30. Li, M. *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333,** 53–58 (2011).
31. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
32. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature Genet.* **41,** 563–571 (2009).
33. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
34. Ren, B. Transcription: Enhancers make non-coding RNA. *Nature* **465,** 173–174 (2010).
35. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–394 (2011).
36. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biol.* (in the press).
37. Hoffman, M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Genome Res.* (in the press).
38. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type specific transcription factor binding. *Genome Res.* (in the press).
39. Kundaje, A. Ubiquitous heterogeneity and asymmetry of the chromatin landscape at transcription regulatory elements. *Genome Res.* (in the press).
40. Miller, B. Pre-programming of chromatin structure across the cell cycle. *Genome Res.* (in the press).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Human Genome Research Institute (NHGRI) production grants U54HG004557, U54HG004555, U54HG004576 and U54HG004558, and by the NHGRI pilot grant R01HG003700. It was also supported by the NHGRI ARRA stimulus grant 1RC2HG005591, the National Science Foundation (SNF) grant 127375, the European Research Council (ERC) grant 249968, a research grant for the RIKEN Omics Science Center from the Japanese Ministry of Education, Culture, Sports, Science and Technology, and grants BIO2011-26205, CSD2007-00050 and INB GNV-1 from the Spanish Ministry of Science. We would also like to thank C. Gunter and W. Spitzer for editorial assistance with the manuscript.

Author Contributions T.R.G., R.G., P.C., B.W., Y.R., M.C.G., G.H., S.E.A., A.R., T.H., M.G. and Y.H. led the project and oversaw the analysis. C.A.D., X.R., B.A.W. and P.C. oversaw or significantly contributed to data generation. S.D., A.Me., A.D., T.L., A.Mo., A.T., J.L., W.L., F.S., C.X., G.K.M., J.K., C.Z., J.R., M.R., F.K. and J.H. made major contributions towards data processing and analysis. R.F.A., T.A., I.A., M.T.B., N.S.B., P.B., K.B., I.B., S.C., X.C., J.Ch., J.Cu., T.D., J.Dr., E.D., J.Du., R.D., E.F., M.F., K.F-T., P.F., S.F., M.J.F., H.Ga., D.G., A.G., H.Gu., C.H., S.J., R.J., P.K., B.K., C.K., O.J.L., E.P., K.P., J.B.P., P.R., B.R., D.R., M.S., L.S., L.-H.S., A.S., J.S., A.M.S., H.Ta., H.Ti., D.T., N.W., H.W., J.W. and Y.Y. were responsible for data production and analysis. T.R.G. and R.G. wrote the manuscript with input from all authors.

Author Information A complete set of data files can be downloaded at GEO under accession codes GSE26284 (CSHL, long RNA), GSE33480 (Caltech, A+ RNA-seq), GSE24565 (CSHL, short RNA), GSE33600 (GIS, RNA-PET) and GSE34448 (RIKEN, CAGE) or are viewable at the UCSC Genome Browser (http://genome-preview.ucsc.edu/ENCODE/). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R.G. (gingeras@cshl.edu) or R.G. (roderic.guigo@crg.eu).

D

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Originally published as:

Landt SG*, Marinov GK*, Kundaje A*, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, Desalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**(9):1813–1831. doi: 10.1101/gr.136184.111.

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] Florencia Pauli,[5] Serafim Batzoglou,[3] Bradley E. Bernstein,[6] Peter Bickel,[7] James B. Brown,[7] Philip Cayting,[1] Yiwen Chen,[8] Gilberto DeSalvo,[2] Charles Epstein,[6] Katherine I. Fisher-Aylor,[2] Ghia Euskirchen,[1] Mark Gerstein,[9] Jason Gertz,[5] Alexander J. Hartemink,[10] Michael M. Hoffman,[11] Vishwanath R. Iyer,[12] Youngsook L. Jung,[13,14] Subhradip Karmakar,[15] Manolis Kellis,[4] Peter V. Kharchenko,[12] Qunhua Li,[16] Tao Liu,[8] X. Shirley Liu,[8] Lijia Ma,[15] Aleksandar Milosavljevic,[17] Richard M. Myers,[5] Peter J. Park,[13,14] Michael J. Pazin,[18] Marc D. Perry,[19] Debasish Raha,[20] Timothy E. Reddy,[5,27] Joel Rozowsky,[9] Noam Shoresh,[6] Arend Sidow,[1,21] Matthew Slattery,[15] John A. Stamatoyannopoulos,[11,22] Michael Y. Tolstorukov,[13,14] Kevin P. White,[15] Simon Xi,[23] Peggy J. Farnham,[24,28] Jason D. Lieb,[25,28] Barbara J. Wold,[2,28] and Michael Snyder[1,28]

[1–25] *[Author affiliations appear at the end of the paper.]*

Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of transcription-factor binding and histone modifications in living cells. Despite its widespread use, there are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use. These practices affect the quality and utility of any global ChIP experiment. Through our experience in performing ChIP-seq experiments, the ENCODE and modENCODE consortia have developed a set of working standards and guidelines for ChIP experiments that are updated routinely. The current guidelines address antibody validation, experimental replication, sequencing depth, data and metadata reporting, and data quality assessment. We discuss how ChIP quality, assessed in these ways, affects different uses of ChIP-seq data. All data sets used in the analysis have been deposited for public viewing and downloading at the ENCODE (http://encodeproject.org/ENCODE/) and modENCODE (http://www.modencode.org/) portals.

[Supplemental material is available for this article.]

Methods for mapping transcription-factor occupancy across the genome by chromatin immunoprecipitation (ChIP) were developed more than a decade ago (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak and Snyder 2002; Weinmann et al. 2002). In ChIP assays, a transcription factor, cofactor, or other chromatin protein of interest is enriched by immunoprecipitation from cross-linked cells, along with its associated DNA. Genomic DNA sites enriched in this manner were initially identified by DNA hybridization to a microarray (ChIP-chip) (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak and Snyder 2002; Weinmann et al. 2002), and more recently by DNA sequencing (ChIP-seq) (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007). ChIP-seq has now been widely used for many transcription factors, histone modifications, chromatin modifying complexes, and other chromatin-associated proteins in a wide variety of organisms. There is, however, much diversity in the way ChIP-seq experiments are designed, executed, scored, and reported. The resulting variability and data quality issues affect not only primary measurements, but also the ability to compare data from multiple studies or to perform integrative analyses across multiple data-types.

The ENCODE and modENCODE consortia have performed more than a thousand individual ChIP-seq experiments for more than 140 different factors and histone modifications in more than 100 cell types in four different organisms (*D. melanogaster*, *C. elegans*, mouse, and human), using multiple independent data production and processing pipelines (The ENCODE Project Consortium 2004, 2011; Celniker et al. 2009). During this work, we developed guidelines, practices, and quality metrics that are applied to all ChIP-seq work done by the Consortium (Park 2009). Here we describe these, together with supporting data and illus-

[26] These authors contributed equally to this work.
[27] Present address: Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA.
[28] Corresponding authors

E-mail mpsnyder@stanford.edu
E-mail woldb@caltech.edu
E-mail jlieb@bio.unc.edu
E-mail pfarnham@usc.edu

Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.136184.111. Freely available online through the *Genome Research* Open Access option.

trative examples. We emphasize issues common to all ChIP-seq studies: immunoprecipitation specificity and quality, impact of DNA sequencing depth, scoring and evaluation of data sets, appropriate control experiments, biological replication, and data reporting.

ChIP overview

The goals of a genome-wide ChIP experiment are to map the binding sites of a target protein with maximal signal-to-noise ratio and completeness across the genome. The basic ChIP-seq procedure is outlined in Figure 1A, and detailed protocols (and data) from our two consortia can be obtained from the ENCODE and modENCODE production groups listed at the UCSC Genome Browser: http://encodeproject.org/ENCODE/ and http://www.modencode.org/, respectively. Cells or tissues are treated with a chemical agent, usually formaldehyde, to cross-link proteins covalently to DNA. This is followed by cell disruption and sonication, or in some cases, enzymatic digestion, to shear the chromatin to a target size of 100–300 bp (Ren et al. 2000; Iyer et al. 2001). The protein of interest (transcription factor, modified histone, RNA polymerase, etc.) with its bound DNA is then enriched relative to the starting chromatin by purification with an antibody specific for the factor. Alternatively, cell lines expressing an epitope-tagged factor can be generated and the fusion protein immunoprecipitated via the epitope tag.

After immuno-enrichment, cross-links are reversed, and the enriched DNA is purified and prepared for analysis. In ChIP-chip, the DNA is fluorescently labeled and hybridized to a DNA microarray, along with differentially labeled reference DNA (Ren et al. 2000; Iyer et al. 2001). In ChIP-seq, the DNA is analyzed by high-throughput DNA sequencing. The ENCODE Consortium chose ChIP-seq for human and mouse experiments because it permits comprehensive coverage of large genomes and increases site resolution (Johnson et al. 2007; Robertson et al. 2007). For organisms with small genomes, the modENCODE Consortium has used both ChIP-chip and ChIP-seq, as modern arrays can provide high-resolution coverage of small genomes (Gerstein et al. 2010; Roy et al. 2010). In all formats, we identified putatively enriched genomic regions by comparing ChIP signals in the experimental sample with a similarly processed reference sample prepared from appropriate control chromatin or a control immunoprecipitation.

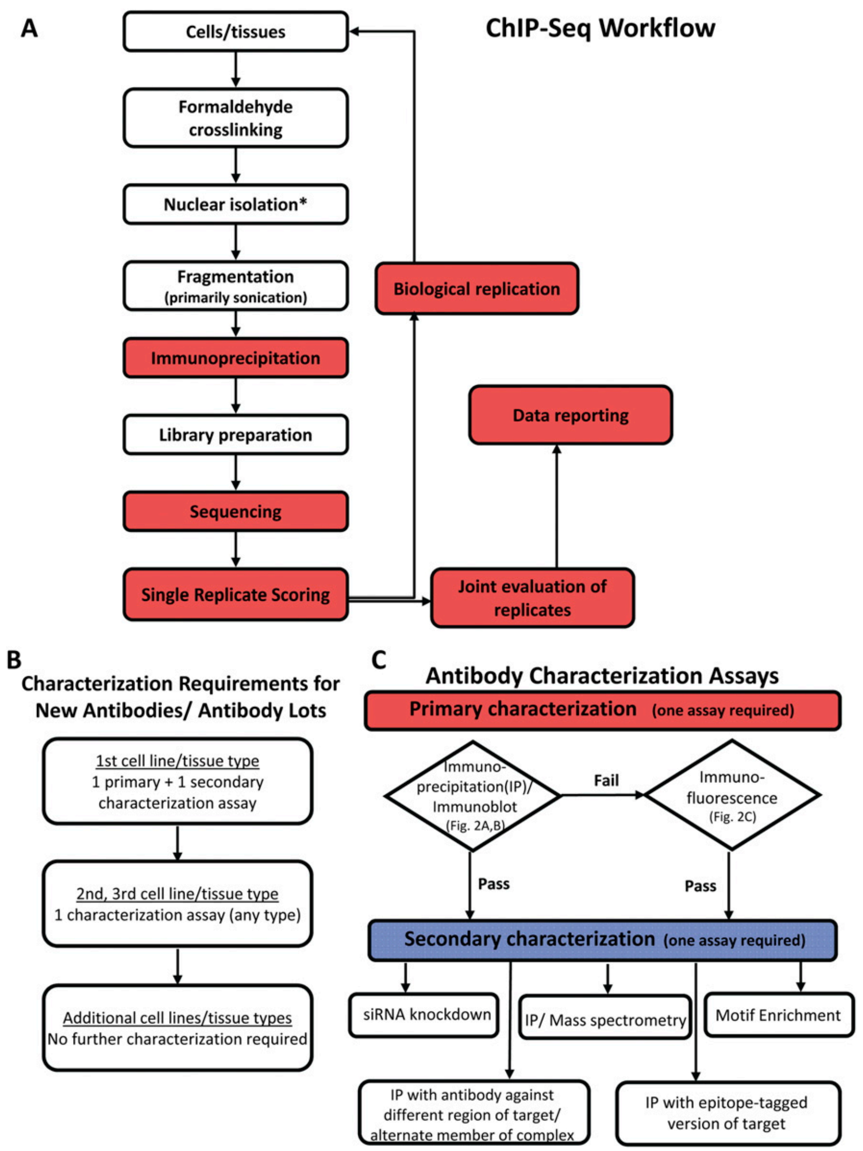

A
ChIP-Seq Workflow
Cells/tissues
Formaldehyde crosslinking
Nuclear isolation*
Fragmentation (primarily sonication)
Immunoprecipitation
Library preparation
Sequencing
Single Replicate Scoring
Biological replication
Data reporting
Joint evaluation of replicates
B
Characterization Requirements for New Antibodies/ Antibody Lots
1st cell line/tissue type
1 primary + 1 secondary characterization assay
2nd, 3rd cell line/tissue type
1 characterization assay (any type)
Additional cell lines/tissue types
No further characterization required
C
Antibody Characterization Assays
Primary characterization (one assay required)
Immuno-precipitation(IP)/ Immunoblot (Fig. 2A,B)
Fail
Immuno-fluorescence (Fig. 2C)
Pass
Pass
Secondary characterization (one assay required)
siRNA knockdown
IP/ Mass spectrometry
Motif Enrichment
IP with antibody against different region of target/ alternate member of complex
IP with epitope-tagged version of target

Figure 1. Overview of ChIP-seq workflow and antibody characterization procedures. (*A*) Steps for which specific ENCODE guidelines are presented in this document are indicated in red. For other steps, standard ENCODE protocols exist that should be validated and optimized for each new cell line/tissue type or sonicator. (*) A commonly used but optional step. (*B*) Flowchart for characterization of new antibodies or antibody lots. (*C*) Flowchart for use of antibody characterization assays.

Different protein classes have distinct modes of interaction with the genome that necessitate different analytical approaches (Pepke et al. 2009):

1. *Point-source* factors and certain chromatin modifications are localized at specific positions that generate highly localized ChIP-seq signals. This class includes most sequence-specific transcription factors, their cofactors, and, with some caveats, transcription start site or enhancer-associated histone marks. These comprise the majority of ENCODE and modENCODE determinations and are therefore the primary focus of this work.
2. *Broad-source* factors are associated with large genomic domains. Examples include certain chromatin marks (H3K9me3, H3K36me3, etc.) and chromatin proteins associated with transcriptional elongation or repression (e.g., ZNF217) (Krig et al. 2007).
3. *Mixed-source* factors can bind in point-source fashion to some locations of

the genome, but form broader domains of binding in others. RNA polymerase II, as well as some chromatin modifying proteins (e.g., SUZ12) behave in this way (Squazzo et al. 2006).

Below, we report our experience with ChIP-seq experimental design, execution, and quality assessment. We offer specific recommendations, based on current experience, as summaries in boxes.

ChIP-seq experimental design considerations

Antibody and immunoprecipitation specificity

The quality of a ChIP experiment is governed by the specificity of the antibody and the degree of enrichment achieved in the affinity precipitation step. The majority of ENCODE/modENCODE ChIP experiments in human cells and in *Drosophila* embryos were performed with antibodies directed against individual factors and histone modifications. A total of 145 polyclonal and 43 monoclonal antibodies had been used to successfully generate ChIP-seq data as of October 2011.

Antibody deficiencies are of two main types: poor reactivity against the intended target and/or cross-reactivity with other DNA-associated proteins. For these reasons, we have developed a set of working standards and reporting guidelines designed to provide measures of confidence that the reagent recognizes the antigen of interest with minimal cross-reactivity toward other chromosomal proteins. Widely accessible methods for measuring antibody specificity and sensitivity range from semiquantitative to qualitative, and each can have noise and interpretation issues. We therefore emphasize reporting of antibody characterization data so that users of the ChIP data, or the reagent itself, can make informed judgments. We also recognize that a successful experiment can be performed with reagents that fail to strictly comply with these guidelines. For example, cross-reacting proteins detected in an immunoblot assay might not interfere in ChIP, because the protein is not attached to chromatin. Secondary tests of diverse types can help to provide confidence concerning the acceptability of an antibody that fails an initial assessment.

Two tests, a primary and a secondary test, are used to characterize each monoclonal antibody or different lots of the same polyclonal antibody. The ordering of the primary and secondary tests are influenced by the effort required to execute each, with the primary assay being easier to perform on large numbers of antibodies. The tests differ for antibodies against transcription factors vs. those against histone modifications. A detailed description of the tests is provided in Box 1, and a typical workflow is presented in Figure 2, B and C. For transcription-factor antigens, we adopted the immunoblot as our primary assay, with immunostaining as the alternative. The former can give more information about cross-reacting material or multiple isoforms; the latter is typically less sensitive, but provides information about nuclear location. Examples of antibodies that pass and fail these tests are shown in Figure 2A.

Our consortia also include one of five criteria as a secondary characterization: (1) factor "knockdown" by mutation or RNAi, (2) independent ChIP experiments using antibodies against more than one epitope on a protein or against different members of the same complex, (3) immunoprecipitation using epitope-tagged constructs, (4) affinity enrichment followed by mass spectrometry, or (5) binding-site motif analysis. Motif enrichment is the easiest assay to perform, but requires pre-existing information about the sequences to which a protein binds and assumes that the motif is uniquely recognized in a given cell source by the factor of interest. ChIP with a second antibody or against an epitope-tagged construct and siRNA experiments coupled with ChIP provide independent evidence that the target sites are bound by the factor of interest. We found that mass spectrometry is particularly useful for cases where multiple or unexpected bands are observed on an immunoblot and the presence of spliced isoforms, post-translational modification, or degradation is suspected. Additionally, it can precisely identify potential alternate sources of ChIP signal, often with novel biological implications, which can be tested by additional ChIP experiments. Due to the significant effort and expense required to perform these assays, our standard for the consortia requires only one secondary assay. We found that ~20% (44 of 227) of the tested commercially available antibodies against transcription factors meet these characterization guidelines and also function in ChIP-seq assays.

To date, 55% of consortia antibodies have been submitted with mass spectrometry data, 28% with ChIP data using a second antibody, epitope tag, or alternate member of a known complex, 10% with data from motif analysis (this standard has only been used by ENCODE for 1 yr), and 7% with siRNA knockdown data. A summary of motif detection for all data sets is in preparation (P Kheradpour and M Kellis, in prep.).

Validating histone modification antibodies involves multiple issues (Egelhofer et al. 2011): (1) specificity with respect to other nuclear/chromatin proteins, (2) specificity with respect to unmodified histones and off-target modified histone residues (e.g., H3K9me vs. H3K27me), (3) specificity with respect to mono-, di-, and trimethylation at the same residue (e.g., H3K9me1, H3K9me2, and H3K9me3), and (4) lot-to-lot variation. For all consortia histone measurements, we set the standard that immunoblot analysis and one of the following secondary criteria are applied: Peptide-binding tests (dot blots), mass spectrometry, immunoreactivity analysis in cell lines containing knockdowns of a relevant histone modification enzyme or mutants histones, or genome annotation enrichment. The details of these standards are in Box 1.

Immunoprecipitation using epitope tagged constructs

Given the challenges in obtaining antibodies for suitable ChIP, an attractive alternative is to tag the factor with an exogenous epitope and immunoprecipitate with a well-characterized monoclonal reagent specific for the tag. Epitope-tagging addresses the problems of antibody variation and cross-reaction with different members of multigene families by using a highly specific reagent that can be used for many different factors. However, this introduces concerns about expression levels and whether tagging alters the activity of the factor. The level of expression is typically addressed by using large clones (usually fosmids and BACs) carrying as much regulatory information as possible to make the level of expression nearly physiological (Poser et al. 2008; Hua et al. 2009). Higher expression is known to result in occupancy of sites not necessarily occupied at physiological levels (DeKoter and Singh 2000; Fernandez et al. 2003). In ENCODE/modENCODE, tagged factors have been used most extensively thus far for *C. elegans* studies, where factors have been tagged with GFP and shown to complement null mutants; six of six tested to date have been found to complement (Zhong et al. 2010; V Reinke, unpubl.). In some cases, information regarding expression is not available and expression from an exogenous promoter has been used (P Farnham, unpubl.)

Box 1: ENCODE guidelines for antibody and immunoprecipitation characterization

Characterization of antibodies directed against transcription factors

Antibodies directed against transcription factors must be characterized using both a primary and secondary characterization; characterizations must be repeated for each new antibody or antibody lot number that is used for ChIP-seq (Fig. 1B,C).

Primary mode of characterization

Antibodies are characterized by one of two primary methods, immunoblot analysis, or immunofluorescence.

Immunoblot analyses

Immunoblot analyses are performed on protein lysates from either whole-cell extracts, nuclear extracts, chromatin preparations, or immunoprecipitated material (before proceeding to ChIP assays, it is helpful to demonstrate that the protein of interest can be efficiently immunoprecipitated from a nuclear extract, see Fig. 2B). We use the guideline that the primary reactive band should contain at least 50% of the signal observed on the blot. Ideally, this band should correspond to the size expected for the protein of interest (Fig. 2A). However, the electrophoretic mobility of many factors can deviate significantly from the expected size due to modifications, isoform differences, or intrinsic properties of the factor. Therefore, antibodies for which the main band differs from the expected size by >20% or for which multiple bands are seen (such that no band represents >50% of the signal) can be used under certain circumstances. In these cases, further criteria must be met, such as (1) the unexpected mobility must have been properly documented in published studies using the same antibody lot, (2) the signal in the band(s) is reduced by siRNA knockdown or mutation, or (3) the factor can be identified in all band(s) by mass spectrometry.

Immunofluorescence

Some antibodies that work well for ChIP do not work well in immunoblots. If immunoblot analysis is not successful, immunofluorescence can be used as an alternative method. Staining should be of the expected pattern (e.g., nuclear and only in cell types or under specific growth conditions that express the factor) (Fig. 2C). Because immunofluorescence does not provide evidence that the antibody detects only one protein, this validation method should be combined with a method that reduces the level of the protein, such as siRNA- or shRNA-mediated knockdown, or used with a knockout cell line or organism (see below).

Secondary mode of characterization

In addition to the primary mode of characterization, the consortia performs at least one of the following five assays as an additional secondary test:

Knockdown or knockout of the target protein

Immunoblots or immunoprecipitations are performed in duplicate using extracts from siRNA or shRNA knockdowns or from knockout mutant cell lines or organisms. We use the guideline that the primary immunoblot (or immunofluorescence) signal, along with additional immunoreactive bands, should be reduced to no more than 30% of the original signal and any signal remaining after genetic mutation, RNAi, or siRNA is noted. As an alternative, knockdown can also be measured with ChIP experiments. ENCODE data can be submitted if reduction of ChIP-chip or ChIP-seq signals by >50% relative to control is observed. A suitable control knockdown (e.g. using "scrambled" siRNA sequences) should also be performed and the data should be submitted; reduction of signal should not be observed in the control knockdown data set. The methodology used for binding-region signal normalization (for instance, normalization against total read counts or using values from reference peaks quantified by qPCR under all experimental conditions) should also be reported.

Immunoprecipitation followed by mass spectrometry

All immunoreactive bands identified by immunoblot analysis are analyzed (Fig. 2D). ENCODE passes such analyses if the protein of interest is identified in such bands; if additional chromosomal proteins are identified in an immunoreactive band, the Consortium accepts the experiment as long as they are present at lower prevalence than the desired protein (as measured by peptide counts or other methods) or can be demonstrated to arise from nonspecific immunoprecipitation (e.g., also present in a control immunoprecipitation). All proteins identified by mass spectrometry and the number of peptide counts for each are reported.

Immunoprecipitation with multiple antibodies against different parts of the target protein or members of the same complex

Different antibodies against different parts of the same protein or other members of a known protein complex can be used in analyzing the specificity of antibodies. In the ENCODE Consortium, results of the different ChIP experiments are compared and significant overlap of enriched loci is expected (ChIP-seq experiments are compared using the IDR-based standards in Box 3). Note that for different proteins that are members of a complex, there may be some functions that are independent of one another. Thus, the targets lists for two different proteins may not entirely overlap. In this case, specific evidence about limited overlap of binding specificity in the literature is presented to justify the significance of the overlap observed between data sets for the factors in question.

Immunoprecipitation with an epitope-tagged version of the protein

An epitope-tagged version of the target protein may be used, preferably expressed from the endogenous gene promoter. ENCODE conducts and analyzes such experiments as described above for the use of multiple antibodies.

Motif enrichment

For transcription factors, if a factor has a well-characterized motif derived from in vitro binding studies or another justifiable method, and if either no paralogs are expressed in the cell lines being analyzed or if the antibody is raised to a unique region of the factor, motif enrichment can be used for validation. Motif analysis can be performed using a defined set of high-quality peaks (a 0.01 IDR threshold is used), and for ENCODE data to be submitted, motifs should be enriched at least fourfold compared with all accessible regions (e.g., DNase hypersensitive regions) and present in >10% of analyzed peaks. Analysis of data sets deposited as of January 2011 identified data sets that meet these standards for 49 of 85 factors (Fig. 2E). We note that due to differences in transcription-factor recruitment mechanisms, failure of a data set to meet the motif enrichment threshold does not necessarily indicate poor quality data.

(*continued*)

Box 1: *Continued*

Other considerations

1. For antibodies directed against members of a multigene family, the best practice is to prepare or obtain antibodies that recognize protein regions unique to individual family members. For an ENCODE validated antibody, any potential cross-reaction is noted when reporting data collected using that antibody.
2. For antibodies that have been previously characterized for one cell type, ENCODE has used only one validation method (such as immunoblot analysis) when the antibody is used to perform ChIP in a new cell type or organism. If an antibody has been validated in at least three different cell types, we do not require further validation for ChIP-seq experiments with additional cell types for ENCODE submission. Similarly, for whole organisms, if the antibody has been characterized in three growth stages, no further characterization is required.
3. If antibodies derived from the same lot are used by different groups in ENCODE, they only need to be characterized once. However, antibodies from different lots of the same catalog number are characterized as if they were new antibodies.

Epitope-tagged proteins

Epitope-tagged factors are introduced into cells by transfection of an expression construct. To help ensure that ChIP-seq results obtained using the tagged factor are comparable to those expected for the endogenous factor, ENCODE uses the criteria that tagged factors are expressed at a comparable amount to the endogenous factor. This is usually achieved by cloning into a low-copy number vector and using the natural promoter to drive expression. If the tagged protein is expressed from a heterologous promoter, data comparing expression levels of the tagged and endogenous proteins (i.e., immunoblots to measure protein levels or qPCR to measure RNA levels) are needed. There are special cases in which ChIP cannot be obtained at endogenous protein levels, and here, elevated expression can provide useful information. ENCODE's recommended control for epitope-tagged measurements is an immunoprecipitation using the same antibody against the epitope tag in otherwise identical cells that do not express the tagged factor.

Histone modifications

For ENCODE data to be submitted, all commercial histone antibodies are validated by at least two independent methods, as described below, and new lots of antibody are analyzed independently. These validations are performed by the ENCODE laboratory performing the ChIP-seq or by the antibody supplier, but only if the supplier provides data for the specific lot of antibody. The tests need only be performed once for each antibody lot.

Primary test

All antibodies used in ENCODE ChIP experiments are checked for reactivity with nonhistone proteins and with unmodified histones by performing immunoblot analysis on total nuclear extract and recombinant histones. To enable visual quantification of reactivity, a concentration series of both extract and recombinant histones are analyzed using recombinant histone levels that are comparable to those of the target histone in nuclear extract. Since cross-reactivity may vary between species, this test is performed using nuclear extracts from each species to be studied by ChIP. To pass the criteria for submission in ENCODE, the specific histone band should constitute at least 50% of the signal in western blots of nuclear extract, show at least 10-fold enrichment relative to any other single band, and show at least 10-fold enriched signal relative to unmodified histone.

Secondary test

In addition to the primary test, antibody specificity is verified by at least one additional test. The pros and cons of each test are described. The first two are the most commonly used.

Peptide binding tests

Peptide binding and peptide competition assays provide a fast method to initially evaluate the specificity and relative binding strength of antibodies to histone tails with different modifications (e.g., H3K9 or H3K27 and me1, me2, and me3 levels of methylation). A potential drawback is that antibodies may differ in their binding specificity toward histone tail peptides in vitro versus toward full-length histones in the context of chromatin in IP experiments. Nevertheless, observing at least a 10-fold enriched binding signal for the modification of interest relative to other modifications provides confidence in the antibody specificity. For these assays, histone tail peptides with particular modifications can be purchased commercially. Alternatively, peptide binding and/or competition assays using the same lot of antibody can be performed by the company from which the antibody is purchased.

Mass spectrometry

For antibodies generated against related and historically problematic modifications, the ability of the antibody to effectively distinguish between similar histone marks (e.g., H3K9me and H3K27me) and between different levels of methylation (e.g., H3K9me1, H3K9me2, and H3K9me3) can be tested by mass spectrometry analysis of material immunoprecipitated from histone preparations. For ENCODE data, the target modification constitutes at least 80% of the immunoprecipitated histone signal. This test may often not be successful because IP for one modification can simultaneously isolate coassociated histones with other modifications. Thus, only a positive result (i.e., a specific modification) is interpretable.

Mutants defective in modifying histones

Strains or cell lines harboring knockouts or catalytically inactive mutants of enzymes responsible for particular histone modifications offer the opportunity to test antibody specificity. Such mutants exist for *S. cerevisiae*, *S. pombe*, *Drosophila*, *C. elegans* and can, in cases where the modifying enzymes are nonredundant, be created for mammalian cells. For submitted ENCODE/modENCODE data, antibody signal is reduced to below 10% of wild-type signal in mutant samples, compared with wild type. RNAi or siRNA depletion of histone modifying activity may be substituted for mutants. Mutant or RNAi or siRNA reduction of signal can be assayed by immunoblot analysis or by immunofluorescence staining. Mutant/RNAi/siRNA tests usually do not allow testing antibodies for the ability to discriminate between mono-, di-, and trimethylation. In cases where more than one enzyme modifies the same residue (e.g., H3K9 methylation in *Drosophila*), double mutants or RNAi may be required. Replicates of this test are encouraged but not required for ENCODE/modENCODE data to be submitted. However, positive controls showing that the antibody works on

(*continued*)

Box 1: *Continued*

wild-type samples processed in parallel, and positive controls showing that the mutant extract is amenable to the assay employed are included for data to be submitted.

Mutant histones

Mutant histones (e.g., histone H3 with Lys4 mutated to Arg or Ala) expressed in yeast provide another avenue to test specificity by immunoblot analysis or even by ChIP. When analyzing a strain containing a mutated histone that cannot be modified, we expect at least a 10-fold reduction in immunoblot or IP signal relative to wild-type histone preparations. Mutant histone tests cannot distinguish whether antibodies discriminate between mono, di, and trimethylation.

Annotation enrichment

Enrichment at annotated features (e.g., transcription start sites) can be used as a validation criterion for certain chromatin-associated modifications and proteins. If a well-characterized modification (e.g. H3K4me3) is analyzed, the observed localization to annotations are expected to be similar to that of known overlap standards derived from the literature or existing ChIP-seq data sets (for point source peaks, overlap with known annotations can be assessed using the IDR guidelines in Box 3).

Use of two different antibodies

Even if antibodies pass the specificity tests described above, observing similar ChIP results with two independent antibodies provides added confidence. We therefore aspire to obtain ChIP-seq data from two independent antibodies whenever possible, providing statistical comparisons of the results and presenting the intersection of the peak sets obtained with the two antibodies. The reasons for a significant discordance can be either biological or technical, and merit further dissection.

Replication, sequencing depth, library complexity, and site discovery

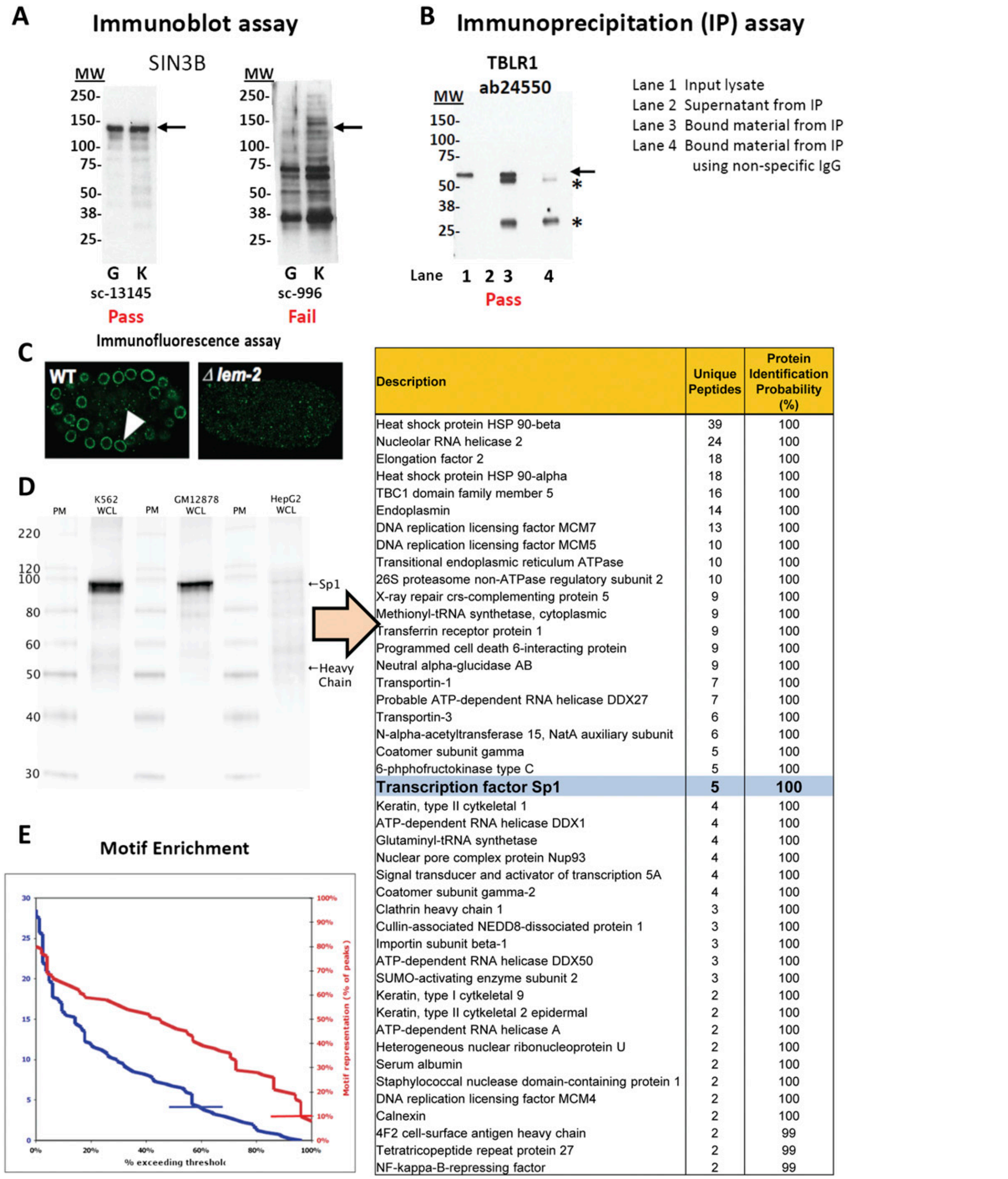
Biological replicate experiments from independent cell cultures, embryo pools, or tissue samples are used to assess reproducibility. Initial RNA polymerase II ChIP-seq experiments showed that more than two replicates did not significantly improve site discovery (Rozowsky et al. 2009). Thus, the ENCODE Consortium set as our standard that all ChIP measurements would be performed on two independent biological replicates. The irreproducible discovery rate (IDR) analysis methodology (Li et al. 2011) is now used to assess replicate agreement and set thresholds (discussed further below). For experiments with poor values for quality metrics described in Section III, additional replicate(s) have been generated.

For a typical point-source DNA-binding factor, the number of ChIP-seq positive sites identified typically increases with the number of sequenced reads (Myers et al. 2011). This result is expected, as studies of numerous factors by ENCODE and by other groups have repeatedly found a continuum of ChIP signal strength, rather than a sharply bounded and discrete set of positive sites (Rozowsky et al. 2009; Myers et al. 2011). Weaker sites can be detected with greater confidence in larger data sets because of the increased statistical power afforded by more reads. Figure 3 shows an analysis of peak calls for 11 human ENCODE ChIP-seq data sets for which deep-sequence data (30–100 million mapped reads) were obtained. Clear saturation of peak counts was observed for one factor with few binding sites, but counts continued to increase at varying rates for all other factors, including a case in which >150,000 peaks were called using 100 million mapped reads. Examination of peak signals reveals that the signal enrichments consistently plateau at greater sequencing depths. At 20 million mapped reads, which we currently use as a minimum for all ENCODE ChIP experiments for point-source transcription factors (Box 2), five- to 13-fold median enrichments are the norm; new peaks identified after 20 million reads give enrichments that are ~20% of the enrichment of the strongest peaks (Fig. 3C). Interestingly, many additional peaks, with enrichment values of three- to sevenfold, can still be found by sequencing to much greater depths. It is likely that many of these regions correspond to low-affinity sites and/or regions of open chromatin that bind TFs less specifically.

The relationship of ChIP signal strength to biological regulatory activity is a current area of active investigation. The biological activity of known enhancers, defined in the literature independently of ChIP data, is distributed quite broadly relative to ChIP-seq signal strength (Ozdemir et al. 2011; G DeSalvo, G Marinov, K Fisher, A Kirilusha, A Mortazavi, B Williams, and B Wold, in prep.). Some highly active transcriptional enhancers reproducibly display modest ChIP signals (Fig. 4B). This means that one cannot a priori set a specific target threshold for ChIP peak number or ChIP signal strength that will assure inclusion of all functional sites (see Discussion). Therefore, a practical goal is to maximize site discovery by optimizing immunoprecipitation and sequencing deeply, within reasonable expense constraints. For point-source factors in mammalian cells, a minimum of 10 million uniquely mapped reads are used by ENCODE for each biological replicate (providing a minimum of 20 million uniquely mapped reads per factor); for worms and flies a minimum of 2 million uniquely mapped reads per replicate is used. For broad areas of enrichment, the appropriate number of uniquely mapped reads is currently under investigation, but at least 20 million uniquely mapped reads per replicate for mammalian cells and 5 million uniquely mapped reads per replicate for worms and flies is currently being produced for most experiments.

Site discovery and reproducibility are also affected by the complexity of a ChIP-seq sequencing library (Fig. 4A). We define library complexity operationally as the fraction of DNA fragments that are nonredundant. With increased depth of sequencing of a library, a point is eventually reached where the complexity will be exhausted and the same PCR-amplified DNA fragments will be sequenced repeatedly. Low library complexity can occur when very low amounts of DNA are isolated during the IP or due to problems with library construction.

A useful complexity metric is the fraction of nonredundant mapped reads in a data set (nonredundant fraction or NRF), which we define as the ratio between the number of positions in the genome that uniquely mappable reads map to and the total number of uniquely mappable reads; it is similar to a recently published redundancy metric (Heinz et al. 2010). NRF decreases with se-



Description	Unique Peptides	Protein Identification Probability (%)
Heat shock protein HSP 90-beta	39	100
Nucleolar RNA helicase 2	24	100
Elongation factor 2	18	100
Heat shock protein HSP 90-alpha	18	100
TBC1 domain family member 5	16	100
Endoplasmin	14	100
DNA replication licensing factor MCM7	13	100
DNA replication licensing factor MCM5	10	100
Transitional endoplasmic reticulum ATPase	10	100
26S proteasome non-ATPase regulatory subunit 2	10	100
X-ray repair crs-complementing protein 5	9	100
Methionyl-tRNA synthetase, cytoplasmic	9	100
Transferrin receptor protein 1	9	100
Programmed cell death 6-interacting protein	9	100
Neutral alpha-glucidase AB	9	100
Transportin-1	7	100
Probable ATP-dependent RNA helicase DDX27	7	100
Transportin-3	6	100
N-alpha-acetyltransferase 15, NatA auxiliary subunit	6	100
Coatomer subunit gamma	5	100
6-phphofructokinase type C	5	100
Transcription factor Sp1	**5**	**100**
Keratin, type II cytkeletal 1	4	100
ATP-dependent RNA helicase DDX1	4	100
Glutaminyl-tRNA synthetase	4	100
Nuclear pore complex protein Nup93	4	100
Signal transducer and activator of transcription 5A	4	100
Coatomer subunit gamma-2	4	100
Clathrin heavy chain 1	3	100
Cullin-associated NEDD8-dissociated protein 1	3	100
Importin subunit beta-1	3	100
ATP-dependent RNA helicase DDX50	3	100
SUMO-activating enzyme subunit 2	3	100
Keratin, type I cytkeletal 9	2	100
Keratin, type II cytkeletal 2 epidermal	2	100
ATP-dependent RNA helicase A	2	100
Heterogeneous nuclear ribonucleoprotein U	2	100
Serum albumin	2	100
Staphylococcal nuclease domain-containing protein 1	2	100
DNA replication licensing factor MCM4	2	100
Calnexin	2	100
4F2 cell-surface antigen heavy chain	2	99
Tetratricopeptide repeat protein 27	2	99
NF-kappa-B-repressing factor	2	99

Figure 2. Representative results from antibody characterization assays. (*A*) Immunoblot analyses of antibodies against SIN3B that (*left*) pass quality control (Santa Cruz sc13145) and (*right*) fail quality control (Santa Cruz sc996). Lanes contain nuclear extract from GM12878 cells (G) and K562 cells (K). Arrows indicate band of expected size of 133 kDa. Molecular weights (MW) are in kilodaltons. (*B*) Immunoblot analysis of an antibody against TBLR1 (Abcam ab24550) that passes quality control and can be used for immunoprecipitation. Immunoprecipitations (IPs) were performed from nuclear lysates of K562 cells. Arrow indicates band of expected size (56 kDa) that is detected in the input lysate (lane *1*) and is efficiently (cf. lanes *3* and *2*) and specifically (absent in lane *4*) immunoprecipitated. (*) IgG light and heavy chains. (*C*) Immunofluorescence analyses of antibodies that pass (*left*) and fail (*right*) quality control. (*D*) Immunoprecipitation/mass spectrometry analysis of an antibody against SP1 (Santa Cruz sc-17824). Whole-cell lysates (WCL) of K562, GM12878, and HepG2 were immunoprecipitated, and a band of expected size (~106 kDa) was detected on a Western blot with SP1 primary antibody. The immunoprecipitation was repeated in K562 WCL, separated on a gel, stained with Coomassie Blue, and the band previously detected on the Western blot was excised and analyzed by mass spectrometry. Peptides were identified using MASCOT (Matrix Science) with probability-based matching at $P < 0.05$. Subsequent analysis was performed in Scaffold (Proteome Software, Inc.) at 0.0% protein FDR and 0.0% peptide FDR. SP1 protein was detected (along with common contaminants that are often obtained in control experiments) (data not shown) and is highlighted in bold and light blue. (*E*) Histogram depicting motif fold-enrichment (blue) for all transcription factors for which ENCODE ChIP-seq data is available (85 factors). Enrichments are relative to all DNase-accessible sites and were corrected for sequence bias using shuffle motifs. Motif searches were conducted with a matching stringency of 4–6. Where multiple data sets are available for a factor, the data set with the highest enrichment was counted. Data sets that meet the ENCODE standard of fourfold enrichment (indicated by blue line) were found for 60% of factors. Motif representation, as a percentage of all analyzed peaks, is shown in red for all factors for which a data set exists that exceeds the enrichment standard. A total of 96% of these data sets meet the ENCODE standard of >10% motif representation (red line). All calculations were carried out on peaks identified by IDR analysis (0.01 cut-off).

quencing depth, and for point source TFs, our current target is NRF ≥0.8 for 10 million (M) uniquely mapped reads (Box 2). We expect that, as sequencing technology improves and read numbers in the hundreds of millions per lane become feasible, even complex libraries from point-source factor libraries may be sequenced at depths greater than necessary. To maximize information that can be obtained for each DNA-sequencing run and to prevent oversequencing, barcoding and pooling strategies can be used (Lefebvre et al. 2010).

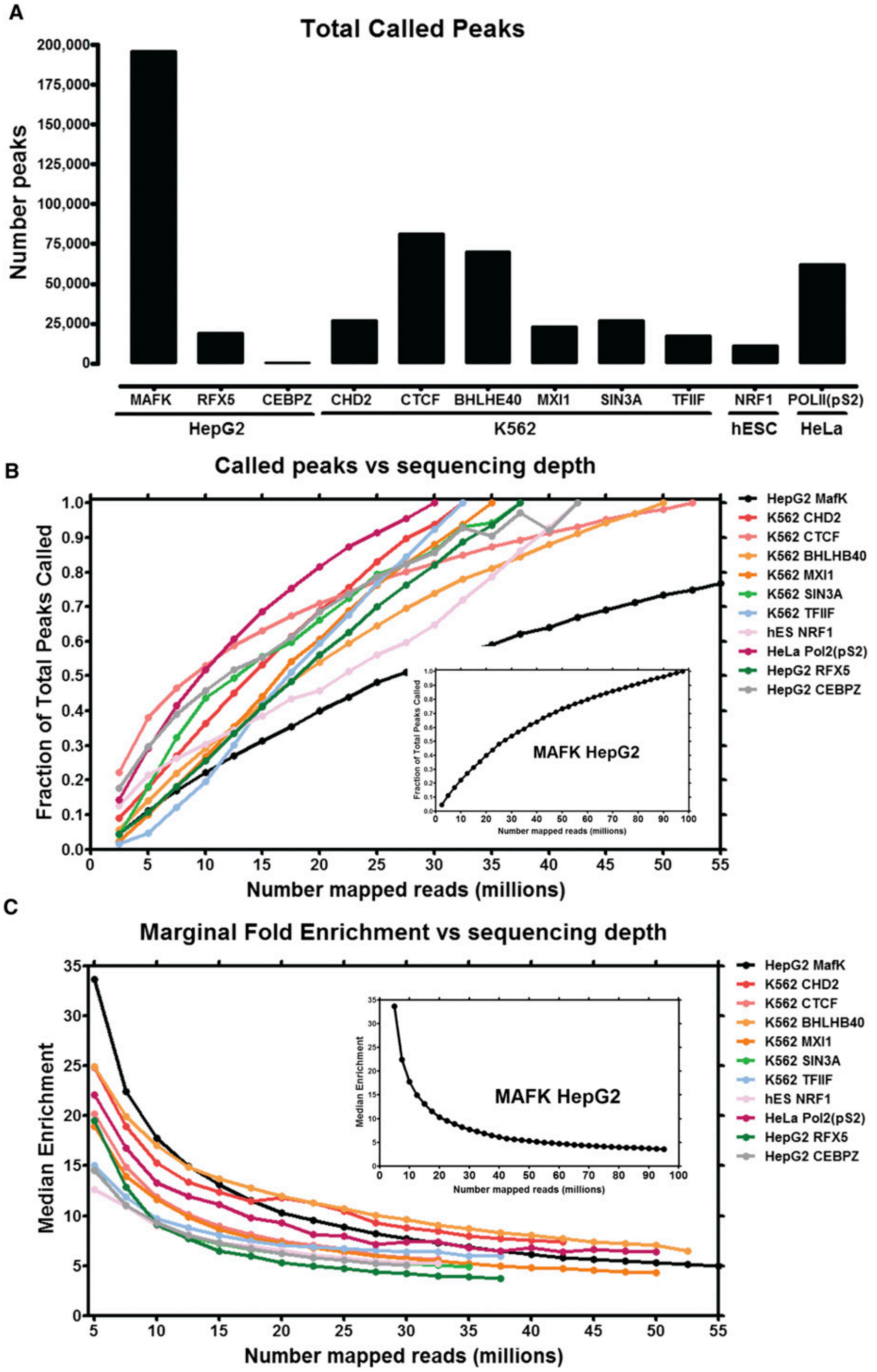


Figure 3. Peak counts depend on sequencing depth. (*A*) Number of peaks called with Peak-seq (0.01% FDR cut-off) for 11 ENCODE ChIP-seq data sets. (*B*) Called peak numbers for 11 ChIP-seq data sets as a function of the number of uniquely mapped reads used for peak calling. (*Inset*) Called peak data for the MAFK data set from HepG2 cells, currently the most deeply sequenced ENCODE ChIP-seq data set (displayed separately due to the significantly larger number of reads relative to the other data sets). Data sets are indicated by cell line and transcription factor (e.g., cell line HepG2, transcription factor MAFK). (*C*) Fold-enrichment for newly called peaks as a function of sequencing depth. For each incremental addition of 2.5 million uniquely mapped reads, the median fold-enrichment for newly called peaks as compared with an IgG control data set sequenced to identical depth is plotted.

Control sample

An appropriate control data set is critical for analysis of any ChIP-seq experiment because DNA breakage during sonication is not uniform. In particular, some regions of open chromatin are preferentially represented in the sonicated sample (Auerbach et al. 2009). There are also platform-specific sequencing efficiency biases that contribute to nonuniformity (Dohm et al. 2008). There are two basic methods to produce control DNA samples, each of which mitigates the effects of these issues on binding-site identification: (1) DNA is isolated from cells that have been cross-linked and fragmented under the same conditions as the immunoprecipitated DNA ("Input" DNA); and (2) a "mock" ChIP reaction is performed using a control antibody that reacts with an irrelevant, non-nuclear antigen ("IgG" control). For both types of controls, ENCODE groups sequence to a depth at least equal to, and preferably larger than, that of the ChIP sample. While the IgG control mimics a ChIP experiment more closely than does an "input" control, it is important that IgG control immunoprecipitations recover enough DNA to build a library of sufficiently high complexity to that of the experimental samples; otherwise, binding-site identifications made using this control can be significantly biased.

Regardless of the type of control used, ENCODE and modENCODE groups perform a separate control experiment for each cell line, developmental stage, and different culture condition/treatment because of known and unknown differences in ploidy, genotype, and epigenetic features that affect chromatin preparation. To serve as a valid control, we use identical protocols to build ChIP and control sequencing libraries (i.e., the same as the number of PCR amplification cycles,

Box 2: ChIP experimental design guidelines

Sequencing and library complexity

For each ChIP-seq point-source library, ENCODE's goal is to obtain ≥10 million uniquely mapping reads per replicate experiment for mammalian genomes, with a target NRF (nonredundancy fraction) ≥0.8 for 10 million reads. The corresponding objective for modENCODE point-source factors is to obtain ≥2 M uniquely mapped reads per replicate, ≥0.8 NRF. The modENCODE target for broad-source ChIP-seq in *Drosophila* is ≥5 million reads, and the ENCODE provisional target for mammalian broad-source histone marks is ≥20 million uniquely mapping reads at NRF ≥0.8. The distribution of NRF values for all ENCODE data sets is shown in Figure 7.

Control libraries

ENCODE generates and sequences a control ChIP library for each cell type, tissue, or embryo collection and sequences the library to the appropriate depth (i.e., at least equal to, and preferably greater than, the most deeply sequenced experimental library). If cost constraints allow, a control library should be prepared from every chromatin preparation and sonication batch, although some circumstances can justify fewer control libraries. Importantly, a new control is always performed if the culture conditions, treatments, chromatin shearing protocol, or instrumentation is significantly modified.

Reproducibility

Experiments are performed at least twice to ensure reproducibility. For ENCODE data to pass criteria for submission, concordance is determined from analysis using the IDR methodology (current ENCODE criteria are in Box 3), and a third replicate is performed if the standard is not reached. Cut-offs for identifying highly reproducible peaks for use in subsequent analyses can be determined by IDR (typically using a 1% threshold).

fragment size, etc.). Although rare in our experience, control libraries with particularly strong sonication biases have been observed and they can adversely affect peak calling (Supplemental Fig. S1). As much as possible, ENCODE/modENCODE groups also generate a separate control for each batch of sonicated samples to control for possible sonication variation.

Peak calling

After mapping reads to the genome, peak calling software is used to identify regions of ChIP enrichment. We have used several peak calling algorithms and corresponding software packages, including SPP, PeakSeq, and MACs (Ji et al. 2008; Valouev et al. 2008; Zhang et al. 2008; Rozowsky et al. 2009). The resulting output of these algorithms generally ranks called regions by absolute signal (read number) or by computed significance of enrichment (e.g., *P*-values and false discovery rates). Because ChIP signal strength is a continuum with many more weak sites than strong ones (Fig. 4B), the composition of the final peak list depends heavily on the specific parameter settings and the algorithm used as well as the quality of the experiment itself. Thresholds that are too relaxed lead to a high proportion of false positives for each replicate, but as discussed below, subsequent analysis can strip false positives from a final joint peak determination. Different peak-calling algorithms rely on different statistical models to calculate *P*-values and false discovery rates (FDR), meaning that significance values from different software packages are not directly comparable. When using standard peak-calling thresholds, successful experiments generally identify thousands to tens of thousands of peaks for most TFs in mammalian genomes, although some exceptions are known (Frietze et al. 2010; Raha et al. 2010). In all cases, it is important to use an appropriate control experiment in peak calling.

Calling discrete regions of enrichment for *Broad-source* factors or *Mixed-source* factors is more challenging and is at an earlier stage of development. Methods to identify such regions are emerging (e.g., ZINBA [Rashid et al. 2011] [installation package at http://code.google.com/p/zinba/], Scripture [Guttman et al. 2010], and MACS2, an updated version of MACS that is specifically designed to process mixed signal types [https://github.com/taoliu/MACS]). Standards for the identification of broad enrichment regions are currently in development.

Evaluating ChIP-seq data

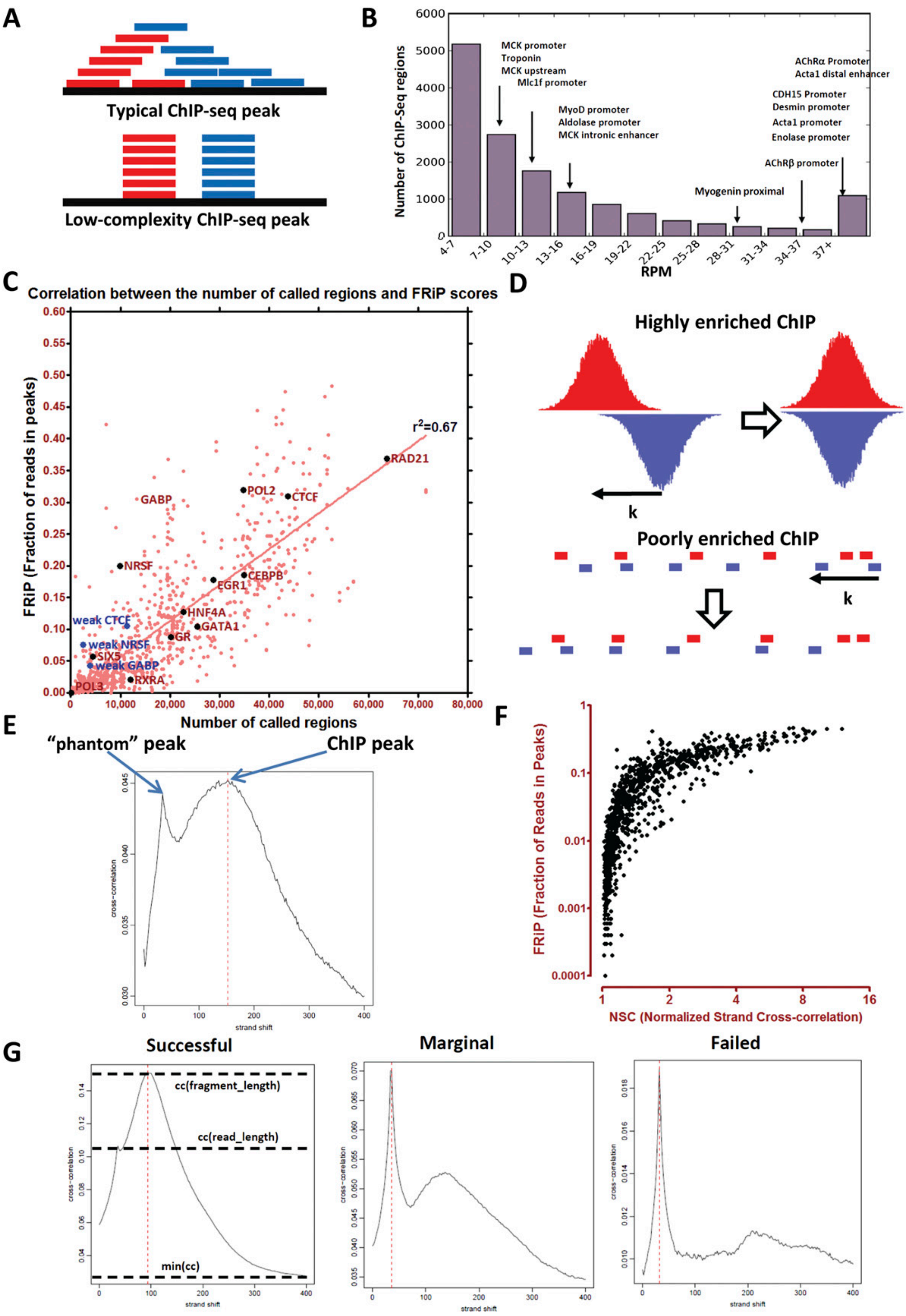
The quality of individual ChIP-seq experiments varies considerably and can be especially difficult to evaluate when new antibodies are being tested or when little is known about the factor and its binding motif. The ENCODE Consortium has developed and uses metrics for several aspects of ChIP-seq quality, together with traditional site-inspection-based evaluation. When applied and interpreted as a group, these metrics and approaches provide a valuable overall assessment of experimental success and data quality.

Browser inspection and previously known sites

A first impression about ChIP-seq quality can be obtained by local inspection of mapped sequence reads using a genome browser. Although not quantitative, this approach is very useful, especially when a known binding location can be examined; read distribution shape and signal strength relative to a control sample can provide a sense of ChIP quality. A true signal is expected to show a clear asymmetrical distribution of reads mapping to the forward and reverse strands around the midpoint (peak) of accumulated reads. This signal should be large compared with the signal of the same region from the control library. Of course it is not feasible to inspect the whole genome in this manner, and evaluating a limited number of the strongest sites may overestimate the quality of the entire data set (Supplemental Fig. S2). The genome-wide metrics discussed below provide more objective and global assessments.

Measuring global ChIP enrichment (FRiP)

For point-source data sets, we calculate the fraction of all mapped reads that fall into peak regions identified by a peak-calling algorithm (Ji et al. 2008). Typically, a minority of reads in ChIP-seq experiments occur in significantly enriched genomic regions (i.e., peaks); the remainder of the read represents background. The fraction of reads falling within peak regions is therefore a useful and simple first-cut metric for the success of the immunoprecipitation, and is called FRiP (fraction of reads in peaks). In general, FRiP values correlate positively and linearly with the number of called regions, although there are exceptions, such as REST (also known as NRSF) and GABP, which yield a more limited number of



$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Figure 4. (Legend on next page)

called regions but display very high enrichment (Fig. 4C). Most (787 of 1052) ENCODE data sets have a FRiP enrichment of 1% or more when peaks are called using MACS with default parameters. The ENCODE Consortium scrutinizes experiments in which the FRiP falls below 1%.

The 1% FRiP guideline works well when there are thousands to tens of thousands of called occupancy sites in a large mammalian genome. However, passing this threshold does not automatically mean that an experiment is successful and a FRiP below the threshold does not automatically mean failure. For example, ZNF274 and human RNA polymerase III have very few true binding sites (Frietze et al. 2010; Raha et al. 2010), and a FRiP of <1% is obtained. At the other extreme, ChIP experiments using antibody/factor pairs capable of generating very high enrichment (such as REST and GABP mentioned above) and/or binding-site numbers (CTCF, RAD21, and others) can result in FRiP scores that exceed those obtained for most factors (Fig. 5C), even for experiments that are suboptimal. Thus, FRiP is very useful for comparing results obtained with the same antibody across cell lines or with different antibodies against the same factor. FRiP is sensitive to the specifics of peak calling, including the way the algorithm delineates regions of enrichment and the parameters and thresholds used. Thus, all FRiP values that are compared should be derived from peaks uniformly called by a single algorithm and parameter set.

Cross-correlation analysis

A very useful ChIP-seq quality metric that is independent of peak calling is strand cross-correlation. It is based on the fact that a high-quality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered around the binding site. As illustrated in Figure 5D, these "true signal" sequence tags are positioned at a distance from the binding site center that depends on the fragment size distribution (Kharchenko et al. 2008). A control experiment, such as sequenced input DNA, lacks this pattern of shifted stranded tag densities (Supplemental Fig. S1). This has made it possible to develop a metric that quantifies fragment clustering (IP enrichment) based on the correlation between genome-wide stranded tag densities (A Kundaje, Y Jung, P Kharchenko, B Wold, A Sidow, S Batzoglou, and P Park, in prep.). It is computed as the Pearson linear correlation between the Crick strand and the Watson strand, after shifting Watson by *k* base pairs (Fig. 5E). This typically produces two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length ("phantom" peak) (Fig. 4E; Heinz et al. 2010; A Kundaje, Y Jung, P Kharchenko, B Wold, A Sidow, S Batzoglou, and P Park, in prep.).

The normalized ratio between the fragment-length cross-correlation peak and the background cross-correlation (normalized strand coefficient, NSC) and the ratio between the fragment-length peak and the read-length peak (relative strand correlation, RSC) (Fig. 4G), are strong metrics for assessing signal-to-noise ratios in a ChIP-seq experiment. High-quality ChIP-seq data sets tend to have a larger fragment-length peak compared with the read-length peak, whereas failed ones and inputs have little or no such peak (Figs. 4G, 5A,B; Fig. 7, below). In general, we observe a continuum between the two extremes, and broad-source data sets are expected to have flatter cross-correlation profiles than point-sources, even when they are of very high quality. As expected, the NSC/RSC and FRiP metrics are strongly and positively correlated for the majority of experiments (Fig. 4F). As with the other quality metrics, even high-quality data sets generated for factors with few genuine binding sites tend to produce relatively low NSCs.

These measures form the basis for one of the current quality standards for ENCODE data sets. We repeat replicates with NSC values <1.05 and RSC values <0.8 and, if additional replicates produce low values, we include a note with the reported data set (Box 3). We illustrate the application of our ChIP-seq quality metrics to a failed pair of replicates in Figure 5, A–E. Initially, two EGR1 ChIP-seq replicates were generated in the K562 cell line. Based on the cross-correlation profiles, FRiP score, and number of called regions, these replicates were flagged as marginal in quality. The experiments were repeated, with all quality control metrics improving considerably. On this basis, the superior measurements replaced the initial ones in the ENCODE database.

Consistency of replicates: Analysis using IDR

As noted above, the modENCODE and ENCODE consortia generate two independent biological replicates, with each experiment passing the basic quality control filters. As another measure of experiment quality, we take advantage of the reproducibility information provided by the duplicates using the IDR (irreproducible discovery rate) statistic that has been developed for ChIP-seq (Li et al. 2011; discussed in detail in A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep.).

Given a set of peak calls for a pair of replicate data sets, the peaks can be ranked based on a criterion of significance, such as the *P*-value, the q-value, the ChIP-to-input enrichment, or the read coverage for each peak (Fig. 6A–E). If two replicates measure the same underlying biology, the most significant peaks, which are likely to be genuine signals, are expected to have high consistency between replicates, whereas peaks with low significance, which are more likely to be noise, are expected to have low consistency. If the consistency between a pair of rank lists that contains both signif-

Figure 4. Criteria for assessing the quality of a ChIP-seq experiment. (*A*) Library complexity. Individual reads mapping to the plus (red) or minus strand (blue) are represented. (*B*) Distribution of functional regulatory elements with respect to the strength of the ChIP-seq signal. ChIP-seq was performed against myogenin, a major regulator of muscle differentiation, in differentiated mouse myocytes. While many extensively characterized muscle regulatory elements exhibit strong myogenin binding, a large number of known functional sites are at the low end of the binding strength continuum. (*C*) Number of called peaks vs. ChIP enrichment. Except in special cases, successful experiments identify thousands to tens of thousands of peaks for most TFs and, depending on the peak finder used, numbers in the hundreds or low thousands indicate a failure. Peaks were called using MACS with default thresholds. (*D*) Generation of a cross-correlation plot. Reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Pearson correlation between the per-position read count vectors for each strand is calculated. Read coverage as wigglegram is represented, not to the same scale in the *top* and *bottom* panels.) (*E*) Two cross-correlation peaks are usually observed in a ChIP experiment, one corresponding to the read length ("phantom" peak) and one to the average fragment length of the library. (*F*) Correlation between the fraction of reads within called regions and the relative cross-correlation coefficient for 1052 human ChIP-seq experiments. (*G*) The absolute and relative height of the two peaks are useful determinants of the success of a ChIP-seq experiment. A high-quality IP is characterized by a ChIP peak that is much higher than the "phantom" peak, while often very small or no such peak is seen in failed experiments.

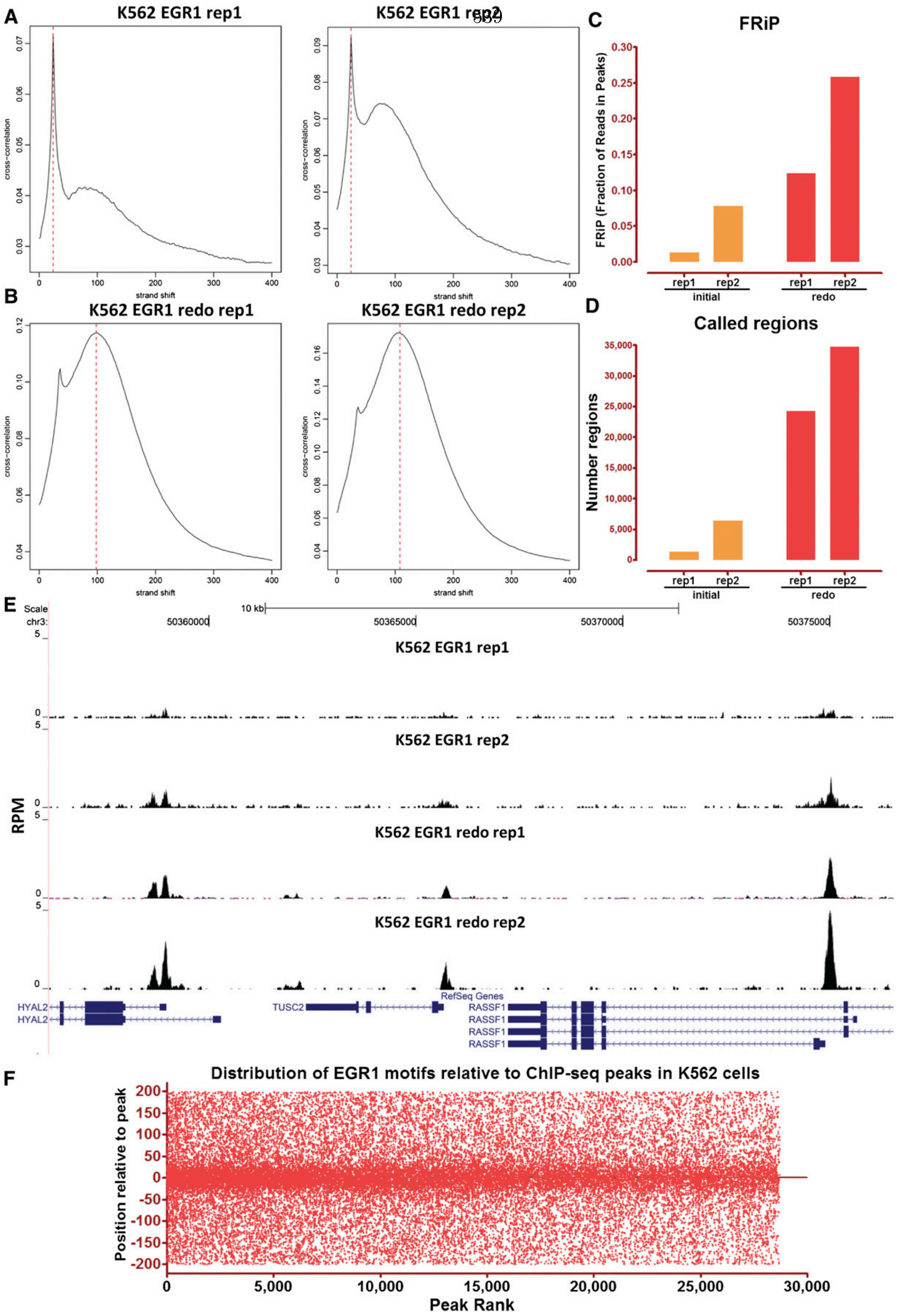


Figure 5. Quality control of ChIP-seq data sets in practice. EGR1 ChIP-seq was performed in K562 cells in two replicates. ChIP enriched regions were identified using MACS. However, the cross-correlation plot profiles (*A*) indicated that both IPs were suboptimal, with one being unacceptable. In agreement with this judgment, ChIP enrichment (*C*) and peak number (*D*) also indicated failure. The ChIP-seq assays were repeated (*B*), with all quality control metrics improving significantly (*B,D*), and many additional EGR1 peaks were identified as a result. (*E*) Representative browser snapshot of the four EGR1 ChIP-seq experiments, showing the much stronger peaks obtained with the second set of replicates. (*F*) Distribution of EGR1 motifs relative to the bioinformatically defined peak position of EGR1-occupied regions derived from ChIP-seq data in K562 cells. Regions are ranked by their confidence scores as called by SPP.

Box 3: ChIP-seq quality assessment guidelines

Within ENCODE, a set of data quality thresholds has been established for submission of ChIP-seq data sets. These have been constructed based on the historical experiences of ENCODE ChIP-seq data production groups with the purpose of balancing data quality with practical attainability and are routinely revised. The current standards are below and the performance of ENCODE data sets against these thresholds is shown in Figure 7.

Cross-correlation analysis

The current ENCODE practice is to calculate and report NSC and RSC for each experiment. For experiments with NSC values below 1.05 and RSC values below 0.8, we currently recommend that an additional replicate be attempted or the experiment explained in the data submission as adequate based on additional considerations.

Irreproducible discovery rate (IDR)

The following guidelines have been established for mammalian cells (optimal parameter may differ for other organisms). Biological replicates are performed for each ChIP-seq data set and subjected to peak calling. IDR analysis is then performed with a 1% threshold. For submission to ENCODE, we currently require that the number of bound regions identified in an IDR comparison between replicates to be at least 50% of the number of regions identified in an IDR comparison between two "pseudoreplicates" generated by pooling and then randomly partitioning all available reads from all replicates (Np/Nt < 2) (Fig. 7). To ensure similar weighting of individual replicates for identifying binding regions, we further recommend that the number of significant peaks identified using IDR on each individual replicate (obtained by partitioning reads into two equal groups for the IDR analysis) be within a factor of 2 of one another (N1/N2 < 2) (Fig. 7). Data sets which fail to meet these criteria may still be deposited by ENCODE experimenters, provided that at least three experimental replicates have been attempted and a note accompanies these data sets explaining which parameters fail to meet the standards and providing any technical information that may explain this failure. This guideline is for point source features; metrics are still being determined for broad peak analyses.

Updated information about the performance of ENCODE data sets against these quality metrics and tools for determining these metrics will be forthcoming through the ENCODE portal (http://encodeproject.org/ENCODE/).

Historical note

A simpler heuristic for establishing reproducibility was previously used as a standard for depositing ENCODE data and was in effect when much of the currently available data was submitted. According to this standard, either 80% of the top 40% of the targets identified from one replicate using an acceptable scoring method should overlap the list of targets from the other replicate, or target lists scored using all available reads from each replicate should share more than 75% of targets in common. As with the current standards, this was developed based on experience with accumulated ENCODE ChIP-seq data, albeit with a much smaller sample size.

icant and insignificant findings is plotted, a transition in consistency is expected (Fig. 6C,F). This consistency transition provides an internal indicator of the change from signal to noise and suggests how many peaks have been reliably detected.

The IDR statistic quantifies the above expectations of consistent and inconsistent groups by modeling all pairs of peaks present in both replicates as belonging to one of two groups: a reproducible group, and an irreproducible group (Li et al. 2011). In general, the signals in the reproducible group are more consistent (i.e., have a larger correlation coefficient) and are ranked higher than the irreproducible group. The proportion of identifications that belong to the "noise" component and the correlation of the significant component are estimated adaptively from the data. The IDR provides a score for each peak, which reflects the posterior probability that the peak belongs to the irreproducible group.

A major advantage of IDR is that it can be used to establish a stable threshold for called peaks that is more consistent across laboratories, antibodies, and analysis protocols (e.g., peak callers) than are FDR measures (A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep.). Increased consistency comes from the fact that IDR uses information from replicates, whereas the FDR is computed on each replicate independently. The application of IDR to real-life data is shown in Figure 6. A pair of high-quality RAD21 ChIP-seq replicates display good consistency between IDR ranks for a large number (~28,000) of highly reproducible peaks (Figs. 6A,B), with a clear inflection between the signal and noise populations near the 1% IDR value (Fig. 6C). In contrast, a pair of SPT20 replicates, which had already been flagged as low-quality based on the individual FRiP and NSC/RSC metrics, display very low IDR reproducibility, with very few significant peaks, and no visible inflection in the IDR curve (Fig. 6F).

It is important that the peak-calling threshold used prior to IDR analysis not be so stringent that the noise component is entirely unrepresented in the data, because the algorithm requires sampling of both signal and noise distributions to separate the peaks into two groups; thus relaxing the default stringency settings when running a given peak caller is advised if IDR analysis will follow.

A caution in applying IDR is that it is dominated by the weakest replicate (A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep.). That is, if one replicate is quite poor, many "good" peaks from the higher quality replicate will be rejected by IDR analysis, because they are not reproducible in the weak replicate. To ensure similar weighting of individual replicates, the number of significant binding regions identified using IDR on each individual replicate (obtained by partitioning reads into two equal groups to allow the IDR analysis) is recommended to be within a factor of 2 for data sets to be submitted to UCSC by ENCODE (Box 3).

ENCODE has begun applying IDR analysis to all ChIP experiments. For all submitted ENCODE ChIP-seq data sets, the number of bound regions identified in an IDR comparison between replicates is at least 50% of the number of regions identified in an IDR comparison between two "pseudoreplicates" generated by randomly partitioning available reads from all replicates (Box 3).

Guidelines for reporting ChIP-seq data

To facilitate data sharing among laboratories, both within and outside the Consortium, and to ensure that results can be reproduced, ENCODE has established guidelines for data sharing in public repositories. Raw data can be submitted to the Short Read Archive (SRA) and ChIP results are submitted to GEO. Through

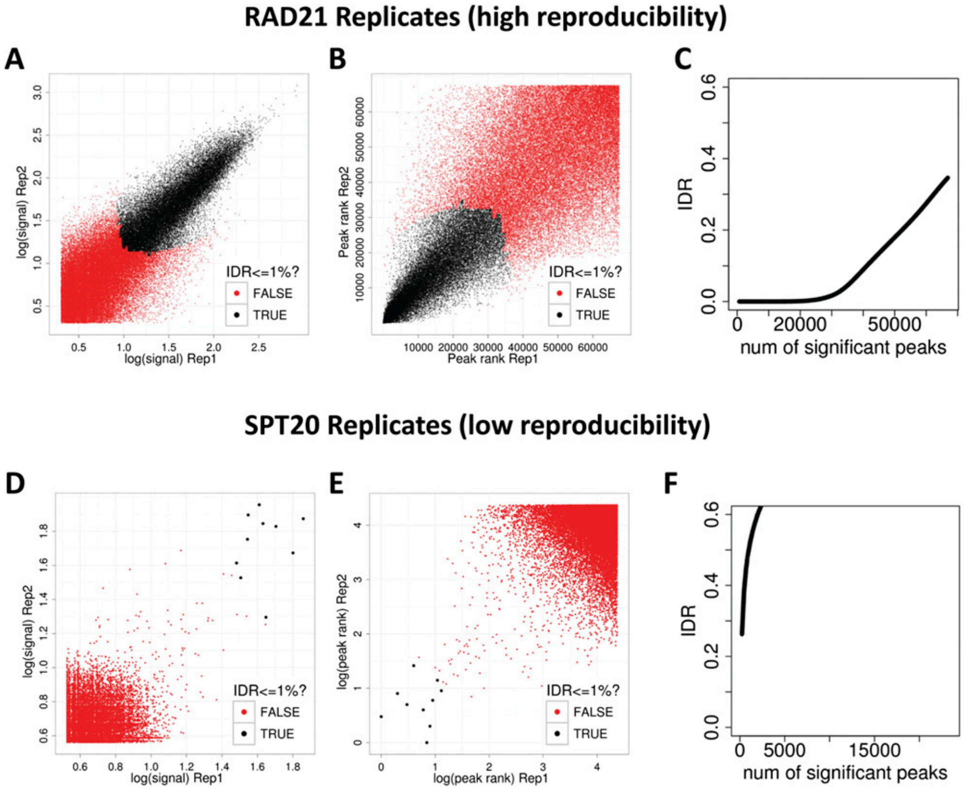


Figure 6. The irreproducible discovery rate (IDR) framework for assessing reproducibility of ChIP-seq data sets. (*A*–*C*) Reproducibility analysis for a pair of high-quality RAD21 ChIP-seq replicates. (*D*,*E*) The same analysis for a pair of low quality SPT20 ChIP-seq replicates. (*A*,*D*) Scatter plots of signal scores of peaks that overlap in each pair of replicates. (*B*,*E*) Scatter plots of ranks of peaks that overlap in each pair of replicates. Note that low ranks correspond to high signal and vice versa. (*C*,*F*) The estimated IDR as a function of different rank thresholds. (*A*,*B*,*D*,*E*) Black data points represent pairs of peaks that pass an IDR threshold of 1%, whereas the red data points represent pairs of peaks that do not pass the IDR threshold of 1%. The RAD21 replicates show high reproducibility with ~30,000 peaks passing an IDR threshold of 1%, whereas the SPT20 replicates show poor reproducibility with only six peaks passing the 1% IDR threshold.

April 2012, 478 ChIP-seq data sets had been submitted to GEO at accession ID PRJNA63441, with submission of all current ENCODE data to be completed by June 2012. UCSC houses the ENCODE data (Rosenbloom et al. 2011) and modMine houses the modENCODE data (Contrino et al. 2011).

Box 4 provides a detailed description of the data and experimental and analytical details to be shared so that others can reproduce both experiments and analyses. Shared information includes the experimental procedures for performing the ChIP, antibody information and validation data, as well as relevant DNA sequencing, peak calling, and analysis details. For ENCODE experiments that do not meet the guidelines described above, data and results may be reported, with a note indicating that the criteria have not been met and explaining why the data are nevertheless released.

Discussion

The ENCODE and ModENCODE standards and practices presented here will be further revised as the protocols, technologies, and our understanding of the assays change. Updated versions will be released and made available at http://encodeproject.org/ENCODE/experiment_guidelines.html. We have begun to address the central but vexing issue of immune reagent specificity and performance by establishing a menu of primary and secondary methods for antibody characterization, including performance-reporting practices. We also developed and applied global metrics to assess the quality of several aspects of an individual ChIP-seq experiment: Library complexity can be measured by the nonredundant fraction (NRF); immunoenrichment can be measured by the fraction of reads in called peaks (FRiP) and by cross-correlation analysis (NSC/RSC); and replicate significance can be measured by IDR. We related these global quality measures to more traditional inspection of ChIP-seq browser tracks (Fig. 5) and discuss below how different aspects of data quality interact with specific uses of ChIP-seq data.

How good can a ChIP-seq experiment be?

Thus far, the most successful point-source factor experiments for ENCODE have FRiP values of 0.2–0.5 (factors such as REST, GABP, and CTCF) (Fig. 4C) and NSC/RSC values of 5–12. Although these quality scores and characteristics were routinely obtained for the best-performing factor/antibody combinations, they are not the rule; for most transcription factors, the ChIP quality metrics were substantially lower and more variable (Fig. 7). We believe that multiple issues contribute to the variability; the quality of antibody (affinity and specificity) is surely important, but epitope availability within fixed chromatin, sensitivity of the antibody to post-translational modifications of the antigen, how long and how often the protein is bound to DNA, and other physical characteristics of the protein–DNA interaction likely also contribute. Further work with epitope-tagged factors, for which the antibody is not a variable, should begin to sort among the possibilities.

Box 4. Data reporting guidelines

Data should be submitted to public repositories. The following information is currently used by ENCODE/modENCODE to submit data to public repositories.

Metadata

For submission of basic experimental data by ENCODE, the following information is minimally included:

- Investigator, organism, or cell line, experimental protocol (or reference to a known protocol).
- Indication as to whether an experiment is a technical or biological replicate.
- Catalog and lot number for any antibody used. If not a commercial antibody, indicate the precise source of the antibody.
- Information used to characterize the antibody, including summary of results (images of immunoblots, immunofluorescence, list of proteins identified by mass spec, etc.).
- Peak calling algorithm[29] and parameters used, including threshold and reference genome used to map peaks.
- A summary of the number of reads and number of targets for each replicate and for the merged data set.
- Criteria that were used to validate the quality of the resultant ChIP-seq data (i.e., overlap results or IDR[29]).
- Experimental validation results (e.g., qPCR).
- Link to the control track that was used.
- An explanation if the experiment fails to meet any of the standards.

High-throughput sequencing data

- Image files from sequencing experiments do not need to be stored.
- Raw data (FASTQ files) should be submitted to both GEO and SRA.
- Each replicate should be submitted independently.
- Target region and peak calling results.

Point source peaks

For point source peaks (e.g. experiments with antibodies to sequence-specific transcription factors), common features that are reported by ENCODE researchers include:

- Peak position, defined as a single base pair.
- Start and end positions, defined as specific base pairs.
- Signal value (e.g., fold enrichment) using an algorithm chosen by the submitter.
- Significance/accuracy measures:
  - □ *P*-value determined using a method chosen by the submitter.
  - □ Q-value (false discovery rate correction) determined using a method chosen by the submitter.
- Metadata, including peak caller approach and genome reference used, plus methods for determining signal values, *P*-values, and Q-values, as applicable.

Broad regions

- Start and end positions, defined as specific base pairs.
- Signal value (e.g., fold enrichment) using an algorithm chosen by the submitter.
- Significance/accuracy measures:
  - □ *P*-value determined using a method chosen by the submitter.
  - □ Q-value (false discovery rate correction) determined using a method chosen by the submitter.
- Metadata, including peak caller approach and genome reference used, plus methods for determining signal values, *P*-values, and Q-values, as applicable.
- Point-source peaks can be called in addition to broad regions (i.e., one can have "peaks" and potentially "valleys" within "regions").

The investigator should determine whether their data best fits the broad region/point source peak data or both.

When measurements differ in quality, the higher-quality replicate often identifies thousands more sites than the lower. Do sites present only in the superior ChIP experiment reflect true occupancy? Motif analysis suggests that many do. In Figure 5F, the position of EGR1 motifs relative to EGR1 ChIP-seq peaks is shown. The known binding motif is prominent and concentrated centrally under the ChIP peaks, as expected if the motif mediates occupancy; importantly, the central location of the motif is observed, even in the low-ranking peaks. The trend continues below the peak-calling cut-offs, suggesting additional true occupancy sites. Depending on the goals of an analysis, users may want to be more or less conservative in defining the threshold for inclusion. Motif presence could be used as one criterion for "rescuing" candidate sites identified in only one experiment.

[29] For uniform peak calling within ENCODE, the MACS peak caller, version 1.4.2 was used. Scripts used for IDR analysis are at https://sites.google.com/site/anshulkundaje/projects/idr.

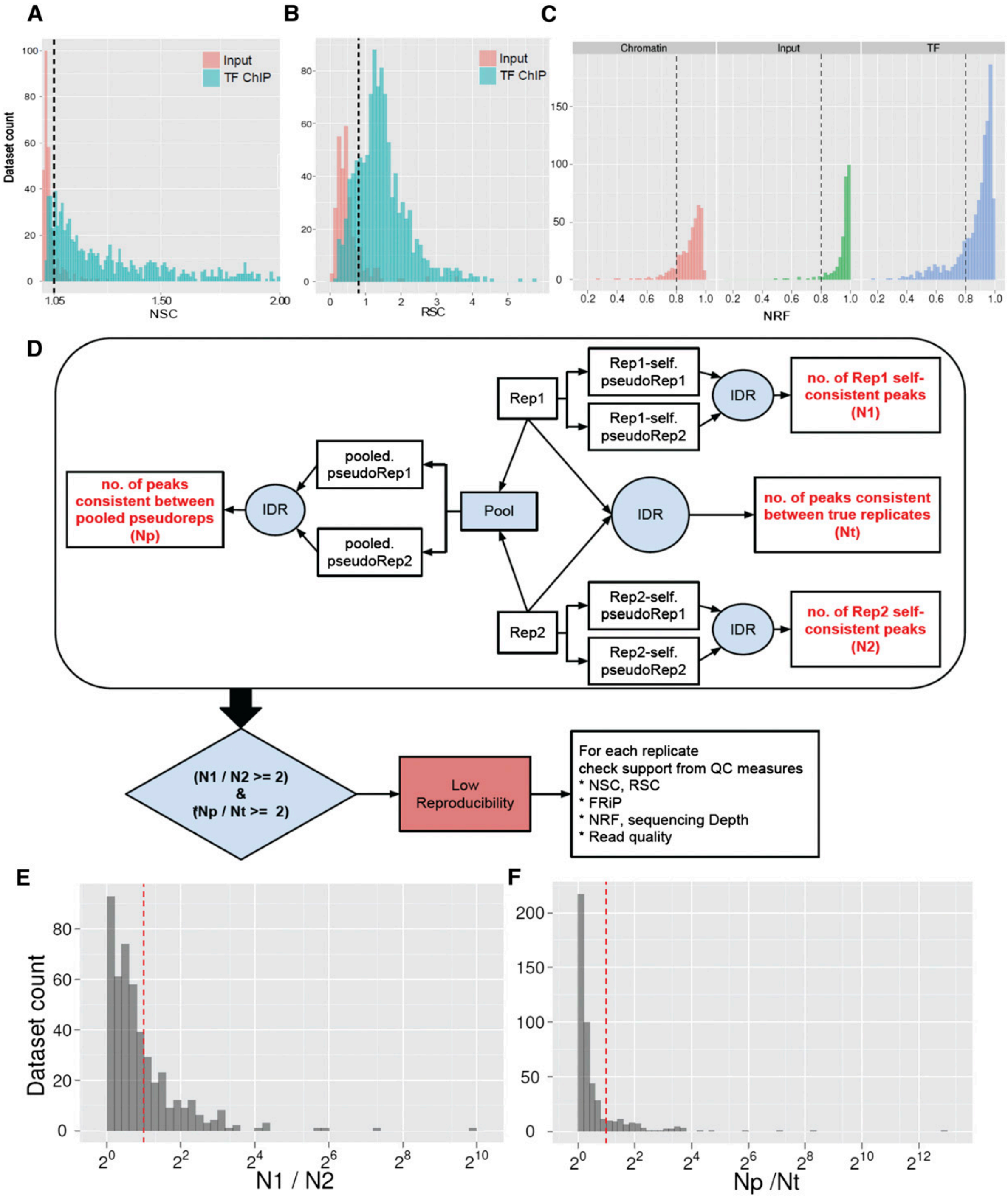


Figure 7. Analysis of ENCODE data sets using the quality control guidelines. (*A–C*) Thresholds and distribution of quality control metric values in human ENCODE transcription-factor ChIP-seq data sets. (*A*) NSC, (*B*) RSC, (*C*) NRF. (*D*) IDR pipeline for assessing ChIP-seq quality using replicate data sets. (*E*,*F*) Thresholds and distribution of IDR pipeline quality control metrics in human ENCODE transcription factor ChIP-seq data sets. (Dashed lines) Current ENCODE thresholds for the given metric, which are NSC > 1.05 (*A*); RSC > 0.8 (*B*); NRF > 0.8, N1/N2 ≥ 2 (where N1 refers to the replicate with higher N) (*E*); Np/Nt ≥ 2 (*F*).

How good does a ChIP-seq experiment need to be?

We have observed that some biologically important sites can have modest ChIP-seq signals (Fig. 4B), while some sites with very high enrichment fail to give positive functional readouts in follow-up experiments. Given this, the best practical guidance for setting thresholds of sensitivity, specificity, and reproducibility will depend on how the data are to be used. Below, we outline four different common ChIP uses, ranging from more relaxed to stringent in their requirements toward data quality and site-calling sensitivity.

Motif analysis

Deriving DNA sequence motifs for a ChIP-assayed factor is relatively simple and has been performed successfully for most ENCODE ChIP-seq data sets (Fig. 2E). Experiments that pass the thresholds we use for NRF, FRiP, and NSC/RSC typically produce thousands to tens of thousands of regions, a sub-sample of which can be readily used to deduce the recognition motif, although more than one motif subfamily is sometimes found by additional analysis (Johnson et al. 2007). Causal motifs are typically centrally positioned and this can be used as a confirming diagnostic (Fig. 6F). Notably, motif derivation can also be successful from marginal quality data that fall below recommended quality metric thresholds (especially if only the top-ranked peaks are used). However, the risk of artifacts increases, and results from such analyses should be cautiously interpreted and stringently validated.

Discovering regions to test for biological function such as transcriptional enhancement, silencing, or insulation

Biologists often use ChIP-seq data to identify candidate regulatory regions at loci of interest. When the goal is to find a few examples of regulatory domains bound by a factor, data of modest quality can still be useful if combined with close inspection of ChIP signals and the corresponding controls before investing in functional and/or mutagenesis studies. However, if the aspiration is to identify a comprehensive collection of all candidate regulatory regions bound by a factor, very high-quality and deeply sequenced data sets are required.

Deducing and mapping combinatoric occupancy

Typical *cis*-acting regulatory modules (CRM) are occupied by multiple factors (Ghisletti et al. 2010; Lin et al. 2010; Wilson et al. 2010; A He et al. 2011; Q He et al. 2011; Tijssen et al. 2011) and associated with multiple histone modifications (Barski et al. 2007; Mikkelsen et al. 2007; Wang et al. 2008). A frequent goal of ChIP-seq studies is to deduce a combination of factors that mediate a common regulatory action at multiple sites in the genome. This is a very quality-sensitive use of ChIP data since the presence of one or more weak data sets that fail to identify significant fractions of the true occupancy sites can seriously confound the analysis; therefore we recommend only the highest quality data sets be used for such analyses.

Integrative analysis

A new frontier of whole-genome analysis is the integration of data from many (hundreds or thousands) experiments with the goal of uncovering complex relationships. These endeavors typically use sophisticated machine learning methods (Ernst and Kellis 2010; Ernst et al. 2011; A Mortazavi, S Pepke, G Marinov, and B Wold, in prep.) with complex and varying sensitivity to ChIP strength; and such efforts can be very sensitive to data quality.

Conclusion

Our goal in developing these current working guidelines for ChIP-seq experiments, now applied over a large number of factors, was to provide information about experimental quality for users of modENCODE and ENCODE data. The strongest ChIP-seq data-sets that readily meet all quality specifications should be especially useful for regulatory network inference and for diverse integrative analyses, including the effects of genetic variation on human traits and disease. The metrics, methods, and thresholds might also be useful to the wider community, although our intention in outlining our approaches was not to imply that ENCODE criteria must be applied rigidly to all studies. As discussed above, some ChIP data and antibodies can and do fall outside these guidelines for varied reasons, yet are highly valuable. In such cases it is critical to try to understand why a data set looks unusual, and to assess the implications for specific uses of those data or reagents. Similar guidelines exist in ENCODE for RNA-seq, DNase-seq, FAIRE-seq, ChIA-PET, and other related assays; the working standards and protocols for these techniques can be found at the ENCODE and modENCODE websites (http://encodeproject.org/ENCODE/experiment_guidelines.html).

Data access

All data sets used in the analysis have been deposited for public viewing and download at the ENCODE (http://encodeproject.org/ENCODE/) and modENCODE (http://www.modencode.org/) portals.

List of affiliations

[1]Department of Genetics, Stanford University, Stanford, California 94305, USA; [2]Division of Biology, California Institute of Technology, Pasadena, California 92116, USA; [3]Department of Computer Science, Stanford University, Stanford, California 94305, USA; [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [5]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [6]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [7]Department of Statistics, University of California, Berkeley, California 94720, USA; [8]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts 02215, USA; [9]Computational Biology & Bioinformatics Program, Yale University, New Haven, Connecticut 06511, USA; [10]Department of Computer Science and Center for Systems Biology, Duke University, Durham, North Carolina 27708, USA; [11]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [12]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas 78701, USA; [13]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA; [14]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Boston, Massachusetts 02115, USA; [15]Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA; [16]Department of Statistics, Penn State University, University Park, Pennsylvania 16802, USA; [17]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA; [18]National Human Genome Research Institute/National Institutes of Health, Rockville, Maryland 20852, USA; [19]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada; [20]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06511, USA; [21]Department of Pathology, Stanford University, Stanford, California 94305, USA; [22]Department of Medicine, University of Washington, Seattle, Washington 98195, USA; [23]University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA; [24]Department of Biochemistry & Molecular Biology,

Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California 90089, USA; [25]Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.

Acknowledgments

We thank the ENCODE and modENCODE project consortia for helpful discussions and making their data available. We also thank Elise Feingold, Peter Good, and Kevin Struhl for advice and helpful discussions, and Leslie Adams and Caroline Kelly for help in preparing the manuscript.

References

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrançois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci* **106:** 14926–14931.

Barski A, Cuddapah S, Cui K, Roh T, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459:** 927–930.

Contrino S, Smith RN, Butano D, Carr A, Hu F, Lyne R, Rutherford K, Kalderimis A, Sullivan J, Carbon S, et al. 2011. modMine: Flexible access to modENCODE data. *Nucleic Acids Res* **40:** D1082–D1088.

DeKoter RP, Singh H. 2000. Regulation of B lymphocyte and macrophage development by graded expression of PU.1. *Science* **288:** 1439–1441.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36:** e105. doi: 10.1093/nar/gkn425.

Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, Cheung M, Day DS, Gadel S, Gorchakov AA, et al. 2011. An assessment of histone-modification antibody quality. *Nat Struct Mol Biol* **18:** 91–93.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Fernandez PC, Frank SR, Wang L, Schroeder M, Liu S, Greene J, Cocito A, Amati B. 2003. Genomic targets of the human c-Myc protein. *Genes Dev* **17:** 1115–1129.

Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ. 2010. ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS ONE* **5:** e15082. doi: 10.1371/journal.pone.0015082.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330:** 1775–1787.

Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei C, et al. 2010. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32:** 317–328.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.

He A, Kong SW, Ma Q, Pu WT. 2011. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci* **108:** 5632–5637.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat Genet* **43:** 414–420.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589.

Horak CE, Snyder M. 2002. ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol* **350:** 469–483.

Hua S, Kittler R, White KP. 2009. Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell* **137:** 1259–1271.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533–538.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26:** 1293–1300.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26:** 1351–1359.

Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, Green R, Farnham PJ. 2007. Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays. *J Biol Chem* **282:** 9703–9712.

Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, et al. 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* **6:** 377. doi: 10.1038/msb.2010.31.

Li Q, Brown J, Huang H, Bickel P. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5:** 1752–1779.

Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28:** 327–334.

Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, Mansson R, Sigvardsson M, Hagman J, Espinoza CA, Dutkowski J, et al. 2010. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat Immunol* **11:** 635–643.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.

Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, et al. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.

Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, Stathopoulos A. 2011. High resolution mapping of Twist to DNA in *Drosophila* embryos: Efficient functional analysis and evolutionary conservation. *Genome Res* **21:** 566–577.

Park PJ. 2009. ChIP-seq: Advantages and challenges of a maturing technology. *Nat Rev Genet* **10:** 669–680.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6:** S22–S32.

Poser I, Sarov M, Hutchins JRA, Hériché J, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW, et al. 2008. BAC TransgeneOmics: A high-throughput method for exploration of protein function in mammals. *Nat Methods* **5:** 409–415.

Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M. 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci* **107:** 3639–3644.

Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* **12:** R67. doi: 10.1186/gb-2011-12-7-r67.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4:** 651–657.

Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, et al. 2011. ENCODE whole-genome data in the UCSC Genome Browser: Update 2012. *Nucleic Acids Res* **40:** D912–D917.

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330:** 1787–1797.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27:** 66–75.

Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang S, Margueron R, Reinberg D, Green R, Farnham PJ. 2006. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* **16:** 890–900.

Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK, et al. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20:** 597–609.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5:** 829–834.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16:** 235–244.

Wilson NK, Foster SD, Wang X, Knezevic K, Schütte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, et al. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7:** 532–544.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137. doi: 10.1186/gb-2008-9-9-r137.

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HYK, Preston E, et al. 2010. Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet* **6:** e1000848. doi: 10.1371/journal.pgen.1000848.

Received December 10, 2011; accepted in revised form May 10, 2012.

E

Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide

Originally published as:

Raskatov JA, Nickols NG, Hargrove AE, Marinov GK, Wold B, Dervan PB. 2012. Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide. *Proc Natl Acad Sci U S A*. **109**(4):16041–16045. doi: 10.1073/pnas.1214267109.

Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide

Jevgenij A. Raskatov[a], Nicholas G. Nickols[a,b], Amanda E. Hargrove[a], Georgi K. Marinov[c], Barbara Wold[c], and Peter B. Dervan[a,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; [b]Department of Radiation Oncology, David Geffen School of Medicine of the University of California Los Angeles, Los Angeles, CA 90095-6951; and [c]Division of Biology, California Institute of Technology, Pasadena, CA 91125

Contributed by Peter B. Dervan, August 20, 2012 (sent for review July 10, 2012)

Gene regulation by DNA binding small molecules could have important therapeutic applications. This study reports the investigation of a DNA-binding pyrrole-imidazole polyamide targeted to bind the DNA sequence 5′-WGGWWW-3′ with reference to its potency in a subcutaneous xenograft tumor model. The molecule is capable of trafficking to the tumor site following subcutaneous injection and modulates transcription of select genes in vivo. An FITC-labeled analogue of this polyamide can be detected in tumor-derived cells by confocal microscopy. RNA deep sequencing (RNA-seq) of tumor tissue allowed the identification of further affected genes, a representative panel of which was interrogated by quantitative reverse transcription-PCR and correlated with cell culture expression levels.

tumor RNA-sequencing | eXpress | in vivo circulation | efficacy

Pyrrole-imidazole (Py-Im) polyamides represent a class of modular DNA minor groove binders with affinity and specificity comparable to the values observed with typical DNA binding proteins (1, 2). Our previous investigations have established a framework for molecular recognition of the minor groove of DNA by polyamides that can target predetermined DNA binding sites (3–5). Cell culture experiments have shown that cellular uptake of Py-Im polyamides targeting six-base pair sequences can be observed (6). Subsequent studies demonstrated that Py-Im polyamides could antagonize DNA binding of transcription factors in live cells. Interrogated transcription factors include the androgen receptor (AR) (7), hypoxia inducible factor 1 alpha (HIF-1α) (8), the glucocorticoid receptor (GR) (9), and nuclear factor kappa B (NF-κB) (10).

Although there is more knowledge to be gained from deeper genome-wide cell culture studies, the next frontier for Py-Im polyamides as medicinally relevant small molecules lies in in vivo applications. Our recent studies demonstrated that the pharmacokinetics and toxicity of Py-Im polyamides in mice depend on architecture (11). Micromolar levels of compounds were observed in mouse plasma for up to 48 h following either intraperitoneal (i.p.) or subcutaneous (s.c.) injection. Efforts of Nagashima et al. established that Py-Im polyamides of different architecture were detectable in rat serum several hours after intravenous (i.v.) administration (12). Matsuda et al. further showed that a Py-Im polyamide targeted to the TGF-β1 promoter affected target gene expression in vivo (rat renal cortex) without evidence of systemic toxicity (13, 14). The present study focuses on the question of whether Py-Im polyamides affect gene expression in vivo, specifically in a xenograft model environment, employing a luciferase-expressing derivative of the commonly used lung nonsmall cell carcinoma line A549.

Results

Acetylated Py-Im Polyamide 1 is More Potent in Cell Culture Than the Analog 2. The first set of experiments compared the in vitro gene regulation activity of Py-Im polyamides **1** and **2**, both targeted to bind to the sequence 5′-WGGWWW-3′ (Fig. 1*A*). Our previous efforts established that the polyamide **2** was capable of modulating a subset of TNF-inducible genes (10). Among the strongly affected genes we had identified *CCL2* and *SERPINE1* as highly repressed targets of **2**.

The basal expression levels of *CCL2* and *SERPINE1* were sufficiently high to enable the study of polyamide effects in the uninduced state. We found that both **1** and **2** reduced the levels of the two transcripts, but the effects exerted by **1** were substantially more pronounced (Fig. 1*B*). In line with the previous study, prolonged incubation times resulted in stronger down-regulation of the target genes—up to fivefold with *CCL2* and 14-fold with *SERPINE1*. Furthermore, **1** was significantly more cytotoxic in vitro than **2** against the chosen cell line with IC_{50} values of 13 ± 5 μM and 33 ± 2 μM, respectively (*SI Text*, Fig. S1*A*). The more potent Py-Im polyamide **1** was, therefore, chosen for in vivo gene regulation experiments. Cellular uptake measurements clearly showed that the FITC-labeled analogue **3** was readily taken up by A549-luc-C8 cells, resulting in characteristic nuclear fluorescence (*SI Text*, Fig. S1*B*).

Py-Im Polyamides 1 and 2 Reach Comparable Plasma Levels with Similar Circulation Times Following S.C. Injection. Prior to conducting in vivo tumor xenograft experiments the pharmacokinetic profiles of **1** and **2** were compared. Our previous investigations showed that **2** could circulate in wild-type mice for several hours at micromolar plasma concentrations but dropped below the limit of detection after 24 h (11). The compound was administered by either the s.c. or the i.p. route and blood collected retro-orbitally. The circulation experiment was conducted for the Py-Im polyamide **1** using subcutaneous administration conditions analogous to those previously reported for **2**. The observed plasma levels compared well with those reported for **2** (Fig. 2 and Fig. S2). Maximum plasma concentrations of 10 μM were attained for both compounds 3 h post injection. The plasma elimination phase appeared slightly shallower for the acetylated Py-Im polyamide **1** than for its close analog **2**, but neither was detectable 24 h post injection.

FITC-Labeled Py-Im Polyamide 3 Can Be Detected in Xenograft-Derived Cell Nuclei. We proceeded to synthesize the fluorescent tagged derivative of **1**, Py-Im polyamide **3** (see *SI Text*, Fig. S1 for structure). Previous experiments had shown that a closely related compound was stable in vivo and circulated in mice for several

Author contributions: J.A.R., and P.B.D. designed research; J.A.R., N.G.N., and A.E.H. performed research; J.A.R., N.G.N., A.E.H., G.K.M., B.W., and P.B.D. analyzed data; and J.A.R., and P.B.D. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE40218).

[1]To whom correspondence should be addressed. E-mail: dervan@caltech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1214267109/-/DCSupplemental.

CHEMISTRY

MEDICAL SCIENCES

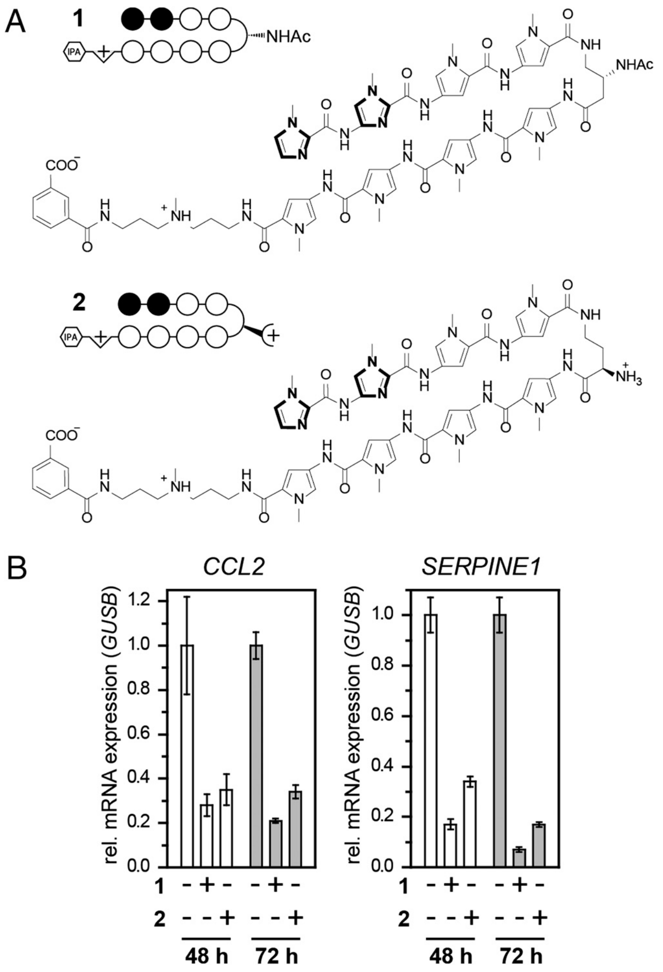


Fig. 1. (*A*) Hairpin Py-Im polyamides **1** and **2**. (*B*) In vitro qRT-PCR (A549-luc-C8 cell culture). Cells were incubated with 10 μM final 1 or 2 for 48 h or 72 h, where indicated. All treatments were conducted with 0.1% DMSO as vehicle.

hours (15). The resultant mouse plasma was found to contain the compound at micromolar concentrations and could be used to produce characteristic nuclear staining of A549 cells in culture (15).

Immunocompromised mice (SCID-beige) were grafted subcutaneously (in the flank) with the commercially available A549-derived luciferase expressing cell line A549-luc-C8 (see *Materials and Methods* for details). In order to ensure that the Py-Im polyamide **3** was entering the tumors through the vascular system, the animals were injected with the polyamide from a site distal to the site of implantation. A representative experiment is depicted in Fig. 3*A*. The tumor-derived cells from the treated animals were found to display strong and characteristic nuclear staining, closely resembling those in the cell culture experiments. Tumors from vehicle-treated mice were prepared and found to be devoid of nuclear fluorescence. This finding provided the impetus to perform treatment of xenografted animals with **1** and investigate whether polyamide treatment could result in gene expression changes of *CCL2* and *SERPINE1* in vivo.

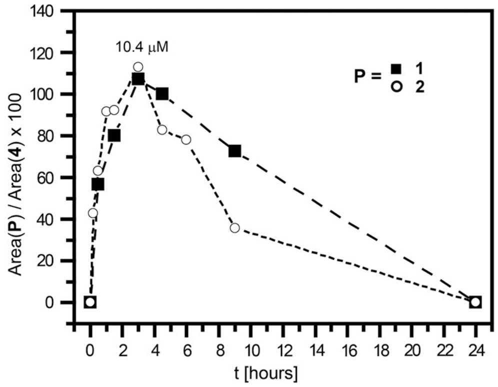


Fig. 2. Plasma values of **1** and **2** as obtained from analytical HPLC traces (C57Bl/6 wild-type mice, four animals per data point, all injections were done subcutaneously at 120 nmol/animal). The levels were normalized to the internal reference **4** (Fig. S2). Datapoints shown for Py-Im polyamide **2** have been previously reported (11).

Py-Im Polyamide 1 Represses *CCL2* and *SERPINE1* Transcription in Vivo. We followed up by testing the potency of **1** to repress *CCL2* and *SERPINE1* in the tumor xenograft setting. To ensure primer selectivity towards human target genes, we isolated total RNA from mouse spleens obtained from the SCID-beige strain and conducted control quantitative reverse transcription-PCR (qRT-PCR) experiments. None of the primers employed in this study exhibited any substantial amplification of mouse RNA.

All experiments were performed in accord with the treatment schedule displayed in Table S1 (*SI Text*) and following the general humane endpoints criteria (see *Materials and Methods*). Mild animal toxicity was observed with an overall weight loss not exceeding 10% as a result of treatment. The transcript levels of *CCL2* and *SERPINE1* were reduced by a factor of 2.3 and 2.0, respectively, by **1** (Fig. 3*B*). Gene expression changes were the same whether normalized to *GUSB* or *PPIA* as the housekeeping gene. Because the IC_{50} of Py-Im polyamide **1** against growth of A549-luc-C8 was 13 ± 5 μM and plasma levels of the compound up to

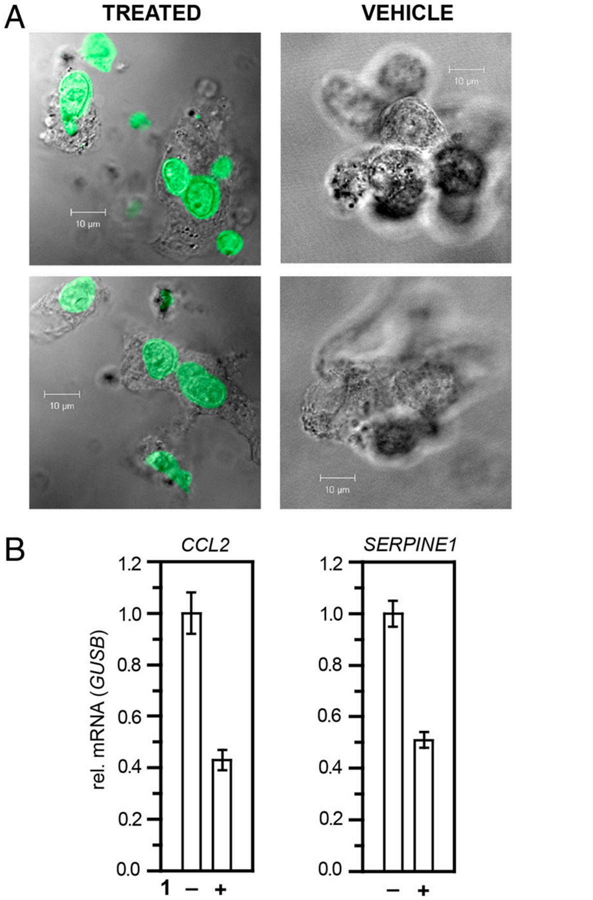


Fig. 3. (*A*) FITC-labeled Py-Im polyamide **3** localizes to engrafted A549-luc-C8 cells (SCID-beige mice). (*B*) qRT-PCR of tumor samples showing repression of *CCL2* and *SERPINE1*. Three independent experiments with *N* = 5 animals per treatment condition (vehicle vs **1**) were averaged.

10 μM were attainable for several hours post injection, it was conceivable that **1** could affect tumor growth. Tumor size was therefore assessed by luciferase imaging as outlined in *Materials and Methods*. A linear correlation between tumor size and photon number over several orders of magnitude has been previously demonstrated for the cell line used (www.caliperls.com/assets/018/7635.pdf). The luciferase output remained within experimental error between the two groups, suggesting that the gene expression changes did not stem from cytotoxicity (*SI Text*, Fig. S3).

Genome-Wide Effects of the Py-Im Polyamide 1. In order to establish the global effects of **1** in a xenograft setting, we measured changes in gene expression using RNA-seq in tumors from treated and untreated mice (see *Materials and Methods* for details). As our RNA-seq libraries contained a mixture of human and mouse RNA derived from the xenograft as well as the host cells infiltrating it, we faced the challenge of accurately determining the transcripts and genes from which sequencing reads originate (Table S2 and discussion in the *SI Text*). We therefore designed an analysis pipeline based upon mapping reads to a combined human and mouse transcriptome and using the recently developed eXpress software package (bio.math.berkeley.edu/eXpress/index.html) to quantify probabilistically transcript abundance for both species simultaneously (Fig. 4). The eXpress output was used as input for differential expression analysis using DESeq (16).

Out of 22,092 genes, 618 (2.8%) experienced a statistically significant change in expression at a confidence level of $p < 0.05$. Within this subpopulation, 115 (0.52%) genes were repressed at least twofold, whereas 53 genes (0.24%) showed at least a twofold up-regulation. For quality control purposes, one replicate was resequenced using paired-end read sequencing with the read length set at 100 nt. High correlation coefficients were determined between the effective counts obtained by single- and paired-end read sequencing, with R^2 values of 0.97 and 0.94 for vehicle and **1**, respectively (see *SI Text*, Fig. S4 for correlation plots).

Comparison of RNA-seq and qRT-PCR for a Panel of Selected Genes in Vivo. A representative panel of genes studied by RNA-seq was further interrogated by qRT-PCR (Fig. 5, *Upper* and Table 1). In addition to *CCL2* and *SERPINE1* that were discussed above, we investigated the effects of **1** on transcription of *NPTX1*, *ROBO1*, *ATM*, *EGFR*, and *MMP28*. The genes were selected so as to range from strongly repressed (*NPTX1*) through weakly down-regulated (*ATM* and *EGFR*) to up-regulated upon polyamide treatment (*MMP28*). *NPTX1* experienced a 3.3-fold repression upon treatment with **1**, whereas the expression of *ATM* was reduced only 1.5-fold. The expression changes in *EGFR* detected by qPCR lie close to the error of the experiment (1.2-fold down). The expression of *MMP28* on the other hand was up-regulated 1.5-fold upon treatment with the Py-Im polyamide **1**. The genes *CCL2*, *NPTX1*, *SERPINE1*, and *MMP28* were categorized as differentially expressed by both techniques (Table 1). Changes in expression of *ATM* and *ROBO1* were only statistically significant assessed by qRT-PCR, not by RNA-seq (*p*-values over 0.05)

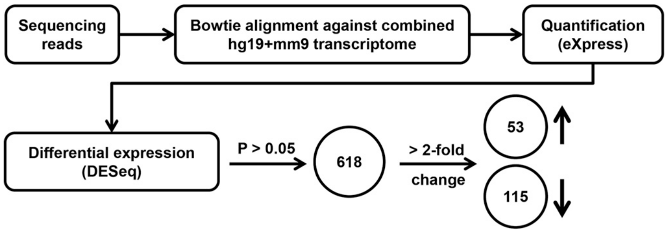

Sequencing reads
Bowtie alignment against combined hg19+mm9 transcriptome
Quantification (eXpress)
Differential expression (DESeq)
P > 0.05
618
> 2-fold change
53
115

Fig. 4. Schematic representation of the pipeline for RNA-seq analysis of tumor-derived RNA. Three independent experiments for each of which $N = 5$ animals per treatment condition (vehicle vs **1**) were averaged, were jointly analyzed.

Comparison of in Vivo and in Vitro Effects of 1 by qRT-PCR on a Panel of Selected Genes. The gene expression changes in the in vivo xenograft setting were compared to those observed in cell culture (Fig. 5, *Lower* and Table S3). Prolonged incubation with Py-Im polyamide **1** in cell culture generally led to more pronounced effects (48 h vs 72 h), the only exception being *MMP28*, for which no effect was observed in cell culture regardless of the incubation time. The correspondence between the in vivo experiment and the cell culture control was found to depend strongly on the transcript interrogated. The in vitro effect of **1** on *NPTX1* expression at 72 h incubation was very close to that observed in vivo (3.5-fold *vs* 3.3-fold), whereas for *CCL2* the gene repression in xenografts resembled more closely the 48 h incubation time point from cell culture experiments (2.3-fold vs 2.2-fold). While *MMP28* expression was unchanged in cell culture, all other interrogated genes were affected more strongly than in the xenograft setting. The largest difference was noted for *SERPINE1*, which was repressed 2.0-fold in vivo but experienced a down-regulation in cell culture amounting to as much as 15.7-fold.

Discussion

The present study shows that the polyamide **1** is capable of trafficking to a xenografted tumor and yielding measurable gene expression changes. Following the establishment of pharmacokinetic properties of Py-Im polyamides targeted to the sequence 5′-WGGWWW-3′ (11), this is the next important step towards the application of Py-Im polyamides in a setting relevant to disease.

Comparison Between Xenografts and Cell Culture. Quantitative correlation between the two settings is of high interest, but differences in exposure times and concentrations of the Py-Im polyamide **1** between cell culture and at the tumor site need to be kept in mind. Typical exposure times in cell culture range from 48 h to 72 h whereas final treatment concentrations do not exceed 10 μM (10). Most of the polyamide remains in the medium so that the concentration is effectively invariant over the experimental timecourse. One fundamental difference in the in vivo experiment is that the serum concentration of **1** does change as a function of postinjection time. Whereas a concentration maximum of approximately 10 μM is typically attained under chosen administration conditions, the circulating levels of **1** drop below the level of detection (high nanomolar) 24 h postinjection. This results in oscillatory compound levels over the course of the 10 d experiment (Fig. 2 and Table S1). Another difference is the inherent heterogeneity of cancerous tissue. Some subpopulations of xenografted cells lie in closer proximity to newly formed blood vessels and hence may be more readily accessible to the drug than others (17, 18). Interactions with the host may also lead to additional complexity (19).

Comparison of the three genes that were most strongly affected in the in vivo experiment to their behavior in cell culture is of interest. Among the genes that were examined by qRT-PCR, *NPTX1* experienced the strongest in vivo repression (3.3-fold down). This was similar to the effects observed in cell culture, namely 2.6-fold and 3.5-fold repression at 48 h and 72 h, respectively. The effect of the Py-Im polyamide **1** against cells in culture was rather similar for both exposure times tested. By contrast, *SERPINE1* was less strongly affected in vivo compared to in vitro. While the in vivo repression amounted to 2.0-fold, the down-regulation was substantially more pronounced in cell culture. Transcription was reduced 8.3-fold after 48 h incubation and 15.7-fold after 72 h. Expression of *CCL2* was down-regulated 2.3-fold in the xenograft experiment whereas the cell culture repression was 2.2-fold (48 h) and 4.4-fold (72 h). This comparative analysis prompts a note of caution, for it is evident that there can be significant variability between gene expression changes observed in vitro and in vivo. We conclude, however, that cell culture data can be used to support in vivo findings in most cases.

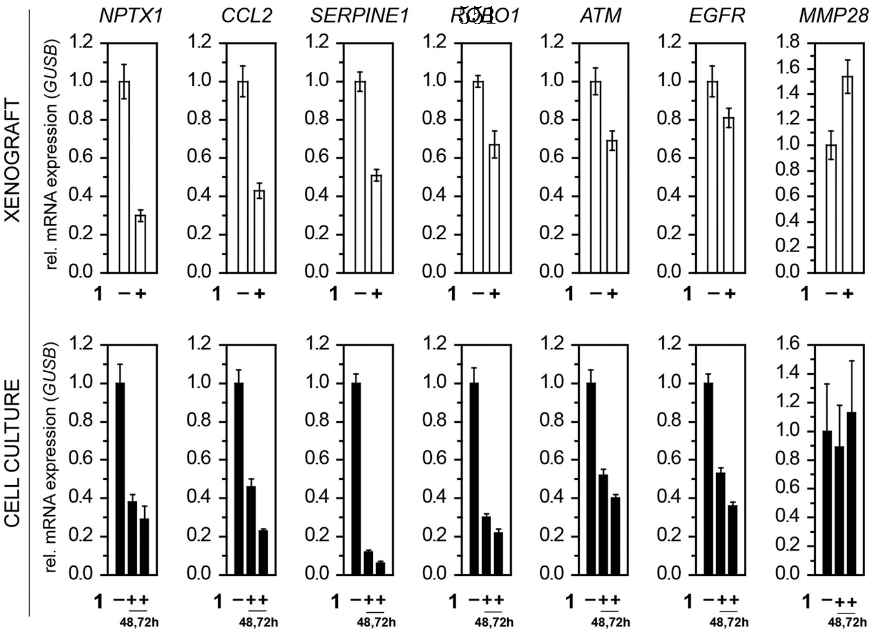


Fig. 5. A panel of genes affected by **1** in an A549-luc-C8 xenograft in SCID-bg animals (*Upper*) and cell culture (*Lower*). Xenograft: three independent experiments with $N = 5$ animals per treatment condition (vehicle vs **1**) were averaged. Cell culture: where indicated, the cells were incubated with Py-Im polyamide **1** at 10 μM final concentration in 0.1% DMSO as vehicle.

Tumor RNA-seq. Because of tumor heterogeneity, stemming mostly from host-derived tumor infiltrating cells, the fraction of sequencing reads unambiguously originating from the human transcriptome was at most only 60%, the rest being mouse-derived (see *SI Text*, Table S2). The computational pipeline described here solves this problem by applying simultaneous probabilistic mapping to both the human and the mouse transcriptome. Moreover, we have confirmed the viability of this approach by conducting qRT-PCR on a representative panel of genes, showing good correlation between the two methods (Table 1) and we expect it to be widely useful to researchers conducting similar types of experiments in different settings. Genome-wide analysis showed a total of 168 genes to be affected by the Py-Im polyamide **1** in xenografts, which corresponds to 0.76% of the NCBI reference sequence (refSeq) annotation ($p < 0.05$, at least twofold change). For comparison, Matsuda et al. reported gene expression changes in rat kidney cortex for 3% of genes interrogated by microarray (14).

Effects on Tumor Size. The tumor sizes were the same (within error) between the animal groups that received repeated injections of Py-Im polyamide **1** and vehicle (Fig. S3). The absence of any significant effect on tumor size could be due to a variety of factors. The compound might not reach sufficient average levels in the tumor. The IC_{50} value of **1** is 13 ± 5 μM (Sulforhodamine B assay, 72 h incubation, 24 h recovery; see also *SI Text*, Fig. S1*A*). Although micromolar levels of **1** can be maintained for several hours postinjection, the overall exposure to the compound may still be too low to produce any measurable effect on size. Treatment efficiency could be enhanced by using more potent Py-Im polyamides or changing the route of administration, e.g., by employing osmotic pumps to maintain steady compound levels over the course of the experiment (20). Alternatively, Py-Im polyamide **1** may not penetrate the tumor to a sufficient depth because of tissue inhomogeneity. Tissue penetration rates can depend on compound lipophilicity and flexibility. Py-Im polyamide substituent variation affords a means to alter binding site preference, affinity, specificity, lipophilicity, and cellular uptake rates (21). Finally, the treatment schedule may be too short. Initial tumor growth is rather slow, the A549-luc-C8 tumors typically entering the exponential growth phase only several weeks after grafting (www.caliperls.com/assets/018/7635.pdf).

Table 1. Comparison of qRT-PCR and RNA-seq of A549-luc-C8 tumor xenograft gene expression levels normalized to GUSB as the housekeeping gene (qRT-PCR). Brackets indicate gene upregulation upon treatment. Three independent experiments with $N = 5$ animals per treatment condition (vehicle vs 1) were averaged. RNA-seq was performed with single-end reads of 50 nt length. See *SI Text* for annotation of these gene products

Gene	Fold change (qPCR)	Fold change (RNA-seq)
ATM	1.5 ± 0.2	1.5 ($p > 0.05$)
NPTX1	3.3 ± 0.6	2.9 ($p < 0.001$)
ROBO1	1.5 ± 0.2	1.7 ($p > 0.05$)
MMP28	[1.5 ± 0.3]	[2.0] ($p < 0.05$)
EGFR	1.2 ± 0.2	1.3 ($p > 0.05$)
CCL2	2.3 ± 0.4	1.7 ($p < 0.001$)
SERPINE1	2.0 ± 0.2	1.8 ($p < 0.001$)

Conclusions

This study reports the ability of Py-Im polyamide **1** and its fluorescent labeled analogue **3** to traffic to the subcutaneously grafted A549-luc-C8 tumor. Unambiguous nuclear staining of tumor-derived cells with the FITC-analogue **3** evidenced the ability of the compound to remain at the site several days after injection. The nonfluorescent parent Py-Im polyamide **1** was capable of affecting gene expression in the tumor, and most trends correlated satisfactorily with cell culture data. From the panel of genes examined by qRT-PCR, the strongest effect was measured for *NPTX1*, which was repressed 3.3-fold. *MMP28* on the other hand experienced a small but significant induction of 1.5-fold upon treatment. It is of the highest importance to increase the potency of a compound at the tumor site, while minimizing its toxic effects to the host. Strategies to that end include testing of Py-Im polyamides targeted to different sequences, incorporating further modifications, development of formulations that would enhance selectivity of delivery and testing of alternative treatment schedules.

Materials and Methods

Polyamide Synthesis and Characterization. The polyamides **1**–**3** were synthesized following modified solid phase synthesis protocols (22). Typically, yields between 25 and 40% were observed. Compound purities were confirmed by analytical HPLC. Compounds **1** and **3** were characterized by MALDI-TOF MS as singly protonated species. Following masses were determined: **1** calculated for $C_{67}H_{79}N_{22}O_{13}$ $[M + H]^+$ 1,399.6, found 1,399.5; **3** calculated for $C_{80}H_{86}N_{23}O_{15}S$ $[M + H]^+$ 1,640.6, found 1,642.3. Analytical data for **2** were in agreement with what has been previously reported (10).

In Vitro Cell Culture Experiments. All experiments were conducted with A549-luc-C8 cells, unless specifically mentioned otherwise. Cells were grown in RPMI medium 1640, which was supplemented with 10% FBS and 1% penicillin/streptomycin, and did not exceed 25 passages. Confocal imaging, cellular proliferation and viability experiments as well as gene expression analyses by quantitative RT-PCR were performed following our previously published protocols (7, 10, 21, 23). Gene expression was normalized against *GUSB* as housekeeping gene. All primers yielded single amplicons as determined by both melting denaturation analysis and agarose gel electrophoresis. The following primer pairs were used. *CCL2*: fwd 5′-AGT GTC CCA AAG AAG CTG TGA-3′ rev. 5′-AAT CCT GAA CCC ACT TCT GCT-3′; *SERPINE1*: fwd. 5′-AGA ACA GGA GGA GAA ACC CA-3′ rev. 5′-AGC TCC TTG TAC AGA TGC CG-3′ *GUSB*: fwd. 5′-CTC ATT TGG AAT TTT GCC GAT T-3′ rev. 5′-CCC AGT GAA GAT CCC CTT TTT A-3′. *ATM*: fwd. 5′-GCT GTG AGA AAA CCA TGG AA-3′ rev. 5′-TTC AAA GGA TTC ATG GTC CAG-3′; *EGFR*: fwd. 5′-GGG CTC TGG AGG AAA AGA AA-3′ rev. 5′-TCC TCT GGA GGC TGA GAA AA-3′; *MMP28*: fwd. 5′-CCT GCA GCT GCT ACT GTG G-3′ rev. 5′-CTT TGG GGA CCT GTT CAT TG-3′; *NPTX1*: fwd. 5′-ACC GAG GAG AGG GTC AAG AT-3′ rev. 5′-GTG GGA ATG TGA GCT GGA AC-3′; *ROBO1*: fwd. 5′-CAA TGC ATC GCT GGA AGT AG-3′ rev. 5′-TTC TTC CAT GAA ATG GTG GG-3′.

Mouse Experiments. ***Pharmacokinetics.*** Analyses for Py-Im polyamide **1** were conducted following our recently established protocols (11). Briefly, the compound was injected subcutaneously into C57/Bl6 mice as a PBS/DMSO solution (4∶1, 200 μL per injection, four animals per group). Blood was collected retro-orbitally at the indicated time points. Plasma was obtained by centrifugation, precleared from protein by methanol precipitation and compound levels determined by analytical HPLC. The plasma levels obtained were compared with those previously reported for **2**. ***Xenografts.*** *Grafting with A549-luc-C8.* Experiments were performed in female SCID-beige mice (Charles River) between 8 and 12 wk of age. Cells were injected into the left flank area of the animals as suspensions of $25 \times 10^6\ mL^{-1}$ in RPMI, 200 μL per injection. *Treatment and tumor proliferation monitoring*. Mice were treated following the schedule delineated in *SI Text* (Table S1). Tumor proliferation was monitored using the XENOGEN imaging device. The animals were anesthetized with 2–5% isoflurane and subsequently transferred to the imaging chamber, whereupon the isoflurane levels were reduced to 1–2.5%. The floor of the imager was heated to +37 °C to avoid hypothermia. Breathing frequency was monitored and not allowed to drop below $1\ s^{-1}$, adjusting the isoflurane levels accordingly at all times. *Endpoint criteria and euthanasia*. Animal endpoint criteria encompassed weight loss of over 15%, restriction of motor function by the engrafted tumor, dehydration of over 10%, and moribund behavior. Where appropriate, the animals were euthanized by asphyxiation in a CO_2 chamber. ***Tumor tissue harvest***. Animals were resected and tumors excised using standard forceps, scissors, and surgical blades. The tumors were combined into one sample per condition and mechanically sheared in TRIzol, employing a specialized device (tissue tearer, model 985370). Total RNA workup was performed following the standard TRIzol procedure, followed by a DNAse digest.

RNA-seq Sample Preparation and Data Processing. Double polyA-selection was used in order to enrich for mRNA. RNA-seq libraries were prepared using standard Illumina reagents and protocols (24) All experiments were carried out in triplicate and 35 million–50 million single-end sequences of 50 bp were generated for each library. One replicate was additionally sequenced as 100 bp paired-end reads for quality control purposes. Sequencing data were mapped to a combined human and mouse transcriptome index (using the hg19 and mm9 refSeq annotations) using Bowtie version 0.12.7 (25) with two mismatches and an unlimited number of locations a read can map to. Alignments were quantified on the transcript level using eXpress 1.0.0 (bio.math.berkeley.edu/eXpress/index.html); for each gene the quantification values of all its transcripts were summed and the eXpress-determined "effective counts" were used as input for differential expression analysis using DESeq (16).

ACKNOWLEDGMENTS. J.A.R. is grateful to the Alexander von Humboldt foundation for the award of a Feodor Lynen postdoctoral fellowship. We thank Michael J. Waring for careful editing of the manuscript. N.G.N. would like to acknowledge the National Institutes of Health (NIH) and the Jonsson Cancer Center Foundation (JCCF) at UCLA for postdoctoral support. A.E.H. thanks the California Tobacco-Related Disease Research Program (19FT-0105) and the NIH (NRSA 1F32CA156833) for postdoctoral support. Sequencing was conducted at the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology. We are grateful for support by The Ellison Medical Foundation (AG-SS-2256-09) and the National Institutes of Health (GM051747).

1. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr Opin Struct Biol* 13:284–299.
2. Hsu CF, et al. (2007) Completion of a programmable DNA-binding small molecule library. *Tetrahedron* 63:6146–6151.
3. White S, Szewczyk JW, Turner JM, Baird EE, Dervan PB (1998) Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands. *Nature* 391:468–471.
4. Kielkopf CL, et al. (1998) A structural basis for recognition of A.T and T.A base pairs in the minor groove of B-DNA. *Science* 282:111–115.
5. Chenoweth DM, Dervan PB (2009) Allosteric modulation of DNA by small molecules. *Proc Natl Acad Sci USA* 106:13175–13179.
6. Edelson BS, et al. (2004) Influence of structural variation on nuclear localization of DNA-binding polyamide-fluorophore conjugates. *Nucleic Acids Res* 32:2802–2818.
7. Nickols NG, Dervan PB (2007) Suppression of androgen receptor-mediated gene expression by a sequence-specific DNA-binding polyamide. *Proc Natl Acad Sci USA* 104:10418–10423.
8. Olenyuk BZ, et al. (2004) Inhibition of vascular endothelial growth factor with a sequencespecific hypoxia response element antagonist. *Proc Natl Acad Sci USA* 101:16768–16773.
9. Muzikar KA, Nickols NG, Dervan PB (2009) Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression. *Proc Natl Acad Sci USA* 106:16598–16603.
10. Raskatov JA, et al. (2012) Modulation of NF-κB-dependent gene transcription using programmable DNA minor groove binders. *Proc Natl Acad Sci. USA* 109:1023–1028.
11. Raskatov JA, Hargrove AE, So AY, Dervan PB (2012) Pharmacokinetics of Py-Im polyamides depend on architecture: Cyclic versus linear. *J Am Chem Soc* 134:7995–7999.
12. Nagashima T, et al. (2009) Pharmacokinetic modeling and prediction of plasma pyrrole-imidazole polyamide concentration in rats using simultaneous urinary and biliary excretion data. *Biol Pharm Bull* 32:921–927.
13. Matsuda H, et al. (2006) Development of gene silencing pyrrole-imidazole polyamide targeting the TGF-beta1 promoter for treatment of progressive renal diseases. *J Am Soc Nephrol* 17:422–432.
14. Matsuda H, et al. (2011) Transcriptional inhibition of progressive renal disease by gene silencing pyrrole-imidazole polyamide targeting of the transforming growth factor-beta 1 promoter. *Kidney Int* 79:46–56.
15. Hargrove AE, Raskatov JA, Meier JL, Montgomery DC, Dervan PB (2012) Characterization and solubilization of pyrrole-imidazole polyamide aggregates. *J Med Chem* 55:5425–5432.
16. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106.
17. Carmeliet P, Jain RK (2000) Angiogenesis in cancer and other diseases. *Nature* 407:249–257.
18. Vaupel P, Kallinowski F, Okunieff P (1989) Blood-flow, oxygen and nutrient supply, and metabolic microenvironment of human-tumors—a review. *Cancer Res* 49:6449–6465.
19. Bankert RB (2001) Human-SCID mouse chimeric models for the evaluation of anti-cancer therapies. *Trends Immunol* 22:386–393.
20. Song P, et al. (2008) Activated cholinergic signaling provides a target in squamous cell lung carcinoma. *Cancer Res* 68:4693–4700.
21. Meier JL, Montgomery DC, Dervan PB (2012) Enhancing the cellular uptake of Py-Im polyamides through next-generation aryl turns. *Nucleic Acids Res* 40:2345–2356.
22. Puckett JW, Green JT, Dervan PB (2012) Microwave assisted synthesis of Py-Im polyamides. *Org Lett* 14:2774–2777.
23. Nickols NG, Jacobs CS, Farkas ME, Dervan PB (2007) Improved nuclear localization of DNA-binding polyamides. *Nucleic Acids Res* 35:363–370.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
25. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

F

Antitumor activity of a pyrrole-imidazole polyamide

Originally published as:

Yang F, Nichols NG, Li BC, Marinov GK, Said JW, Dervan PB. 2013. Antitumor activity of a pyrrole-imidazole polyamide. *Proc Natl Acad Sci U S A* **110**(5):1863–1868. doi: 10.1073/pnas.1222035110.

Antitumor activity of a pyrrole-imidazole polyamide

Fei Yang[a], Nicholas G. Nickols[a,b], Benjamin C. Li[a], Georgi K. Marinov[c], Jonathan W. Said[d], and Peter B. Dervan[a,1]

[a]Division of Chemistry and Chemical Engineering, and [c]Division of Biology, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Radiation Oncology and [d]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine of the University of California, Los Angeles, CA 90095

Contributed by Peter B. Dervan, December 17, 2012 (sent for review October 19, 2012)

Many cancer therapeutics target DNA and exert cytotoxicity through the induction of DNA damage and inhibition of transcription. We report that a DNA minor groove binding hairpin pyrrole-imidazole (Py-Im) polyamide interferes with RNA polymerase II (RNAP2) activity in cell culture. Polyamide treatment activates p53 signaling in LNCaP prostate cancer cells without detectable DNA damage. Genome-wide mapping of RNAP2 binding shows reduction of occupancy, preferentially at transcription start sites, but occupancy at enhancer sites is unchanged. Polyamide treatment results in a time- and dose-dependent depletion of the RNAP2 large subunit RPB1 that is preventable with proteasome inhibition. This polyamide demonstrates antitumor activity in a prostate tumor xenograft model with limited host toxicity.

minor groove binder | small molecule transcription inhibitor | ChIP-Seq

Several chemotherapeutics, including the anthracyclines and cisplatin, exert part of their cytotoxicity through the inhibition of transcription (1). Transformed cells often require constant expression of antiapoptotic genes for survival, making transcription inhibition a relevant therapeutic strategy in oncology (1, 2). Many radio- and chemotherapy treatments that target DNA, including UV irradiation, cisplatin, and the topoisomerase inhibitors, introduce obstacles to RNA polymerase II (RNAP2) elongation by generating bulky or helix-distorting lesions (3–5). In cell culture experiments, transcription blockade has been shown to induce degradation of the RNAP2 large subunit (RPB1), and function as a signal for p53-mediated apoptosis (6, 7). Although many DNA-targeted therapeutics effectively inhibit transcription and induce apoptosis, clinical treatment with genotoxic agents can also damage DNA in normal cells, increasing symptomatic toxicity and potentially leading to secondary cancers (8). The question arises whether high-affinity, noncovalent DNA-binding ligands offer an approach to transcription inhibition without DNA damage.

Hairpin pyrrole-imidazole (Py-Im) polyamides are synthetic oligomers with programmable sequence recognition that bind the minor groove of DNA with high affinity (9). Py-Im polyamide-DNA binding induces allosteric changes in the DNA helix that can interfere with protein–DNA interactions (10, 11). Py-Im polyamides have been used as molecular probes in cell culture to modulate inducible gene-expression pathways (12–15). In rodents, eight-ring hairpin Py-Im polyamides circulate in blood for several hours after administration and affect changes in gene expression in tissues (16–18).

We have previously reported that polyamide **1** (Fig. 1), which targets the sequence 5′-WGWWCW-3′ found in the androgen response element, inhibited a subset of dihydrotestosterone (DHT)-induced genes in LNCaP cells (12). In this article we explore the effects of this polyamide on the RNAP2 transcription machinery. We find that RNAP2 is preferentially reduced from transcription start sites genome-wide without significant perturbation at enhancer loci. This reduction is accompanied by proteasome-dependent degradation of RPB1. Polyamide treatment induces p53 accumulation that is consistent with what is observed for other transcription inhibitors that interact with DNA (4, 5), but without evidence of DNA damage. This polyamide demonstrates efficacy in vivo against prostate cancer xenografts in mice with limited host toxicity.

Results

Effects of Polyamide 1 on Global Occupancy of RNAP2. Polyamide **1** was previously shown to inhibit the induction of a subset of DHT-driven genes in LNCaP cell culture (12). We interrogated the effects of **1** on the RNAP2 transcription machinery by mapping the global occupancy of RNAP2 using ChIP-seq. Under DHT induction, select androgen receptor (AR)-driven genes, such as *KLK3*, showed increased RNAP2 occupancy over genic regions, but this was decreased in the presence of **1** (Fig. 2*A*). Although RNAP2 occupancy across constitutively expressed genes, such as *GAPDH*, did not change with DHT induction, cotreatment with **1** reduced RNAP2 occupancy across these genes (Fig. 2*B*). This reduction in RNAP2 occupancy by **1** was in the context of a global decrease of RNAP2 occupancy across genic regions (Fig. S1), particularly at transcription start sites (Fig. 2*C*). However, **1** did not significantly change RNAP2 occupancy at enhancer loci (Fig. 2*D*), suggesting **1** may affect the active elongation of RNAP2 without disturbing the transcription apparatus anchored at enhancers, and that the observed differences in RNAP2 occupancy are not a result of technical variation in ChIP success between experiments. Reduction in DNA occupancy of RNAP2 has also been reported in cells treated with α-amanitin, a cyclic octapeptide inhibitor of RPB1 (19).

Inhibition of RNAP2 elongation can be caused by a multitude of genotoxic agents and often results in the degradation of the RPB1 subunit (3, 20, 21). Indeed, in addition to reduced RNAP2 DNA occupancy, immunoblot analysis of LNCaP cells treated with **1** shows depletion of RPB1 in a time- and concentration-dependent manner (Fig. 2*E*). To examine if the effect on RPB1 protein were a result of decreased transcription of this gene, we measured levels of RPB1 mRNA (Fig. 2*F*). The expression of RPB1 modestly increased with polyamide treatment, suggesting this depletion is posttranscriptional.

Polyamide Cytotoxicity Is Reduced by Proteasomal Inhibition and Serum Starvation. Inhibition of RNAP2 has been reported to induce apoptosis (4, 6, 22), and may contribute to polyamide cytotoxicity observed in LNCaP cells cultured with **1** (Fig. 3*A*). A previous study with trabectidin, a DNA minor groove alkylator that causes RPB1 degradation, showed the toxicity induced by the molecule can be reduced by cotreatment with the proteasome inhibitor MG132 (22). To evaluate if polyamide-induced toxicity was also reducible by proteasomal inhibition we treated

Author contributions: F.Y., N.G.N., G.K.M., and P.B.D. designed research; F.Y., N.G.N., B.C.L., and J.W.S. performed research; F.Y., N.G.N., and B.C.L. contributed new reagents/analytic tools; F.Y., N.G.N., B.C.L., G.K.M., and P.B.D. analyzed data; and F.Y., N.G.N., and P.B.D. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE43253).

[1]To whom correspondence should be addressed. E-mail: dervan@caltech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1222035110/-/DCSupplemental.

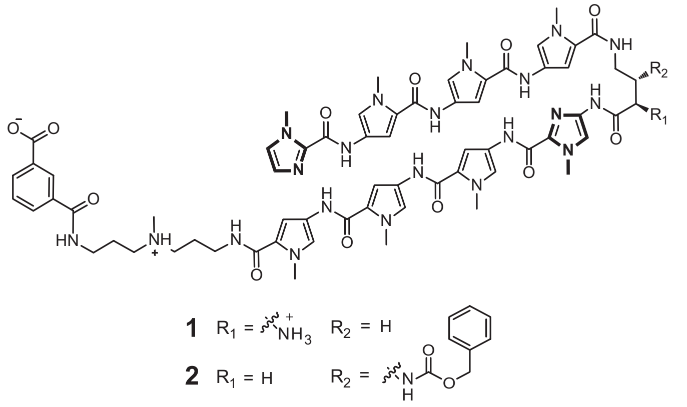


Fig. 1. Structure of polyamides **1** and **2**.

LNCaP cells with **2** in the presence and absence of MG132. We developed analog **2** specifically for this application because prolonged incubation with MG132 alone is cytotoxic, and conjugation of an aryl group to the γ-aminobutyric acid turn have been shown to improve cellular uptake and cytotoxicity of polyamides. Cell-viability experiments showed that **2** induced cell death more rapidly than **1** without significant change to DNA binding (Fig. S2 *A* and *B*). Cell culture experiments revealed coincubation with MG132 reduced cytotoxicity induced by **2** (Fig. 3*B*) and prevented degradation of RPB1 (Fig. 3*C*). Polyamide nuclear uptake was not affected by MG132 (Fig. S2 *C* and *D*). In addition, cytotoxicity studies of cells treated with UV radiation and α-amanitin have shown increased cellular sensitivity to transcription inhibition upon S-phase entry (6, 23). Similarly, **2** was less toxic to LNCaP cells arrested in G_1/G_0 by serum starvation compared with cells grown in normal media (Fig. 3*D* and Fig. S2*E*).

555

Accumulation of p53 and Expression of p53 Targets in the Absence of DNA Damage. Previously published microarray data of LNCaP cells cotreated with DHT and **1** revealed the induction of several p53 target genes (12). Despite depletion of RPB1, treatment of LNCaP cells with **1** alone induced expression of p53 genes that are characteristic of genotoxic stress (Fig. 4*A*) (24). Many of these genes were previously observed to be induced in A549 cells treated with polyamide as well as polyamide-alkylator conjugates (14, 25). To examine if direct DNA damage was contributing to p53 activity, we looked for evidence of DNA damage in LNCaP cells after extended treatment with **1**. Alkaline comet assay showed no evidence of DNA fragmentation (Fig. 4*B*). Additionally, treatment with **1** did not induce cellular markers of DNA damage, including phosphorylation of γH2A.X, ATM, DNA-PKcs, p53, or Chk2 (Fig. 4*C*). However, modest accumulation of p53 and poly(ADP-ribose) polymerase (PARP) cleavage were observed. These data suggest that **1** activates p53 through transcriptional inhibition without DNA damage, a mechanism that has been observed for non-DNA targeting agents that exert transcriptional stress such as the protein kinase inhibitor 5,6-dichlorobenzimidazole (DRB) and α-amanitin (5, 6, 26).

Effects of Polyamide Treatment on Prostate Cancer Xenografts. We recently reported the toxicity and pharmacokinetic (PK) profile of **1** in mice (17). Subcutaneous injection of **1** also results in detectable circulation (Fig. S3). We thus selected this molecule for further testing against xenografts in vivo. Male NOD scid-γ (NSG) mice bearing LNCaP xenografts were treated with either vehicle or 20 nmol (~1 mg/kg) **1** by subcutaneous injection once every 3 d for a cycle of three injections. At the experimental end point, mice treated with **1** had smaller tumors and lower serum

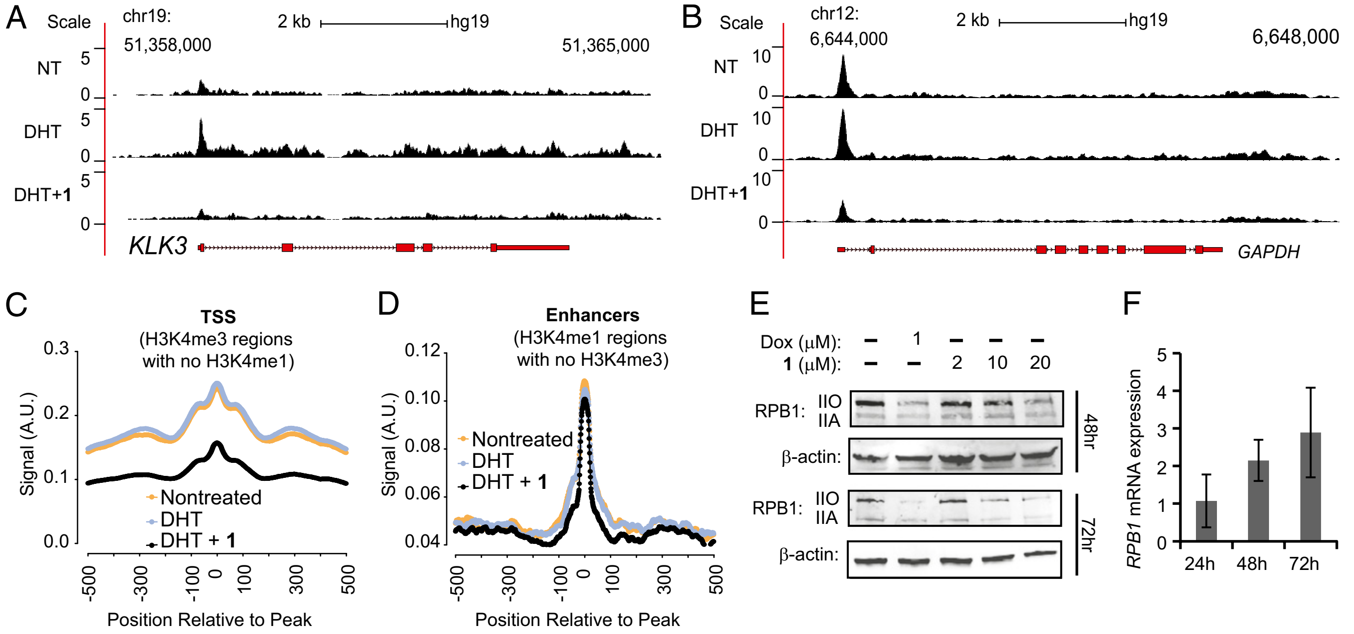


Fig. 2. Global effects of **1** on RNAP2. Genome browser tracks of RPB1 occupancy from untreated, DHT-treated, and DHT + **1**-treated samples over (*A*) an AR-driven gene, *KLK3* (PSA), and (*B*) a housekeeping gene, *GAPDH*. RNAP2 occupancy is mapped as reads per million. (*C*) Genomic RNAP2 occupancy at transcription start sites show comparable levels of enrichment for nontreated and DHT treated samples. Samples treated with DHT + **1** exhibited much lower occupancy. (*D*) Genomic RNAP2 occupancy at enhancer regions is largely unchanged between the three treatment conditions. (*E*) Immunoblot of RPB1 protein in LNCaP cells treated with 1 μM doxorubicin (dox) for 16 h, or **1** at 2 μM, 10 μM, and 20 μM for 48 and 72 h. (*F*) Quantitative RT-PCR measurement of RPB1 transcript levels after LNCaP cells are treated with 10 μM **1** for the indicated times. Relative expression is normalized against nontreated cells. Data represent mean ± SD of biological quadruplicates.

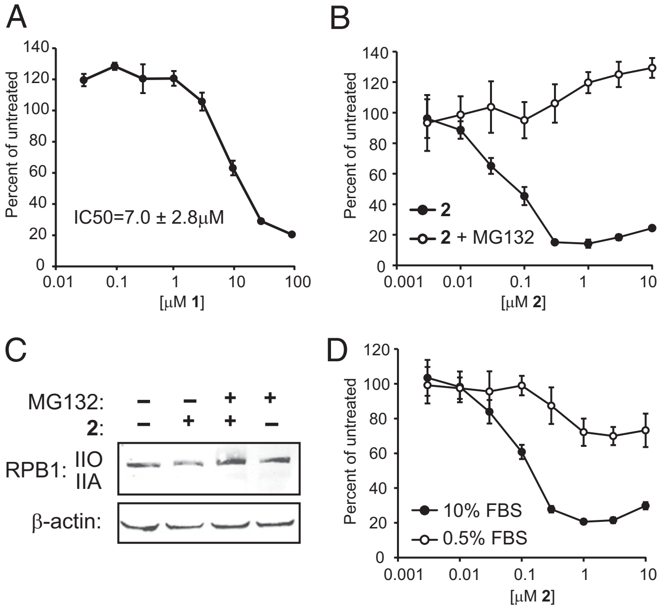


Fig. 3. Cytotoxicity of **1** and **2** and effects on RPB1. (*A*) Cytotoxicity of **1** in LNCaP cells after incubation with **1** for 72 h. Data represent mean ± SD. IC_{50} is calculated from three independent experiments and the error is a 95% confidence intervals. (*B*) Cell viability at 24 h of LNCaP cells treated with varying concentrations **2** with and without proteasome inhibitor MG132 (3 μM, 24 h); proteasome inhibition reduces cytotoxicity of **2**. (*C*) Immunoblot of RPB1 protein in LNCaP cells treated with 10 μM **2** for 12 h followed by 10 μM MG132 for 4 h. (*D*) Cytotoxicity of **2** in LNCaP cells incubated with 10% FBS or with 0.5% FBS for 24 h. Serum starvation decreases percent of cells in the S phase from 8.5% to 4.4% (Fig. S2). Data represent mean ± SD.

prostate-specific antigen (PSA) compared with vehicle controls (Fig. 5 *A* and *B*). Immunohistological analysis of selected tumors showed evidence of cell death by TUNEL stain (Fig. 5*C*). Although tumor-free NSG mice treated with **1** under this regimen showed no signs of distress or weight loss, LNCaP tumor-bearing NSG mice exhibited weight loss by the experimental end point (Fig. S4). This weight loss was accompanied by an elevation in serum uric acid that was not observed in either control group (Fig. 5*D*).

Discussion

DNA targeting agents, including cisplatin, the anthracyclines, minor groove binders, and UV radiation have been demonstrated to affect a multitude of DNA-dependent enzymes, such as the RNA polymerases, DNA polymerase, topoisomerases, and helicases (21, 27). Our research group and others have used polyamides as molecular tools to modulate gene-expression programs (12–15). The programmable sequence specificity of Py-Im polyamides offers a unique mechanism to target specific transcription factor–DNA interfaces and thereby modulate particular gene-expression pathways. In previous studies we have focused our analysis on specific changes to inducible pathways of gene expression. For example, we have shown polyamide **1** affects ∼30% of the DHT-induced transcripts in LNCaP cells, which may result from inhibition of the transcription factor AR-DNA interface (12). However, the cellular cytotoxicity of this polyamide may not only be a result of inhibition of DHT-induced gene expression because analogs of **1** exhibit toxicity in a variety of cancer cells (28). It is more likely that polyamides perturb multiple DNA-dependent cellular processes (transcription, replication) that contribute to cytotoxicity. In this study we show that **1** interferes with RNAP2 elongation resulting in the degradation of RPB1, activation of p53, and triggering of apoptosis, without detectable genomic damage.

Our previous study has shown polyamide **1** decreased the expression of a large number of genes in LNCaP cells (12). To examine the effect of **1** on the transcription machinery, we performed genome-wide mapping of RNAP2 occupancy by ChIP-seq. We found that although DHT induction increased RNAP2 occupancy at select AR-driven genes, cotreatment with **1** caused a genome-wide decrease of RNAP2 occupancy across genic regions. The effect was most pronounced at transcription start sites. Interestingly, RNAP2 occupancy at enhancer loci, where the transcription assemblies may be attached via contacts through other proteins, was not significantly affected by polyamide treatment. This finding suggests polyamide **1** may preferentially affect RNAP2 loading at regions where RNAP2 is actively engaged, a mechanism that has been previously proposed for the gene regulatory activity of polyamides (29).

The displacement of RNAP2 from DNA is caused by many DNA damaging agents that pose an impediment to RNAP2 elongation. This effect is normally coupled with the degradation of the large RNAP2 subunit RPB1. Indeed, the cellular level of

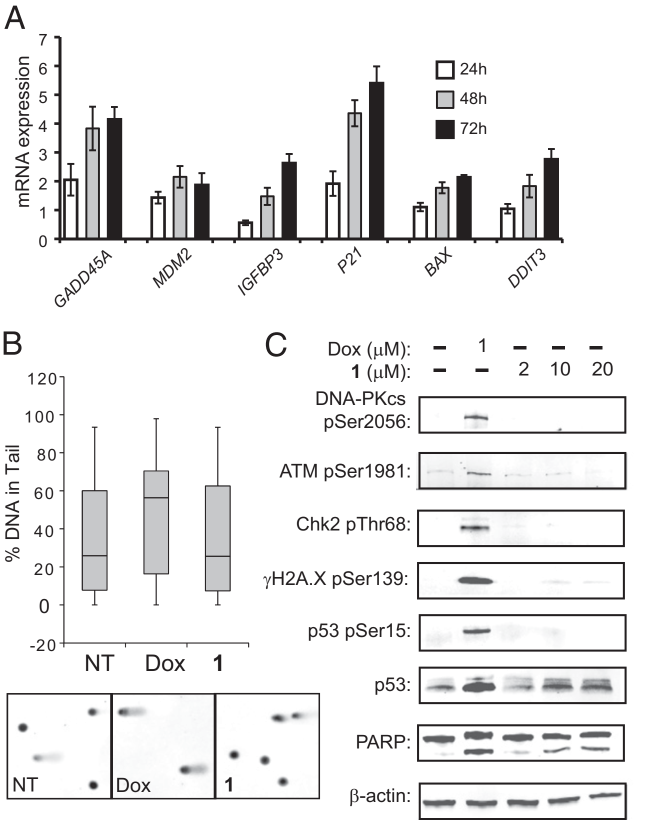


Fig. 4. Induction of p53 activity without evidence of DNA damage. (*A*) Induction of p53 target genes (*GADD45A*, *MDM2*, *IGFBP3*, *P21*, *BAX*) and DNA damage-inducible transcript 3 (DDIT3), by **1** (10 μM) at 24, 48, and 72 h. Data represent the mean of four biological replicates and error bars represent SD. (*B*) Alkaline comet assay of LNCaP cells treated with vehicle, dox (5 μM, 4 h), **1** (10 μM, 48 h). Error bars represents maximum and minimum; boxes represents the upper and lower quartiles and median. Representative comets for each treatment are shown. Effects of **1** are indistinguishable from the nontreated control, but dox treatment significantly increases comet-tail percent of DNA. $P = 0.00043$. (*C*) DNA damage markers after treatment of LNCaP cells with **1**. There is no significant phosphorylation of DNA-PKcs, ATM, Chk2, p53, or γH2A.X. Accumulation p53 and PARP cleavage are observed. Data are representative of biological triplicates except for DNA-PKcs, which was in replicate.

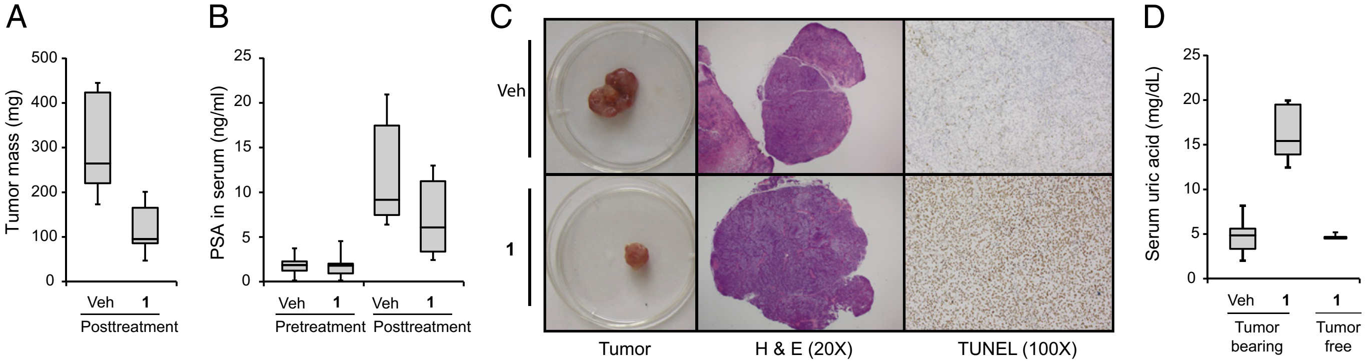


Fig. 5. Polyamide **1** demonstrates antitumor activity in prostate cancer xenografts. (*A*) Male immunocompromised mice were engrafted with LNCaP cells and observed until tumors reached ~100 mm^3. Tumor-bearing mice were then treated with 20 nmol **1** ($n = 12$) or vehicle ($n = 13$) by subcutaneous injections into the flank distal to the tumor once every 3 d for a total of three injections. Mice were killed and tumors resected and weighed 2 d after the final injection. Tumors from mice treated with **1** were smaller (mean: 112 mg; median: 94 mg; range: 47–201 mg) than those of vehicle treated mice (mean: 310 mg; median: 292 mg; range: 173–440 mg). Error bars represents maximum and minimum; boxes represents the upper and lower quartiles and median. $P = 1.6E-5$. (*B*) Serum PSA measured by ELISA pre- and posttreatment. Serum PSA is lower in the posttreatment serum of mice treated with **1** compared with vehicle. $P = 0.024$. (*C*) Selected tumors and histological stains of tumor cross-sections from mice treated with vehicle or **1**. (*D*) Treatment of LNCaP tumor bearing mice with **1** increases serum uric acid compared with vehicle controls and polyamide-treated, nontumor-bearing mice. $P = 3.2E-9$.

RPB1 in LNCaP cells was found to decrease in both a time- and concentration-dependent manner when treated with polyamide **1**. Polyamide **2**, a more cytotoxic analog of **1**, also reduced cellular RPB1 in LNCaP cells and induced cell death. Cotreatment of **2** with a proteasomal inhibitor MG132 was able to prevent the degradation of RPB1 and reduce the toxicity of **2** in cell culture. In addition, the cytotoxic effects of other RNAP2 inhibitors are reported to be attenuated by preventing S-phase entry. LNCaP cells arrested in G_0/G_1 by serum starvation also exhibited reduced sensitivity to **2** compared with cells grown in normal media. The finding that cytotoxicity is partially rescued by MG132 treatment and G_0/G_1 arrest suggests RPB1 degradation contributes to cytotoxicity; however, contributions from other DNA-dependent processes are not ruled out.

Although transcription inhibition can activate p53 signaling, both events can be caused by DNA damage. Analysis of previously published microarray data revealed the induction of several p53 target genes in LNCaP cells cotreated with DHT and **1** (12). Further validation of transcript levels of the genes in this study also showed a time-dependent increase in the expression of *GADD45A*, *MDM2*, *IGFBP3*, *P21*, *BAX*, and *DDIT3* (Fig. 4*A*). Because these genes are also markers of genotoxic stress (24) and were found to be induced in A549 cells treated with alkylating polyamide derivatives (25), we searched for signs of DNA damage to determine if it was causing transcription inhibition and p53 activation. Interestingly, both comet assay and immunblot analysis of cellular DNA damage markers showed no significant signs of DNA damage. Although faint phosphorylation of γH2A.X was visible, it is likely caused by cellular apoptosis as indicated by the concurrent PARP cleavage. These data are consistent with studies in yeast mutants that are hypersensitive to DNA damage, which showed no increased sensitivity to polyamide treatment, suggesting these reversible DNA binders do not compromise genomic integrity (30).

The activation of p53 by transcription inhibition in the absence of DNA damage has been observed for DNA-independent inhibitors of RNAP2, such as DRB, α-amanitin, and various RNAP2-targeted antibodies (5, 6, 26). Distamycin A, the natural product that provided the structural inspiration for Py-Im polyamides, inhibits the initiation of RNA synthesis in cell-free assays (27). In cell culture, distamycin also induces degradation of RPB1 and activates p53 (31, 32). However, low antitumor potency and poor stability limit its utility.

To assess the therapeutic potential of polyamide **1** as an antitumor agent, LNCaP xenografts in a murine model were treated with **1** or PBS vehicle. After three rounds of treatment, tumor growth was reduced by 64% in the treated group. Although treatment with **1** alone did not cause changes in animal body weight or obvious signs of toxicity in tumor-free animals, treatment in tumor-bearing animals resulted in weight loss after three treatments. The accompanied elevation in serum uric acid may be an indication of tumor lysis syndrome (33), which is associated with rapid tumor cell turnover upon polyamide treatment. We anticipate that Py-Im polyamides could also demonstrate efficacy in additional xenograft models.

Methods

Compounds and Reagents. Py-Im polyamides **1**, **2**, and **3** were synthesized on oxime resin, as described previously (28, 34, 35). (R)-MG132 (MG132) was from Santa Cruz Biotechnology.

Cell Viability Assays. LNCaP cells were plated in clear bottom 96-well plates at 5,000–7,500 cells per well. The cells were allowed to adhere for 24–36 h before compounds were added in fresh media. Cell viability was determined by the WST-1 assay (Roche) for **1** and **2** after 24- or 72-h incubation with cells. Cells in cytotoxicity rescue experiments were treated with **2** alone or with 3 μM MG132 for 24 h. For cell-cycle arrest experiments, LNCaP cells were seeded at 2,500–5,000 cells per well in normal media and allowed to adhere for 24–36 h. The media was replaced with normal media or media supplemented with 0.5% (vol/vol) FBS and incubated for 48 h before treatment with compound.

In Vivo Xenograft Experiments. All mice experiments were conducted under an approved protocol by the Institutional Animal Care and Use Committee of the California Institute of Technology. Male NSG mice were purchased from The Jackson Laboratory. The animals were individually caged and maintained on a standard light-dark cycle. NSG mice were engrafted with LNCaP cells (2.5 million cells) in a mixture of 1:1 media and matrigel in the left flank. Tumors were grown to ~100 mm^3 (L × W^2) before beginning treatment with compound or vehicle. Py-Im polyamide **1** was administered once every 3 d at 20 nmol per animal (~1 mg/kg) in a 5% (vol/vol) DMSO:PBS vehicle solution until the experiment endpoint.

Serum Measurements. To investigate if polyamide **1** could be detected in peripheral blood after subcutaneous injections, 120 nmol of **1** [in 5% (vol/vol) DMSO/PBS] was injected into the right flank of four C57BL/6J mice. Blood was collected from anesthetized mice via retroorbital collection at 5 min, 4 h, and 12 h after injection, then processed by methods previously described and analyzed by HPLC (36). For measurement of serum PSA (KLK3) and uric acid, blood was collected from anesthetized mice via retroorbital collection at experimental endpoint and serum was separated from blood by centrifugation. Serum PSA (KLK3) was measured by ELISA (R&D Systems) according to the manufacturer's instructions. Uric acid was measured as previously described (37).

Chromatin Immunoprecipitation. Genomic occupancy of RNAP2 was determined by ChIP with the 4H8 antibody (Abcam). LNCaP cells were plated at 35 million cells per plate in RPMI supplemented with 10% (vol/vol) CTFBS and allowed to adhere for 24–36 h. The cells were treated with compound **1** in fresh media (10% CTFBS) for 48 h. Cells treated and untreated with **1** were incubated with 1 nM DHT for 6 h. Two-step cross-linking was performed as previously described (38). After DSG removal, chromatin was immunoprecipitated by previously published methods (39). DNA was harvested by phenol chloroform extraction and purified with the QIAquick purification kit (Qiagen). Quantitative PCR was used to validate enrichment at the GAPDH transcription start site (Primers: F-GGTTTCTCTCCGCCCGTCTT, R-TGTTCGACAGTCAGCCGCAT) compared with an internal negative locus (Primers: F-TAGAAGGGGGATAGGGGAAC, R-CCAGAAAACTGGCTCCTTCTT). Each sample was immunoprecipated as five technical replicates. The three most consistent samples were combined and submitted for sequencing on an Illumina genome analyzer. Biological replicates were acquired.

Data Processing and Analysis. Sequencing reads were trimmed down to 36 bp and then mapped against the male set of human chromosomes (excluding all random chromosomes and haplotypes) using the hg19 version of the human genome as a reference. Bowtie 0.12.7 was used for aligning reads (40), with the following settings: "-v 2 -t–best–strata". Signal profiles over genomic locations were generated using custom written python scripts; the refSeq annotation was used for gene coordinates. Enhancers and promoters were defined using previously published histone marker data (41). ChIP-seq peaks were called using MACS2 with default settings (42). Enhancers were defined as H3K4me1$^+$ regions that did not intersect with H3K4me3$^+$ regions and promoters as H3K4me3$^+$ regions that did not intersect with H3K4me1$^+$ regions. Clustering was performed with Cluster 3.0 (43) and visualized with Java TreeView (44).

Comet Assay. LNCaP cells were plated at 1 million cells per 10-cm plate and allowed to adhere for 24–36 h. Cells were then incubated with either 10 μM **1** for 48 h or 5 μM doxorubicin for 4 h. DNA damage was assayed using the Trevigen CometAssay system and samples were prepared from harvested cells according to the manufacture protocol. Comets were imaged on a confocal microscope (Exciter, Zeiss) at 10× magnification. Percentage of DNA in the tail was determined using Comet Assay Lite IV (Perceptive Instruments). More than 100 comets were scored for each condition.

Immunoblot Assay. Samples for immunoblot analysis were prepared by plating LNCaP or DU145 cells at 1 million cells per 10-cm plate. Cells were allowed to adhere for 24–36 h before incubation with compound. After the appropriate incubation time, cells were washed once with ice-cold PBS and harvested in ice-cold 125 μL lysis buffer (50 mM Tris•HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% Triton X 100) containing protease inhibitor mixture (Roche), 1 mM PMSF (Sigma), and phosphatase inhibitors (Sigma). Samples were incubated on ice for 10 min with vortexing once every 3 min. Cellular debris was pelleted by spinning at 21,000 × *g* for 15 min to collect the supernatant. Samples were then quantified for protein content with the Bradford assay (Bio-Rad) and boiled with 4× sample buffer (Li-Cor) for 5 min. Protein electrophoresis was performed in 4–20% precast Tris•glycine SDS gels (Bio-Rad) and transferred to PVDF membranes. Membrane blocking was done with Odyssey Blocking Buffer (Li-Cor). The following antibodies used to probe changes in protein levels or phosphorylation states: RBP1 (Santa Cruz Biotechnology; N20), p53 (Santa Cruz Biotechnology; DO1), phospho-Chk2-Thr68 (Cell Signaling Technology), Phospho-p53-Ser15 (Cell Signaling Technology), phospho-H2A.X-Ser139 (Cell Signaling Technology), phosphor-ATM-Ser1981 (Abcam), phospho-DNA-PKcs-Ser2056 (Abcam), and β-actin (Abcam). Near-IR secondary antibodies (Li-Cor) were used for imaging. Experiments were performed in biological triplicate except for DNA-PKcs (replicate).

Flow Cytometry. To determine cell cycle distribution of LNCaP cells grown in normal media or under serum-starved conditions, 1 million cells were seeded to each 10-cm plate and allowed to adhere for 24–36 h. Media was then replaced with fresh normal media [10% (vol/vol) FBS] or serum-starved media [0.5% (vol/vol) FBS] and incubated for an additional 48 h. Cells were then trypsinized and prepared for analysis as previously described (45). Samples were analyzed in biological triplicate on a FACSCalibur (Becton-Dickinson) instrument. Data analysis was performed using FlowJo 7.6.5.

Quantitative RT-PCR. RNA was extracted using RNEasy columns (Qiagen) according to the manufacturer's protocols. cDNA was generated from RNA by reverse transcriptase (Transcriptor First Strand cDNA kit; Roche). Quantitative real-time RT-PCR was performed using SYBR Green PCR Master Mix (Applied Biosystems) on an ABI 7300 instrument. mRNA was measured relative to β-glucuronidase as an endogenous control. Experiments were performed in biological quadruplicates. For primer sequences see Table S1.

Confocal Microscopy. Cells were plated in 35-mm optical dishes (MatTek) and dosed with polyamide **3** at 2 μM for 24 h with or without 3μM MG132. Cells were then washed with PBS and imaged on a confocal microscope (Exciter; Zeiss) using a 63× oil immersion lens. Confocal imaging was performed following established protocols (34).

Histology and Immunohistochemistry. Tumors were resected immediately after euthanasia and fixed in neutral buffered formalin. Selected samples were embedded in paraffin, sectioned and stained with H&E. Selected sections were assessed by TUNEL, as previously described (46).

Thermal Denaturation Assays. Polyamides **1** and **2** were incubated with duplex DNA 5′-CGATGTTCAAGC-3′, which contains the predicted target site for these compounds (underlined). Melting temperature analyses were performed on a Varian Cary 100 spectrophotometer as described (47). Melting temperatures were defined as a maximum of the first derivative of absorbance at 260 nm over the range of temperatures.

Statistical Analysis. Statistical significance was calculated using the Student *t* test with two tailed variance. Results were considered significant when $P < 0.05$.

ACKNOWLEDGMENTS. We thank Rochelle Diamond of the Caltech Flow Cytometry Cell Sorting Facility for help with flow cytometry experimental setup and data acquisition; and Dr. Janet Baer, Dr. Karen L. Lencioni, and Gwen E. Williams for helpful discussions and technical assistance with animal experiments. This work was supported in part by the National Institutes of Health Grants GM27681 and HG004576, and by the Prostate Cancer Foundation. N.G.N. acknowledges the Jonsson Cancer Center Foundation at University of California at Los Angeles for continued support.

1. Derheimer FA, Chang CW, Ljungman M (2005) Transcription inhibition: A potential strategy for cancer therapeutics. *Eur J Cancer* 41(16):2569–2576.
2. Koumenis C, Giaccia A (1997) Transformed cells require continuous activity of RNA polymerase II to resist oncogene-induced apoptosis. *Mol Cell Biol* 17(12):7306–7316.
3. Jung Y, Lippard SJ (2006) RNA polymerase II blockage by cisplatin-damaged DNA. Stability and polyubiquitylation of stalled polymerase. *J Biol Chem* 281(3):1361–1370.
4. Ljungman M, Zhang FF (1996) Blockage of RNA polymerase as a possible trigger for u.v. light-induced apoptosis. *Oncogene* 13(4):823–831.
5. Ljungman M, Zhang FF, Chen F, Rainbow AJ, McKay BC (1999) Inhibition of RNA polymerase II as a trigger for the p53 response. *Oncogene* 18(3):583–592.
6. Arima Y, et al. (2005) Transcriptional blockade induces p53-dependent apoptosis associated with translocation of p53 to mitochondria. *J Biol Chem* 280(19):19166–19176.
7. Nguyen VT, et al. (1996) In vivo degradation of RNA polymerase II largest subunit triggered by alpha-amanitin. *Nucleic Acids Res* 24(15):2924–2929.
8. Arseneau JC, et al. (1972) Nonlymphomatous malignant tumors complicating Hodgkin's disease. Possible association with intensive therapy. *N Engl J Med* 287(22):1119–1122.
9. Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr Opin Struct Biol* 13(3):284–299.
10. Chenoweth DM, Dervan PB (2009) Allosteric modulation of DNA by small molecules. *Proc Natl Acad Sci USA* 106(32):13175–13179.
11. Chenoweth DM, Dervan PB (2010) Structural basis for cyclic Py-Im polyamide allosteric inhibition of nuclear receptor binding. *J Am Chem Soc* 132(41):14521–14529.
12. Nickols NG, Dervan PB (2007) Suppression of androgen receptor-mediated gene expression by a sequence-specific DNA-binding polyamide. *Proc Natl Acad Sci USA* 104 (25):10418–10423.
13. Nickols NG, Jacobs CS, Farkas ME, Dervan PB (2007) Modulating hypoxia-inducible transcription by disrupting the HIF-1-DNA interface. *ACS Chem Biol* 2(8):561–571.
14. Muzikar KA, Nickols NG, Dervan PB (2009) Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression. *Proc Natl Acad Sci USA* 106(39): 16598–16603.
15. Raskatov JA, et al. (2012) Modulation of NF-κB-dependent gene transcription using programmable DNA minor groove binders. *Proc Natl Acad Sci USA* 109(4):1023–1028.

16. Matsuda H, et al. (2011) Transcriptional inhibition of progressive renal disease by gene silencing pyrrole-imidazole polyamide targeting of the transforming growth factor-β1 promoter. *Kidney Int* 79(1):46–56.
17. Synold TW, et al. (2012) Single-dose pharmacokinetic and toxicity analysis of pyrrole-imidazole polyamides in mice. *Cancer Chemother Pharmacol* 70(4):617–625.
18. Raskatov JA, et al. (2012) Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide. *Proc Natl Acad Sci USA* 109(40):16041–16045.
19. Palstra RJ, et al. (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS ONE* 3(2):e1661.
20. Bregman DB, et al. (1996) UV-induced ubiquitination of RNA polymerase II: A novel modification deficient in Cockayne syndrome cells. *Proc Natl Acad Sci USA* 93(21): 11586–11590.
21. Ratner JN, Balasubramanian B, Corden J, Warren SL, Bregman DB (1998) Ultraviolet radiation-induced ubiquitination and proteasomal degradation of the large subunit of RNA polymerase II. Implications for transcription-coupled DNA repair. *J Biol Chem* 273(9):5184–5189.
22. Aune GJ, et al. (2008) Von Hippel-Lindau-coupled and transcription-coupled nucleotide excision repair-dependent degradation of RNA polymerase II in response to trabectedin. *Clin Cancer Res* 14(20):6449–6455.
23. McKay BC, Becerril C, Spronck JC, Ljungman M (2002) Ultraviolet light-induced apoptosis is associated with S-phase in primary human fibroblasts. *DNA Repair (Amst)* 1(10):811–820.
24. el-Deiry WS (1998) Regulation of p53 downstream genes. *Semin Cancer Biol* 8(5): 345–357.
25. Kashiwazaki G, et al. (2012) Synthesis and biological properties of highly sequence-specific-alkylating N-methylpyrrole-N-methylimidazole polyamide conjugates. *J Med Chem* 55(5):2057–2066.
26. Derheimer FA, et al. (2007) RPA and ATR link transcriptional stress to p53. *Proc Natl Acad Sci USA* 104(31):12778–12783.
27. Puschendorf B, Petersen E, Wolf H, Werchau H, Grunicke H (1971) Studies on the effect of distamycin A on the DNA dependent RNA polymerase system. *Biochem Biophys Res Commun* 43(3):617–624.
28. Meier JL, Montgomery DC, Dervan PB (2012) Enhancing the cellular uptake of Py-Im polyamides through next-generation aryl turns. *Nucleic Acids Res* 40(5):2345–2356.
29. Carlson CD, et al. (2010) Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci USA* 107(10):4544–4549.
30. Marini NJ, et al. (2003) DNA binding hairpin polyamides with antifungal activity. *Chem Biol* 10(7):635–644.
31. Zhang Z, et al. (2009) Tanshinone IIA triggers p53 responses and apoptosis by RNA polymerase II upon DNA minor groove binding. *Biochem Pharmacol* 78(10): 1316–1322.
32. Hirota M, Fujiwara T, Mineshita S, Sugiyama H, Teraoka H (2007) Distamycin A enhances the cytotoxicity of duocarmycin A and suppresses duocarmycin A-induced apoptosis in human lung carcinoma cells. *Int J Biochem Cell Biol* 39(5):988–996.
33. Coiffier B, Altman A, Pui CH, Younes A, Cairo MS (2008) Guidelines for the management of pediatric and adult tumor lysis syndrome: An evidence-based review. *J Clin Oncol* 26(16):2767–2778.
34. Best TP, Edelson BS, Nickols NG, Dervan PB (2003) Nuclear localization of pyrrole-imidazole polyamide-fluorescein conjugates in cell culture. *Proc Natl Acad Sci USA* 100(21):12063–12068.
35. Puckett JW, Green JT, Dervan PB (2012) Microwave assisted synthesis of Py-Im polyamides. *Org Lett* 14(11):2774–2777.
36. Raskatov JA, Hargrove AE, So AY, Dervan PB (2012) Pharmacokinetics of Py-Im polyamides depend on architecture: Cyclic versus linear. *J Am Chem Soc* 134(18): 7995–7999.
37. Dai KS, et al. (2005) An evaluation of clinical accuracy of the EasyTouch blood uric acid self-monitoring system. *Clin Biochem* 38(3):278–281.
38. Nowak DE, Tian B, Brasier AR (2005) Two-step cross-linking method for identification of NF-kappaB gene network by chromatin immunoprecipitation. *Biotechniques* 39(5): 715–725.
39. Reddy TE, et al. (2009) Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res* 19(12):2163–2171.
40. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
41. Yu JD, et al. (2010) An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* 17(5):443–454.
42. Zhang Y, et al. (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9(9): R137.
43. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20(9):1453–1454.
44. Saldanha AJ (2004) Java Treeview—Extensible visualization of microarray data. *Bioinformatics* 20(17):3246–3248.
45. Diamond RA, DeMaggio S (2000) *In Living Color: Protocols in Flow Cytometry and Cell Sorting* (Springer, Berlin, New York), pp xxv, 800 pp.
46. Zisman A, et al. (2003) LABAZ1: A metastatic tumor model for renal cell carcinoma expressing the carbonic anhydrase type 9 tumor antigen. *Cancer Res* 63(16): 4952–4959.
47. Dose C, Farkas ME, Chenoweth DM, Dervan PB (2008) Next generation hairpin polyamides with (R)-3,4-diaminobutyric acid turn unit. *J Am Chem Soc* 130(21): 6859–6866.

G

Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state

Originally published as:

Le Thomas A*, Rogers AK*, Webster A*, Marinov GK*, Liao SE, Perkins EM, Hur JK, Aravin AA, Fejes-Tóth K. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes & Development.* **27**:390–399. doi: 10.1101/gad.209841.112.

Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state

Adrien Le Thomas,[1,2,3] Alicia K. Rogers,[1,3] Alexandre Webster,[1,3] Georgi K. Marinov,[1,3] Susan E. Liao,[1] Edward M. Perkins,[1] Junho K. Hur,[1] Alexei A. Aravin,[1,4] and Katalin Fejes Tóth[1,4]

[1]California Institute of Technology, Pasadena, California 91125, USA; [2]Université Pierre et Marie Curie, Ecole Doctorale Complexité du Vivant, 75005 Paris, France

In the metazoan germline, piwi proteins and associated piwi-interacting RNAs (piRNAs) provide a defense system against the expression of transposable elements. In the cytoplasm, piRNA sequences guide piwi complexes to destroy complementary transposon transcripts by endonucleolytic cleavage. However, some piwi family members are nuclear, raising the possibility of alternative pathways for piRNA-mediated regulation of gene expression. We found that *Drosophila* Piwi is recruited to chromatin, colocalizing with RNA polymerase II (Pol II) on polytene chromosomes. Knockdown of Piwi in the germline increases expression of transposable elements that are targeted by piRNAs, whereas protein-coding genes remain largely unaffected. Derepression of transposons upon Piwi depletion correlates with increased occupancy of Pol II on their promoters. Expression of piRNAs that target a reporter construct results in a decrease in Pol II occupancy and an increase in repressive H3K9me3 marks and heterochromatin protein 1 (HP1) on the reporter locus. Our results indicate that Piwi identifies targets complementary to the associated piRNA and induces transcriptional repression by establishing a repressive chromatin state when correct targets are found.

[*Keywords*: piRNA; Piwi; chromatin; RNA polymerase II; transcription; transposon]

Supplemental material is available for this article.

Received November 8, 2012; revised version accepted January 14, 2013.

Diverse small RNA pathways function in all kingdoms of life, from bacteria to higher eukaryotes. In eukaryotes, several classes of small RNA associate with members of the Argonaute protein family, forming effector complexes in which the RNA provides target recognition by sequence complementarity, and the Argonaute provides the repressive function. Argonaute–small RNA complexes have been shown to regulate gene expression both transcriptionally and post-transcriptionally. Post-transcriptional repression involves cleavage of target RNA through either the endonucleolytic activity of Argonautes or sequestering targets into cytoplasmic ribonucleoprotein (RNP) granules (Hutvagner and Simard 2008).

The mechanism of transcriptional repression by small RNAs has been extensively studied in fission yeast and plants. Several studies showed that Argonaute–small RNA complexes induce transcriptional repression by tethering chromatin modifiers to target loci. In fission yeast, the effector complex containing the Argonaute and the bound siRNA associates with the histone H3 Lys 9 (H3K9) methyltransferase Clr4 to install repressive H3K9-dimethyl marks at target sites (Nakayama et al. 2001; Maison and Almouzni 2004; Sugiyama et al. 2005; Grewal and Jia 2007). Methylation of histone H3K9 leads to recruitment of the heterochromatin protein 1 (HP1) homolog Swi6, enhancing silencing and further promoting interaction with the Argonaute complex. The initial association of Ago with chromatin, however, requires active transcription (Ameyar-Zazoua et al. 2012; Keller et al. 2012). Plants also use siRNAs to establish repressive chromatin at repetitive regions. Contrary to yeast, heterochromatin in plants is marked by DNA methylation, although repression also depends on histone methylation by a Clr4 homolog (Soppe et al. 2002; Onodera et al. 2005). Although siRNA-mediated gene silencing is predominant on repetitive sequences, it is not limited to these sites. Constitutive expression of dsRNA mapping to promoter regions results in production of corresponding siRNAs, de novo DNA methylation, and gene silencing (Mette et al. 2000; Matzke et al. 2004).

In metazoans, small RNA pathways are predominantly associated with post-transcriptional silencing. One class

[3]These authors contributed equally to this work.
[4]Corresponding authors
E-mail kft@caltech.edu
E-mail aravin@caltech.edu
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.209841.112.

of small RNA, microRNA, regulates expression of a large fraction of protein-coding genes (Friedman et al. 2009). In *Drosophila*, siRNAs silence expression of transposable elements (TEs) in somatic cells (Chung et al. 2008; Ghildiyal et al. 2008) and target viral genes upon infection (Galiana-Arnoux et al. 2006; Wang et al. 2006; Zambon et al. 2006). Another class of small RNAs, Piwi-interacting RNAs (piRNAs), associates with the Piwi clade of Argonautes and acts to repress mobile genetic elements in the germline of both *Drosophila* and mammals (Siomi et al. 2011). Analysis of piRNA sequences in *Drosophila* revealed a very diverse population of small RNAs that primarily maps to transposon sequences and is derived from a number of heterochromatic loci called piRNA clusters, which serve as master regulators of transposon repression (Brennecke et al. 2007). Additionally, a small fraction of piRNAs seems to be processed from the mRNA of several host protein-coding genes (Robine et al. 2009; Saito et al. 2009). The *Drosophila* genome encodes three piwi proteins: Piwi, Aubergine (AUB), and Argonaute3 (AGO3). In the cytoplasm, AUB and AGO3 work together to repress transposons through cleavage of transposon transcripts, which are recognized through sequence complementarity by the associated piRNAs (Vagin et al. 2006; Agger et al. 2007; Brennecke et al. 2007; Gunawardane et al. 2007).

In both *Drosophila* and mammals, one member of the Piwi clade proteins localizes to the nucleus. Analogously to small RNA pathways in plants, the mouse piRNA pathway is required for de novo DNA methylation and silencing of TEs (Carmell et al. 2007; Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008); however, the exact mechanism of this process is unknown. In *Drosophila*, DNA methylation is absent; however, several studies indicate that elimination of Piwi from the nucleus causes changes in histone marks on TEs (Klenov et al. 2011; Pöyhönen et al. 2012), yet a genome-wide analysis of Piwi's effect on chromatin marks and transcription is lacking.

Here we show that Piwi interacts with chromatin on polytene chromosomes in nurse cell nuclei. We found that Piwi exclusively represses loci that are targeted by piRNAs. We show that Piwi-mediated silencing occurs through repression of transcription and correlates with installment of repressive chromatin marks at targeted loci.

Results

To analyze the role of Piwi in the nucleus, we generated transgenic flies expressing a GFP-tagged Piwi protein (GFP-Piwi) under the control of its native regulatory region. GFP-Piwi was expressed in the ovary and testis in a pattern indistinguishable from the localization of native Piwi and was able to rescue the piwi-null phenotype as indicated by ovarian morphology, fertility, transposon expression, and piRNA levels. GFP-Piwi was deposited into the mature egg and localized to the pole plasm; however, contrary to a previous observation (Brower-Toland et al. 2007), we did not detect Piwi expression outside of the ovary and testis in third instar larvae or adult flies. We also did not observe the association of Piwi with polytene chromosomes in salivary gland cells of third instar larvae. In both follicular and germline cells of the *Drosophila* ovary, GFP-Piwi localized exclusively in the nucleus, with slightly higher concentrations apparent in regions enriched for DAPI, indicating a possible interaction with chromatin. To gain further insight into Piwi localization in the nucleus, we took advantage of the fact that nurse cell chromosomes are polytenized and can be visualized on the *otu* mutant background (Mal'ceva et al. 1997). Analysis of polytene chromosomes from nurse cells demonstrated that GFP-Piwi associates with chromatin in a specific banding pattern. Interestingly, coimmunostaining showed that a GFP-Piwi signal on polytene chromosomes generally overlaps with the RNA polymerase II (Pol II) signal, which marks sites of active transcription (Fig. 1A).

In order to identify factors that might be responsible for targeting Piwi to chromatin, we immunoprecipitated Piwi complexes from the *Drosophila* ovary and analyzed Piwi interaction partners by mass spectrometry. We purified Piwi complexes from ovaries of three different transgenic lines expressing GFP-Piwi, myc-Piwi, or Flag-Piwi using antibodies against each respective tag. As a control, we used flies expressing free GFP in the ovary.

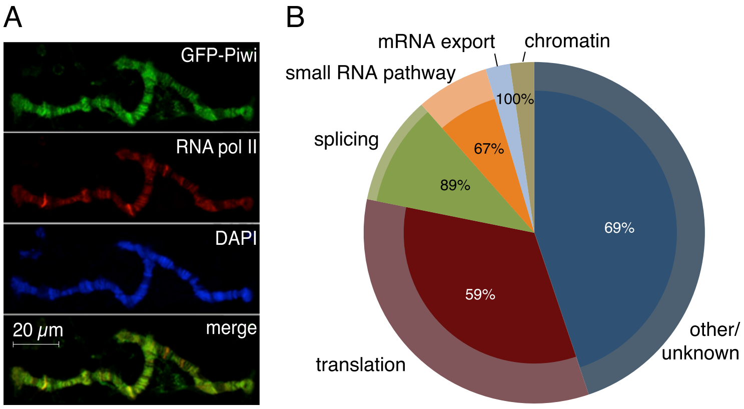


Figure 1. Piwi associates with chromatin and nuclear transcripts. (*A*) Polytene chromosomes from *Drosophila* nurse cells expressing GFP-Piwi on the *otu[7]/otu[11]* background. Piwi pattern on chromosomes correlates with Pol II staining. (*B*) Mass spectrometry analysis of Piwi interaction partners. Piwi complexes were precipitated in the presence and absence of RNase A. The outer circle represents classification of Piwi-associated proteins based on GO term analysis. The inner pies represent the fraction of each group whose association with Piwi depends on RNA (percentage indicated). Note that chromatin, splice, and mRNA export factors are virtually absent after RNase A treatment.

We identified >50 factors that showed significant enrichment in all three Piwi purifications but were absent in the control. We were unable to identify chromatin-associated factors that directly associate with Piwi but identified several RNA-binding proteins that associate with nascent transcripts, such as splicing (Rm62, Pep, Ref1, Yps, CG9684, CG31368, CG5728, and Mago) and nuclear export (Tho2 and Hpr1) factors (Fig. 1B). Upon RNase A treatment prior to immunoprecipitation, the presence of most of these RNA-binding proteins in purified Piwi complexes was eliminated.

Piwi proteins are believed to find their targets through sequence complementarity of the associated piRNA. In fact, it has been proposed that lack of the associated piRNA leads to destabilization of piwi proteins and to Piwi's inability to localize to the nucleus (Saito et al. 2009; Haase et al. 2010; Olivieri et al. 2010; Handler et al. 2011; Ishizu et al. 2011). On the other hand, Piwi has been proposed to have functions that are independent of its role in transposon control by regulating stem cell niche development (Cox et al. 1998; Klenov et al. 2011). To address the role of piRNA in translocation of Piwi into the nucleus and its function, we generated transgenic flies expressing a point mutant Piwi—referenced as Piwi-YK—that is deficient in piRNA binding due to a substitution of two conserved amino acid residues (Y551L and K555E) in the 5′ phosphate-binding pocket (Kiriakidou et al. 2007; Djuranovic et al. 2010). The Piwi-YK mutant was expressed in *Drosophila* follicular and germ cells at levels similar to that of wild-type Piwi but was completely devoid of associated piRNA (Fig. 2A). In contrast to wild-type Piwi, Piwi-YK could be found in the cytoplasm, supporting the existence of a quality control mechanism that prevents entrance of unloaded Piwi into the nucleus (Ishizu et al. 2011). Nevertheless, a significant amount of piRNA-deficient Piwi localized to the nucleus (Fig. 2B). Similar to wild-type Piwi, Piwi-YK seemed to associate with chromatin, as indicated by its localization in DAPI-stained regions of the nuclei, and this is consistent with fluorescence loss in photobleaching (FLIP) experiments that demonstrated reduced nuclear mobility compared with free diffusion (Supplemental Fig. S1). Based on sterility and ovarian morphology, the *piwi-YK* transgene was unable to rescue the piwi-null phenotype despite its nuclear localization (Fig. 2C), indicating that while piRNA binding is not absolutely essential for stability and nuclear localization of Piwi, it is required for Piwi function.

To directly test the function of Piwi in the nucleus, we analyzed the effect of Piwi deficiency on gene expression and chromatin state on a genome-wide scale. Piwi mutant females have atrophic ovaries caused by Piwi deficiency in somatic follicular cells (Lin and Spradling 1997; Cox et al. 1998), which precludes analysis of Piwi function in null mutants. Instead, we used RNAi knockdown to deplete Piwi in germ cells while leaving it functionally intact in somatic follicular cells. The Piwi knockdown flies did not exhibit gross morphological defects in the ovary; however, they showed drastic reduction in GFP-Piwi expression in germ cells and were sterile (Fig. 3A,B).

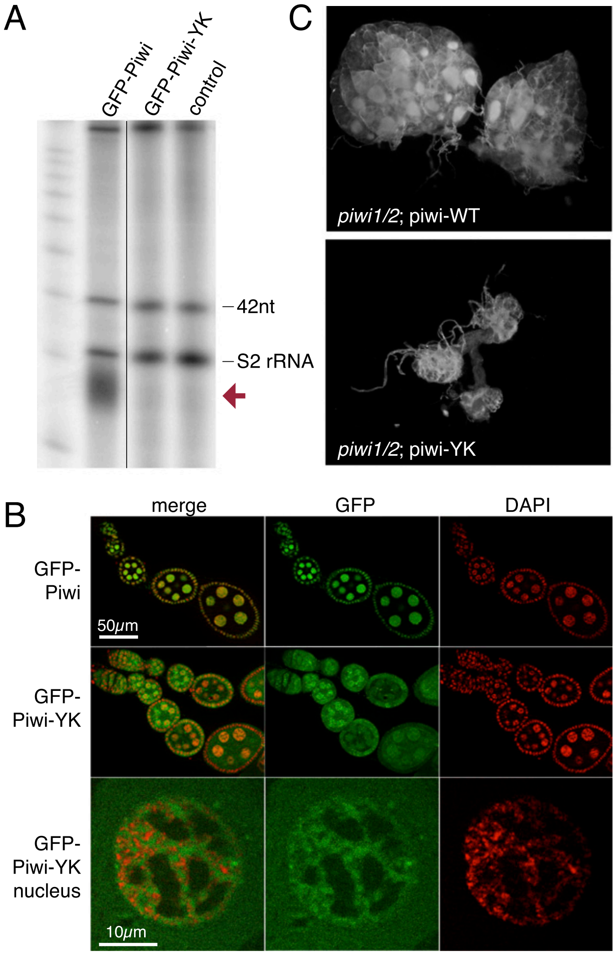


Figure 2. Piwi function, but not its nuclear localization, requires piRNA association. (*A*) The Piwi-YK mutant does not associate with piRNA. Immunoprecipitation of Piwi–piRNA complexes was performed with GFP antibody on ovaries from GFP-Piwi and GFP-Piwi-YK transgenic flies and a control strain. Small RNAs were isolated, 5′-labeled, and resolved on a denaturing gel. The same amount of 42-nucleotide RNA oligonucleotides was spiked into all samples prior to RNA isolation to control for loss of RNA during isolation and labeling. piRNAs (red arrow) are absent in the Piwi-YK complex. (*B*) GFP-Piwi-YK is present in the nuclei of nurse cells and colocalizes with chromatin (DAPI-stained areas). (*C*) The Piwi-YK mutant does not rescue the morphological changes caused by the piwi-null mutation. Dark-field images of ovaries where either the wild-type *piwi* or the *piwi-YK* transgene has been backcrossed onto the piwi-null background.

To analyze the effect of Piwi deficiency on the steady-state transcriptome as well as the transcription machinery, we performed RNA sequencing (RNA-seq) and Pol II chromatin immunoprecipitation (ChIP) combined with deep sequencing (ChIP-seq) experiments from Piwi knockdown and control flies.

In agreement with previous observations that implicated Piwi in transposon repression (Saito et al. 2006; Aravin et al. 2007; Brennecke et al. 2007), we found that steady-state transcript levels of several TEs were increased

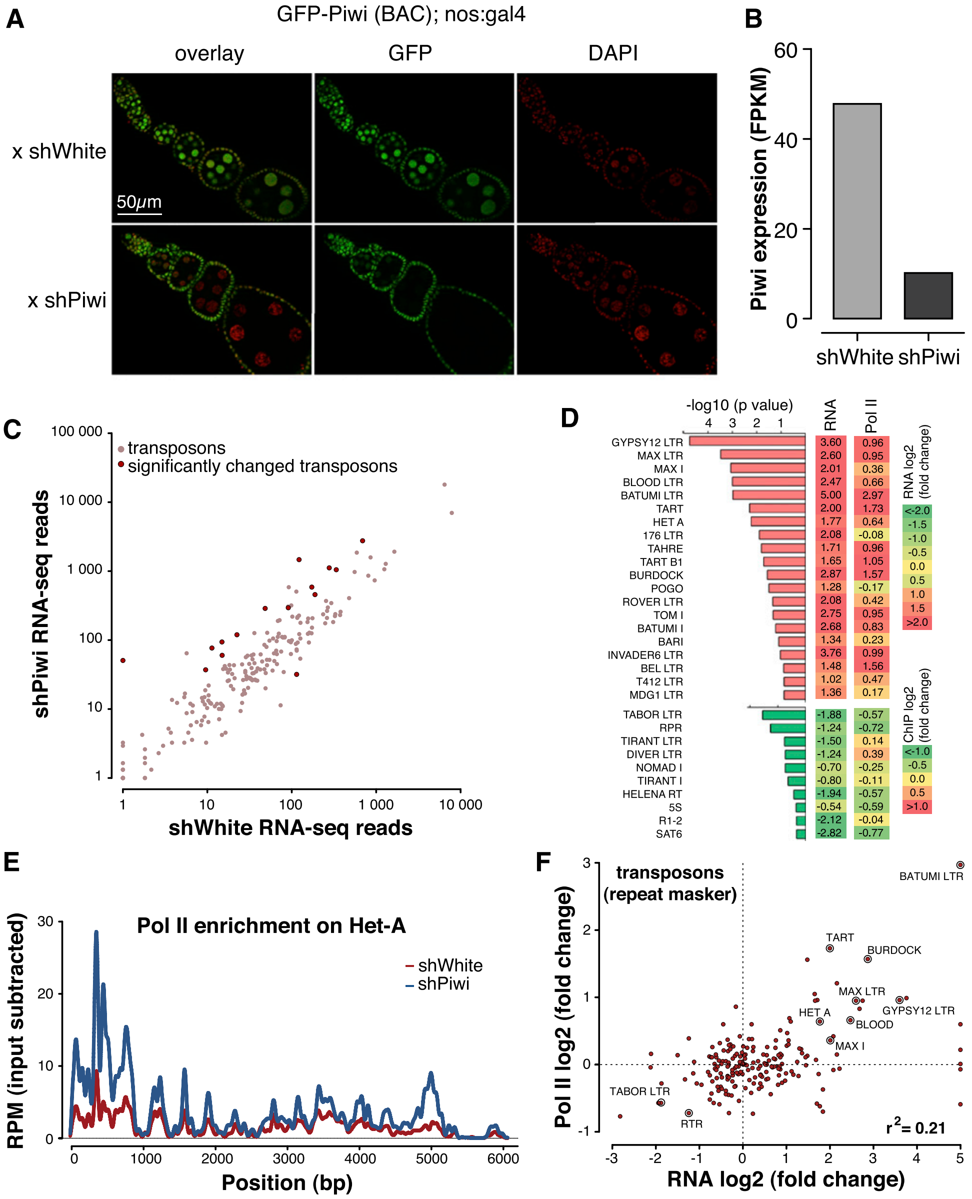


Figure 3. Piwi transcriptionally represses TEs. (*A*) Piwi knockdown is efficient and specific to ovarian germ cells as indicated by GFP-Piwi localization. GFP-Piwi; Nanos-Gal4-VP16 flies were crossed to control shRNA (shWhite) or shPiwi lines. Piwi is specifically depleted in germ cells and not in follicular cells, consistent with expression of the Nanos-Gal4-VP16 driver. (*B*) Piwi expression as measured by RNA-seq in the Piwi knockdown and control lines. Note that Piwi expression is unaffected in follicular cells, leading to relatively weak apparent knockdown in RNA-seq libraries from whole ovaries. (*C*) Effect of Piwi knockdown on the expression of TEs. Two biological replicate RNA-seq experiments were carried out, and differential expression was assessed using DESeq. Transposons that show significant change ($P < 0.05$) are indicated by dark-red circles. Out of 217 individual RepeatMasker-annotated TEs, 15 show a significant increase in expression upon Piwi knockdown. (*D*) The change in the levels of TE transcripts and Pol II occupancy on their promoters upon Piwi knockdown. Twenty up-regulated and 10 down-regulated transposons with the most significant changes in expression level are shown. Note the low statistical significance for down-regulated transposons. For a complete list of transposons, see Supplemental Figure S2. (*E*) Pol II signal over the Het-A retrotransposon in control flies (shWhite; red) and upon Piwi knockdown (shPiwi; blue). (*F*) Increased abundance of transposon transcripts upon Piwi depletion correlates with increased Pol II occupancy over their promoters ($r^2 = 0.21$). Note that the majority of elements do not show significant change in either RNA abundance or Pol II occupancy.

upon Piwi knockdown in germ cells (Fig. 3C,D; Supplemental Fig. S2). We found little to no change of RNA levels for transposons whose activity is restricted to follicular cells of the ovary, indicating that the observed changes are indeed due to loss of Piwi in the germline (Supplemental Fig. S2). The analysis of Pol II ChIP-seq showed that Pol II occupancy increased over promoters of multiple TEs (Fig. 3D–F; Supplemental Fig. S3). Indeed, the change in steady-state levels of transposon transcripts upon Piwi depletion correlated with changes of Pol II occupancy (Fig. 3F). This result demonstrates that Piwi ensures low levels of transposon transcripts through a repressive effect on the transcription machinery.

To test whether Piwi-mediated transcriptional repression is accompanied by a corresponding change in chromatin state, we used ChIP-seq to analyze the genome-wide distribution of the repressive H3K9me3 mark in the ovary upon Piwi knockdown. We identified 705 genomic loci at which the level of H3K9me3 significantly decreased. More than 90% of the regions that show a decrease in the H3K9me3 mark upon Piwi depletion overlapped TE sequences, compared with the 33% that is expected from random genome sampling (Fig. 4A). Furthermore, these regions tend to be located in the heterochromatic portions of the genome that are not assembled on the main chromosomes (Fig. 4B). Only 20 of the identified regions localized to the euchromatic parts of the genome. Of these, 15 (75%) contained potentially active annotated copies of transposons. Taken together, our results indicate that Piwi is required for installment of repressive H3K9me3 chromatin marks on TE sequences of the genome.

While the vast majority of protein-coding host genes did not show significant changes in transcript level or Pol II occupancy upon Piwi knockdown, the expression of a small set of protein-coding genes (150 genes with a P-value <0.05) was significantly increased (Fig. 5A; Supplemental Table 1). There are several possible explanations for Piwi's effect on host gene expression. First, failure in the piRNA pathway might cause up-regulation of several genes that generate piRNAs in wild-type ovaries (Robine et al. 2009; Saito et al. 2009). However, the genes up-regulated in Piwi-deficient ovaries were not enriched in piRNAs compared with other genes. Second, H3K9me3 marks installed on TE sequences in a Piwi-dependent manner might spread into neighboring host genes and repress their transcription, as was recently demonstrated in a follicular cell culture model (Sienski et al. 2012). To address this possibility, we analyzed genomic positions of the genes whose expression was increased upon Piwi knockdown relative to genomic regions that showed a decrease in H3K9me3 marks. We found that up-regulated genes did not show a significant change in the H3K9me3 mark (Fig. 5B; Supplemental Fig. S4). Furthermore, the few genes located close to the regions that show a decrease in H3K9me3 signal had unaltered expression levels upon Piwi knockdown. Next, we analyzed the functions of up-regulated genes using gene ontology (GO) term classifications and found significant enrichment for proteins involved in protein turnover and stress and DNA damage response pathways (Fig. 5C). Particularly, we found that 31 subunits of the proteasome complex were overexpressed. Therefore, our analysis indicates that up-regulation of specific host genes is likely a secondary response to elevated transposon levels and genomic damage.

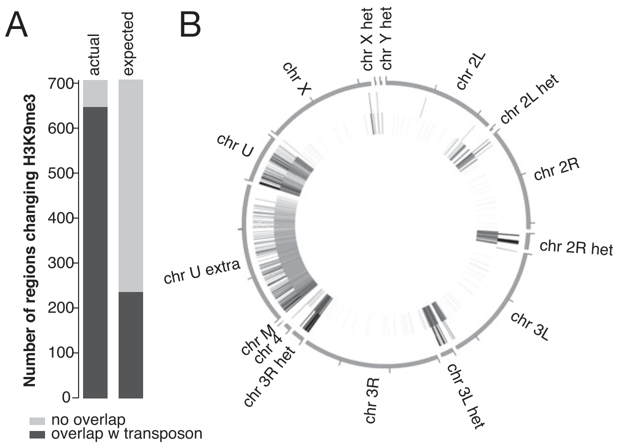


Figure 4. Piwi-induced transcriptional repression correlates with establishment of a repressive chromatin state. (*A*) Overlap between genomic regions of H3K9me3 depletion upon Piwi knockdown and TEs. Two replicates of H3K9me3 ChIP-seq experiments were carried out on control and Piwi-depleted ovaries, and enriched regions were identified using DESeq (see the Materials and Methods for details). A total of 705 regions show significant ($P < 0.05$) decrease in H3K9me3 occupancy upon Piwi knockdown, while only 30 regions showed a similarly significant increase. Out of the 705 regions that show a decrease in H3K9me3 marks upon Piwi knockdown, 91% (646) overlap with TE sequences compared with the 33% expected from random genome sampling. (*B*) Genomic positions of H3K9me3-depleted regions upon Piwi depletion (outer circle) and RepeatMasker-annotated transposons (inner circle). Note that almost all regions are localized in heterochromatic and repeat-rich portions of the genome (Het, chrU, and chrUExtra chromosomes).

In contrast to host genes, transcripts of TEs are targeted by piRNA. To directly address the role of piRNA in Piwi-mediated transcriptional silencing, we took advantage of a fly strain that expresses artificial piRNAs against the *lacZ* gene, which are loaded into Piwi complexes and are able to repress *lacZ* reporter expression in germ cells (Fig. 6A; Josse et al. 2007; Muerdter et al. 2012). Expression of piRNAs that are antisense to the reporter gene caused transcriptional silencing of the *lacZ* gene as measured by Pol II occupancy (Fig. 6B). Furthermore, we found that piRNA-induced silencing of the reporter gene was associated with an increase in the repressive H3K9me3 mark and HP1 occupancy and a decrease in the abundance of the active H3K4me2/3 marks at the reporter locus (Fig. 6C). This result is in good agreement with the genome-wide effect of Piwi depletion on distribution of the H3K9me3 mark and suggests that transcriptional silencing correlates with the establishment of a repressive chromatin structure and is mediated by piRNAs that match the target locus.

Discussion

Little is known about the function of nuclear piwi proteins. The nuclear piwi in mice (Miwi2) affects DNA methylation of TEs (Carmell et al. 2007; Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008). Several recent reports implicate *Drosophila* Piwi in regulation of chromatin marks on transposon sequences (Lin and Yin 2008; Klenov et al. 2011; Wang and Elgin 2011; Sienski et al. 2012). The mechanism of these processes is unknown in

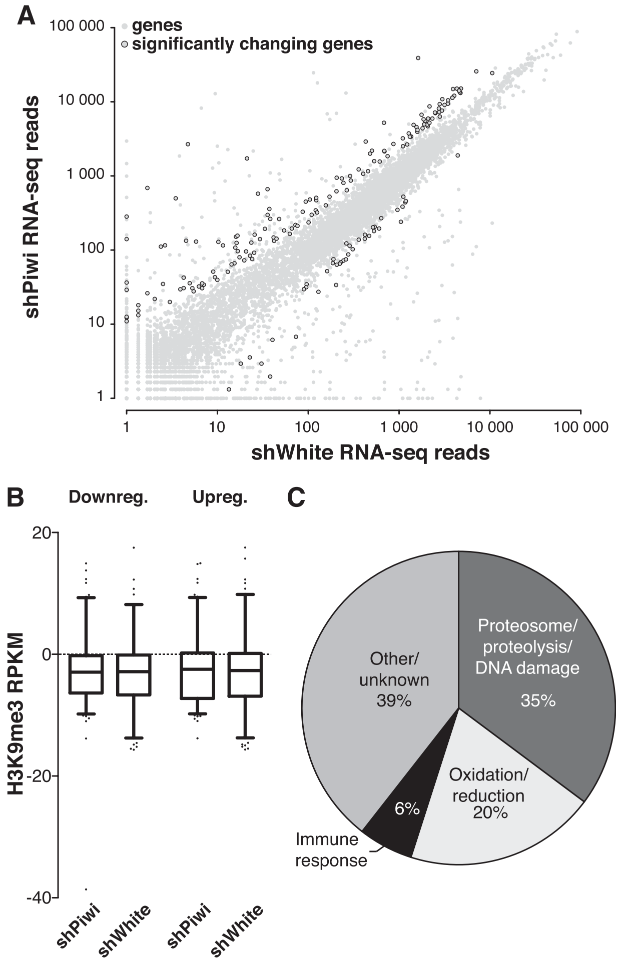


Figure 5. Piwi does not directly repress protein-coding genes. (*A*) Effect of Piwi knockdown on the expression of genes. Two replicate RNA-seq experiments were carried out, and differential expression was assessed using DESeq. Genes that show significant change ($P < 0.05$) are indicated by black circles. The vast majority of genes does not change significantly upon germline Piwi knockdown (shPiwi) compared with control (shWhite). (*B*) H3K9me3 mark density does not change over genes that show a significant change in expression upon Piwi knockdown (see Fig. 3C). Up-regulated and down-regulated genes are plotted separately. Signal indicated is after background subtraction. (*C*) Functional analysis of up-regulated genes by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) reveals activation of the protein degradation and DNA damage response pathways. Percentages of all up-regulated genes are indicated.

both organisms. Previously, Piwi was shown to associate with polytene chromosomes in salivary gland cells and colocalize with HP1, a chromodomain protein that binds to heterochromatin and a few loci in euchromatin, suggesting that HP1 mediates Piwi's interaction with chromatin (Brower-Toland et al. 2007). However, recent results showed that the putative HP1-binding site on Piwi is dispensable for Piwi-mediated transposon silencing (Wang and Elgin 2011).

We did not detect Piwi expression outside of the ovary and testis, including in salivary gland cells, using a GFP-Piwi transgene expressed under native regulatory elements. We detected GFP-Piwi on polytene chromosomes in ovarian nurse cells that have a germline origin; however, it localizes in a pattern that largely does not overlap with HP1. FLIP experiments with GFP-Piwi indicated a relatively fast rate of fluorescence redistribution as compared with histone H2A (Supplemental Fig. S1), implying a transient interaction of Piwi with chromatin. Our proteomic analysis of Piwi complexes isolated from *Drosophila* ovaries did not identify chromatin-associated factors but revealed several RNA-binding proteins, such as splicing and nuclear export factors that bind nascent RNA transcripts (Fig. 1B). Importantly, the interaction of most of these RNA-binding proteins with Piwi was dependent on RNA, indicating that Piwi associates with nascent transcripts. As Piwi itself lacks DNA- and RNA-binding domains (beyond the piRNA-binding domain),

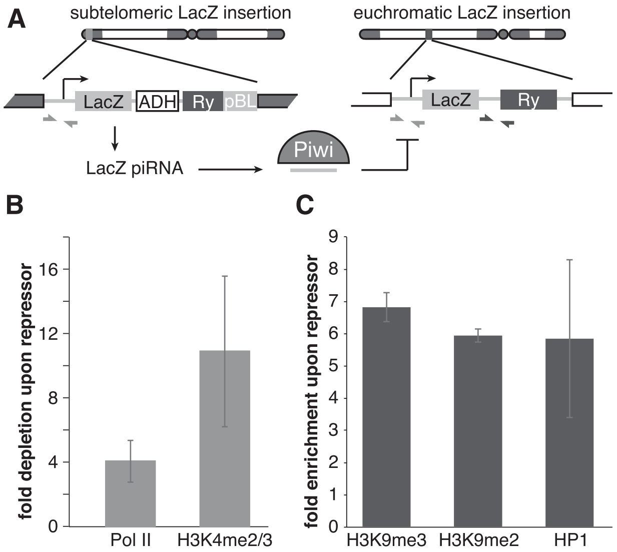


Figure 6. piRNA-dependent targeting of Piwi to a reporter locus leads to establishment of a repressive chromatin state and transcriptional silencing. (*A*) The mechanism of *trans*-silencing mediated by artificial piRNA and a schematic representation of the repressor and reporter *lacZ* constructs. The repressor construct is inserted in a subtelomeric piRNA cluster, leading to generation of piRNA from its sequence. Primers mapping to both constructs used for the Pol II and H3K4me2/3 ChIP-quantitative PCR (qPCR) are shown by light-gray arrows; primers specific to the reporter locus used for the H3K9me3, H3K9me2, and HP1 ChIP-qPCR are indicated by dark-gray arrows. (*B*) piRNAs induce transcriptional repression of the *lacZ* reporter. Pol II and H3K4me2/3 signals decreased on the *lacZ* promoter in the presence of artificial piRNAs as measured by ChIP-qPCR. Shown is the fold depletion of signal in flies that carry both repressor and reporter constructs compared with control flies that have only the reporter construct. The signal was normalized to RP49. (*C*) piRNAs induce an increase in H3K9me3 and H3K9me2 marks and HP1 binding as measured by ChIP-qPCR. Shown is the fold increase of corresponding ChIP signals downstream from the *lacZ* reporter in flies that carry both repressor and reporter constructs compared with control flies that have only reporter construct. The signal was normalized to RP49.

it is likely that the recruitment of Piwi to chromatin is through interactions with other RNA-binding proteins or sequence-specific interactions between Piwi-bound piRNA and nascent transcripts.

Using specific Piwi knockdown in germ cells of the *Drosophila* ovary, we analyzed the effect of Piwi depletion on gene expression, the transcription machinery, and H3K9me3 chromatin marks genome-wide. In agreement with previous results (Klenov et al. 2011), we found up-regulation of several TEs upon Piwi knockdown (Fig. 3C). The TEs that did not change their expression upon germline knockdown of Piwi might be expressed exclusively in somatic follicular cells of the ovary, such as the *gypsy* retrotransposon. Alternatively, some elements present in the genome might not have transcriptionally active copies, or the cytoplasmic AUB/AGO3 proteins may efficiently silence them at the post-transcriptional level.

The increase in steady-state levels of RNA upon Piwi depletion strongly correlates with an increase in Pol II occupancy on the promoters of transposons (Fig. 3D,F; Supplemental Fig S2). This result suggests that Piwi represses transposon expression at the transcriptional level, although we cannot completely exclude the possibility of an additional post-transcriptional effect. It was shown previously that depletion or mutation of Piwi leads to depletion of the repressive H3K9me3 mark and an increase in the active H3K4me2/3 marks on several transposon sequences (Klenov et al. 2011; Wang and Elgin 2011). Our ChIP-seq data extend these results to a genome-wide scale, proving that transposons are indeed the sole targets of Piwi, and demonstrate that changes in histone marks directly correlate with transcriptional repression.

Piwi depletion in the germline does not affect expression of the majority of host genes, although a small fraction of genes changes expression (Fig. 5A). One possible mechanism of the effect Piwi has on host genes is the spreading of repressive chromatin structure from transposon sequences to adjacent host genes. Indeed, such a spreading and the resulting repression of host gene transcription were observed in an ovarian somatic cell (OSC) culture model (Sienski et al. 2012). However, we did not find significant changes in the H3K9me3 mark for genes that are up-regulated upon germline depletion of Piwi, arguing against this mechanism playing a major role in host gene regulation. Instead, we found that the majority of host genes whose expression is increased as a result of Piwi depletion participate in protein turnover (e.g., proteasome subunits) and stress and DNA damage response pathways, indicating that they might be activated as a secondary response to cellular damage induced by transposon activation. The different effect of Piwi depletion on host gene expression in ovary and cultured cells might be explained by the fact that silencing of host genes due to transposon insertion would likely have a strong negative effect on the fitness of the organism but could be tolerated in cultured cells. Accordingly, new transposon insertions that cause repression of adjacent host genes should be eliminated from the fly population but can be detected in cultured cells. In agreement with this explanation, the majority of cases of repressive chromatin spreading in OSCs were observed for new transposon insertions that are absent in the sequenced *Drosophila* genome. Indeed, it was shown that the vast majority of new transposon insertions is present at a low frequency in the *Drosophila* population, likely due to strong negative selection (Petrov et al. 2003). Such selection was primarily attributed to the ability of TE sequences to cause recombination and genomic rearrangements. We propose that in addition to the effects on recombination, the selection against transposons can be driven by their negative impact on host gene expression in the germline linked to Piwi-mediated chromatin silencing.

How does Piwi discriminate its proper targets—transposons—from host genes? In the case of cytoplasmic Piwi proteins AUB and AGO3, recognition and post-transcriptional destruction of TE transcripts is guided by associated piRNAs. Our results indicate that piRNAs provide guidance for transcriptional silencing by the nuclear Piwi protein as well. First, in contrast to host genes that are not targeted by piRNAs, TE transcripts, which are regulated by Piwi, are recognized by antisense Piwi-bound piRNA (Brennecke et al. 2007). Second, a Piwi mutant that is unable to bind piRNA failed to rescue the piwi-null mutation despite its ability to enter the nucleus. Finally, expression of artificial piRNAs that target a reporter locus induced transcriptional silencing associated with an increase in repressive H3K9me3 and HP1 chromatin marks and a decrease in the active H3K4me2/3 marks (Fig. 6B,C). In contrast, the tethering of Piwi to chromatin in a piRNA-independent fashion by fusing Piwi with the lacI DNA-binding domain that recognizes lacO sequences inserted upstream of a reporter gene did not lead to silencing of the reporter (data not shown). Together, our results demonstrate that piRNAs are the essential guides of Piwi to recognize its targets for transcriptional repression.

It is tempting to propose that, similar to Argonautes in fission yeast, *Drosophila* Piwi directly recruits the enzymatic machinery that establishes the repressive H3K9me3 mark on its targets. Establishment of repressive marks can lead to stable chromatin-based transcriptional silencing that does not require further association of Piwi with target loci. This model explains why we found that Piwi is relatively mobile in the nucleus, indicative of only a transient interaction with chromatin. The Piwi-mediated transcriptional silencing has an interesting parallel in *Caenorhabditis elegans*, where the Piwi protein PRG-1 and associated 21U RNAs are able to induce stable transgenerational repression that correlates with formation of silencing chromatin marks on target loci. Interestingly, PRG-1 and 21U RNAs are necessary only for initial establishment of silencing, while continuing repression depends on siRNA and the WAGO group of Argonautes (Ashe et al. 2012; Bagijn et al. 2012; Buckley et al. 2012; Shirayama et al. 2012). Future studies should reveal the pathway that leads to transcriptional repression downstream from Piwi in *Drosophila* and the differences from and similarities to other species.

Materials and methods

Drosophila *stocks*

Nanos-Gal4-VP16 (BL4937), *UASp-shWhite* (BL33623), *UASp-shPiwi* (BL 33724), and Chr. I and II Balancer (BL7197) were purchased from the Bloomington Stock Center. GFP-Piwi-expressing flies (see below) were backcrossed onto the *piwi1/piwi2* (available from Bloomington Stock Center) background or the *otu7/otu11* (available from Bloomington Stock Center) background, respectively. *LacZ* reporter lines were a generous gift from S. Ronsseray.

Generation of transgenic fly lines

The GFP-Piwi, 3xFlag-HA-Piwi, and myc-Piwi constructs were generated using bacterial recombineering (Gene Bridges Counter Selection kit) to insert the respective tag after the start codon of the Piwi genomic region cloned in BAC clone BACN04M10. The KpnI–XbaI genomic fragment that contains the Piwi gene and flanking sequences was transferred to corresponding sites of the pCasper4 vector to create pCasper4/tagged Piwi.

The pCasper4/GFP-Piwi construct was used to generate pCasper4/GFP-Piwi-YK with two point mutations, Y551I and K555E. Mutations were introduced by PCR, amplifying products corresponding to a 3.1-kb upstream fragment and a 2.58-kb downstream fragment. The upstream fragment included a unique XbaI site at the 5′ end of the amplicon and overlapped 39 base pairs (bp) with the downstream fragment, which included a unique BamHI site at its 3′ end. The single XbaI–BamHI fragment was generated by overlap PCR with outside primers and cloned into corresponding sites of pCasper4/GFP-Piwi to replace the wild-type fragment. Transgenic flies were generated by P-element-mediated transformation (BestGene).

Immunoprecipitation of Piwi proteins and RNA gel of piRNA

Dissected ovaries were lysed in lysis buffer (20 mM HEPES at pH 7.0, 150 mM KCl, 2.5 mM MgCl, 0.5% Triton X-100, 0.5% Igepal, 100 U/mL RNasin [Promega], EDTA-free Complete Protease Inhibitor Cocktail [Roche]) and supernatant clarified by centrifugation. Supernatant was incubated with anti-eGFP polyclonal antibody (Covance) conjugated to Protein-G Dynabeads at 4°C. Beads were spiked with 5 pmol of synthesized 42-nucleotide RNA oligomer to assess purification efficiency, proteinase K-digested, and phenol-extracted. Isolated RNA was CIP-treated, radiolabeled using PNK and γ-P32-labeled ATP, and run on a 15% urea-PAGE gel. Western blots of ovary lysate and anti-eGFP immunoprecipitates were obtained from 8% SDS-PAGE gels and probed with polyclonal rabbit anti-eGFP antibody to confirm expression of the full-length transgene.

Mass spectrometric analysis of Piwi interaction partners

Lysis and clarification of ovary samples were performed as described above using lysis buffer with reduced detergent (0.1% Triton X-100, 0.1% Igepal). Piwi proteins with Flag, Myc, or GFP tag were purified from *Drosophila* ovaries using corresponding antibodies covalently coupled to M-270 epoxy Dynabeads (Invitrogen) (Cristea et al. 2005). Immunoprecipitation of free GFP from GFP-expressing ovaries was used as a negative control. Immunoprecipitations were performed in the presence or absence of RNase A (100 μg/mL; 30 min at 25C). Piwi and copurified interacting proteins were resolved on NuPAGE Novex 4%–12% Bis-Tris gels and stained with colloidal Coomassie blue. Gel fragments that contained protein bands were excised and in-gel-trypsinized, and the peptides were extracted following the standard protocol of the Proteome Exploration Laboratory at California Institute of Technology. Peptide analyses were performed on an LTQ-FT Ultra (Thermo Fisher Scientific) equipped with a nanoelectrospray ion source (Thermo Fisher Scientific) connected to an EASY-nLC. Fractionation of peptides was performed on a 15-cm reversed-phase analytical column (75-μm internal diameter) in-house-packed with 3-μm C18 beads (ReproSil-Pur C18-AQ medium; Dr. Maisch GmbH). Acquired spectra were searched against the *Drosophila melanogaster* proteome using the search engine Mascot (Matrix Science, version 2.2.06), and protein inferences were performed using Scaffold (Proteome Software, version 3). For an Excel file of Piwi interaction partners, see the Supplemental Material.

ChIP, ChIP-seq, and RNA-seq

ChIP was carried out using standard protocols (Moshkovich and Lei 2010). ChIP-seq and RNA-seq library construction and sequencing were carried out using standard protocols following the general principles described by Johnson et al. (2007) and Mortazavi et al. (2008), respectively. Data analysis was carried out using a combination of publicly available software tools and custom-written python scripts. Additional details regarding high-throughput data analysis are described in the Supplemental Material. For quantitative PCR (qPCR) primers, see Supplemental Table 2. GO term analysis of genes up-regulated upon Piwi knockdown was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang et al. 2009a,b) and FlyBase for additional assignment of GO terms. Sequencing data is available through Gene Expression Omnibus (accession no. GSE43829).

Antibodies

eGFP antibody (rabbit polyclonal serum; Covance) was affinity-purified in our laboratory. Anti-myc (Millipore), anti-Flag (Sigma), Pol II (ab5408), and Pol II pSer5 (ab5131) are commercially available.

Imaging of ovaries

Ovaries were fixed in 4% PFA in PBS for 20 min, permeabilized in 1% Triton X-100 in PBS, DAPI-stained (Sigma-Aldrich), washed, and mounted in 50% glycerol/PBS. Images were captured using an AxioImager microscope; an Apotome structured illumination system was used for optical sections (Carl Zeiss).

FLIP

FLIP time series were captured on an LSM510 confocal microscope equipped with a 40×/0.9 NA Imm Corr multi-immersion objective. Ovaries were dissected into halocarbon 700 oil (Sigma) and mounted under a 0.17-mm coverslip (Carl Zeiss) immediately before imaging. Two initial baseline images were captured, followed by 80–100 iterations consisting of two bleach iterations at 100% laser power (488 nm or 543 nm for GFP- and RFP-tagged proteins, respectively), followed by two images with reduced illumination intensity. FLIP series were cropped and median-filtered with a 2-pixel radius to reduce noise using FIJI (Schindelin et al. 2012) and the "Rigid Body" function of the StackReg plugin (Thévenaz et al. 1998) to correct drift when needed. Using Matlab software (The Mathworks), images were background-subtracted and corrected for acquisition bleaching. A value representing the true loss of intensity relative to the initial prebleach images, where 0 indicates no change in intensity and 1 represents complete photobleaching, was calculated for each pixel and each bleach/capture cycle and plotted with a color lookup table and calibration bar. Scale bars and annotations were made in Inkscape (http://inkscape.org).

Preparation of polytene squashes for immunofluorescence

Flies carrying the GFP-Piwi BAC construct were backcrossed onto the *otu[7]* and *otu[11]* background. Progeny from the cross of the two lines were grown at 18°C. Stage 7–12 egg chambers were separated and transferred to a polylysine-coated microscopic slide into PBST. From here, the "smush" protocol was followed (Johansen et al. 2009), but PFA cross-linking was reduced to 10 min. Slides were imaged using an AxioImager microscope and a 63× oil immersion objective (Carl Zeiss).

Acknowledgments

We are grateful to Evelyn Stuwe from the Aravin laboratory for purifying the GFP antibody; I. Antoshechkin of the Millard and Muriel Jacobs Genetics and Genomics Laboratory for sequencing; D. Trout, H. Amrhein, and S. Upchurch for computational assistance; the Bloomington Stock Center for fly stocks; and S. Hess, B. Graham, and M Sweredoski from the Proteome Exploration Laboratory at the Beckmann Institute, California Institute of Technology, for assistance with the mass spectrometry experiments. We thank members of the Aravin laboratory for critical comments on the manuscript. We thank Barbara Wold and members of the Wold laboratory for helpful discussions on ChIP protocols and analysis. A.K.R. and E.M.P. are supported by the Institutional Training Grant NIH/NRSA 5T32 GM07616, and E.M.P. is additionally supported by the Gordon Ross Medical Foundation. G.K.M. is supported by The Beckman Foundation, the Donald Bren Endowment, and NIH grant U54 HG004576. This work was supported by grants from the National Institutes of Health (R01 GM097363, R00 HD057233, and DP2 OD007371A to A.A.A.), the Searle Scholar Award (to A.A.A.), and the Ellison Medical Foundation New Scholar in Aging Award (to K.F.T.).

References

Agger K, Cloos P, Christensen J, Pasini D, Rose S, Rappsilber J, Issaeva I, Canaani E, Salcini A, Helin K. 2007. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature* **449:** 731–734.

Ameyar-Zazoua M, Rachez C, Souidi M, Robin P, Fritsch L, Young R, Morozova N, Fenouil R, Descostes N, Andrau J-C et al. 2012. Argonaute proteins couple chromatin silencing to alternative splicing. *Nat Struct Mol Biol* **19:** 998–1004.

Aravin A, Hannon G, Brennecke J. 2007. The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318:** 761–764.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31:** 785–799.

Ashe A, Sapetschnig A, Weick E-M, Mitchell J, Bagijn M, Cording A, Doebley A-L, Goldstein L, Lehrbach N, Le Pen J et al. 2012. piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. *Cell* **150:** 88–99.

Bagijn M, Goldstein L, Sapetschnig A, Weick E-M, Bouasker S, Lehrbach N, Simard M, Miska E. 2012. Function, targets, and evolution of *Caenorhabditis elegans* piRNAs. *Science* **337:** 574–578.

Brennecke J, Aravin A, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon G. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128:** 1089–1103.

Brower-Toland B, Findley S, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin S, Lin H. 2007. *Drosophila* PIWI associates with chromatin and interacts directly with HP1a. *Genes Dev* **21:** 2300–2311.

Buckley B, Burkhart K, Gu S, Spracklin G, Kershner A, Fritz H, Kimble J, Fire A, Kennedy S. 2012. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489:** 447–451.

Carmell M, Girard A, van de Kant H, Bourc'his D, Bestor T, de Rooij D, Hannon G. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12:** 503–514.

Chung W-J, Okamura K, Martin R, Lai E. 2008. Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. *Curr Biol* **18:** 795–802.

Cox D, Chao A, Baker J, Chang L, Qiao D, Lin H. 1998. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev* **12:** 3715–3727.

Cristea IM, Williams R, Chait BT, Rout MP. 2005. Fluorescent proteins as proteomic probes. *Mol Cell Proteomics* **4:** 1933–1941.

Djuranovic S, Zinchenko M, Hur J, Nahvi A, Brunelle J, Rogers E, Green R. 2010. Allosteric regulation of Argonaute proteins by miRNAs. *Nat Struct Mol Biol* **17:** 144–150.

Friedman R, Farh K, Burge C, Bartel D. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19:** 92–105.

Galiana-Arnoux D, Dostert C, Schneemann A, Hoffmann J, Imler J-L. 2006. Essential function in vivo for Dicer-2 in host defense against RNA viruses in *Drosophila*. *Nat Immunol* **7:** 590–597.

Ghildiyal M, Seitz H, Horwich M, Li C, Du T, Lee S, Xu J, Kittler E, Zapp M, Weng Z et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* **320:** 1077–1081.

Grewal S, Jia S. 2007. Heterochromatin revisited. *Nat Rev Genet* **8:** 35–46.

Gunawardane L, Saito K, Nishida K, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi M. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315:** 1587–1590.

Haase A, Fenoglio S, Muerdter F, Guzzardo P, Czech B, Pappin D, Chen C, Gordon A, Hannon G. 2010. Probing the initiation and effector phases of the somatic piRNA pathway in *Drosophila*. *Genes Dev* **24:** 2499–2504.

Handler D, Olivieri D, Novatchkova M, Gruber F, Meixner K, Mechtler K, Stark A, Sachidanandam R, Brennecke J. 2011. A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *EMBO J* **30:** 3977–3993.

Huang DW, Sherman B, Lempicki R. 2009a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1–13.

Huang DW, Sherman B, Lempicki R. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Hutvagner G, Simard M. 2008. Argonaute proteins: Key players in RNA silencing. *Nat Rev Mol Cell Biol* **9:** 22–32.

Ishizu H, Nagao A, Siomi H. 2011. Gatekeepers for Piwi–piRNA complexes to enter the nucleus. *Curr Opin Genetic Dev* **21:** 484–490.

Johansen K, Cai W, Deng H, Bao X, Zhang W, Girton J, Johansen J. 2009. Polytene chromosome squash methods for studying transcription and epigenetic chromatin modification in *Drosophila* using antibodies. *Methods* **48:** 387–397.

Johnson D, Mortazavi A, Myers R, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Josse T, Teysset L, Todeschini A-L, Sidor C, Anxolabéhère D, Ronsseray S. 2007. Telomeric *trans*-silencing: An epigenetic repression combining RNA silencing and heterochromatin formation. *PLoS Genetic* **3:** 1633–1643.

Keller C, Adaixo R, Stunnenberg R, Woolcock KJ, Hiller S, Buhler M. 2012. HP1(Swi6) mediates the recognition and destruction of heterochromatic RNA transcripts. *Mol Cell* **47:** 215–227.

Kiriakidou M, Tan G, Lamprinaki S, De Planell-Saguer M, Nelson P, Mourelatos Z. 2007. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* **129:** 1141–1151.

Klenov M, Sokolova O, Yakushev E, Stolyarenko A, Mikhaleva E, Lavrov S, Gvozdev V. 2011. Separation of stem cell maintenance and transposon silencing functions of Piwi protein. *Proc Natl Acad Sci* **108:** 18760–18765.

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri T et al. 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev* **22:** 908–917.

Lin H, Spradling A. 1997. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* **124:** 2463–2476.

Lin H, Yin H. 2008. A novel epigenetic mechanism in *Drosophila* somatic cells mediated by Piwi and piRNAs. *Cold Spring Harb Symp Quant Biol* **73:** 273–281.

Maison C, Almouzni G. 2004. HP1 and the dynamics of heterochromatin maintenance. *Nat Rev Mol Cell Biol* **5:** 296–304.

Mal'ceva N, Belyaeva E, King R, Zhimulev I. 1997. Nurse cell polytene chromosomes of *Drosophila melanogaster* otu mutants: Morphological changes accompanying interallelic complementation and position effect variegation. *Dev Genetic* **20:** 163–174.

Matzke M, Aufsatz W, Kanno T, Daxinger L, Papp I, Mette M, Matzke A. 2004. Genetic analysis of RNA-mediated transcriptional gene silencing. *Biochim Biophys Acta* **1677:** 129–141.

Mette M, Aufsatz W, van der Winden J, Matzke M, Matzke A. 2000. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J* **19:** 5194–5201.

Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* **5:** 621–628.

Moshkovich N, Lei E. 2010. HP1 recruitment in the absence of argonaute proteins in *Drosophila*. *PLoS Genetics* **6:** e1000880.

Muerdter F, Olovnikov I, Molaro A, Rozhkov N, Czech B, Gordon A, Hannon G, Aravin A. 2012. Production of artificial piRNAs in flies and mice. *RNA* **18:** 42–52.

Nakayama J, Rice J, Strahl B, Allis C, Grewal S. 2001. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292:** 110–113.

Olivieri D, Sykora M, Sachidanandam R, Mechtler K, Brennecke J. 2010. An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in *Drosophila*. *EMBO J* **29:** 3301–3317.

Onodera Y, Haag J, Ream T, Costa Nunes P, Pontes O, Pikaard C. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120:** 613–622.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol* **20:** 880–892.

Pöyhönen M, de Vanssay A, Delmarre V, Hermant C, Todeschini A, Teysset L, Ronsseray S. 2012. Homology-dependent silencing by an exogenous sequence in the *Drosophila* germline. *G3 (Bethesda)* **2:** 331–338.

Robine N, Lau N, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower M, Lai E. 2009. A broadly conserved pathway generates 3′UTR-directed primary piRNAs. *Curr Biol* **19:** 2066–2076.

Saito K, Nishida K, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi M. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev* **20:** 2214–2222.

Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, Kotani H, Asai K, Siomi H, Siomi M. 2009. A regulatory circuit for piwi by the large Maf gene traffic jam in *Drosophila*. *Nature* **461:** 1296–1299.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B et al. 2012. Fiji: An open-source platform for biological-image analysis. *Nat Methods* **9:** 676–682.

Shirayama M, Seth M, Lee H-C, Gu W, Ishidate T, Conte D, Mello C. 2012. piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell* **150:** 65–77.

Sienski G, Dönertas D, Brennecke J. 2012. Transcriptional silencing of transposons by piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* **151:** 964–980.

Siomi M, Sato K, Pezic D, Aravin A. 2011. PIWI-interacting small RNAs: The vanguard of genome defence. *Nat Rev Mol Cell Biol* **12:** 246–258.

Soppe WJ, Jasencakova Z, Houben A, Kakutani T, Meister A, Huang M, Jacobsen S, Schubert I, Fransz P. 2002. DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *EMBO J* **21:** 6549–6559.

Sugiyama T, Cam H, Verdel A, Moazed D, Grewal S. 2005. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proc Natl Acad Sci* **102:** 152–157.

Thévenaz P, Ruttimann U, Unser M. 1998. A pyramid approach to subpixel registration based on intensity. *IEEE Trans Image Process* **7:** 27–41.

Vagin V, Sigova A, Li C, Seitz H, Gvozdev V, Zamore P. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313:** 320–324.

Wang S, Elgin S. 2011. *Drosophila* Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. *Proc Natl Acad Sci* **108:** 21164–21169.

Wang X-H, Aliyari R, Li W-X, Li H-W, Kim K, Carthew R, Atkinson P, Ding S-W. 2006. RNA interference directs innate immunity against viruses in adult *Drosophila*. *Science* **312:** 452–454.

Zambon RA, Vakharia VN, Wu LP. 2006. RNAi is an antiviral immune response against a dsRNA virus in *Drosophila melanogaster*. *Cell Microbiol* **8:** 880–889.

H

Genome-Wide Analysis Reveals Coating of the Mitochondrial Genome by TFAM

Originally published as:

Wang YE, Marinov GK, Wold BJ, Chan DC. 2013. Genome-Wide Analysis Reveals Coating of the Mitochondrial Genome by TFAM. *PLoS ONE*. 2013. **8**(8):e74513. doi: 10.1371/journal.pone.0074513.

Genome-Wide Analysis Reveals Coating of the Mitochondrial Genome by TFAM

Yun E. Wang[1], Georgi K. Marinov[1], Barbara J. Wold[1], David C. Chan[1,2]*

1 Division of Biology, California Institute of Technology, Pasadena, California, United States of America, **2** Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California, United States of America

Abstract

Mitochondria contain a 16.6 kb circular genome encoding 13 proteins as well as mitochondrial tRNAs and rRNAs. Copies of the genome are organized into nucleoids containing both DNA and proteins, including the machinery required for mtDNA replication and transcription. The transcription factor TFAM is critical for initiation of transcription and replication of the genome, and is also thought to perform a packaging function. Although specific binding sites required for initiation of transcription have been identified in the D-loop, little is known about the characteristics of TFAM binding in its nonspecific packaging state. In addition, it is unclear whether TFAM also plays a role in the regulation of nuclear gene expression. Here we investigate these questions by using ChIP-seq to directly localize TFAM binding to DNA in human cells. Our results demonstrate that TFAM uniformly coats the whole mitochondrial genome, with no evidence of robust TFAM binding to the nuclear genome. Our study represents the first high-resolution assessment of TFAM binding on a genome-wide scale in human cells.

Citation: Wang YE, Marinov GK, Wold BJ, Chan DC (2013) Genome-Wide Analysis Reveals Coating of the Mitochondrial Genome by TFAM. PLoS ONE 8(8): e74513. doi:10.1371/journal.pone.0074513

Editor: Yidong Bai, University of Texas Health Science Center at San Antonio, United States of America

Received June 27, 2013; **Accepted** August 5, 2013; **Published** August 26, 2013

Funding: This work was supported by NIH grants GM062967 (DCC), U54 HG004576 (BJW), U54 HG006998 (BJW), the Beckman Foundation, and the Donald Bren Endowment. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: dchan@caltech.edu

Introduction

Mitochondria are essential eukaryotic organelles, serving as the epicenter of ATP production in the cell through oxidative phosphorylation. To perform this bioenergetic function, mitochondria utilize gene products encoded by the mitochondrial genome, a circular DNA that is 16.6 kb long. This genome is organized into DNA/protein structures termed nucleoids [1]. Mitochondrial DNA (mtDNA) encodes thirteen components of the electron transport chain, as well as 22 tRNAs and two ribosomal RNA genes. These gene products are essential for the proper function of the respiratory chain, and therefore maintenance of mtDNA levels and sequence fidelity is essential for cellular bioenergetics. In a human cell, there are hundreds to thousands of copies of the mtDNA genome [2,3]. Damage or depletion of mtDNA causes numerous inherited disorders, including Alpers' Disease, ataxia neuropathy spectrum, and progressive external ophthalmoplegia [4,5]. Furthermore, loss and damage to mtDNA has been implicated in cardiovascular disease [6–9], diabetes [10–12], neurodegenerative disorders such as Alzheimer's [13,14], and aging [15,16]. Strikingly, increasing mtDNA copy number promotes cell survival or function in many models of disease associated with decreased mtDNA abundance, such as diabetes [12,17], aging [18], Alzheimer's [19], and Parkinson's [20,21]. Thus, it is critical to understand how mtDNA copy number and integrity are maintained.

Mitochondrial transcription factor A (TFAM) is a DNA binding protein that plays multiple roles in regulating mtDNA function. As a sequence-specific transcription factor, it binds upstream of the light strand promoter (LSP) and heavy strand promoter 1 (HSP1) to activate initiation of transcription. At these sites, the footprint of TFAM binding is ~22 bp long [22,23]. As a result, TFAM is essential for production of gene products from the mitochondrial genome. In addition, TFAM is required for normal mtDNA copy number, because RNA primers generated from LSP are used to prime mtDNA replication [24,25]. Mice heterozygous for a knockout of TFAM exhibit not only an expected reduction (22%) in mitochondrial transcript levels in the heart and kidney, but also a universal 34% reduction in mtDNA copy number across all assayed tissues. Furthermore, homozygous knockout mice have no detectable levels of mtDNA and die during embryogenesis [26], highlighting the importance of TFAM in maintenance of mtDNA levels and in cellular and organismal viability.

Apart from its sequence-specific functions, TFAM is thought to organize the mtDNA genome by coating it in a nonspecific manner. Although how TFAM packages mtDNA is not well-understood, it is known to bind nonspecifically to DNA [27] and is estimated to be sufficiently abundant to coat the genome completely [28–30]. One model suggests that nonspecific binding radiates from the TFAM LSP binding site, which acts as a nucleation site for subsequent cooperative binding in a phased pattern to yield an inter-genome homogeneous pattern of binding [31,32]. The packaging function of TFAM appears to have important consequences for maintenance of the mtDNA genome. A TFAM variant that is deficient in transcriptional activation but competent in DNA binding is capable of preventing mtDNA depletion [33]. Therefore, as a prominent component of mtDNA nucleoids, TFAM appears to coat the mitochondrial genome, perhaps protecting it from turnover or deleterious damage.

Despite the importance of the associations of TFAM with mtDNA in the maintenance of mtDNA integrity and in cellular viability, these interactions have only been visualized in vivo at low resolution [34]. Therefore, to capture a high-resolution profile of TFAM-mtDNA interactions across the entire mitochondrial genome, we performed chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) for TFAM in human HeLa cells.

Results

Detection of TFAM-DNA interactions using ChIP-seq

To characterize TFAM binding to both the mitochondrial and nuclear genomes in an unbiased manner, we performed ChIP-seq targeting TFAM in HeLa cells. Because ChIP-seq data is highly dependent on the use of high-quality antibodies, we generated two new TFAM monoclonal antibodies (20G2C12 and 20F8A9) that efficiently immunoprecipitated TFAM (Figure 1A). Both of these antibodies gave clean mitochondrial and nucleoid signals in immunofluorescence experiments with cultured HeLa cells (Figure 1C,D). The 20G2C12 antibody also performed well in Western blots of whole-cell lysates, recognizing a single protein band of ~23 kDa (Figure 1B).

Given the high efficiency of 20G2C12 in immunoprecipitating TFAM, as well as its high specificity, we used it to capture TFAM-associated DNA fragments for ChIP-seq analysis. DNA was sonicated prior to immunoenrichment and size-selected prior to library building so that the average fragment length of the final library was centered around 200 bp, a fragment distribution allowing for high-resolution deconvolution of binding events. We generated 3 replicates and matching controls. The sequencing depth of all samples was between 18 million and 48 million mappable reads, which is generally sufficient for comprehensive identification of transcription factor binding sites [35].

A common concern with ChIP-seq datasets is the variability of enrichment for true binding events as compared to background. In a typical ChIP-seq experiment, a minority of sequencing reads originates from binding events, with the majority representing random genomic DNA. Even for the same DNA binding factor, large variations in the strength of enrichment can be observed, and therefore it is critical to assess the degree of enrichment before downstream analysis. A number of ChIP-seq quality control metrics have been developed [35] for nuclear transcription factors. However, TFAM is expected to bind to the mitochondrial genome, which has very different characteristics from the nuclear genome. In addition, it is predicted to bind both in the classical localized manner [36] as well as broadly across the mitochondrial genome. As a result, metrics for evaluating nuclear transcription factors are not well-suited for analysis of TFAM binding data. We therefore examined the fraction of sequencing reads in our libraries mapping to the mitochondria as a proxy for the enrichment of TFAM binding events. Strikingly, between 30% and 75% of TFAM ChIP-seq reads mapped to the mitochondrial genome, while less than 2% of reads mapped to the mitochondrial genome in the input samples, indicating that our TFAM ChIP-seq datasets are indeed highly enriched for TFAM binding events (Figure 1B). We note that 75% ChIP enrichment is extremely high (in fact, practically unprecedented) for any transcription factor dataset [35], thus underscoring the high experimental quality of our datasets.

Because partial copies of the mitochondrial genome are also present in the nuclear genome, not all reads originating from mtDNA can be mapped uniquely. Therefore, we characterized TFAM binding to mtDNA and to the nuclear genome separately. We analyzed mitochondrial binding events by aligning sequencing reads to the mitochondrial genome alone (restricting our analysis to reads mapping perfectly without any mismatches to further increase mapping accuracy), and analyzed binding to the nuclear genome by aligning only the reads which did not map to the mitochondrial genome, as outlined in Figure 2A. For a standard nuclear transcription factor, this approach may cause some reads originating from the nuclear genome to artificially map to the mitochondrial genome. However, given that TFAM is known to bind to the mitochondrial genome and the extremely high enrichment for TFAM binding to mtDNA in our TFAM ChIP-seq libraries, this should not be a significant confounding factor.

TFAM coats the mitochondrial genome

As discussed above, TFAM has not only been proposed to bind specifically to well-defined binding sites in the D-loop, but has also been suggested to play a nonspecific packaging role in the nucleoid that is essential for mtDNA integrity. However, little is known about the pattern of non-specific binding of TFAM to the mitochondrial genome. Localized binding at the D-loop and diffuse binding across the rest of the genome are expected to result in distinct ChIP-seq signal profiles. Localized, “point-source” binding to DNA results in an asymmetric distribution of reads mapping to the forward and reverse strand around the binding site of the protein [36,37], while diffuse binding does not produce such strand asymmetry.

To characterize TFAM binding to mtDNA, we examined the forward and reverse strand read distribution after mapping TFAM ChIP-seq and input library reads to the mitochondrial genome. Strikingly, we did not observe regions of obvious enrichment and strand asymmetry in the D-loop; in particular, we did not see specific binding at the predicted HSP1 and LSP sites. On the whole, the TFAM ChIP-seq signal was broadly distributed over the whole mitochondrial chromosome, and

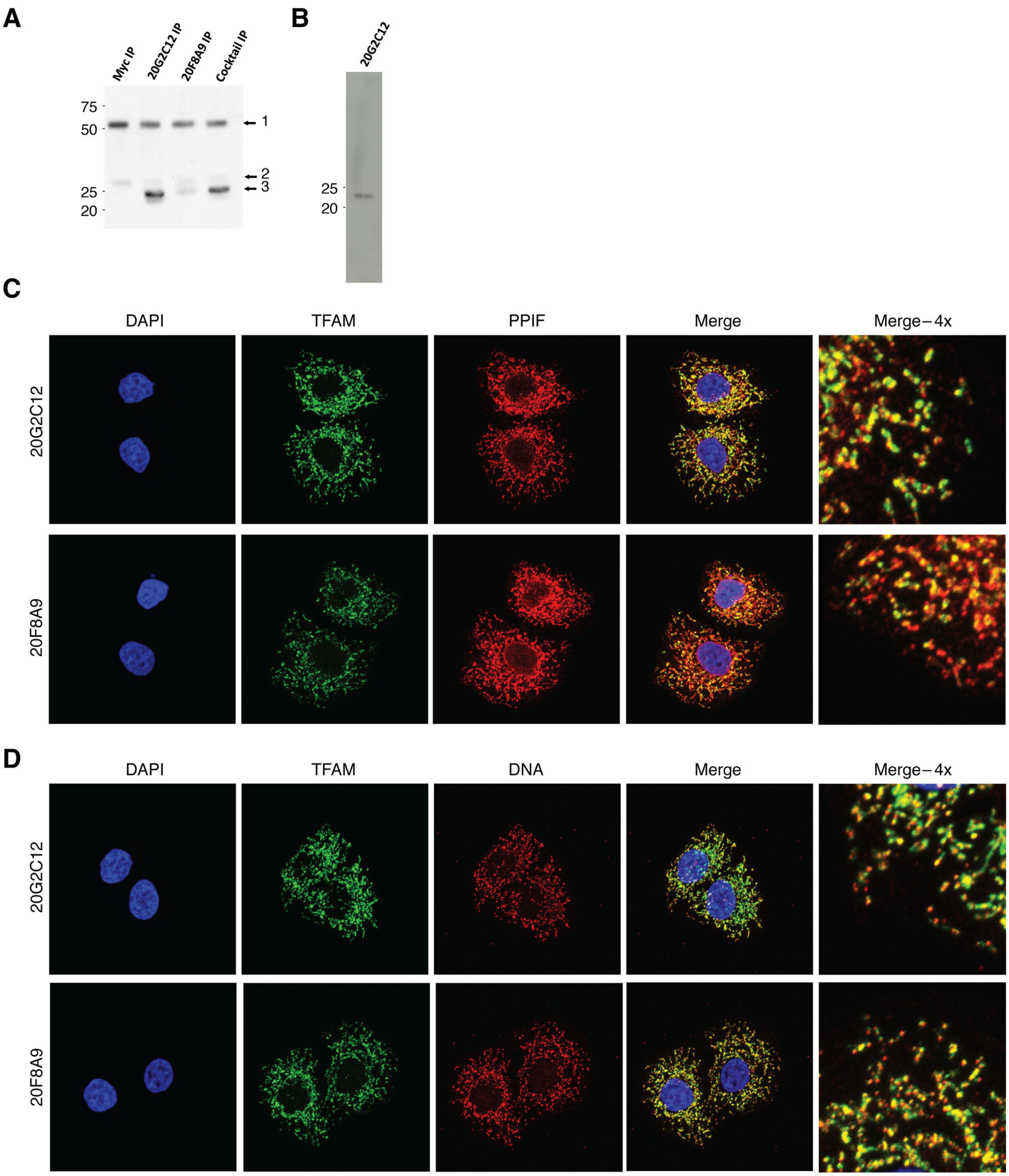


Figure 1. Characterization of TFAM monoclonal antibodies. (A) Immunoprecipitation of TFAM from cell lysates. HeLa cell lysate was applied to sheep anti-mouse Dynabeads conjugated to anti-Myc, 20G2C12 TFAM antibody, 20F8A9 TFAM antibody, or a 50/50 mixture of 20G2C12 and 20F8A9 TFAM antibodies. The labeled bands are: 1) Antibody heavy chain; 2) antibody light chain; 3) TFAM. (B) Western blot using the 20G2C12 antibody detects a ~23kDa band. (C and D) Immunocytochemistry showing TFAM localization. Mitochondria were identified by PPIF staining; mtDNA was identified by anti-DNA staining. There was no evidence for nuclear localization of TFAM using either antibody.

doi: 10.1371/journal.pone.0074513.g001

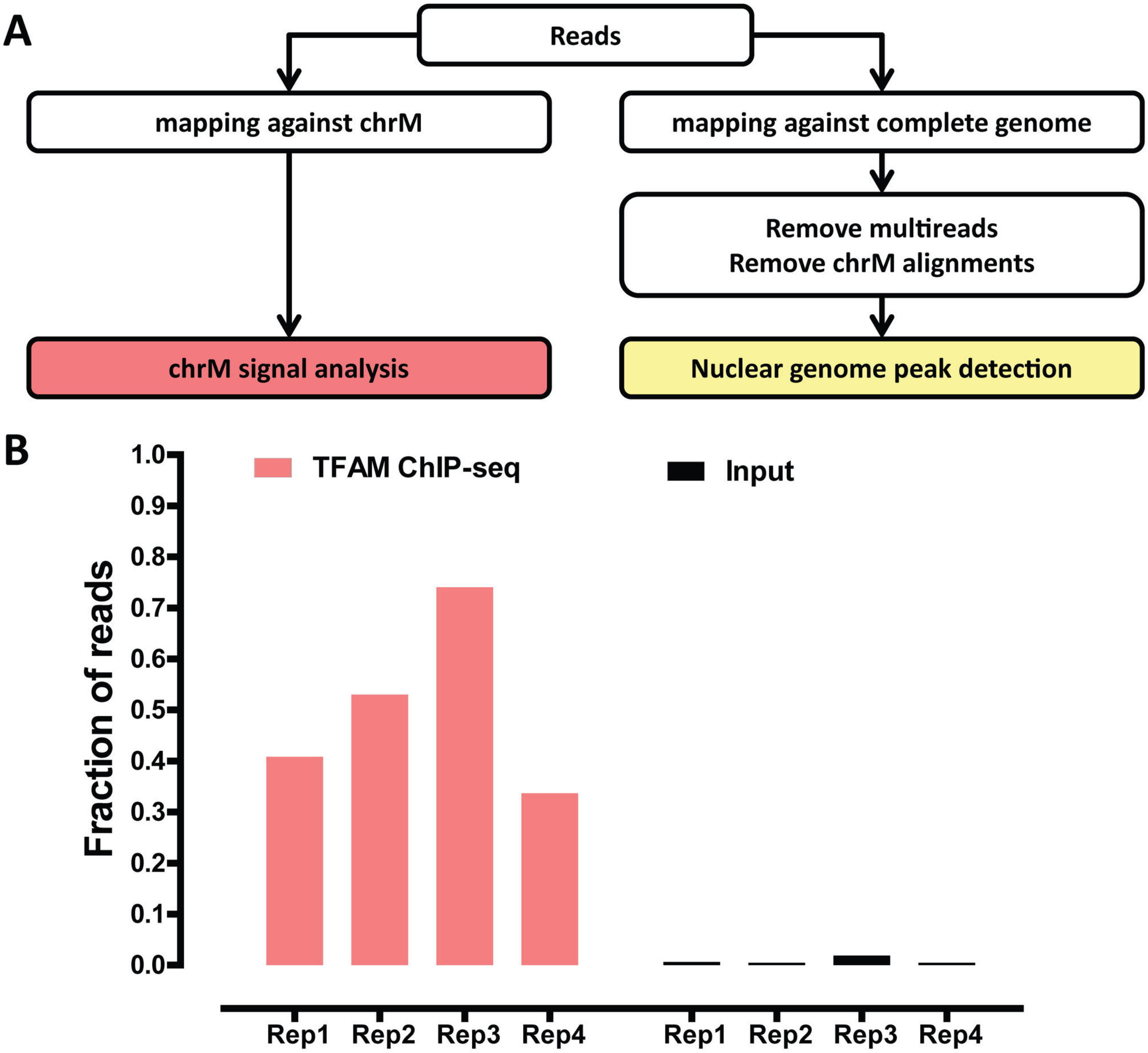


Figure 2. ChIP-seq analysis of genome-wide TFAM binding. (A) Overview of computational processing of data. Reads were trimmed to 36 bp and then either mapped against the mitochondrial genome (ChrM), or the complete hg19 version of the genome. After removing multireads and alignments to the mitochondrial genome, peaks in the nuclear genome were called using MACS2. (B) The proportion of sequencing reads mapping to chrM in ChIP and input datasets. All replicates of the ChIP-seq resulted in at least 30% of reads mapping to the mitochondrial genome, much greater than the 0.4-1.9% of reads mapping to mtDNA in the input datasets. Replicates 1-3 were performed using the 20G2C12 antibody, while Replicate 4 was performed using the 20F8A9 antibody.

doi: 10.1371/journal.pone.0074513.g002

while coverage was not perfectly uniform, the amplitude of the non-uniformity was not significant, and the signal profile closely tracked that of the input sample (Figure 3). The low level of non-uniformity likely results from sequencing biases, which has been documented to skew coverage [38,39]. Because our libraries were carefully size-selected for fragments in the 200 bp range, discrete TFAM binding sites would be expected to yield discrete signal localizations. Therefore, we interpret these results as evidence for the uniform coating of the whole mitochondrial genome by TFAM. We observed one region of apparent localized enrichment exhibiting strand asymmetry in the ND2 ORF near the origin of light strand replication (O_L) (Figure 3F), which we discuss in the Discussion section.

To further verify our results, we carried out ChIP-seq against TFAM with a second TFAM monoclonal antibody, 20F8A9. We obtained similar results (Figure S1) and found significant

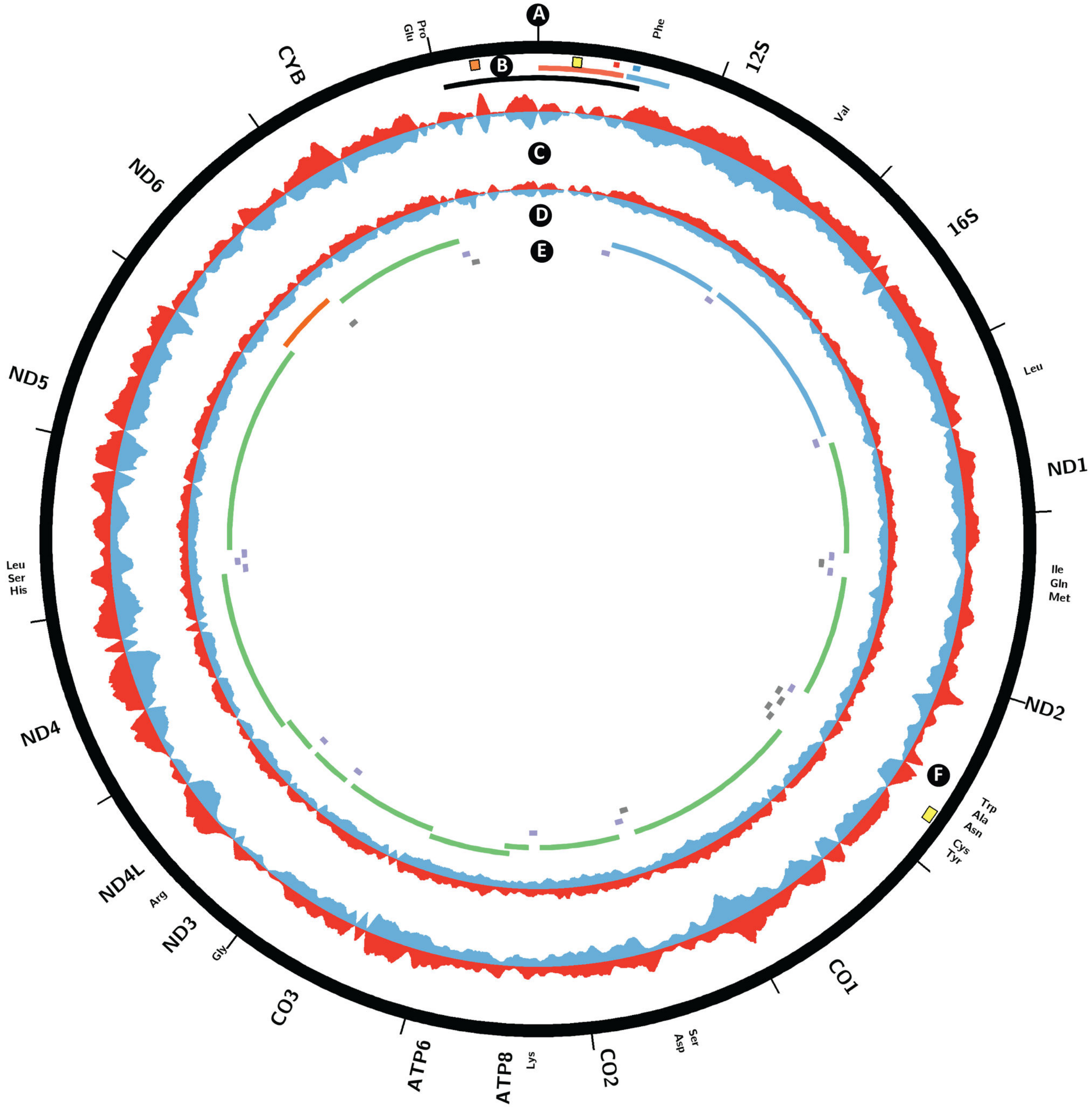


Figure 3. Coating of the mitochondrial genome by TFAM in HeLa cells. Circos plot of plus strand and minus strand TFAM ChIP-seq and input read density signal over chrM. (A, E) Annotation of protein coding (green on forward/heavy strand, red on reverse/light strand), ribosomal RNA (blue) and tRNA (blue on forward/heavy strand, grey on reverse/light strand) transcripts. (B) D-loop (black), LSP promoter (large red tile), known LSP TFAM binding site (small red tile), HSP promoter (large blue tile), known HSP1 TFAM binding site (small blue tile), and origins of heavy strand replication (Ori-b, orange tile; O_H, yellow tile). (C) TFAM ChIP-seq signal on forward (red) and reverse (blue) strands. (D) Input signal on forward (red) and reverse (blue) strands. (F) Origin of light strand replication (yellow tile). Note that the input signal is exaggerated 60-fold relative to the ChIP-seq signal in order to visualize coverage irregularities. The signal from the TFAM ChIP-seq largely follows that of the input, indicating generalized binding across the mitochondrial genome.

doi: 10.1371/journal.pone.0074513.g003

correlation between the 20F8A9 dataset and the three datasets obtained from the 20G2C12 antibody datasets ($p < 0.0001$).

No evidence for binding to the nuclear genome

Previous studies have suggested that TFAM can be found in the nucleus and that it modulates the transcription of nuclear genes. In rat neonatal cardiac myocytes, TFAM was found to bind to the promoter of SERCA2, the homolog of human sarco(endo) plasmic reticulum calcium-ATPase 2 (ATP2A2), and was implicated in regulating its transcription [40]. Given the extremely high degree of TFAM binding enrichment in our datasets, any robust nuclear TFAM binding events should be readily detectable. To analyze nuclear binding, we excluded all sequencing reads mapping to the mitochondrial genome and used the resulting set of reads to identify putative TFAM binding sites. We first looked for significant global read clustering using cross-correlation between reads mapping to the forward and the reverse DNA strands [35,36]. Cross-correlation plots for input samples and for TFAM ChIP-seq datasets were indistinguishable from each other (Figure 4A,B). Next, we called putative TFAM binding sites using MACS2 [41]. Using default settings (corresponding to a q-value cut-off of 10^{-2}), we identified 72, 137 and 153 sites respectively for the three replicates generated with antibody 20G2C12, and a single site for the 20F8A9 antibody. However, manual inspection of each of the identified sites revealed that all were likely to represent artifacts, mostly associated with repetitive DNA sequences, as none had the expected strand asymmetry of read distribution around a binding site. Instead, the two strand profiles at each site were identical (summarized in Figure 4D, with the classic nuclear transcription factor NRSF shown for comparison in Figure 4C), and numerous unmappable regions and repetitive elements were present in the immediate vicinity of many of the called sites. Inspection of the ATP2A2 gene revealed no TFAM enrichment neither in the promoter region nor anywhere else in the neighborhood of the gene (Figure 4E). Furthermore, we do not detect nuclear localization of TFAM in our cells (Figure 1C). Therefore, in HeLa cells under normal growth conditions, we find no evidence for specific binding of TFAM to nuclear target genes.

Discussion

Previous in vitro studies have suggested that TFAM binds specifically to LSP and HSP1, and that it may also bind nonspecifically in a phased manner. Furthermore, evidence has been presented for its nuclear localization and action as a canonical nuclear transcription factor in rat neonatal cardiac myocytes. However, no direct genome-wide measurements of TFAM binding have been previously reported. Our TFAM ChIP-seq data reveal very high enrichment for reads mapping to the mitochondrial genome, but a binding pattern that largely mirrors the read distribution observed in the input DNA, suggesting broad, non-specific binding to mitochondrial genome. This pattern is highly reproducible, indicating that the average population-wide state of TFAM-mtDNA interactions is stable. We found no correlation between irregularities in TFAM signal distribution and characteristics of the mitochondrial genome such as GC content (data not shown). Thus, we conclude that TFAM binds to the mitochondrial genome nonspecifically and without bias when cells are grown under typical culture conditions. Although we do not observe the synchronized phased binding seen in in vitro studies, we cannot rule out a model where individual mtDNAs have such a pattern of binding initiating from a non-universal nucleation site.

Strikingly, we did not observe localized enrichment of binding at the known LSP and HSP1 TFAM binding sites. Peak patterns mirrored that of the input in these regions, and no ChIP-seq peaks displaying the canonical strand asymmetry in read distribution were observed. This finding can be explained by a model in which the interaction of TFAM with the LSP and HSP1 binding sites is relatively transient and infrequent compared to a more stable non-specific association with the genome in its packaging state.

We did detect one site in the genome exhibiting the characteristics of a specific, localized ChIP-seq peak, centered at 5175 bp in the ND2 ORF. The localized nature of the ChIP signal at this site suggests higher occupancy of TFAM. This peak localizes to 546 bp upstream of the O_L. Strikingly, TFAM has previously been localized 520 bp upstream of the O_L of rat mtDNA [42–44]. We found no sequence similarity between the rat and human sites, and in general this region of the mtDNA genome shows low homology between the two species. Further work will be required to understand the significance of this putative TFAM binding site.

Finally, analysis of all datasets for TFAM binding to the nuclear genome yielded no hits distinguishable from common ChIP-seq artifacts. Although Watanabe et al. observed regulation of the SERCA2 gene in rat myocytes, we did not detect TFAM binding at the promoter of its ortholog in humans. Previous studies have shown nuclear localization of TFAM in rat hepatoma cells [45], as well as an alternate isoform of TFAM in mouse testis nuclei [46]. We have thus far been unable to detect nuclear TFAM localization in HeLa cells (Figure 1C), suggesting that nuclear localization and transcriptional regulation may be cell type or perhaps species-dependent. ChIP-seq in different cell lines may be able to detect such nuclear interactions.

We demonstrate here the first high-resolution ChIP-seq analysis of TFAM binding to the mitochondrial genome. Aside from generalized, largely non-specific binding across the mitochondrial genome, we detected a putative specific binding site upstream of the origin of light strand replication. We do not observe the expected binding at the known HSP1 and LSP sites, nor did we identify any nuclear binding sites. An area that remains to be explored is the dynamic nature of TFAM-DNA interactions with respect to both the nuclear and mitochondrial genomes. ChIP-chip on the yeast mitochondrial genome has shown that metabolic changes can lead to differential binding of the yeast TFAM homolog, Abf2p [47]. It is possible that such remodeling also occurs in the mammalian system, and further studies will provide insight into the dynamic nature of the mtDNA-protein interactions within the nucleoid that serve to protect its integrity.

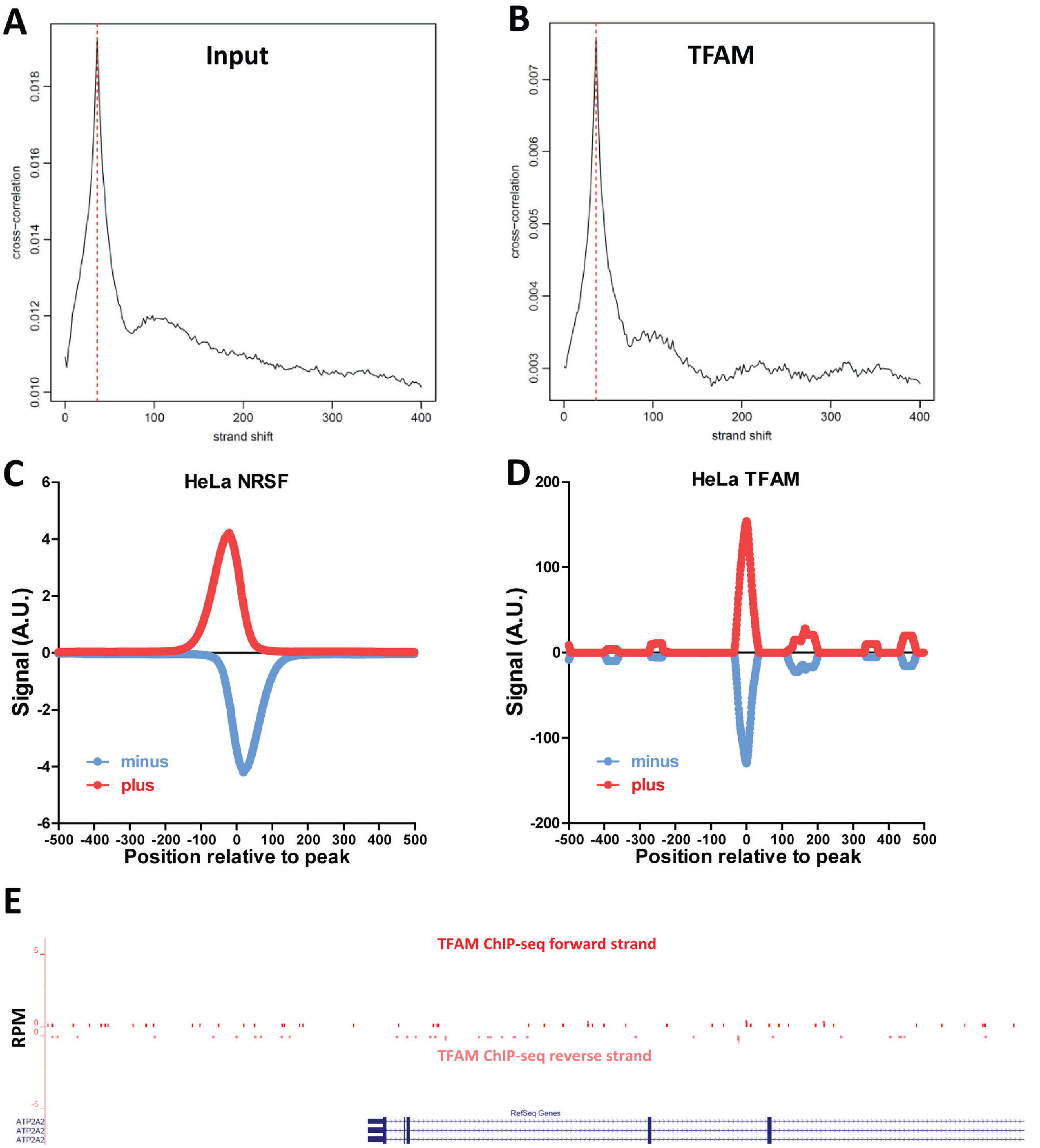


Figure 4. Absence of TFAM binding to the nuclear genome. (A) Cross-correlation plot of input DNA computed over the nuclear genome. (B) Cross-correlation plot of TFAM ChIP-seq computed over the nuclear genome. (C) Distribution of ChIP-seq reads mapping to the plus and minus strand around called binding sites in a ChIP-seq dataset for the NRSF transcription factor [51] in HeLa cells, generated by the ENCODE consortium [52]. (D) Distribution of TFAM ChIP-seq reads mapping to the plus and minus strand around called binding sites indicates lack of real binding sites. (E) No ChIP-seq enrichment around the promoter of the SERCA2/ATP2A2 gene, previously suggested to be a TFAM target.

doi: 10.1371/journal.pone.0074513.g004

Materials and Methods

Cell growth and treatment

HeLaS3 cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Invitrogen #11995) containing 10% bovine serum (Invitrogen #16170), penicillin and streptomycin, and additional L-glutamine (2mM). Cells were fed 24 hours before harvest for ChIP-seq, which was performed at 80-90% confluency.

Antibody Production and characterization

Antibodies were produced by the Caltech Monoclonal Antibody Facility and raised against the full-length TFAM protein in mouse. Immunoprecipitation with 20G2C12 and 20F8A9 TFAM antibodies and Myc antibody (Santa Cruz #sc-40) was performed according to established protocols using M-280 sheep anti-mouse Dynabeads (Invitrogen #11201D). Immunoblotting of IP products was performed using a monoclonal TFAM 18G102B2E11 antibody, also custom generated, at 1:2000, with goat anti-mouse HRP antibody (1:10,000, Jackson ImmunoResearch #115-056-003). Immunoblotting of HeLa whole cell lysate with 20G2C12 was performed at a 1:200 dilution and with goat anti-mouse HRP antibody.

Immunocytochemistry

HeLa cells cultured as described above were plated onto poly-lysine coated glass coverslips 48 hours prior to fixation in formaldehyde and permeabilization with 0.1% Triton X-100. For colocalization of TFAM to mitochondria, 20G2C12 or 20F8A9 antibodies were used at 1:10 in conjunction with PPIF at 1:200 (ProteinTech #18466-1-AP). Secondary antibodies were goat anti-mouse AF488 (1:500, Invitrogen #A11001) and donkey anti-rabbit AF546 (1:500, Invitrogen #A10040). Cells were also stained with DAPI to visualize nuclei. Immunocytochemistry to visualize colocalization of mitochondrial nucleoids and TFAM was performed sequentially due to both antibodies being raised in mouse. Sequential immunostaining yielded no background fluorescence due to cross-antibody reactivity (data not shown). Order was as follows: anti-TFAM antibody (1:10); goat anti-mouse AF488 (1:500, Invitrogen #A11001); anti-DNA antibody (1:25, Millipore #CBL186); goat anti-mouse AF555 (1:500, Invitrogen #A21426), DAPI. Images were acquired with a Zeiss LSM 710 confocal microscope with PlanApochromat 63X/1.4 oil objective. Z-stack acquisitions were converted to maximum z-projections using ImageJ software.

Chromatin immunoprecipitation and sequencing

ChIP experiments and preparation of DNA for sequencing were performed following standard procedures [48] with some modifications. Cells were fixed for 10min at RT in 1% formaldehyde, harvested using a cell scraper, washed once in ice-cold PBS, and resuspended in RIPA buffer with protease inhibitor. The sample was then sonicated using a 3.2mm microtip (QSonica Sonicator 4000) at 30s on/30s off intervals and 40% amplitude for 180min while in a -30°C 3:1 isopropanol and water bath containing dry ice. Subsequent steps were performed as per the standard protocol. DNA was size-selected during library building to an average fragment size of 200bp. Libraries were sequenced using Illumina GAIIx and Illumina HiSeq 2000. Sequencing data is available under GEO accession record GSE48176.

Sequencing data processing and analysis

Sequencing reads were trimmed down to 36 bp and then mapped against either the female set of human chromosomes (excluding the Y chromosome and all random chromosomes and haplotypes) or the mitochondrial genome alone, using the hg19 version of the human genome as a reference. Bow tie 0.12.7 [49] was used for aligning reads, not allowing for any mismatches between the reads and the reference. ChIP-seq peaks were called using MACS2 [41] with default settings except for the mfold parameter, which was lowered to (2,30). Circos plots were generated using Circos version 0.60 [50]. Additional data processing was carried out using custom-written python scripts. ENCODE data was downloaded from the UCSC browser (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs) and its use here complies with its terms of usage. Pearson correlation coefficient, t-test, and p values were calculated using embedded and custom Microsoft Excel functions.

Supporting Information

Figure S1. Comparison of profiles of TFAM binding to mitochondrial genome.
Circos plots of TFAM ChIP-seq experiments: (1) 20F8A9 antibody ChIP-Seq; (2) 20G2C12 replicate 1; (3) 20G2C12 replicate 2; (4) 20G2C12 replicate 3. Read profiles are very similar across replicates and antibodies.
(TIF)

Acknowledgements

We thank Igor Antoshechkin and the Millard and Muriel Jacobs Genetics and Genomics Laboratory for assistance with library building and high-throughput sequencing. We thank Henry Amrhein and Diane Trout for computational assistance, Susan Ou for monoclonal antibody production, and Elizabeth Nelson for help with antibody characterization.

Author Contributions

Conceived and designed the experiments: YEW DCC. Performed the experiments: YEW. Analyzed the data: YEW GKM. Wrote the manuscript: YEW GKM BJW DCC.

References

1. Bogenhagen DF, Rousseau D, Burke S (2008) The layered structure of human mitochondrial DNA nucleoids. J Biol Chem 283(6): 3665-3675. PubMed: 18063578.
2. Bogenhagen D, Clayton DA (1974) The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. J Biol Chem 249(24): 7991-7995. PubMed: 4473454.
3. Satoh M, Kuroiwa T (1991) Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res 196(1): 137-140. doi:10.1016/0014-4827(91)90467-9. PubMed: 1715276.
4. Suomalainen A, Isohanni P (2010) Mitochondrial DNA depletion syndromes - many genes, common mechanisms. Neuromuscul Disord 20(7): 429-437. doi:10.1016/j.nmd.2010.03.017. PubMed: 20444604.
5. Stumpf JD, Saneto RP, Copeland RC (2013) Clinical and Molecular Features of POLG-Related Mitochondrial Disease. Cold Spring Harb Perspect Biol 4(5): a011395. PubMed: 23545419.
6. Sugiyama S, Hattori K, Hayakawa M, Ozawa T (1991) Quantitative analysis of age-associated accumulation of mitochondrial DNA with deletion in human age-associated accumulation of mitochondrial DNA with deletion in human hearts. Biochem Biophys Res Commun 180(2): 894-899. doi:10.1016/S0006-291X(05)81149-0. PubMed: 1953759.
7. Ide T, Tsutsui H, Hayashidani S, Kang D, Suematsu N et al. (2001) Mitochondrial DNA damage and dysfunction associated with oxidative stress in failing hearts after myocardial infarction. Circ Res 88(5): 529-535. doi:10.1161/01.RES.88.5.529. PubMed: 11249877.
8. Karamanlidis G, Nascimben L, Couper GS, Shekar PS, del Monte F et al. (2010) Defective DNA replication impairs mitochondrial biogenesis in human failing hearts. Circ Res 106(9): 1541-1548. doi:10.1161/CIRCRESAHA.109.212753. PubMed: 20339121.
9. Karamanlidis G, Bautista-Hernandez FV, Fynn-Thompson F, Del Nido P, Tian R (2011) Impaired mitochondrial biogenesis precedes heart failure in right ventricular hypertrophy in congenital heart disease. Circ Heart Fail 4(6): 707-713. doi:10.1161/CIRCHEARTFAILURE.111.961474. PubMed: 21840936.
10. Maassen JA, 'T Hart LM, Van Essen E, Heine RJ, Nijpels G et al. (2004) Mitochondrial diabetes: molecular mechanisms and clinical presentation. Diabetes 53 Suppl 1: S103-S109. doi:10.2337/diabetes.53.2007.S103. PubMed: 14749274.
11. Simmons RA, Suponitsky-Kroyter I, Selak MA (2005) Progressive accumulation of mitochondrial DNA mutations and decline in mitochondrial function lead to beta-cell failure. J Biol Chem 280(31): 28785-28791. doi:10.1074/jbc.M505695200. PubMed: 15946949.
12. Gauthier BR, Wiederkehr A, Baquié M, Dai C, Powers AC et al. (2009) PDX1 deficiency causes mitochondrial dysfunction and defective insulin secretion through TFAM suppression. Cell Metab 10(2): 110-118. doi:10.1016/j.cmet.2009.07.002. PubMed: 19656489.
13. Coskun PE, Beal MF, Wallace DC (2004) Alzheimer’s brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. Proc Natl Acad Sci USA 101(29): 10726-10731. doi:10.1073/pnas.0403649101. PubMed: 15247418.
14. Coskun P, Wyrembak J, Schriner SE, Chen HW, Marciniack C et al. (2012) A mitochondrial etiology of Alzheimer and Parkinson disease. Biochim Biophys Acta 1820(5): 553-564. doi:10.1016/j.bbagen.2011.08.008. PubMed: 21871538.
15. Corral-Debrinski M, Shoffner JM, Lott MT, Wallace DC (1992) Association of mitochondrial DNA damage with aging and coronary atherosclerotic heart disease. Mutat Res 275(3-6): 169-180. doi:10.1016/0921-8734(92)90021-G. PubMed: 1383759.
16. Trifunovic A, Larsson NG (2008) Mitochondrial dysfunction as a cause of ageing. J Intern Med 263(2): 167-178. doi:10.1111/j.1365-2796.2007.01905.x. PubMed: 18226094.
17. Suarez J, Hu Y, Makino A, Fricovsky E, Wang H et al. (2008) Alterations in mitochondrial function and cytosolic calcium induced by hyperglycemia are restored by mitochondrial transcription factor A in cardiomyocytes. Am J Physiol Cell Physiol 295(6): 1561-1568. doi:10.1152/ajpcell.00076.2008. PubMed: 19060297.
18. Hayashi Y, Yoshida M, Yamato M, Ide T, Wu Z et al. (2008) Reverse of age-dependent memory impairment and mitochondrial DNA damage in microglia by an overexpression of human mitochondrial transcription factor a in mice. J Neurosci 28(34): 8624-8634. doi:10.1523/JNEUROSCI.1957-08.2008. PubMed: 18716221.
19. Xu S, Zhong M, Zhang L, Wang Y, Zhou Z et al. (2009) Overexpression of TFAM protects mitochondria against beta-amyloid-induced oxidative damage in SH-SY5Y cells. FEBS J 276(14): 3800-3809. doi:10.1111/j.1742-4658.2009.07094.x. PubMed: 19496804.
20. Keeney PM, Quigley CK, Dunham LD, Papageorge CM, Iyer S et al. (2009) Mitochondrial gene therapy augments mitochondrial physiology in a Parkinson's disease cell model. Hum Gene Ther 20(8): 897-907. doi:10.1089/hum.2009.023. PubMed: 19374590.
21. Piao Y, Kim HG, Oh MS, Pak YK (2012) Overexpression of TFAM, NRF-1 and myr-AKT protects the MPP(+)-induced mitochondrial dysfunctions in neuronal cells. Biochim Biophys Acta:1820(5): 577-585. doi:10.1016/j.bbagen.2011.08.007. PubMed: 21856379.
22. Fisher RP, Clayton DA (1988) Purification and characterization of human mitochondrial transcription factor 1. Mol Cell Biol 8(8): 3496-3509. PubMed: 3211148.
23. Ngo HB, Kaiser JT, Chan DC (2011) The mitochondrial transcription and packaging factor TFAM imposes a U-turn on mitochondrial DNA. Nat Struct Mol Biol 18(11): 1290-1296. doi:10.1038/nsmb.2159. PubMed: 22037171.
24. Chang DD, Clayton DA (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. Cell 36(3): 635-643. doi:10.1016/0092-8674(84)90343-X. PubMed: 6697390.
25. Chang DD, Clayton DA (1985) Priming of human mitochondrial DNA replication occurs at the light-strand promoter. Proc Natl Acad Sci USA 82(2): 351-355. doi:10.1073/pnas.82.2.351. PubMed: 2982153.
26. Larsson NG, Wang J, Wilhelmsson H, Oldfors A, Rustin P et al. (1998) Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. Nat Genet 18(3): 231-236. doi:10.1038/ng0398-231. PubMed: 9500544.
27. Fisher RP, Parisi MA, Clayton DA (1989) Flexible recognition of rapidly evolving promoter sequences by mitochondrial transcription factor 1. Genes Dev 3(12b):2202-17
28. Alam TI, Kanki T, Muta T, Ukaji K, Abe Y et al. (2003) Human mitochondrial DNA is packaged with TFAM. Nucleic Acids Res 31(6): 1640-1645. doi:10.1093/nar/gkg251. PubMed: 12626705.
29. Ekstrand MI, Falkenberg M, Rantanen A, Park CB, Gaspari M et al. (2004) Mitochondrial transcription factor A regulates mtDNA copy number in mammals. Hum Mol Genet 13(9): 935-944. doi:10.1093/hmg/ddh109. PubMed: 15016765.
30. Kaufman BA, Durisic N, Mativetsky JM, Costantino S, Hancock MA et al. (2007) The mitochondrial transcription factor TFAM coordinates the assembly of multiple DNA molecules into nucleoid-like structures. Mol Biol Cell 18(9): 3225-3236. doi:10.1091/mbc.E07-05-0404. PubMed: 17581862.
31. Fisher RP, Lisowsky T, Parisi MA, Clayton DA (1992) DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. J Biol Che 267(5): 3358-3367. PubMed: 1737790.
32. Ghivizzani SC, Madsen CS, Nelen MR, Ammini CV, Hauswirth WW (1994) In organello footprint analysis of human mitochondrial DNA: human mitochondrial transcription factor A interactions at the origin of replication. Mol Cell Biol 14(12): 7717-7730. PubMed: 7969115.
33. Kanki T, Ohgaki K, Gaspari M, Gustafsson CM, Fukuoh A et al. (2004) Architectural role of mitochondrial transcription factor A in maintenance of human mitochondrial DNA. Mol Cell Biol 24(22): 9823-9834. doi:10.1128/MCB.24.22.9823-9834.2004. PubMed: 15509786.
34. Ohgaki K, Kanki T, Fukuoh A, Kurisaki H, Aoki Y et al. (2007) The C-terminal tail of mitochondrial transcription factor a markedly strengthens its general binding to DNA. J Biochem 141(2): 201-211. PubMed: 17167045.
35. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F et al. (2011) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22(9): 1813-1831. PubMed: 22955991.
36. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359. doi:10.1038/nbt.1508. PubMed: 19029915.
37. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6(11 Suppl): S22-S32. doi:10.1038/nmeth.1371. PubMed: 19844228.
38. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36(16): e105. doi:10.1093/nar/gkn425. PubMed: 18660515.
39. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ et al. (2013) Characterizing and measuring bias in sequence data. Genome Biol 14(5): R51. doi:10.1186/gb-2013-14-5-r51. PubMed: 23718773.
40. Watanabe A, Arai M, Koitabashi N, Niwano K, Ohyama Y et al. (2011) Mitochondrial transcription factors TFAM and TFB2M regulate Serca2 gene transcription. Cardiovasc Res 90(1): 57-67. doi:10.1093/cvr/cvq374. PubMed: 21113058.
41. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9(9): R137. doi:10.1186/gb-2008-9-9-r137. PubMed: 18798982.

42. Gadaleta G, D'Elia D, Capaccio L, Saccone C, Pepe G (1996) Isolation of a 25-kDa protein binding to a curved DNA upstream the origin of the L strand replication in the rat mitochondrial genome. J Biol Chem 271(23): 13537-13541. doi:10.1074/jbc.271.23.13537. PubMed: 8662779.
43. Cingolani G, Capaccio L, D'Elia D, Gadaleta G (1997) In organelle footprinting analysis of rat mitochondrial DNA: protein interaction upstream of the Ori-L. Biochem Biophys Res Commun 231(3): 856-860. doi:10.1006/bbrc.1997.6203. PubMed: 9070910.
44. Pierro P, Capaccio L, Gadaleta G (1999) The 25 kDa protein recognizing the rat curved region upstream of the origin of the L-strand replication is the rat homologue of the human mitochondrial transcription factor A. FEBS Lett 457(3): 307-310. doi:10.1016/S0014-5793(99)01055-8. PubMed: 10471798.
45. Dong X, Ghoshal K, Majumder S, Yadav SP, Jacob ST (2002) Mitochondrial transcription factor A and its downstream targets are up-regulated in a rat hepatoma. J Biol Chem 277(45): 43309-43318. doi:10.1074/jbc.M206958200. PubMed: 12198131.
46. Larsson NG, Garman JD, Oldfors A, Barsh GS, Clayton DA (1996) A single mouse gene encodes the mitochondrial transcription factor A and a testis-specific nuclear HMG-box protein. Nat Genet 13(3): 296-302. doi:10.1038/ng0796-296. PubMed: 8673128.
47. Kucej M, Kucejova B, Subramanian R, Chen XJ, Butow RA (2008) Mitochondrial nucleoids undergo remodeling in response to metabolic cues. J Cell Sci 121(11): 1861-1868. doi:10.1242/jcs.028605. PubMed: 18477605.
48. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830): 1497-1502. doi:10.1126/science.1141319. PubMed: 17540862.
49. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3): R25. doi:10.1186/gb-2009-10-3-r25. PubMed: 19261174.
50. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9): 1639-1645. doi:10.1101/gr.092759.109. PubMed: 19541911.
51. Schoenherr CJ, Anderson DJ (1995) The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific gene. Science 267(5202): 1360-1363. doi:10.1126/science.7871435. PubMed: 7871435.
52. ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLOS Biol 9(4):e1001046. PubMed: 21526222.

I

Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps

Originally published as:

Mortazavi A*, Pepke S*, Jansen C, Marinov GK, Ernst J, Kellis M, Hardison R, Myers RM, Wold B. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Research* 2013. **23**:2136-2148. doi: 10.1101/gr.158261.113.

Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps

Ali Mortazavi,[1,2,12,13] Shirley Pepke,[3,4,12] Camden Jansen,[1,2] Georgi K. Marinov,[4] Jason Ernst,[5] Manolis Kellis,[6,7] Ross C. Hardison,[8,9] Richard M. Myers,[10] and Barbara J. Wold[4,11,13]

[1]Department of Developmental and Cell Biology, University of California, Irvine, California 92697, USA; [2]Center for Complex Biological Systems, University of California, Irvine, California 92697, USA; [3]Center for Advanced Computing Research, California Institute of Technology, Pasadena, California 91125, USA; [4]Division of Biology, California Institute of Technology, Pasadena, California 91125, USA; [5]Department of Biological Chemistry, University of California, Los Angeles, California 90095, USA; [6]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA; [7]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [8]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [9]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [10]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [11]Beckman Institute, California Institute of Technology, Pasadena, California 91125, USA

We tested whether self-organizing maps (SOMs) could be used to effectively integrate, visualize, and mine diverse genomics data types, including complex chromatin signatures. A fine-grained SOM was trained on 72 ChIP-seq histone modifications and DNase-seq data sets from six biologically diverse cell lines studied by The ENCODE Project Consortium. We mined the resulting SOM to identify chromatin signatures related to sequence-specific transcription factor occupancy, sequence motif enrichment, and biological functions. To highlight clusters enriched for specific functions such as transcriptional promoters or enhancers, we overlaid onto the map additional data sets not used during training, such as ChIP-seq, RNA-seq, CAGE, and information on *cis*-acting regulatory modules from the literature. We used the SOM to parse known transcriptional enhancers according to the cell-type-specific chromatin signature, and we further corroborated this pattern on the map by EP300 (also known as p300) occupancy. New candidate cell-type-specific enhancers were identified for multiple ENCODE cell types in this way, along with new candidates for ubiquitous enhancer activity. An interactive web interface was developed to allow users to visualize and custom-mine the ENCODE SOM. We conclude that large SOMs trained on chromatin data from multiple cell types provide a powerful way to identify complex relationships in genomic data at user-selected levels of granularity.

[Supplemental material is available for this article.]

Sequence-based functional genomics assays are generating vast amounts of data that map the occupancy of specific transcription factors, the chemical status (such as acetylation and methylation), and positions of chromatin components such as core histones, the loading of RNA polymerases, and domains of DNase I hypersensitivity across the human genome at high resolution (Barski et al. 2007; Johnson et al. 2007; Mortazavi et al. 2008; Hesselberth et al. 2009; for review, see Pepke et al. 2009). Such measurements are now being made for a myriad of cell types, states, and tissues by individual laboratories and by large consortia such as ENCODE and the Epigenome Roadmap (Bernstein et al. 2010; The ENCODE Project Consortium 2012). This wealth of data contains rich, complex, combinatoric information about the inputs and outputs of gene regulatory networks (GRNs) that define each cell type and state. However, it is not yet easy to extract and distill biologically meaningful relationships, especially not on the multiple scales that range from broad global relationships to fine-grained ones that affect small groups of similarly behaving genes or subgenic regulatory elements.

Numerous prior studies have focused on understanding the relationship between an increasingly complex histone modification "code" and the activity state of DNA elements, such as transcriptional enhancers, insulators, promoters, and more or less vigorously transcribed regions for a given cell type or tissue (for review, see Hon et al. 2009). Furthermore, apparent cross talk between context-dependent histone modifications suggests a complex grammar (for review, see Lee et al. 2010). Pioneering analyses focused on specific ad hoc combinations of modifications found in the proximity of transcription start sites (TSS) or in selected distal intergenic regions (Barski et al. 2007; Wang et al. 2008).

More recent approaches have been more general and agnostic, dividing the entire genome systematically, either at regular intervals or based on the data (i.e., "segmenting" the genome) and then classifying the resulting genome segments (regions) into five to 100 states of chromatin mark combinations (classes) by applying statistical or machine learning methods such as Hidden Markov

[12]These authors contributed equally to this work.
[13]Corresponding authors
E-mail woldb@caltech.edu
E-mail ali.mortazavi@uci.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.158261.113. Freely available online through the *Genome Research* Open Access option.

Models (HMMs) or Dynamic Bayesian Networks (e.g., Ernst and Kellis 2010; Hoffman et al. 2012). The resulting machine-derived "states" are then semi-manually annotated to relate them to functions such as gene activation or repression. However, it is not clear a priori if the limited numbers of states used in these analyses, partly for ease of interpretation, fully or optimally capture the biological richness in the data, especially for the much larger and more diverse collections of data sets now being generated by projects such as the ENCODE and NIH Roadmap Epigenomics Projects.

The self-organizing map (SOM) is an unsupervised machine-learning method that was developed to cluster and visualize high-dimensional data (for review, see Kohonen 2001). It projects high-dimensional data onto a two-dimensional map composed of many units, each of which can be regarded as a mini-cluster, defined by its associated prototype vector of component weights. SOMs capture similarity relationships present in the training data as map topology, such that individual neighboring hex-units can subsequently be clustered after training into "metaclusters" as appropriate. This is analogous to the way biologists typically interact with RNA expression patterns and subpatterns in a classic two-way hierarchical clustering (Eisen et al. 1998). Indeed, SOMs with modest map sizes of less than 100 units have been used for more than a decade for clustering gene expression data (Golub et al. 1999; Milone et al. 2010; Newman and Cooper 2010; Spencer et al. 2011) or modest numbers of other genomic data sets (Moorman et al. 2006; Suzuki et al. 2011). While SOMs with small map sizes produce results that are generally equivalent to K-means, SOMs with thousands of units on boundaryless maps can show emergent behavior (Ultsch 1999). We reasoned that large SOMs should be able to capture a greater variety of combined chromatin mark patterns compared with methods that find a relatively small number of chromatin states, and that the resulting organization could be more readily visualized and ultimately mined in an intuitive way. Specifically, we anticipated that a large SOM, constructed from multiple genome-wide data types, collected across biologically distinct ENCODE cell types, would begin to reveal patterns of active, cell-type-specific transcriptional control elements based on their associated chromatin signatures.

As a first test of these possibilities, the trained ENCODE chromatin SOM presented here displayed distinct spatial organization that reveals how combinations of histone marks, DNase I hypersensitivity, and RNA polymerase occupancy correlate with gene features and activity, such as a relatively large supercluster of transcription start sites (TSS) that are active in one or more cell types, or a cluster of genes repressed in another cell type or types. We show how additional ChIP-seq, RNA-seq, transcription factor binding motifs, and other functional data can be placed on the chromatin map to identify and interpret cell-type-specific regulatory elements and transcription start sites. We then hierarchically cluster the SOM hex-units to explore global relationships of the different data sets on the SOM. Gene Ontology (GO) analysis reveals distinct enrichments in individual, often neighboring, units on the map related to cell-type-specific gene regulation. Finally, we introduce an interactive web interface to facilitate further mining of the ENCODE SOM and apply it to the analysis of cell-type-specific EP300 (also known as p300)–enriched units.

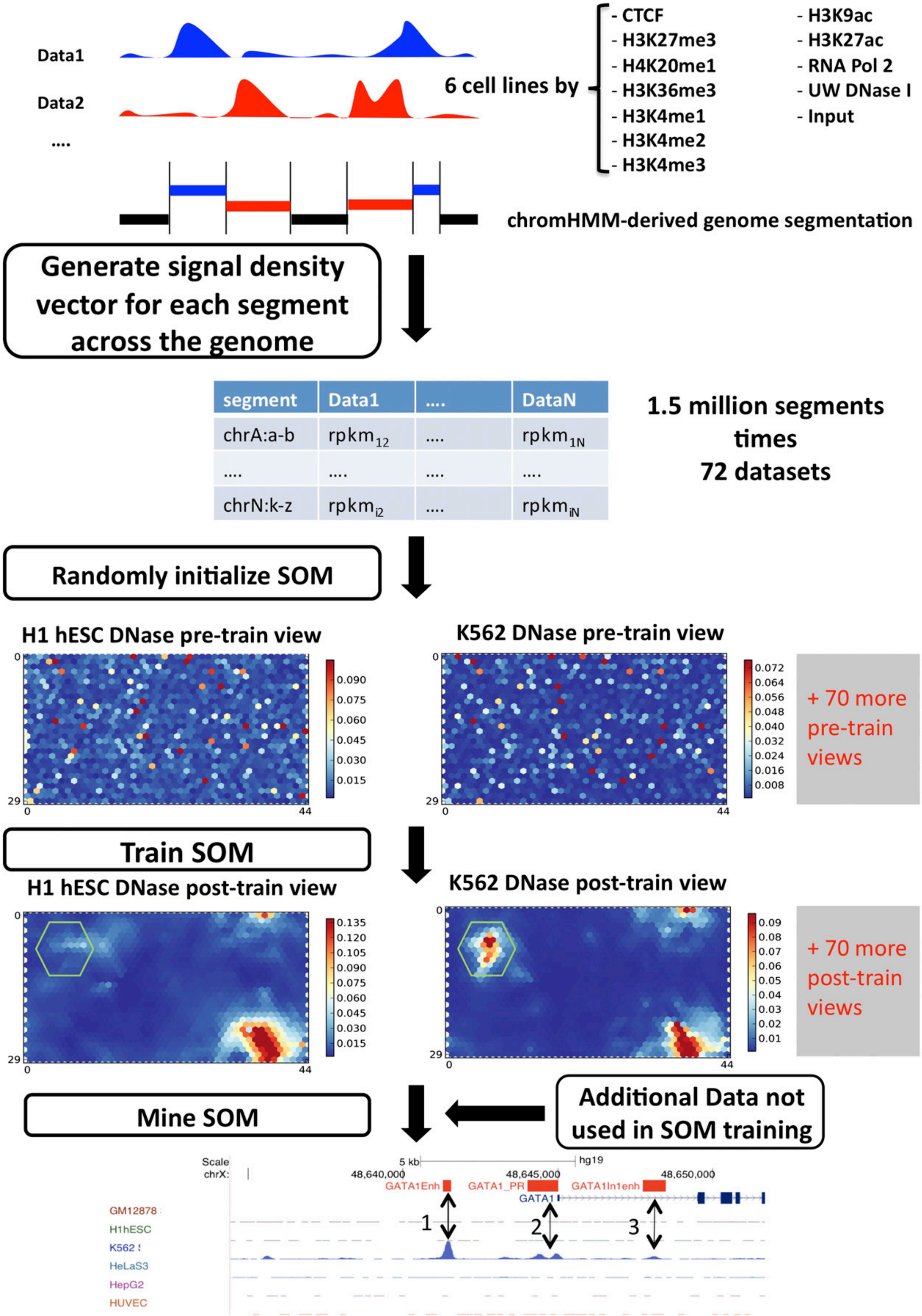
Results

Chromatin SOM construction and overall organization

The workflow for building a chromatin-based SOM begins with primary data mapping and genome segmentation and ends with visualization and data mining (Fig. 1). Briefly, the first step is to computationally break the genome into "segments" based on the data. The goal of segmentation is to define, across the entire genome, DNA segments that share the presence and absence of marks in the input data. To coordinate our results with other ENCODE Project Consortium work (The ENCODE Project Consortium 2012), we used a specific genome segmentation generated on 84 preselected data sets of eight histone modifications, RNA polymerase II, and CTCF from ChIP-seq, ChIP input control, and three open chromatin assays across six cell types using a "stacked" segmentation generated with ChromHMM (Ernst and Kellis 2010). We then constructed a training matrix consisting of the signal density for 72 of these data sets for each of the 1.5 million individual genome segments using only one of the DNase-seq assays to represent open chromatin. The Methods and Supplemental Figure S1 describe how the stacked segmentation differs from other segmentations of the same data.

We used the resulting matrix of 1.5 million 72-dimensional data vectors to train a SOM with map size of 30 rows of 45 columns (1350 units), and selected the best out of 10 maps based on the lowest quantization error (Methods) (Supplemental Fig. S2). The size of the map was selected to allow us to recover at least a thousand distinct states, if they were present in the data. In a uniformly distributed untrained map, we would expect 1170 segments/unit and 2.2 Mb/unit, on average. This map is a toroid, meaning that the top units on the map are seamlessly connected to the bottom units, and that the same applies to the leftmost and rightmost units (Supplemental Fig. S3). We chose the toroid form because it has no boundaries, which should prevent it from compressing clusters into map corners. To display a toroid map in two dimensions, we "slice it open," and some clusters are therefore visually split; that is, they "wrap around" the top edge to the bottom and from the left edge to the right, as indicated by the arrows (Supplemental Fig. S3). All assignments of segments to SOM hex-units are available for this SOM as a single bed file (Supplemental Table S1).

The distribution of DNA segments and nucleotides on the untrained map was without pattern and relatively even, while the trained map was much more uneven (Figs. 1, 2). This is expected because the segments on the trained map have been organized into clusters that contain differing segment numbers and nucleotide densities. For example, many of the larger DNA segments had little to no signal for any data set, and they were sequestered into a relatively small fraction of the SOM; on this 30-by-45 map, 48 contiguous units (3.5% of all units) captured 38% of the entire genome sequence, and is shown as high nucleotide density and segment count in Figure 2, A and B. The remainder of this map is dedicated to more finely parsing segments that have some signal in at least one of the training data sets. These overall organizational properties were not specific to this particular instance of the SOM nor to the ENCODE chromatin data. The top-scoring ENCODE SOM was very similar to the next nine best-scoring SOMs, each trained independently on the same input data, but from different random initializations. Specifically, we found that, for all of the units and regions of the SOM discussed below, segments within the same unit were clustered on the other nine maps within the same unit or adjoining units >80% of the time (Fig. 2C). We further analyzed the effect of leaving individual data sets out by retraining SOMs with 72 combinations of 71 data sets each and repeating the reproducibility analysis. We found that map reproducibility was robust to the removal of any one of 29 data sets (listed in Supplemental Table S2). While no single group of data sets was completely re-


Data1
Data2
....
6 cell lines by
- CTCF
- H3K27me3
- H4K20me1
- H3K36me3
- H3K4me1
- H3K4me2
- H3K4me3
- H3K9ac
- H3K27ac
- RNA Pol 2
- UW DNase I
- Input
chromHMM-derived genome segmentation
Generate signal density vector for each segment across the genome
1.5 million segments times 72 datasets
Randomly initialize SOM
H1 hESC DNase pre-train view
K562 DNase pre-train view
+ 70 more pre-train views
Train SOM
H1 hESC DNase post-train view
K562 DNase post-train view
+ 70 more post-train views
Mine SOM
Additional Data not used in SOM training
Scale
chrX:
5 kb
hg19
48,640,000
48,645,000
48,650,000
GATA1Enh
GATA1_PR
GATA1In1enh
GATA1
1
2
3
GM12878
H1hESC
K562
HeLaS3
HepG2
HUVEC

segment	Data1	....	DataN
chrA:a-b	rpkm_{12}	....	rpkm_{1N}
....	....	....	....
chrN:k-z	rpkm_{i2}	....	rpkm_{iN}

Figure 1. Training the self-organizing map and general overview of data analysis. The genome is first segmented based on the signal density of input data sets. Any segmentation approach can be applied; in this case, the ChromHMM-derived segmentation in the primary publications by The ENCODE Project Consortium was used. The signal density is calculated for each segment and each data set, resulting in an input matrix of $M \times N$ dimensions, where M is the number of segments and N the number of data sets. The SOM is then initialized randomly from the input matrix, and trained. Additional data sets, not used for training, can then be mapped to the SOM, and these mappings and the distribution of segments on the trained SOM can be mined for interesting biological relationships.

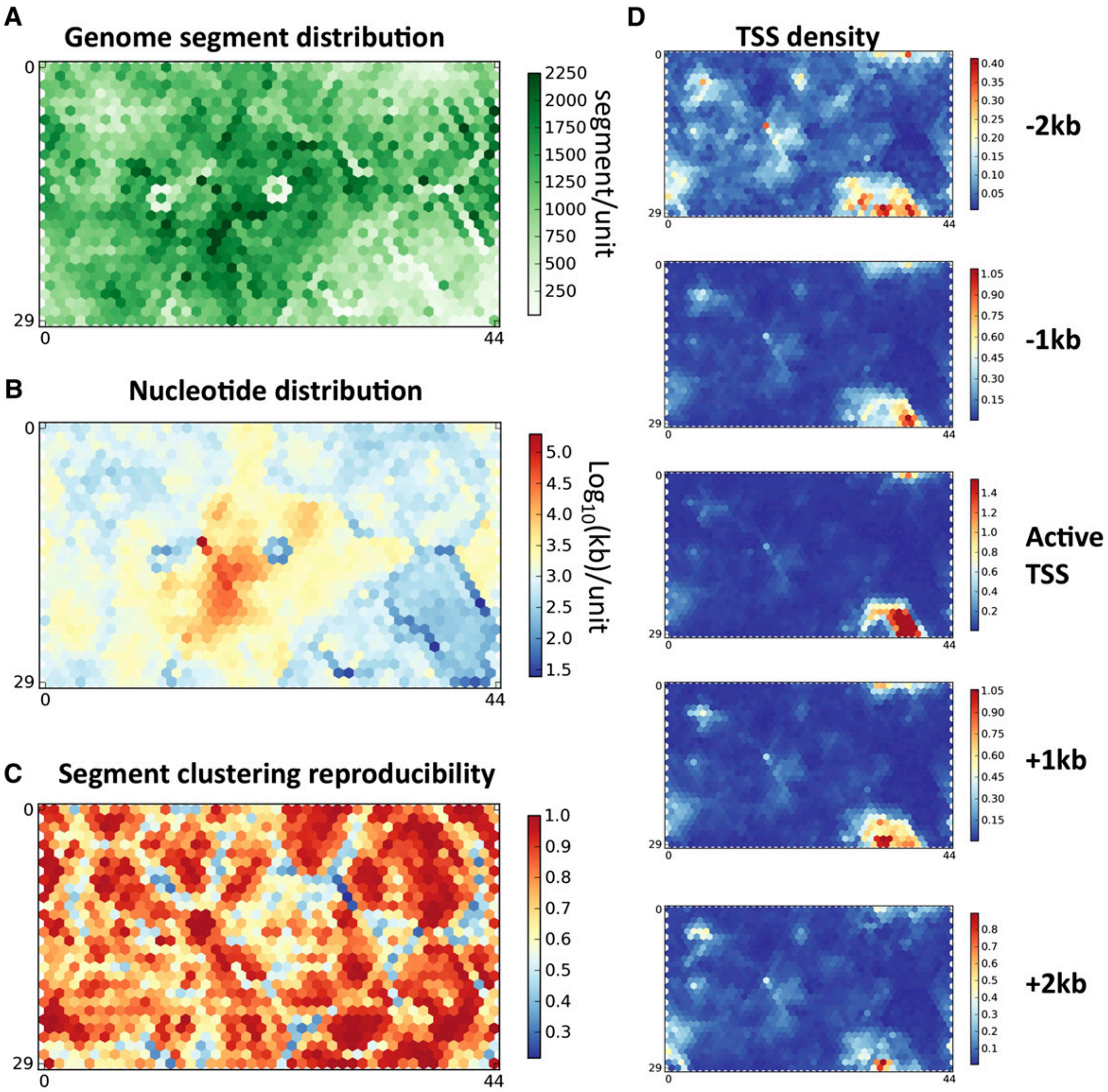


Figure 2. Map organization. (*A*) The segment count distribution over the map is uneven. While the average number of segments per unit is 1170, individual units range from 30 to 9334 segments. Note the distinct 1-unit-wide boundaries that contain very few segments separating denser regions. (*B*) The nucleotide distribution reflects the segment count, with the units with the most segments also containing the most nucleotides. These segments are also larger, thus accounting for the large portion of the genome that has little to no signal. (*C*) Reproducibility of clustering of two segments in the same unit or adjoining units as described in the text. (*D*) TSS-centric organization of active proximal promoters. The unit densities of points −2 kb, −1 kb, 0 bp, +1 kb, and +2 kb of GENCODE 7 TSS show the distinct organization of active promoters driven primarily by a common set of genes expressed in more than one cell type.

dundant, we found that three groups of data sets (H3K9ac, H3K36me3, and Control) were redundant in four out of six cell types, whereas another group of data sets (RNA Pol II, DNase I, and H3K4me3) was redundant in only one of six cell lines. Interestingly, the removal of these apparently redundant data sets still affected the reproducibility of a distinct subset of units, suggesting that they still contributed to the organization of the SOM in restricted regions of the map. These results argue that our SOM is robust and stable, and that segments with similar signatures are stably located near each other on the map, even though such segments do not always fall into a single hex-unit on independently trained SOMs. Local differences of the latter kind are expected for a nondeterministic method and can be discriminated from major differences, as shown below.

The SOM displayed several distinctive, very-low-segment-count "boundaries," usually just one unit wide and with as few as 30 segments/unit (Fig. 2A,B). These are, in effect, boundary units that separate clusters located on either side and that are characterized by distinct mark profiles. For example, H3K4me3-enriched segments are segregated from CTCF-associated ones in an adjacent map region (Supplemental Fig. S4).

We next explored where transcription start sites (TSS) map on the ENCODE SOM. No explicit information on annotated TSSs was used in building this SOM. Our expectations were that active TSSs would share a set of features present in the training data, including high DNase I hypersensitivity, RNA polymerase II occupancy (in varying intensities), H3K9ac, and H3K4me3 marks. This predicts that active TSSs would generally cluster together some-where on the SOM. In contrast, inactive TSSs were expected to lack these marks and, additionally, they might or might not show a repressive mark signature. We therefore expected inactive TSSs to occur elsewhere on the map, sequestered into one or a few clusters, depending on whether they have no other data from the training set or contain repressive mark data. A further expectation was that the SOM would detect and subcluster segments according to the intensity of their active-TSS signatures, since we had not reduced the data to simple present–absent calls for signal, but had retained all the quantitative information in the primary data. Finally, we expected that the SOM would subcluster active TSSs according to the cell type or combinations of types in which they were active.

All of the above expectations were met. A prominent region of the map, having relatively low segment and nucleotide density, showed the highest fractional enrichment in the number of GENCODE 7 (Harrow et al. 2012) TSS, with 27 units passing a threshold of 0.8 TSS/segment (Fig. 2D). Note that each TSS in this analysis was mapped as a single nucleotide, and was therefore assigned to only one DNA segment, even if there were several neighboring segments with very similar histone mark data. For this reason, we do not expect every DNA segment with an active TSS histone mark signature to score positive in this tally. As expected, the prominent TSS domain in the lower-right quadrant of the SOM corresponded with a domain of maximal DNase I hypersensitivity, as illustrated by comparing this with H1-hESC DNase-seq data (cf. Fig. 1 DNase I panels with Fig. 2D).

We next asked how DNA sequences located at varying distances from the nearest active TSS are organized on the map and found that 35 units are enriched in segments within 2 kb of these TSSs. We expected that near an active TSS, the chromatin signature would be very similar to the TSS point nucleotide for many segments, but that some segments would now display "mixed" chromatin signatures that retain some qualities of a pure TSS and add some characteristics of nearby chromatin. Such a "neighborhood" effect reflects properties of the original ChromHMM segmentation process as well as the biology of the histone mark pattern in each input cell type. As the distance from the TSS increases into the gene body or into the upstream promoter region, the histone signatures changed. On average, the distinct enrichments of single nucleotides that are located at −2 kb, −1 kb, +1 kb, and +2 kb from the TSSs in neighboring units demonstrates that the map has spatially clustered active promoters and their immediate upstream and downstream regions (Fig. 2D).

The prototype vectors for the units in the active-TSSs region revealed that most DNA segments at the center of this region possess signatures of expression in more than one cell type, although some adjacent clusters are cell-type-specific. When examined for RNA expression pattern and GO terms, the shared ones were housekeeping and other genes common to the cell types in this study, as expected. Investigating even more closely, we observed that individual units parse the levels of associated chromatin marks (e.g., high vs. medium vs. low H3K4me3) and the magnitude of the RNA polymerase signal, in different data sets and cell types. As discussed below, a user can drill even further down to select and extract DNA segments from hex-units with particular signature characteristics by using the SOM viewer and its associated DNA segment database.

Inspection of the SOM also reveals that multiple histone modification marks, previously shown to be associated with active transcription or active repression, drove the organization of the majority of the map (e.g., H3K4 mono-, di-, and tri-methylation, and H3K27me3 for activation and repression, respectively). This emphasis was expected, as several histone marks associated with active transcription tend to produce strong ChIP signals that are localized over relatively short DNA regions. The information-rich map regions typically show distinctive quantitative and qualitative combinations of marks. Most component planes, such as the ones shown for RNA polymerase II or H3K4me3 occupancy in the cell line GM12878 (Supplemental Fig. S4), form a single, internally connected cluster for their respective signal densities on the toroid. However, several other marks such as H3K4me2 and H3K27me3 have more than one distinct cluster on the map. This pattern suggests that they are found together with at least one other different additional chromatin profiles(s), or that regions rich in these marks are distinctive for individual cell types, or both (all component weights are displayed in Supplemental Figs. S5–S10). We return to dissecting the more complex patterns below.

Interactive SOM viewer for visualization and mining

We created an interactive JavaScript web-based SOM viewer with an associated map segment database to facilitate these explorations (http://woldlab.caltech.edu/ENCODESOM). It allows users to visualize and compare units on the map with respect to any input data set or to additional data types (see below), to find properties of different regions of the map, such as Gene Ontology enrichments, and to mine the segments in a given hex-unit or cluster. The interface for version 1.0 consists of five tabs: Training Data, TSS, GO, Other Data, and Clusters, which correspond to the results in this manuscript. A tool for highlighting groups of hex-units in one view and then seeing that outline on any subsequent view aids in evaluating the relatedness of one distribution (RNA polymerase II, for example) with another (TSS annotation or CAGE tags). Users can click on individual units and find the associated segments, genes, and GO-enriched genes. They can also select their own set of units and flag them across the different views of the data. This allows users, for example, to highlight a cluster of interest in the Cluster tab and see the clustering reproducibility of those highlighted units in the Other tab.

By using the viewer to ask how data from the input data sets are clustered and how those clusters relate to each other, one immediately sees the overlaps of units high in DNase I hypersensitivity, H3K9ac, H3K27ac, H3K4me2, and H3K4me3. Had we not known prior to this study that these chromatin signatures are affiliated with active promoters, the SOM would have allowed us to readily discover these relationships. Even knowing these general relationships, the SOM allows us to mine for fine structure that includes more complicated profiles of cell type specificity.

In contrast, we detected little overall change in H4K20me1 across the cell types and little affiliation of this mark with other signals, which leads segments high in those marks to cluster in a single location (upper-left quadrant of the map, Supplemental Fig. S11). Finally, we saw that the RNA Pol II component plane enrichments showed a gradient of RNA Pol II signal centered on a single unit that has the highest signal, which emphasizes that the SOM is clustering on the presence of the signal and also on its intensity. Units immediately around it have lower RNA Pol II intensity, and a user could then mine these, asking what additional information (possibly other marks and/or cell-type patterns) are distinguishing them from the single peak RNA Pol II unit.

Overlaying other ChIP-seq and functional data to find additional relationships

The SOM can also be used to test predictions, mine associations, and map relationships for data sets that were not used to train the

SOM. We began by exploring evidence for cell-type-specific *cis*-regulatory modules (CRMs) in the erythroid/monocyte lineage (K562) and in embryonic stem cells (H1-hESC) (Fig. 3). The transcription factors GATA2 and SPI1 (also known as PU.1) are important in erythroid differentiation, while POU5F1 (also known as OCT4) and NANOG are critical for defining embryonic stem cells. ENCODE ChIP-seq occupancy data for each factor was mapped onto the SOM (Fig. 3E–J). Occupancy for each factor was concentrated in two cell-type-specific clusters, one in the upper-left quadrant, and the other in the lower right (wrapping around to the top right, due to the continuous structure of the map). We then asked how these clusters relate to each other within each cell type, across cell types, and with underlying histone-mark signatures.

In K562 and H1-hESC cells, the upper-left quadrant of the SOM was prominent for the concentration of histone marks H3K27ac and H3K4me1, which have been affiliated with active enhancers and some promoters in previous studies. When H3K4me1 domains are outlined for K562 and H1-hESC (hexagon and triangle, respectively), prominent cell-type specificity is shown by the fact that they are largely separated (Fig. 3C,D). However, there is also a small domain of overlap, reflecting a few units in which similar chromatin signatures exist in both cell types.

We next asked how SOM domains of enhancer-associated histone marks are related to transcription factor occupancy data. We used well-studied factors that regulate hematopoetic target genes (GATA2 and SPI1) in K562 cells, and factors that regulate pluripotence target genes (NANOG and OCT4) in H1-hESC cells. When we overlaid the H3K4me1 chromatin outlines onto these individual factor ChIP-seq data views (Fig. 3E–H), the factors clearly coclustered with the enhancer histone marks in a cell-type-appropriate manner.

These transcription factors, plus PAX5 and SPI1 in the cell line GM12878 (Supplemental Fig. S12), also display some concentration of ChIP-seq signal in the lower-right portion of the map, where active TSS and their adjacent promoters are concentrated (Fig. 2D) and where H3K4me3, a mark of active and poised promoters, is strongly concentrated (Fig. 3A,B). This active TSS and peri-TSS domain of the SOM had especially prominent signals for SPI1 and NANOG, suggesting that these factors are associated by direct binding at or near promoters, or that they are otherwise physically engaged with promoter/TSS bound proteins (i.e., through protein:protein interactions that are recovered in ChIP). It is notable that there is a much weaker concentration of GATA2 in this SOM region. Taken at face value, this suggests that GATA2 is mainly associated with nonpromoter CRMs rather than with the peri-TSS domains, and that SPI1 has the opposite preference in K562.

Another expectation is that functionally active transcription factor occupancy will be marked with enhancer signatures (H3K4me1, H3K4me2, H3K27AC, and DNase I hypersensitivity). Active transcription factor occupancy is expected to be a subset of all sites of occupancy that should overlap with independently validated *cis*-regulatory modules (CRMs). We therefore asked where known CRMs are located on the SOM by taking advantage of a manually curated set of 118 erythroid CRMs. This set contains both distant enhancers and promoters. The CRMs localized prominently to the enhancer- and TSS-proximate zones of the map in K562 cells (Fig. 3K), with those in the enhancer area showing clear preference for the GATA2-enriched cluster of units (Fig. 3E). As would be predicted, the erythroid CRM map units are also enriched for K562-specific active enhancer histone marks and EP300 occupancy (Fig. 3C,I) that do not overlap with H1-hESC-specific enhancer marks and EP300 (Fig. 3D,J). A single hex-unit containing 979 genomic segments was most prominent for known erythroid CRMs, and we investigated it further (Fig. 3M,N). Remarkably, this single unit contained 11% of all high-confidence EP300 ChIP-seq peaks in the genome for K562 (P-value < 10^{-100}), and these overlapped strongly with segments also occupied by GATA2. The contents of this unit can now be further mined and tested to learn whether features lacking EP300 occupancy nevertheless contain active enhancers.

Functional CRMs are also expected to contain conserved sequence motifs that are targets for direct DNA binding. We used motifs curated from the literature for PAX5 and GATA2, along with closely related ones derived from ChIP-seq data, as defined by The ENCODE Project Consortium (The ENCODE Project Consortium 2012). We used phastCons conservation scores (Siepel et al. 2005) to compile a set of conserved motifs for each factor. We then mapped the locations of conserved instances of these motifs onto the SOM. As many transcription factor motifs in eukaryotes are short, they can occur within conserved domains for reasons other than being part of CRMs (i.e., being located with the coding portion of genes). Other instances of the motif are expected to be conserved on account of functioning in cell types or states other than this one. For these reasons, a dispersed map is expected. Nevertheless, NANOG motifs (Fig. 3L) and GATA motifs exhibited clear clustering, concentrated around the stem-cell-specific and erythroid-specific enhancer clusters of units.

Although we are herein primarily concerned with analyzing the ChromHMM-derived segmentation, we have also tested the behavior of the SOM using a naïve, 200-bp segmentation, as described in the Methods. We found that the map shows anisotropy, with enhancer-like and repressed regions more likely to cocluster, but with significant differences in some of the promoter regions. We conclude that the details of the segmentation do matter to a certain extent and that the particulars of each segmentation will interact differently in a way that depends on the data itself.

Taken together, these observations demonstrate the ability of a multi-cell chromatin SOM to concentrate and reveal cell-type-specific regulatory regions, and to allow users to visualize important patterns and relationships between transcription factor occupancy, candidate binding sites, chromatin signatures, and curated functional elements. Other relationships not shown in this set, but strongly visible in the data, include DNase I hypersensitivity and RNA Pol II occupancy. The ENCODE SOM-viewer allows users to explore these relationships by selecting views and marking the boundaries of one or more areas of interest based on more than 96 data sets.

SOM metaclusters capture regional and global properties of histone mark combinations

In addition to fine-grained unit-level clustering of relatively small numbers of segments into each unit done by the SOM itself, we can further cluster the unit prototype vectors across the entire map into metaclusters. We expect this level of analysis to be useful for further probing global genome-scale organization captured by the structure of the SOM. This clustering emphasizes more complex combinatoric chromatin signatures and thus augments the way we have already observed groups of units that cluster together based on the component plane of one training set (e.g., H3K4me1).

The full phylogenetic ordering of all units (Fig. 4A) is fine-grained, and it can be interpreted by a user visually in much the same manner as a phylogenetic ordering of genes. We also per-

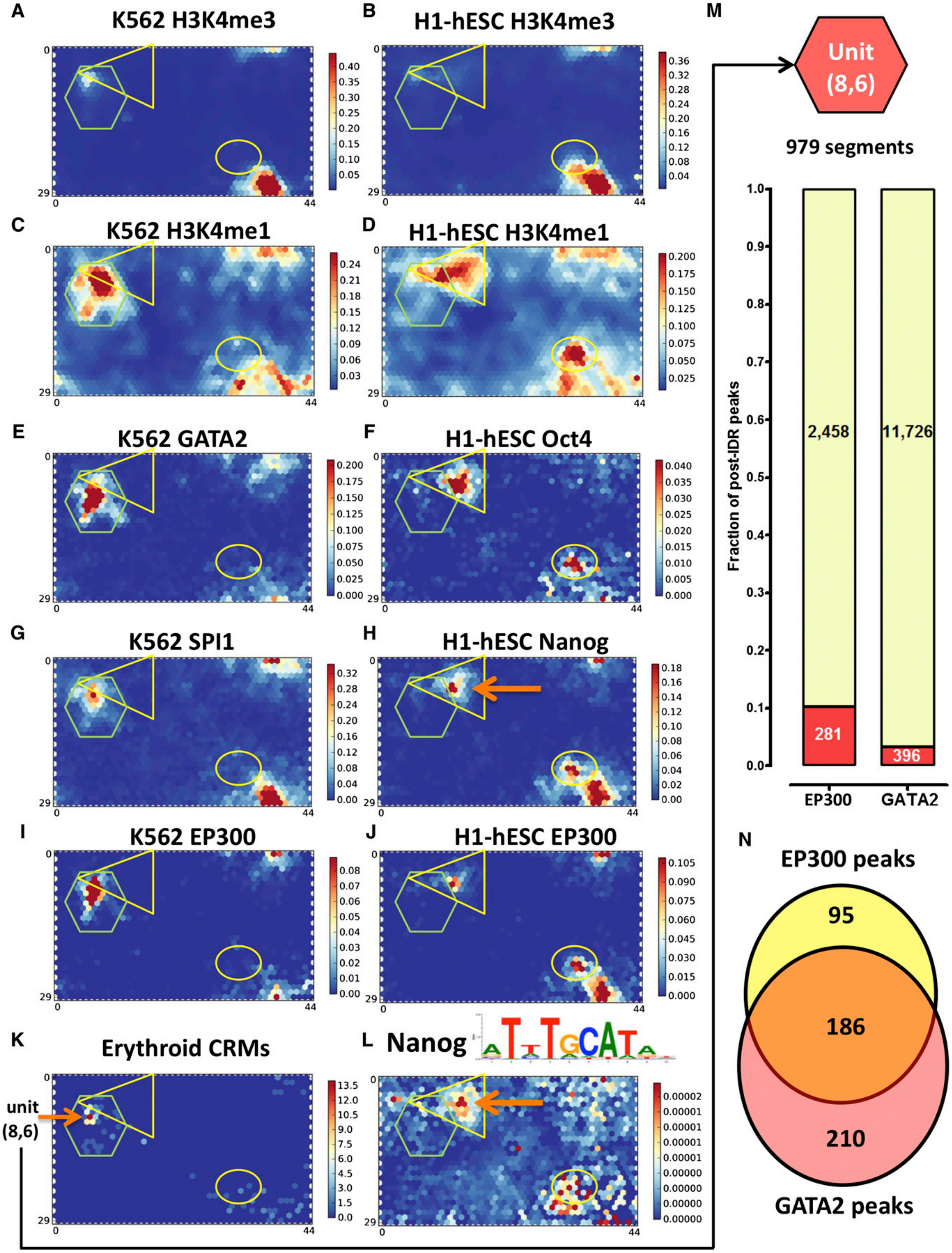


Figure 3. (Legend on next page)

formed an automated clustering to produce a nonsupervised set of boundaries for metaclusters of SOM units that are more similar to each other (based on their unit vector) than they are to other SOM units (see Methods). As with phylogenetic clustering of a single measurement, such as gene expression, we expect the phylogenetic ordering to be composed of graded similarity groups, rather than homogeneous and starkly bounded clusters. This is what we observed when we surveyed a stepped series of similarity thresholds versus metacluster number. The internal data structure identified several natural discontinuities as a function of clustering threshold, and we then selected three of these for full clusterings (Supplemental Fig. S13) to provide users with choices. Prominent driving relationships for the 126 cluster set that we found to be the most useful in our mining are shown in Figure 4B. Finally, we show the specific composition of each cluster for the 126-cluster instance (Supplemental Fig. S14).

The metaclusters showed enrichment patterns that are either cell-type-specific or common across multiple cell types. For example, cluster 1 contains 12 units that have high H3K36me3, RNA Pol II, and H4K20me1 in HUVEC cells (Fig. 4C,D). Different units within cluster 1 differ from each other based on which additional data sets are enriched in that unit. For example, two of the 12 units also show an additional enrichment for H3K36me3 and RNA Pol II in H1-hESC cells. The metaclustering captured features described in earlier sections, such as the active TSS region, and the K562-specific TSS with SPI1 region that corresponds to specific metaclusters, respectively.

Overall, the marks generally associated with active transcription, either at promoters or distant transcriptional enhancers, such as H4K4me1/2/3, H3K9ac, H3K27ac, and DNase I hypersensitivity, clustered in a cell-type-specific manner, whereas H3K36me3 and H4K20me1 clustered together by data type (Fig. 4E). The repressive mark H3K27me3 component planes also clustered together to form an outgroup. The SOM shows that while there is a strong common core of units shared by all six CTCF component planes, they each have more specific enriched units at the periphery. Whether these reflect cell-type-specific CTCF binding or have an alternative explanation such as changed chromatin marks near consistently CTCF-occupied sites is uncertain, and both could be at work. Interestingly, CTCF and RNA Pol II both displayed some clustering by cell type, and some that joined with other active marks from the same cell type.

Some Gene Ontology terms have distinctive chromatin mark signatures

We asked if any Gene Ontology (GO) functional terms are enriched in individual SOM units. Two hundred and twenty-eight GO terms displayed statistically significant enrichment following a Bonferroni correction (P-value < 10^{-10}) at the unit level (Supplemental Table S3). As might be expected, these included enrichments in GO terms that correspond to actively transcribed genes, or to actively repressed genes (for example, neuron-specific genes in non-neuronal cells). Most GO terms (164) were enriched in <1% of the map (13 units or less), and some of these are very specific. For example, "extracellular matrix" is enriched in five neighboring units (Fig. 5), and further inspection suggested that this enrichment is driven by genes that are much more highly expressed in HUVEC than in other cells. The regional GO enrichments typically correlated with metacluster boundaries of the SOM. In the case of "extracellular matrix" (Fig. 5A), four of the five units are part of cluster 1 (Fig. 4C). Another 30 GO terms were enriched in >5% of the map units, and these were typified by broad categories relating to the housekeeping functions of the cell such as "cell cycle." These GO terms are particularly associated with units that are high in H3K36me3 in one or more cell lines. Thirty-four GO terms were enriched in 1%–5% of the map, and these were typically much more specific, developmental terms in units with particular histone mark combinations. The enrichment in specific units for "GTPase activator activity," for example, is driven by gene families that show similar signal profiles across cell lines; the top two hex-units correspond to segments that have a high ratio of H3K4me1 over H3K4me2 in HUVECs that are candidate HUVEC-specific regulatory elements. Similarly, "sequence-specific transcription factor activity" (Fig. 5B) is enriched primarily in units that have cell-type-specific H3K27me3, whether in all cell types or in only some, such as H1-hESC cells and HUVEC. The two units with the most enrichment in Figure 5B have many additional associated developmental GO terms (Fig. 5C) and differ based on the presence of H3K27me3 signal in embryonic stem (ES) cells for segments in both units, but only H3K27me3 signal in HUVEC cells for one unit. This fine parsing by the SOM is nicely illustrated within the HOXD cluster, where the anterior and posterior parts of the cluster are split between these two units (Fig. 5D).

EP300 ChIP-seq overlay and cell-type-specific candidate enhancer segments

We extended our analysis of ENCODE EP300 data sets from K562 by including GM12878, H1-hESC, and HepG2 cells to identify 45 cell-type-specific and common EP300-high units, accounting for 1.4% of the genome and 1.9% of the segments. We found that each cell type had its own specific set of units with high EP300 occupancy, whereas only a few units showed EP300 signal in more than two cell types (Fig. 6). These common EP300 units correspond to the common TSS region, whereas the cell-type-specific clusters are primarily more than 2 kb from the TSSs (Fig. 2D). We showed earlier (Fig. 3) that we found K562 EP300 ChIP-seq signal in

Figure 3. Organization of genomic functional elements on the SOM. A triangle, hexagon, and ellipse are superimposed to allow comparison between maps. (*A*,*B*) H3K4me3 signal density in K562 and H1-hESC. (*C*) The hexagon encompasses the K562 units high in H3K4me1. (*D*) The triangle and hexagon capture the two disjoint regions that are high in H3K4me1 in H1-hESC. (*E*) GATA2 signal, which was not used in the training, is high in a subset of the H3K4me10high units in C. (*F*) Similarly, POU5F1 is primarily found overlapping the H3K4me1 high units. (*G*,*H*) In contrast to GATA2 and POU5F1, SPI1 and NANOG are found primarily in units that are high in H3K4me3 (to the *lower right* of the ellipse) with less signal found at H3K4me1 high units. (*I*,*J*) EP300 signal (also not used in the training) is found either primarily at enhancers in K562, but promoters in H1-hESC. (*K*) More than one-third of known erythroid CRMs cluster into a single unit with coordinates (8, 6). (*L*) Conserved NANOG motifs (motif derived from NANOG ChIP-seq data). ChIP-seq occupancy and motif occurrences were defined by the uniform ENCODE ChIP-seq binding site and motif calling pipelines. Conservation was assessed using the 46-way vertebrate phastCons scores for hg19 downloaded from the UCSC Genome Browser. The scores for each unit in the motif maps were normalized for the total number of base pairs in the unit to avoid the map being dominated by units with very high number of base pairs in them. (*M*) Ten percent of EP300 ChIP-seq calls and 3.2% of GATA2 calls in K562 fall within the top erythroid-CRM enriched unit (8, 6). (*N*) Sixty-six percent of the EP300 peaks in unit (8, 6) overlap a GATA2 peak.

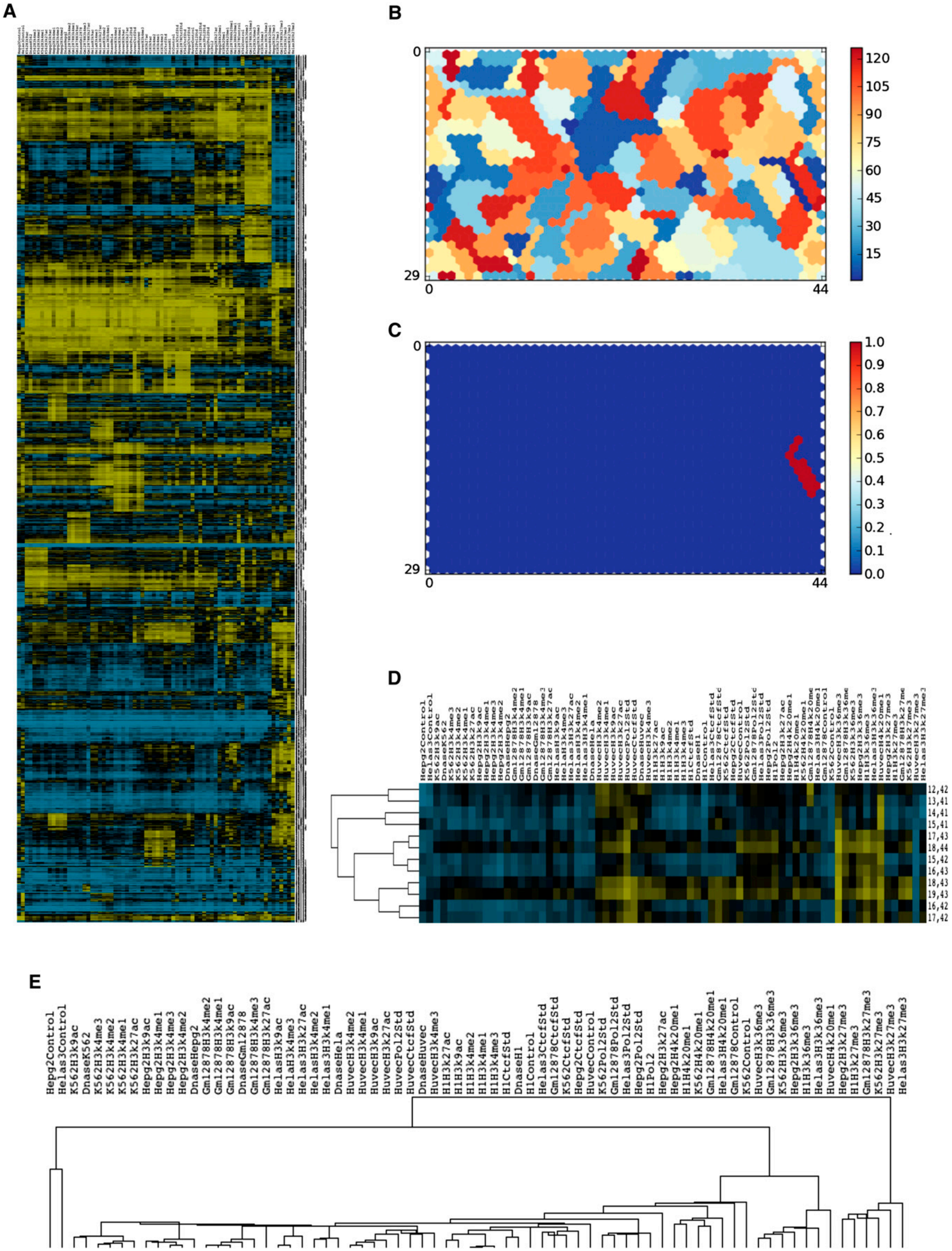


Figure 4. Metaclustering of the SOM. (*A*) Hierarchical clustering of the ranked unit weights (rows) and components (columns) shows both the large-scale and fine structure of the SOM unit ranked weights (yellow, high enrichment rank; blue, low enrichment rank). (*B*) Metaclustering of the SOM into ~120 clusters based on a consistency threshold of 2.6. (*C*) Twelve units make up metacluster 1. (*D*) Ranked component weights of metacluster 1. All 12 units share enrichment in HUVEC RNA Pol II, H3K36me3, and H4K20me1. Individual units show additional distinct enrichments, which distinguish them from one another. (*E*) Clustering of the component columns of Figure 5A, showing the relationships of the data sets to one another.

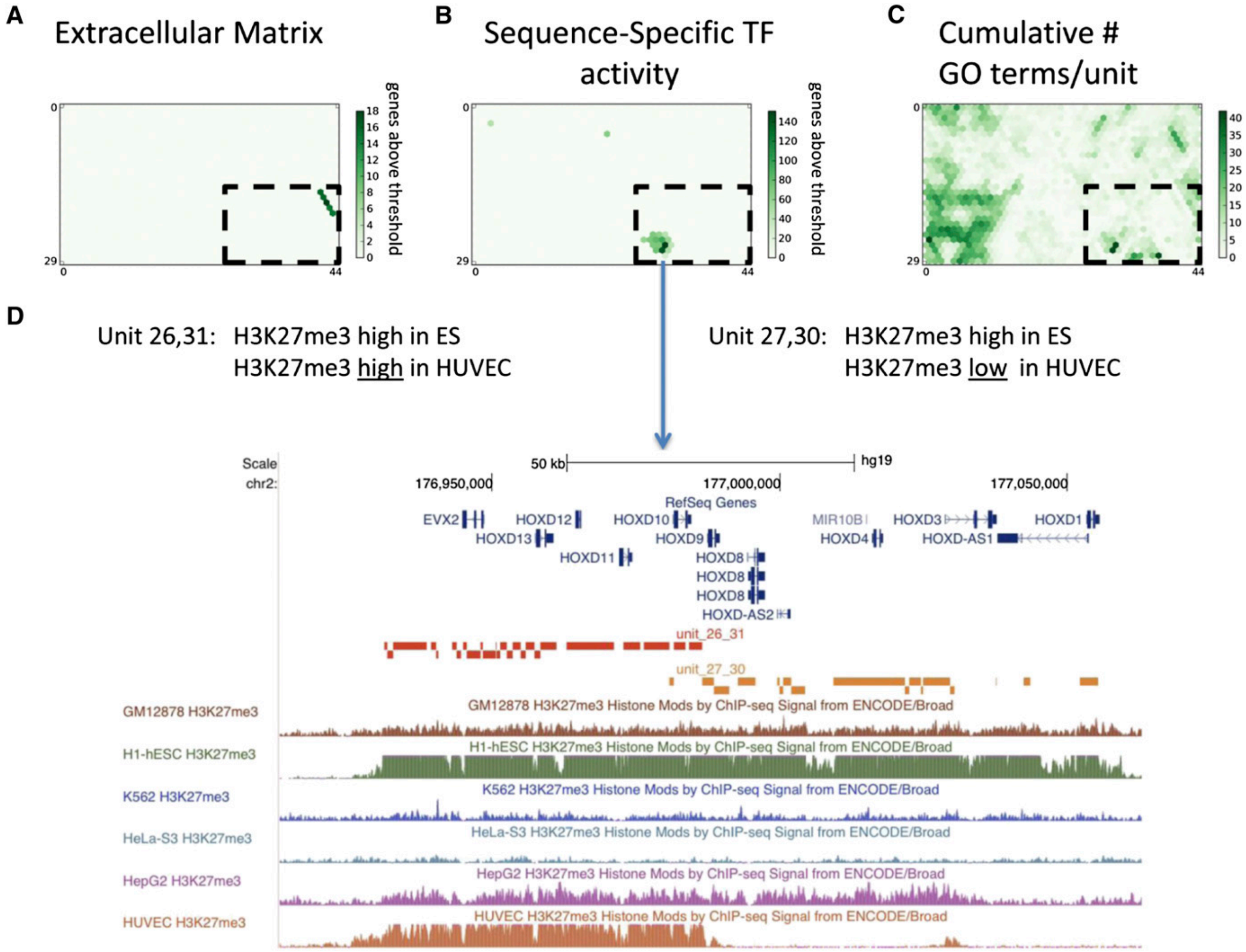
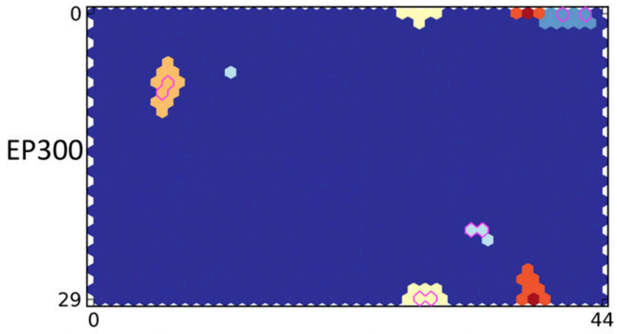


Figure 5. Specific patterns of GO enrichment over the SOM. (*A*) Specific GO terms such as "extracellular matrix" are highly enriched in portions of the map because of activity in one or more cell types. (*B*) Other GO terms are enriched because of their pattern of repression over the map. (*C*) The map has overall highly uneven distribution of GO enrichments away from the regions with the highest nucleotide density. (*D*) An example of the different patterns of H3K27me3 distribution across cell lines captured by neighboring units in the map in the HOXD cluster.

a cluster of units in the upper-left quadrant of the map that did not correspond to TSSs, but that did overlap with validated erythroid CRMs. These units are high in H3K4me1 and H3K27ac that are specific to each cell type. We then asked whether the segments within these units show functional enrichment. For example, three of the GM12878-specific units are enriched with the GO term "immune response." We can easily extend the analysis of the SOM by pooling segments from multiple units and analyzing them using tools such as GREAT (McLean et al. 2010) that associate *cis*-regulatory regions with genes for enrichment in many functional annotations besides GO. Applying GREAT to pooled segments from the cell-specific enriched EP300 units returned a wealth of enriched functional annotations that are predictably associated with the cell-type tissue of origin (Fig. 6). We illustrate this by showing enrichments in Pathway annotations for each cell type. Whereas the units with EP300 signal in more than two cell types are enriched in housekeeping pathways, the GM12878 units show the most enrichment in "immune system" and "interferon signaling," which nicely captures the biology of the cells. This functional enrichment of neighboring units on the map suggests richness of the SOM.

Discussion

Rapidly growing bodies of functional genomics data require methods to integrate and mine large numbers of data sets of multiple kinds. We constructed a self-organized map (SOM) of ENCODE chromatin data from 72 ChIP-seq and DNase-seq data sets from six ENCODE cell lines. Subsequent analyses and mining were facilitated by an interactive web-based SOM-viewer (http://woldlab.caltech.edu/ENCODESOM), which allows users to extend the analysis and extract groups of DNA segments that have characteristics of interest for further computational or wet-bench analysis. While most prior studies of global chromatin data have focused on a specific cell type or tissue, the ENCODE collection allowed us to explore relationships among multiple cell types in a single coherent analysis. By projecting high-dimensional chromatin data onto the two-dimensional SOM, we identified clusters of units



🟨 HepG2

FOXA transcription factor networks	7.0e-46
FOXA2 & FOXA3 transcription factor networks	1.3e-45
Integrins in angiogenesis	6.4e-27

🟧 K562

IL2-mediated signaling events	2.0e-29
KitReceptor	1.5e-18
ATF-2 transcription factor network	6.45e-18

🟦 GM12878

Immune System	6.3e-140
Interferon Signaling	4.9e-91
Interferon alpha/beta signaling	7.1e-80

🟧🟥 Multiple / All

Gene Expression	1.8e-76
Metabolism of RNA	1.1e-61
Metabolism of mRNA	2.3e-53

🟦 H1 hESC

Wnt	1.3e-7

Figure 6. EP300 enrichment highlights cell-type-specific enrichments. ChIP-seq signals of the transcriptional coactivator EP300 in four ENCODE cell types were overlaid on the SOM. While some of the signal is common to multiple/all cell types (orange/brown), each EP300 ChIP-seq data set highlights a different set of adjoining units on the map that is specifically enriched based on the cell type. These cell-type-specific units are also high in H3K4me1 and H3K27ac, which suggest that they hold cell-type-specific enhancers. Segments from each of the colored clusters were pooled and analyzed for functional enrichment with GREAT such as pathways (*top* three terms per cluster shown). While the units common to multiple cell types are enriched in genes involved in housekeeping pathways, those in the cell-type-specific regions are enriched in pathways that are known to be relevant to the biology of those cells.

with chromatin mark combinations corresponding to promoter activity and transcriptional enhancer activity. These were further parsed into smaller clusters that were either cell-type-specific or more ubiquitous. By overlaying data for specific transcription factor binding, enhancer activity, and transcription start sites onto the SOM, we show that the user can discover relationships and mine corresponding genome segments of interest. This was demonstrated for known and candidate erythroid CRMs (Fig. 3). To our knowledge, this is the first use of self-organizing maps for multicell data integration and mining. Although we used a specific, "stacked" genome segmentation generated by ChromHMM, the overall approach can be applied to any segmentation. As discussed below, we expect that the choice of segmentation strategy and the mixture and quality of data sets used in training will affect the resulting SOM.

We mined the SOM to address specific classes of questions. First, individual training data sets revealed clusters that are cell-type-specific or shared for individual marks. The same was true for certain shared sets of marks. Second, units of the SOM were hierarchically clustered based on their prototype vectors, to investigate how multiple mark densities interact with each other. Third, additional data not used in training were projected onto the SOM to map their enrichment in one or more areas, and to relate the underlying chromatin characteristics to map units and clusters where other specific data features are concentrated. In this way, we investigated how individual sequence-specific regulatory factor occupancy for GATA2, SPI1, OCT4, and NANOG, their DNA binding motifs, and the EP300 coactivator are related to each other and to underlying chromatin signatures. Fourth, we mined the SOM for specific functional classes using transcription start sites (TSSs) as the best-defined test case, followed by a curated set of CRMs. The SOM segregated TSSs that are commonly expressed in multiple cell types from the TSSs with cell-type-specific activity into subclusters. Finally, we found that some individual GO terms are preferentially affiliated with different chromatin signatures. To facilitate exploration of the ENCODE SOM by users, we provide a web interface SOM viewer that allows users to explore all the data sets mapped here and to mine out the DNA segment coordinates in any hex-cell or group of cells. We expect this web interface to be the primary means by which users interact with the SOM results.

At the highest level, most observations agreed with conclusions of previous studies using other methods to integrate chromatin data such as hidden Markov models, which were applied to these ENCODE data (The ENCODE Project Consortium 2012). The SOM, however, provided an additional level of granularity that is not accommodated by a relatively small number of states. The SOM also lent itself well to visualizing relationships between the chromatin data and additional data of any type that can be mapped to specific points or intervals on the genome (and hence to the DNA segments in the map). The fine structure of the SOM allowed us to identify distinct combinations of marks and mark intensities shared by only a small number of genomic regions, and did so without any a priori decision about the number of states. For example, the SOM easily separated the variety of different types of TSS into a major cluster of active TSSs versus inactive ones. The active TSSs were internally more finely parsed, based on levels of H3K4me3, as well as distinct cell-type-specific units.

A summary analysis of new candidate transcriptional enhancers is shown in Figure 6. This aggregate analysis is the same one performed for K562 cells (Fig. 3) and uses EP300 signal from each cell type to further concentrate and focus on units active in individual cell types, as well as units that correspond to activity in multiple cell types. Just two units displayed activity in all participating cell types, while a surrounding set of units is variously multitype. Analysis of these units by GREAT showed that those active in all cell types are enriched for well-known housekeeping functions such as protein synthesis. The cell-type-specific units were enriched according to cell type (B lymphocyte, hepatocyte, embryonic stem cell), just as K562 showed erythroid and monocyte categories.

While much of the map organization was driven by histone marks associated with active promoters and enhancers, we point out that this is partly the result of the histone marks used in the ENCODE study for genome segmentation and SOM training. Our input histone marks to the ENCODE SOM clearly favored a fine parsing of active regions over passive ones, and important repressive marks such as H3K9me3 were not included. This makes the ability of this SOM to parse differences in H3K27me3 in different cell lines quite remarkable. Overall, the ENCODE integration efforts showed that a relatively small number of HMM-derived states can capture the broad landscape of active and repressed regions in the ENCODE cell lines (The ENCODE Project Consortium 2012), while the SOM detailed here does this and also gives the biologist access to a wealth of increased resolution and specificity that we coupled with visualization and mining tools. We anticipate that this kind of analysis will be even more useful as the number of cell types and diversity of chromatin marks increase in future studies, making the challenge of combinatoric signatures and their functional correlates greater. In a similar way, as transcription factor location data for many more factors accumulates, the SOM approach and tools developed here will enable end users to better identify and stratify the functionally important and interesting minority of occupied sites that are active in various subsets of cell types.

Methods

Rationale for training matrix design

The joint analysis of multiple cell types presents additional challenges beyond the analysis of multiple data sets in a single cell line. If each cell line is analyzed separately, one is left with the difficult task of trying to reconcile the states found for each with different definitions, before proceeding to analyze state changes between cell lines. Alternatively, one can "concatenate" the data from multiple cell lines (Ernst et al. 2011). Concatenation has the great advantage that the states defined will be consistent across cell lines, but this approach still requires intensive post-processing to extract the segments that change states across cell lines; assuming that a concatenated HMM had seven states in six cell lines, any given genomic segment could be in one of 7^6 = 117,649 combinations of states. Another solution, which we implement here, is to train on all data jointly as a "stack" to learn a single set of states with a single set of genomic boundaries. In this case, one is then left only with the problem of how to interpret the states, whose definitions are virtually certain to involve nonintuitive, complex combinations of marks in one or more cell types and requires additional methods to mine the results in a systematic and intuitive way.

"Stacked" training matrix implementation

To train the SOM, we first built a training matrix composed of signal densities of all 72 data sets (columns) over all segments (rows). The segments were taken from a ChromHMM segmentation of a "stacked" training set of 84 data sets (ChIP-seq for eight histone modifications, RNA Pol II, and CTCF; and three open chromatin data sets for each of six cell lines) using 25 states. We set aside two of the open chromatin data sets to avoid overtraining on open chromatin, and only used the UW DNase-seq data to represent open chromatin as the three experiments are effectively redundant. We converted uniformly processed signal densities of the remaining 72 data sets used for the SOM training into RPKM (reads per kilobase per million reads) for every segment on each training data set using the ERANGE 3.3 getDensity.py script. The training matrix was built using the ERANGE 3.3 buildMatrix.sh script, with a maximum threshold of 100 RPKM and the rescale option.

Training the SOM

The self-organizing maps were trained and analyzed using ERANGE v3.3. For every SOM instance, we shuffled the training set, randomly initialized the toroid map of hexagonal units from the training set, and incrementally trained a SOM with map size 30 by 45 using 5 million iterations, which is equivalent to going through the entire data set 3.3 times, starting with an update bubble radius of 15 and a learning rate of 0.2, both of which decreased exponentially over the course of training. Each segment was assigned to its best matching unit based on the Euclidean distance. We selected for analysis the best of 10 trials based on the lowest quantization error, which is defined as the average Euclidean distance of all segments to the prototype vector of their assigned unit. The other nine instances were used to evaluate the reproducibility of the map by analyzing the fraction of segments from each unit of our best map that resided in the same unit or adjoining units in the other nine map instances.

While we decided to use the entire training matrix for training for the SOM discussed in the main text, the software supports training on the training set and scoring on a distinct test set. In particular, we trained 10 SOMs with half of the segments from the 200-bp naïve segmentation (i.e., half of 1.5 million segments) for 25 million iterations, selected the best one based on the scoring of the other half of the segments, and rescored the best SOM with the ChromHMM segmentation to provide directly comparable genomic coordinates.

There are no theoretical limits to the number of data sets, segments, or map size that could be analyzed with the SOM. However, the ERANGE implementation of the SOM was designed for compatibility with the rest of the package rather than for scalability or performance and will be significantly slower on much larger data sets or number of training iterations. The final training run for the main ENCODE SOM above took a couple of hours, while the naïve segmentation run took 1 d. The per-unit gene-level analysis took significantly longer.

Gene-level analysis

We recovered the identity of the nearest gene within 20 kb of each segment within a unit using the NCBI gene annotation, which is conservative and means that in lower gene-density areas of the genome, many segments were not affiliated with any gene. We then analyzed every unit for Gene Ontology (GO) enrichment as previously described (Mortazavi et al. 2006), adjusting for multiple-hypotheses testing by applying a Bonferroni correction for both the number of tested Gene Ontology terms and the map size.

Metaclustering methods

The unit prototype vectors were automatically aggregated into the larger clusters using standard hierarchical clustering, subject to the constraint that only adjacent clusters on the SOM could be aggregated. A centered correlation distance and centroid linkage were used. Prior to the hierarchical clustering, the prototype vector values along each dimension were replaced with rank values normalized to range between −1 and 1. Heat map visualizations of the hierarchical clustering were rendered using Java Treeview (Saldanha 2004). The clustering itself and the SOM visualizations of it were done using custom C++ and Python code (available at http://woldlab.caltech.edu/~spepke/somclustering/).

Partitionings of the hierarchical clustering at varying levels of detail were generated using the branch length inconsistency criterion implemented in SciPy (depth = 6). The inconsistency of a branch is the ratio of its length to the average length of branches to clusters less then a specified depth below it. For a specified threshold value *t*, the hierarchical clustering is cut at branches that exhibit an inconsistency coefficient greater than *t*. Partitioning of the unit vectors was performed over a broad range of values of *t* up to that for which no branch's inconsistency criterion exceeded *t*, i.e., only one cluster resulted. Sharp drops in the number of clusters as a function of the threshold value occur and are typically followed by plateaus that show little or no change in cluster number. Such behavior suggests partitionings that are relatively robust with respect to the threshold value (see Supplemental Fig. S13).

Acknowledgments

We gratefully acknowledge Ewan Birney, Ian Dunham, Eric Mjolsness, and Paul Sternberg for general discussion of SOMs and their applications to functional genomics; and Diane Trout, Henry Amrhein, and Anna Abelin for computational assistance. At Caltech this work was supported by grants to B.J.W. from the Beckman Foundation, the Donald Bren Endowment, the Gordon Moore Cell Center at Caltech, NIH U54HG004576, NIH U54HG006998, and RC2HG005573; at Hudson Alpha Institute by grant NIH U54HG004576 and NIH U54HG006998 to R.M.M.; and at Penn State University by NIH R01DK065806, U54HG006998, and RC2HG005573 to R.C.H. A.M. was partly supported as a Caltech Beckman Fellow and a Moore Cell Center Fellow while at Caltech. The Mortazavi laboratory at UC Irvine was supported by grant NIH U54HG006998 as well as EU-FP7 project STATegra (306000) to A.M.

References

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28:** 1045–1048.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis of genome-wide expression patterns. *Proc Natl Acad Sci* **95:** 14863–14868.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for the ENCODE Project. *Genome Res* **22:** 1760–1774.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6:** 283–289.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9:** 473–476.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18:** R195–R201.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316:** 1497–1502.

Kohonen T. 2001. *Self-organizing maps*, 3rd ed. Springer, New York.

Lee JS, Smith E, Shilatifard A. 2010. The language of histone crosstalk. *Cell* **142:** 682–685.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501.

Milone DH, Stegmayer GS, Kamenetzky L, Lopez M, Lee JM, Giovannoni JJ, Carrari F. 2010. *omeSOM: A software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinformatics* **11:** 438.

Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu X, White KP, Bussemaker HJ, et al. 2006. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster. Proc Natl Acad Sci* **103:** 12027–12032.

Mortazavi A, Leeper Thompson EC, Garcia ST, Myers RM, Wold B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire. *Genome Res* **16:** 1208–1221.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* **5:** 621–628.

Newman AM, Cooper JB. 2010. AutoSOME: A clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC Bioinformatics* **11:** 117.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6:** S22–S32.

Saldanha AJ. 2004. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20:** 3246–3248.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, et al. 2011. A spatial and temporal map of *C. elegans* gene expression. *Genome Res* **21:** 325–341.

Suzuki M, Oda M, Ramos MP, Pascual M, Lau K, Stasiek E, Agyiri F, Thompson RF, Glass JL, Jing Q, et al. 2011. Late-replicating heterochromatin is characterized by decreased cytosine methylation in the human genome. *Genome Res* **21:** 1833–1840.

Ultsch A. 1999. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen maps* (ed. Oja E), pp. 33–46. Elsevier Science, New York.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903.

Received March 29, 2013; accepted in revised form October 7, 2013.

J

From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing

Originally published as:

Marinov GK*, Williams BA*, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* 2014. **24**:496-510. doi: 10.1101/gr.161034.113.

From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing

Georgi K. Marinov,[1,4] Brian A. Williams,[1,4] Ken McCue,[1] Gary P. Schroth,[2] Jason Gertz,[3] Richard M. Myers,[3] and Barbara J. Wold[1,5]

[1]*Division of Biology, California Institute of Technology, Pasadena, California 91125, USA;* [2]*Illumina, Inc., Hayward, California 94545, USA;* [3]*HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA*

Single-cell RNA-seq mammalian transcriptome studies are at an early stage in uncovering cell-to-cell variation in gene expression, transcript processing and editing, and regulatory module activity. Despite great progress recently, substantial challenges remain, including discriminating biological variation from technical noise. Here we apply the SMART-seq single-cell RNA-seq protocol to study the reference lymphoblastoid cell line GM12878. By using spike-in quantification standards, we estimate the absolute number of RNA molecules per cell for each gene and find significant variation in total mRNA content: between 50,000 and 300,000 transcripts per cell. We directly measure technical stochasticity by a pool/split design and find that there are significant differences in expression between individual cells, over and above technical variation. Specific gene coexpression modules were preferentially expressed in subsets of individual cells, including one enriched for mRNA processing and splicing factors. We assess cell-to-cell variation in alternative splicing and allelic bias and report evidence of significant differences in splice site usage that exceed splice variation in the pool/split comparison. Finally, we show that transcriptomes from small pools of 30–100 cells approach the information content and reproducibility of contemporary RNA-seq from large amounts of input material. Together, our results define an experimental and computational path forward for analyzing gene expression in rare cell types and cell states.

[Supplemental material is available for this article.]

Gene expression levels can differ widely between superficially similar cells. One source of variation is stochastic transcriptional "bursting" (Elowitz et al. 2002; Ozbudak et al. 2002; Blake et al. 2003; Raser and O'Shea 2005; Kaufmann and van Oudenaarden 2007). Those studies mainly used fluorescent protein fusion genes to monitor the expression of one or a few genes. They revealed dynamic fluctuations through time that are seen as "salt-and-pepper" variation across a cell population at any given time. In addition to this bursting behavior, individual cells are expected to display controlled and coordinated differences in the expression of genes engaged in dynamic physiologic processes, such as cell cycle phase progression, paracrine or autocrine signaling response, or stress response. Beyond such already appreciated heterogeneity lie currently unknown cell-to-cell differences with biological implications for defining cell states, metabolic function, and, in complex tissues, cell identity.

Measuring RNA transcripts in single cells is now done in multiple ways, and similar conclusions about variability are emerging from the higher sensitivity methods. For individual genes, single molecule RNA fluorescence in situ hybridization (SM-RNA FISH) is highly informative (Femino et al. 1998; Raj et al. 2008), and multiplexed versions now enable multiple genes to be measured in parallel (Lubeck and Cai 2012). In principle, an advantage of SM-RNA FISH is the ability to accurately count the absolute number of transcripts in a cell. A second and older approach is multiplexed single-cell RT-qPCR (Cornelison and Wold 1997), which has now been advanced to increasingly high-throughput formats (White et al. 2011; Sanchez-Freire et al. 2012, Livak et al. 2013). It produces semiquantitative relative comparisons between individual cells. However, neither SM-RNA FISH nor the current forms of multiplex RT-qPCR cover the entire transcriptome or have the single-nucleotide resolution needed to study fine-structure features of gene expression such as allele specificity, RNA editing, and alternative splicing.

To address these and other limitations, elegant methods have recently been developed for performing RNA-seq with very small amounts of RNA, down to the level of individual cells. These are broadly referred to as "single-cell RNA-seq" (Tang et al. 2009, 2010, 2011; Ozsolak et al. 2010; Islam et al. 2011; Brouilette et al. 2012; Cann et al. 2012; Hashimshony et al. 2012; Pan et al. 2012; Qiu et al. 2012; Ramsköld et al. 2012). Despite these significant advances, there are substantial shortcomings in these methods, and a robust method for comprehensive and accurate measurement of the transcriptome of a single cell is not yet available.

A particular challenge for single-cell methods is the efficiency and uniformity with which each mRNA is copied into cDNA and ultimately represented in the library. This challenge intersects in crucial ways with transcriptome structure. Specifically, thousands of genes are expressed in the range of 1 to 30 mRNA copies per cell, including many essential mRNAs (for example, key transcription factors) (Zenklusen et al. 2008). Even lower transcript levels, averaging <1 mRNA per cell on the population level, are now being reliably detected by RNA-seq. This raises questions whether very rare RNAs represent background biological noise, or alternatively, are functional in only a small fraction of cells. Single-cell RNA-seq has the potential to address these issues,

[4]These authors contributed equally to this work.
[5]Corresponding author
E-mail woldb@caltech.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.161034.113. Freely available online through the *Genome Research* Open Access option.

but their resolution depends on how faithfully and efficiently RNAs are captured and represented in sequencing libraries (referred to throughout as the "single-molecule capture efficiency," p_{smc}). In addition, the uniformity of transcript coverage in early single-cell RNA-seq protocols has typically been heavily biased toward the 3′ end, which affects both gene expression estimates and the ability to analyze alternative splicing, RNA editing, and allelic bias.

A second major use for single-cell RNA-seq is the transcriptomic characterization of rare cells. The human body consists of hundreds of distinct cell types, plus large numbers of neuronal and transient developmental cell types. Many of these are numerically minor components of complex tissues, making them inaccessible to standard methods relying on large RNA inputs. Isolation of single cells based on the cell surface markers or using microdissection coupled with single-cell RNA-seq could fill this gap in complex multicellular organisms. However, the feasibility of this approach also depends on the experimental robustness of single-cell RNA-seq protocols. Alternatively, single-cell resolution may not be absolutely required for this purpose, and small pools of cells may be sufficient to characterize rare cell-type transcriptomes. An open unresolved question is how small such pools can be to adequately meet that goal.

In this study, we address the issues highlighted above. We used the SMART-seq protocol (Ramsköld et al. 2012) to measure the transcriptome of single cells and small cell pools from the GM12878 lymphoblastoid cell line. This line is derived from the NA12878 individual, for which a fully sequenced genome with completely phased heterozygous single nucleotide polymorphisms (SNPs) and indels is available (The 1000 Genomes Project Consortium 2012). GM12878 cells have also been the subject of an extensive functional genomic characterization by the ENCODE Consortium (The ENCODE Project Consortium 2011, 2012) and have been used in prior population-level studies of allele-biased gene expression and transcription factor occupancy (Rozowsky et al. 2011; Reddy et al. 2012).

Using spike-in quantification standards of known abundance (Mortazavi et al. 2008), we derive estimates for the absolute number of transcript copies for each gene in each cell and directly measure the average value of p_{smc}. "Pool/split" experiments (consisting of pooling RNA from multiple single cells, splitting the pool into the same number of separate reactions and building libraries from them) allowed us to measure the extent of and control for technical variation. We find that the p_{smc} value is quite low: ~0.1. An analysis framework accounting for technical stochasticity is described and used to assess variability in gene expression, allelic bias, and alternative splicing between single cells. Distinct from prior studies, our approach allowed us to parse findings into those that are just as likely to be of technical origins and those that are more likely to be of biological interest.

We report evidence of significant variability in the total number of mRNA molecules per cell, and identify biologically coherent modules of coexpressed genes specifically expressed in individual cells or groups of cells. These include expected variation associated with cell cycle phases, and an unexpected module enriched for mRNA processing and splicing genes. We observe evidence of higher levels of autosomal allelic exclusion on the single-cell level, potentially associated with transcription bursts; however, it is at present difficult to confidently distinguish from technical variability. In contrast, we find much stronger evidence for widespread major splice site usage switches between individual cells. Finally, our analysis of similarly constructed small cell pools (30–100 cells) reveals a high robustness and reproducibility, approaching that of bulk RNA measurements. This presents a reliable path forward toward the future comprehensive transcriptomic characterization of rare cell types.

Results

In silico examination of major variables affecting informativeness of single-cell and small cell-pool RNA-seq

We began this study with two goals: first, to study gene expression heterogeneity in GM12878 cells on the single-cell level, and second, to determine the minimal optimal size of a cell pool that is informative of the characteristics of the larger cell population, with the goal of applying that approach to rare cell types in future studies. How well these goals are achieved depends on several parameters affecting biological and technical stochasticity and detection sensitivity, the values of which were unknown. To understand their influence, we carried out a simulation of single-cell and cell-pool transcriptomes (see Supplemental Methods for details) by varying the following parameters:

1. Single-molecule capture efficiency p_{smc}. In contrast to bulk RNA-seq libraries, an individual cell contains a very limited total number of mRNA molecules. Individual genes can be present in single-digit transcript numbers. If only a fraction of mRNAs are successfully represented in a library, a technical stochasticity component is introduced. Depending on its magnitude, data interpretability can be significantly affected due to false negatives and a distortion of relative gene abundance estimates. The p_{smc} parameter is the probability that any given original RNA molecule is captured in the final library. We examined the effect on expression quantification of p_{smc} ranging from 0.01 to 1.
2. Total number of mRNA molecules per cell. The impact of low p_{smc} on expression measurements will be more severe if fewer mRNA molecules are present in a cell. The average total number of mRNA molecules in a single cell is not known for most cell types, but it is expected to vary with cell size, metabolic status, and even cell cycle phase. This means that single-cell expression measurements in some cell types are likely to be more robust to technical noise than in others. We varied the total number of mRNAs from 50,000 to 1,000,000 (while keeping the number of genes expressed constant).
3. Frequency of expression of individual genes in single cells. From prior studies we expect that some genes will be expressed in all or most cells, while others will be expressed in only a subset of cells. Genes detected at lower levels in bulk RNA-seq are the most obvious candidates to be expressed in a subset of cells in a population, although we do not know what fraction of low-abundance RNAs behave in such a way. This is particularly relevant to cell pools: a gene expressed at 50 copies per cell but only in 10% of cells would still be stochastically represented in a pool of 10 cells even if p_{smc} is high. In the absence of reliable data on this, we modeled the probability of expression in a given single cell with a distribution centered around very high values for genes highly expressed in bulk RNA-seq measurements, and progressively lower values with decreasing expression levels (details in Supplemental Methods).

The simulation results are summarized in Figure 1, A–C and Supplemental Figures 1–25. As expected, low p_{smc} has a profoundly negative impact on gene expression quantification accuracy and reliability, leading to frequent false negatives (Fig. 1A; Supplemental

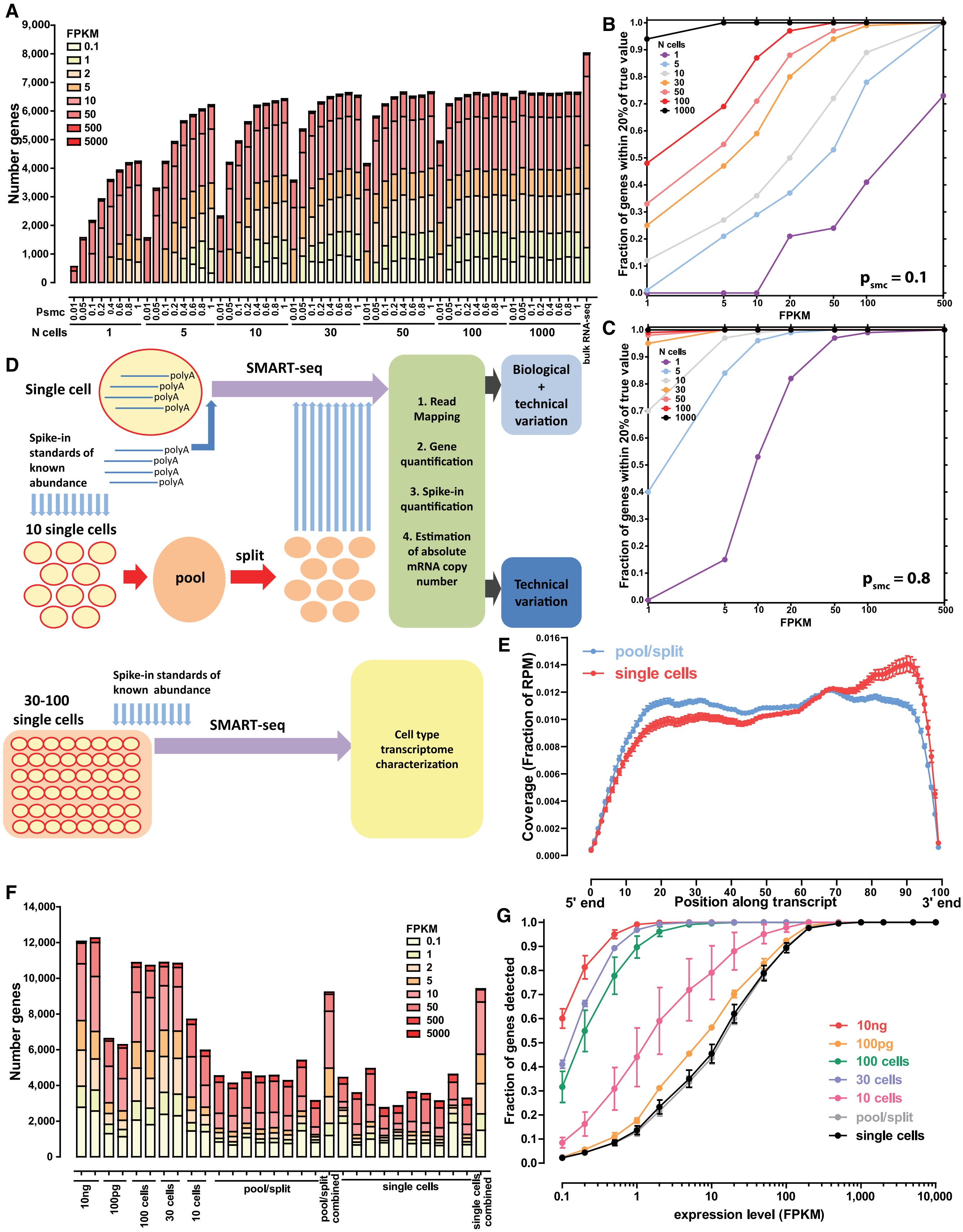


Figure 1. (Legend on next page)

Fig. 1), and to poor estimates of expression levels. For example, in a single cell with 100,000 mRNAs, $p_{smc} = 0.1$ results in only 40% of genes expressed at 100 FPKM receiving FPKMs within 20% of the true value (Supplemental Fig. 1C), but this fraction rises to nearly 100% if $p_{smc} = 0.8$ (Supplemental Fig. 1G). The quantification of relative expression levels is similarly affected, with only the most highly expressed genes being consistently well-quantified relative to each other at low p_{smc} (Supplemental Figs. 12–25).

In contrast, our simulation results indicate that cell pools are much more robust to technical noise, with 90% of genes expressed at 10 FPKM receiving FPKM estimates within 20% of their true value (Supplemental Fig. 1C) at $p_{smc} = 0.1$ in a pool of 100 cells. They also represent the expression profiles of the general population reasonably well (Supplemental Fig. 1), even at low p_{smc}, starting from a size of ~30 cells (10-cell pools seem not to be sufficient to achieve this). Finally, as expected, the larger the number of total mRNA molecules per cell, the greater is the buffer against technical noise, resulting in more robust quantification (Supplemental Figs. 2–11).

Transcriptome measurements of individual single cells and companion pool/splits

The simulation results informed our experimental design, which aimed to gain a firm grasp on technical stochasticity in two ways (Fig. 1D). First, we generated single-cell RNA-seq libraries and in parallel carried out "pool/split" experiments. In a pool/split, multiple cells are pooled and lysed together, then split into the same number of reactions, from which libraries are built. Variation between these libraries should be purely technical (with stochastic splitting possibly playing a role at the low end). Variation observed at similar levels in both single cells and pool/splits cannot be confidently attributed to biological differences, although the stringency of this criterion may cause some true biological variation to be obscured. However, variation above the pool/split level can be identified and ascribed to biological sources with high confidence.

We generated single-cell RNA-seq libraries from 15 single GM12878 cells and from two pairs of 10-cell pool/split experiments. We also sequenced replicates of pools of multiple cells (10, 30, and 100 cells), as well as 100-pg and 10-ng samples of bulk RNA (corresponding to ~10 and ~1000 cells), to assess the stability of measurements as a function of the amount of starting material.

We used the SMART-seq protocol (Supplemental Fig. 12; Ramsköld et al. 2012) to generate our libraries. A detailed description of the protocol, as we implemented it, is presented in Supplemental Methods. We obtained nearly uniform full-length transcript coverage (Fig. 1E; Supplemental Fig. 29). Uniformity of coverage, which depends on the intactness of RNAs and the successful copying of full-length molecules, is highly desirable for several reasons. First, RNA-seq data quantification using the RPKM/FPKM metric (Mortazavi et al. 2008; Trapnell et al. 2010) makes an implicit assumption of full coverage. Second, it enables the analysis of alternative splicing and allelic bias, as read coverage of 5′-proximal splice sites and heterozygous positions is ensured.

We added spike-in quantification standards of known abundance (in absolute number of RNA copies) (Supplemental Table 2) at the very beginning of cDNA synthesis. This allows us to, first, estimate p_{smc}, and second, derive gene expression estimates in absolute numbers of copies per cell. The latter is important because while FPKM is useful for comparing expression levels within a library, it can only be used to compare directly across different libraries when the total amount of RNA in each starting sample is roughly the same (Anders and Huber 2010). This assumption is usually only mildly violated when working with bulk samples, but when single cells are compared, it becomes significantly more problematic as the variation in the total amount of RNA in each cell is expected to be much larger.

Figures 1 and 2 summarize the technical characterization of the SMART-seq protocol applied to GM12878 cells. In addition to the mostly complete coverage along transcript length, sequencing libraries were also highly enriched for exonic sequences (Supplemental Fig. 28), indicating a high efficiency of enrichment for polyadenylated molecules.

Gene detection in single cells versus pools of varied sizes

We compared single-cell and pool/split libraries, as well as cell pools, with bulk RNA samples from GM12878 cells (Fig. 1F). In bulk RNA libraries, we detect about 12,000 genes expressed at more than 0.1 FPKM. A lower number of genes, between 4000 and 5000, is detected in both single-cell and pool/split libraries. These differences between single cells and bulk libraries are due mostly to genes expressed at low levels. Genes expressed at more than 100 FPKM in 10-ng bulk RNA samples are detected in almost all libraries, while only ~30% of genes expressed at ~10 FPKM and 10% of genes expressed at ~1 FPKM were detected in any given single-cell library (Fig. 1G). Notably, the number of genes detected in both 100-cell and 30-cell pools was similar to that detected in the 10-ng libraries (~11,000). In contrast, in the 10-cell pools and 100-pg libraries, lower numbers of genes were detected, between

Figure 1. Simulated and measured transcriptome profiles from individual cells and small cell pools. (*A*) Number of detected genes in simulated data sets as a function of the number of cells pooled and the single molecule capture efficiency (p_{smc}) (assuming 100,000 mRNA molecules per cell). See Supplemental Figure 1 for full details. (*B,C*) Accuracy of gene expression estimation as a function of the number of cells pooled and the single molecule capture efficiency; $p_{smc} = 0.1$ in *B* and $p_{smc} = 0.8$ in *C*, 100,000 mRNA molecules per cell assumed. Shown is the fraction of genes at the indicated expression levels in FPKM, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value, after modeling the stochasticity due to the single-molecule capture efficiency of the library-building protocol. See the Methods section and Supplemental Figures 2–11 for full details. Note that the simulation is intended to illuminate the relative effects of the various parameters studied, and the absolute numbers of genes should not be directly compared to the real-life data shown in *G*. (*D*) Experimental design. Single cells are combined with spike-in quantification standards and SMART-seq libraries are generated. In parallel, multiple single cells are pooled together and combined with spikes, then lysed and split into the same number of reactions and converted into SMART-seq libraries. Libraries are then sequenced, data processed computationally, and estimates for the absolute number of copies per cell are derived based on the spikes. Variation in pool/split experiments is due to technical stochasticity, while variation in single-cell libraries is a combination of biological variation and technical noise. (*E*) Uniformity of transcript coverage. Shown is the average coverage along the length of an mRNA for single cells and pool/split experiments. Only mRNAs longer than 1 kb from genes with a single annotated isoform in the RefSeq annotation set were included. See Supplemental Figure 29 for more details. (*F*) Number of detected protein-coding genes for libraries built from 10 ng and 100 pg of poly(A) RNA, pools of 100, 30, and 10 cells, representative pool/split experiments (individually and summed across all libraries), and representative single cells (individually and summed across all libraries). (*G*) Fraction of genes from 100-ng bulk poly(A)$^+$ RNA libraries that were detected in pools of 100, 30, or 10 cells, 100 pg of poly(A)$^+$ RNA, pools/split experiments, and single cells. FPKM is shown on the *x*-axis.

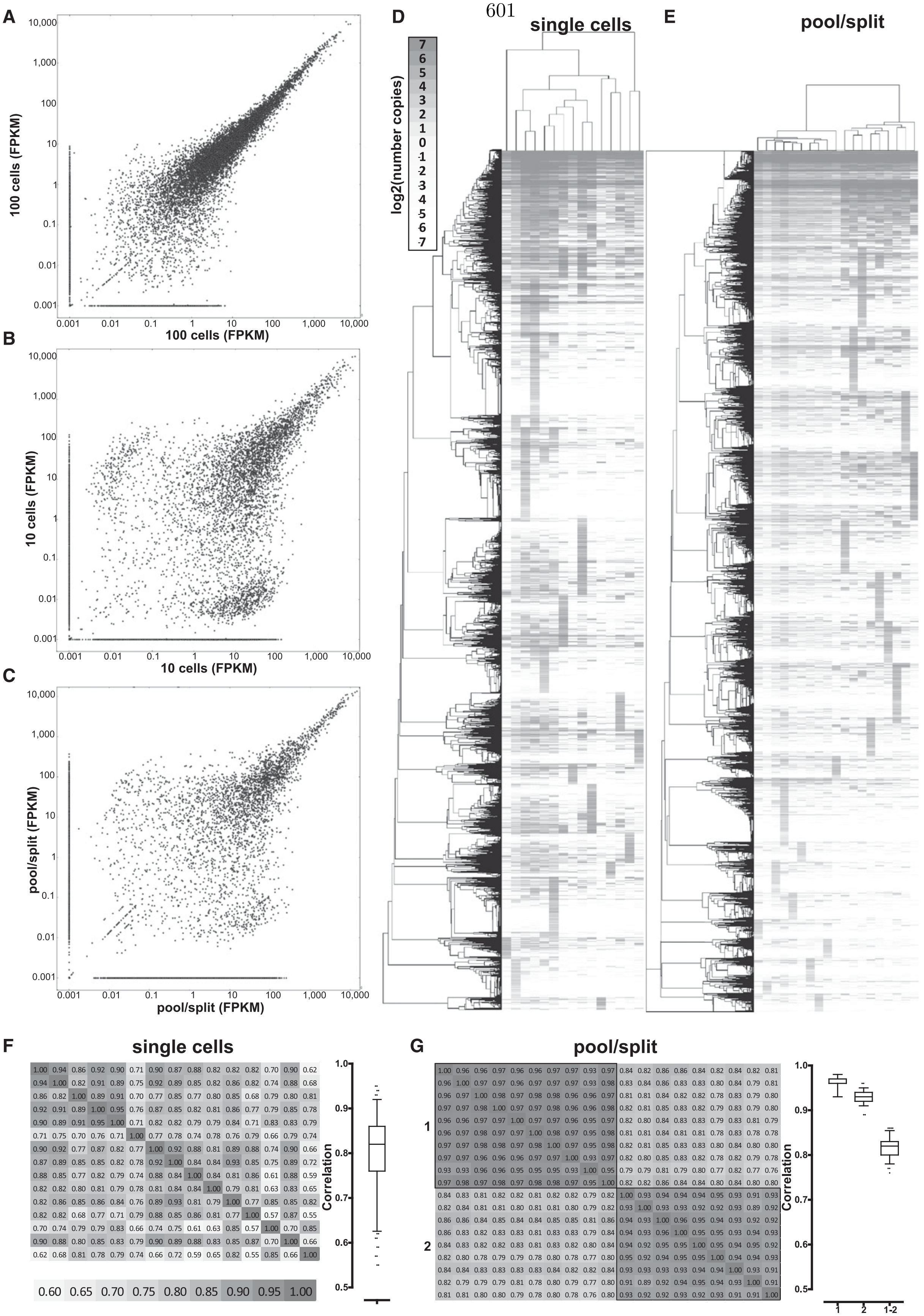


Figure 2. (Legend on next page)

6000 and 7000. This is consistent with simulation results suggesting that 30 cells is the lower limit of cell number at which the transcriptome library complexity begins to approach that of the larger cell population. This is corroborated by the correlation between the expression levels of replicate measurements (Fig. 2A; Supplemental Fig. 50). In contrast, a sizable population of genes present at high levels in one replicate and at very low levels or completely absent in the other appears in 10-cell pools (Fig. 2B) and especially in pool/split libraries (Fig. 2C). Finally, union sets of genes detected in all individual cell libraries and in all pool/split libraries was ~10,000, which was in the range seen for 30- to 100-cell pools.

Pool/splits measure technical variation and reveal biological variation among single cells

The observed variations in gene expression levels and detection can be explained as a combination of some genes not being expressed in each and every cell and low p_{smc} resulting in large numbers of false negatives. We calculated the average p_{smc} across all libraries based on the detection of spike-ins (details in Methods). This number is in our estimates: ~0.1. We also estimate that for GM12878 single cells, one transcript copy corresponds to ~10 FPKM on average. This agrees well with the observation that detection of genes becomes unstable below ~100 FPKM (Fig. 2B,C), which is also consistent with previous observations (Ramsköld et al. 2012).

We next compared expression measurements in single-cell and pool/split libraries. Hierarchical clustering of each group is shown in Figure 2, D and E (with two independent biological replicate pool/split experiments shown in Fig. 2E). The distances between the expression profiles within the same pool/split experiment were significantly smaller than those for individual single cells (branch lengths in Fig. 2D,E), and average correlations between single cells were, accordingly, lower than those between libraries from the same pool/split (Fig. 2F,G; Supplemental Fig. 32). A notable feature of the data is small clusters of genes present at high levels in only one library. These are more prominent in single cells than in pool/splits, yet they are clearly present in all samples. In single cells, this is due to a mixture of stochastic capture effects and real biological variation. In pool/splits, stochastic capture is the predominant source. It is important to note that given the low p_{smc}, it is difficult to determine the cause of variation for any given gene. Nevertheless, the major conclusion at the transcriptome level is that there are biological differences between single cells because the technical stochasticity in pool/splits is significantly less than variation across single cells.

Estimating absolute transcript levels in single cells

Absolute transcript counts are the biologically relevant values ideally obtained from a single-cell gene expression profiling experiment because, as discussed above, FPKM is a poor metric for comparing gene expression levels between individual cells if the total amount of RNA varies substantially. We derive transcript number estimates for each gene based on the FPKM values of spike-ins. We observed good agreement between the input number of spike-in RNA copies and the corresponding FPKM values in the final libraries (Supplemental Figs. 30, 31).

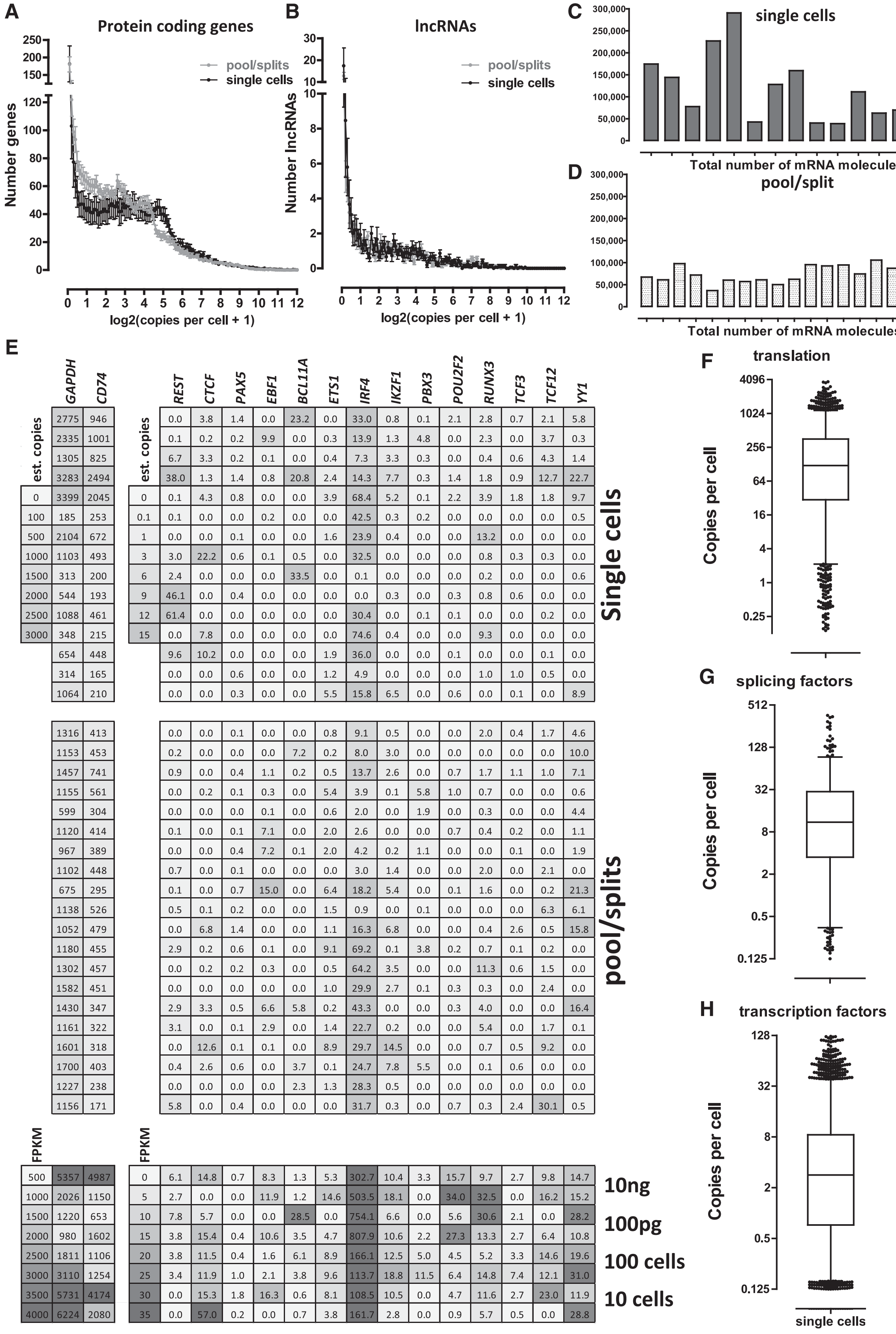
We use the transcripts-per-cell estimates for all subsequent analyses. Previous studies have reported that genes can be separated into two distinct groups based on their expression levels—one group expressed at high (>1 FPKM) levels and one at very low (<<1 FPKM) levels (Hebenstreit et al. 2011). We examined the distribution of estimated copies per cell in single cells and in pool/splits (Fig. 3A). We find that in individual cells, most protein-coding genes are expressed at levels between 1 and ~50 copies per cell. The distribution suggests a roughly equal number of genes at each level except for a larger group of transcripts with fractional transcript-per-cell values. Obviously, single-cell determinations are constrained in a way that population level measurements cannot be: One transcript per cell is the minimum nonzero value possible. The lower values likely represent a combination of mapping artifacts (due to high sequence homology of paralogs) and RNAs that were present at low levels to begin with and then poorly represented in the final library (due, for example, to the fragmentation of a single original RNA molecule resulting in artificially low FPKMs as a result of coverage only at the 3′ end). The distribution of estimated copies in pool/split libraries exhibited a more linear decrease in the number of more highly expressed genes, consistent with averaging of variation between cells.

We also examined the distribution of the expression levels of long noncoding RNAs (lncRNAs) (Guttman et al. 2009). Consistent with previous observations (Ramsköld et al. 2009; Guttman et al. 2010; Djebali et al. 2012), lncRNAs have generally much lower expression levels compared to protein-coding genes (Fig. 3B).

We were also able to directly assess the total number of mRNAs present in each cell (Fig. 3C,D). Based on the average mass of RNA in each cell (derived from bulk RNA samples from a known number of cells) and the average length of mRNAs in the human genome, we estimated that each GM12878 cell contains, on average, ~80,000 mRNAs. However, we observed striking cell-to-cell differences in the total transcript number of single cells, with some cells expressing <50,000 mRNAs and others almost 300,000. In contrast, pool/split experiments exhibited remarkable uniformity (between 50,000 and 100,000 transcripts) and agree well with prior expectations. It is therefore unlikely that the observed cell-to-cell variability is due to technical noise.

Because transcriptional regulators play a crucial role in defining the gene expression state of cells, we examined the expression of several well-known general transcription factors as well as major regulators of B-cell differentiation (Fig. 3E). Remarkably, except for *IRF4*, which was usually expressed at several dozen copies, most factors were detected at <10 copies per cell, and were often not detected at all. We stress that this does not mean that they are not expressed. Given the 10% p_{smc} of the protocol, these

Figure 2. Technical and biological variation in single-cell RNA-seq measurements of gene expression. (*A*) Correlation between expression levels (in FPKM) between two pools of 100 cells. (*B*) Correlation between expression levels (in FPKM) between two pools of 10 cells. (*C*) Correlation between expression levels (in FPKM) between two representative pool/split libraries. A pseudocount of 0.001 was added to each data point in the scatter plots for visualization purposes. (*D*,*E*) Hierarchical clustering of estimated copies-per-cell values for protein-coding genes in single-cell (*D*) and pool/split (*E*) libraries. Pearson correlation was used as a distance metric, and only genes expressed at a level of at least one estimated copy in at least one library were included. (*F*,*G*) Correlation between estimated copies-per-cell values for protein-coding genes in single-cell libraries (*F*) and pool/split libraries (*G*). Two sets of pool/split experiments (1 and 2) are shown and "1-2" in the boxplot refers to correlations between the two sets, while "1" and "2" refer to correlation within each experiment. Similar plots, but using the Spearman correlation, are shown in Supplemental Figure 32.



	GAPDH	*CD74*	*REST*	*CTCF*	*PAX5*	*EBF1*	*BCL11A*	*ETS1*	*IRF4*	*IKZF1*	*PBX3*	*POU2F2*	*RUNX3*	*TCF3*	*TCF12*	*YY1*
Single cells	2775	946	0.0	3.8	1.4	0.0	23.2	0.0	33.0	0.8	0.1	2.1	2.8	0.7	2.1	5.8
	2335	1001	0.1	0.2	0.2	9.9	0.0	0.3	13.9	1.3	4.8	0.0	2.3	0.0	3.7	0.3
	1305	825	6.7	3.3	0.2	0.1	0.0	0.4	7.3	3.3	0.3	0.0	0.4	0.6	4.3	1.4
	3283	2494	38.0	1.3	1.4	0.8	20.8	2.4	14.3	7.7	0.3	1.4	1.8	0.9	12.7	22.7
	3399	2045	0.1	4.3	0.8	0.0	0.0	3.9	68.4	5.2	0.1	2.2	3.9	1.8	1.8	9.7
	185	253	0.1	0.0	0.0	0.2	0.0	0.0	42.5	0.3	0.2	0.0	0.0	0.0	0.0	0.5
	2104	672	0.0	0.0	0.1	0.0	0.0	1.6	23.9	0.4	0.0	0.0	13.2	0.0	0.0	0.0
	1103	493	3.0	22.2	0.6	0.1	0.5	0.0	32.5	0.0	0.0	0.0	0.8	0.3	0.3	0.0
	313	200	2.4	0.0	0.0	0.0	33.5	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.6
	544	193	46.1	0.0	0.4	0.0	0.0	0.0	0.0	0.3	0.0	0.3	0.8	0.6	0.0	0.0
	1088	461	61.4	0.0	0.0	0.0	0.0	0.0	30.4	0.0	0.1	0.1	0.0	0.0	0.2	0.0
	348	215	0.0	7.8	0.0	0.0	0.0	0.0	74.6	0.4	0.0	0.0	9.3	0.0	0.0	0.0
	654	448	9.6	10.2	0.0	0.0	0.0	1.9	36.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0
	314	165	0.0	0.0	0.6	0.0	0.0	1.2	4.9	0.0	0.0	0.0	1.0	1.0	0.5	0.0
	1064	210	0.0	0.0	0.3	0.0	0.0	5.5	15.8	6.5	0.0	0.6	0.0	0.1	0.0	8.9
pool/splits	1316	413	0.0	0.0	0.1	0.0	0.0	0.8	9.1	0.5	0.0	0.0	2.0	0.4	1.7	4.6
	1153	453	0.2	0.0	0.0	0.0	7.2	0.2	8.0	3.0	0.0	0.0	0.0	0.0	0.0	10.0
	1457	741	0.9	0.0	0.4	1.1	0.2	0.5	13.7	2.6	0.0	0.7	1.7	1.1	1.0	7.1
	1155	561	0.0	0.2	0.1	0.3	0.0	5.4	3.9	0.1	5.8	1.0	0.7	0.0	0.0	0.6
	599	304	0.0	0.0	0.0	0.1	0.0	0.6	2.0	0.0	1.9	0.0	0.3	0.0	0.0	4.4
	1120	414	0.1	0.0	0.1	7.1	0.0	2.0	2.6	0.0	0.0	0.7	0.4	0.2	0.0	1.1
	967	389	0.0	0.0	0.4	7.2	0.1	2.0	4.2	0.2	1.1	0.0	0.0	0.1	0.0	1.9
	1102	448	0.7	0.0	0.1	0.0	0.0	0.0	3.0	1.4	0.0	0.0	2.0	0.0	2.1	0.0
	675	295	0.1	0.0	0.7	15.0	0.0	6.4	18.2	5.4	0.0	0.1	1.6	0.0	0.2	21.3
	1138	526	0.5	0.1	0.2	0.0	0.0	1.5	0.9	0.0	0.1	0.0	0.0	0.0	6.3	6.1
	1052	479	0.0	6.8	1.4	0.0	0.0	1.1	16.3	6.8	0.0	0.0	0.4	2.6	0.5	15.8
	1180	455	2.9	0.2	0.6	0.1	0.0	9.1	69.2	0.1	3.8	0.2	0.7	0.1	0.2	0.0
	1302	457	0.0	0.2	0.2	0.3	0.0	0.5	64.2	3.5	0.0	0.0	11.3	0.6	1.5	0.0
	1582	451	0.0	0.0	0.0	0.0	0.0	1.0	29.9	2.7	0.1	0.3	0.3	0.0	2.4	0.0
	1430	347	2.9	3.3	0.5	6.6	5.8	0.2	43.3	0.0	0.0	0.3	4.0	0.0	0.0	16.4
	1161	322	3.1	0.0	0.1	2.9	0.0	1.4	22.7	0.2	0.0	0.0	5.4	0.0	1.7	0.1
	1601	318	0.0	12.6	0.1	0.1	0.0	8.9	29.7	14.5	0.0	0.0	0.7	0.5	9.2	0.0
	1700	403	0.4	2.6	0.6	0.0	3.7	0.1	24.7	7.8	5.5	0.0	0.1	0.6	0.0	0.0
	1227	238	0.0	0.0	0.0	0.0	2.3	1.3	28.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	1156	171	5.8	0.0	0.4	0.0	0.0	0.0	31.7	0.3	0.0	0.7	0.3	2.4	30.1	0.5
10ng	5357	4987	6.1	14.8	0.7	8.3	1.3	5.3	302.7	10.4	3.3	15.7	9.7	2.7	9.8	14.7
	2026	1150	2.7	0.0	0.0	11.9	1.2	14.6	503.5	18.1	0.0	34.0	32.5	0.0	16.2	15.2
100pg	1220	653	7.8	5.7	0.0	0.0	28.5	0.0	754.1	6.6	0.0	5.6	30.6	2.1	0.0	28.2
	980	1602	3.8	15.4	0.4	10.6	3.5	4.7	807.9	10.6	2.2	27.3	13.3	2.7	6.4	10.8
100 cells	1811	1106	3.8	11.5	0.4	1.6	6.1	8.9	166.1	12.5	5.0	4.5	5.2	3.3	14.6	19.6
	3110	1254	3.4	11.9	1.0	2.1	3.8	9.6	113.7	18.8	11.5	6.4	14.8	7.4	12.1	31.0
10 cells	5731	4174	0.0	15.3	1.8	16.3	0.6	8.1	108.5	10.5	0.0	4.7	11.6	2.7	23.0	11.9
	6224	2080	0.0	57.0	0.2	0.0	0.7	3.8	161.7	2.8	0.0	0.9	5.7	0.5	0.0	28.8

Figure 3. (Legend on next page)

observations are consistent with simple technical failure to detect them. It is also possible that there are no mRNA copies in some cells at the moment of harvest, especially if they are infrequently transcribed. Extending these observations to other functional groups, we assessed proteins involved in translation (as a major group of genes with housekeeping functions) (Fig. 3F), splicing regulators (Fig. 3G), and all transcription factors (Fig. 3H). The median number of copies per cell was ∼100 for translation proteins, ∼10 for splicing regulators, and strikingly, only ∼3 for transcription factors. Beyond their biological interest, these large expression differences between functional gene categories mean that quantification is inherently less robust and less informative for some biological functions than it is for others.

Identification of modules of coexpressed genes

Cell-to-cell gene expression variability may occur on the level of individual genes, but it can also occur in a coordinated fashion. A well-studied example is cell cycle phase-specific gene expression. In an asynchronous culture, groups of genes expressed highly at specific times during the cell cycle should be present in a fraction of cells that is roughly proportional to the time cells spend in each identified phase. Population data do not, however, predict that most cells will be in a "pure" phase state nor that they will express phase-class genes at peak levels.

To test whether we are able to identify cell cycle-associated variation, and to search for any novel functional modules, we carried out weighted gene coexpression network analysis (WCGNA) (Zhang and Horvath 2005) using the copies per cell estimates for single cells and removing genes that were highly variant in pool/split libraries in order to minimize technical noise (see Methods; Supplemental Figs. 33, 34). We identified 19 coexpression modules containing ≥10 genes each (Supplemental Fig. 35). The expression patterns of these modules were mostly well-differentiated among single cells and were absent from pool/split libraries (Fig. 4B; Supplemental Fig. 34).

We then determined the Gene Ontology (GO) category enrichment of each module. The largest module (module 1) was highly enriched for GO categories relating to housekeeping and anabolic gene functions (Table 1; Supplemental Table 3). This included some enrichment for the G_1- and S-phase GO terms, and also contained most genes that are generally highly expressed (Fig. 4A). Module 6 was enriched for genes involved in the M phase of the cell cycle. A single cell from the sample of 15 showed strong coordinated expression of genes from the M-phase GO categories enriched in this module. Transcripts from these M-phase genes were not similarly coordinated in other individual cells or in pool/split samples. We measured the fraction of unsynchronized GM12878 cells in the $G_0 + G_1$, S, and M phases of the cell cycle using flow cytometry (Fig. 4B). About 14% of cells were in M phase, and the probability of capturing exactly one such cell out of 15 is 0.25; that is, these observations are consistent with this cell being in the peak of M phase.

A more surprising observation was that the second largest module (module 2) was enriched for genes involved in splicing and mRNA processing. It is driven by an individual cell and two additional cells with a somewhat similar expression profile. The signature cell, however, was not an outlier when splice site usage patterns were compared between individual cells (data not shown). A simple interpretation of these observations is a general up-regulation of splicing and mRNA processing factors in that cell that does not result in a distinctive alternative splicing program.

Module 3 was enriched for metabolic cofactor and iron-sulfur cluster binding proteins, including proteins involved in mitochondrial respiratory chains. This is an intriguing observation, as module 3 was mostly driven by the two cells exhibiting the highest total number of mRNA molecules per cell (Fig. 3C; fourth and fifth columns in clustergram in Fig. 4A), consistent with a generally elevated metabolic state.

We also carried out a mirrored WCGNA analysis in which the pool/splits were treated as single cells and vice versa. We did not observe significant GO enrichment beyond a few trivial terms in the largest modules (Supplemental Fig. 54; Supplemental Table 4). This is in contrast to the much more specific GO enrichment seen in single cells.

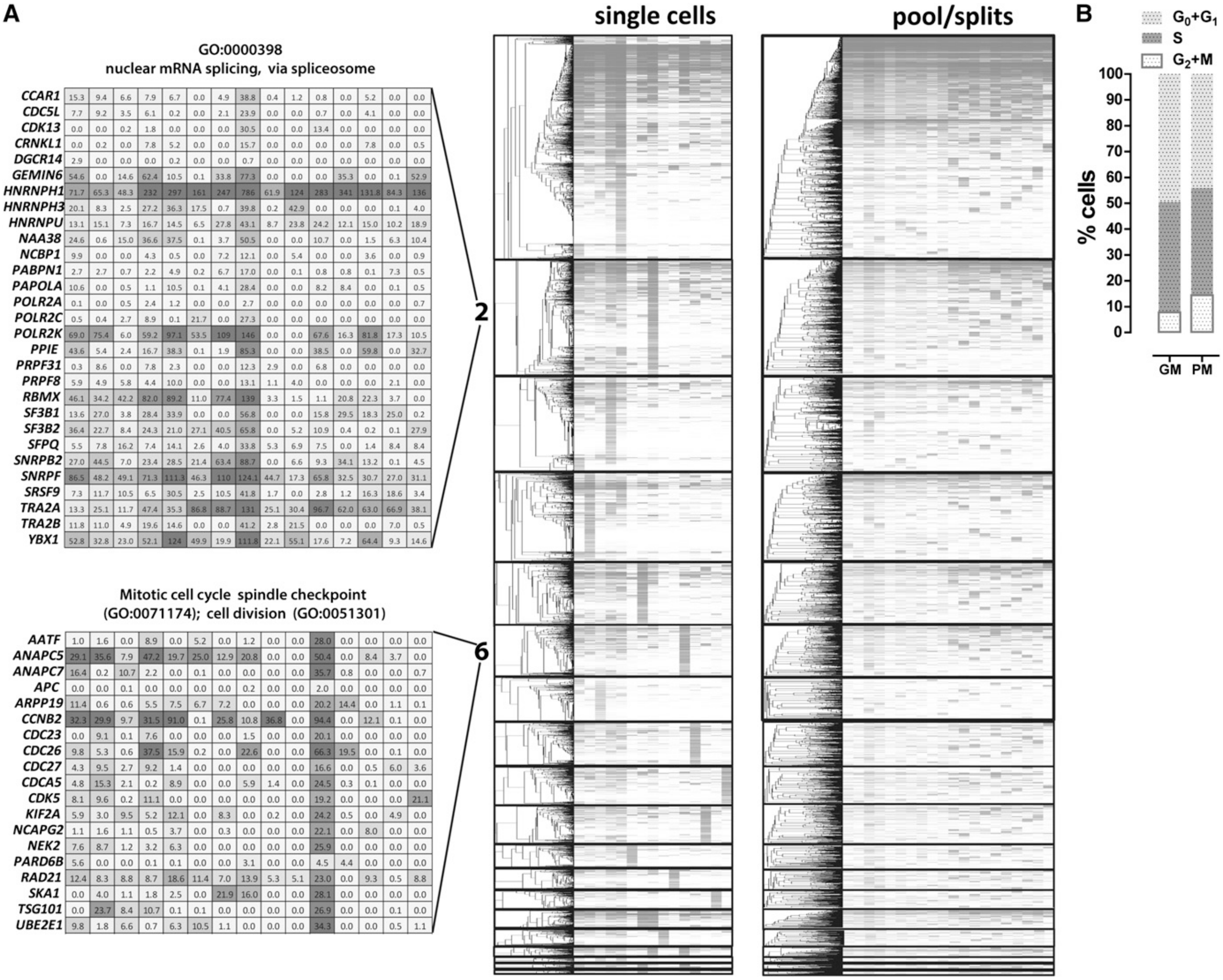
In addition to the coexpression analysis, we also examined the relationship between the expression variability of genes and various genomic data about their promoters, including long-range chromatin interactions, DNA methylation status, histone marks, transcription start site sequence elements, and CpG islands. No robust explanatory correlation was evident (Supplemental Figs. 46–-50), and we expect that data with less technical stochasticity will be needed to illuminate relationships of this kind.

Allele-biased expression at the single-cell level

Allele-specific gene expression (either monoallelic or highly biased toward one autosomal allele) has been previously reported to be widespread (Gimelbrant et al. 2007; Zhang and Borevitz 2009; McManus et al. 2010; Pickrell et al. 2010; Rozowsky et al. 2011; Reddy et al. 2012). An intriguing phenomenon observed for hundreds of genes in clonal lymphoblastoid cell lines (Gimelbrant et al. 2007; Chess 2012) is the random monoallelic expression of autosomal genes. However, those studies were conducted on large pools of cells, producing a snapshot of average allelic bias in the population, and leaving open the possibility that monoallelic expression is even more widespread on the single-cell level.

GM12878 cells are a good system for addressing this issue, as the fully phased heterozygous NA12878 genome sequence is available (The 1000 Genomes Project Consortium 2012). We aligned

Figure 3. Absolute expression levels at the single-cell level. FPKM values converted to estimated copies per cell using the spike-in quantification standards are shown. (*A*) Distribution of expression levels of RefSeq protein-coding genes in estimated copies per cell in single cells and pool/split experiments. (*B*) Distribution of expression levels of GENCODE v13 lncRNA protein-coding genes in estimated copies per cell in single cells and pool/split experiments. (*C*) Total number of mRNA copies per cell in single cells. (*D*) Total number of mRNA copies in pool/split experiments. (*E*) Expression levels of housekeeping and highly expressed genes (*GAPDH*, *CD74*, *left* panel), and general (*CTCF*, *REST*, *YY1*) and B-cell regulatory (*PAX5*, *EBF1*, *BCL11A*, *ETS1*, *IRF4*, *IKZF1*, *PBX3*, *POU2F2*, *RUNX3*, *TCF3*, *TCF12*) transcription factors (*right* panel). *Upper* and *middle* panels show the estimated copies-per-cell numbers for single cells and pool/splits, respectively. The *lower* panel shows FPKM values for cell pools and bulk RNA libraries. (*F–H*) Distribution of absolute expression levels in copies per cell in single cells for translation initiation, elongation, and termination proteins (*F*), splicing regulators (*G*), and transcription factors (*H*). The list of translation proteins was retrieved from the corresponding GO category annotations downloaded from FuncAssociate 2.0 (Berriz et al. 2009). The list of splicing regulators was obtained from the SpliceAid-F database of human splicing factors (Giulietti et al. 2013). The list of transcription factors used was the one from Vaquerizas et al. (2009). Note that only values ≥0.1 estimated copies per cell were included in these plots, i.e., libraries in which the genes were not detected were excluded.



GO:0000398
nuclear mRNA splicing, via spliceosome

CCAR1	15.3	9.4	6.6	7.9	6.7	0.0	4.9	38.8	0.4	1.2	0.8	0.0	5.2	0.0	0.0
CDC5L	7.7	9.2	3.5	6.1	0.2	0.0	2.1	23.9	0.0	0.0	0.7	0.0	4.1	0.0	0.0
CDK13	0.0	0.0	0.2	1.8	0.0	0.0	0.0	30.5	0.0	0.0	13.4	0.0	0.0	0.0	0.0
CRNKL1	0.0	0.2	0.0	7.8	5.2	0.0	0.0	15.7	0.0	0.0	0.0	0.0	7.8	0.0	0.5
DGCR14	2.9	0.0	0.0	0.0	0.2	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GEMIN6	54.6	0.0	14.6	62.4	10.5	0.1	33.8	77.3	0.0	0.0	0.0	35.3	0.0	0.1	52.9
HNRNPH1	71.7	65.3	48.3	232	297	161	247	786	61.9	124	283	341	131.8	84.3	136
HNRNPH3	20.1	8.3	2.5	27.2	36.3	17.5	0.7	39.8	0.2	42.9	0.0	0.0	0.0	0.1	4.0
HNRNPU	13.1	15.1	7.3	16.7	14.5	6.5	27.8	43.1	8.7	23.8	24.2	12.1	15.0	10.2	18.9
NAA38	24.6	0.6	15.0	36.6	37.5	0.1	3.7	50.5	0.0	0.0	10.7	0.0	1.5	6.3	10.4
NCBP1	9.9	0.0	0.0	4.3	0.5	0.0	7.2	12.1	0.0	5.4	0.0	0.0	3.6	0.0	0.9
PABPN1	2.7	2.7	0.7	2.2	4.9	0.2	6.7	17.0	0.0	0.1	0.8	0.8	0.1	7.3	0.5
PAPOLA	10.6	0.0	0.5	1.1	10.5	0.1	4.1	28.4	0.0	0.0	8.2	8.4	0.0	0.1	0.5
POLR2A	0.1	0.0	0.5	2.4	1.2	0.0	0.0	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.7
POLR2C	0.5	0.4	2.7	8.9	0.1	21.7	0.0	27.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
POLR2K	69.0	75.4	6.0	59.2	97.1	53.5	109	146	0.0	0.0	67.6	16.3	81.8	17.3	10.5
PPIE	43.6	5.4	2.4	16.7	38.3	0.1	1.9	85.3	0.0	0.0	38.5	0.0	59.8	0.0	32.7
PRPF31	0.3	8.6	0.0	7.8	2.3	0.0	0.0	12.3	2.9	0.0	6.8	0.0	0.0	0.0	0.0
PRPF8	5.9	4.9	5.8	4.4	10.0	0.0	0.0	13.1	1.1	4.0	0.0	0.0	0.0	2.1	0.0
RBMX	46.1	34.2	42.2	82.0	89.2	11.0	77.4	139	3.3	1.5	1.1	20.8	22.3	3.7	0.0
SF3B1	13.6	27.0	3.8	28.4	33.9	0.0	0.0	56.8	0.0	0.0	15.8	29.5	18.3	25.0	0.2
SF3B2	36.4	22.7	8.4	24.3	21.0	27.1	40.5	65.8	0.0	5.2	10.9	0.4	0.2	0.1	27.9
SFPQ	5.5	7.8	16.2	7.4	14.1	2.6	4.0	33.8	5.3	6.9	7.5	0.0	1.4	8.4	8.4
SNRPB2	27.0	44.5	7.0	23.4	28.5	21.4	63.4	88.7	0.0	6.6	9.3	34.1	13.2	0.1	4.5
SNRPF	86.5	48.2	49.1	71.3	111.3	46.3	110	124.1	44.7	17.3	65.8	32.5	30.7	27.0	31.1
SRSF9	7.3	11.7	10.5	6.5	30.5	2.5	10.5	41.8	1.7	0.0	2.8	1.2	16.3	18.6	3.4
TRA2A	13.3	25.1	11.7	47.4	35.3	86.8	88.7	131	25.1	30.4	96.7	62.0	63.0	66.9	38.1
TRA2B	11.8	11.0	4.9	19.6	14.6	0.0	0.0	41.2	2.8	21.5	0.0	0.0	0.0	7.0	0.5
YBX1	52.8	32.8	23.0	52.1	124	49.9	19.9	111.8	22.1	55.1	17.6	7.2	64.4	9.3	14.6

Mitotic cell cycle spindle checkpoint
(GO:0071174); cell division (GO:0051301)

AATF	1.0	1.6	0.0	8.9	0.0	5.2	0.0	1.2	0.0	0.0	28.0	0.0	0.0	0.0	0.0
ANAPC5	29.1	35.6	7.9	47.2	19.7	25.0	12.9	20.8	0.0	0.0	50.4	0.0	8.4	3.7	0.0
ANAPC7	16.4	0.2	10.7	2.2	0.0	0.1	0.0	0.0	0.0	0.0	35.7	0.8	0.0	0.0	0.7
APC	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	2.0	0.0	0.0	0.0	0.0
ARPP19	11.4	0.6	0.6	5.5	7.5	6.7	7.2	0.0	0.0	0.0	20.2	14.4	0.0	1.1	0.1
CCNB2	32.3	29.9	9.7	31.5	91.0	0.1	25.8	10.8	36.8	0.0	94.4	0.0	12.1	0.1	0.0
CDC23	0.0	9.1	0.1	7.6	0.0	0.0	0.0	1.5	0.0	0.0	20.1	0.0	0.0	0.0	0.0
CDC26	9.8	5.3	0.6	37.5	15.9	0.2	0.0	22.6	0.0	0.0	66.3	19.5	0.0	0.1	0.0
CDC27	4.3	9.5	2.7	9.2	1.4	0.0	0.0	0.0	0.0	0.0	16.6	0.0	0.5	6.0	3.6
CDCA5	4.8	15.3	2.1	0.2	8.9	0.0	0.0	5.9	1.4	0.0	24.5	0.3	0.1	0.0	0.0
CDK5	8.1	9.6	0.2	11.1	0.0	0.0	0.0	0.0	0.0	0.0	19.2	0.0	0.0	0.0	21.1
KIF2A	5.9	3.0	9.5	5.2	12.1	0.0	8.3	0.0	0.2	0.0	24.2	0.5	0.0	4.9	0.0
NCAPG2	1.1	1.6	1.1	0.5	3.7	0.0	0.3	0.0	0.0	0.0	22.1	0.0	8.0	0.0	0.0
NEK2	7.6	8.7	1.2	3.2	6.3	0.0	0.0	0.0	0.0	0.0	25.9	0.0	0.0	0.0	0.0
PARD6B	5.6	0.0	0.0	0.1	0.1	0.0	0.0	3.1	0.0	0.0	4.5	4.4	0.0	0.0	0.0
RAD21	12.4	8.3	8.8	8.7	18.6	11.4	7.0	13.9	5.3	5.1	23.0	0.0	9.3	0.5	8.8
SKA1	0.0	4.0	1.1	1.8	2.5	0.0	21.9	16.0	0.0	0.0	28.1	0.0	0.0	0.0	0.0
TSG101	0.0	23.7	8.4	10.7	0.1	0.1	0.0	0.0	0.0	0.0	26.9	0.0	0.0	0.1	0.0
UBE2E1	9.8	1.8	6.6	0.7	6.3	10.5	1.1	0.0	0.0	0.0	34.3	0.0	0.0	0.5	1.1

Figure 4. Gene coexpression modules derived from single GM12878 cells. Weighted gene correlation networks were constructed using the WCGNA R package (Langfelder and Horvath 2008). (*A*) Expression levels and hierarchical clustering of genes within modules (modules are sorted by number, which corresponds to their size) in single cells and pool/split experiments. Only genes are clustered (dendrograms on the *left*), and the identity of the cells and pool/split experiments is the same in each column (two *right* panels). The absolute expression values of genes belonging to representative GO categories associated with cell cycle phases (modules 1 and 6) and mRNA processing and splicing (module 2) are also shown. (*B*) Distribution of cell cycle states in a representative GM12878 cell population, in growth media (GM), and picking media (PM). The fraction of cells in M phase is consistent with one such cell being picked in a sample of 15.

RNA-seq reads in an allele-specific manner to the heterozygous GM12878 transcriptome and calculated allelic bias for each gene as the fraction of reads mapping to the maternal allele. As detailed in the Methods and Supplemental Methods, we applied very stringent criteria for determining statistically significant allele-biased expression events based on the absolute transcript number estimates and taking into account the challenges presented by the nature of single-cell RNA-seq data. We observed good reproducibility of allelic bias profiles in 10-ng bulk RNA libraries (Supplemental Fig. 37A), with most genes being expressed from both alleles (Supplemental Fig. 37D). Allelic bias was also highly reproducible in 30-cell and 100-cell pools (Supplemental Fig. 51). In contrast, allelic bias profiles of single cells correlated poorly with each other, and a large fraction of genes were apparently monoallelically expressed from different alleles in different cells (Supplemental Fig. 37B). The majority of highly expressed genes (≥100 copies per cell) exhibited biallelic expression, while most genes at low expression levels were measured as monoallelically expressed (Supplemental Fig. 37F). We then compared allelic bias variability for individual genes across individual single cells, focusing only on cells in which statistically significant allelic bias was observed, and observed frequent "switching" between the two alleles (Supplemental Figs. 37G, 38A).

These observations can be explained as a combination of biological and technical factors. First, it has been previously reported that allelic bias at the population level is more common among genes expressed at low levels (Gimelbrant et al. 2007, Reddy et al. 2012). A second explanation is the phenomenon of "transcriptional bursting" (Raj and van Oudenaarden 2008; Dar et al. 2012). A single transcription burst produces several mRNA molecules from a single allele. If all mRNAs from a gene in a given cell at a given moment are the product of one or a linked series of such bursts, all

Table 1. **Representative Gene Ontology categories enriched in coexpressed gene modules**

Adjusted *P*-value	GO attrib ID	Attrib name
Module 1		
<0.001	GO:0006415	Translational termination
<0.001	GO:0006414	Translational elongation
<0.001	GO:0070469	Respiratory chain
<0.001	GO:0071845	Cellular component disassembly at cellular level
<0.001	GO:0004129	Cytochrome-c oxidase activity
<0.001	GO:0022904	Respiratory electron transport chain
<0.001	GO:0030964	NADH dehydrogenase complex
<0.001	GO:0072413	Signal transduction involved in mitotic cell cycle checkpoint
0.019	GO:0006626	Protein targeting to mitochondrion
<0.001	GO:0048002	Antigen processing and presentation of peptide antigen
<0.001	GO:0010467	Gene expression
<0.001	GO:0006839	Mitochondrial transport
0.007	GO:0006458	De novo protein folding
<0.001	GO:0016071	mRNA metabolic process
<0.001	GO:0000216	M/G1 transition of mitotic cell cycle
0.014	GO:0000502	Proteasome complex
0.005	GO:0060333	Interferon-gamma-mediated signaling pathway
<0.001	GO:0000084	S phase of mitotic cell cycle
<0.001	GO:0000082	G1/S transition of mitotic cell cycle
0.005	GO:0000209	Protein polyubiquitination
<0.001	GO:0008380	RNA splicing
Module 2		
<0.001	GO:0000398	Nuclear mRNA splicing, via spliceosome
0.017	GO:0005681	Spliceosomal complex
<0.001	GO:0006397	mRNA processing
Module 3		
<0.001	GO:0051186	Cofactor metabolic process
0.002	GO:0051539	Four iron, four sulfur cluster binding
0.021	GO:0051536	Iron-sulfur cluster binding
Module 6		
0.027	GO:0005680	Anaphase-promoting complex
0.001	GO:0007094	Mitotic cell cycle spindle assembly checkpoint

Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al. 2009). The full list of enriched categories is available in Supplemental Table 3.

copies would originate from one allele. Finally, stochastic effects due to the low single-molecule capture efficiency of the protocol undoubtedly play a role. The fewer founder molecules are captured, the more likely it is that they come from only one allele. To help parse these sources of variation, we performed the same analyses on pool/split libraries and observed a broadly similar (although always lower) fraction of genes passing all significance tests for allelic bias (Supplemental Figs. 37C,E, 38). The quantitative trend within the pool/split comparison suggests there is a component of allelic RNA bias between cells that is biological in origin but that there is also a large technical variation component. The widespread occurrence of random monoallelic expression at the single-cell level should therefore be viewed as a provisional conclusion.

Alternative splicing at the single-cell level

Previous studies have suggested that most genes in mammalian genomes undergo some alternative splicing (Mortazavi et al. 2008; Wang et al. 2008; Djebali et al. 2012). At present, however, the biological relevance of the majority of these alternative isoforms is still uncertain, and stochastic noise in the splicing machinery is one explanation (Sorek et al. 2004; Melamud and Moult 2009). Characterizing alternative splicing at the single-cell level should bring clarity to the population-based observations, and perhaps offer clues about the mechanistic origin of the multiple isoforms observed within cell types.

We quantified alternative splicing using the intron-centric splice inclusion ψ score approach (Pervouchine et al. 2013). Details of our mapping and analysis pipeline are described in the Supplemental Methods. For reasons given there, we adopted a conservative approach and only analyzed novel splice junctions for which at least one of the donor or acceptor sites has already been annotated in GENCODE v13 (Harrow et al. 2012), thus avoiding library-building artifacts.

We detected between 200 and 2000 novel splice junctions satisfying these criteria in each individual cell (Supplemental Fig. 43). This number is certainly an underestimate, given the low p_{smc}. About 35% of novel junctions connected two annotated exons (Fig. 5A); most of these represent novel exon skipping events. In another 60%, the unannotated donor or acceptor site was internal to the gene. These were concentrated close to already annotated splice sites (Supplemental Fig. 40B,C). In particular, novel acceptor sites peaked at the +3 and −3 position downstream from annotated sites representing mostly instances of NAGNAG tandem acceptor sites (Hiller et al. 2004; Bradley et al. 2012). Novel 5′ donor sites were fewer in number and peaked at +4 and −4 positions relative to annotated donor sites, thus shifting the coding frame of the transcript. This is a phenomenon we have previously also observed in bulk RNA-seq data (observations of the present study's authors), the significance of which is at present not clear. The proportions observed were independent of the read coverage and estimated number of copies per cell thresholds applied (Supplemental Fig. 40A).

We also examined the distribution of unannotated splices across individual single cells and found that the majority were detected in only a single cell, with <10% found in two cells, and very few in three or more cells (Fig. 5B). While this result could be greatly affected by p_{smc} issues, it was independent of the read and estimated transcript copies threshold used (Supplemental Fig. 40), suggesting that most novel splices are indeed only present in a small fraction of cells.

We asked how often multiple alternative splice sites are used at individual single cells. In bulk RNA-seq at a threshold of 15 distinct read fragments, a numeric minority of ψ scores was equal to 1 (i.e., exclusive use of only one donor-acceptor pair). The presence of alternative splice sites is thus widespread in the cell population. Nevertheless, in most cases, ψ was close to 1, suggesting quantitative dominance of one isoform. The vast majority of novel splices received very low inclusion scores (Fig. 5C) and would generally be considered to be the result of biological noise in the splicing system. In contrast, in single cells, one dominant splice site was the norm for annotated junctions, except for very highly expressed genes (≥100 copies per cell), for which a wide diversity of splice site usage was seen (Fig. 5D; details in Supplemental Fig. 42). As this observation was true even for genes expressed at ≥50 copies per cell, we believe it is not a p_{smc} artifact. It is an interesting and open question why very highly expressed genes (enriched for genes with housekeeping function) exhibit very high levels of alternative splicing in single cells. These results differ significantly from the same analysis carried out on novel splice junctions (Fig. 5E; Supplemental Fig. 43). Somewhat surprisingly, we found that

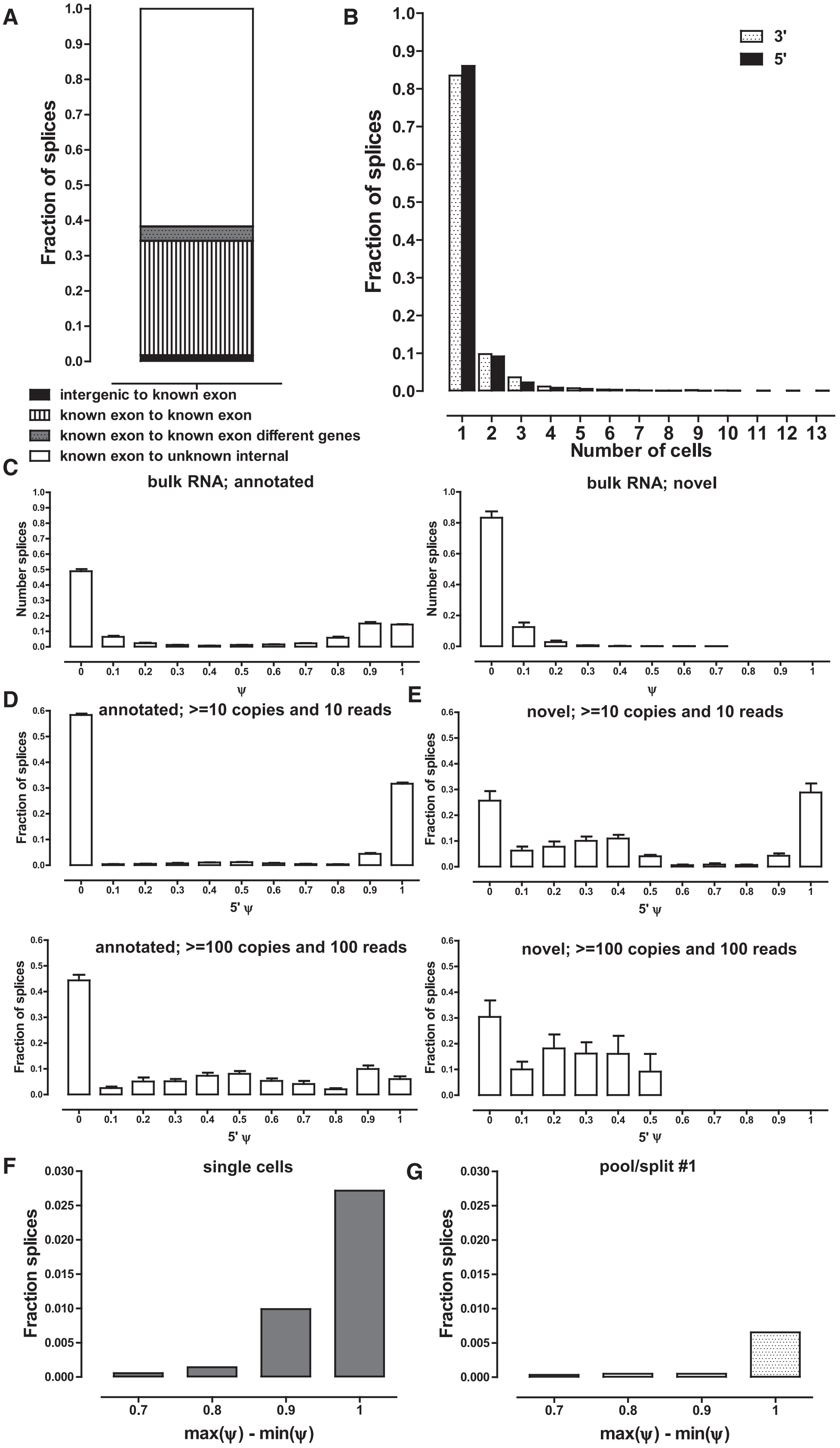


Figure 5. (Legend on next page)

a significant proportion of novel splices had ψ scores of 1 in single cells. This was true, however, only for genes expressed at lower levels (≤50 copies), where p_{smc} artifacts are a likely cause. In contrast, in highly expressed genes, no novel junctions received a dominant (≥0.5) ψ score. However, the scores were still consistently higher than what is observed for novel splices in bulk RNA samples.

Finally, we evaluated the consistency of splice site usage between individual cells. We applied a statistical framework similar to the one used to analyze allelic bias and derived a list of dominant splice junctions in each cell, taking into account the estimated absolute number of copies and the stochastic capture effects. We asked how often the dominant splice site changes between different cells. We found 282 such genes in single cells, suggesting the phenomenon may be widespread. The genes involved were enriched for ribosomal and translation proteins, and also, intriguingly, for proteins involved in RNA splicing and processing (Supplemental Table 6). We tested this single-cell variation against pool/split experiments, in which we found very few genes with different dominant splice sites across libraries. (Fig. 5F,G; Supplemental Fig. 44). This argues that much alternative splicing variation is in fact due to biological differences between cells, and is in agreement with the bimodality of splicing in individual mouse immune cells reported recently (Shalek et al. 2013).

Discussion

Two major goals for single-cell RNA-seq are to obtain high-resolution transcriptomes for rare cell types or states and to measure the differences in RNA expression and processing between individual cells. Here, we showed that the first goal can be achieved by studying 30- to 100-cell pool samples even in the absence of perfect capture of each and every individual RNA molecule. Our conclusion is consistent with independent 80-cell measurements (Ramsköld et al. 2012). The pools reproduce the expression profiles (Supplemental Fig. 53) and allelic-bias patterns (Supplemental Fig. 51) of the larger population, and similar numbers of genes and splice junctions are detected in them (Supplemental Figs. 52, 53). The approach is applicable to cells collected by laser-capture (to be presented elsewhere), micromanipulation (used here), or cell sorting based on molecular markers or reporter-gene expression. This defines a general and relatively economical path forward for the transcriptomic characterization of many previously inaccessible rare cell types and states, including transient cell types in embryonic development, diverse neuronal types in the brain, and cells in tumors.

In agreement with previous single-cell RNA-seq studies, we observed high cell-to-cell variability in gene expression levels in GM12878 B-cells. We found that some of this variation was attributable to coordinated differences in the expression of biologically coherent sets of genes: for example, genes associated with the M phase of the cell cycle or with mRNA processing and splicing.

Despite good data quality, evidenced by complete and relatively uniform coverage across the mRNA length spectrum, our results were similar to other published data in displaying significant stochasticity. Stochasticity is expected to arise from a combination of biological variation and technical measurement variation. We present experimental and analytical approaches for measuring and accounting for technical stochasticity. We introduced and measured single-molecule capture efficiency, the key parameter influencing technical stochasticity, and found that its value was around 0.1 with the current SMART-seq protocol. This low capture efficiency provides a parsimonious explanation for the level of variation between single-cell measurements that is technical in origin. We also measured technical variation by carrying out pool/split experiments. This empirical test for nonbiological variation in the system is a stringent one, which includes capture efficiency, PCR effects, and any other unspecified sources. We then used the pool/split results to help parse biological variation between cells that is detectable over and above variation in pool/split measurements.

We observed unexpected levels of cell-to-cell variation in autosomal allelic expression bias and alternative splicing. The observation of allele switching between single cells could be explained as a technical artifact, given that a similar, although always lower, level of switching was observed in pool/split libraries. We therefore consider this a provisional result in need of further investigation with improved experimental protocols. The observed frequency of major splice switching in single cells is a stronger effect, and based on comparison with pool/split experiments, it is unlikely to be the sole result of technical stochasticity. It has also been independently reported in a different system (Shalek et al. 2013).

Transcriptional bursting suggests an attractive biological explanation for these observations. If a gene is expressed in a series of infrequent (relative to the half life of its mRNAs) such bursts, at any given time the population of mRNAs in the cell is likely to originate from only one allele. Such bursting could also be the source of cell-to-cell variation in alternative splicing. It is possible that the same set of factors influencing splice-site choice maintain physical association with the gene during a transcriptional burst, leading to a particular splicing pattern being highly favored locally in space and time, even if factors supporting a different splice choice are present within the same nucleus. Alternatively, isoform choice could be driven by temporal switching of factors and would operate regardless of bursting behavior. These are testable alternatives for future studies.

Many specific biological processes, especially regulatory ones, involve genes whose transcript levels are in the range highly affected by technical variation, as shown by our survey of transcription factors. While measurements with current methods can give some important clues about coherent biological variation, especially when large numbers of individual cells are assayed, our results argue that considerable improvement in the single-

Figure 5. Alternative splicing at the single-cell level. (*A*) Classification of new junctions connecting known splice sites. (*B*) Frequency of detection of novel splice junctions. Novel junctions for which neither the donor nor acceptor site has been annotated were excluded for reasons described in the main text in both *A* and *B*. A threshold of 10 estimated copies and a coverage of 10 reads was applied, but results are essentially the same, independent of the thresholds used (Supplemental Fig. 40A). (*C*) Distribution of ψ scores in bulk RNA samples for annotated and novel splice junctions. A threshold of 15 reads combined for all splice junctions in which a donor or acceptor site participates was applied. Note that for each ψ_1 score there is at least one matching $\psi_2 \leq 1 - \psi_1$ score corresponding to the other alternative junction; in some cases, more than two alternative donor or acceptor sites exist; thus the relative height of the $0 \leq \psi \leq 0.1$ bar. (*D*, *upper* and *lower*). Distribution of 5′ ψ scores for annotated splice junctions at two different detection thresholds in single-cell libraries (see Supplemental Fig. 41 for more detail). (*E*, *upper* and *lower*) Distribution of 5′ ψ scores for novel splice junctions at two different detection thresholds in single-cell libraries (see Supplemental Fig. 42 for more detail). (*F*, *G*) Frequency of major splice site usage switches between individual cells (*F*) and individual libraries in a pool/split experiment (*G*). Note the strong support for major splice site use switching across the collection of single cells.

molecule capture efficiency would profoundly advance the field. Based on our simulations and results from pool/split experiments, we estimate that an increase in p_{smc} from 0.1 to 0.5 would be a major leap forward, while $p_{smc} \geq 0.8$ would provide sufficient reliability for virtually any biological use. We anticipate that this empirical and analytical framework will be useful for evaluating future improvements in protocols, such as the recently described SMART-seq2 protocol (Picelli et al. 2013).

Finally, we found that the amount of mRNA per cell is highly variable between individual cells. Beyond biological interest, these differences in mRNA number are important for analysis pipelines. RPKM-type metrics are not reliable when there are large differences in total RNA per cell (Lin et al. 2012; Lovén et al. 2012). At present, the direct relationship between the absolute number of mRNA copies per cell and the number of sequencing reads in a library is lost due to the fragmentation of amplified cDNA molecules that is a common feature of available protocols, resulting in multiple distinct but overlapping sequencing fragments for each founder RNA molecule. mRNA copy number can be estimated back from FPKMs with the help of spike-in sequences. However, this is far from ideal, as it depends on the accuracy of quantification of the spike-ins and assumes the absence of systemic differences between spike-in RNAs and endogenous RNAs. If these assumptions are wrong, we expect a systematic error in the calculated number of mRNAs per cell, although the more interesting and important differences between individual cells versus pool/splits would remain. The above considerations make it very clear that a future ideal single-cell RNA-seq assay would combine a very high single-molecule capture efficiency with an amplification-free, and preferably also reverse transcription-free, direct RNA sequencing method to achieve direct counting of intact transcripts. Emerging sequencing technologies (Branton et al. 2008; Schadt et al. 2010) already hold promise for such radical improvements.

Methods

Cell growth and single-cell RNA-seq library construction

Individual GM12878 cells grown according to standard ENCODE protocols were picked with a glass micropipette, deposited into lysis buffer, and frozen. Cells were later lysed in reaction buffer, and single-cell SMART cDNA was generated following the SMART-seq protocol (Ramsköld et al. 2012) with the following modifications: (1) Carrier yeast tRNA was added in the lysis buffer to reduce handling losses and help maintain the integrity of the mRNA; (2) spikes of known copy number were introduced; and (3) the PCR cycle number was empirically titrated to accommodate the relatively small GM12878 cells. The SMART cDNA was tagmented using Illumina/Nextera reagents as described in Gertz et al. (2012). A detailed description of experimental protocols is provided in the Supplemental Methods.

Sequence alignment and gene expression quantification

Reads were aligned against a combined Bowtie index of the human genome and spike-in sequences using TopHat (Trapnell et al. 2009, 2012). Gene expression was quantified using Cufflinks (Trapnell et al. 2010, 2012). FPKMs were converted to copies-per-cell estimates using the input and measured spike-in abundances.

Single-molecule capture efficiency estimation

We estimated the average p_{smc} based on the number of libraries with 0 FPKM for each spike and the number of input molecules (accounting for the fact that the number of successful captures is not known but only the number of complete failures; a detailed description of the procedure is provided in the Supplemental Methods). The average p_{smc} for all spikes for which libraries with 0 FPKMs were observed was used, which is ~0.01.

Analysis of allele-biased expression

We used the diploid (May 2011 release) NA12878 genome containing phased SNPs and indels based on the NCBI build 36 (hg18) version of the human genome (downloaded from http://sv.gersteinlab.org/NA12878_diploid/). Heterozygous transcriptomes containing two copies of each transcript were built, and reads were aligned using Bowtie (Langmead et al. 2009) (version 0.12.7) with zero mismatches allowed. Identical reads were collapsed, and reads were assigned to an allele if they mapped only to one of the alleles of a gene. Allele-biased expression was assessed by accounting for all of the following: (1) significance of allelic bias on the level of reads; (2) significance of allelic bias on the level of estimated copies per cell for each allele (derived from the total number of estimated copies for the gene); this is necessary, as a common feature of all current single-cell protocols is the production of multiple overlapping fragments from each original molecule; and (3) the possibility that the observed allelic bias is due to differential stochastic capture of the two alleles. A detailed description of the procedure is provided in the Supplemental Methods.

Alternative splicing analysis

We carried out alternative splicing analysis using the 5′ and 3′ splicing inclusion ψ scores described by Pervouchine et al. (2013), and applying the same statistical procedure we used to assess allelic expression bias to determine statistically significant splice variant exclusion. A detailed description of the splicing analysis procedure is provided in the Supplemental Methods.

Gene expression clustering and weighted correlation network analysis

We used the WGCNA R package (Langfelder and Horvath 2008) to carry out the weighted correlation network analysis. Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al. 2009). Gene expression clustering was carried out using Cluster 3.0 (de Hoon et al. 2004) and visualized using TreeView (Saldanha 2004).

Data access

BAM files containing aligned and unaligned sequencing reads have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE44618.

Acknowledgments

We thank Henry Amrhein, Diane Trout, and Sean Upchurch for computational assistance, and members of the Wold laboratory for helpful discussions. This work has been supported by NIH grants U54 HG004576 and U54 HG006998, the Simons Foundation, and the McDonnell Foundation. In addition, G.K.M., B.A.W., and B.J.W. are supported by the Beckman Foundation and the Donald Bren Endowment.

References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106.

Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. 2009. Next generation software for functional trend analysis. *Bioinformatics* **25:** 3043–3044.

Blake WJ, Kaern M, Cantor CR, Collins JJ. 2003. Noise in eukaryotic gene expression. *Nature* **422:** 633–637.

Bradley RK, Merkin J, Lambert NJ, Burge CB. 2012. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol* **10:** e1001229.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26:** 1146–1153.

Brouilette S, Kuersten S, Mein C, Bozek M, Terry A, Dias KR, Bhaw-Rosun L, Shintani Y, Coppen S, Ikebe C, et al. 2012. A simple and novel method for RNA-seq library preparation of single cell cDNA analysis by hyperactive Tn5 transposase. *Dev Dyn* **241:** 1584–1590.

Cann GM, Gulzar ZG, Cooper S, Li R, Luo S, Tat M, Stuart S, Schroth G, Srinivas S, Ronaghi M, et al. 2012. mRNA-Seq of single prostate cancer circulating tumor cells reveals recapitulation of gene expression and pathways found in prostate cancer. *PLoS ONE* **7:** e49144.

Chess A. 2012. Mechanisms and consequences of widespread random monoallelic expression. *Nat Rev Genet* **13:** 421–428.

Cornelison DD, Wold BJ. 1997. Single-cell analysis of regulatory gene expression in quiescent and activated mouse skeletal muscle satellite cells. *Dev Biol* **191:** 270–283.

Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, Simpson ML, Weinberger LS. 2012. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc Natl Acad Sci* **109:** 17454–17459.

de Hoon MJ, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. *Bioinformatics* **20:** 1453–1454.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108.

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic gene expression in a single cell. *Science* **297:** 1183–1186.

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Femino AM, Fay FS, Fogarty K, Singer RH. 1998. Visualization of single RNA transcripts in situ. *Science* **280:** 585–590.

Gertz J, Varley KE, Davis NS, Baas BJ, Goryshin IY, Vaidyanathan R, Kuersten S, Myers RM. 2012. Transposase mediated construction of RNA-seq libraries. *Genome Res* **22:** 134–141.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318:** 1136–1140.

Giulietti M, Piva F, D'Antonio M, D'Onorio De Meo P, Paoletti D, Castrignanò T, D'Erchia AM, Picardi E, Zambelli F, Principato G, et al. 2013. SpliceAid-F: A database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res* **41:** D125–D131.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.

Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2:** 666–673.

Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. 2011. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7:** 497.

Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M. 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet* **36:** 1255–1257.

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21:** 1160–1167.

Kaufmann BB, van Oudenaarden A. 2007. Stochastic gene expression: From single molecules to the proteome. *Curr Opin Genet Dev* **17:** 107–112.

Langfelder P, Horvath S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9:** 559.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA. 2012. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151:** 56–67.

Livak KJ, Wills QF, Tipping AJ, Datta K, Mittal R, Goldson AJ, Sexton DW, Holmes CC. 2013. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* **59:** 71–79.

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. 2012. Revisiting global gene expression analysis. *Cell* **151:** 476–482.

Lubeck E, Cai L. 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* **9:** 743–748.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res* **20:** 816–825.

Melamud E, Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res* **37:** 4873–4886.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. 2002. Regulation of noise in the expression of a single gene. *Nat Genet* **31:** 69–73.

Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, Milos PM. 2010. Digital transcriptome profiling from attomole-level RNA samples. *Genome Res* **20:** 519–525.

Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, Marjani SL, Euskirchen G, Ma C, Lamotte RH, et al. 2012. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci* **110:** 594–599.

Pervouchine DD, Knowles DG, Guigó R. 2013. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29:** 273–274.

Picelli S, Björklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* doi: 10.1038/nmeth.2639.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.

Qiu S, Luo S, Evgrafov O, Li R, Schroth GP, Levitt P, Knowles JA, Wang K. 2012. Single-neuron RNA-Seq: Technical feasibility and reproducibility. *Front Genet.* **3:** 124.

Raj A, van Oudenaarden A. 2008. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135:** 216–226.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5:** 877–879.

Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5:** e1000598.

Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30:** 777–782.

Raser JM, O'Shea EK. 2005. Noise in gene expression: Origins, consequences, and control. *Science* **309:** 2010–2013.

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22:** 860–869.

Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al. 2011. AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7:** 522.

Saldanha AJ. 2004. Java Treeview–extensible visualization of microarray data. *Bioinformatics* **20:** 3246–3248.

Sanchez-Freire V, Ebert AD, Kalisky T, Quake SR, Wu JC. 2012. Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat Protoc* **7:** 829–838.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19:** R227–R240.

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498:** 236–240.

Sorek R, Shamir R, Ast G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet* **20:** 68–71.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6:** 377–382.

Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, Lao K, Surani MA. 2010. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* **5:** 516–535.

Tang F, Lao K, Surani MA. 2011. Development and applications of single-cell transcriptome analysis. *Nat Methods* (Suppl) **8:** S6–S11.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31:** 46–53.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet* **10:** 252–263.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, Hansen CL. 2011. High-throughput microfluidic single-cell RT-qPCR. *Proc Natl Acad Sci* **108:** 13999–14004.

Zhang X, Borevitz JO. 2009. Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* **182:** 943–954.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4:** 17.

Zenklusen D, Larson DR, Singer RH. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* **15:** 1263–1271.

Received May 24, 2013; accepted in revised form November 20, 2013.

K

Large-scale quality analysis of published ChIP-seq data

Originally published as:

Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-scale quality analysis of published ChIP-seq data. *G3.* 2014. **4**(2):209-223. doi: 10.1534/g3.113.008680.

Large-Scale Quality Analysis of Published ChIP-seq Data

Georgi K. Marinov,* Anshul Kundaje,,††,[1] Peter J. Park,†,‡,§ and Barbara J. Wold*,[2]**
*Division of Biology, California Institute of Technology, Pasadena, California 91125, †Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, ‡Informatics Program, Children's Hospital Boston, Boston, Massachusetts 02115, §Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, **Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and ††The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142

ABSTRACT ChIP-seq has become the primary method for identifying *in vivo* protein–DNA interactions on a genome-wide scale, with nearly 800 publications involving the technique appearing in PubMed as of December 2012. Individually and in aggregate, these data are an important and information-rich resource. However, uncertainties about data quality confound their use by the wider research community. Recently, the Encyclopedia of DNA Elements (ENCODE) project developed and applied metrics to objectively measure ChIP-seq data quality. The ENCODE quality analysis was useful for flagging datasets for closer inspection, eliminating or replacing poor data, and for driving changes in experimental pipelines. There had been no similarly systematic quality analysis of the large and disparate body of published ChIP-seq profiles. Here, we report a uniform analysis of vertebrate transcription factor ChIP-seq datasets in the Gene Expression Omnibus (GEO) repository as of April 1, 2012. The majority (55%) of datasets scored as being highly successful, but a substantial minority (20%) were of apparently poor quality, and another ~25% were of intermediate quality. We discuss how different uses of ChIP-seq data are affected by specific aspects of data quality, and we highlight exceptional instances for which the metric values should not be taken at face value. Unexpectedly, we discovered that a significant subset of control datasets (*i.e.*, no immunoprecipitation and mock immunoprecipitation samples) display an enrichment structure similar to successful ChIP-seq data. This can, in turn, affect peak calling and data interpretation. Published datasets identified here as high-quality comprise a large group that users can draw on for large-scale integrated analysis. In the future, ChIP-seq quality assessment similar to that used here could guide experimentalists at early stages in a study, provide useful input in the publication process, and be used to stratify ChIP-seq data for different community-wide uses.

KEYWORDS
ChIP-seq
chromatin immunoprecipitation
cross-correlation
quality assessment
transcription factor

Chromatin immunoprecipitation (ChIP) (Gilmour and Lis 1984; Gilmour and Lis 1985; Solomon *et al.* 1988) experiments identify sites of occupancy by specific transcription factors (TFs), cofactors, and other chromatin-associated proteins as well as histone modifications. Such proteins are concentrated at specific loci via direct binding to DNA or by indirect binding mediated by other proteins or RNA molecules. In most ChIP protocols, proteins are first cross-linked to DNA, most often using formaldehyde. The fixed chromatin is sheared, and an antibody specific for the protein or histone modification of interest is used to retrieve protein:DNA complexes from which the DNA segments are released and then assayed. The assay was first applied to individual TF/promoter complexes by using qPCR to detect enrichment over specific DNA segments (Hecht *et al.* 1996). Subsequent adaptations extended it to large sets of promoters or other genomic regions by using microarrays (ChIP-on-Chip/ChIP-Chip) (Ren *et al.*

doi: 10.1534/g3.113.008680
Manuscript received September 29, 2013; accepted for publication November 21, 2013; published Early Online December 17, 2013.

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.113.008680/-/DC1

A global analysis of published ChIP-seq data addressing a subset of the quality control issues presented here was published while this paper was under final review (Mendoza-Parra *et al.* 2013).

[1] Present address: Department of Genetics, Department of Computer Science, Stanford University, Stanford, CA 94305.

[2] Corresponding author: Barbara J. Wold. Division of Biology, California Institute of Technology, 1200 E. California Boulevard, MC 156-29, Pasadena, CA 91125.

2000; Iyer *et al.* 2001; Lieb *et al.* 2001; Horak and Snyder 2002; Weinmann *et al.* 2002). Ultimately, the entire genome became accessible with the advent of high-throughput sequencing and the development of ChIP-seq (Johnson *et al.* 2007; Barski *et al.* 2007; Mikkelsen *et al.* 2007; Robertson *et al.* 2007).

In all cases, preferential enrichment of a given immunoprecipitated DNA segment is detected and quantified by comparing it with a control experiment in which there is no specific antibody enrichment step. These controls can be generated from sonicated DNA before immunoprecipitation (input) or a mock immunoprecipitation with an unrelated antibody (IgG). Sequencing-based ChIP has become the method of choice because it enables genome-wide coverage, even for large genomes, and because of its superior signal-to-noise characteristics compared to alternative methods. Since its initial development, ChIP-seq has been used in hundreds of publications (778 in PubMed as of December 18, 2012), including by the ENCODE consortium (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012), to map occupancy over 100 human TFs and cofactors in a diverse collection of cell lines (Gerstein *et al.* 2012; Wang *et al.* 2012).

A basic question for any ChIP-seq experiment is, how successful was it? It has taken several years for the field to develop objective ways to quantify key aspects of success in immunoprecipitation enrichment, library building, and final sequencing. Poor datasets that have high false-negative rates in peak calling are a predictable pitfall that has significant downstream consequences for some kinds of biological and computational analyses. For example, when lower-quality datasets are used for integrative analyses that are sensitive to false-negative rates, incorrect inferences and conclusions become likely (see *Discussion*). In estimating data quality, the traditional approach of visual inspection at a limited number of sites (often previously well-characterized using low-throughput approaches) is inefficient, subjective, and ultimately can be deceptive. It is also possible (and commonly observed in practice) that sites, the biological importance of which has been defined by independent functional assays, can decrease to below the sensitivity threshold of a poor or mediocre ChIP-seq experiment. Moreover, there is no current way to predict, *a priori*, the number of sites in the genome that should be detectable for a given factor and cell type. Most TFs studied thus far reproducibly occupy thousands to tens of thousands of sites (ENCODE Project Consortium 2012; Landt *et al.* 2012). Thus, a dataset for which several thousand sites have been called might in fact be capturing a minority of true positive interactions, or it might encompass virtually all biologically pertinent sites. To help address the problem of data assessment as part of the ENCODE project, we and others developed a set of ChIP-seq quality control (QC) metrics and guidelines (Landt *et al.* 2012) that were adopted and applied to all of its datasets. Substandard datasets were consequently replaced, flagged as substandard, and/or removed from analysis (ENCODE Project Consortium 2012; Landt *et al.* 2012).

Incorporating published datasets into an ongoing study can bring new biological insights and avoid unnecessary duplication of work. Variable quality of published data can be a significant barrier to these uses of existing data. They are the products of work from many different laboratories with invaluable expertise in specific biological systems, but they also use many variations of ChIP-seq experimental protocols and bioinformatics treatments. The extent and nature of the variations have not been assessed globally and systematically. In this work, we examined the GEO submission series containing vertebrate TF ChIP-seq datasets and found that ~20% of datasets scored as being of low quality, with an additional ~25% exhibiting intermediate ChIP enrichment. We also noticed that approximately one-third of studies have control datasets with a high degree of read clustering that is normally expected only in ChIP-seq datasets. This was observed more often for the IgG control design than for input DNA controls. These and related observations argue for data quality measures routine characterization and reporting of ChIP-seq data quality measures.

MATERIALS AND METHODS

Sequencing read alignment

Raw sequencing reads for all non-ENCODE GEO series containing ChIP-seq datasets against TFs and chromatin-modifying proteins (submitted before April 1, 2012) were downloaded from GEO in SRA format and converted to FASTQ format using the fastq-dump program in the sratoolkit (version 2.1.9). Reads were aligned using Bowtie (Langmead *et al.* 2009) version 0.12.7 with the following setting: "-v 2 -t -k 2 -m 1 –best–strata," which– allows for two mismatches relative to the reference and only retains unique alignments. Human datasets were mapped against the male set of chromosomes (excluding all random chromosomes and haplotypes) for version hg19 of the human genome; the mm9 version of the mouse genome was used for mouse data, rn5 was used for rat data, danRer7 was used for zebrafish data, susScr2 was used for pig data, and xenTro3 was used for the clawed frog *Xaenopus tropicalis* data, and all assemblies were downloaded from the UCSC genome browser (Kent at al. 2002).

ChIP quality assessment

ChIP quality assessment was performed on both ChIP and input datasets using the general strategy described by Landt *et al.* (2012). Because a library may score as an "unsuccessful ChIP" for reasons other than IP failure (e.g. being performed in a knockout background, in si/shRNA-treated cells, or in conditions under which the factor is not expressed or not bound to DNA), the following additional criteria were used to determine whether each library is expected to score positively in the QC assessment:

1. All experiments claimed to be successful by authors are expected to exhibit high level of read clustering.
2. All inputs (sonicated DNA and IgG mock IPs) are expected to exhibit minimal read clustering (QC tag of −2 or −1).
3. All ChIP-seq experiments performed in a knockout background for the factor are expected to exhibit minimal read clustering (QC tag of −2 or −1).
4. Because knockdown efficiency varies and because it is unknown what protein levels would be sufficiently high for the factor to be successfully ChIP-ed, ChIP-seq experiments performed in cells treated with si/shRNAs targeting the factor are set aside as "unknown" and assessed for library complexity and sequencing depth but not for ChIP quality.
5. Experiments against factors known to bind to DNA on some stimulus performed in unstimulated cells are also tagged as "unknown" because lower-level binding in unstimulated cells cannot be ruled out (and is, in fact, often observed).
6. Experiments performed in conditions that may result in the factor not binding to DNA (time courses, knockdowns, or knockouts for other factors that may affect binding of the targeted factor) are also tagged as "unknown."
7. Other experiments not matching any of these categories are expected to exhibit high levels of read clustering.

Cross-correlation analysis was performed using version 1.10.1 of SPP (Kharchenko *et al.* 2008) and the following parameter: "−s = 0:2:400." QC scores were assigned based on the relative strand correlation (RSC) values (integers ranging from −2 to −2, $RSC \in \{0, 0.25\} \Rightarrow QC \leftarrow -2$, $RSC \in \{0.25, 0.50\} \Rightarrow QC \leftarrow -1$, $RSC \in \{0.50, 1.00\} \Rightarrow QC \leftarrow 0$, $RSC \in \{1, 1.50\} \Rightarrow QC \leftarrow +1$, $RSC \geq 1.5 \Rightarrow QC \leftarrow +2$, with −2 corresponding to minimal read clustering and 2 corresponding to a highly clustered library) and used as a measure of ChIP quality. These scores capture the extent of read clustering in a ChIP-seq experiment in organisms whose genomes have similar size and structure to those of mammals. We point out that these scores may not be appropriate in genomes with very different size and/or structure. This motivated us to discard data from nonvertebrate model organisms for this analysis. Different values than those used here for RSC or normalized strand correlation (NSC) coefficients may be needed for such genomes, and this is a topic for future investigation. Cross-correlation plots were manually examined to ensure no artifactual QC scores were included because of size selection issues (such as, for example, a library being fragmented to an average size close to the read length and confusing the automated fragment peak assignment). In general, we recommend manual examination of cross-correlation plots in all cases. This presents a deeper and more detailed view of the characteristics of the dataset because the cross-correlation profile provides not only information regarding ChIP enrichment but also regarding the fragment length distribution in the datasets. For example, a dataset might exhibit periodicity in the distribution of fragment size lengths, presenting itself as numerous smaller peaks along the curve (often seen when chromatin is enzymatically digested rather than sonicated), or it can deviate from the standard unimodal pattern (aside from the phantom peak) indicating issues with size selection. The code for running SPP and assigning QC scores is available at https://code.google.com/p/phantompeakqualtools/.

MyoD and myogenin ChIP-seq peak calling

MyoD and myogenin datasets were generated by the Wold laboratory and are available under GEO accession number GSE44824. We note that the apparent weakness of the "myogenin 2" ChIP dataset is most likely attributable to undersequencing and would be elevated to high-quality status if sequenced deeper; undersequencing is one possible reason for suboptimal quality metrics (A. Kundaje *et al.*, unpublished data). Reads were mapped as described above and peaks were called using ERANGE3.2 (Johnson *et al.* 2007) with the following settings: "−minimum 2 −ratio 3 −shift learn −revbackground −listPeak." ChIP-seq peak calls were counted as overlapping if their summits were within 200 bp of each other. Read mapping statistics and QC metrics for these datasets can be found in Supporting Information, Table S2.

RESULTS

Dataset collection, data processing, and quality metrics

We downloaded all GEO series containing ChIP-seq datasets for vertebrate TFs or chromatin-modifying and remodeling proteins, along with their corresponding control libraries, submitted before April 1, 2012. We excluded ENCODE datasets because they have previously been subjected to this quality assessment (ENCODE Project Consortium 2012). We provide here a summary of ENCODE TF ChIP-seq data quality from the two main production groups in Figure S9 and Figure S10 (Landt *et al.* 2012).

For several reasons, we also excluded histone modifications and RNA Polymerase II datasets. First, in our experience, ChIP-seq against these targets is very robust to experimental variation and the success rate is reliably high (provided the antibody reagents used are of high quality). Second, an especially large proportion of published data are for histone marks. The effect of including all of these in the survey is to obscure or skew what is happening in the information-rich sample set that includes diverse TFs and cofactors. Finally, the currently available QC metrics were designed and are best suited for TF data that produce highly localized "point-source" occupancy (as they quantify the extent of read clustering in the genome). This means that the metrics themselves need to be interpreted differently if they are applied to, for example, repressive histone marks such as H3K9me3 and H3K27me3, which form large "broad-source" regions of enrichment (Pepke *et al.* 2009). Arguably, these data will need their own metrics and this will be a challenge for the future.

The final collection of datasets contained 191 GEO series containing a total of 917 ChIP-seq and 292 control libraries. Except for a limited number of cases in which a GEO series was associated with multiple publications, two or three GEO series were associated with the same publication, or a GEO series has not yet been used in a publication, and there is a one-to-one relationship between GEO series and published articles in the literature (Robertson *et al.* 2007; Chen *et al.* 2008; Marson *et al.* 2008; Bilodeau *et al.* 2009; Cheng *et al.* 2009; De Santa *et al.* 2009; Lister *et al.* 2009; Nishiyama *et al.* 2009; Visel *et al.* 2009; Welboren *et al.* 2009; Wilson *et al.* 2009; Yu *et al.* 2009; Yuan *et al.* 2009; Barish *et al.* 2010; Blow *et al.* 2010; Blow *et al.* 2010; Cao *et al.* 2010; Chi *et al.* 2010; Chia *et al.* 2010; Chicas *et al.* 2010; Corbo *et al.* 2010; Cuddapah *et al.* 2009; Durant *et al.* 2010; Fortschegger *et al.* 2010; Gotea *et al.* 2010; Gu *et al.* 2010; Han *et al.* 2010; Heinz *et al.* 2010; Heng *et al.* 2010; Ho *et al.* 2009; Hollenhorst *et al.* 2009; Hu *et al.* 2010; Johannes *et al.* 2010; Jung *et al.* 2010; Kagey *et al.* 2010; Kassouf *et al.* 2010; Kim *et al.* 2010; Kong *et al.* 2010; Kouwenhoven *et al.* 2010; Krebs *et al.* 2010; Kunarso *et al.* 2010; Kwon *et al.* 2009; Law *et al.* 2010; Lee *et al.* 2010; Lefterova *et al.* 2010; Li *et al.* 2010; Lin *et al.* 2010; Liu *et al.* 2010; Ma *et al.* 2010; MacIsaac *et al.* 2010; Mahony *et al.* 2010; Martinez *et al.* 2010; Palii *et al.* 2010; Qi *et al.* 2010; Rada-Iglesias *et al.* 2010; Rahl *et al.* 2010; Ramagopalan *et al.* 2010; Ramos *et al.* 2010; Schlesinger *et al.* 2010; Schnetz *et al.* 2010; Sehat *et al.* 2010; Steger *et al.* 2010; Tallack *et al.* 2010; Tang *et al.* 2010; Vermeulen *et al.* 2010; Verzi *et al.* 2010; Vivar *et al.* 2010; Wei *et al.* 2010; Woodfield *et al.* 2010; Yang *et al.* 2010; Yao *et al.* 2010; Yu *et al.* 2010; An *et al.* 2011; Ang *et al.* 2011; Bergsland *et al.* 2011; Bernt *et al.* 2011; Botcheva *et al.* 2011; Brown *et al.* 2011; Bugge *et al.* 2011; Ceol *et al.* 2011; Ceschin *et al.* 2011; Costessi *et al.* 2011; Ebert *et al.* 2011; Fang *et al.* 2011; Handoko *et al.* 2011; He *et al.* 2011; Heikkinen *et al.* 2011; Holmstrom *et al.* 2011; Horiuchi *et al.* 2011; Hu *et al.* 2011; Joseph *et al.* 2010; Kim *et al.* 2011; Klisch *et al.* 2011; Koeppel *et al.* 2011; Kong *et al.* 2011; Little *et al.* 2011; Liu *et al.* 2011; Lo *et al.* 2011; Marban *et al.* 2011; Mazzoni *et al.* 2011; McManus *et al.* 2011; Mendoza-Parra *et al.* 2011; Meyer *et al.* 2012; Miyazaki *et al.* 2011; Mullen *et al.* 2011; Mullican *et al.* 2011; Nakayamada *et al.* 2011; Nitzsche *et al.* 2011; Norton *et al.* 2011; Novershtern *et al.* 2011; Quenneville *et al.* 2011; Rao *et al.* 2011; Rey *et al.* 2011; Sahu *et al.* 2011; Schmitz *et al.* 2011; Seitz *et al.* 2011; Shen *et al.* 2011; Shukla *et al.* 2011; Siersbæk *et al.* 2011; Smeenk *et al.* 2011; Smith *et al.* 2011; Soccio *et al.* 2011; Stadler *et al.* 2011; Sun *et al.* 2011; Tan *et al.* 2011a; Tan *et al.* 2011b; Teo *et al.* 2011; Tijssen *et al.* 2011; Tiwari *et al.* 2011a; Tiwari *et al.* 2011b; Trompouki *et al.* 2011; van Heeringen *et al.* 2011; Verzi *et al.* 2011; Wang *et al.* 2011a; Wang *et al.* 2011b; Wei *et al.* 2011; Whyte *et al.* 2011; Wu *et al.* 2011a; Wu *et al.* 2011b; Xu *et al.* 2011; Yang *et al.* 2011; Yildirim *et al.* 2011; Yoon *et al.* 2011; Zhang *et al.* 2011; Zhao *et al.* 2011a; Zhao *et al.* 2011b; Avvakumov *et al.* 2012; Barish *et al.* 2012; Boergesen *et al.* 2012; Bugge *et al.* 2012; Canella *et al.* 2012; Cardamone *et al.* 2012; Cheng *et al.* 2012; Chlon *et al.* 2012; Cho *et al.* 2012; Doré *et al.* 2012; Fan *et al.* 2012; Feng *et al.* 2011; Fong *et al.* 2012; Gao *et al.* 2012; Gowher *et al.* 2012; Hunkapiller *et al.* 2012; Hutchins *et al.* 2012; Li

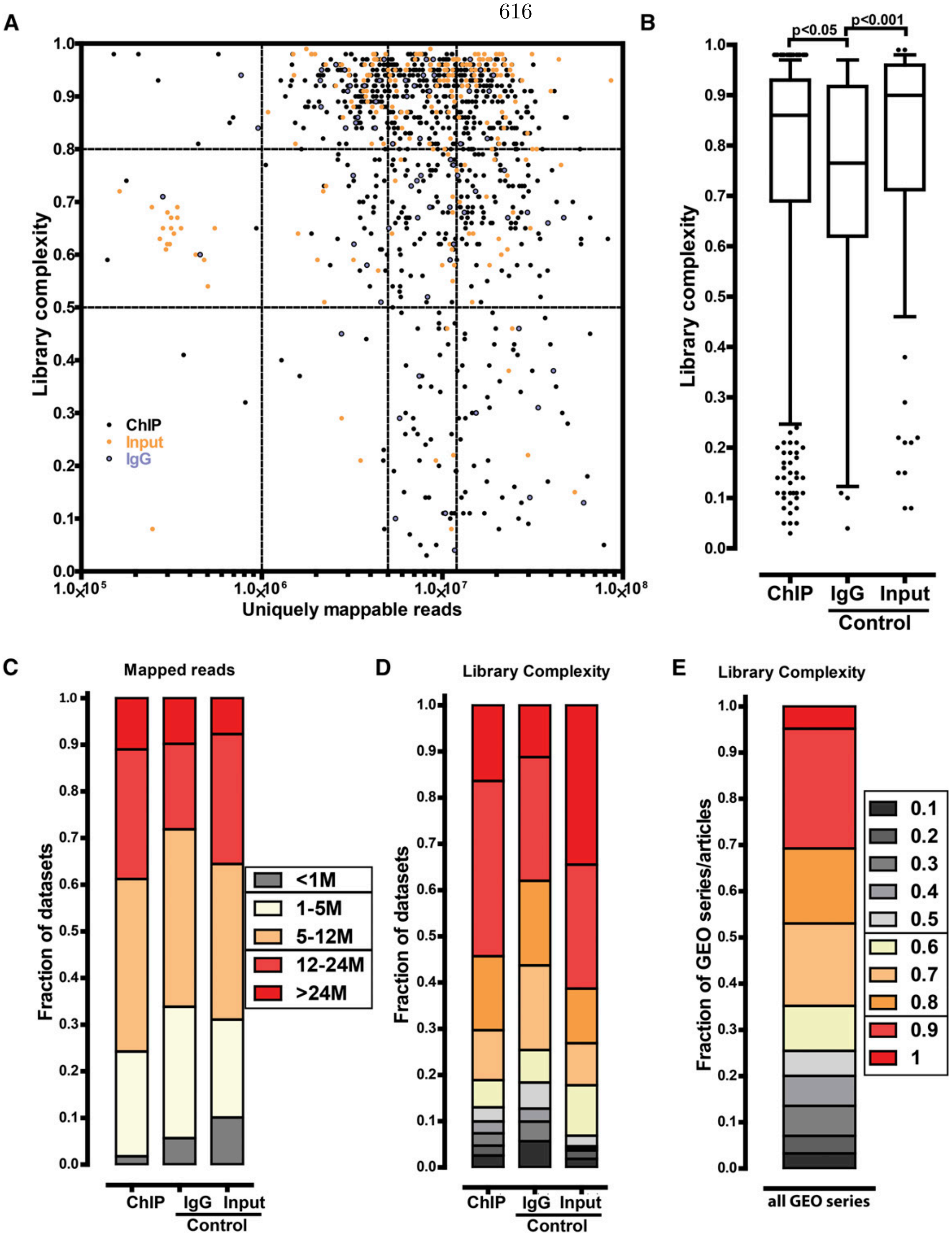


Figure 1 Sequencing library characteristics. (A) Joint distribution of library complexity and sequencing depth for all datasets examined. Vertical lines are drawn at 1 million, 5 million, and 12 million reads. Horizontal and vertical lines indicate quality classes discussed in the text. The upper right domain (number of uniquely mappable reads ≥12 million and library complexity ≥0.8) passes current quality thresholds. (B) Distribution of library complexity for ChIP-seq datasets, IgG controls, and inputs. (C) Distribution of sequencing depth for ChIP-seq datasets, IgG controls, and sonicated inputs. (D) Fraction of ChIP-seq, IgG, and input datasets exhibiting high, medium, and low complexity. (E) Fraction of studies containing libraries of high, medium, and low complexity (the distribution of the minimum library complexity observed is shown)

et al. 2012; Lu *et al.* 2012; Miller *et al.* 2011; Ntziachristos *et al.* 2012; Pehkonen *et al.* 2012; Ptasinska *et al.* 2012; Remeseiro *et al.* 2012; Sadasivam *et al.* 2012; Sakabe *et al.* 2012; Schödel *et al.* 2012; Trowbridge *et al.* 2012; Vilagos *et al.* 2012; Wu *et al.* 2012; Xiao *et al.* 2012; Yu *et al.* 2012; unpublished at the time of completion of this manuscript are the following GEO accession numbers: GSE33346, GSE33850, GSE36561, GSE30919, GSE33128, GSE35109, GSE25426, GSE31951, GSE26711, GSE23581, GSE26136, GSE26680, GSE15844, GSE21916, GSE22303, and GSE29180; direct links to all GEO series can be found in Table S1).

We discuss IgG and input controls separately because, to the best of our knowledge, any potential general differences between the two types of controls have not been investigated systematically in the context of ChIP-seq (Peng *et al.* 2007 addressed these questions for ChIP-Chip data; however, the nature of the background is substantially different for microarrays).

We mapped all reads with uniform settings (see *Materials and Methods* for details) and examined library and ChIP QC metrics for each dataset. These criteria have already been discussed by Landt *et al.* (2012), and a detailed treatment of cross-correlation is presented elsewhere (Kundaje *et al.*, unpublished data). Here, we provide a brief overview of each.

Sequencing depth: If a ChIP-seq experiment achieves successful immune enrichment and the resulting library adequately represents the sample, then greater sequencing depth will produce a more complete map of TF occupancy (Landt *et al.* 2012). At a greater depth, the measurement will identify a larger number of reproducible sites containing the corresponding DNA-binding sequence motif. Undersequencing of an otherwise successful library will lead to false-negative results. It has been difficult to establish a universal minimal sequencing depth because of differences between factors. Any threshold is going to be somewhat arbitrary but, in general, the major cost/benefit trade-off is between sequencing individual samples more deeply and generating more replicates; for most contemporary purposes, an independent duplicate measurement of 12 million reads arguably adds greater overall value than a single determination with 24 million reads, even though the higher number of reads will increase sensitivity. The number of mapped reads less than 1–2 million for a typical TF will usually be inadequate for capturing the complexity of an interactome for a mammalian-size genome. Many datasets now in the public domain were generated when sequencing throughput was lower than it is now and costs were higher (between 2007 and 2013, sequencing throughput has increased by approximately two orders of magnitude). As a consequence, many early ChIP-seq libraries were sequenced to a depth of only a few million reads. We therefore divided datasets into sequencing bins by using thresholds of 1 million, 5 million, 12 million, and 24 million uniquely mapped reads (taking into account sequencing depths recommended in the past by the ENCODE consortium for TFs). Libraries having less than 1 million reads are considered severely undersequenced, and those with more than 12 million are considered reasonably deeply sequenced.

Library complexity: A second characteristic that influences the quality of a ChIP-seq measurement is the sequence fragment diversity of the sequencing library. This is often referred to as library complexity, and low complexity is undesirable, although we note that much better IP enrichment than what is now obtained could, in the future, lead to very high-quality datasets with low library complexity. Currently, low-complexity libraries mainly result from experimental deficiencies: either too few starting molecules at the end of the immunoprecipitation step or inefficient steps in subsequent library building. As a result, the same starting molecules are sequenced repeatedly. Very-low-complexity libraries will not contain enough information to effectively sample the true positive occupancy sites and they distort the signal position and intensity. This can confuse peak callers (especially if the algorithm does not collapse presumptive PCR duplicates), leading to peak calling artifacts (Landt *et al.* 2012). We calculate the following metric as an indicator of library complexity (Landt *et al.* 2012):

$$\text{Library complexity} = \frac{\text{Number positions in the genome with uniquely mappable reads in dataset}}{\text{Number uniquely mappable reads in dataset}} \tag{1}$$

Estimated in this simple way, library complexity is expected to decrease eventually with increased sequencing depth because even highly complex libraries become exhausted by very deep sequencing. Reduced apparent complexity would also be observed with extremely successful ChIP-seq experiments for TFs that bind to the genome in a highly discriminative fashion to a limited number of locations. In such libraries, the majority of reads would originate from the limited genomic subspace around binding sites, resulting in low library complexity. With current methods, this is a largely theoretical consideration; in practice, in most ChIP-seq libraries only a minority of reads originates from factor-bound sites, with the rest (the majority) representing genomic background. Because the majority of libraries we examined were in the sequencing depth range over which these values represent library complexity reasonably well (Figure 1A and Figure S2), we separated datasets into the following complexity groups: high complexity (apparent library complexity $\geq .8$); medium to low complexity (apparent library complexity between 0.5 and 0.8); and very low complexity (apparent library complexity $\leq .5$). We also note that in substantially smaller genomes, the apparent library complexity is expected to be lower because the number of positions from which sequencing library fragments can originate is smaller.

Cross-correlation analysis of read clustering and ChIP enrichment: Because the majority of sequencing reads in a ChIP-seq library represent nonspecific genomic backgrounds, these reads are expected to be distributed randomly over the genome, to a first approximation. In contrast, reads originating from specific occupancy events cluster around the sites of protein–DNA interactions, where they are distributed in a characteristic asymmetric pattern on the plus and minus strands (Kharchenko *et al.* 2008). Cross-correlation analysis is an effective way of measuring the extent of this clustering. It also captures additional global features of the data, such as the average fragment length and fragment length distribution (Kharchenko *et al.* 2008; Landt *et al.* 2012). Specifically, the read coverage profiles on the two strands are shifted relative to the other over a range of shift values and the correlation between the profiles is calculated at each shift (Kharchenko *et al.* 2008). The resulting plot has one ("phantom") peak corresponding to the read length and another peak corresponding to the average fragment length; the height of the fragment-length peak is highly informative of the extent of read clustering in the library and, in turn, of the success of a ChIP-seq experiment. This feature is best captured by the NSC and RSC metrics discussed by Landt *et al.* (2012).

We applied SPP (Kharchenko *et al.* 2008) to perform cross-correlation analysis for all libraries in our survey. We then used the RSC cross-correlation metric to assign integer QC tag values in the $\{-2, 2\}$ range to datasets, with QC values of 2 corresponding to very highly clustered (and most likely, also successful) datasets and QC values of -2 to datasets exhibiting no to minimal read clustering; negative values are

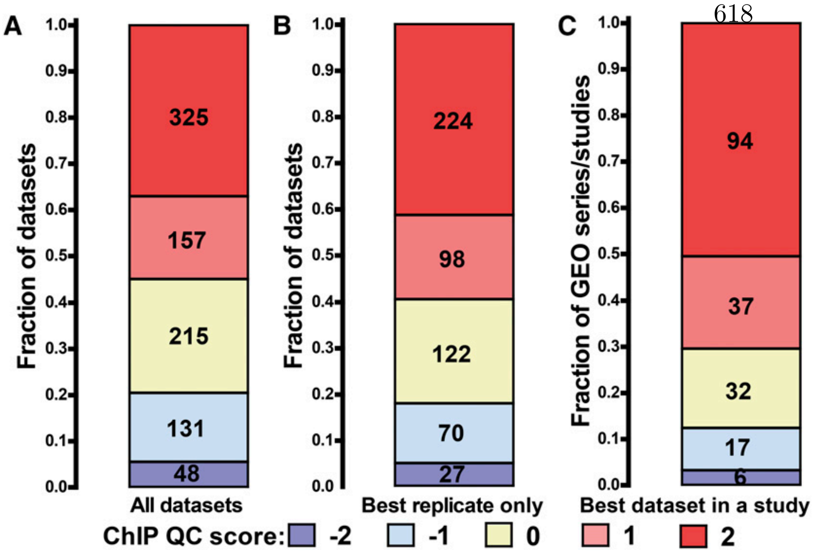


Figure 2 ChIP QC assessment summary. The numbers in each box indicate the total number of datasets/studies belonging to it. SPP QC scores of +1 and +2 indicate a high degree of read clustering in a dataset. (A) Distribution of SPP QC scores for all ChIP-seq datasets examined. (B) Distribution of SPP QC scores for the best replicates for a factor/condition combination in each study. (C) Distribution of the maximum SPP QC scores for all ChIP-seq datasets in a study.

expected for input datasets. The RSC metric captures well the extent of read enrichment in vertebrate genomes similar in size and structure to humans, which this study focuses on. We provide representative examples of cross-correlation plots for each of the five QC categories in Figure S1A, and we use these tags as convenient general proxies for ChIP quality throughout the following analysis. We note that the discretization thresholds are not intended to be absolute determinants of quality, but they do enable one to rapidly scan very large numbers of datasets. In practice, examining the cross-correlation plots and the continuously distributed NSC and RSC values and using those together with information about sequencing depth and library complexity are always more informative and can provide valuable nuances for understanding specific datasets. Direct examination of plots allows one to detect datasets with odd cross-correlation profiles (we show a few representative examples in Figure S11). It is possible in theory for low-complexity libraries to produce artificially high cross-correlation scores if stacks of reads on opposite strands are located close to each other in regions of enrichment; however, the Pearson correlation between library complexity scores and RSC values in the collection of ChIP datasets surveyed here was 0.0084, indicating that such cases do not feature significantly in this analysis.

An additional major component of the ChIP-seq QC pipeline developed by the ENCODE consortium is reproducibility analysis of replicates, based on the irreproducible discovery rate (IDR) statistic (Li *et al.* 2011). However, because many of the studies we surveyed did not have replicates, we only evaluated datasets on the level of individual experiments. Single dataset evaluation is almost always a valuable precursor to evaluation of replicates because, typically, a second replicate is generated after a successful first one. The full list of datasets, mapping, and QC statistics is provided in Table S1.

Sequencing depth and library complexity

Figure 1A shows the distribution of sequencing depth and library complexity for ChIP-seq and control datasets. The upper right domain, bounded by 12 million reads per sample and a complexity value of 0.8, is an arbitrary but useful definition of high quality according to these measures. A majority of datasets had reasonably good complexity and severely undersequenced libraries were rare (Figure 1C). A minority (38.8%) of datasets had more than 12 million mapped reads; however, as discussed, this is not unexpected, because a large fraction of the datasets we surveyed were generated in times of significantly higher sequencing cost and lower throughput. Strikingly, the median complexity of IgG control datasets was less than 0.8 and considerably lower than that of either ChIP-seq or sonicated input libraries (Figure 1B). This is not a result of IgG datasets having been sequenced much more deeply than the other two groups; in fact, the median sequencing depth of IgG controls is lower (Figure S2). The concern that some individual IgG inputs might provide insufficient DNA mass to build highly complex libraries has been raised before (Landt *et al.* 2012), and our observations are consistent with this, although it is not a characteristic of all IgG controls.

Slightly more than half (54.3%) of ChIP-seq datasets had library complexity more than 0.8, whereas very-low-complexity (< 0.5) libraries comprised 12.9% of datasets; the fraction of very-low-complexity libraries was higher and lower for IgG and input datasets, respectively (Figure 1D). Because most GEO series contained multiple libraries, we also asked, how common is the presence of low-complexity libraries in individual studies? Figure 1E shows the distribution of the minimum library complexity in each such series (for all types of datasets). One-quarter (25.4%) of all studies contained very-low-complexity libraries.

Cross-correlation quality assessment of ChIP-seq datasets

Next, we examined the distribution of SPP QC scores for ChIP-seq datasets. Before doing this, we excluded a minority of datasets for which there was a good reason to think high ChIP enrichment should not be expected. For example, experiments executed in knockouts, knockdowns, or settings in which the factor is not expressed are not expected to produce a high-scoring measurement. And in a few cases, the factor in question might be known to bind to only a small number of sites in the genome; this has been proposed, for example, for some ZNF TFs and Pol3 and its associated factors (Landt *et al.* 2012). Our detailed criteria for inclusion are described in *Materials and Methods*.

Figure 2A shows the QC score distribution for all ChIP-seq datasets we retained. Strikingly, only 55% (482 out of 876) of datasets had QC scores of 1 or 2, *i.e.*, they were likely to be highly successful. An additional 24.5% (215 out of 876) had a score of 0, indicating that they were of intermediate quality, and 20.4% (179 out of 876) had low-quality scores of -1 and -2. Sometimes multiple replicates for a factor were submitted but only one scored poorly, so we also compiled a second set of ChIP-seq experiments that only included the best available replicate for each factor and condition (Figure 2B). This

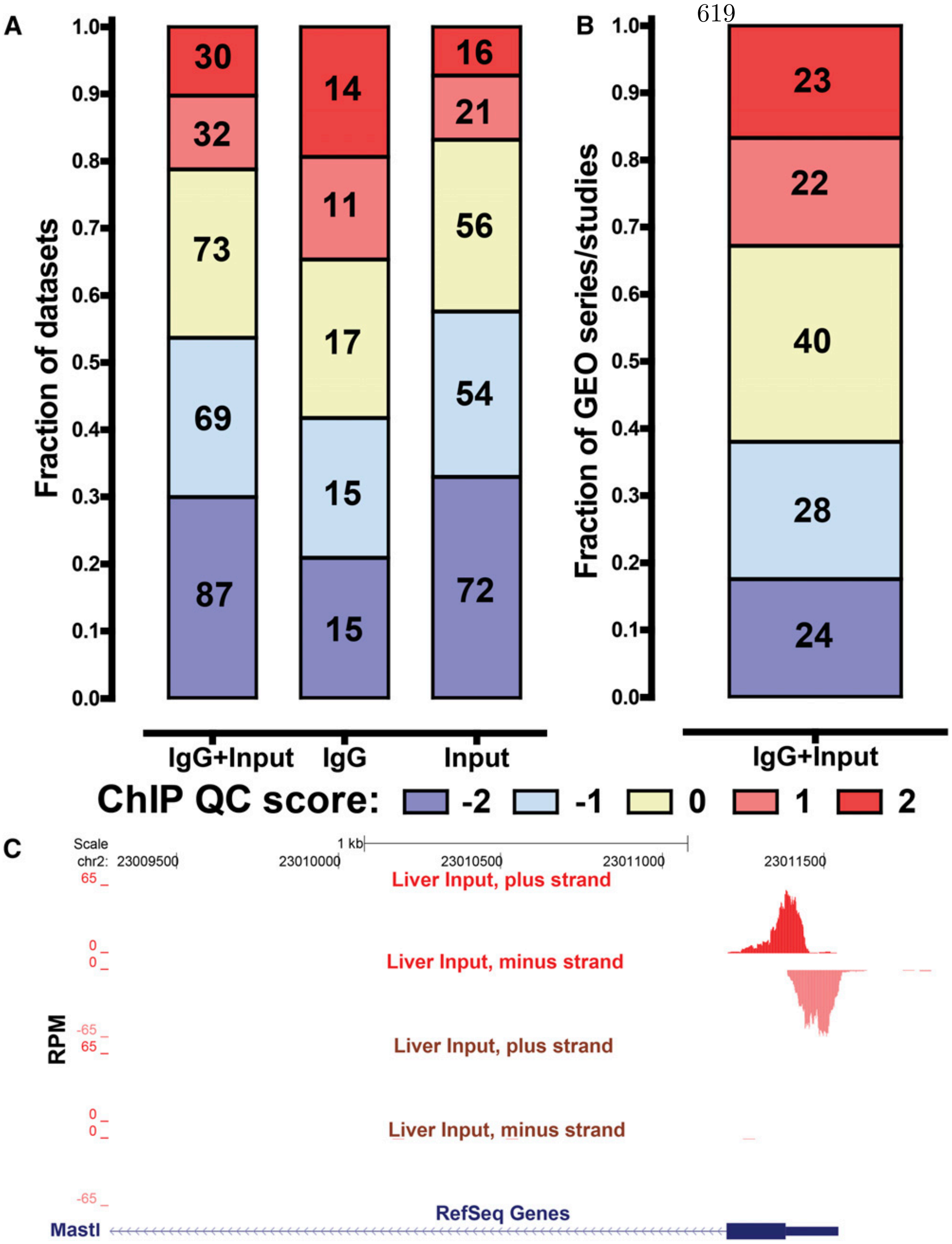


Figure 3 Assessment of read clustering in control datasets. The numbers in each box indicate the total number of datasets/studies belonging to it. SPP QC scores of 1 and 2 indicate a high degree of read clustering in a dataset. (A) Distribution of SPP QC scores for all control datasets (IgG + input), IgG/mock IP controls (IgG), and sonicated inputs (inputs). (B) Fraction of studies containing highly clustered inputs. The distribution of the maximum SPP QC score for all inputs in a dataset is shown. (C) Examples of a highly clustered input [mouse liver, upper two tracks, (MacIsaac *et al.* 2010), QC score of 2] and an input that does not show high extent of read clustering [mouse liver, lower two tracks (Soccio *et al.* 2011), QC score of −1). The promoter of the *MASTL* gene is shown. All tracks are shown to the same scale and reads mapping to the plus and minus strands are displayed separately for better visualization of the cross-correlation between the two.

set included 322 datasets (59%) with QC scores of 2 or 1. The fraction of intermediate-quality or low-scoring datasets in this group decreased as expected. However, the decrease was modest with 18% (97 out of 541) of the best available replicates scoring −1 or −2, and 22.5% (122 out of 541) scoring 0.

We then examined the distribution of the maximum QC score for each study, regardless of the target identity (Figure 3C). The fraction of low scores decreased further, though only 70.4% of studies (131 out of 186) had a score of 1 or 2 for their best experiment. Finally, we compiled a list of the top-scoring datasets from all studies that assayed only a single TF; 19.7% (19 out of 96) of these studies had scores of −1 or −2, 25% (24 of 96) had a score of 0, and 55.2% (53 of 96) were marked as likely to be successful, with scores of 1 and 2 (Figure S3C).

Read clustering in control datasets

Control datasets serve the important purpose of helping to distinguish read enrichment attributable to the immunoprecipitation step from artifactual read clustering attributable to other experimental factors, both known and unknown. It is, for example, well-appreciated that differential chromatin shearing efficiency can lead to the overrepresentation of areas of open chromatin (usually immediately surrounding transcribed promoters) in sequencing libraries. This has been termed the "Sono-seq" effect when attributed to sonication (Auerbach *et al.* 2009). In addition, unknown copy number variants relative to the reference genome or sequence composition biases can give false-positive occupancy calls. In particular, specifics of the amplification step in sequencing platforms can introduce bias due to GC content (Ho *et al.* 2011).

In general, control datasets are not expected to exhibit a pattern of significant read clustering similar in strength to that of successful ChIP-seq datasets. In our own practice, under standard cross-linking protocols, most do not. However, we noticed that a minority of control datasets produce positive ChIP QC metric scores along with prominent cross-correlation peaks. Figure S1B shows examples of cross-correlation plots for individual control datasets with all possible QC scores, from

−2 to 2, and Figure 3C shows a browser snapshot of a region with strong read enrichment in a highly clustered (QC score of 2) input library. No such enrichment was observed in a different control library from a similar biological source having a QC score of −1.

We asked how general this phenomenon is by examining the distribution of QC scores of both IgG and input control datasets (Figure 3A). Surprisingly, only 53.6% (156 out of 291) of control datasets had QC scores of −2 or −1 and 25% (73 of 291) had a score of 0, whereas 21.3% (62 of 291) exhibited a very high degree of read clustering and received scores of 1 or 2. The highly clustered inputs were notably more common among IgG controls than among input chromatin controls (Figure 3A). Moreover, high read clustering was more often found in low-complexity libraries (which are themselves more common among IgG controls) (Figure S4, A and B).

We also examined how widespread control sample clustering is on the level of individual GEO series/studies to see if the phenomenon is restricted to a few larger studies. Figure 3B shows the distribution of the maximal control sample QC score for all studies. Of the studies for which control datasets were available, 32.8% (45 of 123) contained at least one highly clustered control with a score of 1 or 2, and 29.2% (40 of 123) contained a control with a score of 0. Thus, control datasets surprisingly often exhibit a high extent of read clustering similar to that of ChIP-seq datasets. This is even more striking considering that formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) data (an assay that is based on the preferential enrichment of open chromatin in sonicated DNA and aims to achieve high read clustering) from ENCODE usually have QC scores between −2 and 0, Moreover, the Sono-seq datasets published by Auerbach *et al.* (2009) all have scores of −2.

We note that unless this effect is very strong and is associated with notable genomic features such as promoters of genes, it can be difficult to detect by the usual methods of visual inspection of signal tracks on a genome browser. It is, however, readily apparent in cross-correlation analysis and our results raise awareness of its existence. As mentioned, one candidate explanation for this phenomenon is the previously described "Sono-seq" effect. Using standard experimental protocols, this effect has been rare in our experience; however, under more aggressive cross-linking conditions, we have observed increased read clustering in control samples (Figure S5). Notably, the original "Sono-seq" description focused on promoter regions, but we have also observed it over distal regulatory elements, where its strength was even higher than at promoters (Figure S5). Thus, variation in the extent of fixation, as well as sonication, might be a substantial contributor to variation in read clustering across the broader data collection. Another potential contributing factor is sequencing depth. Although the average sequencing depth for highly clustered IgG and input controls is higher than that of controls with negative QC scores (Figure S4, C and D) this by no means explains all the clustering observed in controls. There are many examples of more deeply sequenced input and IgG libraries with no significant cross-correlation peaks and very few of them were sequenced especially deeply (only eight control libraries had $>4 \times 10^7$ reads not desirable. Finally, "Sono-seq" need not be the only explanation. Whereas a number of control datasets with QC scores of 2 exhibited higher read coverage around promoters, others did not (Figure S6), suggesting at least one additional source of unexplained read enrichment in control samples. Because rich annotation of functional genomic elements outside promoter regions was not available for many cell types in our survey, this phenomenon is a subject for future analyses.

DISCUSSION

We performed a systematic survey of ChIP quality for publicly available vertebrate ChIP-seq datasets and found that more than half score as high quality by our measures. This group comprises a set that we believe can be used with confidence for integrative analyses. This conclusion carries the important caveat that we could not assess the specificity of the immune reagents used to perform the experiments. which powerfully affects the biological meaning of the data.

A substantial minority of published datasets (between 20% and 45% of those examined) were of low or intermediate quality by our metrics. This was true not only for individual libraries but also for the best replicates from each study. In addition, we observed a substantial number of low-complexity datasets and an unexpected group of highly clustered control datasets. These observations underscore the widespread variation in published ChIP-seq data. They also raised questions about which kinds of conclusions in primary publications are more or less sensitive to these aspects of data quality. In particular, global quality analysis is useful for guiding subsequent re-use of published data that require higher quality than was needed or achieved in the source study.

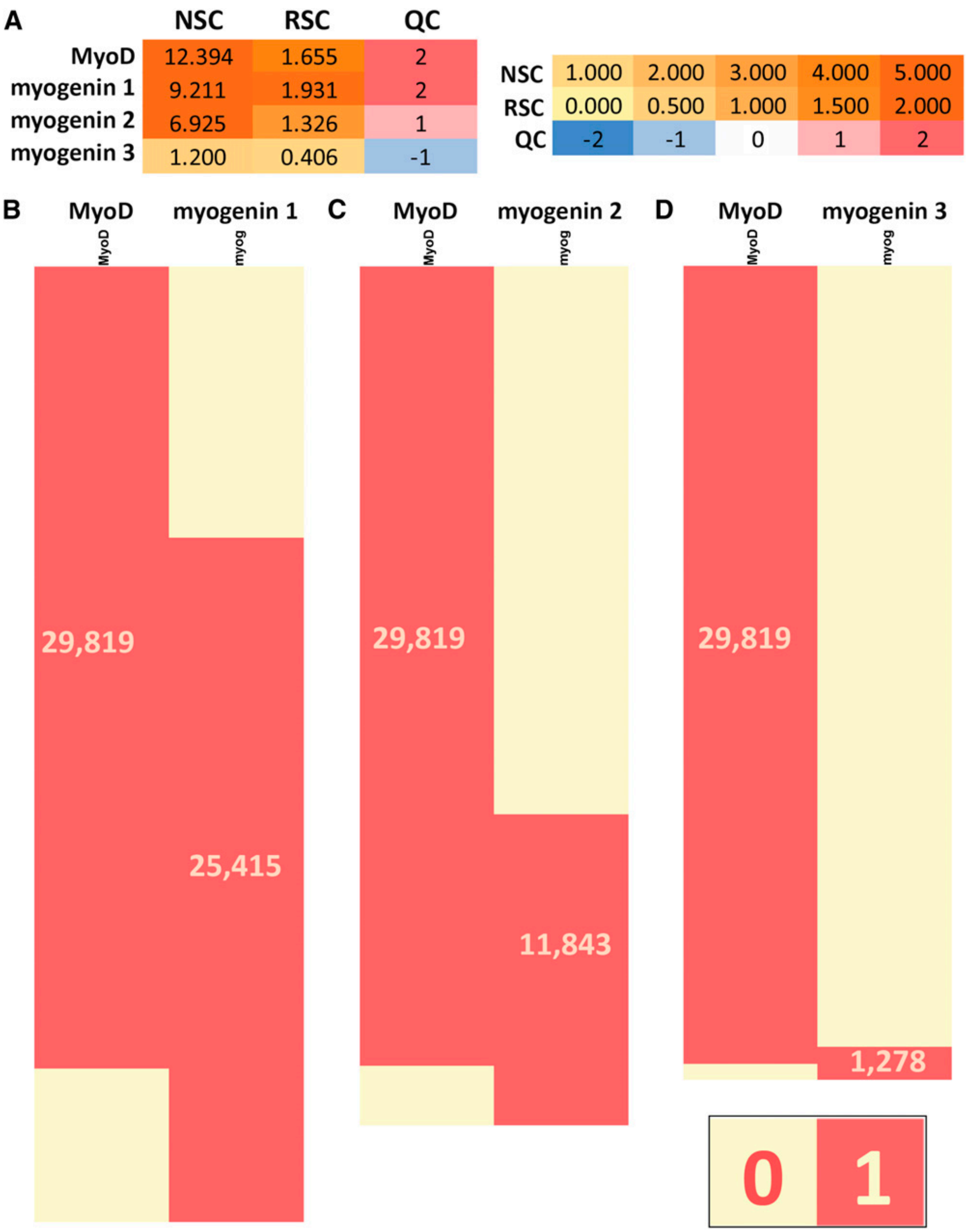
Data quality varied widely across "impact" levels. We separated datasets into groups according to the 2011 Thomson Reuters Impact Factor for the journal in which the corresponding article was published and examined the distribution of QC scores in each group (Figure S8). The group with highest impact factor (≥25) contained the largest fraction of datasets with a low QC score of −2 or −1. We also examined the distribution of QC scores with respect to the year of publication and found that the fraction of datasets with low scores has stabilized in the past 3 yr at approximately 20% (Figure S7).

We emphasize that datasets scoring as low quality by the metrics used here can, nevertheless, produce important biological discoveries. For this reason, it would be an error to set a rigid "standard" that every published dataset must meet. Instead, routine QC analysis can make it easy to see when there is reason for concern about a given dataset. It can also provide a first tier of guidance about what uses are likely to be appropriate for a given dataset. As discussed previously, the appropriate level of QC stringency depends on the specific goals of the experiment and methods of analysis (Landt *et al.* 2012). In particular, some analyses that are sensitive to false-negative results are particularly vulnerable to inclusion of low-scoring datasets. For example, trying to derive combinatorial TF occupancy rules is seriously compromised and even misleading if a subset of the datasets included is suboptimal.

We illustrate this with a simple example from our own experience (Figure 4). The MyoD and myogenin TFs are well-known regulators of muscle differentiation (Yun and Wold 1996) and C2C12 cells (Yaffe and Saxel 1977) have been widely used to study the process because they can be propagated in an undifferentiated myoblast state and easily induced to differentiate into myocytes and myotubes. We have performed several ChIP-seq experiments with these factors in differentiated and undifferentiated C2C12 cells (G. DeSalvo *et al.*, unpublished data; A. Kirilusha *et al.*, unpublished data; K. Fisher-Aylor *et al.*, unpublished data), some of which have been highly successful, whereas others were of poor or intermediate quality. Here, we examined the effect of weaker ChIP-seq datasets on combinatorial occupancy analysis using a MyoD ChIP-seq dataset with very high QC metrics and three myogenin datasets with very high, moderately good, and very low metrics (Figure 4A). Using the best myogenin dataset, we found a high degree of overlap between the binding sites of the two factors (Figure 4B). When the medium-quality myogenin dataset was used instead, a sizable group of MyoD-only sites emerged (Figure 4C) and the erroneous conclusion that a substantial number of MyoD sites lack myogenin binding could be reached if this was the only dataset available for analysis. Finally, the poor-quality myogenin dataset contains very few called peaks and, as a result, almost all MyoD sites show no myogenin binding when it is used for analysis (Figure 4D).

Recently, IDR analysis of replicate datasets (Li *et al.* 2011; ENCODE Project Consortium 2012; Landt *et al.* 2012) emerged as a robust method for deriving lists of reproducible occupancy sites from ChIP-seq datasets. IDR is based on differences in the consistency of ranking (usually by signal strength as measured by read enrichment or by statistical significance) for all identified peaks in a pair of ChIP-seq replicates. A virtue of this approach is that it allows a statistically robust set of binding sites to be derived largely independent of thresholds and settings specific to a particular peak-calling algorithm. Ideally, IDR would be used in conjunction with the quality metrics used here (ENCODE Project Consortium 2012; Landt *et al.* 2012). However, replicate measurements do not exist for many of the datasets in our survey of the historic. We expect that IDR will become common practice as sequencing costs decline. Even when that happens, measurements of the quality of individual datasets will remain important because they capture specific information in addition to reproducibility and because IDR analysis is sensitive to the presence of poor-quality replicates. An asymmetric pair consisting of one high-quality and one poorer-quality dataset is dominated in IDR by the weaker replicate, resulting in a shorter list of sites and a high false-negative rate. Care should be exercised in such cases. Although the best approach is to obtain a second high-quality replicate, but if this is not possible, special strategies for treating asymmetric replicates have been devised (Landt *et al.* 2012).

The most perplexing observation was that a subset of control datasets have extensive read clustering in the same range as successful ChIP-seq experiments. In our own practice, we have rarely encountered such libraries and, to the best of our knowledge, there has been no extensive treatment of this issue or its influence on data analysis in the literature. The phenomenon occurred more frequently in IgG controls than in input chromatin controls, although it is by no means limited to the former. In theory, an IgG control should be a superior representation of the true background noise in a ChIP-seq sample because it incorporates biases introduced by the entire



A	NSC	RSC	QC
MyoD	12.394	1.655	2
myogenin 1	9.211	1.931	2
myogenin 2	6.925	1.326	1
myogenin 3	1.200	0.406	-1

NSC	1.000	2.000	3.000	4.000	5.000
RSC	0.000	0.500	1.000	1.500	2.000
QC	-2	-1	0	1	2

Figure 4 Effect of suboptimal datasets on combinatorial occupancy analysis. The muscle-regulatory factors MyoD and myogenin were assayed in C2C12 myocytes at 60 hr after differentiation. Shown are a single, highly successful MyoD ChIP-seq dataset and three myogenin ChIP-seq datasets, one of which is similarly highly successful ("myogenin 1"), a second weaker one ("myogenin 2"), and a third one that is an experimental failure ("myogenin 3"). (A) Quality control metrics. (B, C, D) The extent of overlap of MyoD and myogenin-binding sites as determined using each of the three myogenin datasets (see *Materials and Methods* for data processing details). MyoD and myogenin are mostly found to bind to the same sites when interactome determinations of comparable strength are used. (B) A sizable group of apparently MyoD-only sites emerges when the medium-strength myogenin dataset is used because of a large number of false-negative myogenin calls. (C) Finally, the unsuccessful myogenin ChIP reveals that most MyoD are not shared by myogenin. (D) Numbers listed in the red blocks corresponding to each set of peak calls indicate size.

immunoprecipitation process, in addition to any enrichments or biases created by chromatin shearing. Using this logic, a simple interpretation is that high read clustering in these controls correctly identifies artifacts in the IP process. When high background sample clustering is observed in control sample, we suggest that it merits immediate investigation of its replicability and its impact on peak-calling for the corresponding ChIP. samples. The fact that we also observed a large number of IgG controls (Figure 3A) that showed no such clustering, argues that this is not a general feature.

A crucial issue is the extent to which clustering in controls is also present as experimental noise in ChIP libraries from the same material. In other words, how well-matched are the control samples with the corrresponding experimental samples, and how robust are the controls? For example, a very strong Sono-seq effect in a control sample is expected to give ChIP-seq libraries with high read clustering that is a combination of true ChIP (antibody-specific) signal plus Sono-seq-derived noise that covers promotors and enhancers in a nonspecific manner. Whereas most contemporary peak callers normalize for enrichment in controls, very strong background noise will diminish the signal-to-noise ratio and adversely affect sensitivity. How severely this affects the results will depend on the overlap between true factor occupancy sites and regions of artifactual read enrichment (for some factors this overlap may be negligible because they do not bind to Sono-seq regions); on the magnitude of the Sono-seq effect; and on the strength of the ChIP itself (sufficiently strong determinations are not greatly affected). Conversely, if a ChIP-seq library has a strong Sono-seq component and peak calling is performed against an imperfectly matched "control" sample in which the Sono-seq effect is of significantly lower magnitude, false-positive peak calls will increase. Unfortunately, in practice such cases are difficult to detect. They are not flagged directly by current quality metrics and are best detected by analyses specific to each study and factor, including specific motif enrichment. especially when little is known about the expected true-positive rates. Similar reasoning applies if the noise source is something other than Sono-seq.

Uniform retrospective quality assessment is resource-intensive and will not be practically feasible because the number of ChIP-seq datasets is growing exponentially. Retrospective analysis also comes too late to influence the experiments themselves or to contribute to the review process. A reasonable path forward would be to incorporate routine data quality assessment into experimental analysis, review for publication, and submission to public repositories, as a matter of community practice. However, our results also strongly caution against the blind and arbitrary application of our metrics (or others) in the absence of experimental and biological context. The character of the metrics used here reflects contemporary technology and the quality scale has been calibrated based on factors and co-factors most studied to date. We have seen that it is possible for good datasets to receive low QC scores in certain special situations (*e.g.*, very few sites of occupancy in the genome). It is also possible for some poor or mediocre datasets to receive high QC scores. For example, this can happen as a side-product of strongly clustered backgrounds of the kind discussed above. Some examples of datasets in which this might be the case are shown in Figure S11. For factors that ChIP extremely well, even datasets that are substantially suboptimal score highly. For example, CTCF ChIP-seq datasets routinely identify 35,000–40,000 reproducible binding sites and have QC scores of 2; a dataset that identifies only 15,000 sites is suboptimal given that knowledge; yet it will still receive a positive QC score. For these reasons, the current quality metrics are best used in the context of what is known about the factor, the biological system, and the questions being asked.

Despite important nuances of interpretation, we suggest that using ChIP quality metrics and making the results readily accessible will facilitate better-informed data use by the wider community. An important adjunct to routine QC annotation would be the ability, in major public data repositories, to flag and explain the exceptional cases for which QC scores should not be taken at face value. Finally, quality metrics themselves will continue to improve as the field's understanding of data structure, experimental artifacts, and the underlying biology all become more sophisticated. Provisions will be needed for incorporating such advances into routine dataset annotation while still achieving comparability through time.

ACKNOWLEDGMENTS

We thank members of the ENCODE consortium and members of the Wold laboratory for helpful discussions, and Henry Amrhein, Diane Trout, and Sean Upchurch for computational assistance. G.K.M. and B.J.W. are supported by the Beckman Foundation, the Donald Bren Endowment, and National Institutes of Health grants U54 HG004576 and U54 HG006998.

LITERATURE CITED

An, C. I., Y. Dong, and N. Hagiwara, 2011 Genome-wide mapping of Sox6 binding sites in skeletal muscle reveals both direct and indirect regulation of muscle terminal differentiation by Sox6. BMC Dev. Biol. 11: 59.

Ang, Y. S., S. Y. Tsai, D. F. Lee, J. Monk, J. Su *et al.*, 2011 Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell 145: 183–197.

Auerbach, R. K., G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi *et al.*, 2009 Mapping accessible chromatin regions using Sono-Seq. Proc. Natl. Acad. Sci. USA 106: 14926–14931.

Avvakumov, N., M. E. Lalonde, N. Saksouk, E. Paquet, K. C. Glass *et al.*, 2012 Conserved molecular interactions within the HBO1 acetyltransferase complexes regulate cell proliferation. Mol. Cell. Biol. 32: 689–703.

Barish, G. D., R. T. Yu, M. Karunasiri, C. B. Ocampo, J. Dixon *et al.*, 2010 Bcl-6 and NF-κB cistromes mediate opposing regulation of the innate immune response. Genes Dev. 24: 2760–2765.

Barish, G. D., R. T. Yu, M. S. Karunasiri, D. Becerra, J. Kim *et al.*, 2012 The Bcl6-SMRT/NCoR cistrome represses inflammation to attenuate atherosclerosis. Cell Metab. 15: 554–562.

Barski, A., S. Cuddapah, K. Cui, T. Roh, D. E. Schones *et al.*, 2007 High-resolution profiling of histone methylations in the human genome. Cell 129: 823837.

Bergsland, M., D. Ramsköld, C. Zaouter, S. Klum, R. Sandberg *et al.*, 2011 Sequentially acting Sox transcription factors in neural lineage development. Genes Dev. 25: 2453–2464.

Bernt, K. M., N. Zhu, A. U. Sinha, S. Vempati, J. Faber *et al.*, 2011 MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. Cancer Cell 20: 66–78.

Bilodeau, S., M. H. Kagey, G. M. Frampton, P. B. Rahl, and R. A. Young, 2009 SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev. 23: 2484–2489.

Blow, M. J., D. J. McCulley, Z. Li, T. Zhang, J. A. Akiyama *et al.*, 2010 ChIP-Seq identification of weakly conserved heart enhancers. Nat. Genet. 42: 806–810.

Boergesen, M., T. Å. Pedersen, B. Gross, S. J. van Heeringen, D. Hagenbeek *et al.*, 2012 Genome-wide profiling of liver X receptor, retinoid X receptor, and peroxisome proliferator-activated receptor a in mouse liver reveals extensive sharing of binding sites. Mol. Cell. Biol. 32: 852–867.

Botcheva, K., S. R. McCorkle, W. R. McCombie, J. J. Dunn, C. W. Anderson *et al.*, 2011 Distinct p53 genomic binding patterns in normal and cancer-derived human cells. Cell Cycle 10: 4237–4249.

Brown, S., A. Teo, S. Pauklin, N. Hannan, C. H. Cho *et al.*, 2011 Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. Stem Cells 29: 1176–1185.

Bugge, A., D. Feng, L. J. Everett, E. R. Briggs, S. E. Mullican *et al.*, 2011 Rev-erbα and Rev-erbβ coordinately protect the circadian clock and normal metabolic function. Genes Dev. 26: 657–667.

Canella, D., D. Bernasconi, F. Gilardi, G. LeMartelot, E. Migliavacca *et al.*, 2012 A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. Genome Res. 22: 666–680.

Cao, L., Y. Yu, S. Bilke, R. L. Walker, L. H. Mayeenuddin *et al.*, 2010 Genome-wide identification of PAX3-FKHR binding sites in rhabdomyosarcoma reveals candidate target genes important for development and cancer. Cancer Res. 70: 6497–6508.

Cardamone, M. D., A. Krones, B. Tanasa, H. Taylor, L. Ricci *et al.*, 2012 A protective strategy against hyperinflammatory responses requiring the nontranscriptional actions of GPS2. Mol. Cell 46: 91–104.

Ceol, C. J., Y. Houvras, J. Jane-Valbuena, S. Bilodeau, D. A. Orlando *et al.*, 2011 The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. Nature 471: 513–517.

Ceschin, D. G., M. Walia, S. S. Wenk, C. Duboé, C. Gaudon *et al.*, 2011 Methylation specifies distinct estrogen-induced binding site repertoires of CBP to chromatin. Genes Dev. 25: 1132–1146.

Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss *et al.*, 2008 Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106–1117.

Cheng, Y., W. Wu, S. A. Kumar, D. Yu, W. Deng *et al.*, 2009 Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. Genome Res. 19: 2172–2184.

Cheng, B., T. Li, P. B. Rahl, T. E. Adamson, N. B. Loudas *et al.*, 2012 Functional association of Gdown1 with RNA polymerase II poised on human genes. Mol. Cell 45: 38–50.

Chi, P., Y. Chen, L. Zhang, X. Guo, J. Wongvipat *et al.*, 2010 ETV1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours. Nature 467: 849–853.

Chia, N. Y., Y. S. Chan, B. Feng, X. Lu, Y. L. Orlov *et al.*, 2010 A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468: 316–320.

Chicas, A., X. Wang, C. Zhang, M. McCurrach, Z. Zhao *et al.*, 2010 Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. Cancer Cell 17: 376–387.

Chlon, T. M., L. C. Doré, and J. D. Crispino, 2012 Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. Mol. Cell 47: 608–621.

Cho, H., X. Zhao, M. Hatori, R. T. Yu, G. D. Barish *et al.*, 2012 Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β. Nature 485: 123–127.

Corbo, J. C., K. A. Lawrence, M. Karlstetter, C. A. Myers, M. Abdelaziz *et al.*, 2010 CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. Genome Res. 20: 1512–1525.

Costessi, A., N. Mahrour, E. Tijchon, R. Stunnenberg, M. A. Stoel *et al.*, 2011 The tumour antigen PRAME is a subunit of a Cul2 ubiquitin ligase and associates with active NFY promoters. EMBO J. 30: 3786–3798.

Cuddapah, S., R. Jothi, D. E. Schones, T. Y. Roh, K. Cui *et al.*, 2009 Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. 19: 24–32.

De Santa, F., V. Narang, Z. H. Yap, B. K. Tusi, T. Burgold *et al.*, 2009 Jmjd3 contributes to the control of gene expression in LPS-activated macrophages. EMBO J. 28: 3341–3352.

Doré, L. C., T. M. Chlon, C. D. Brown, K. P. White, and J. D. Crispino, 2012 Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. Blood 119: 3724–3733.

Durant, L., W. T. Watford, H. L. Ramos, A. Laurence, G. Vahedi *et al.*, 2010 Diverse targets of the transcription factor STAT3 contribute to T cell pathogenicity and homeostasis. Immunity 32: 605–615.

Ebert, A., S. McManus, H. Tagoh, J. Medvedovic, G. Salvagiotto *et al.*, 2011 The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. Immunity 34: 175–187.

ENCODE Project Consortium, 2011 A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9: e1001046.

ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Fan, R., S. Bonde, P. Gao, B. Sotomayor, C. Chen *et al.*, 2012 Dynamic HoxB4-regulatory network during embryonic stem cell differentiation to hematopoietic cells. Blood 119: e139–e147.

Fang, X., J. G. Yoon, L. Li, W. Yu, J. Shao *et al.*, 2011 The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. BMC Genomics 12: 11.

Feng, D., T. Liu, Z. Sun, A. Bugge, S. E. Mullican *et al.*, 2011 A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism. Science 331: 1315–1319.

Fong, A. P., Z. Yao, J. W. Zhong, Y. Cao, W. L. Ruzzo *et al.*, 2012 Genetic and epigenetic determinants of neurogenesis and myogenesis. Dev. Cell 22: 721–735.

Fortschegger, K., P. de Graaf, N. S. Outchkourov, F. M. van Schaik, H. T. Timmers *et al.*, 2010 PHF8 targets histone methylation and RNA polymerase II to activate transcription. Mol. Cell. Biol. 30: 3286–3298.

Gao, Z., J. Zhang, R. Bonasio, F. Strino, A. Sawai *et al.*, 2012 PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol. Cell 45: 344–356.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan *et al.*, 2012 Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91–100.

Gilmour, D. S., and J. T. Lis, 1984 Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. Proc. Natl. Acad. Sci. USA 81: 4275–4279.

Gilmour, D. S., and J. T. Lis, 1985 In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. Mol. Cell. Biol. 5: 2009–2018.

Gotea, V., A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio *et al.*, 2010 Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 20: 565–577.

Gowher, H., K. Brick, R. D. Camerini-Otero, and G. Felsenfeld, 2012 Vezf1 protein binding sites genome-wide are associated with pausing of elongating RNA polymerase II. Proc. Natl. Acad. Sci. USA 109: 2370–2375.

Gu, F., H. K. Hsu, P. Y. Hsu, J. Wu, Y. Ma *et al.*, 2010 Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. BMC Syst. Biol. 4: 170.

Han, J., P. Yuan, H. Yang, J. Zhang, B. S. Soh *et al.*, 2010 Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature 463: 1096–1100.

Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew *et al.*, 2011 CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. 43: 630–638.

He, A., S. W. Kong, Q. Ma, and W. T. Pu, 2011 Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc. Natl. Acad. Sci. USA 108: 5632–5637.

Hecht, A., S. Strahl-Bolsinger, and M. Grunstein, 1996 Spreading of transcriptional repressor SIR3 rom telomeric heterochromatin. Nature 383: 92–96.

Heikkinen, S., S. Väisänen, P. Pehkonen, S. Seuter, V. Benes *et al.*, 2011 Nuclear hormone 1α, 25-dihydroxyvitamin D_3 elicits a genome-wide shift in the locations of VDR chromatin occupancy. Nucleic Acids Res. 39: 9181–9193.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin *et al.*, 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38: 576–589.

Heng, J. C., B. Feng, J. Han, J. Jiang, P. Kraus *et al.*, 2010 The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. Cell Stem Cell 6: 167–174.

Ho, J. W., E. Bishop, P. V. Karchenko, N. Négre, K. P. White *et al.*, 2011 ChIP-chip *vs.* ChIP-seq: lessons for experimental design and data analysis. BMC Genomics 12: 134.

Ho, L., R. Jothi, J. L. Ronan, K. Cui, K. Zhao *et al.*, 2009 An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. Proc. Natl. Acad. Sci. USA 106: 5187–5191.

Hollenhorst, P. C., K. J. Chandler, R. L. Poulsen, W. E. Johnson, N. A. Speck *et al.*, 2009 DNA specificity determinants associate with distinct transcription factor functions. PLoS Genet. 5: e1000778.

Holmstrom, S. R., T. Deering, G. H. Swift, F. J. Poelwijk, D. J. Mangelsdorf *et al.*, 2011 LRH-1 and PTF1-L coregulate an exocrine pancreas-specific transcriptional network for digestive function. Genes Dev. 25: 1674–1679.

Horak, C. E., and M. Snyder, 2002 ChIP-chip: A genomic approach for identifying transcription factor binding sites. Methods Enzymol. 350: 469483.

Horiuchi, S., A. Onodera, H. Hosokawa, Y. Watanabe, T. Tanaka *et al.*, 2011 Genome-wide analysis reveals unique regulation of transcription of Th2-specific genes by GATA3. J. Immunol. 186: 6378–6389.

Hu, M., J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. S. Qin, 2010 On the detection and refinement of transcription factor binding sites using ChIP-Seq data. Nucleic Acids Res. 38: 2154–2167.

Hu, G., D. E. Schones, K. Cui, R. Ybarra, D. Northrup *et al.*, 2011 Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. Genome Res. 21: 1650–1658.

Hunkapiller, J., Y. Shen, A. Diaz, G. Cagney, D. McCleary *et al.*, 2012 Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. PLoS Genet. 8: e1002576.

Hutchins, A. P., S. Poulain, and D. Miranda-Saavedra, 2012 Genome-wide analysis of STAT3 binding in vivo predicts effectors of the anti-inflammatory response in macrophages. Blood 119: e110–e119.

Iyer, V. R., C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder *et al.*, 2001 Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533538.

Johannes, F., R. Wardenaar, M. Colomé-Tatché, F. Mousson, P. de Graaf *et al.*, 2010 Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. Bioinformatics 26: 1000–1006.

Johnson, D. S., A. Mortazavi, R. M. Myers, and B. Wold, 2007 Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502.

Joseph, R., Y. L. Orlov, M. Huss, W. Sun, S. L. Kong *et al.*, 2010 Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. Mol. Syst. Biol. 6: 456.

Jung, H., J. Lacombe, E. O. Mazzoni, K. F. Liem, Jr, J. Grinstein *et al.*, 2010 Global control of motor neuron topography mediated by the repressive actions of a single hox gene. Neuron 67: 781–796.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando *et al.*, 2010 Mediator and cohesin connect gene expression and chromatin architecture. Nature 467: 430–435.

Kassouf, M. T., J. R. Hughes, S. Taylor, S. J. McGowan, S. Soneji *et al.*, 2010 Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res. 20: 1064–1083.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. Genome Res. 12: 996–1006.

Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park, 2008 Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26: 1351–1359.

Kim, S. W., S. J. Yoon, E. Chuong, C. Oyolu, A. E. Wills *et al.*, 2011 Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Dev. Biol. 357: 492–504.

Kim, T. K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear *et al.*, 2010 Widespread transcription at neuronal activity-regulated enhancers. Nature 465: 182–187.

Klisch, T. J., Y. Xi, A. Flora, L. Wang, W. Li *et al.*, 2011 In vivo Atoh1 targetome reveals how a proneural transcription factor regulates cerebellar development. Proc. Natl. Acad. Sci. USA 108: 3288–3293.

Koeppel, M., S. J. van Heeringen, D. Kramer, L. Smeenk, E. Janssen-Megens *et al.*, 2011 Crosstalk between c-Jun and TAp73α/β contributes to the apoptosis-survival balance. Nucleic Acids Res. 39: 6069–6085.

Kong, S. L., G. Li, S. L. Loh, W. K. Sung, and E. T. Liu, 2011 Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. Mol. Syst. Biol. 7: 526.

Kouwenhoven, E. N., S. J. van Heeringen, J. J. Tena, M. Oti, B. E. Dutilh *et al.*, 2010 Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. PLoS Genet. 6: e1001065.

Krebs, A. R., J. Demmers, K. Karmodiya, N. C. Chang, A. C. Chang *et al.*, 2010 ATAC and Mediator coactivators form a stable complex and regulate a set of non-coding RNA genes. EMBO Rep. 11: 541–547.

Kunarso, G., N. Y. Chia, J. Jeyakani, C. Hwang, X. Lu *et al.*, 2010 Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. 42: 631–634.

Kwon, H., D. Thierry-Mieg, J. Thierry-Mieg, H. P. Kim, J. Oh *et al.*, 2009 Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors. Immunity 31: 941–952.

Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli *et al.*, 2012 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22: 1813–1831.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

Law, M. J., K. M. Lower, H. P. Voon, J. R. Hughes, D. Garrick *et al.*, 2010 ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell 143: 367–378.

Lee, B. K., A. A. Bhinge, and V. R. Iyer, 2010 Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. Nucleic Acids Res. 39: 3558–3573.

Lefterova, M. I., D. J. Steger, D. Zhuo, M. Qatanani, S. E. Mullican *et al.*, 2010 Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. Mol. Cell. Biol. 30: 2078–2089.

Li, L., R. Jothi, K. Cui, J. Y. Lee, T. Cohen *et al.*, 2010 Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. Nat. Immunol. 12: 129–136.

Li, M., Y. He, W. Dubois, X. Wu, J. Shi *et al.*, 2012 Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol. Cell 46: 30–42.

Li, Q., J. Brown, H. Huang, and P. Bickel, 2011 Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. 5: 17521779.

Lieb, J. D., X. Liu, D. Botstein, and P. O. Brown, 2001 Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat. Genet. 28: 327334.

Lin, Y. C., S. Jhunjhunwala, C. Benner, S. Heinz, E. Welinder *et al.*, 2010 A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat. Immunol. 11: 635–643.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon *et al.*, 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.

Little, G. H., H. Noushmehr, S. K. Baniwal, B. P. Berman, G. A. Coetzee *et al.*, 2011 Genome-wide Runx2 occupancy in prostate cancer cells suggests a role in regulating secretion. Nucleic Acids Res. 40: 3538–3547.

Liu, W., B. Tanasa, O. V. Tyurina, T. Y. Zhou, R. Gassmann *et al.*, 2010 PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. Nature 466: 508–512.

Liu, Z., D. R. Scannell, M. B. Eisen, and R. Tjian, 2011 Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell 146: 720–731.

Lo, K. A., M. K. Bauchmann, A. P. Baumann, C. J. Donahue, M. A. Thiede *et al.*, 2011 Genome-wide profiling of H3K56 acetylation and transcription factor binding sites in human adipocytes. PLoS ONE 6: e19778.

Lu, F., K. Tsai, H. S. Chen, P. Wikramasinghe, R. V. Davuluri *et al.*, 2012 Identification of host-chromosome binding sites and candidate gene targets for Kaposi's sarcoma-associated herpesvirus LANA. J. Virol. 86: 5752–5762.

Ma, Z., T. Swigut, A. Valouev, A. Rada-Iglesias, J. Wysocka *et al.*, 2010 Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. Nat. Struct. Mol. Biol. 18: 120–127.

MacIsaac, K. D., K. A. Lo, W. Gordon, S. Motola, T. Mazor *et al.*, 2010 A quantitative model of transcriptional regulation reveals the influence of binding location on expression. PLOS Comput. Biol. 6: e1000773.

Mahony, S., E. O. Mazzoni, S. McCuine, R. A. Young, H. Wichterle *et al.*, 2010 Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis. Genome Biol. 12: R2.

Marban, C., T. Su, R. Ferrari, B. Li, D. Vatakis *et al.*, 2011 Genome-wide binding map of the HIV-1 Tat protein to the human genome. PLoS ONE 6: e26894.

Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink *et al.*, 2008 Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134: 521–533.

Martinez, P., M. Thanasoula, A. R. Carlos, G. Gómez-López, A. M. Tejera *et al.*, 2010 Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites. Nat. Cell Biol. 12: 768–780.

Mazzoni, E. O., S. Mahony, M. Iacovino, C. A. Morrison, G. Mountoufaris *et al.*, 2011 Embryonic stem cell-based mapping of developmental transcriptional programs. Nat. Methods 8: 1056–1058.

McManus, S., A. Ebert, G. Salvagiotto, J. Medvedovic, Q. Sun *et al.*, 2011 The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. EMBO J. 30: 2388–2404.

Mendoza-Parra, M. A., M. Walia, M. Sankar, and H. Gronemeyer, 2011 Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. Mol. Syst. Biol. 7: 538.

Mendoza-Parra, M. A., M. Van Gool, M. A. Mohamed Saleem, D. G. Ceschin, and H. Gronemeyer, 2013 A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. 41: e196.

Meyer, M. B., P. D. Goetsch, and J. W. Pike, 2012 VDR/RXR and TCF4/β-catenin cistromes in colonic cells of colorectal tumor origin: impact on c-FOS and c-MYC gene expression. Mol. Endocrinol. 26: 37–51.

Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman *et al.*, 2007 Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.

Miller, T. W., J. M. Balko, E. M. Fox, Z. Ghazoui, A. Dunbier *et al.*, 2011 ERα-dependent E2F transcription can mediate resistance to estrogen deprivation in human breast cancer. Cancer Discov. 1: 338–351.

Miyazaki, M., R. R. Rivera, K. Miyazaki, Y. C. Lin, Y. Agata *et al.*, 2011 The opposing roles of the transcription factor E2A and its antagonist Id3 that orchestrate and enforce the naive fate of T cells. Nat. Immunol. 12: 992–1001.

Mullen, A. C., D. A. Orlando, J. J. Newman, J. Lovén, R. M. Kumar *et al.*, 2011 Master transcription factors determine cell-type-specific responses to TGFβ signaling. Cell 147: 565–576.

Mullican, S. E., C. A. Gaddis, T. Alenghat, M. G. Nair, P. R. Giacomin *et al.*, 2011 Histone deacetylase 3 is an epigenomic brake in macrophage alternative activation. Genes Dev. 25: 2480–2488.

Nakayamada, S., Y. Kanno, H. Takahashi, D. Jankovic, K. T. Lu *et al.*, 2011 Early Th1 cell differentiation is marked by a Tfh cell-like transition. Immunity 35: 919–931.

Nishiyama, A., L. Xin, A. A. Sharov, M. Thomas, G. Mowrer *et al.*, 2009 Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. Cell Stem Cell 5: 420–433.

Nitzsche, A., M. Paszkowski-Rogacz, F. Matarese, E. M. Janssen-Megens, N. C. Hubner *et al.*, 2011 RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. PLoS ONE 6: e19470.

Norton, L., M. Fourcaudot, M. A. Abdul-Ghani, D. Winnier, F. F. Mehta *et al.*, 2011 Chromatin occupancy of transcription factor 7-like 2 (TCF7L2) and its role in hepatic glucose metabolism. Diabetologia 54: 3132–3142.

Novershtern, N., A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining *et al.*, 2011 Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144: 296–309.

Ntziachristos, P., A. Tsirigos, P. van Vlierberghe, J. Nedjic, T. Trimarchi *et al.*, 2012 Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. Nat. Med. 18: 298–301.

Palii, C. G., C. Perez-Iratxeta, Z. Yao, Y. Cao, F. Dai *et al.*, 2010 Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. EMBO J. 30: 494–509.

Pehkonen, P., L. Welter-Stahl, J. Diwo, J. Ryynänen, A. Wienecke-Baldacchino *et al.*, 2012 Genome-wide landscape of liver X receptor chromatin binding and gene regulation in human macrophages. BMC Genomics 13: 50.

Peng, S., A. A. Alekseyenko, E. Larschan, M. I. Kuroda, and P. J. Park, 2007 Normalization and experimental design for ChIP-chip data. BMC Bioinformatics 8: 219.

Pepke, S., B. Wold, and A. Mortazavi, 2009 Computation for ChIP-seq and RNA-seq studies. Nat. Methods 6: S22–S32.

Ptasinska, A., S. A. Assi, D. Mannari, S. R. James, D. Williamson *et al.*, 2012 Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. Leukemia 26: 1829–1841.

Qi, H. H., M. Sarkissian, G. Q. Hu, Z. Wang, A. Bhattacharjee *et al.*, 2010 Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. Nature 466: 503–507.

Quenneville, S., G. Verde, A. Corsinotti, A. Kapopoulou, J. Jakobsson *et al.*, 2011 In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Mol. Cell 44: 361–372.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn *et al.*, 2010 A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470: 279–283.

Rahl, P. B., C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine *et al.*, 2010 c-Myc regulates transcriptional pause release. Cell 141: 432–445.

Ramagopalan, S. V., A. Heger, A. J. Berlanga, N. J. Maugeri, M. R. Lincoln *et al.*, 2010 A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res. 20: 1352–1360.

Ramos, Y. F., M. S. Hestand, M. Verlaan, E. Krabbendam, Y. Ariyurek *et al.*, 2010 Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. Nucleic Acids Res. 38: 5396–5408.

Rao, N. A., M. T. McCalman, P. Moulos, K. J. Francoijs, A. Chatziioannou *et al.*, 2011 Coactivation of GR and NFKB alters the repertoire of their binding sites and target genes. Genome Res. 21: 1404–1416.

Remeseiro, S., A. Cuadrado, G. Gómez-López, D. G. Pisano, and A. Losada, 2012 A unique role of cohesin-SA1 in gene regulation and development. EMBO J. 31: 2090–2102.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings *et al.*, 2000 Genome-wide location and function of DNA binding proteins. Science 290: 2306–2309.

Rey, G., F. Cesbron, J. Rougemont, H. Reinke, M. Brunner *et al.*, 2011 Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. PLoS Biol. 9: e1000595.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao *et al.*, 2007 Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods 4: 651–657.

Sadasivam, S., S. Duan, and J. A. DeCaprio, 2012 The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. Genes Dev. 26: 474–489.

Sahu, B., M. Laakso, K. Ovaska, T. Mirtti, J. Lundin *et al.*, 2011 Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. EMBO J. 30: 3962–3976.

Sakabe, N. J., I. Aneas, T. Shen, L. Shokri, S. Y. Park *et al.*, 2012 Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. Hum. Mol. Genet. 21: 2194–2204.

Schödel, J., C. Bardella, L. K. Sciesielski, J. M. Brown, C. W. Pugh *et al.*, 2012 Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. Nat. Genet. 44:420–425, S1–S2.

Schlesinger, J., M. Schueler, M. Grunert, J. J. Fischer, Q. Zhang *et al.*, 2010 The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. PLoS Genet. 7: e1001313.

Schmitz, S. U., M. Albert, M. Malatesta, L. Morey, J. V. Johansen *et al.*, 2011 Jarid1b targets genes regulating development and is involved in neural differentiation. EMBO J. 30: 4586–4600.

Schnetz, M. P., L. Handoko, B. Akhtar-Zaidi, C. F. Bartels, C. F. Pereira *et al.*, 2010 CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. PLoS Genet. 6: e1001023.

Sehat, B., A. Tofigh, Y. Lin, E. Trocmé, U. Liljedahl *et al.*, 2010 SUMOylation mediates the nuclear translocation and signaling of the IGF-1 receptor. Sci. Signal. 3: ra10.

Seitz, V., P. Butzhammer, B. Hirsch, J. Hecht, I. Gütgemann *et al.*, 2011 Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma. PLoS ONE 6: e26837.

Shen, T., I. Aneas, N. Sakabe, R. J. Dirschinger, G. Wang *et al.*, 2011 Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. J. Clin. Invest. 121: 4640–4654.

Shukla, S., E. Kavak, M. Gregory, M. Imashimizu, B. Shutinoski *et al.*, 2011 CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479: 74–79.

Siersbæk, R., R. Nielsen, S. John, M. H. Sung, S. Baek *et al.*, 2011 Extensive chromatin remodelling and establishment of transcription factor hotspots' during early adipogenesis. EMBO J. 30: 1459–1472.

Smeenk, L., S. J. van Heeringen, M. Koeppel, B. Gilbert, E. Janssen-Megens *et al.*, 2011 Role of p53 serine 46 in p53 target gene regulation. PLoS ONE 6: e17574.

Smith, E. R., C. Lin, A. S. Garrett, J. Thornton, N. Mohaghegh *et al.*, 2011 The little elongation complex regulates small nuclear RNA transcription. Mol. Cell 44: 954–965.

Soccio, R. E., G. Tuteja, L. J. Everett, Z. Li, M. A. Lazar *et al.*, 2011 Species-specific strategies underlying conserved functions of metabolic transcription factors. Mol. Endocrinol. 25: 694–706.

Solomon, M. J., P. L. Larsen, and A. Varshavsky, 1988 Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell 53: 937–947.

Stadler, M. B., R. Murr, L. Burger, R. Ivanek, F. Lienert *et al.*, 2011 DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480: 490–495.

Steger, D. J., G. R. Grant, M. Schupp, T. Tomaru, M. I. Lefterova *et al.*, 2010 Propagation of adipogenic signals through an epigenomic transition state. Genes Dev. 24: 1035–1044.

Sun, J., H. Pan, C. Lei, B. Yuan, S. J. Nair *et al.*, 2011 Genetic and genomic analyses of RNA polymerase II-pausing factor in regulation of mammalian transcription and cell growth. J. Biol. Chem. 286: 36248–36257.

Tallack, M. R., T. Whitington, W. S. Yuen, E. N. Wainwright, J. R. Keys *et al.*, 2010 A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res. 20: 1052–1063.

Tan, P. Y., C. W. Chang, K. R. Chng, K. D. Wansa, W. K. Sung *et al.*, 2011a Integration of regulatory networks by NKX3–1 promotes androgen-dependent prostate cancer survival. Mol. Cell. Biol. 32: 399–414.

Tan, S. K., Z. H. Lin, C. W. Chang, V. Varang, K. R. Chng *et al.*, 2011b AP-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. EMBO J. 30: 2569–2581.

Tang, C., X. Shi, W. Wang, D. Zhou, J. Tu *et al.*, 2010 Global analysis of in vivo EGR1-binding sites in erythroleukemia cell using chromatin immunoprecipitation and massively parallel sequencing. Electrophoresis 31: 2936–2943.

Teo, A. K., S. J. Arnold, M. W. Trotter, S. Brown, L. T. Ang *et al.*, 2011 Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes Dev. 25: 238–250.

Tijssen, M. R., A. Cvejic, A. Joshi, R. L. Hannah, R. Ferreira *et al.*, 2011a Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev. Cell 20: 597–609.

Tiwari, V. K., L. Burger, V. Nikoletopoulou, R. Deogracias, S. Thakurela *et al.*, 2011b Target genes of Topoisomerase IIβ regulate neuronal survival and are defined by their chromatin state. Proc. Natl. Acad. Sci. USA 109: E934–E943.

Tiwari, V. K., M. B. Stadler, C. Wirbelauer, R. Paro, D. Schübeler *et al.*, 2011 A chromatin-modifying function of JNK during stem cell differentiation. Nat. Genet. 44: 94–100.

Trompouki, E., T. V. Bowman, L. N. Lawton, Z. P. Fan, D. C. Wu *et al.*, 2011 Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell 147: 577–589.

Trowbridge, J. J., A. U. Sinha, N. Zhu, M. Li, S. A. Armstrong *et al.*, 2012 Haploinsufficiency of Dnmt1 impairs leukemia stem cell function through derepression of bivalent chromatin domains. Genes Dev. 26: 344–349.

van Heeringen, S. J., W. Akhtar, U. G. Jacobi, R. C. Akkers, Y. Suzuki *et al.*, 2011 Nucleotide composition-linked divergence of vertebrate core promoter architecture. Genome Res. 21: 410–421.

Vermeulen, M., H. C. Eberl, F. Matarese, H. Marks, S. Denissov *et al.*, 2010 Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. Cell 142: 967–980.

Verzi, M. P., H. Shin, H. H. He, R. Sulahian, C. A. Meyer *et al.*, 2010 Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. Dev. Cell 19: 713–726.

Verzi, M. P., H. Shin, L. L. Ho, X. S. Liu, and R. A. Shivdasani, 2011 Essential and redundant functions of caudal family proteins in activating adult intestinal genes. Mol. Cell. Biol. 31: 2026–2039.

Vilagos, B., M. Hoffmann, A. Souabni, Q. Sun, B. Werner *et al.*, 2012 Essential role of EBF1 in the generation and function of distinct mature B cell types. J. Exp. Med. 209: 775–792.

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama *et al.*, 2009 ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854–858.

Vivar, O. I., X. Zhao, E. F. Saunier, C. Griffin, O. S. Mayba *et al.*, 2010 Estrogen receptor beta binds to and regulates three distinct classes of target genes. J. Biol. Chem. 285: 22059–22066.

Wang, D., I. Garcia-Bassets, C. Benner, W. Li, X. Su *et al.*, 2011a Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474: 390–394.

Wang, H., J. Zou, B. Zhao, E. Johannsen, T. Ashworth *et al.*, 2011b Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. Proc. Natl. Acad. Sci. USA 108: 14908–14913.

Wang, J., J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield *et al.*, 2012 Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 22: 1798–1812.

Wei, L., G. Vahedi, H. W. Sun, W. T. Watford, H. Takatori *et al.*, 2010 Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. Immunity 32: 840–851.

Wei, G., B. J. Abraham, R. Yagi, R. Jothi, K. Cui *et al.*, 2011 Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. Immunity 35: 299–311.

Weinmann, A. S., P. S. Yan, M. J. Oberley, T. H. Huang, and P. J. Farnham, 2002 Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev. 16: 235244.

Welboren, W. J., M. A. van Driel, E. M. Janssen-Megens, S. J. van Heeringen, F. C. Sweep *et al.*, 2009 ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. EMBO J. 28: 1418–1428.

Whyte, W. A., S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton *et al.*, 2011 Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature 482: 221–225.

Wilson, N. K., D. Miranda-Saavedra, S. Kinston, N. Bonadies, S. D. Foster *et al.*, 2009 The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. Blood 113: 5456–5465.

Woodfield, G. W., Y. Chen, T. B. Bair, F. E. Domann, and R. J. Weigel, 2010 Identification of primary gene targets of TFAP2C in hormone responsive breast carcinoma cells. Genes Chromosomes Cancer 49: 948–962.

Wu, H., A. C. D'Alessio, S. Ito, K. Xia, Z. Wang *et al.*, 2011a Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 473: 389–393.

Wu, H., A. C. D'Alessio, S. Ito, Z. Wang, K. Cui *et al.*, 2011b Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. 25: 679–684.

Wu, J. Q., M. Seay, V. P. Schulz, M. Hariharan, D. Tuck *et al.*, 2012 Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. PLoS Genet. 8: e1002565.

Xiao, S., D. Xie, X. Cao, P. Yu, X. Xing *et al.*, 2012 Comparative epigenomic annotation of regulatory DNA. Cell 149: 1381–1392.

Xu, C., Z. P. Fan, P. Müller, R. Fogley, A. DiBiase *et al.*, 2011 Nanog-like regulates endoderm formation through the Mxtx2-Nodal pathway. Dev. Cell 22: 625–638.

Yaffe, D., and O. Saxel, 1977 Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle. Nature 270: 725–727.

Yang, Y., Y. Lu, A. Espejo, J. Wu, W. Xu *et al.*, 2010 TDRD3 is an effector molecule for arginine-methylated histone marks. Mol. Cell 40: 1016–1023.

Yang, X. P., K. Ghoreschi, S. M. Steward-Tharp, J. Rodriguez-Canales, J. Zhu *et al.*, 2011 Opposing regulation of the locus encoding IL-17 through direct, reciprocal actions of STAT3 and STAT5. Nat. Immunol. 12: 247–254.

Yao, H., K. Brick, Y. Evrard, T. Xiao, and R. D. Camerini-Otero, 2010 Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. Genes Dev. 24: 2543–2555.

Yildirim, O., R. Li, J. H. Hung, P. B. Chen, X. Dong *et al.*, 2011 Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. Cell 147: 1498–1510.

Yoon, S. J., A. E. Wills, E. Chuong, R. Gupta, and J. C. Baker, 2011 HEB and E2A function as SMAD/FOXH1 cofactors. Genes Dev. 25: 1654–1661.

Yu, M., L. Riva, H. Xie, Y. Schindler, T. B. Moran *et al.*, 2009 Insights into GATA-1-mediated gene activation *vs.* repression via genome-wide chromatin occupancy analysis. Mol. Cell 36: 682–695.

Yu, M., T. Mazor, H. Huang, H. T. Huang, K. L. Kathrein *et al.*, 2012 Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. Mol. Cell 45: 330–343.

Yu, S., K. Cui, R. Jothi, D. M. Zhao, X. Jing *et al.*, 2010 GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells. Blood 117: 2166–2178.

Yuan, P., J. Han, G. Guo, Y. L. Orlov, M. Huss *et al.*, 2009 Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes Dev. 23: 2507–2520.

Yun, K., and B. Wold, 1996 Skeletal muscle determination and differentiation: story of a core regulatory network and its context. Curr. Opin. Cell Biol. 8: 877–889.

Zhang, Y., E. V. Laz, and D. J. Waxman, 2011 Dynamic, sex-differential STAT5 and BCL6 binding to sex-biased, growth hormone-regulated genes in adult mouse liver. Mol. Cell. Biol. 32: 880–896.

Zhao, B., J. Zou, H. Wang, E. Johannsen, C. W. Peng *et al.*, 2011a Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. Proc. Natl. Acad. Sci. USA 108: 14902–14907.

Zhao, L., E. A. Glazov, D. R. Pattabiraman, F. Al-Owaidi, P. Zhang *et al.*, 2011b Integrated genome-wide chromatin occupancy and expression analyses identify key myeloid pro-differentiation transcription factors repressed by Myb. Nucleic Acids Res. 39: 4664–4679.

Communicating editor: T. R. Hughes

L

Evidence for site-specific occupancy of the mitochondrial genome by nuclear transcription factors

Originally published as:

Marinov GK*, Wang YE*, Wold BJ, Chan DC. Evidence for site-specific occupancy of the mitochondrial genome by nuclear transcription factors. *PLoS ONE*. 2014. **9**(1):e84713. doi: 10.1371/journal.pone.0084713.

Evidence for Site-Specific Occupancy of the Mitochondrial Genome by Nuclear Transcription Factors

Georgi K. Marinov[1*☯], Yun E. Wang[1☯], David Chan[1,2], Barbara J. Wold[1]

1 Division of Biology, California Institute of Technology, Pasadena, California, United States of America, 2 Howard Hughes Medical Institute, Pasadena, California, United States of America

Abstract

Mitochondria contain their own circular genome, with mitochondria-specific transcription and replication systems and corresponding regulatory proteins. All of these proteins are encoded in the nuclear genome and are post-translationally imported into mitochondria. In addition, several nuclear transcription factors have been reported to act in mitochondria, but there has been no comprehensive mapping of their occupancy patterns and it is not clear how many other factors may also be found in mitochondria. Here we address these questions by using ChIP-seq data from the ENCODE, mouseENCODE and modENCODE consortia for 151 human, 31 mouse and 35 *C. elegans* factors. We identified 8 human and 3 mouse transcription factors with strong localized enrichment over the mitochondrial genome that was usually associated with the corresponding recognition sequence motif. Notably, these sites of occupancy are often the sites with highest ChIP-seq signal intensity within both the nuclear and mitochondrial genomes and are thus best explained as true binding events to mitochondrial DNA, which exist in high copy number in each cell. We corroborated these findings by immunocytochemical staining evidence for mitochondrial localization. However, we were unable to find clear evidence for mitochondrial binding in ENCODE and other publicly available ChIP-seq data for most factors previously reported to localize there. As the first global analysis of nuclear transcription factors binding in mitochondria, this work opens the door to future studies that probe the functional significance of the phenomenon.

Citation: Marinov GK, Wang YE, Chan D, Wold BJ (2014) Evidence for Site-Specific Occupancy of the Mitochondrial Genome by Nuclear Transcription Factors. PLoS ONE 9(1): e84713. doi:10.1371/journal.pone.0084713

Editor: Frances M. Sladek, Univeristy of California Riverside, United States of America

Received August 20, 2013; **Accepted** November 18, 2013; **Published** January 20, 2014

Funding: This study has been supported by the Beckman Institute Functional Genomics Center, the Donald Bren Endowment, and NIH grants U54 HG004576, U54 HG006998, and GM062967. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: georgi@caltech.edu

☯ These authors contributed equally to this work.

Introduction

Mitochondria are the primary site of ATP production through oxidative phosphorylation and are therefore critical to eukaryotic cells. It is widely accepted that they arose as the result of an endosymbiotic event [63] between the ancestor of modern eukaryotes and a member of the α-proteobacteria clade [82]. Reflective of the organelle's prokaryotic ancestry, mitochondria retain their own reduced circular genome [55], although its size has been greatly reduced in many eukaryotes through transfer of genes to the eukaryotic nucleus. After transcription and translation of nuclear components of the separate mitochondrial transcription, replication and regulatory machineries, a number of which retain evidence of their prokaryotic origin [74], the protein products are then imported back into the mitochondria to modulate organellar function.

The mitochondrial genome in mammals encodes 13 proteins, all of which are components of the electron transport chain, as well as 22 tRNAs and two rRNAs [3,5]. Mitochondrial DNA (mtDNA) is organized in cells as macromolecular DNA-protein complexes called nucleoids. Mitochondrial genes are densely packed along the genome with the notable exception of the non-coding displacement loop (D-loop) regulatory region [66], which is located within the non-coding region (NCR). Transcription initiates in the D-loop, is carried out by the mitochondrial-specific RNA polymerase POLRMT, and results in long polycistronic transcripts from each strand (called the Heavy- or H-strand and the Light- or L-strand), from the light strand promoter (LSP) and two Heavy strand promoters (HSP1 and HSP2) [9,52]. In addition, the transcription factors mtTFA/TFAM [27,28] and mtTFB2/TFB2M as well as the methyltransferase mtTFB1/TFB1M [26,29,49] are required for initiation and regulation of transcription [69]. Unlike many of the proteins involved in regulation of the mitochondrial genome, these transcription factors are generally accepted as not being of prokaryotic origin. Instead, they are genes of eukaryotic ancestry, appropriated for their function through co-evolution of the organellar and cellular genomes and imported into mitochondria to regulate mtDNA transcription.

In addition to these well-characterized regulators of mitochondrial transcription, multiple reports have suggested that transcription factors that typically act in the nucleus might also have regulatory functions in mitochondrial transcription [44,73]. The glucocorticoid receptor (GR) was the first such factor reported to localize to mitochondria and to interact with mtDNA [18,19,40,59]. A 43 kDa isoform of the thyroid hormone T_3 receptor $T_3R\alpha1$ called p43 has been found to directly control mitochondrial transcription [11,24,25,81]. Cyclic-AMP Response element Binding protein (CREB) has been shown to localize to

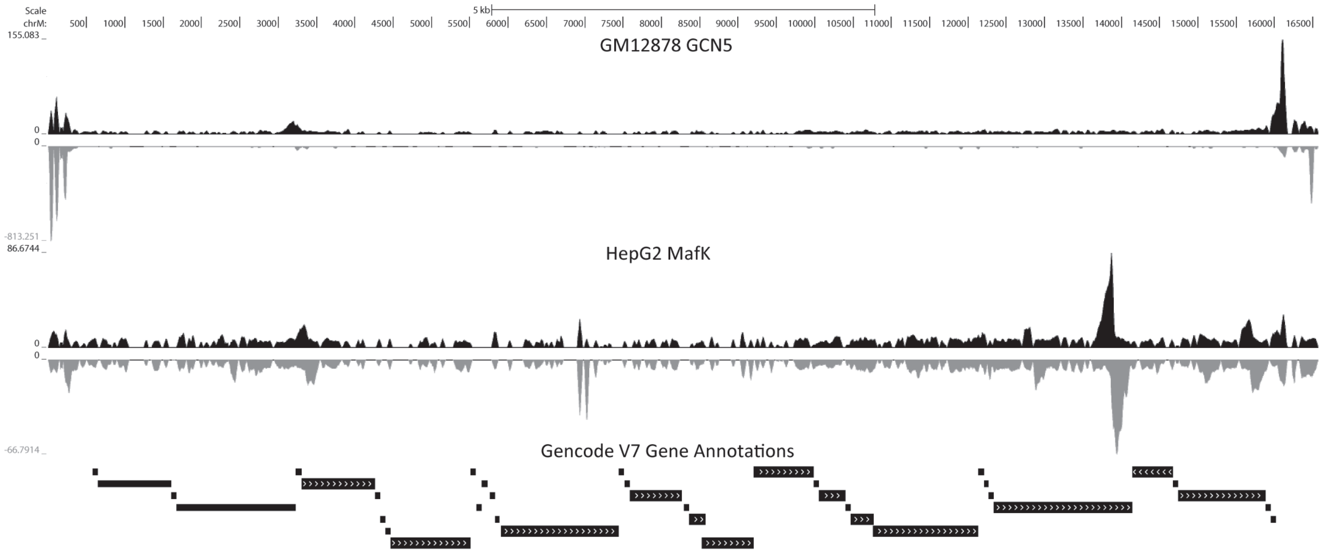


Figure 1. Representative USCS Genome Browser snapshots of nuclear transcription factor ChIP-seq datasets exhibiting strong enrichment in the mitochondrial genome. (A) GM12878 GCN5 shows high signal intensity in the D-loop (the region between coordinates 16030 and 580, i.e. the non-coding regions on the left and right ends of the snapshot) representative of the D-loop enrichment observed for a large number of transcription factors (B) In contrast, a large MafK peak is observed in a coding region outside of the D-loop in HepG2 cells. Upper track (black) shows reads aligning to the forward strand, lower track (gray) shows read aligning to the reverse strand.
doi:10.1371/journal.pone.0084713.g001

mitochondria and suggested to bind to the D-loop [8,17,43,62]. The tumor suppressor transcription factor p53 has been implicated in mtDNA repair and regulation of gene expression through interactions with TFAM [1,34,47,48,83]. It has also been proposed to play a proapoptotic role through association with the outer mitochondrial membrane [76]. A similar role has been also ascribed to the IRF3 transcription factor [12,46]. The mitochondrial localization of the estrogen receptor (ER) is also well established, for both its ERα and ERβ isoforms, and it too has been suggested to bind to the D-loop [13,51]. NFκB and IκBα have been found in mitochondria and have been proposed to regulate mitochondrial gene expression [16,36]. The AP-1 and PPARγ2 transcription factors have been proposed to localize to mitochondria and bind to the genome. [10,57,58] and the MEF2D transcription factor was found to regulate the expression of the ND6 gene by binding to a consensus sequence recognition motif within it [67]. Finally, the presence of STAT3 in mitochondria has been found to be important for the function of the electron transport chains and also to be necessary for TNF-induced necroptosis [32,68,71,72,79], although direct mtDNA binding has not been established. Mitochondrial localization has also been reported for STAT1 and STAT5 [6,14].

However, direct *in vivo* chromatin immunoprecipitation evidence for the binding of these factors to mtDNA exists only for CREB [43], p53 [1] and MEF2D [67], and with the exception of MEF2D characterization is limited to the D-loop region. No prior studies have assayed transcription factor occupancy across the entire mitochondrial genome in vivo with modern high resolution techniques such as ChIP-seq (Chromatin Immunoprecipitation coupled with deep sequencing, [35]). As a result, the precise nature, and in many instances the existence, of the proposed binding events remains unknown. The limited sampling of transcription factors in previous studies also leaves uncertain how common or rare localization to mitochondria and binding to mtDNA is for nuclear transcription factors in general.

Here we survey the large compendium of ChIP-seq and other functional genomic data made publicly available by the ENCODE, mouseENCODE and modENCODE Consortia [22,23,30,50,54] to identify transcription factors that associate directly with mtDNA and to characterize the nature of these interactions. We identify eight human and three mouse transcription factors for which robust evidence of site-specific occupancy in the mitochondrial genome exists. These sites exhibit the strand asymmetry typical of nuclear transcription factor binding sites, usually contain the recognition motifs for the factors in question, and are typically the strongest (as measured by ChIP-seq signal strength) binding sites found in both the nuclear and mitochondrial genome by a wide margin. Notably, these interactions are all found outside of the non-coding D-loop region. The D-loop region itself exhibits widespread sequencing read enrichment for dozens of transcription factors. However, it does not show the aforementioned feature characteristics of true binding events. Though not observed in control datasets generated from sonicated input DNA, the high ChIP-seq signal over the D-loop is frequently seen in control datasets generated using mock immunoprecipitation, suggesting that it is likely to represent an experimental artifact. Examination of available ChIP-seq data for the transcription factors previously proposed to play a role in mitochondria (GR, ERα, CREB, STAT3, p53) revealed no robust binding sites except for enrichment in the D-loop. Resolving the functional significance of the identified occupancy sites in future studies should provide exciting insights into the biology of both mitochondrial and nuclear transcriptional regulation.

Results

In the course of a study of TFAM occupancy in the mitochondrial and nuclear genomes [78], we noticed that a number of nuclear transcription factors exhibit localized enrichment in certain areas of the mitochondrial genome in ChIP-seq data (Figure 1). These events could be divided in two classes: high ChIP-seq signal over the NCR, and localized high read density over regions outside of it. Given prior reports suggesting that nuclear transcription factors might act in mitochondria, this

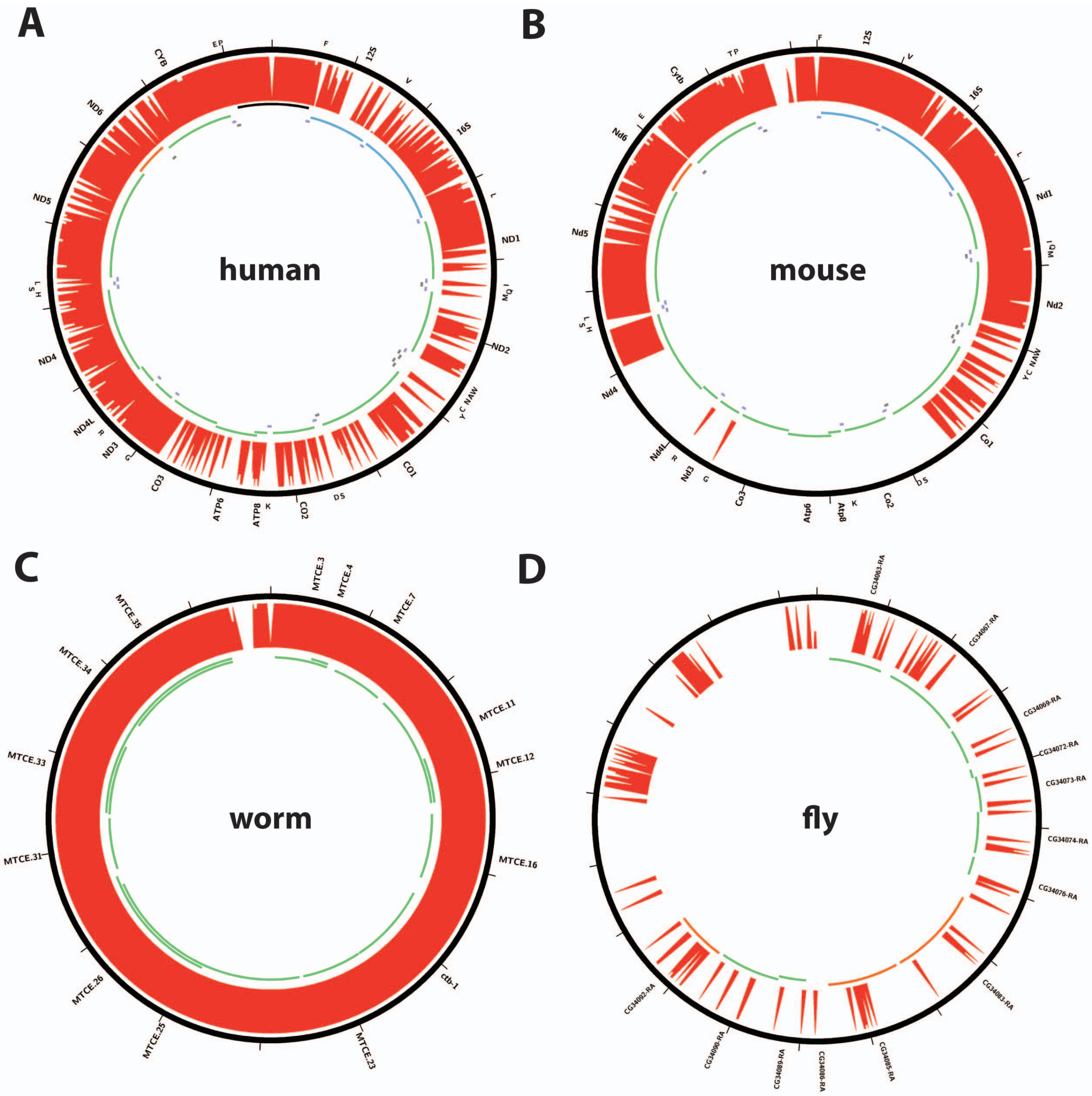


Figure 2. Unique mappability of the mitochondrial genome (chrM) in ENCODE and modENCODE species. (A) human; (B) mouse; (C) *C. elegans*; (D) *D. melanogaster*. The 36 bp mappability track (see Methods for details) is shown. The annotated protein coding and rRNA and tRNA genes are shown in the inner circles as follows: forward-strand genes are shown as green lines, while reverse-strand genes are shown as red lines, with the exception of mouse and human rRNA and tRNAs (blue). The D-loop region in human is shown in black. Gene annotations were obtained from ENSEMBL (version 66). Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g002

prompted us to determine the general prevalence of the phenomenon among transcription factors and investigate evidence of occupancy in detail, as the power and resolution of ChIP-seq have not previously been brought to bear on this somewhat mysterious phenomenon. We took advantage of the wide compendium of human, mouse, fly and worm functional genomics data generated by the ENCODE [22,23], mouseENCODE [54] and modENCODE [30,50] consortia.

Identifying transcription factor binding events in the mitochondrial genome

We downloaded publicly available (as of February 2012) ENCODE and mouseENCODE ChIP-seq and control data from the UCSC Genome Browser and modENCODE data from ftp://ftp.modencode.org, including ChIP-seq data for 151 transcription factors in human cell lines [77], 31 in mouse and 35 in *C.elegans* (see discussion on *D. melanogaster* below). We also downloaded DNase hypersensitvity (both DNase-seq [75] and Digital Genomic

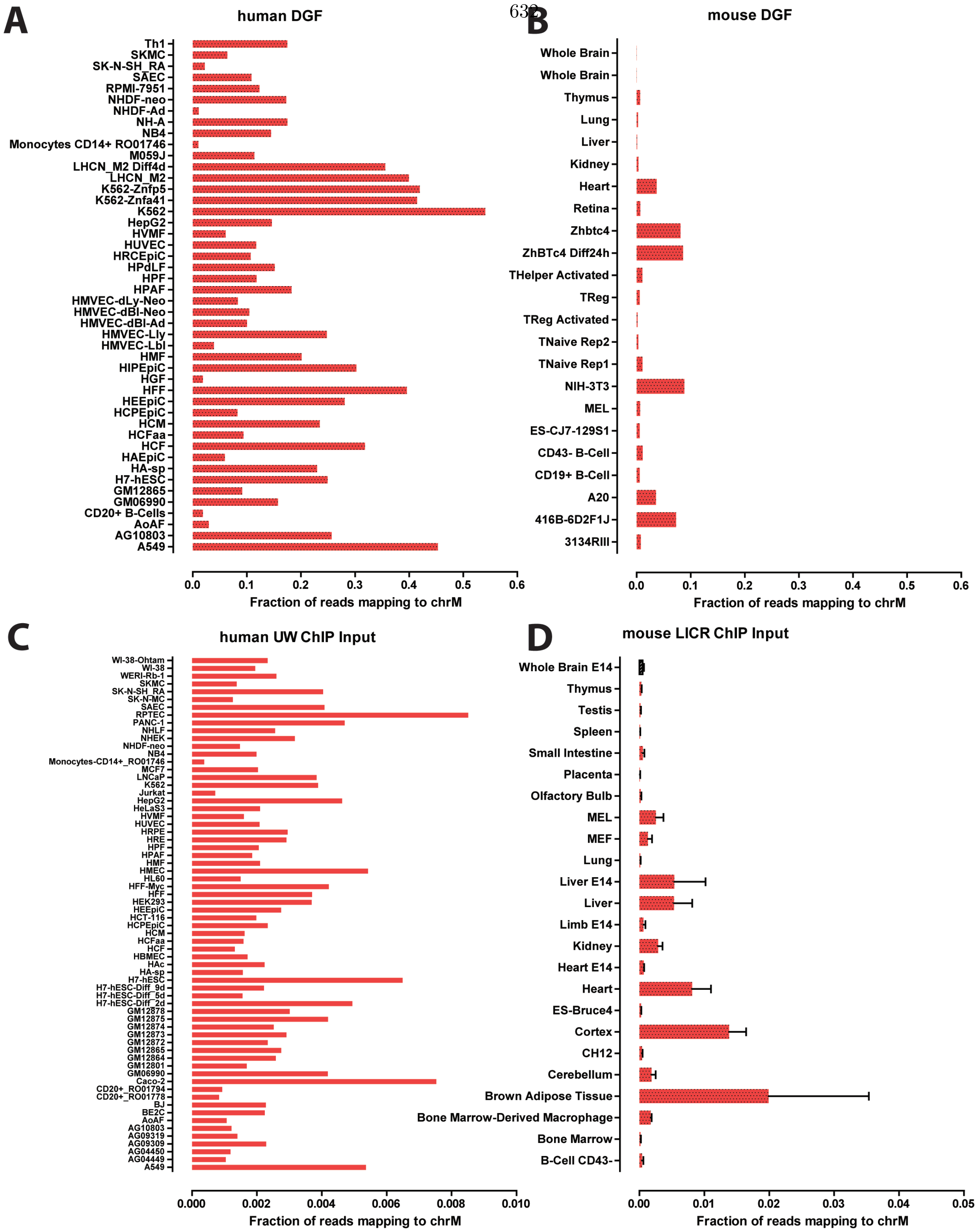


Figure 3. Variation in mitochondrial DNA copy number in cell lines and tissues. The fraction of reads mapping to the mitochondrial genome (chrM) is shown. (A,B) UW human (A) and mouse (B) UW ENCODE digital genomic footprinting (DGF) data; (C) UW human ChIP input datasets; (D) LICR mouse ChIP input datasets. ''UW'' and ''LICR'' refers to the ENCODE production groups that generated the data. Inputs from the UW and LICR groups were chosen because they are the largest ENCODE sets in terms of number of cell lines/tissues assayed by the same production groups, thus avoiding possible variation between different laboratories. A general positive correlation between the expected metabolic demand of the tissue type and the relative amount of reads mapping to chrM is observed.
doi:10.1371/journal.pone.0084713.g003

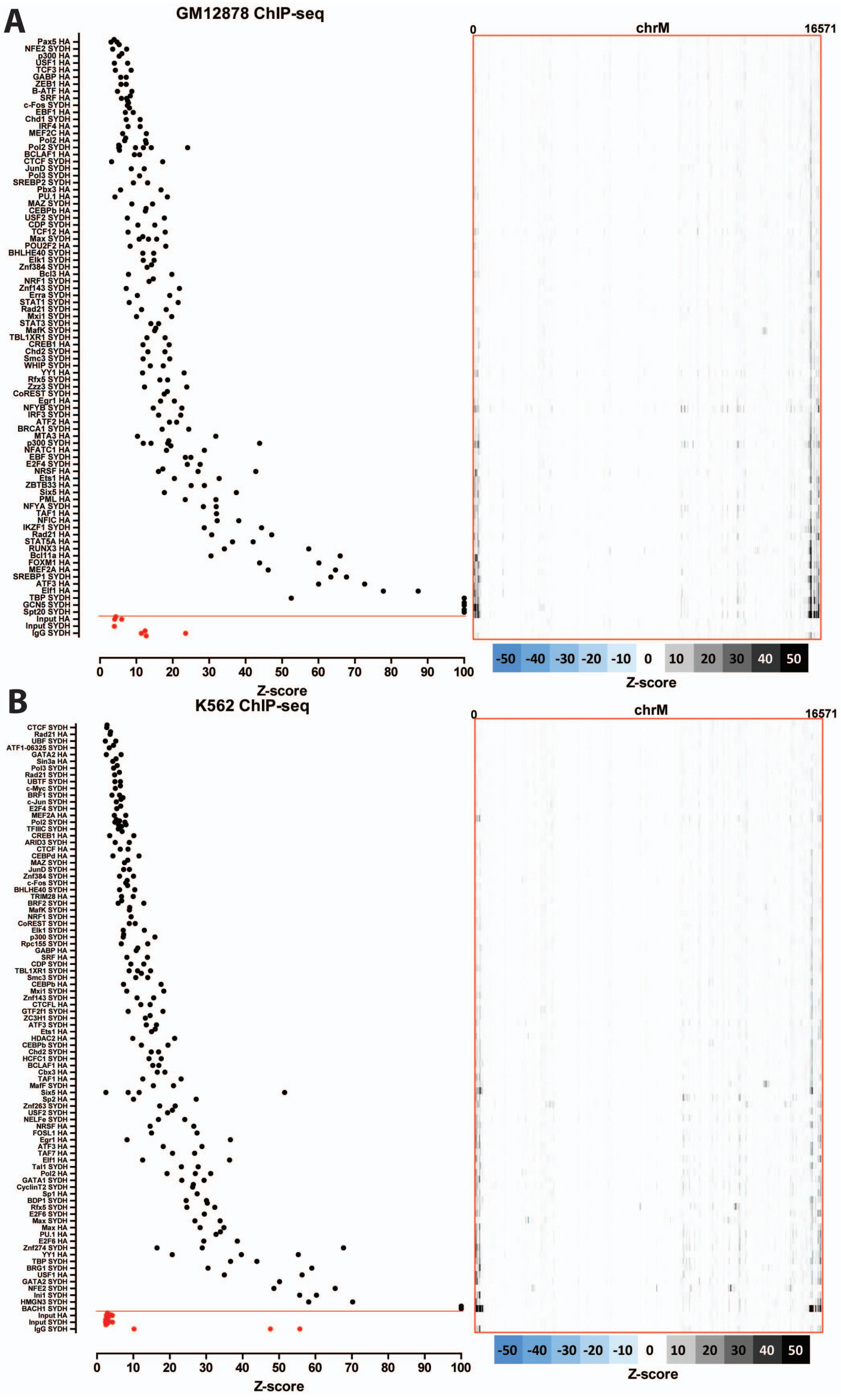

A
GM12878 ChIP-seq
Pax5 HA
NFE2 SYDH
p300 HA
USF1 HA
TCF3 HA
GABP HA
ZEB1 HA
B-ATF HA
SRF HA
c-Fos SYDH
EBF1 HA
Chd1 SYDH
IRF4 HA
MEF2C HA
Pol2 HA
Pol2 SYDH
BCLAF1 HA
CTCF SYDH
JunD SYDH
Pol3 SYDH
SREBP2 SYDH
Pbx3 HA
PU.1 HA
MAZ SYDH
CEBPb HA
USF2 SYDH
CDP SYDH
TCF12 HA
Max SYDH
POU2F2 HA
BHLHE40 SYDH
Elk1 SYDH
Znf384 SYDH
Bcl3 HA
NRF1 SYDH
Znf143 SYDH
Erra SYDH
STAT1 SYDH
Rad21 SYDH
Mxi1 SYDH
STAT3 SYDH
MafK SYDH
TBL1XR1 SYDH
CREB1 HA
Chd2 SYDH
Smc3 SYDH
WHIP SYDH
YY1 HA
Rfx5 SYDH
Zzz3 SYDH
CoREST SYDH
Egr1 HA
NFYB SYDH
IRF3 SYDH
ATF2 HA
BRCA1 SYDH
MTA3 HA
p300 SYDH
NFATC1 HA
EBF SYDH
E2F4 SYDH
NRSF HA
Ets1 HA
ZBTB33 HA
Six5 HA
PML HA
NFYA SYDH
TAF1 HA
NFIC HA
IKZF1 SYDH
Rad21 HA
STAT5A HA
RUNX3 HA
Bcl11a HA
FOXM1 HA
MEF2A HA
SREBP1 SYDH
ATF3 HA
Elf1 HA
TBP SYDH
GCN5 SYDH
Spt20 SYDH
Input HA
Input SYDH
IgG SYDH
0
10
20
30
40
50
60
70
80
90
100
Z-score
0
chrM
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score
B
K562 ChIP-seq
CTCF SYDH
Rad21 HA
UBF SYDH
ATF1-06325 SYDH
GATA2 HA
Sin3a HA
Pol3 SYDH
Rad21 SYDH
UBTF SYDH
c-Myc SYDH
BRF1 SYDH
c-Jun SYDH
E2F4 SYDH
MEF2A HA
Pol2 SYDH
TFIIIC SYDH
CREB1 HA
ARID3 SYDH
CTCF HA
CEBPd HA
MAZ SYDH
JunD SYDH
Znf384 SYDH
c-Fos SYDH
BHLHE40 SYDH
TRIM28 HA
BRF2 SYDH
MafK SYDH
NRF1 SYDH
CoREST SYDH
Elk1 SYDH
p300 SYDH
Rpc155 SYDH
GABP HA
SRF HA
CDP SYDH
TBL1XR1 SYDH
Smc3 SYDH
CEBPb HA
Mxi1 SYDH
Znf143 SYDH
CTCFL HA
GTF2f1 SYDH
ZC3H1 SYDH
ATF3 SYDH
Ets1 HA
HDAC2 HA
CEBPb SYDH
Chd2 SYDH
HCFC1 SYDH
BCLAF1 HA
Cbx3 HA
TAF1 HA
MafF SYDH
Six5 HA
Sp2 HA
Znf263 SYDH
USF2 SYDH
NELFe SYDH
NRSF HA
FOSL1 HA
Egr1 HA
ATF3 HA
TAF7 HA
Elf1 HA
Tal1 SYDH
Pol2 HA
GATA1 SYDH
CyclinT2 SYDH
Sp1 HA
BDP1 SYDH
Rfx5 SYDH
E2F6 SYDH
Max SYDH
Max HA
PU.1 HA
E2F6 HA
Znf274 SYDH
YY1 HA
TBP SYDH
BRG1 SYDH
USF1 HA
GATA2 SYDH
NFE2 SYDH
Ini1 SYDH
HMGN3 SYDH
BACH1 SYDH
Input HA
Input SYDH
IgG SYDH
0
10
20
30
40
50
60
70
80
90
100
Z-score
0
chrM
16571
-50
-40
-30
-20
-10
0
10
20
30
40
50
Z-score

Figure 4. Signal distribution over the mitochondrial genome in human ChIP-seq datasets. The maximum z-score for each individual TF ChIP-seq replicate in each cell line is shown on the left (factors are sorted by average z-score, with control datasets always shown on the bottom in red, below the red horizontal line). The z-score profile along the mitochondrial chromosome for the replicate with the highest z-score is shown on the right. "SYDH" and "HA" refer to the ENCODE production groups which generated the data. Z-scores ≥100 are shown as equal to 100. (A) GM12878 cells; (B) K562 cells.
doi:10.1371/journal.pone.0084713.g004

Footprinting (DGF) [56]), FAIRE-seq (Formaldehyde Assisted Isolation of Regulatory Elements) [70] and MNase-seq data as these datasets provide valuable orthogonal information about potentially artifactual patterns of read enrichment over the mitochondrial genome.

It is well known that the nuclear genome contains partial copies of the mitochondrial genome (**NU**clear **M**i**T**ochondrial sequences or NUMTs) [20,33]. Depending on their levels of divergence from the mitochondrial sequence, they can present an informatics challenge for distinguishing binding events to the true mitochondrial genome from binding events to NUMTs. For this reason, we aligned reads simultaneously against the nuclear and mitochondrial genomes. We then retained only reads that map uniquely, and with no mismatches, relative to the reference for further analysis (see Methods for details). As a consequence this stringent mapping strategy, regions of the mitochondrial genome that are also present as perfectly identical copies in the nuclear genome are "invisible" to our analysis; this was a necessary compromise in order to focus only on a maximally stringent set of putative mitochondrial binding events. However, before proceeding, we examined how widely affected the mitochondrial genome is by this treatment in the four relevant species by generating mappability tracks (shown in Figure 2). The human mitochondrial genome contains numerous small islands of unmappable sequence, particularly concentrated between the ND1 and CO3 genes, but it displays no large completely unmappable segments (Figure 2A). The mouse genome contains a large unmappable stretch between the CO1 and ND4 genes (Figure 2B). The *C. elegans* mitochondrial genome is almost completely uniquely mappable (Figure 2C). In contrast, the *D. melanogaster* genome is almost completely unmappable, indicating the presence of very recent insertions into the nuclear genome with high sequence similarity. We therefore excluded fly datasets from further analysis and focused on human, mouse and worm data.

Mammalian cells typically contain hundreds to thousands of copies of mtDNA, with the precise number varying depending on the metabolic needs of the particular cell type [7,64,80]. This variation is relevant to our analysis because the relative read density over the mitochondrial genome is expected to scale with the mtDNA:nuclear DNA ratio for a given cell. Thus, cell types with very high mtDNA copy number are expected to display correspondingly elevated background read density over the mitochondrial genome. Several types of ENCODE data provide a rough proxy for the relative mitochondrial genome copy number per cell. In particular, the fraction of reads originating from the mitochondrial genome in DNase hypersensitivity and ChIP control datasets is expected to scale accordingly. We examined the distribution of this fraction in ENCODE and mouseENCODE DGF datasets and observed very large differences between different cell lines and tissues (Figure 3). For example, about half of reads in K562 DGF data originated from mitochondria, while the fraction was less than 2% in CD20+ B-cells (Figure 3A). Notably, these differences are in many cases (though not always) consistent with what is known about the cell lines, with certain cancer cell lines (such as K562 and A549) and muscle cells (LHCN) showing the largest number of mitochondrial reads, while primary cells with small volumes of cytoplasm such as B-cells showed the least.

Mouse DGF data was available mostly for tissues, and the fraction of mitochondrial reads in these was much smaller compared to both the human cell lines and the few mouse cell lines assayed (Figure 3B). This is consistent with a significant proportion of cells in tissues being in a less active metabolic state than cell lines in culture. Still, we observed expected differences between tissues. For example, one of the tissues that was most enriched for reads mapping to the mitochondrial genome was the heart. We observed similarly large differences in ChIP control datasets (Figure 3CD), although the absolute number of reads was much lower than it was in DGF data. Again, the mouse tissues with the highest number of mitochondrial reads were the more metabolically active ones, such as brown adipose tissue, cortex, and heart.

These large differences in background read coverage between different cells lines/tissues have two consequences for the analysis of putative transcription factor binding to the mitochondrial genome. First, peak calling algorithms usually used to identify transcription factor binding sites from ChIP-seq data may not work equally well in different cell lines due to the highly variable background read density. Second, these differences render comparing the strength of binding across cell lines difficult.

We therefore devised a normalization procedure (described in Methods) to convert read coverage to signal intensity z-scores reflecting how strongly regions of enrichment stand out compared to the average background read density along the mitochondrial genome for each dataset. We then used the maximum z-scores for each dataset to identify datasets with very strong such enrichment, which we then examined manually in detail.

Nuclear transcription factor binding to the mitochondrial genome in human cell lines

The distribution of read density z-scores for transcription factor ChIP-seq and control datasets in seven ENCODE human cell lines (GM1278, K562, HepG2, HeLa, H1-hESC, IMR90 and A549) is shown in Figures 4, 5 and 6. A wide range in the values of the maximum z-score is observed, from less than 5, to more than 100. Strikingly, most factors exhibit high read density in the NCR. One obvious explanation for this observation is that it represents an experimental artifact. This is likely, as the NCR contains the D-loop [66], the unique triple-strand structure of which could conceivably either cause overrepresentation of DNA fragments originating from it in sequencing libraries or it could be non-specifically bound by antibodies during the immunoprecipitation process. To distinguish between these possibilities, we carried out the same analysis on DNase, FAIRE and MNase data. As these assays do not involve an immunoprecipitation step, they are a proper control for sequencing artifacts. We did not observe significant localized read enrichment in these datasets (Figure 7), suggesting that the observed read enrichment over the D-loop is not due to sequencing biases or overrepresentation of D-loop fragments in ChIP libraries. Similarly, we did not observe enrichment in the matched sonicated input ChIP-seq control datasets. However, a number of mock-immunoprecipitation (IgG) control datasets did exhibit high z-scores (up to >50 in K562 cells)

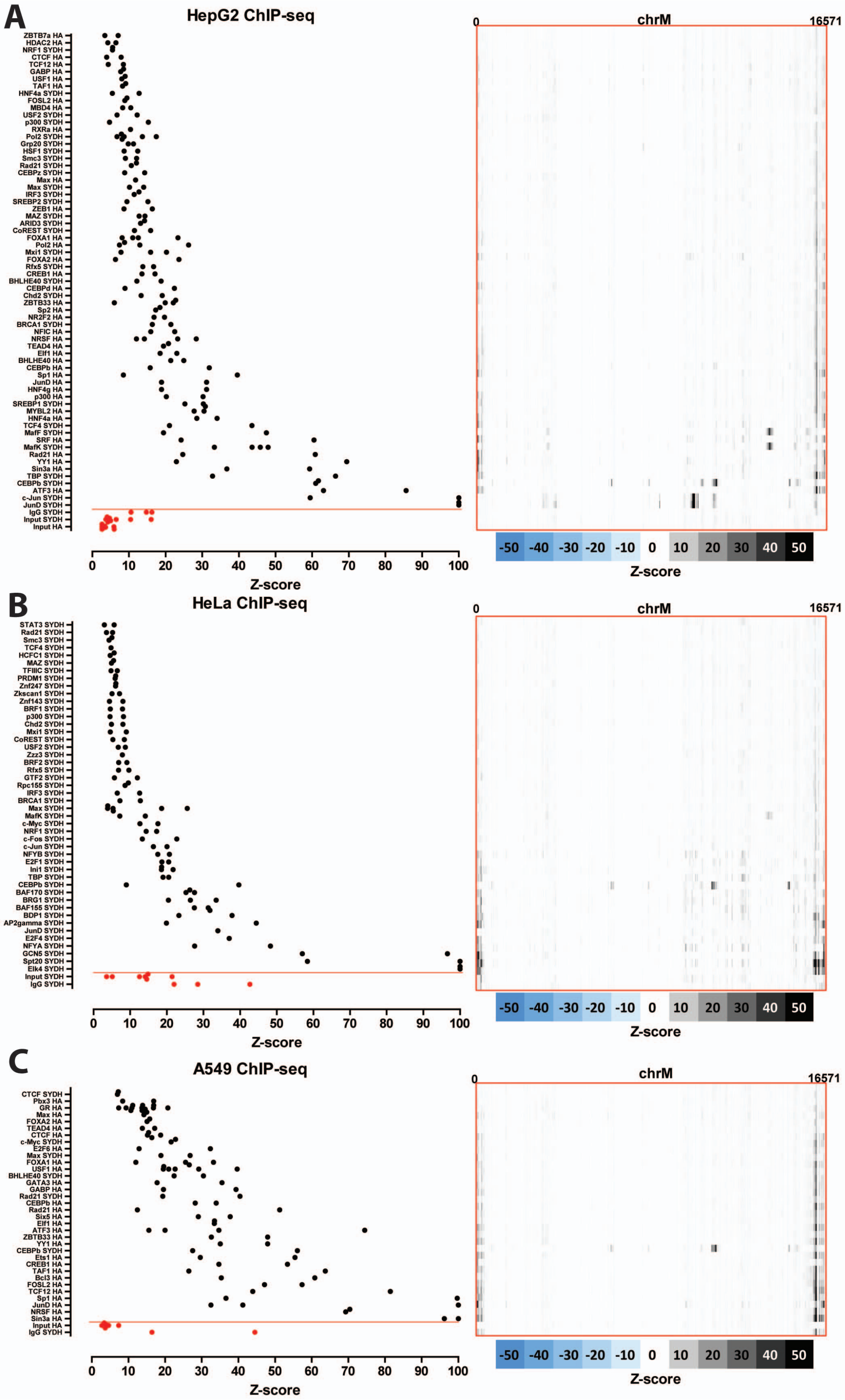

A
HepG2 ChIP-seq
ZBTB7a HA
HDAC2 HA
NRF1 SYDH
CTCF HA
TCF12 HA
GABP HA
USF1 HA
TAF1 HA
HNF4a SYDH
FOSL2 HA
MBD4 HA
USF2 SYDH
p300 SYDH
RXRa HA
Pol2 SYDH
Grp20 SYDH
HSF1 SYDH
Smc3 SYDH
Rad21 SYDH
CEBPz SYDH
Max HA
Max SYDH
IRF3 SYDH
SREBP2 SYDH
ZEB1 HA
MAZ SYDH
ARID3 SYDH
CoREST SYDH
FOXA1 HA
Pol2 HA
Mxi1 SYDH
FOXA2 HA
Rfx5 SYDH
CREB1 HA
BHLHE40 SYDH
CEBPd HA
Chd2 SYDH
ZBTB33 HA
Sp2 HA
NR2F2 HA
BRCA1 SYDH
NFIC HA
NRSF HA
TEAD4 HA
Elf1 HA
BHLHE40 HA
CEBPb HA
Sp1 HA
JunD HA
HNF4g HA
p300 HA
SREBP1 SYDH
MYBL2 HA
HNF4a HA
TCF4 SYDH
MafF SYDH
SRF HA
MafK SYDH
Rad21 HA
YY1 HA
Sin3a HA
TBP SYDH
CEBPb SYDH
ATF3 HA
c-Jun SYDH
JunD SYDH
IgG SYDH
Input SYDH
Input HA
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score
B
HeLa ChIP-seq
STAT3 SYDH
Rad21 SYDH
Smc3 SYDH
TCF4 SYDH
HCFC1 SYDH
MAZ SYDH
TFIIIC SYDH
PRDM1 SYDH
Znf247 SYDH
Zkscan1 SYDH
Znf143 SYDH
BRF1 SYDH
p300 SYDH
Chd2 SYDH
Mxi1 SYDH
CoREST SYDH
USF2 SYDH
Zzz3 SYDH
BRF2 SYDH
Rfx5 SYDH
GTF2 SYDH
Rpc155 SYDH
IRF3 SYDH
BRCA1 SYDH
Max SYDH
MafK SYDH
c-Myc SYDH
NRF1 SYDH
c-Fos SYDH
c-Jun SYDH
NFYB SYDH
E2F1 SYDH
Ini1 SYDH
TBP SYDH
CEBPb SYDH
BAF170 SYDH
BRG1 SYDH
BAF155 SYDH
BDP1 SYDH
AP2gamma SYDH
JunD SYDH
E2F4 SYDH
NFYA SYDH
GCN5 SYDH
Spt20 SYDH
Elk4 SYDH
Input SYDH
IgG SYDH
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score
C
A549 ChIP-seq
CTCF SYDH
Pbx3 HA
GR HA
Max HA
FOXA2 HA
TEAD4 HA
CTCF HA
c-Myc SYDH
E2F6 HA
Max SYDH
FOXA1 HA
USF1 HA
BHLHE40 SYDH
GATA3 HA
GABP HA
Rad21 SYDH
CEBPb HA
Rad21 HA
Six5 HA
Elf1 HA
ATF3 HA
ZBTB33 HA
YY1 HA
CEBPb SYDH
Ets1 HA
CREB1 HA
TAF1 HA
Bcl3 HA
FOSL2 HA
TCF12 HA
Sp1 HA
JunD HA
NRSF HA
Sin3a HA
Input HA
IgG SYDH
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score

Figure 5. Signal distribution over the mitochondrial genome in human ChIP-seq datasets. The maximum z-score for each individual TF ChIP-seq replicate in each cell line is shown on the left (factors are sorted by average z-score, with control datasets always shown on the bottom in red, below the red horizontal line). The z-score profile along the mitochondrial chromosome for the replicate with the highest z-score is shown on the right. "SYDH" and "HA" refer to the ENCODE production groups which generated the data. Z-scores ≥100 are shown as equal to 100. (A) HepG2 cells; (B) HeLa cells; (C) A549 cells.
doi:10.1371/journal.pone.0084713.g005

and closely matched the signal profile over the D-loop of ChIP-seq datasets (Figure 8B). We also examined the forward and reverse strand read distribution in the NCR (Figure 8). Site-specific transcription factor binding events display a characteristic asymmetry in the distribution of reads mapping to the forward and reverse strands, with reads on the forward strand showing a peak to the left of the binding site and reads on the reverse strand showing a peak to the right of it [39] (Figure 8C). Such read asymmetry was not observed in the D-loop region (average profile shown in Figure 8A, individual dataset profile shown in Figure 1).

These results suggest that while immunoprecipitation is necessary for high enrichment over the D-loop, the enrichment might not be mediated by the proteins targeted by the primary antibody. This does not explain why a large number of factors show little enrichment over the D-loop (Figures 4, 5 and 6) and why some factors show enrichment that is much higher than that observed in K562 IgG controls, with z-scores of up to 300 (compared to a maximum of 50 for the most highly enriched IgG controls). Still, given the lack of clear hallmarks of site-specific occupancy, and the IgG control results, enrichment over the D-loop has to be provisionally considered to be primarily the result of an experimental artifact, even if it cannot be ruled that at least in some cases it is the result of real biochemical association with nuclear transcriptional regulators.

In contrast to the widespread, but likely artifactual, read enrichment over the D-loop, we observed strong enrichment, exhibiting the canonical characteristics of a ChIP-seq peak over a true transcription factor binding site, in other regions of the

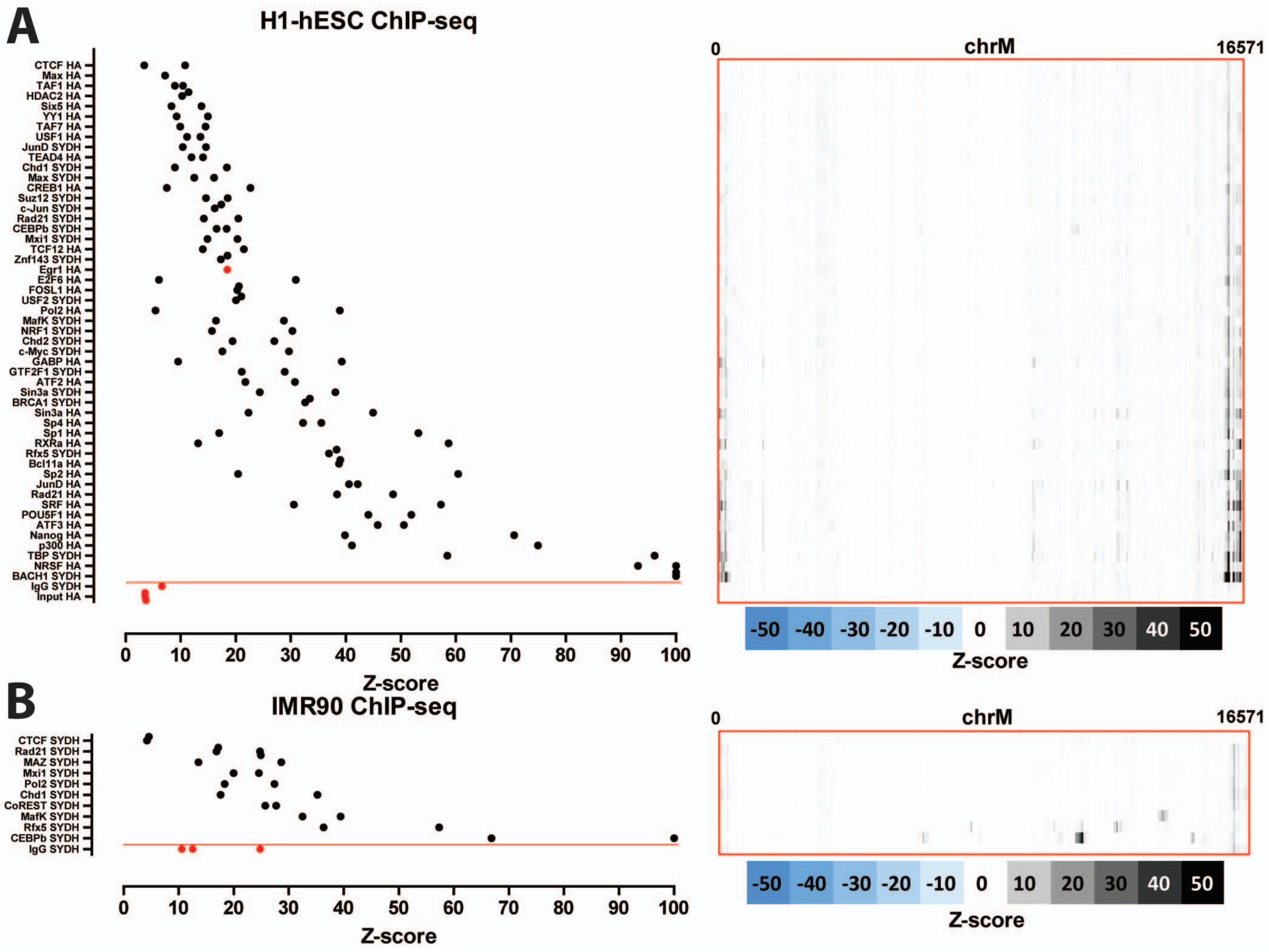


Figure 6. Signal distribution over the mitochondrial genome in human ChIP-seq datasets. The maximum z-score for each individual TF ChIP-seq replicate in each cell line is shown on the left (factors are sorted by average z-score, with control datasets always shown on the bottom in red, below the red horizontal line). The z-score profile along the mitochondrial chromosome for the replicate with the highest z-score is shown on the right. "SYDH" and "HA" refer to the ENCODE production groups which generated the data. Z-scores ≥100 are shown as equal to 100. (A) H1-hESC cells; (B) IMR90.
doi:10.1371/journal.pone.0084713.g006

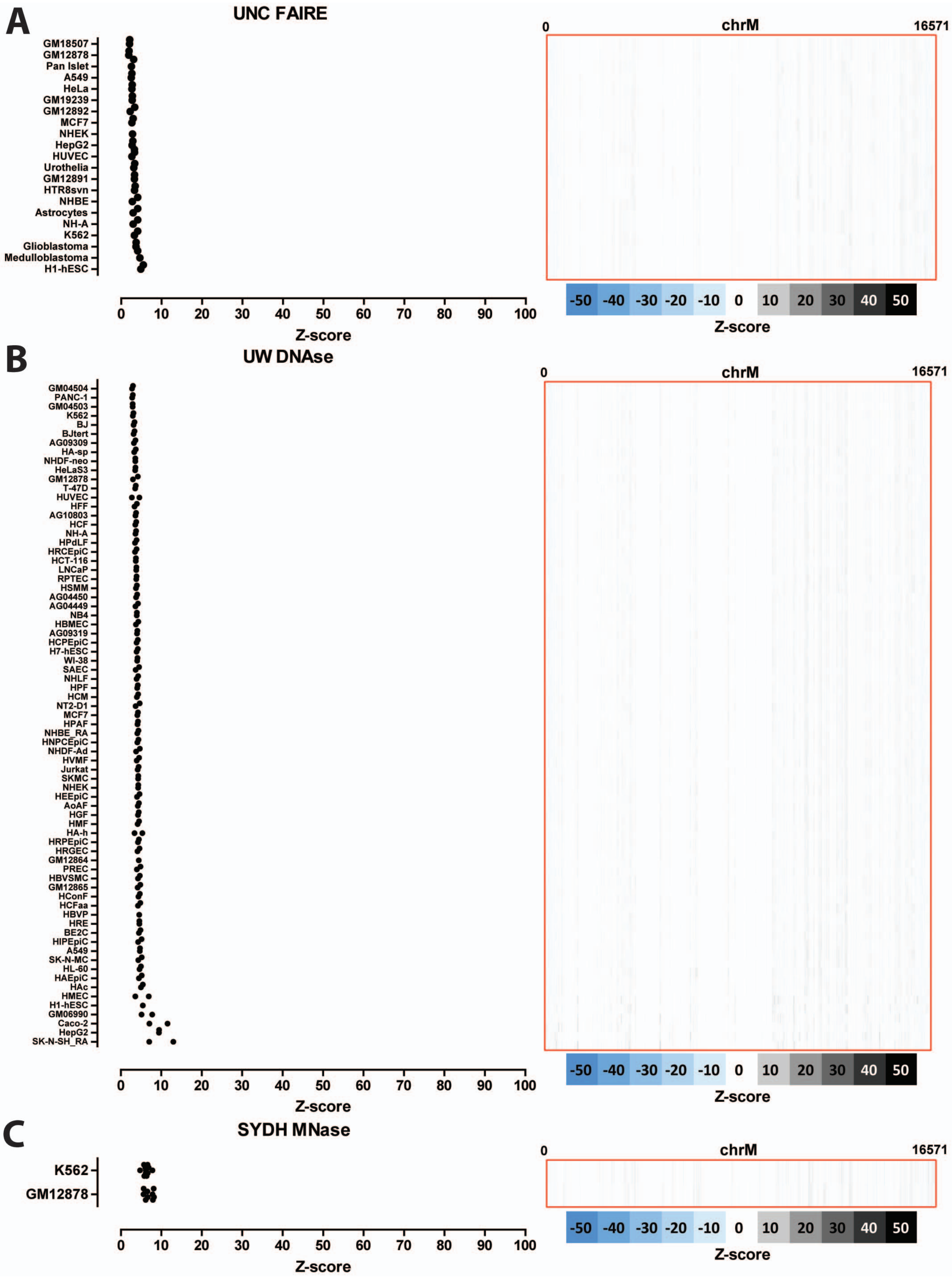

A
UNC FAIRE
GM18507
GM12878
Pan Islet
A549
HeLa
GM19239
GM12892
MCF7
NHEK
HepG2
HUVEC
Urothelia
GM12891
HTR8svn
NHBE
Astrocytes
NH-A
K562
Glioblastoma
Medulloblastoma
H1-hESC
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score
B
UW DNAse
GM04504
PANC-1
GM04503
K562
BJ
BJtert
AG09309
HA-sp
NHDF-neo
HeLaS3
GM12878
T-47D
HUVEC
HFF
AG10803
HCF
NH-A
HPdLF
HRCEpiC
HCT-116
LNCaP
RPTEC
HSMM
AG04450
AG04449
NB4
HBMEC
AG09319
HCPEpiC
H7-hESC
WI-38
SAEC
NHLF
HPF
HCM
NT2-D1
MCF7
HPAF
NHBE_RA
HNPCEpiC
NHDF-Ad
HVMF
Jurkat
SKMC
NHEK
HEEpiC
AoAF
HGF
HMF
HA-h
HRPEpiC
HRGEC
GM12864
PREC
HBVSMC
GM12865
HConF
HCFaa
HBVP
HRE
BE2C
HIPEpiC
A549
SK-N-MC
HL-60
HAEpiC
HAc
HMEC
H1-hESC
GM06990
Caco-2
HepG2
SK-N-SH_RA
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score
C
SYDH MNase
K562
GM12878
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score

Figure 7. Signal distribution over the mitochondrial genome in human FAIRE-seq, DNAse-seq and MNAse-seq datasets. Shown is the maximum z-score for each individual replicate for each cell line (left) and the z-score profile along the mitochondrial chromosome for the replicate with the highest z-score (right). (A) FAIRE data; (B) DNAse data; (C) MNAse data. "UNC", "UW" and "SYDH" refer to the ENCODE production groups which generated the data. Z-scores larger than 100 are shown as 100. No read enrichment over the D-loop is observed, suggesting that the D-loop signal found in TF ChIP-seq datasets is not due to sequencing biases but is a result of the immunoprecipitation process.
doi:10.1371/journal.pone.0084713.g007

human mitochondrial genome for eight of the examined transcription factors using a minimum z-score threshold of 20: CEBPβ, c-Jun, JunD, MafF, MafK, Max, NFE2 and Rfx5. Figures 9 and 10 show the forward and reverse strand read distribution for representative replicates of each factor in each assayed cell line, as well as the occurrences of the corresponding explanatory motifs (identified from the top 500 ChIP-seq peaks in the nuclear genome, see Methods for details). The putative binding sites outside of the D-loop are characterized by an asymmetric forward and reverse strand read distribution, and in most cases, the presence of the explanatory motif in a position consistent with binding by the factor. We identified multiple binding sites for CEBPβ: a strong site of enrichment around the 5′ end of the CYB gene, what seems to be two closely clustered sites in the ND4 gene, a weaker site in the ND4L gene, and two other regions of enrichment over CO2 and CO1 (Figure 9D). A single very strong binding site over the ND3 gene was observed for c-Jun, as well as two weaker sites, one coinciding with the ND4 CEBPβ sites and one near the 5′ end of ATP6 (Figure 9B); the strong ND3 site was also observed for JunD in HepG2 cells. Max exhibited two putative binding sites: one in the middle of the 16S rRNA gene, containing a cluster of Max motifs, and another one around the 5′ end of CO3, which also contains a cluster of Max motifs but is in a region of poor mappability. A common and very strong MafK and MafF binding site is present near the 3′ end of ND5, though it does not contain the common explanatory motif for both factors (Figure 10AB). Several putative binding sites were identified for NFE2: one close to the CEBPβ site in the 5′end of CYB, one over the tRNA cluster between ND4 and ND5, one in the 5′ end of ATP6 and one in the 16S rRNA gene (Figure 10C). Finally, two putative binding sites ar observed for Rfx5, at the 5′ end of ND5 and in the middle of CO2 (Figure 10D). Intriguingly, these binding events are not always present in all cell lines. For example, CEBPβ binding around CYB was absent in K562, A549 and H1-hESC cells, while the MafK ND5 binding site was absent in GM18278 and H1-hESC cells, but present in the other cell lines for which data is available.

Nuclear transcription factor occupancy to the mitochondrial genome in model organisms

We carried out the same analysis as described above on mouse and *C. elegans* ChIP-seq datasets. Figure 11 shows the distribution of read density z-scores in mouse CH12 and MEL cells. Similarly to the human data, we observe widespread but probably artifactual read enrichment over the D-loop. In addition to that, we saw that three transcription factors (Max, MafK, and USF2) also exhibit strong enrichment elsewhere in the mitochondrial genome (Figure 12). We observe a single MafK binding site, containing the explanatory motif and situated over the tRNA cluster between the ND2 and CO1 genes (Figure 12A). Max displayed a strong binding site (possibly a cluster of closely spaced binding sites) in the ND4 gene, and a weaker binding site near the 5′ end of ND5; both sites contained the explanatory motif (Figure 12B). Finally, a single site, also containing the explanatory motif for the factor and situated near the ND5 Max site, was present in CH12 USF2 datasets (but not in MEL cells) (Figure 12C). MafK and Max were also assayed in human cells, and, as discussed above, putative mitochondrial sites were identified there for both, though not at obviously orthologous to those found in the mouse data positions in the genome. We also analyzed available ChIP-seq data for the mouse orthologs of c-Jun and JunD, which in human cells exhibited putative mitochondrial binding sites. In contrast to observation in human, we did not detect strong sites for either protein in mouse.

Unlike the mouse and human datasets, most *C. elegans* ChIP-seq datasets did not show very strong enrichment over the mitochondrial genome (Figure 13A), with the exception of DPY-27 and W03F9.2. Of these, only W03F9.2 exhibited regions of enrichment with the characteristics of transcription factor binding sites (Figure 13B); however, very little is known about this protein and the significance of its binding to the mitochondrial genome is unclear.

ChIP-seq signal is significantly stronger over mitochondrial occupancy sites than it is over nucleus sites

The occupancy observations reported above for human and mouse mitochondria do not formally rule out the possibility that there are unannotated NUMTs in the genomes of the cell lines in which binding is detected in our analysis and the observed binding is in fact nuclear. Such an explanation is superficially likely, given that binding to the mitochondrial genome was observed in some cell lines and not in others. However, closer examination reveals that this hypothesis would require different NUMTs in different cell lines as the cell lines that lack binding are not the same for all factors. For example, MafF and MafK binding is very prominent in K562 cells but CEBPβ and c-Jun seem not to bind to mtDNA in those cells. While still possible, we consider the independent insertion of multiple partial NUMTs in different cell lines to be an unlikely explanation for the observed binding patterns.

Each chromosome in the nuclear genome exists as only two copies in diploid cells, as compared to the hundreds of mitochondria, each of which may contain multiple copies of the mitochondrial genome [7,64], and although cancer cells may exhibit various aneuploidies and copy number variants, the number of mtDNA copies is still expected to be much higher. Thus, higher read density over mitochondrial transcription factor binding sites than over nuclear ones is expected, assuming similar occupancy rates. We therefore used the strength of ChIP-seq signal over mitochondrial occupancy sites in order to test the hypothesis that they are in fact nuclear, and not mitochondrial in origin. We compared the peak height (in **R**eads **P**er **M**illion, RPM) of the top 10 nuclear peaks (peak calls generated by the ENCODE consortium were downloaded from the UCSC Genome Browser) with that of the putatively mitochondrial binding sites (Figure 14). We found that the mitochondrial binding sites are usually the strongest binding sites by a wide margin, or at least within the top three of all peaks. For example, while the strongest nuclear MafK peak in mouse CH12 cells has a peak height of 14.5 RPM, the mitochondrial binding site has a peak height of 290 RPM. These observations are difficult to explain as being the result of binding to unannotated NUMTs in the nuclear genome, but are entirely consistent with the hypothesis that these

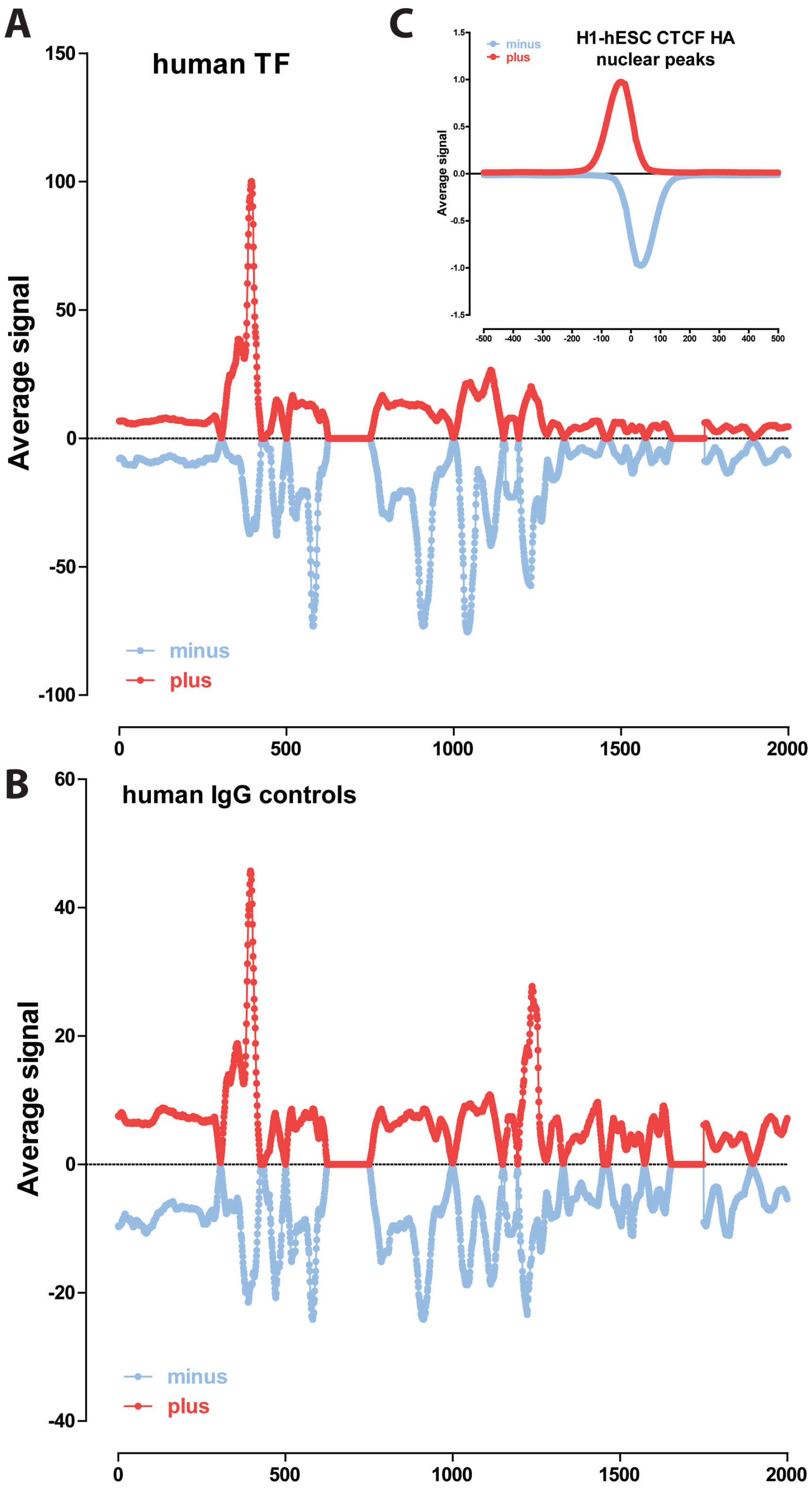


Figure 8. Combined signal distribution profile for the forward and reverse strand in the D-loop region. Shown is the average signal (in RPM) for each strand in human ChIP-seq datasets with z-scores ≥20 (A) and human IgG controls (B). Also shown for comparison is the plus and minus strand read distribution around nuclear CTCF binding sites in H1-hESC cells (C).

doi:10.1371/journal.pone.0084713.g008

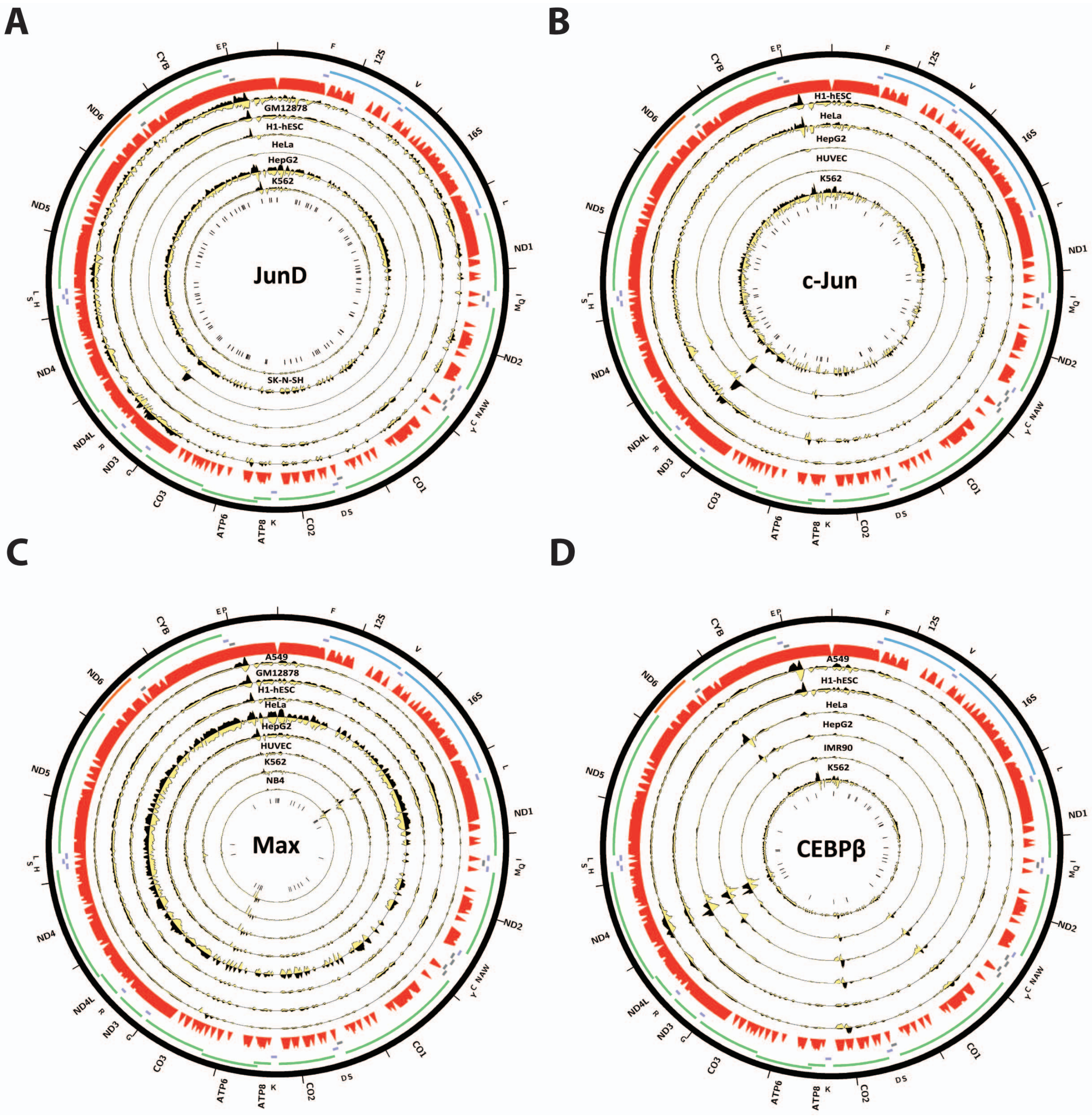


Figure 9. Human transcription factors with canonical ChIP-seq peaks (displaying the typical strand asymmetry in read distribution around the putative binding site) outside of the D-loop. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) JunD (B) c-Jun; (C) Max; (D) CEBPβ. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g009

factors indeed bind to the large number of copies of the mitochondrial genome present in each cell.

Evidence for localization of transcription factors to mitochondria

If the observed binding sites in ChIP-seq data are the result of actual association of nuclear transcription factors with mtDNA, then these transcription factors should exhibit mitochondrial localization. We directly tested this by performing immunocytochemistry (ICC) for MafK in HepG2 cells (Figure 15). It is important to note that such an assay for localization to mitochondria is potentially difficult to interpret if binding is the result of only a few protein molecules entering mitochondria, which would not yield sufficient signal for interpretation via ICC.

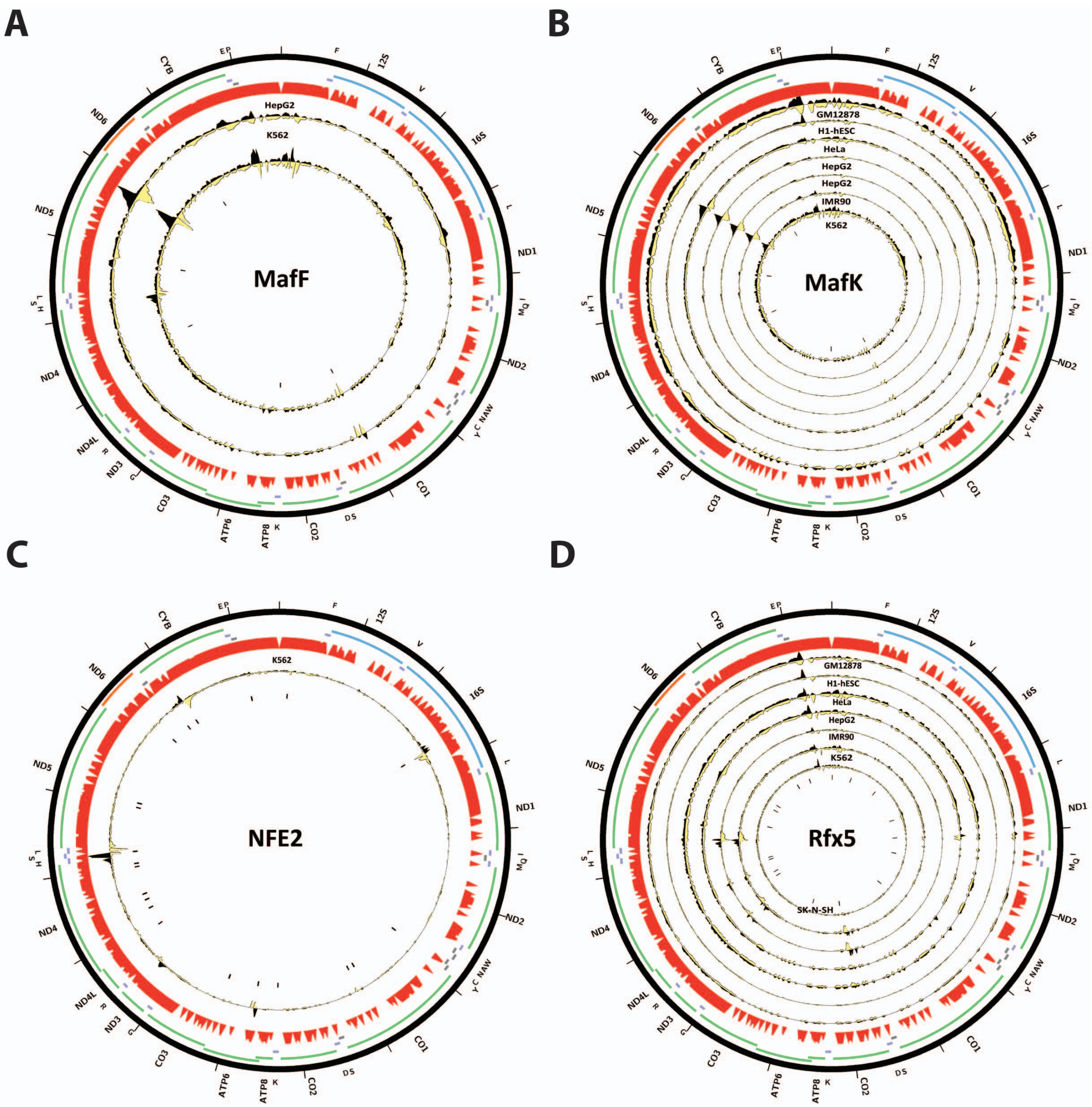


Figure 10. Human transcription factors with canonical ChIP-seq peaks (displaying the typical strand asymmetry in read distribution around the putative binding site) outside of the D-loop. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) MafF; (B) MafK (note that MafK has been assayed using two different antibodies in HepG2, both of which are shown); (C) NFE2; (D) Rfx5. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g010

However, strikingly, we observe clear colocalization of MafK to mitochondira in 60% of cells ($n = 124$). These observations provide independent corroboration for the mtDNA binding events identified through ChIP-seq.

No robust mitochondrial occupancy in ChIP-seq data for most previously reported mitochondrially targeted nuclear factors

We note that none of the factors previously reported to be localized to mitochondria and to bind to mtDNA was retrieved by our analysis, even though CREB, GR, ERα, IRF3, NFκB,

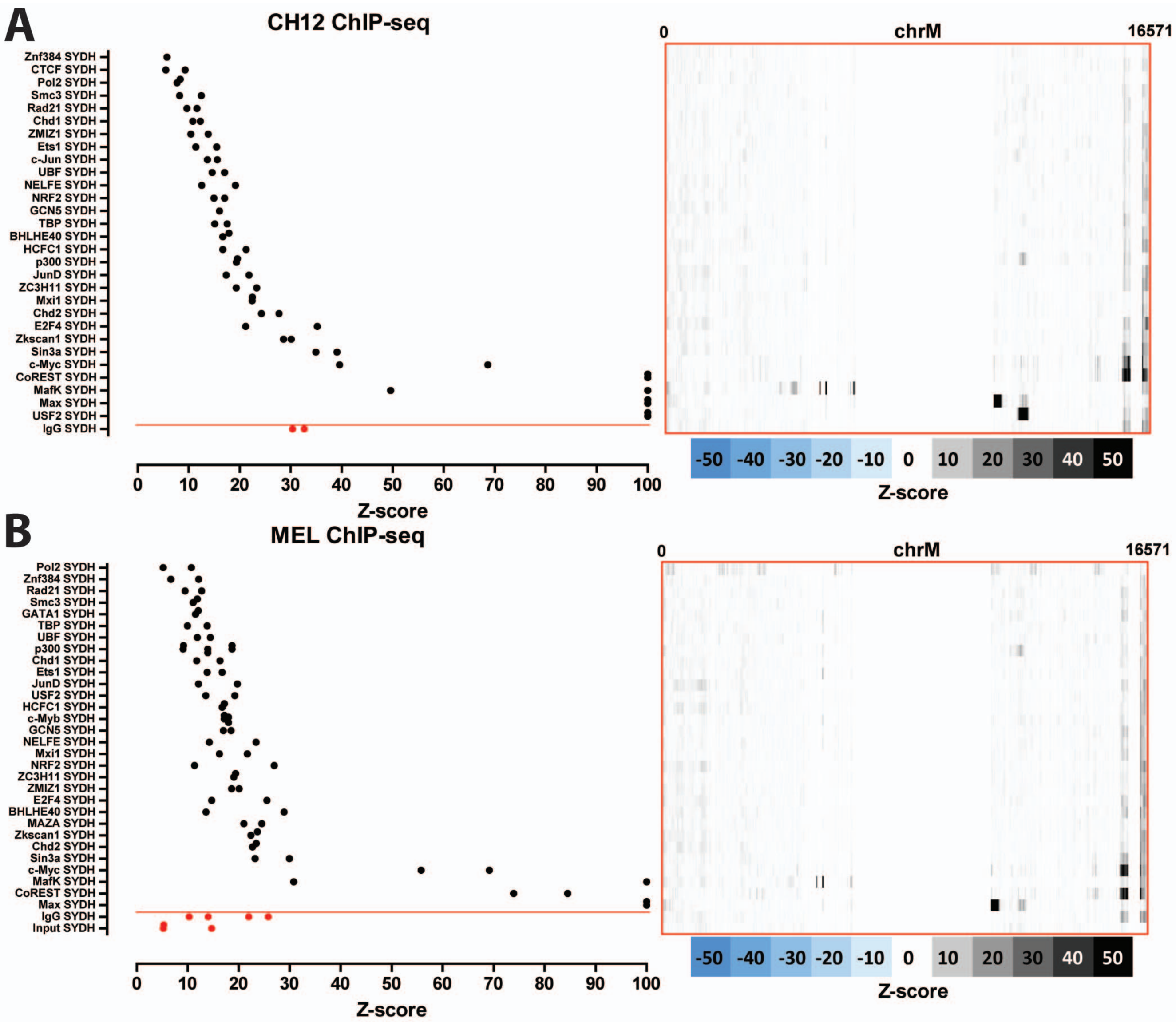


Figure 11. Signal distribution over the mitochondrial genome in mouse ChIP-seq datasets. Shown is the maximum z-score for each individual replicate for each cell line (left) and the z-score profile along the mitochondrial chromosome for the replicate with the highest z-score (right). Control datasets are shown in red on the bottom, below the red horizontal line. (A) CH12 cells; (B) MEL cells.
doi:10.1371/journal.pone.0084713.g011

STAT1, STAT5A and STAT3 were assayed by the ENCODE Consortium. This failure could be attributed to the use of too stringent a z-score threshold when selecting datasets with significant enrichment. We therefore examined available ChIP-seq data against these factors more carefully (Figure 16, Figure S1). We also performed the same analysis on published mouse and human p53 ChIP-seq data [2,38,45] (Figure 17). Again, we did not observe any major sites of enrichment outside of the D-loop. For these factors, the D-loop region exhibits the same putatively artifactual pattern discussed previously. And for STAT3 and p53, even the enrichment over the D-loop was low. The one factor for which binding to mtDNA is confirmed by ChIP-seq is MEF2D, data for two of the isoforms of which in mouse C2C12 myoblasts was recently published [65] (Figure 18). It exhibits a very complex binding pattern over large portions of the mouse mitochondrial genome, which is not straightforward to interpet, but nevertheless a number of locations exhibit strand asymmetry and contain the MEF2 sequence recognition motif. Notably, most of these are outside the ND6 gene.

It is at present not clear how to interpret these discrepancies. It is not surprising that some of these factors do not exhibit binding to mtDNA, as they were reported to play a role in mitochondrial biology through mechanisms other than regulating gene expression (for example, IRF3 and STAT3). However, this is not the case for all of them. One possibility is that many prior studies reporting physical association of transcription factors with the D-loop suffered from the same artifactual read enrichment over that region that we observe, but this would not have been noticeable using the methods of the time. This would not be surprising, as it is only apparent that D-loop enrichment is likely to be artifactual when the high spatial resolution of ChIP-seq is combined with the joint analysis of input and mock immunoprecipitation controls. However, the mitochondrial localization of these factors has been carefully documented in a number of cases [8,11,17]. Another

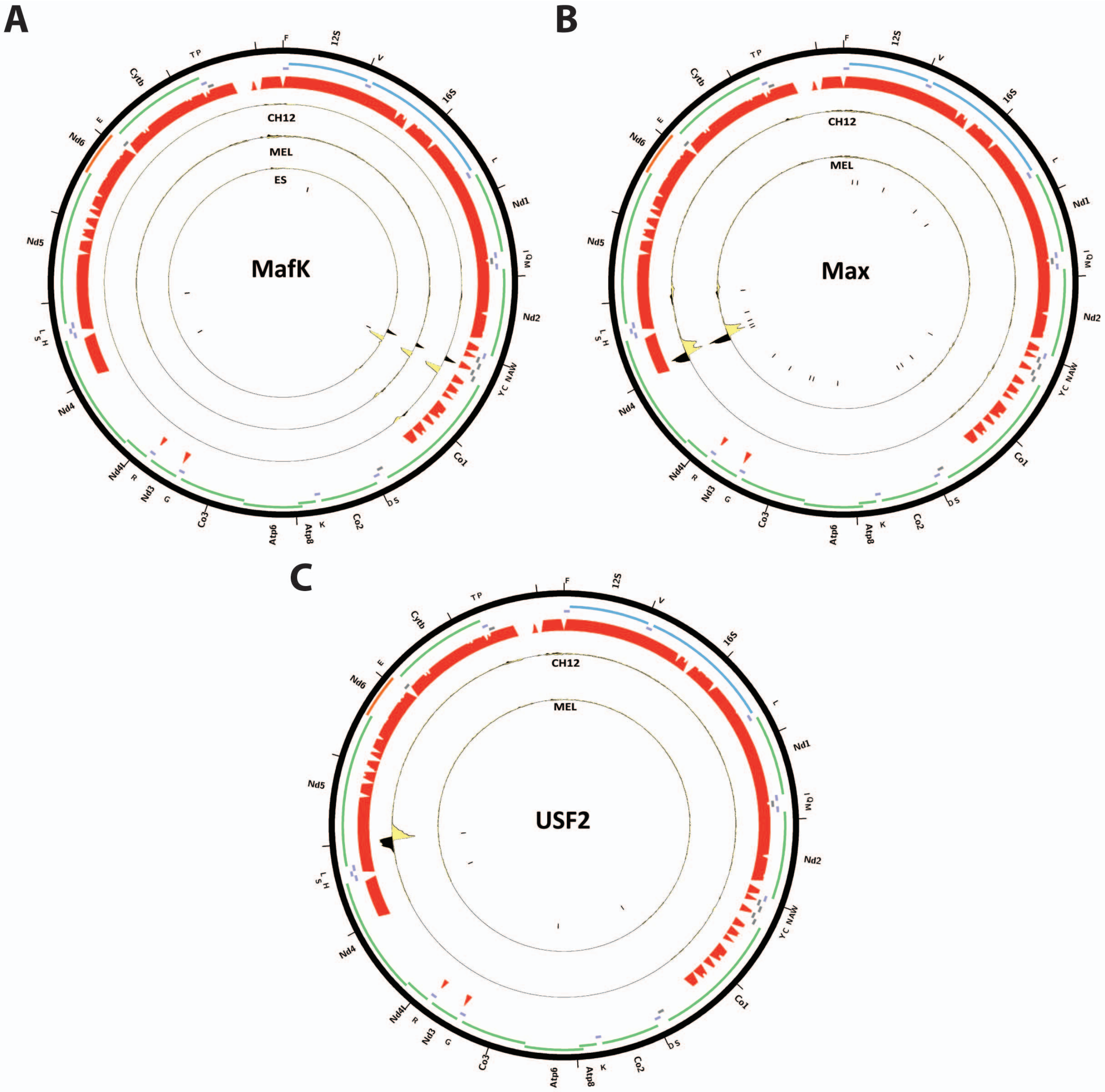


Figure 12. Mouse transcription factors with canonical ChIP-seq peaks (displaying the typical strand asymmetry in read distribution around the putative binding site) outside of the D-loop. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) MafK (note that the putative binding site is found in a region that is not completely mappable, thus the read profiles loses the canonical shape but the strand asymmetry is nevertheless apparent and a motif is present); (B) Max; (C) USF2. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g012

possiblity is that binding to mtDNA only occurs under certain physiological conditions and the factors were assayed using ChIP-seq only in cellular states not matching those. Further analysis of ChIP-seq data collected over a wide range of conditions should help resolve these issues.

Discussion

We report here the first large-scale characterization of the association of nuclear transcription factors along the entire mitochondrial genome by utilizing the vast ChIP-seq data resource made publicly available by the ENCODE and modENCODE consortia. We find two classes of signal enrichment events, neither

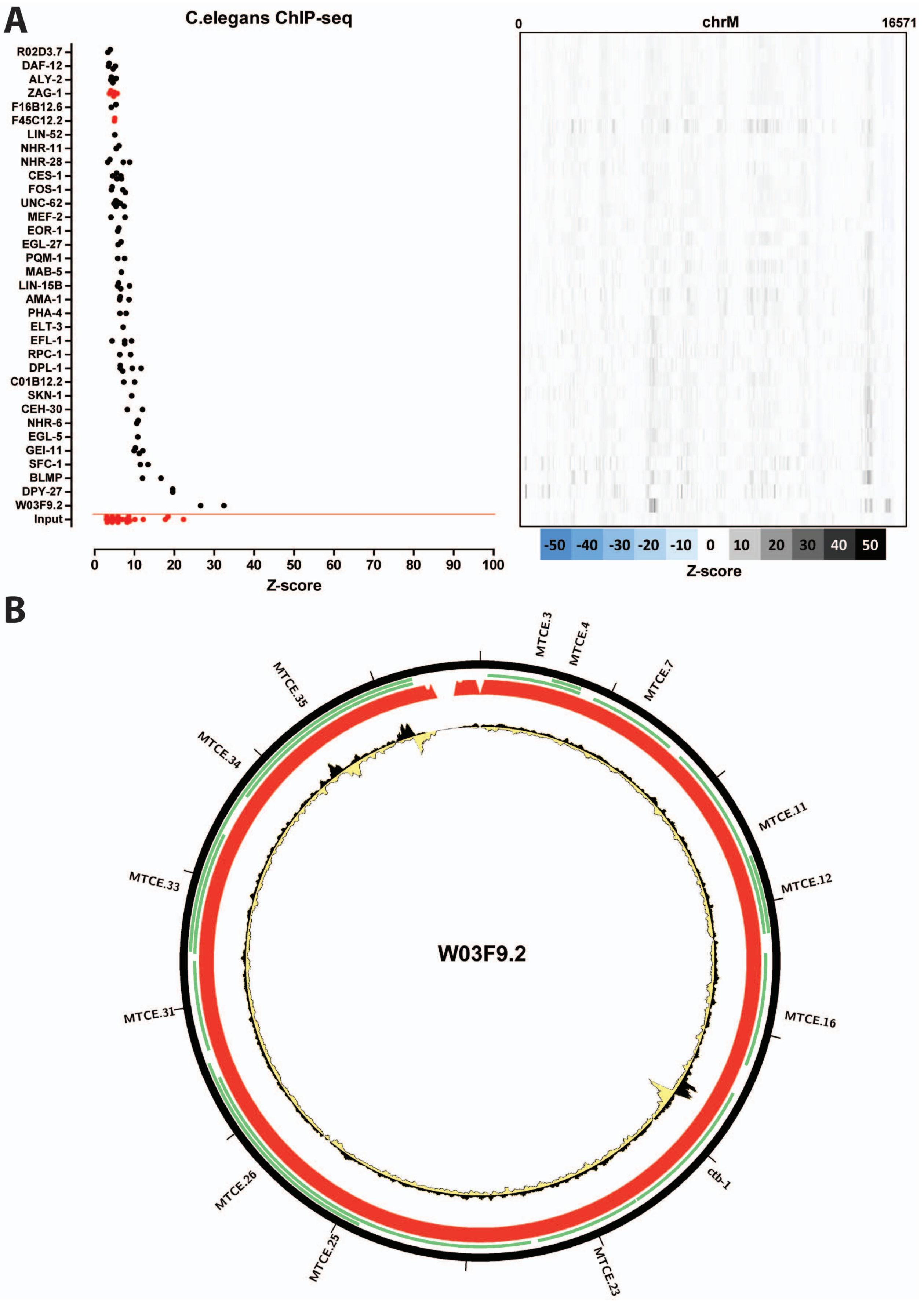

A
C.elegans ChIP-seq
R02D3.7
DAF-12
ALY-2
ZAG-1
F16B12.6
F45C12.2
LIN-52
NHR-11
NHR-28
CES-1
FOS-1
UNC-62
MEF-2
EOR-1
EGL-27
PQM-1
MAB-5
LIN-15B
AMA-1
PHA-4
ELT-3
EFL-1
RPC-1
DPL-1
C01B12.2
SKN-1
CEH-30
NHR-6
EGL-5
GEI-11
SFC-1
BLMP
DPY-27
W03F9.2
Input
0 10 20 30 40 50 60 70 80 90 100
Z-score
0
chrM
16571
-50 -40 -30 -20 -10 0 10 20 30 40 50
Z-score
B
MTCE.3
MTCE.4
MTCE.7
MTCE.11
MTCE.12
MTCE.16
ctb-1
MTCE.23
MTCE.25
MTCE.26
MTCE.31
MTCE.33
MTCE.34
MTCE.35
W03F9.2

Figure 13. Signal distribution over the mitochondrial genome in *C.elegans* ChIP-seq datasets. (A) Shown is the maximum z-score for each individual replicate for each cell line (left) and the z-score profile along the mitochondrial chromosome for the replicate with the highest z-score (right). Control datasets are shown in red on the bottom, below the red horizontal line; (B) Forward and reverse strand read distribution over the *C.elegans* mitochondrial genome for W03F9.2 (''Young Adult'' stage). Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Plots generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g013

of which is detected in high-throughput sequencing datasets that do not involve immunoprecipitation and therefore they are not due to sequencing biases. First, the majority of factors for which we detect strong read enrichment over the mitochondrial genome display high ChIP-seq signal only over the D-loop non-coding region in both human and mouse datasets. However, these signals do not have the characteristics of sequence specific occupancy and are present in a number of mock-immunoprecipitation control datasets. They are thus best explained as experimental artifacts, although it remains possible that they represent real non-canonical association with the D-loop for some factors. Second, for a subset of factors, specific ChIP-seq peaks are observed outside of the D-loop, and these display the additional hallmark characteristics of sequence specific occupancy.

Nuclear transcription factors previously reported to localize to mitochondria either did not exhibit significant enrichment in the available ChIP-seq datasets or, when they did, it was over the D-loop region with similar non-specific read distribution shape as other factors. In contrast, applying conservative thresholds we found eight human and three mouse transcription factors (two in common between the two species) that strongly occupy sites outside of the D-loop. They display the strand asymmetry pattern around the putative binding site that typifies true nuclear ChIP-seq peaks. Even more convincing is the fact that the explanatory motif for the factor is usually found under the observed enrichment peaks, further suggesting that they correspond to true in vivo biochemical events.

There are three main explanations for our observations. First, it is possible that despite our considerable bioinformatic precautions the observed binding events are in fact nuclear, originating from NUMTs present in the genomes of the cell lines assayed, but absent from the reference genome sequence. We believe that this is very unlikely. An experimental argument against unknown NUMTs comes from the strength of the ChIP-seq signal we see in the mitochondrial genome. These signals are much higher than even the strongest peaks in the nuclear genome for the same factor in the same dataset. This is expected for true mitochondrial genome binding because of the presence of many copies of the mitochondrial genome per cell, in contrast to the presence of only two copies of the nuclear genome. Second, it is possible that mitochondria are sometimes lysed in vivo, with mitochondrial DNA spilling into the cytoplasm where transcription factors could then bind. This cannot be ruled out based on the ChIP data alone but we consider it unlikely, as this would need to happen with a sufficient frequency to explain the remarkable strength of mitochondrial occupancy sites. The third and most plausible interpretation is that these nuclear transcription factors indeed translocate to the mitochondria and interact with the genome, as has been observed for the D-loop in some previous studies for other factors. Indeed, immunocytochemistry experiments in our study confirm the presence of MafK in mitochondria in a majority of HepG2 cells.

Several major questions are raised by our results. First, it is not clear how these nuclear transcription factors are targeted to the mitochondria. Mitochondrial proteins are typically imported into the mitochondrial matrix through the TIM/TOM protein translocator complex, and are targeted to the organelle by a mitochondrial localization sequence, which is cleaved upon import. We scanned both human and mouse versions of our factors for mitochondrial target sequences (MTS) with both Mitoprot [15] and TargetP [21] (using default settings), but we were unable to identify significant matches using either. This seems to be a common feature of nuclear transcription factors previously found to localize to mitochondria, most of which lack import sequences and are instead imported through other means [11,73]. Posttranslational modifications may be important for

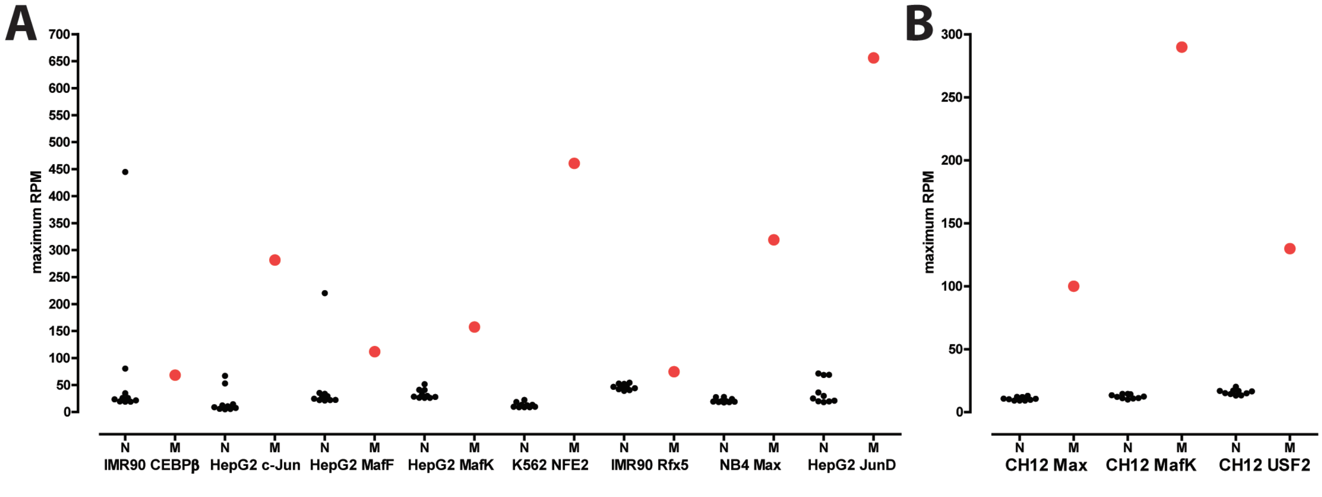


Figure 14. Mitochondrial ChIP-seq peaks are generally significantly stronger than nuclear peaks. Shown is the maximum signal (in RPM) for the top 10 nuclear peaks (''N'', smaller black dots), and the maximum signal intensity (also in RPM) in the mitochondrial genome (''M'', larger red dot) for representative ChIP-seq datasets for each factor. (A) Mouse datasets (B) Human datasets.
doi:10.1371/journal.pone.0084713.g014

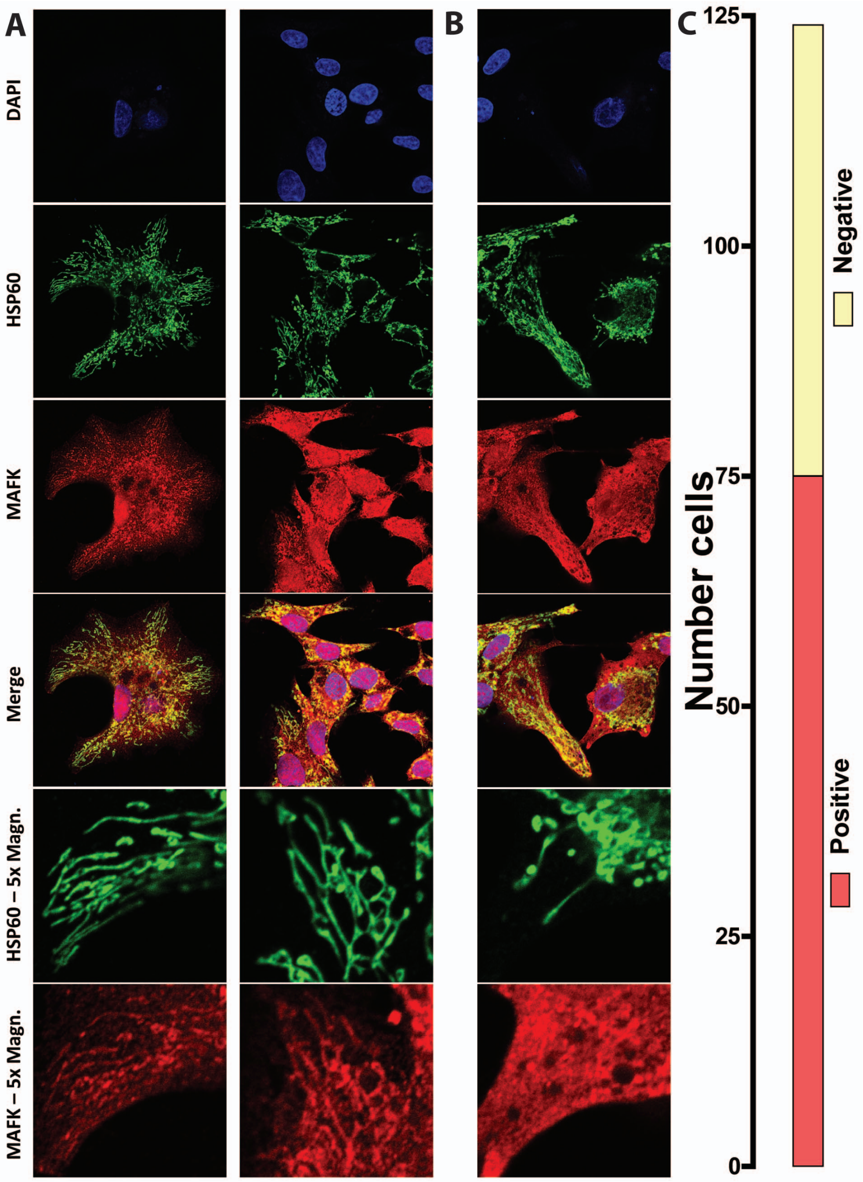


Figure 15. Localization of MafK to the mitochondria (A) Immunocytochemistry showing MafK localization in HepG2 cells. Mitochondria were identified by HSP60 staining. Shown are two representative images of cells showing that MAFK localizes strongly to the nucleus and mitochondria, and exhibits diffuse staining in the cytoplasm. In 60% of cells (C), there is colocalization of HSP60 with MAFK staining at an intensity higher than that of the surrounding cytoplasm. (B) An example of a cell exhibiting only nuclear and cytoplasmic MAFK localization.
doi:10.1371/journal.pone.0084713.g015

import, as has been demonstrated for STAT3 in TNF-induced necroptosis [68].

Second, it is unclear why the same factor binds detectably to the mitochondrial genome in some cell types but not in others. It is certainly possible that different splice isoforms or post-translationally modified proteins are present in different cell types, with only some capable of being imported into mitochondria, or that import into mitochondria only happens under certain physiological conditions only met in some cell lines.

Third, the question of the biochemical reality of transcription factor binding at the D-loop remains open. Previous studies understandably focused on the D-loop, given its well-appreciated importance in regulating mitochondrial transcription. As a consequence, the literature supporting a role for some nuclear factors in mitochondria suggests that they do so through binding to the D-loop. Our analysis of ChIP-seq data, which was carried out in an agnostic manner, revealed that dozens of transcription factors – many more than had been studied locally at the D-loop alone – also show high level of enrichment over the D-loop. However, the observed enrichment has characteristics suggesting that these signals are mainly due to experimental artifacts. In support of this judgment, the explanatory motifs for most of these factors were generally not found under the area of strongest enrichment in the D-loop. Therefore a conservative interpretation is that enrichment over the D-loop is an artifact in most cases.

Finally, and most importantly, the functional significance of factor occupancy observed by ChIP-seq remains unknown. It is entirely possible that it represents biochemical noise, with transcription factors entering the mitochondria because they have the right biochemical properties necessary to be imported, then

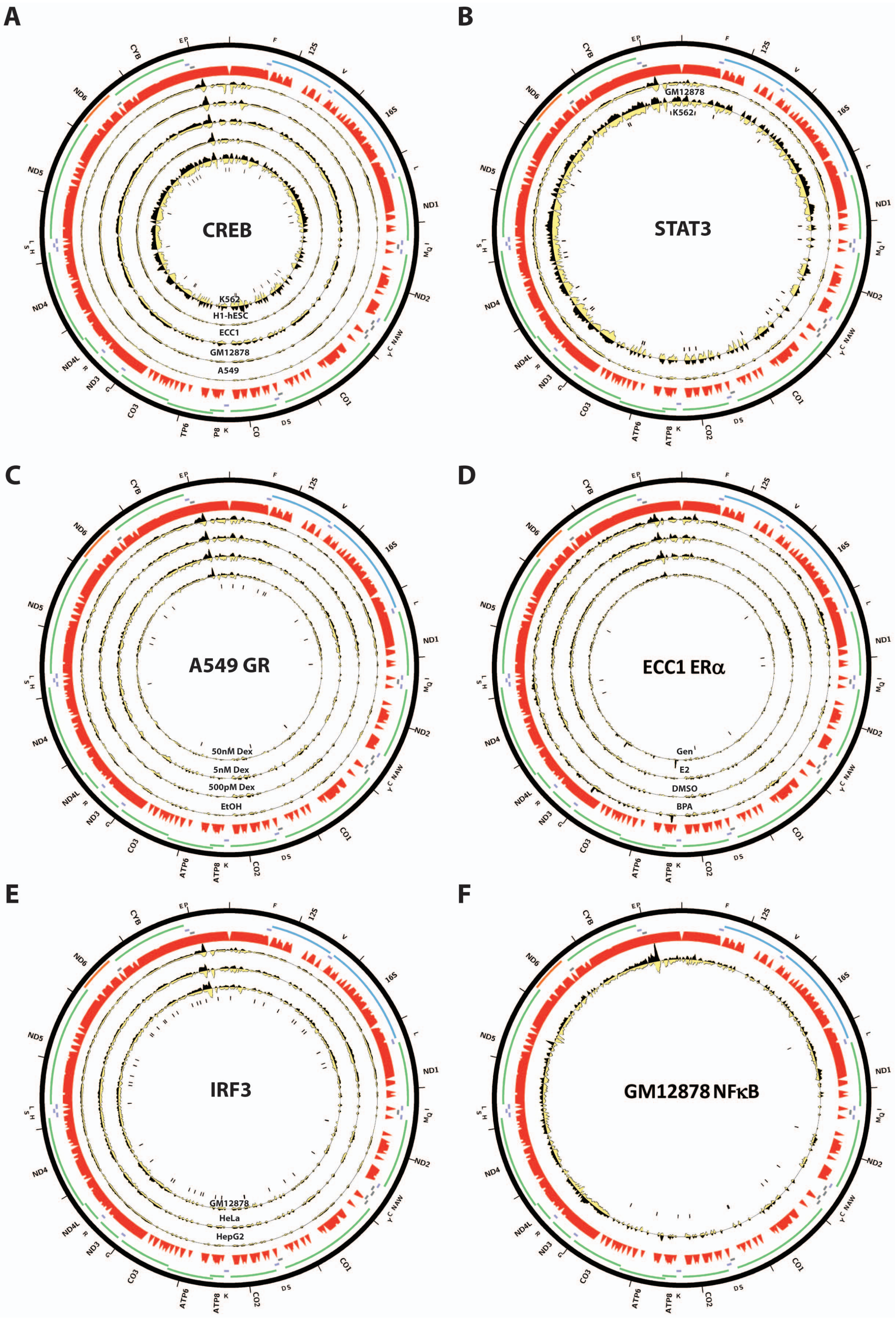

A
CREB
K562
H1-hESC
ECC1
GM12878
A549
CYB
EP
F
12S
V
16S
L
ND1
MQI
ND2
WAN CY
CO1
DS
CO
K
P8
TP6
CO3
G
ND3
R
ND4L
ND4
HS L
ND5
ND6
B
STAT3
GM12878
K562
CYB
EP
F
12S
V
16S
L
ND1
MQI
ND2
WAN CY
CO1
DS
CO2
K
ATP8
ATP6
CO3
G
ND3
R
ND4L
ND4
HS L
ND5
ND6
C
A549 GR
50nM Dex
5nM Dex
500pM Dex
EtOH
CYB
EP
F
12S
V
16S
L
ND1
MQI
ND2
WAN CY
CO1
DS
CO2
K
ATP8
ATP6
CO3
G
ND3
R
ND4L
ND4
HS L
ND5
ND6
D
ECC1 ERα
Gen
E2
DMSO
BPA
CYB
EP
F
12S
V
16S
L
ND1
MQI
ND2
WAN CY
CO1
DS
CO2
K
ATP8
ATP6
CO3
G
ND3
R
ND4L
ND4
HS L
ND5
ND6
E
IRF3
GM12878
HeLa
HepG2
CYB
EP
F
12S
V
16S
L
ND1
MQI
ND2
WAN CY
CO1
DS
CO2
K
ATP8
ATP6
CO3
G
ND3
R
ND4L
ND4
HS L
ND5
ND6
F
GM12878 NFκB
CYB
EP
F
12S
V
16S
L
ND1
MQI
ND2
WAN CY
CO1
DS
CO2
K
ATP8
ATP6
CO3
G
ND3
R
ND4L
ND4
HS L
ND5
ND6

Figure 16. Distribution of reads over the human mitochondrial genome for factors previously reported to bind to mitochondria in ENCODE ChIP-seq data. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) CREB; (B) STAT3; (C) GR in A549 cells treated with different concentrations of dexamethasone (Dex) [60,61]; (D) ERα in untreated (DMSO) ECC1 cells and ECC1 cells treated with bisphenol A (BPA), genistein (Gen) or 17β-estradiol (E2) [31]; (E) IRF3; (F) NFκB in GM12878 cells treated with TNFα [37]. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g016

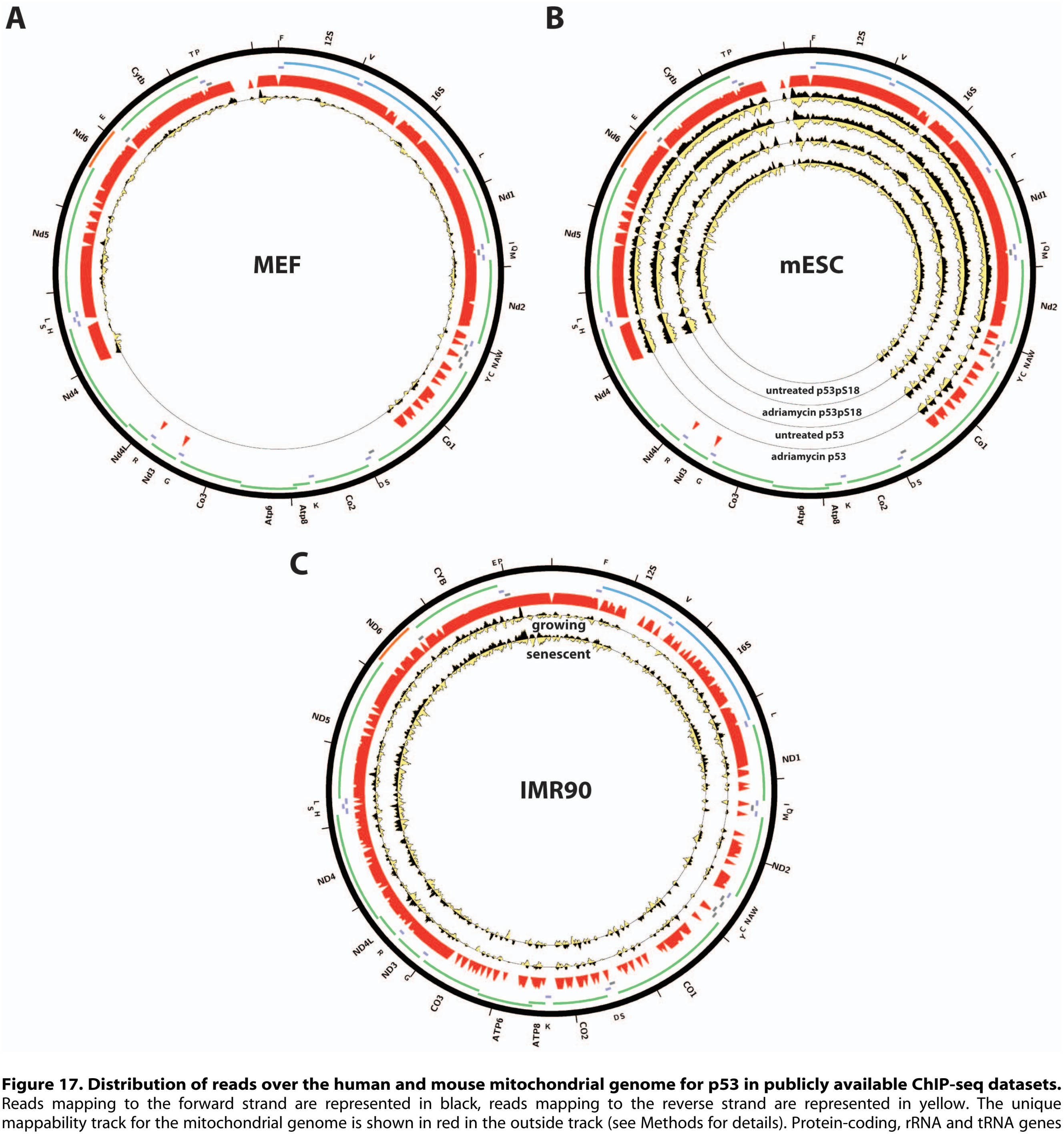


Figure 17. Distribution of reads over the human and mouse mitochondrial genome for p53 in publicly available ChIP-seq datasets. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) p53 in mouse embryionic fibroblasts (MEFs), data from [38], GSE46240. (B) p53 in mouse embryonic stem cells (mESC), data from [45], GSE26361; (C) p53 in human IMR90 cells, data from [2], GSE42728. The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g017

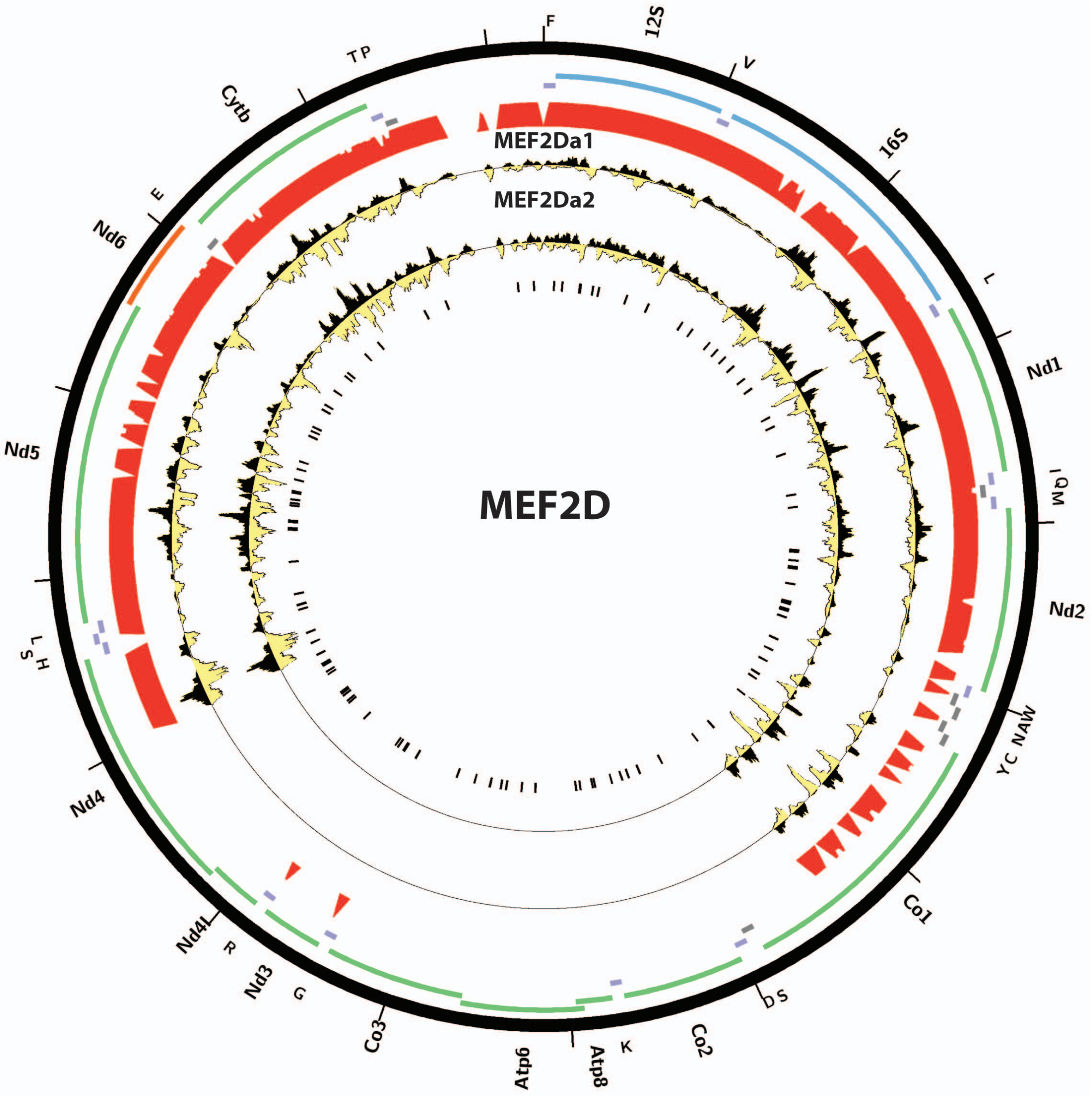


Figure 18. Distribution of reads over the mouse mitochondrial genome for MEF2D isoforms MEF2Da1 and MEF2Da2 in C2C12 myoblasts. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the MEF2D motif occurrences in the mitochondrial genome as black vertical bars. Data was obtained from [65], GSE43223. Plots were generated using Circos version 0.60 [41].
doi:10.1371/journal.pone.0084713.g018

binding to mtDNA but with little functional consequence. Alternatively, nuclear transcription factors may in fact be playing a regulatory role in mtDNA. It is difficult to imagine the exact mechanisms through which they might be acting, aside from interactions with the regulatory D-loop. While we do observe pairs of related factor such as c-Jun and JunD, and MafK and MafF binding to the same sites, binding events are overall widely dispersed over the mitochondrial genome and are found outside of the known regulatory regions. Plausible regulatory relationships are therefore not obvious and our results suggest that biological noise should be the working null hypothesis explaining the data. The functional regulatory role of these nuclear transcription factors in mitochondria is a very exciting possibility but it will have to be demonstrated in subsequent studies. Direct functional tests are the golden standard for establishing regulatory relationships, using gain and loss of function experiments and genetic manipulation of putative regulatory sites. The latter is at present not possible for mitochondria while the former are difficult to interpret in the case of the role of nuclear transcription factors in mitochondrial gene regulation, as it is not easy to separate the direct effects of binding to mtDNA from the indirect effects of transcriptional changes in the nucleus. Thus, it may be some time before definitive answers to these questions are obtained. In the meantime, larger compendia of transcription factor ChIP-seq data such as those expected to be generated by the next phase of the ENCODE project will be a primary source of further insight by providing binding data for additional nuclear transcription factors that will clarify allowed or preferred occupancy patterns across the mitochondrial genome.

Materials and Methods

Except for where indicated otherwise, all analysis was carried out using custom-written python scripts.

Sequencing read alignment

Raw sequencing reads were downloaded from the UCSC genome browser for ENCODE and mouseENCODE [54] data, and from ftp://ftp.modencode.org for modENCODE data [30,50] (data current as of February 2012). ChIP-seq data for p53 was obtained rom GEO series GSE26361 [45], GSE46240 [38] and GSE42728 [2]. Reads were aligned using Bowtie [42], version 0.12.7. Human data was mapped against either the female or the male set of human chromosomes (excluding the Y chromosome and/or all random chromosomes and haplotypes) depending on the sex of the cell line (where the sex was known, otherwise the Y chromosome was included), genome version hg19. Mouse data was mapped against the mm9 version of the mouse genome. modENCODE *D. melanogaster* data was mapped against the dm3 version of the fly genome. modENCODE data for *C. elegans* was mapped against the ce10 version of the worm genome. Reads were mapped with the following settings: "-v 2 -k 2 -m 1 -t –best –strata", which allow for two mismatches relative to the reference, however for all downstream analysis only reads mapping uniquely and with zero mismatches were considered, to eliminate any possible mapping artifacts.

Mappability track generation

Mappability was assessed as follows. Sequences of length N bases were generated starting at each position in the mitochondrial genome. The resulting set of "reads" was then mapped against the same bowtie index used for mapping real data. Positions covered by N reads were considered fully mappable. In this case, $N = 36$ as this is the read length for most of the sequencing data analyzed in this study.

Signal normalization of ChIP-seq data over the mitochondrial genome

Because the number of mitochondria per cell varies from one cell line/tissue to another, direct comparisons between datasets based on the absolute magnitude of the signal in RPM are not entirely valid. For this reason, we normalized the signal as follows. For each dataset, we fit a Gamma distribution over the RPM coverage scores for the bottom F_b percentile of fully mappable position on the mitochondrial chromosome. The estimated parameters were then used to rescale the raw signal over all position to a z-score. This results in datasets with strong peaks receiving low z-scores over most of the mappable mitochondrial genome, and very high z-scores over the regions with highly localized enrichment. We used $F = 0.8$ for our analysis. As this procedure is sensitive to datasets with very low total read coverage over the mitochondrial genome, we restricted our analysis to datasets with at least 5000 uniquely mappable reads (and with no mismatches to the reference), i.e. $\geq 10x$ coverage. We used a z-score cutoff of 20 to select datasets with high enrichment over the mitochondrial genome, as it was the highest z-score observed in sonicated input samples

Motif analysis

The peak calls for human and mouse ENCODE data available from the USCS Genome Browser were used to find de novo motifs for transcription factors from ChIP-seq data. The sequence around the peak summit (using a 50 bp radius) was retrieved for the top 500 called peaks for each factor in each cell line and motifs were called using the MEME program in the MEME SUITE, version 4.6.1 [4]. The MEME-defined position weight matrix was then used to scan the mitochondrial genome for motif matches following the approach described in [53].

Cell growth and immunocytochemistry

HepG2 cells were grown following the standard ENCODE protocol (DMEM media, 4 mM L-glutamine, 4.5 g/L glucose, without sodium pyruvate, with 10% FBS (Invitrogen 10091-148) and penicillin-streptomycin). Cells were fixed in 10% formalin (Sigma-Aldrich HT501128-4L) for 10 min, permeabilized with 0.1% Triton X-100, and blocked in 5% FBS. Primary antibodies used were MafK (1:100, Abcam, ab50322) and Hsp60 (1:125, Santa Cruz, sc-1052). Secondary antibodies used were donkey anti-goat AF488 (Invitrogen A11055) and donkey anti-rabbit AF546 (Invitrogen A10040). Imaging on a Zeiss LSM 710 confocal microscope with PlanApochromat 63X/1.4 oil objective, and 0.7 µm optical sections were acquired.

Supporting Information

Figure S1 Distribution of reads over the human mitochondrial genome for STAT1 and STAT5A in ENCODE ChIP-seq data. Reads mapping to the forward strand are represented in black, reads mapping to the reverse strand are represented in yellow. The unique mappability track for the mitochondrial genome is shown in red in the outside track (see Methods for details). Protein-coding, rRNA and tRNA genes are shown as colored bars. The innermost circle shows the motif occurrences in the mitochondrial genome for each factor as black vertical bars. (A) STAT1; (B) STAT5A; The reads per million (RPM) tracks are shown, scaled to the maximum signal level (for both strands) for each dataset. Plots were generated using Circos version 0.60 [41].
(PDF)

Acknowledgments

The authors thank Henry Amrhein, Diane Trout, Sean Upchurch and Sreeram Balasubramanian for technical and computational assistance.

Author Contributions

Conceived and designed the experiments: GKM YEW BJW. Performed the experiments: GKM YEW. Analyzed the data: GKM YEW. Wrote the paper: GKM YEW DC BJW.

References

1. Achanta G, Sasaki R, Feng L, Carew JS, Lu W, et al. (2005) Novel role of p53 in maintaining mitochondrial genetic stability through interaction with DNA Polγ. EMBO Journal 24(19):3482–92.
2. Aksoy O, Chicas A, Zeng T, Zhao Z, McCurrach M, et al. (2012) The atypical E2F family member E2F7 couples the p53 and RB pathways during cellular senescence. Genes Dev 26(14):1546–57.
3. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290(5806):457–465.
4. Bailey TL, Bodén M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208.

5. Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA (1981) Sequence and gene organization of mouse mitochondrial DNA. Cell 26(2 Pt 2):167–180.
6. Boengler K, Hilfiker-Kleiner D, Heusch G, Schulz R (2010) Inhibition of permeability transition pore opening by mitochondrial STAT3 and its role in myocardial ischemia/reperfusion. Basic Res Cardiol 105(6):771–785.
7. Bogenhagen D, Clayton DA (1974) The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. J Biol Chem 249(24):7991–5.
8. Cammarota M, Paratcha G, Bevilaqua LR, Levi de Stein M, et al. (1999) Cyclic AMP-responsive element binding protein in brain mitochondria. J Neurochem 72(6):2272–7.
9. Cantatore P, Attardi G (1980) Mapping of nascent light and heavy strand transcripts on the physical map of HeLa cell mitochondrial DNA. Nucleic Acids Res 8(12):2605–2625.
10. Casas F, Domenjoud L, Rochard P, Hatier R, Rodier A, et al. (2000) A 45 kDa protein related to PPARγ2, induced by peroxisome proliferators, is located in the mitochondrial matrix. FEBS Lett 478(1–2):4–8.
11. Casas F, Rochard P, Rodier A, Cassar-Malek I, Marchal-Victorion S, et al. (1999) A variant form of the nuclear triiodothyronine receptor c-ErbAα1 plays a direct role in regulation of mitochondrial RNA synthesis. Mol Cell Biol 19(12):7913–24.
12. Chattopadhyay S, Marques JT, Yamashita M, Peters KL, Smith K, et al. (2010) Viral apoptosis is induced by IRF-3-mediated activation of Bax. EMBO J 29(10):1762–1773.
13. Chen JQ, Delannoy M, Cooke C, Yager JD (2004) Mitochondrial localization of ERα and ERβ in human MCF7 cells. Am J Physiol Endocrinol Metab 286(6):E1011–22.
14. Chueh FY, Leong KF, Yu CL (2010) Mitochondrial translocation of signal transducer and activator of transcription 5 (STAT5) in leukemic T cells and cytokine-stimulated cells. Biochem Biophys Res Commun 402(4):778–783.
15. Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 241(3):779–786.
16. Cogswell PC, Kashatus DF, Keifer JA, Guttridge DC, Reuther JY, et al. (2003) NFκB and IκBα are found in the mitochondria. Evidence for regulation of mitochondrial gene expression by NFκB. J Biol Chem 278(5):2963–2968.
17. De Rasmo D, Signorile A, Roca E, Papa S (2009) cAMP response element-binding protein (CREB) is imported into mitochondria and promotes protein synthesis. FEBS J 276(16):4325–33.
18. Demonacos C, Tsawdaroglou NC, Djordjevic-Markovic R, Papalopoulou M, Galanopoulos V, et al. (1993) Import of the glucocorticoid receptor into rat liver mitochondria in vivo and in vitro. J Steroid Biochem Mol Biol 46(3):401–13.
19. Demonacos C, Djordjevic-Markovic R, Tsawdaroglou N, Sekeris CE (1995) The mitochondrion as a primary site of action of glucocorticoids: the interaction of the glucocorticoid receptor with mitochondrial DNA sequences showing partial similarity to the nuclear glucocorticoid responsive elements. J Steroid Biochem Mol Biol 55(1):43–55.
20. du Buy H, Riley F (1967) Hybridization between the nuclear and kinetoplast DNA's of Leishmania enriettii and between nuclear and mitochondrial DNA's of mouse liver. Proc Natl Acad Sci U S A 57(3):790–797.
21. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971.
22. ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9(4):e1001046. doi: 10.1371/journal.pbio.1001046
23. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74.
24. Enríquez JA, Fernández-Sílva P, Montoya J (1999) Autonomous regulation in mammalian mitochondrial DNA transcription. Biol Chem 380(7–8):737–747.
25. Enríquez JA, Fernández-Silva P, Garrido-Pérez N, López-Pérez MJ, Pérez-Martos A, et al. (1999) Direct regulation of mitochondrial RNA synthesis by thyroid hormone. Mol Cell Biol 19(1):657–70.
26. Falkenberg M, Gaspari M, Rantanen A, Trifunovic A, Larsson NG, et al. (2002) Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. Nat Genet 31(3):289–294.
27. Fisher RP, Clayton DA (1985) A transcription factor required for promoter recognition by human mitochondrial RNA polymerase. Accurate initiation at the heavy- and light-strand promoters dissected and reconstituted in vitro. J Biol Chem 260(20):11330–11338.
28. Fisher RP, Clayton DA (1988) Purification and characterization of human mitochondrial transcription factor 1. Mol Cell Biol 8(8):3496–3509.
29. Gaspari M, Falkenberg M, Larsson NG, Gustafsson CM (2004) The mitochondrial RNA polymerase contributes critically to promoter specificity in mammalian cells. EMBO J 23(23):4606–4614.
30. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330(6012):1775–1787.
31. Gertz J, Reddy TE, Varley KE, Garabedian MJ, Myers RM (2012) Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. Genome Res 22(11):2153–62.
32. Gough DJ, Corlett A, Schlessinger K, Wegrzyn J, Larner AC, et al. (2009) Mitochondrial STAT3 supports Ras-dependent oncogenic transformation. Science 324(5935):1713–1716.
33. Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6(2):e1000834. doi: 10.1371/journal.pgen.1000834.
34. Heyne K, Mannebach S, Wuertz E, Knaup KX, Mahyar-Roemer M, et al. (2004) Identification of a putative p53 binding sequence within the human mitochondrial genome. FEBS Lett 578(1–2):198–202.
35. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830):1497–502.
36. Johnson RF, Witzel II, Perkins ND (2011) p53-dependent regulation of mitochondrial energy production by the RelA subunit of NF-κB. Cancer Res 71(16):5588–5597.
37. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, et al. (2010) Variation in transcription factor binding among humans. Science 328(5975):232–235.
38. Kenzelmann Broz D, Spano Mello S, Bieging KT, Jiang D, Dusek RL, et al. (2013) Global genomic profiling reveals an extensive p53-regulated autophagy program contributing to key p53 responses. Genes Dev 27(9):1016–31.
39. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351–1359.
40. Koufali MM, Moutsatsou P, Sekeris CE, Breen KC (2003) The dynamic localization of the glucocorticoid receptor in rat C6 glioma cell mitochondria. Mol Cell Endocrinol 209(1–2):51–60.
41. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–45.
42. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25.
43. Lee J, Kim CH, Simon DK, Aminova LR, Andreyev AY, et al. (2005) Mitochondrial cyclic AMP response element-binding protein (CREB) mediates mitochondrial gene expression and neuronal survival. J Biol Chem 280(49):40398–401.
44. Leigh-Brown S, Enriquez JA, Odom DT (2010) Nuclear transcription factors in mammalian mitochondria. Genome Biol 11(7):215.
45. Li M, He Y, Dubois W, Wu X, Shi J, et al. (2012) Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol Cell 46(1):30–42.
46. Liu XY, Wei B, Shi HX, Shan YF, Wang C (2010) Tom70 mediates activation of interferon regulatory factor 3 on mitochondria. Cell Res 20(9):994–1011.
47. Marchenko ND, Wolff S, Erster S, Becker K, Moll UM (2007) Monoubiquitylation promotes mitochondrial p53 translocation. EMBO J 26(4):923–934.
48. Marchenko ND, Zaika A, Moll UM (2000) Death signal-induced localization of p53 protein to mitochondria. A potential role in apoptotic signaling. J Biol Chem 275(21):16202–12.
49. Metodiev MD, Lesko N, Park CB, Cámara Y, Shi Y, et al. (2009) Methylation of 12S rRNA is necessary for in vivo stability of the small subunit of the mammalian mitochondrial ribosome. Cell Metab 9(4):386–397.
50. modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, et al. (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330(6012):1787–1797.
51. Monje P, Boland R (2001) Subcellular distribution of native estrogen receptor α and β isoforms in rabbit uterus and ovary. J Cell Biochem 82(3):467–79.
52. Montoya J, Christianson T, Levens D, Rabinowitz M, Attardi G (1982) Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. Proc Natl Acad Sci 79(23):7195–7199.
53. Mortazavi A, Leeper Thompson EC, Garcia ST, Myers RM, Wold B (2006) Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. Genome Res 16(10):1208–1221.
54. Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, et al. (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13(8):418.
55. Nass S, Nass MM, Hennix U (1965) Deoxyribonucleic acid in isolated rat-liver mitochondria. Biochim Biophys Acta 95:426435.
56. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489(7414):83–90.
57. Ogita K, Fujinami Y, Kitano M, Yoneda Y (2003) Transcription factor activator protein-1 expressed by kainate treatment can bind to the non-coding region of mitochondrial genome in murine hippocampus. J Neurosci Res 73(6):794–802.
58. Ogita K, Okuda H, Kitano M, Fujinami Y, Ozaki K, et al. (2002) Localization of activator protein-1 complex with DNA binding activity in mitochondria of murine brain after in vivo treatment with kainate. J Neurosci 22(7):2561–70.
59. Psarra AM, Solakidi S, Sekeris CE (2006) The mitochondrion as a primary site of action of steroid and thyroid hormones: presence and action of steroid and thyroid hormone receptors in mitochondria of animal cells. Mol Cell Endocrinol 246(1–2):21–33.
60. Reddy TE, Gertz J, Crawford GE, Garabedian MJ, Myers RM (2012) The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. Mol Cell Biol 32(18):3756–67.
61. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, et al. (2009) Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res 19(12):2163–71.

62. Ryu H, Lee J, Impey S, Ratan RR, Ferrante RJ (2005) Antioxidants modulate mitochondrial PKA and increase CREB binding to D-loop DNA of the mitochondrial genome in neurons. Proc Natl Acad Sci U S A 102(39):13915–20.
63. Sagan L (1967) On the origin of mitosing cells. J Theor Biol 14(3):255–274.
64. Satoh M, Kuroiwa T (1991) Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res 196:137140
65. Sebastian S, Faralli H, Yao Z, Rakopoulos P, Palii C, et al. (2013) Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. Genes Dev 27(11):1247–1259.
66. Shadel GS, Clayton DA (1997) Mitochondrial DNA maintenance in vertebrates. Annu Rev Biochem 66:409–435.
67. She H, Yang Q, Shepherd K, Smith Y, Miller G, et al. (2011) Direct regulation of complex I by mitochondrial MEF2D is disrupted in a mouse model of Parkinson disease and in human patients. J Clin Invest 121(3):930–940.
68. Shulga N, Pastorino JG (2012) GRIM-19-mediated translocation of STAT3 to mitochondria is necessary for TNF-induced necroptosis. J Cell Sci 125(Pt 12):2995–3003.
69. Shutt TE, Bestwick M, Shadel GS (2011) The core human mitochondrial transcription initiation complex: It only takes two to tango. Transcription 2(2):55–59.
70. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21(10):1757–1767.
71. Szczepanek K, Chen Q, Derecka M, Salloum FN, Zhang Q, et al. (2011) Mitochondrial-targeted Signal transducer and activator of transcription 3 (STAT3) protects against ischemia-induced changes in the electron transport chain and the generation of reactive oxygen species. J Biol Chem 286(34):29610–29620.
72. Szczepanek K, Chen Q, Larner AC, Lesnefsky EJ (2012) Cytoprotection by the modulation of mitochondrial electron transport chain: the emerging role of mitochondrial STAT3. Mitochondrion 12(2):180–189.
73. Szczepanek K, Lesnefsky EJ, Larner AC (2012). Multi-tasking: nuclear transcription factors with novel roles in the mitochondria. Trends Cell Biol 22(8):429–437.
74. Szklarczyk R, Huynen MA (2010) Mosaic origin of the mitochondrial proteome. Proteomics 10(22):4012–4024.
75. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. (2012) The accessible chromatin landscape of the human genome. Nature 489(7414):75–82.
76. Vaseva AV, Moll UM (2009) The mitochondrial p53 pathway. Biochim Biophys Acta 1787(5):414–420.
77. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 22(9):1798–1812.
78. Wang YE, Marinov GK, Wold BJ, Chan DC (2013) Genome-wide analysis reveals coating of the mitochondrial genome by TFAM. PLoS ONE 8(8):e74513. doi: 10.1371/journal.pone.0074513
79. Wegrzyn J, Potla R, Chwae YJ, Sepuri NB, Zhang Q, et al. (2009) Function of mitochondrial Stat3 in cellular respiration. Science 323(5915):793–797.
80. Williams RS (1986) Mitochondrial gene expression in mammalian striated muscle. Evidence that variation in gene dosage is the major regulatory event. J Biol Chem 261(26):12390–4.
81. Wrutniak C, Cassar-Malek I, Marchal S, Rascle A, Heusser S, et al. (1995) A 43-kDa protein related to c-Erb A α1 is located in the mitochondrial matrix of rat liver. J Biol Chem 270(27):16347–54.
82. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR (1985) Mitochondrial origins. Proc Natl Acad Sci U S A 82(13):4443–4447.
83. Yoshida Y, Izumi H, Torigoe T, Ishiguchi H, Itoh H, et al. (2003) P53 physically interacts with mitochondrial transcription factor A and differentially regulates binding to damaged DNA. Cancer Res 63(13):3729–34.

M

Defining functional DNA elements in the human genome

Originally published as:

Kellis M*, Hardison RC*, Wold BJ*, Snyder MP*, Bernstein BE*, Kundaje A‡, Marinov GK‡, Ward LD‡, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski L, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard TJP, Kent WJ, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos J, Weng Z, White KP, Members of the ENCODE Consortium. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**(17):6131-6138. doi: 10.1073/pnas.1318948111.

Defining functional DNA elements in the human genome

Manolis Kellis[a,b,1,2], Barbara Wold[c,2], Michael P. Snyder[d,2], Bradley E. Bernstein[b,e,f,2], Anshul Kundaje[a,b,3], Georgi K. Marinov[c,3], Lucas D. Ward[a,b,3], Ewan Birney[g], Gregory E. Crawford[h], Job Dekker[i], Ian Dunham[g], Laura L. Elnitski[j], Peggy J. Farnham[k], Elise A. Feingold[j], Mark Gerstein[l], Morgan C. Giddings[m], David M. Gilbert[n], Thomas R. Gingeras[o], Eric D. Green[j], Roderic Guigo[p], Tim Hubbard[q], Jim Kent[r], Jason D. Lieb[s], Richard M. Myers[t], Michael J. Pazin[j], Bing Ren[u], John A. Stamatoyannopoulos[v], Zhiping Weng[i], Kevin P. White[w], and Ross C. Hardison[x,1,2]

[a]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Broad Institute, Cambridge, MA 02139; [c]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125; [d]Department of Genetics, Stanford University, Stanford, CA 94305; [e]Harvard Medical School and [f]Massachusetts General Hospital, Boston, MA 02114; [g]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; [h]Medical Genetics, Duke University, Durham, NC 27708; [i]Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01605; [j]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; [k]Biochemistry and Molecular Biology, University of Southern California, Los Angeles, CA 90089; [l]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520; [m]Marketing Your Science, LLC, Boise, ID 83702; [n]Department of Biological Science, Florida State University, Tallahassee, FL 32306; [o]Functional Genomics Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; [p]Bioinformatics and Genomics Program, Center for Genome Regulation, E-08003 Barcelona, Catalonia, Spain; [q]Medical and Molecular Genetics, King's College London and Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; [r]Biomolecular Engineering, University of California, Santa Cruz, CA 95064; [s]Lewis Sigler Institute, Princeton University, Princeton, NJ 08544; [t]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806; [u]Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA 92093; [v]Genome Sciences and Medicine, University of Washington, Seattle, WA 98195; [w]Human Genetics, University of Chicago, Chicago, IL 60637; and [x]Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved January 29, 2014 (received for review October 16, 2013)

With the completion of the human genome sequence, attention turned to identifying and annotating its functional DNA elements. As a complement to genetic and comparative genomics approaches, the Encyclopedia of DNA Elements Project was launched to contribute maps of RNA transcripts, transcriptional regulator binding sites, and chromatin states in many cell types. The resulting genome-wide data reveal sites of biochemical activity with high positional resolution and cell type specificity that facilitate studies of gene regulation and interpretation of noncoding variants associated with human disease. However, the biochemically active regions cover a much larger fraction of the genome than do evolutionarily conserved regions, raising the question of whether nonconserved but biochemically active regions are truly functional. Here, we review the strengths and limitations of biochemical, evolutionary, and genetic approaches for defining functional DNA segments, potential sources for the observed differences in estimated genomic coverage, and the biological implications of these discrepancies. We also analyze the relationship between signal intensity, genomic coverage, and evolutionary conservation. Our results reinforce the principle that each approach provides complementary information and that we need to use combinations of all three to elucidate genome function in human biology and disease.

Quest to Identify Functional Elements in the Human Genome

Completing the human genome reference sequence was a milestone in modern biology. The considerable challenge that remained was to identify and delineate the structures of all genes and other functional elements. It was quickly recognized that nearly 99% of the ~3.3 billion nucleotides that constitute the human genome do not code for proteins (1). Comparative genomics studies revealed that the majority of mammalian-conserved and recently adapted regions consist of noncoding elements (2–10). More recently, genome-wide association studies have indicated that a majority of trait-associated loci, including ones that contribute to human diseases and susceptibility, also lie outside protein-coding regions (11–16). These findings suggest that the noncoding regions of the human genome harbor a rich array of functionally significant elements with diverse gene regulatory and other functions.

Despite the pressing need to identify and characterize all functional elements in the human genome, it is important to recognize that there is no universal definition of what constitutes function, nor is there agreement on what sets the boundaries of an element. Both scientists and nonscientists have an intuitive definition of function, but each scientific discipline relies primarily on different lines of evidence indicative of function. Geneticists, evolutionary biologists, and molecular biologists apply distinct approaches, evaluating different and complementary lines of evidence. The genetic approach evaluates the phenotypic consequences of perturbations, the evolutionary

Author contributions: M.K., B.W., M.P.S., B.E.B., and R.C.H. designed research; M.K., B.W., M.P.S., B.E.B., A.K., G.K.M., L.D.W., and R.C.H. performed research; A.K., G.K.M., and L.D.W. contributed computational analysis and tools; M.K., B.W., M.P.S., B.E.B., E.B., G.E.C., J.D., I.D., L.L.E., P.J.F., E.A.F., M.G., M.C.G., D.M.G., T.R.G., E.D.G., R.G., T.H., J.K., J.D.L., R.M.M., M.J.P., B.R., J.A.S., Z.W., K.P.W., and R.C.H. contributed to manuscript discussions and ideas; and M.K., B.W., M.P.S., B.E.B., and R.C.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: In addition to data already released via the ENCODE Data Coordinating Center, the erythroblast DNase-seq data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE55579, GSM1339559, and GSM1339560).

Authored by members of the ENCODE Consortium.

[1] To whom correspondence may be addressed. E-mail: manoli@mit.edu or rch8@psu.edu.

[2] M.K., B.W., M.P.S., B.E.B., and R.C.H. contributed equally to this work.

[3] A.K., G.K.M., and L.D.W. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318948111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1318948111

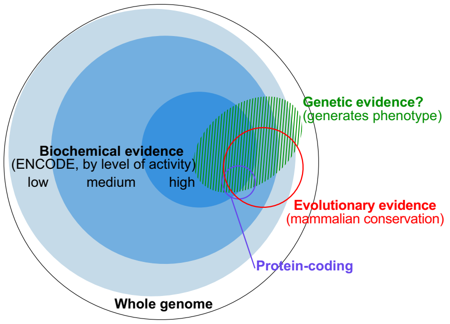


Fig. 1. The complementary nature of evolutionary, biochemical, and genetic evidence. The outer circle represents the human genome. Blue discs represent DNA sequences acted upon biochemically and partitioned by their levels of signal [combined 10th percentiles of different ENCODE data types for high, combined 50th percentiles for medium, and all significant signals for low (see *Reconciling Genetic, Evolutionary, and Biochemical Estimates* and Fig. 2)]. The red circle represents, at the same scale, DNA with signatures of evolutionary constraint (GERP++ elements derived from 34mammal alignments). Overlaps among the sequences having biochemical and evolutionarily evidence were computed in this work (Fig. 3 and *SI Methods*). The small purple circle represents protein-coding nucleotides (Gencode). The green shaded domain conceptually represents DNA that produces a phenotype upon alteration, although we lack well-developed summary estimates for the amount of genetic evidence and its relationship with the other types. This summary of our understanding in early 2014 will likely evolve substantially with more data and more refined experimental and analytical methods.

approach quantifies selective constraint, and the biochemical approach measures evidence of molecular activity. All three approaches can be highly informative of the biological relevance of a genomic segment and groups of elements identified by each approach are often quantitatively enriched for each other. However, the methods vary considerably with respect to the specific elements they predict and the extent of the human genome annotated by each (Fig. 1).

Some of these differences stem from the fact that function in biochemical and genetic contexts is highly particular to cell type and condition, whereas for evolutionary measures, function is ascertained independently of cellular state but is dependent on environment and evolutionary niche. The methods also differ widely in their false-positive and false-negative rates, the resolution with which elements are defined, and the throughput with which they can be surveyed. Moreover, each approach remains incomplete, requiring continued method development (both experimental and analytical) and increasingly large datasets (additional species, assays, cell types, variants, and phenotypes). It is thus not surprising that the methods vary considerably with respect to the specific elements they identify. However, the extent of the difference is much larger than simply technical limitations would suggest, challenging current views and definitions of genome function.

Many examples of elements that appear to have conflicting lines of functional evidence were described before the Encyclopedia of DNA Elements (ENCODE) Project, including elements with conserved phenotypes but lacking sequence-level conservation (17–20), conserved elements with no phenotype on deletion (21, 22), and elements able to drive tissue-specific expression but lacking evolutionary conservation (23, 24). However, the scale of the ENCODE Project survey of biochemical activity (across many more cell types and assays) led to a significant increase in genome coverage and thus accentuated the discrepancy between biochemical and evolutionary estimates. This discrepancy led to much debate both in the scientific literature (25–31) and in online forums, resulting in a renewed need to clarify the challenges of defining function in the human genome and to understand the sources of the discrepancy.

To address this need and provide a perspective by ENCODE scientists, we review genetic, evolutionary, and biochemical lines of evidence, discuss their strengths and limitations, and examine apparent discrepancies between the conclusions emanating from the different approaches.

Genetic Approach. Genetic approaches, which rely on sequence alterations to establish the biological relevance of a DNA segment, are often considered a gold standard for defining function. Mutations can be naturally occurring and identified by screening for phenotypes generated by sequence variants (13, 32) or produced experimentally by targeted genetic methods (33) or nongenetic interference (34). Transfection studies that use reporter assays in cell lines (35, 36) or embryos (37) can also be used to identify regulatory elements and measure their activities. Genetic approaches tend to be limited by modest throughput, although speed and efficiency is now increasing for some methods (36, 38–40). The approach may also miss elements whose phenotypes occur only in rare cells or specific environmental contexts, or whose effects are too subtle to detect with current assays. Loss-of-function tests can also be buffered by functional redundancy, such that double or triple disruptions are required for a phenotypic consequence. Consistent with redundant, contextual, or subtle functions, the deletion of large and highly conserved genomic segments sometimes has no discernible organismal phenotype (21, 22), and seemingly debilitating mutations in genes thought to be indispensible have been found in the human population (41).

Evolutionary Approach. Comparative genomics provides a powerful approach for detecting noncoding functional elements that show preferential conservation across evolutionary time. A high level of sequence conservation between related species is indicative of purifying selection, whereby disruptive mutations are rejected, with the corresponding sequence deemed to be likely functional. Evidence of function can also come from accelerated evolution across species or within a particular lineage, revealing elements under positive selection for recently acquired changes that increase fitness; such an approach gains power by incorporating multiple closely related genomes because each species provides information about sequence constraint. Multispecies comparisons have been used in studies of diverse clades, ranging from yeast to mammals. Methods that detect sequences likely under selection have had success in recognizing protein-coding regions, structural RNAs, gene regulatory regions, regulatory motifs, and specific regulatory elements (3, 42–48). The comparative genomics approach can also incorporate information about mutational patterns that may be characteristic of different types of elements.

Although powerful, the evolutionary approach also has limitations. Identification of conserved regions depends on accurate multispecies sequence alignments, which remain a substantial challenge. Alignments are generally less effective for distal-acting regulatory regions, where they may be impeded by regulatory motif turnover, varying spacing constraints, and sequence composition biases (17, 49). Analyzing aligned regions for conservation can be similarly challenging. First, most transcription factor-binding sequences are short and highly degenerate, making them difficult to identify. Second, because detection of neutrally evolving elements requires sufficient phylogenetic distance, the approach is well suited for detecting mammalian-conserved elements, but it is less effective for primate-specific elements and essentially blind to human-specific elements. Third, certain types of functional elements such as immunity genes may be prone to rapid evolutionary turnover even among closely related species. More generally, alignment methods are not well suited to capture substitutions that preserve function, such as compensatory changes preserving RNA structure, affinity-preserving substitutions within regulatory motifs, or mutations whose effect is buffered by redundancy or epistatic effects. Thus, absence of conservation cannot be interpreted as evidence for the lack of function.

Finally, although the evolutionary approach has the advantage that it does not require a priori knowledge of what a DNA element does or when it is used, it is unlikely to reveal the molecular mechanisms under selection or the relevant cell types or physiological processes. Thus, comparative genomics requires complementary studies.

Biochemical Approach. The biochemical approach for identifying candidate functional genomic elements complements the other approaches, as it is specific for cell type, condition, and molecular process. Decades of detailed studies of gene regulation and RNA metabolism have defined major classes of functional noncoding elements, including promoters, enhancers, silencers, insulators, and noncoding RNA genes such as microRNAs, piRNAs, structural RNAs, and regulatory RNAs (50–53). These noncoding functional elements are associated with distinctive chromatin structures that display signature patterns of histone modifications, DNA methylation, DNase accessibility, and transcription factor occupancy (37, 54–66). For example, active enhancers are marked by specific histone modifications and DNase-accessible chromatin and are occupied by sequence-specific transcription factors, coactivators such as EP300, and, often, RNA polymerase II. Although the extent to which individual features contribute to function remains to be determined, they provide a useful surrogate for annotating candidate enhancers and other types of functional elements.

The ENCODE Project was established with the goal of systematically mapping functional elements in the human genome at high resolution and providing this information as an open resource for the research community (67, 68). Most data acquisition in the project thus far has taken the biochemical approach, using evidence of cellular or enzymatic processes acting on a DNA segment to help predict different classes of functional elements. The recently completed phase of ENCODE applied a wide range of biochemical assays at a genome-wide scale to study multiple human cell types (69). These assays identified genomic sequences (*i*) from which short and long RNAs, both nuclear and cytoplasmic, are transcribed; (*ii*) occupied by sequence-specific transcription factors, cofactors, or chromatin regulatory proteins; (*iii*) organized in accessible chromatin; (*iv*) marked by DNA methylation or specific histone modifications; and (*v*) physically brought together by long-range chromosomal interactions.

An advantage of such functional genomics evidence is that it reveals the biochemical processes involved at each site in a given cell type and activity state. However, biochemical signatures are often a consequence of function, rather than causal. They are also not always deterministic evidence of function, but can occur stochastically. For example, GATA1, whose binding at some erythroid-specific enhancers is critical for function, occupies many other genomic sites that lack detectable enhancer activity or other evidence of biological function (70). Likewise, although enhancers are strongly associated with characteristic histone modifications, the functional significance of such modifications remains unclear, and the mere presence of an enhancer-like signature does not necessarily indicate that a sequence serves a specific function (71, 72). In short, although biochemical signatures are valuable for identifying candidate regulatory elements in the biological context of the cell type examined, they cannot be interpreted as definitive proof of function on their own.

What Fraction of the Human Genome Is Functional?

Limitations of the genetic, evolutionary, and biochemical approaches conspire to make this seemingly simple question difficult to answer. In general, each approach can be used to lend support to candidate elements identified by other methods, although focusing exclusively on the simple intersection set would be much too restrictive to capture all functional elements. However, by probing quantitative relationships in data from the different approaches, we can begin to gain a more sophisticated picture of the nature, identity, and extent of functional elements in the human genome.

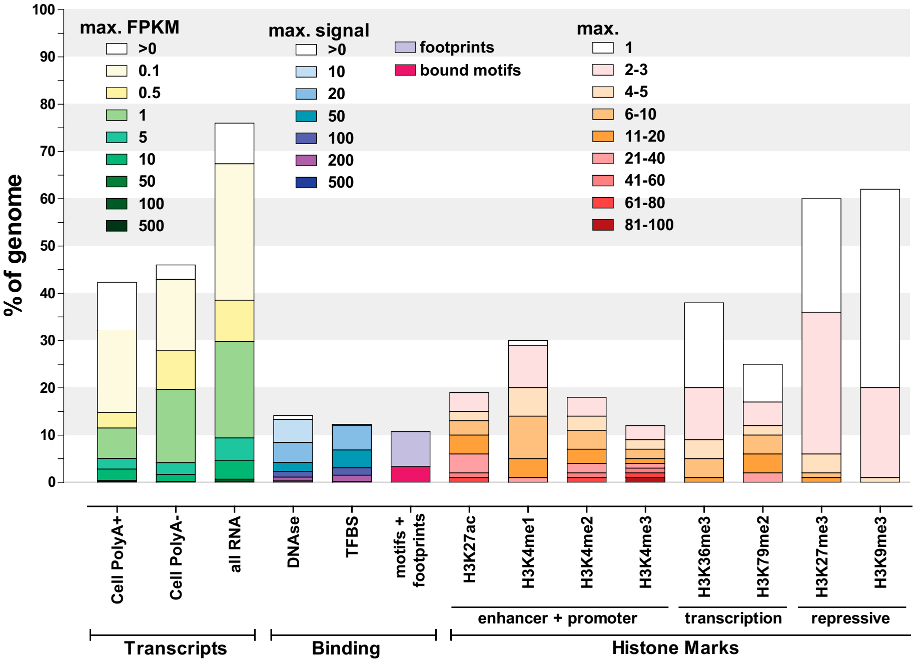


Fig. 2. Summary of the coverage of the human genome by ENCODE data. The fraction of the human genome covered by ENCODE-detected elements in at least one cell line or tissue for each assay is shown as a bar graph. All percentages are calculated against the whole genome, including the portion that is not uniquely mappable with short reads and thus is invisible to the analysis presented here (see Fig. S1). A more detailed summary can be found in Fig. S2. For transcripts, coverage was calculated from RNA-seq–derived contigs (104) using the count of read fragments per kilobase of exon per million reads (FPKM) and separated into abundance classes by FPKM values. Note that FPKMs are not directly comparable among different subcellular fractions, as they reflect relative abundances within a fraction rather than average absolute transcript copy numbers per cell. Depending on the total amount of RNA in a cell, one transcript copy per cell corresponds to between 0.5 and 5 FPKM in PolyA+ whole-cell samples according to current estimates (with the upper end of that range corresponding to small cells with little RNA and vice versa). "All RNA" refers to all RNA-seq experiments, including all subcellular fractions (Fig. S2). DNAse hypersensitivity and transcription-factor (TFBS) and histone-mark ChIP-seq coverage was calculated similarly but divided according to signal strength. "Motifs+footprints" refers to the union of occupied sequence recognition motifs for transcription factors as determined by ChIP-seq and as measured by digital genomic footprinting, with the fuscia portion of the bar representing the genomic space covered by bound motifs in ChIP-seq. Signal strength for ChIP-seq data for histone marks was determined based on the *P* value of each enriched region (the –log10 of the *P* value is shown), using peak-calling procedures tailored to the broadness of occupancy of each modification (*SI Methods*).

Case for Abundant Junk DNA. The possibility that much of a complex genome could be nonfunctional was raised decades ago. The C-value paradox (27, 73, 74) refers to the observation that genome size does not correlate with perceived organismal complexity and that even closely related species can have vastly different genome sizes. The estimated mutation rate in protein-coding genes suggested that only up to ∼20% of the nucleotides in the human genome can be selectively maintained, as the mutational burden would be otherwise too large (75). The term "junk DNA" was coined to refer to the majority of the rest of the genome, which represent segments of neutrally evolving DNA (76, 77). More recent work in population genetics has further developed this idea by emphasizing how the low effective population size of large-bodied eukaryotes leads to less efficient natural selection, permitting proliferation of transposable elements and other neutrally evolving DNA (78). If repetitive DNA elements could be equated with nonfunctional DNA, then one would surmise that the human genome contains vast nonfunctional regions because nearly 50% of nucleotides in the human genome are readily recognizable as repeat elements, often of high degeneracy. Moreover, comparative genomics studies have found that only 5% of mammalian genomes are under strong evolutionary constraint across multiple species (e.g., human, mouse, and dog) (2, 3).

Case for Abundant Functional Genomic Elements. Genome-wide biochemical studies, including recent reports from ENCODE, have revealed pervasive activity over an unexpectedly large fraction of the genome, including noncoding and nonconserved regions and repeat elements (58–60). Such results greatly increase upper bound estimates of candidate functional sequences (Fig. 2 and Fig. S2). Many human genomic regions previously assumed to be nonfunctional have recently been found to be teeming with biochemical activity, including portions of repeat elements, which can be bound by transcription factors and transcribed (79, 80), and are thought to sometimes be exapted into novel regulatory regions (81–84). Outside the 1.5% of the genome covered by protein-coding sequence, 11% of the genome is associated with motifs in transcription factor-bound regions or high-resolution DNase footprints in one or more cell types (Fig. 2), indicative of direct contact by regulatory proteins. Transcription factor occupancy and nucleosome-resolution DNase hypersensitivity maps overlap greatly and each cover approximately 15% of the genome. In aggregate, histone modifications associated with promoters or enhancers mark ∼20% of the genome, whereas a third of the genome is marked by modifications associated with transcriptional elongation. Over half of the genome has at least one repressive histone mark. In agreement with prior findings of pervasive transcription (85, 86), ENCODE maps of polyadenylated and total RNA cover in total more than 75% of the genome. These already large fractions may be underestimates, as only a subset of cell states have been assayed. However, for multiple reasons discussed below, it remains unclear what proportion of these biochemically annotated regions serve specific functions.

The lower bound estimate that 5% of the human genome has been under evolutionary constraint was based on the excess conservation observed in mammalian alignments (2, 3, 87) relative to a neutral reference (typically ancestral repeats, small introns, or fourfold degenerate codon positions). However, estimates that incorporate alternate references, shape-based constraint (88), evolutionary turnover (89), or lineage-specific constraint (90) each suggests roughly two to three times more constraint than previously (12–15%), and their union might be even larger as they each correct different aspects of alignment-based excess constraint. Moreover, the mutation rate estimates of the human genome are still uncertain and surprisingly low (91) and not inconsistent with a larger fraction of the genome under relatively weaker constraint (92). Although still weakly powered, human population studies suggest that an additional 4–11% of the genome may be under lineage-specific constraint after specifically excluding protein-coding regions (90, 92, 93), and these numbers may also increase as our ability to detect human constraint increases with additional human genomes. Thus, revised models, lineage-specific constraint, and additional datasets may further increase evolution-based estimates.

Results of genome-wide association studies might also be interpreted as support for more pervasive genome function. At present, significantly associated loci explain only a small fraction of the estimated trait heritability, suggesting that a vast number of additional loci with smaller effects remain to be discovered. Furthermore, quantitative trait locus (QTL) studies have revealed thousands of genetic variants that influence gene

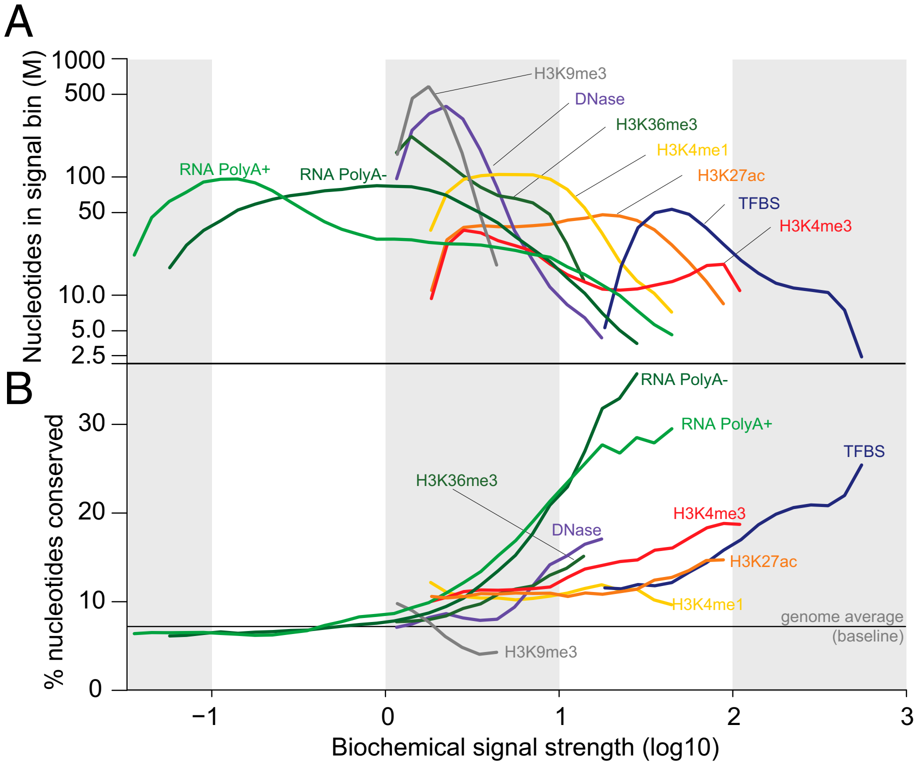


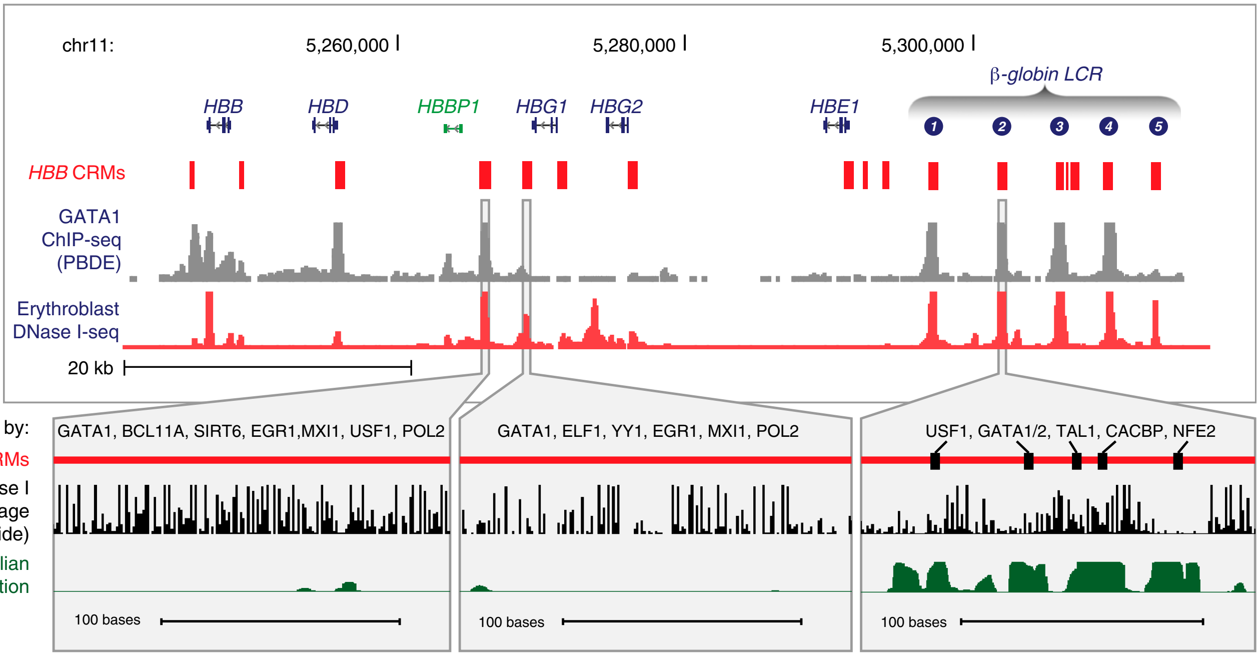
Fig. 3. Relationship between ENCODE signals and conservation. Signal strength of ENCODE functional annotations were defined as follows: log10 of signal intensity for DNase and TFBS, log10 of RPKM for RNA, and log10 of –log10 *P* value for histone modifications. Annotated regions were binned by 0.1 units of signal strength. (*A*) The number of nucleotides in each signal bin was plotted. (*B*) The fraction of the genome in each signal bin covered by conserved elements (by genomic evolutionary rate profiling) (115) was plotted.

expression and regulatory activity (94–98). These observations raise the possibility that functional sequences encompass a larger proportion of the human genome than previously thought.

Reconciling Genetic, Evolutionary, and Biochemical Estimates

The proportion of the human genome assigned to candidate functions varies markedly among the different approaches, with estimates from biochemical approaches being considerably larger than those of genetic and evolutionary approaches (Fig. 1). These differences have stimulated scientific debate regarding the interpretation and relative merits of the various approaches (26–29). We highlight below caveats of each approach and emphasize the importance of integration and new high-throughput technologies for refining estimates and better understanding the functional segments in the human genome.

Although ENCODE has expended considerable effort to ensure the reproducibility of detecting biochemical activity (99), it is not at all simple to establish what fraction of the biochemically annotated genome should be regarded as functional. The dynamic range of biochemical signals differs by one or more orders of magnitude for many assays, and the significance of the differing levels is not yet clear, particularly for lower levels. For example, RNA transcripts of some kind can be detected from ∼75% of the genome, but a significant portion of these are of low abundance (Fig. 2 and Fig. S2). For polyadenylated RNA, where it is possible to estimate abundance levels, 70% of the documented coverage is below approximately one transcript per cell (100–103). The abundance of complex nonpolyadenylated RNAs and RNAs from subcellular fractions, which account for half of the total RNA coverage of the genome, is likely to be even lower, although their absolute quantification is not yet achieved. Some RNAs, such as lncRNAs, might be active at very low levels. Others might be expressed stochastically at higher levels in a small fraction of the cell population (104), have hitherto unappreciated architectural or regulatory functions, or simply be biological noise of various kinds. At present, we cannot distinguish which low-abundance transcripts are functional, especially for RNAs that lack the defining characteristics of known protein coding, structural, or regulatory RNAs. A priori, we should not expect the transcriptome to consist exclusively of functional RNAs. Zero tolerance for errant transcripts would come at high cost in the proofreading machinery needed to perfectly gate RNA polymerase and splicing activities, or to instantly eliminate spurious transcripts. In general, sequences encoding RNAs transcribed by noisy transcriptional machinery are expected to be less constrained, which is consistent with data shown here for very low abundance



Evidence for function:	*HBG1*-D BCL11a	*HBG1* 3′ enhancer	β-globin LCR HS2
Genetics	–	+	+
Biochemical	+	+	+
Evolution	–	–	+

Fig. 4. Epigenetic and evolutionary signals in *cis*-regulatory modules (CRMs) of the *HBB* complex. (*Upper*) Many CRMs (red rectangles) (106) have been mapped within the cluster of genes encoding β-like globins expressed in embryonic (*HBE1*), fetal (*HBG1* and *HBG2*), and adult (*HBB* and *HBD*) erythroid cells. All are marked by DNase hypersensitive sites and footprints (Gene Expression Omnibus accession nos. GSE55579, GSM1339559, and GSM1339560), and many are bound by GATA1 in peripheral blood derived erythroblasts (PBDEs). (*Lower*, *Left*) A DNA segment located between the *HBG1* and *HBD* genes is one of the DNA segments bound by BCL11A (109, 110) and several other proteins (ENCODE uniformly processed data) to negatively regulate *HBG1* and *HBG2*. It is sensitive to DNase I but is not conserved across mammals. (*Center*) An enhancer located 3′ of the *HBG1* gene (red line) (108) is bound by several proteins in PBDEs and K562 cells (from the ENCODE uniformly processed data) and is sensitive to DNase I, but shows almost no signal for mammalian constraint. (*Right*) The enhancer at hypersensitive site (HS)2 of the locus control region (LCR) (red line) (107) is bound by the designated proteins at the motifs indicated by black rectangles. High-resolution DNase footprinting data (116) show cleavage concentrated between the bound motifs, which are strongly constrained during mammalian evolution, as shown on the mammalian phastCons track (48).

RNA (Fig. 3). Similarly, a majority of the genome shows reproducible evidence of one or more chromatin marks, but some marks are in much lower abundance, are preferentially associated with nonconserved heterochromatin regions (e.g., H3K9me3; Fig. 3*B*), or are known to act at a distance by spreading (105). Indeed, for any given biochemical assay, the proportion of the genome covered is highly dependent on the signal threshold set for the analysis (Fig. 2 and Fig. S2). Regions with higher signals generally exhibit higher levels of evolutionarily conservation (Fig. 3 and Fig. S3). Thus, one should have high confidence that the subset of the genome with large signals for RNA or chromatin signatures coupled with strong conservation is functional and will be supported by appropriate genetic tests. In contrast, the larger proportion of genome with reproducible but low biochemical signal strength and less evolutionary conservation is challenging to parse between specific functions and biological noise.

Another major variable underlying the differences in genome coverage is assay resolution. Biochemical methods, such as ChIP or DNase hypersensitivity assays, capture extended regions of several hundred bases, whereas the underlying transcription factor-binding elements are typically only 6–15 bp in length. Regulatory motifs and DNase footprints within bound regions show much stronger evidence of constraint than surrounding nucleotides that nevertheless fall within the region. Functional elements predicted from chromatin-state annotations tend to span even larger regions (e.g., the median length of enhancer states is ∼600 bp), although the driver nucleotides can be similarly few. Biochemical activity may also spread from neighboring regions, in genomic coordinates or 3D genome organization, making it even more difficult to establish the potential nucleotide drivers. Nonetheless, immediately consigning a biochemically marked region to the nonfunctional bin for lack of a driver motif would be premature. Genetic tests by deletion or sequence substitution are needed to resolve the question of their functional significance.

Thus, unanswered questions related to biological noise, along with differences in the resolution, sensitivity, and activity level of the corresponding assays, help to explain divergent estimates of the portion of the human genome encoding functional elements. Nevertheless, they do not account for the entire gulf between constrained regions and biochemical activity. Our analysis revealed a vast portion of the genome that appears to be evolving neutrally according to our metrics, even though it shows reproducible biochemical activity, which we previously referred to as "biochemically active but selectively neutral" (68). It could be argued that some of these regions are unlikely to serve critical functions, especially those with lower-level biochemical signal. However, we also acknowledge substantial limitations in our current detection of constraint, given that some human-specific functions are essential but not conserved and that disease-relevant regions need not be selectively constrained to be functional. Despite these limitations, all three approaches are needed to complete the unfinished process of inferring functional DNA elements, specifying their boundaries, and defining what functions they serve at molecular, cellular, and organismal levels.

Functional Genomic Elements and Human Disease

Presently, ∼4,000 genes have been associated with human disease, a likely underestimate given that the majority of disease-associated mutations have yet to be mapped. There is overwhelming evidence that variants in the regulatory sequences associated with such genes can lead to disease-relevant phenotypes. Biochemical approaches provide a rich resource for understanding disease-relevant functional elements, but they are most powerful as part of a multifaceted body of evidence for establishing function. Three specific examples from the β-globin locus illustrate how biochemical data can be integrated with evolutionary constraint and genetic assays of function (Fig. 4). The expression of globin genes at progressive stages of development is controlled by transcription factors binding at multiple *cis*-regulatory modules (CRMs) (106), but these CRMs differ dramatically in epigenetic signals and evolutionary history. For example, the independently acting enhancer LCR hypersensitive site 2 (HS2) (107) shows strong constraint on the motifs bound by transcription factors and strong DNase footprints. A second CRM, *HBG1* 3′ enhancer (108), is also bound in vivo by GATA1 (and other proteins) and is active as an enhancer, but shows almost no constraint over mammalian evolution. Last, a third location, *HBG1*-D (109, 110), shows DNase hypersensitivity but lacks biological activity in enhancer assays. Rather, binding of this and other CRMs in the locus by BCL11A leads to a reorganization of the chromatin interactions and repression of genes encoding the fetally expressed γ-globins in adult erythroid cells. This CRM is virtually devoid of evidence of mammalian constraint, at least in part because the adult-stage silencing of γ-globin genes is specific to primates. These vignettes illustrate the complementary nature of genetic, evolutionary, and biochemical approaches for understanding disease-relevant genomic elements and also the importance of data integration, as no single assay identifies all functional elements.

Conclusion

In contrast to evolutionary and genetic evidence, biochemical data offer clues about both the molecular function served by underlying DNA elements and the cell types in which they act, thus providing a launching point to study differentiation and development, cellular circuitry, and human disease (14, 35, 69, 111, 112). The major contribution of ENCODE to date has been high-resolution, highly-reproducible maps of DNA segments with biochemical signatures associated with diverse molecular functions. We believe that this public resource is far more important than any interim estimate of the fraction of the human genome that is functional.

By identifying candidate genomic elements and placing them into classes with shared molecular characteristics, the biochemical maps provide a starting point for testing how these signatures relate to molecular, cellular, and organismal function. The data identify very large numbers of sequence elements of differing sizes and signal strengths. Emerging genome-editing methods (113, 114) should considerably increase the throughput and resolution with which these candidate elements can be evaluated by genetic criteria. Given the limitations of our current understanding of genome function, future work should seek to better define genome elements by integrating all three methods to gain insight into the roles they play in human biology and disease.

1 Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

2 Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

3 Lindblad-Toh K, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.

4 Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? *Genome Res* 21(11):1769–1776.

5 Jones FC, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484(7392):55–61.

6 Grossman SR, et al.; 1000 Genomes Project (2013) Identifying recent adaptations in large-scale genomic data. *Cell* 152(4):703–713.

7 Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Res* 23(7):1089–1096.

8 Jeong S, et al. (2008) The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell* 132(5):783–793.

9 Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134(1):25–36.

10 Chan YF, et al. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327(5963):302–305.

11 Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am J Hum Genet* 76(1):8–32.

12 Kleinjan DA, Lettice LA (2008) Long-range gene control and genetic disease. *Adv Genet* 61:339–388.

13 Hindorff LA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106(23):9362–9367.

14 Maurano MT, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099): 1190–1195.

15 Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. *Genome Res* 22(9):1748–1759.

16 Ward LD, Kellis M (2012) HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40(Database issue):D930–D934.

17 Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: Conservation and turnover. *Mol Biol Evol* 19(7):1114–1121.

18 Costas J, Casares F, Vieira J (2003) Turnover of binding sites for transcription factors involved in early Drosophila development. *Gene* 310:215–220.

19 Moses AM, et al. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLOS Comput Biol* 2(10):e130.

20 Ludwig MZ, Patel NH, Kreitman M (1998) Functional analysis of eve stripe 2 enhancer evolution in Drosophila: Rules governing conservation and change. *Development* 125(5):949–958.

21 Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302(5644): 413.

22 Ahituv N, et al. (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol* 5(9):e234.

23 McGaughey DM, et al. (2008) Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res* 18(2):252–260.

24 Vakhrusheva OA, Bazykin GA, Kondrashov AS (2013) Genome-Level Analysis of Selective Constraint without Apparent Sequence Conservation. *Genome Biol Evol* 5(3):532–541.

25 Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* 110(14):5294–5300.

26 Graur D, et al. (2013) On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5(3):578–590.

27 Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. *Curr Biol* 22(21):R898–R899.

28 Eddy SR (2013) The ENCODE project: Missteps overshadowing a success. *Curr Biol* 23(7):R259–R261.

29 Mattick JS, et al. (2013) The extent of functionality in the human genome. *HUGO J* 7(1):2.

30 Niu DK, Jiang L (2012) Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun* 430(4): 1340–1343.

31 Germain PL, Ratti E, Boem F (2014) Junk or functional DNA?: ENCODE and the function controversy. *Biology & Philosophy*, 10.1007/s10539-014-9441-3.

32 Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33(Database issue):D514–D517.

33 Amsterdam A, et al. (1999) A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev* 13(20):2713–2724.

34 Berns K, et al. (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 428(6981): 431–437.

35 Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43–49.

36 Kheradpour P, et al. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 23(5):800–811.

37 Visel A, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457(7231):854–858.

38 Patwardhan RP, et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30(3): 265–270.

39 Melnikov A, et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3):271–277.

40 Pfeiffer BD, et al. (2008) Tools for neuroanatomy and neurogenetics in Drosophila. *Proc Natl Acad Sci USA* 105(28): 9715–9720.

41 MacArthur DG, et al.; 1000 Genomes Project Consortium (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335(6070):823–828.

42 Stark A, et al.; Harvard FlyBase curators; Berkeley Drosophila Genome Project (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450(7167):219–232.

43 Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.

44 Xie X, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* 434(7031):338–345.

45 Thomas JW, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424(6950): 788–793.

46 Cliften P, et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301(5629):71–76.

47 Boffelli D, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299(5611):1391–1394.

48 Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8): 1034–1050.

49 Elnitski L, et al. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res* 13(1):64–72.

50 Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. *Cell* 136(2):215–233.

51 Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166.

52 Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318(5851):761–764.

53 Olovnikov I, Aravin AA, Fejes Toth K (2012) Small RNA in the nucleus: The RNA-chromatin ping-pong. *Curr Opin Genet Dev* 22(2): 164–171.

54 Grosveld F, van Assendelft GB, Greaves DR, Kollias G (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* 51(6):975–985.

55 Agarwal S, Rao A (1998) Long-range transcriptional regulation of cytokine gene expression. *Curr Opin Immunol* 10(3):345–352.

56 Lakshmanan G, Lieuw KH, Grosveld F, Engel JD (1998) Partial rescue of GATA-3 by yeast artificial chromosome transgenes. *Dev Biol* 204(2):451–463.

57 Noonan JP, McCallion AS (2010) Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* 11:1–23.

58 Nardone J, Lee DU, Ansel KM, Rao A (2004) Bioinformatics for the 'bench biologist': How to find regulatory regions in genomic DNA. *Nat Immunol* 5(8):768–774.

59 Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57:159–197.

60 Li CC, Ramirez-Carrozzi VR, Smale ST (2006) Pursuing gene regulation 'logic' via RNA interference and chromatin immunoprecipitation. *Nat Immunol* 7(7):692–697.

61 Weinmann AS, Farnham PJ (2002) Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* 26(1):37–47.

62 Johnson KD, Bresnick EH (2002) Dissecting long-range transcriptional mechanisms by chromatin immunoprecipitation. *Methods* 26(1):27–36.

63 Rada-Iglesias A, et al. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333):279–283.

64 Creyghton MP, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107(50):21931–21936.

65 Ozsolak F, et al. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev* 22(22):3172–3183.

66 Horak CE, Snyder M (2002) Global analysis of gene expression in yeast. *Funct Integr Genomics* 2(4-5):171–180.

67 ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696):636–640.

68 Birney E, et al.; ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799–816.

69 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.

70 Cheng Y, et al. (2009) Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* 19(12): 2172–2184.

71 Henikoff S, Shilatifard A (2011) Histone modification: Cause or cog? *Trends Genet* 27(10):389–396.

72 Weiner A, et al. (2012) Systematic dissection of roles for chromatin regulators in a yeast stress response. *PLoS Biol* 10(7):e1001369.

73 Thomas CA, Jr. (1971) The genetic organization of chromosomes. *Annu Rev Genet* 5:237–256.

74 Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc* 76(1):65–101.

75 Keightley PD (2012) Rates and fitness consequences of new mutations in humans. *Genetics* 190(2):295–304.

76 Ehret CF, De Haller G (1963) Origin, development and maturation of organelles and organelle systems of the cell surface in Paramecium. *J Ultrastruct Res* 23(Suppl 6):1–42.

77 Ohno S (1972) So much "junk" DNA in our genome. *Brookhaven Symp Biol* 23:366–370.

78 Lynch M (2007) *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA).

79 Kamal M, Xie X, Lander ES (2006) A large family of ancient repeat elements in the human genome is under strong selection. *Proc Natl Acad Sci USA* 103(8):2740–2745.

80 Lowe CB, Bejerano G, Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci USA* 104(19):8005–8010.

81 Lowe CB, et al. (2011) Three periods of regulatory innovation during vertebrate evolution. *Science* 333(6045):1019–1024.

82 McClintock B (1956) Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* 21:197–216.

83 de Souza FS, Franchini LF, Rubinstein M (2013) Exaptation of transposable elements into novel cis-regulatory elements: Is the evidence always strong? *Mol Biol Evol* 30(6):1239–1251.

84 Nishihara H, Smit AF, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16(7):864–874.

85 Clark MB, et al. (2011) The reality of pervasive transcription. *PLoS Biol*, 9(7):e1000625, discussion e1001102.

86 Jacquier A (2009) The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* 10(12):833–844.

87 Lindblad-Toh K, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–819.

88 Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324(5925):389–392.

89 Meader S, Ponting CP, Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* 20(10):1335–1343.

90 Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675–1678.

91 Scally A, Durbin R (2012) Revising the human mutation rate: Implications for understanding human evolution. *Nat Rev Genet* 13(10):745–753.

92 Lohmueller KE, et al. (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* 7(10):e1002326.

93 Ward LD, Kellis M (2013) Response to comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions" *Science* 340(6133):682.

94 Dimas AS, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325(5945): 1246–1250.

95 Montgomery SB, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464(7289):773–777.

96 Battle A, et al. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24(1):14–24.

97 Degner JF, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482(7385):390–394.

98 Pickrell JK, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289):768–772.

99 Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5(3):27.

100 Lovén J, et al. (2012) Revisiting global gene expression analysis. *Cell* 151(3):476–482.

101 Islam S, et al. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 21(7):1160–1167.

102 Marinov GK, et al. (2014) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res*.

103 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.

104 Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.

105 Talbert PB, Henikoff S (2006) Spreading of silent chromatin: Inaction at a distance. *Nat Rev Genet* 7(10):793–803.

106 King DC, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* 15(8): 1051–1060.

107 Tuan DY, Solomon WB, London IM, Lee DP (1989) An erythroid-specific, developmental-stage-independent enhancer far upstream of the human "beta-like globin" genes. *Proc Natl Acad Sci USA* 86(8):2554–2558.

108 Bodine DM, Ley TJ (1987) An enhancer element lies 3′ to the human A gamma globin gene. *EMBO J* 6(10):2997–3004.

109 Xu J, et al. (2010) Transcriptional silencing of gamma-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev* 24(8):783–798.

110 Sankaran VG, et al. (2011) A functional element necessary for fetal hemoglobin silencing. *N Engl J Med* 365(9):807–814.

111 Gerstein MB, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91–100.

112 Trynka G, et al. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45(2): 124–130.

113 Ran FA, et al. (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154(6): 1380–1389.

114 Carr PA, Church GM (2009) Genome engineering. *Nat Biotechnol* 27(12):1151–1162.

115 Davydov EV, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput Biol* 6(12):e1001025.

116 Hesselberth JR, et al. (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 6(4):283–289.

N

Other publications:

ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology.* **9**(4):e1001046. doi: 10.1371/journal.pbio.1001046.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G, Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE, Weissman SM, Cayting P, Hariharan M, Lian J, Cheng Y, Landt SG, Ma Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C, Kumar SA, Mishra T, Jain D, Byrska-Bishop M, Blankenberg D, Lajoie1 BR, Jain G, Sanyal A, Chen KB, Denas O, Taylor J, Blobel GA, Weiss MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI, Desalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L, McCleary D, Kuan S, Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S, Xue C, Dobin A, Lin W, Fastuca M, Wang H, Guigo R, Djebali S, Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM, Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin M, Lowdon RF, Adams LB. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**(8):418. doi:10.1186/gb-2012-13-8-418.

Part VII

Bibliography

17

Bibliography

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422):56–65.

Abdella PM, Smith PK, Royer GP. 1979. A new cleavable reagent for cross-linking and reversible immobilization of proteins. *Biochem Biophys Res Commun* **87**(3):734–742.

Abdallah F, Salamini F, Leister D. 2000. A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci* **5**(4):141–142.

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**(5669):441–445.

Achanta G, Sasaki R, Feng L, Carew JS, Lu W, Pelicano H, Keating MJ, Huang P. 2005. Novel role of p53 in maintaining mitochondrial genetic stability through interaction with DNA Pol gamma. *EMBO Journal* **24**(19):3482–3492.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu

X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**(5461):2185–2195.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**(5013):1651–1656.

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**(6547 Suppl):3–174.

Adl SM, Simpson AG, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, Le Gall L, Lynn DH, McManus H, Mitchell EA, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick RS, Schoch CL, Smirnov A, Spiegel FW. 2012. The revised classification of eukaryotes. *J Eukaryot Microbiol* **59**(5):429–493.

Adli M, Bernstein BE. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc* 6(10):1656–1668.

Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**(7232):1028–1032.

Agger K, Cloos P, Christensen J, Pasini D, Rose S, Rappsilber J, Issaeva I, Canaani E, Salcini A, Helin K. 2007. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature* **449**:731-734.

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**(9):e234.

Akerman M, Mandel-Gutfreund Y. 2006. Alternative splicing regulation at tandem 3' splice sites. *Nucleic Acids Res* **34**(1):23–31.

Aksoy O, Chicas A, Zeng T, Zhao Z, McCurrach M, Wang X, Lowe SW. 2012. The atypical E2F family member E2F7 couples the p53 and RB pathways during cellular senescence. *Genes Dev* **26**(14):1546–1557.

Ala U, Karreth FA, Bosia C, Pagnani A, Taulli R, Léopold V, Tay Y, Provero P, Zecchina R, Pandolfi PP. 2013. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc Natl Acad Sci U S A* **110**(18):7154–7159.

Alam TI, Kanki T, Muta T, Ukaji K, Abe Y, Nakayama H, Takio K, Hamasaki N, Kang D. 2003. Human mitochondrial DNA is packaged with TFAM. *Nucleic Acids Res* **31**(6):1640–1645.

Alcolea PJ, Alonso A, Gómez MJ, Moreno I, Domínguez M, Parro V, Larraga V. 2010. Transcriptomics throughout the life cycle of *Leishmania infantum*: high down-regulation rate in the amastigote stage. *Int J Parasitol* **40**(13):1497–1516.

Aldridge S, Watt S, Quail MA, Rayner T, Lukk M, Bimson MF, Gaffney D, Odom DT. 2013. AHT-ChIP-seq: a completely automated robotic protocol for high-throughput chromatin immunoprecipitation. *Genome Biol* **14**(11):R124.

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**(1):61–65.

Allen TA, Von Kaenel S, Goodrich JA, Kugel JF. 2004. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. *Nat Struct Mol Biol* **11**(9):816–821.

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**(7458):360–363.

Altmann R. 1890. Die Elementarorganismen und ihre Beziehungen zu den Zellen. *Veit & Comp, Liepzig*

Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *Plant Cell* **23**(7):2499–2513.

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. *Science* **319**(5871):1787–1789.

Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* **342**(6165):1241089

Ameyar-Zazoua M, Rachez C, Souidi M, Robin P, Fritsch L, Young R, Morozova N, Fenouil R, Descostes N, Andrau JC, Mathieu J, Hamiche A, Ait-Si-Ali S, Muchardt

C, Batsché E, Harel-Bellan A. 2012. Argonaute proteins couple chromatin silencing to alternative splicing. *Nat Struct Mol Biol* **19**(10):998–1004.

Ammermann D. 1971. Morphology and development of the macronuclei of the ciliates *Stylonychia mytilus* and *Euplotes aediculatus*. *Chromosoma* **33**(2):209–238.

Ameur A, Wetterbom A, Feuk L, Gyllensten U. 2010. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol* **11**(3):R34.

An CI, Dong Y, Hagiwara N. 2011. Genome-wide mapping of Sox6 binding sites in skeletal muscle reveals both direct and indirect regulation of muscle terminal differentiation by Sox6. *BMC Dev Biol* **11**:59.

Anders L, Guenther MG, Qi J, Fan ZP, Marineau JJ, Rahl PB, Lovén J, Sigova AA, Smith WB, Lee TI, Bradner JE, Young RA. 2014. Genome-wide localization of small molecules. *Nat Biotechnol* **32**(1):92–96.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**(10):R106.

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**(10):2008–2017.

Andersen JL, Kornbluth S. 2013. The tangled circuitry of metabolism and apoptosis. *Mol Cell* **49**(3):399-410.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290**(5806):457–465.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**(7062):1149–1152.

Ang YS, Tsai SY, Lee DF, Monk J, Su J, Ratnakumar K, Ding J, Ge Y, Darr H, Chang B, Wang J, Rendl M, Bernstein E, Schaniel C, Lemischka IR. 2011. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145**(2):183–197.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**(5585):1301–1310.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**(6814):796–815.

Arany Z, Sellers WR, Livingston DM, Eckner R. 1994. E1A-associated p300 and CREB-associated CBP belong to a conserved family of coactivators. *Cell* **77**(6):799–800.

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, Chien M, Russo JJ, Ju J, Sheridan R, Sander C, Zavolan M, Tuschl T. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**(7099):203–207.

Aravin AA, Hannon GJ, Brennecke J. 2007a. The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**(5851):761–764.

Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* **5**:337-350.

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007b. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**(5825):744–747.

Aravin AA, Sachidanandam R, Bourchis D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31**:785-799.

Archibald JM, Keeling PJ. 2002. Recycled plastids: a ’green movement’ in eukaryotic evolution. *Trends Genet* **18**(11):577–584.

Archibald JM, Lane CE. 2009. Going, going, not quite gone: nucleomorphs as a case study in nuclear genome reduction. *J Hered* **100**(5):582–590.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Krger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**(5693):79–86.

Arnberg A, van Bruggen EF, Borst P. 1971. The presence of DNA molecules with a displacement loop in standard mitochondrial DNA preparations. *Biochim Biophys Acta* **246**(2):353–357.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**(6123):1074–1077.

Aronica L, Bednenko J, Noto T, DeSouza LV, Siu KW, Loidl J, Pearlman RE, Gorovsky MA, Mochizuki K. 2008. Study of an RNA helicase implicates small RNA-noncoding RNA interactions in programmed DNA elimination in *Tetrahymena. Genes Dev* **22**:2228-2241

Aschoff M, Hotz-Wagenblatt A, Glatting KH, Fischer M, Eils R, König R. 2013. Splicing-Compass: differential splicing detection using RNA-seq data. *Bioinformatics* **29**(9):1141–1148.

Ashe A, Sapetschnig A, Weick EM, Mitchell J, Bagijn MP, Cording AC, Doebley AL, Goldstein LD, Lehrbach NJ, Le Pen J, Pintacuda G, Sakaguchi A, Sarkies P, Ahmed S, Miska EA. 2012. piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans. Cell* **150**(1):88–99.

Asin-Cayuela J, Gustafsson CM. 2007. Mitochondrial transcription and its regulation in mammalian cells. *Trends Biochem Sci* **32**(3):111–117.

Au KF, Jiang H, Lin L, Xing Y, Wong WH. 2010. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* **38**(14):4570–4578.

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, van Bakel H, Schadt EE, Reijo-Pera RA, Underwood JG, Wong WH. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci U S A* **110**(50):E4821–4830.

Au KF, Underwood JG, Lee L, Wong WH. 2012. Improving PacBio long read accuracy by short read alignment. *PLoS One* **7**(10):e46679.

Auer TO, Duroure K, De Cian A, Concordet JP, Del Bene F. 2014. Highly efficient CRISPR/Cas9-mediated knock-in in zebrafish by homology-independent DNA repair. *Genome Res* 24(1):142–153.

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefranois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* **106**(35):14926–14931.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouël A, Lepère G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Bétermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia. Nature* **444**(7116):171–178.

Avery OT, Macleod CM, McCarty M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from *Pneumococcus* Type III. *J Exp Med* **79**(2):137-158.

Avvakumov N, Lalonde ME, Saksouk N, Paquet E, Glass KC, Landry AJ, Doyon Y, Cayrou C, Robitaille GA, Richard DE, Yang XJ, Kutateladze TG, Côté J. 2012. Conserved molecular interactions within the HBO1 acetyltransferase complexes regulate cell proliferation. *Mol Cell Biol* **32**(3):689–703.

Ayliffe MA, Scott NS, Timmis JN. 1998. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. *Mol Biol Evol* **15**:738745

Ayub M, Bayley H. 2012. Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore. *Nano Lett* **12**(11):5637–5643.

Ayub M, Hardwick SW, Luisi BF, Bayley H. 2013. Nanopore-based identification of individual nucleotides for direct RNA sequencing. *Nano Lett* **13**(12):6144–6150.

Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. 2007. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* **318:**798-801.

Bachvaroff TR, Place AR. 2008. From stop to start: Tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae. PLoS One* **3**:e2929

Bagijn M, Goldstein L, Sapetschnig A, Weick E-M, Bouasker S, Lehrbach N, Simard M, Miska E. 2012. Function, targets, and evolution of *Caenorhabditis elegans* piRNAs. *Science* **337**:574-578.

Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**(1):142–150.

Bailey TL, Bodén M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Res* **37**:W202–W208.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**(7374):534–537.

Baird SE, Fino GM, Tausta SL, Klobutcher LA. 1989. Micronuclear genome organization in *Euplotes crassus*: a transposonlike element is removed during macronuclear development. *Mol Cell Biol* **9**(9):3793–3807.

Balakirev ES, Ayala FJ. 2003. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37**:123–151.

Balhoff JP, Wray GA. 2005. Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites. *Proc Natl Acad Sci U S A* **102**(24):8591–8596.

Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol* **10**(7):R79.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**:299-308

Banerji J, Olson L, Schaffner W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**:729-740.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**(5):455–477.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**(6114):1587–1593.

Bao E, Jiang T, Girke T. BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences. *Bioinformatics* **29**(10):1250–1259.

Bao H, Xiong Y, Guo H, Zhou R, Lu X, Yang Z, Zhong Y, Shi S. 2009. MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads. *BMC Genomics* **10**(Suppl 3):S13.

Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RE, Howe CJ. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae. Plant Mol Biol* **79**(4-5):347–357.

Barbrook AC, Howe CJ. 2000. Minicircular plastid DNA in the dinoflagellate *Amphidinium operculatum. Mol Gen Genet* **263**:152-158

Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. 2012. Organization and expression of organellar genomes. *Philos Trans R Soc Lond B Biol Sci* **365**(1541):785–797.

Barish GD, Yu RT, Karunasiri M, Ocampo CB, Dixon J, Benner C, Dent AL, Tangirala RK, Evans RM. 2010. Bcl-6 and NF-κB cistromes mediate opposing regulation of the innate immune response. *Genes Dev* **24**(24):2760–2765.

Barish GD, Yu RT, Karunasiri MS, Becerra D, Kim J, Tseng TW, Tai LJ, Leblanc M, Diehl C, Cerchietti L, Miller YI, Witztum JL, Melnick AM, Dent AL, Tangirala RK, Evans RM. 2012. The Bcl6-SMRT/NCoR cistrome represses inflammation to attenuate atherosclerosis. *Cell Metab* **15**(4):554–562.

Barski A, Cuddapah S, Cui K, Roh T, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:**823-837.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2):281–297.

Bártfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, Treeck M, Gilberger TW, Françoijs KJ, Stunnenberg HG. 2010. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLoS Pathog* **6**(12):e1001223.

Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S, Conte D Jr, Luo S, Schroth GP, Carrington JC, Bartel DP, Mello CC. 2008. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol Cell* **31**(1):67–78.

Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* **23**(1):169–180.

Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. 2011. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**(1):107–114.

Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Bétermier M. 2009. PiggyMac, a domesticated *piggyBac* transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev* **23**(21):2478–2483.

Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC, Pinello L, Sabo PJ, Vierstra J, Voit RA, Yuan GC, Porteus MH, Stamatoyannopoulos JA, Lettre G, Orkin SH. 2013. An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science* **342**(6155):253–257.

Bayer T, Aranda M, Sunagawa S, Yum LK, Desalvo MK, Lindquist E, Coffroth MA, Voolstra CR, Medina M. 2012. *Symbiodinium* transcriptomes: genome insights into the dinoflagellate symbionts of reef-building corals. *PLoS One* **7**(4):e35269.

Beaton MJ, Cavalier-Smith T. 1999. Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonad genomes. *Proc Biol Sci* **266**(1433):2053–2059.

Behe MJ. 2003 A functional pseudogene: an open letter to Nature. `http://www.discovery.org/a/1448`

Behr J, Bohnert R, Zeller G, Schweikert G, Hartmann L, Rätsch G. 2010. Next generation genome annotation with mGene.ngs. *BMC Bioinformatics* **11**(Suppl 10):O8

Behr J, Kahles A, Zhong Y, Sreedharan VT, Drewe P, Rätsch G. 2013. MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics* **29**(20):2529–2538.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**(7089):87–90.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**(5675):1321–1325.

Bendich AJ. 1993. Reaching for the ring: the study of mitochondrial genome structure. *Curr Genet* **24**:279-290

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. 1986. Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**(6):819–826.

Bente M, Harder S, Wiesgigl M, Heukeshoven J, Gelhaus C, Krause E, Clos J, Bruchhaus I. 2003. Developmentally induced changes of the proteome in the protozoan parasite *Leishmania donovani.* *Proteomics* **3**(9):1811–1829.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley

NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218):53–59.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**(8):3171–3175.

Bergman J. 2001. The functions of introns: from junk DNA to designed DNA. *Perspectives on Science and Christian Faith* **53**:170–178.

Bergsland M, Ramsköld D, Zaouter C, Klum S, Sandberg R, Muhr J. 2011. Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev* **25**(23):2453–2464.

Berkes CA, Tapscott SJ. 2005. MyoD and the transcriptional control of myogenesis. *Semin Cell Dev Biol* **16**(4-5):585–595.

Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R. 2004. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**(6981):431–437.

Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS, Kouzarides T, Schreiber SL. 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A* **99**:8695-8700.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. *Cell* **128**(4):669–681.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**(10):1045–1048.

Bernt KM, Zhu N, Sinha AU, Vempati S, Faber J, Krivtsov AV, Feng Z, Punt N, Daigle A, Bullinger L, Pollock RM, Richon VM, Kung AL, Armstrong SA. 2011. MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. *Cancer Cell* **20**(1):66–78.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Bhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser

A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM. 2005. The genome of the African trypanosome *Trypanosoma brucei.* *Science* **309**(5733):416–422.

Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. 2009. Next generation software for functional trend analysis. *Bioinformatics* **25**(22):3043–3044.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705):2242–2246.

Bestor T, Laudano A, Mattaliano R, Ingram V. 1988. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* **203**(4):971–983.

Bhattacharjee Y. The Vigilante. *Science* **343**(6177):1306–1309

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA. 1981. Sequence and gene organization of mouse mitochondrial DNA. *Cell* **26**(2 Pt 2):167–180.

Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ. 2012. Introns in UTRs: why we should stop ignoring them. *Bioessays* **34**(12):1025–1034.

Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA. 2009. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**(21):2484–2489.

Bird AP. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321**(6067):209–213.

Blahnik KR, Dou L, O'Geen H, McPhillips T, Xu X, Cao AR, Iyengar S, Nicolet CM, Lud'"ascher B, Korf I, Farnham PJ. 2010. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res* **38**(3):e13.

Blake CC. 1979. Exons encode protein functional units. *Nature* **277**(5698):598

Blake WJ, Kærn M, Cantor CR, Collins JJ. 2003. Noise in eukaryotic gene expression. *Nature* **422**(6932):633–637.

Blau HM, Blakely BT. 1999. Plasticity of cell fate: insights from heterokaryons. *Semin Cell Dev Biol* **10**(3):267–272.

Blau HM, Chiu CP, Webster C. 1983. Cytoplasmic activation of human nuclear genes in stable heterocaryons. *Cell* **32**:1171–1180

Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. 2013. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat Protoc* **8**(3):539–554.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Bristow J, Ren B, Black BL, Rubin EM, Visel A, Pennacchio LA. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**(9):806–810.

Blum B, Bakalara N, Simpson L. 1990. A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* **60**(2):189–198.

Boengler K, Hilfiker-Kleiner D, Heusch G, Schulz R. 2010. Inhibition of permeability transition pore opening by mitochondrial STAT3 and its role in myocardial ischemia/reperfusion. *Basic Res Cardiol* **105**(6):771–785.

Boergesen M, Pedersen TÅ, Gross B, van Heeringen SJ, Hagenbeek D, Bindesbøll C, Caron S, Lalloyer F, Steffensen KR, Nebb HI, Gustafsson JÅ, Stunnenberg HG, Staels B, Mandrup S. 2012. Genome-wide profiling of liver X receptor, retinoid X receptor, and peroxisome proliferator-activated receptor a in mouse liver reveals extensive sharing of binding sites. *Mol Cell Biol* **32**(4):852–

867.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**(5611):1391–1394.

Bogenhagen D, Clayton DA. 1974. The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. *J Biol Chem* **249**(24):7991–7995.

Bogenhagen DF, Rousseau D, Burke S. 2008. The layered structure of human mitochondrial DNA nucleoids. *J Biol Chem* **283**(6):3665–3675.

Bogorad L. 2008. Evolution of early eukaryotic cells: genomes, proteomes, and compartments. *Photosynth Res* **95**(1):11–21.

Bohnert R, Behr J, Rätsch G. 2009. Transcript quantification with RNA-Seq data. *BMC Bioinformatics* **10**(Suppl 13):P5.

Bohnert R, Rätsch G. 2010. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res* **38**(Web Server issue):W348–351.

Bolduc N, Yilmaz A, Mejia-Guerra MK, Morohashi K, O'Connor D, Grotewold E, Hake S. 2012. Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes Dev* **26**(15):1685–1690.

Boley N, Stoiber MH, Booth BW, Wan KH, Hoskins RA, Bickel PJ, Celniker SE, Brown JB. 2014. Genome-guided transcript assembly by integrative analysis of RNA sequence data. *Nat Biotechnol* Nat Biotechnol **32**(4):341–346.

Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S. 2012. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**(6083):934–937.

Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D, Capra V, Grompe M, Pizzuti A, Muzny D, Lawrence C, Willard HF, Avner P, Ballabio A. 1991. Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**(6324):325–329.

Botcheva K, McCorkle SR, McCombie WR, Dunn JJ, Anderson CW. 2011. Distinct p53 genomic binding patterns in normal and cancer-derived human cells. *Cell Cycle* **10**(24):4237–4249.

Bougdour A, Braun L, Cannella D, Hakimi MA. 2010. Chromatin modifications: implications in the regulation of gene expression in *Toxoplasma gondii. Cell Microbiol* **12**(4):413–423.

Boveri TH 1904. Ergebnisse über die Konstitution der chromatischen Substanz des Zelkerns. *Fisher, Jena.*

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kröger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jézéquel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**(7219):239–244.

Bowmaker M, Yang MY, Yasukawa T, Reyes A, Jacobs HT, Huberman JA, Holt IJ. 2003. Mammalian mitochondrial DNA replicates bidirectionally from an initiation zone. *J Biol Chem* **278**(51):50961–50969.

Boxma B, de Graaf RM, van der Staay GW, van Alen TA, Ricard G, Gabaldón T, van Hoek AH, Moon-van der Staay SY, Koopman WJ, van Hellemond JJ, Tielens AG, Friedrich T, Veenhuis M, Huynen MA, Hackstein JH. 2005. An anaerobic mitochondrion that produces hydrogen. *Nature* **434**(7029):74–79.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**(6):947–956.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinfor-*

matics **24**(21):2537–2538.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**(9):1790–1767.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* **21**(3):456–464.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**(18):3710–3715.

Bradford JR, Farren M, Powell SJ, Runswick S, Weston SL, Brown H, Delpuech O, Wappett M, Smith NR, Carr TH, Dry JR, Gibson NJ, Barry ST. 2013. RNA-Seq Differentiates Tumour and Host mRNA Expression Changes Induced by Treatment of Human Tumour Xenografts with the VEGFR Tyrosine Kinase Inhibitor Cediranib. *PLoS ONE* **8**(6):e66003.

Bradley RK, Merkin J, Lambert NJ, Burge CB. 2012. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol* **10**(1):e1001229.

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol* **8**(3):e1000343.

Brady JJ, Li M, Suthram S, Jiang H, Wong WH, Blau HM. 2013. Early role for IL-6 signalling during generation of induced pluripotent stem cells revealed by heterokaryon RNA-Seq. *Nat Cell Biol* **15**(10):1244–1252.

Brainerd EL, Slutz SS, Hall EK, Phillis RW. 2001. Patterns of genome size evolution in tetraodontiform fishes. *Evolution* **55**:2363–2368.

Brandt R, Salla-Martret M, Bou-Torrent J, Musielak T, Stahl M, Lanz C, Ott F, Schmid M, Greb T, Schwarz M, Choi SB, Barton MK, Reinhart BJ, Liu T, Quint M, Palauqui JC, Martínez-García JF, Wenkel S. 2012. Genome-wide binding-site analysis of REVOLUTA reveals a link between leaf patterning and light-mediated growth responses. *Plant J* **72**(1):31–42.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**(10):1146–1153.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**(6):1089–1103.

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**(5906):1387–1392.

Brenner S. 1998. Refuge of spandrels. *Curr Biol* **8**:R669

Brenner S, Stretton AO, Kaplan S. 1965. Genetic code: the 'nonsense' triplets for chain termination and their suppression. *Nature* **206**(988):994–998.

Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG. 2010. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **52**(3):232–236.

Britten R, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165**(3891):349-357.

Britten RJ. Davidson EH. 1968. Repetitive and Non-Repetitive DNA Sequences and a Speculation on Origins of Evolutionary Novelty. *Q Rev Biol* **46**(2):111–138.

Britten RJ, Kohne DE. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**(3841):529–540.

Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, Ribacke U, Volkman S, Duraisingh M, Wirth D, Sabeti PC, Rinn JL. 2011. A global transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol* **12**(6):R56.

Brouilette S, Kuersten S, Mein C, Bozek M, Terry A, Dias KR, Bhaw-Rosun L, Shintani Y, Coppen S, Ikebe C, Sawhney V, Campbell N, Kaneko M, Tano N, Ishida H, Suzuki K, Yashiro K. 2012. A simple and novel method for RNA-seq library preparation of single cell cDNA analysis by hyperactive Tn5 transposase. *Dev Dyn* **241**(10):1584–1590.

Brower-Toland B, Findley S, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin S, Lin H. 2007. *Drosophila* PIWI associates with chromatin and interacts directly with HP1a. *Genes Dev* **21**:2300-2311.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**(6304):38–44.

Brown D, Boytchev H. 2012 Sep 5. Junk DNA concept debunked by new analysis of human genome. *The Washington Post*.

Brown S, Teo A, Pauklin S, Hannan N, Cho CH, Lim B, Vardy L, Dunn NR, Trotter M, Pedersen R, Vallier L. 2011. Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. *Stem Cells* **29**(8):1176–1185.

Brownell JE, Zhou J, Ranalli T, Kobayashi R, Edmondson DG, Roth SY, Allis CD. 1996. *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* **84**(6):843–851.

Bryant DW Jr, Shen R, Priest HD, Wong WK, Mockler TC. 2010. Supersplat - spliced RNA-seq alignment. *Bioinformatics* **26**(12):1500–1505.

Buckley B, Burkhart K, Gu S, Spracklin G, Kershner A, Fritz H, Kimble J, Fire A, Kennedy S. 2012. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489**:447-451.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**(12):1213–1218.

Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, Jaenisch R. 2012. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**(6):1209–1222.

Buganim Y, Faddah DA, Jaenisch R. 2013. Mechanisms and models of somatic cell reprogramming. *Nat Rev Genet* **14**(6):427–439.

Bugge A, Feng D, Everett LJ, Briggs ER, Mullican SE, Wang F, Jager J, Lazar MA. 2011. Rev-erbα and Rev-erbβ coordinately protect the circadian clock and normal metabolic function. *Genes Dev* **26**(7):657–667.

Bullerwell CE, Leigh J, Forget L, Lang BF. 2003. A comparison of three fission yeast mitochondrial genomes. *Nucleic Acids Res* **31**(2):759–768.

Buratowski S, Hahn S, Guarente L, Sharp PA. 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**(4):549–561.

Buratowski S. 2009. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **36**(4):541–546.

Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. 2010. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet* **6**(12):e1001233.

Burda P, Laslo P, Stopka T. 2010. The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia* **24**(7):1249–1257.

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell* **2**(6):773–785.

Burger G, Forget L, Zhu Y, Gray MW, Lang BF. 2003. Unique mitochondrial genome architecture in unicellular relatives of animals. *Proc Natl Acad Sci U S A* **100**(3):892–897.

Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol* **5**(2):418–438.

Burger G, Gray MW, Lang BF. 2003. Mitochondrial genomes: anything goes. *Trends Genet* **19**(12):709–716.

Burke T, Kadonaga J. 1996. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **10**:711-724

Buske OJ, Hoffman MM, Ponts N, Le Roch KG, Noble WS. 2011. Exploratory analysis of genomic segmentations with Segtools. *BMC Bioinformatics* **12**:415.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **18**(5):810–820.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**(5396):2012–2018.

Caballero A. 1994. Developments in the prediction of effective population size. *Heredity* **73**:657-679.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**(18):1915–1927.

Cairns J, Spyrou C, Stark R, Smith ML, Lynch AG, Tavaré S. 2011. BayesPeak – an R package for analysing ChIP-seq data. *Bioinformatics* **27**(5):713–714.

Cammarota M, Paratcha G, Bevilaqua LR, Levi de Stein M, Lopez M, Pellegrino de Iraldi A, Izquierdo I, Medina JH. 1999. Cyclic AMP-responsive element binding protein in brain mitochondria. *J Neurochem* **72**(6):2272–2277.

Campbell CT, Kolesar JE, Kaufman BA. 2012. Mitochondrial transcription factor A regulates mitochondrial transcription initiation, DNA packaging, and genome copy number. *Biochim Biophys Acta* **1819**(9-10):921–929.

Campbell DA, Thomas S, Sturm NR. 2003. Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect* **5**(13):1231–1240.

Canella D, Bernasconi D, Gilardi F, LeMartelot G, Migliavacca E, Praz V, Cousin P, Delorenzi M, Hernandez N; CycliX Consortium. 2012. A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. *Genome Res* **22**(4):666–680.

Cann GM, Gulzar ZG, Cooper S, Li R, Luo S, Tat M, Stuart S, Schroth G, Srinivas S, Ronaghi M, Brooks JD, Talasaz AH. 2012. mRNA-Seq of single prostate cancer circulating tumor cells reveals recapitulation of gene expression and pathways found in prostate cancer. *PLoS ONE* **7**(11):e49144.

Cantatore P, Attardi G. 1980. Mapping of nascent light and heavy strand transcripts on the physical map of HeLa cell mitochondrial DNA. *Nucleic Acids Res* **8**(12):2605–2625.

Cao Y, Yao Z, Sarkar D, Lawrence M, Sanchez GJ, Parker MH, MacQuarrie KL, Davison J, Morgan MT, Ruzzo WL, Gentleman RC, Tapscott SJ. 2010. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell* **18**(4):662–674.

Cao L, Yu Y, Bilke S, Walker RL, Mayeenuddin LH, Azorsa DO, Yang F, Pineda M, Helman LJ, Meltzer PS. 2010. Genome-wide identification of PAX3-FKHR binding sites in rhabdomyosarcoma reveals candidate target genes important for development and cancer. *Cancer Res* **70**(16):6497–6508.

Capanna E, Manfredi Romanini MG. 1971. Nuclear DNA content and morphology of the karyotype in certain palearctic Microchiroptera. *Caryologia* **24**:471–482.

Capel B, Swain A, Nicolis S, Hacker A, Walter M, Koopman P, Goodfellow P, Lovell-Badge R. 1993. Circular transcripts of the testis-determining gene *Sry* in adult mouse testis. *Cell* **73**(5):1019–1030.

Cardamone MD, Krones A, Tanasa B, Taylor H, Ricci L, Ohgi KA, Glass CK, Rosenfeld MG, Perissi V. 2012. A protective strategy against hyperinflammatory responses requiring the nontranscriptional actions of GPS2. *Mol Cell* **46**(1):91–104.

Carey BW, Markoulaki S, Beard C, Hanna J, Jaenisch R. 2010. Single-gene transgenic mouse strains for reprogramming adult somatic cells. *Nat Methods* **7**(1):56–59.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007a. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol* **7**:192

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007b. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res* **17**:1034-1044.

Carmell M, Girard A, van de Kant H, Bourchis D, Bestor T, de Rooij D, Hannon G. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**:503-514.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). 2005. The transcriptional landscape of the mammalian genome. *Science* **309**(5740):1559–1563.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrm PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**(6):626–635.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**(1):25–36.

Casas F, Rochard P, Rodier A, Cassar-Malek I, Marchal-Victorion S, Wiesner RJ, Cabello G, Wrutniak C. 1999. A variant form of the nuclear triiodothyronine receptor c-ErbAα1 plays a direct role in regulation of mitochondrial RNA synthesis. *Mol Cell Biol* **19**(12):7913–7924.

Casas F, Domenjoud L, Rochard P, Hatier R, Rodier A, Daury L, Bianchi A, Kremarik-Bouillaud P, Becuwe P, Keller J, Schohn H, Wrutniak-Cabello C, Cabello G, Dauça M. 2000. A 45 kDa protein related to PPARγ2, induced by peroxisome proliferators, is located in the mitochondrial matrix. *FEBS Lett* **478**(1-2):4–8.

Cavalier-Smith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J Cell Sci* **34**:247-278.

Cavalier-Smith T. 2002. Nucleomorphs: enslaved algal nuclei. *Current Opinion in Microbiology* **5**(6):612–619.

Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot* **95**(1):147–175.

Cech TR. 1986. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* Cech TR. 1986. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* **44**:207-210

Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH; modENCODE Consortium. 2009. Unlocking the secrets of the genome. *Nature* **459**(7249):927–930.

Ceol CJ, Houvras Y, Jane-Valbuena J, Bilodeau S, Orlando DA, Battisti V, Fritsch L, Lin WM, Hollmann TJ, Ferré F, Bourque C, Burke CJ, Turner L, Uong A, Johnson LA, Beroukhim R, Mermel CH, Loda M, Ait-Si-Ali S, Garraway LA, Young RA, Zon LI. 2011. The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* **471**(7339):513–517.

Ceschin DG, Walia M, Wenk SS, Duboé C, Gaudon C, Xiao Y, Fauquier L, Sankar M, Vandel L, Gronemeyer H. 2011. Methylation specifies distinct estrogen-induced binding site repertoires of CBP to chromatin. *Genes Dev* **25**(11):1132–1146.

Chan SS, Copeland WC. 2009. DNA polymerase gamma and mitochondrial disease: understanding the consequence of *POLG* mutations. *Biochim Biophys Acta* **1787**(5):312–319.

Chang DD, Clayton DA. 1984. Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. *Cell* **36**(3):635–643.

Chang DD, Clayton DA. 1985. Priming of human mitochondrial DNA replication occurs at the light-strand promoter. *Proc Natl Acad Sci U S A* **82**(2):351–355.

Chang DD, Hauswirth WW, Clayton DA. 1985. Replication priming and transcription initiate from precisely the same site in mouse mitochondrial DNA. *EMBO J* **4**(6):1559–1567.

Chang GS, Noegel AA, Mavrich TN, Müller R, Tomsho L, Ward E, Felder M, Jiang C, Eichinger L, Glöckner G, Schuster SC, Pugh BF. 2012. Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in *Dictyostelium*. *Genome Res* **22**(6):1098–1106.

Chang N, Sun C, Gao L, Zhu D, Xu X, Zhu X, Xiong JW, Xi JJ. 2013. Genome editing with RNA-guided Cas9 nuclease in zebrafish embryos. *Cell Res* **23**(4):465–472.

Chang YF, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76:**51–74.

Chang YJ, Land M, Hauser L, Chertkov O, Del Rio TG, Nolan M, Copeland A, Tice H, Cheng JF, Lucas S, Han C, Goodwin L, Pitluck S, Ivanova N, Ovchinikova G, Pati A, Chen A, Palaniappan K, Mavromatis K, Liolios K, Brettin T, Fiebig A, Rohde M, Abt B, Gker M, Detter JC, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP, Lapidus A. 2011. Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium *Ktedonobacter racemifer* type strain (SOSP1-21). *Stand Genomic Sci* **5**(1):97–111.

Chattopadhyay S, Marques JT, Yamashita M, Peters KL, Smith K, Desai A, Williams BR, Sen GC. 2010. Viral apoptosis is induced by IRF-3-mediated activation of Bax. *EMBO J* **29**(10):1762–1773.

Chen C, Ara T, Gautheret D. 2009. Using *Alu* elements as polyadenylation sites: A case of retroposon exaptation. *Mol Biol Evol* **26**(2):327–334.

Chen C, Fenk LA, de Bono M. 2013. Efficient genome editing in *Caenorhabditis elegans* by CRISPR-targeted homologous recombination. *Nucleic Acids Res* **41**(20):e193.

Chen JQ, Delannoy M, Cooke C, Yager JD. 2004. Mitochondrial localization of ERα and ERβ in human MCF7 cells. *Am J Physiol Endocrinol Metab* **286**(6):E1011–1022.

Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol Biol Evol* [Epub ahead of print]

Chen LY, Wei KC, Huang AC, Wang K, Huang CY, Yi D, Tang CY, Galas DJ, Hood LE. 2012. RNASEQR – a streamlined and accurate RNA-seq sequence analysis program. *Nucleic Acids Res* **40**(6):e42

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**(6):1106–1117.

Chen Z, Duan X. 2011. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol Biol* **733**:931-103.

Cheng B, Li T, Rahl PB, Adamson TE, Loudas NB, Guo J, Varzavand K, Cooper JJ, Hu X, Gnatt A, Young RA, Price DH. 2012. Functional association of Gdown1 with RNA polymerase II poised on human genes. *Mol Cell* **45**(1):38–50.

Cheng CY, Vogt A, Mochizuki K, Yao MC. 2010. A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*. *Mol Biol Cell* **21**(10):1753–1762.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**(5725):1149–1154.

Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D, Giardine B, Schuster SC, Miller W, Chiaromonte F, Zhang Y, Blobel GA, Weiss MJ, Hardison RC. 2009. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19**(12):2172–2184.

Chenoweth DM, Dervan PB. 2009. Allosteric modulation of DNA by small molecules. *Proc Natl Acad Sci U S A* **106**(32):13175–13179.

Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at Åprecision. *Nat Biotechnol* **30**(4):344–348.

Chess A. 2012. Mechanisms and consequences of widespread random monoallelic expression. *Nat Rev Genet* **13**(6):421–428.

Chi P, Chen Y, Zhang L, Guo X, Wongvipat J, Shamu T, Fletcher JA, Dewell S, Maki RG, Zheng D, Antonescu CR, Allis CD, Sawyers CL. 2010. ETV1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours. *Nature* **467**(7317):849–853.

Chi SW, Zang JB, Mele A, Darnell RB. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**(7254):479–486.

Chia NY, Chan YS, Feng B, Lu X, Orlov YL, Moreau D, Kumar P, Yang L, Jiang J, Lau MS, Huss M, Soh BS, Kraus P, Li P, Lufkin T, Lim B, Clarke ND, Bard F, Ng HH. 2010. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* **468**(7321):316–320.

Chicas A, Wang X, Zhang C, McCurrach M, Zhao Z, Mert O, Dickins RA, Narita M, Zhang M, Lowe SW. 2010. Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. *Cancer Cell* **17**(4):376–387.

Chiu H, Schwartz HT, Antoshechkin I, Sternberg PW. 2013. Transgene-free genome editing in *Caenorhabditis elegans* using CRISPR-Cas. *Genetics* **195**(3):1167–1171.

Chlon TM, Doré LC, Crispino JD. 2012. Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. *Mol Cell* **47**(4):608–621.

Cho H, Zhao X, Hatori M, Yu RT, Barish GD, Lam MT, Chong LW, DiTacchio L, Atkins AR, Glass CK, Liddle C, Auwerx J, Downes M, Panda S, Evans RM. 2012. Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β. *Nature* **485**(7396):123–127.

Chong JA, Tapia-Ramírez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G. 1995. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**(6):949–957.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**(1):1–8.

Chow MH, Yan KT, Bennett MJ, Wong JT. 2010. Birefringence and DNA condensation of liquid crystalline chromosomes. *Eukaryot Cell* **9**(10):1577–1587.

Christov CP, Gardiner TJ, Szts D, Krude T. 2006. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol Cell Biol* **26**(18):6993–7004.

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. 2011. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44**(4):667–678.

Chu HT, Hsiao WW, Chen JC, Yeh TJ, Tsai MH, Lin H, Liu YW, Lee SA, Chen CC, Tsao TT, Kao CY. 2013. EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. *Bioinformatics* **29**(8):1004-01010.

Chueh FY, Leong KF, Yu CL. 2010. Mitochondrial translocation of signal transducer and activator of transcription 5 (STAT5) in leukemic T cells and cytokine-stimulated cells. *Biochem Biophys Res Commun* **402**(4):778–783.

Chung JH, Whiteley M, Felsenfeld G. 1993. A 5' element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell* **74**(3):505–514.

Chung W-J, Okamura K, Martin R, Lai E. 2008. Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. *Curr Biol* **18**:795-802.

Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**(7330):368–373.

Cingolani G, Capaccio L, D'Elia D, Gadaleta G. 1997. In organelle footprinting analysis of rat mitochondrial DNA: protein interaction upstream of the Ori-L. *Biochem Biophys Res Commun* **231**(3):856–60.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS. 2011. The reality of pervasive transcription. *PLoS Biol* **9**(7):e1000625.

Clark SJ, Harrison J, Paul CL, Frommer M. 1994. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res* **22**(15):2990–2997.

Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J. 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* **40**(4):e29.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**(4):265–270.

Claros MG, Vincens P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* **241**(3):779–786.

Clayton DA. 1982. Replication of animal mitochondrial DNA. *Cell* **28**(4):693–705.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**(5629):71–76.

Clifton SW, Minx P, Fauron CM, Gibson M, Allen JO, Sun H, Thompson M, Barbazuk WB, Kanuganti S, Tayloe C, Meyer L, Wilson RK, Newton KJ. 2004. Sequence and comparative analysis of the maize NB mitochondrial genome. *Plant Physiol* **136**(3):3486–3503.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**(7):613–619.

Cocquerelle C, Daubersies P, Majérus MA, Kerckaert JP, Bailleul B. 1992. Splicing with inverted order of exons occurs proximal to large introns. *EMBO J* **11**(3):1095–1098.

Cocquerelle C, Mascrez B, Hétuin D, Bailleul B. 1993. Mis-splicing yields circular RNA molecules. *FASEB J* **7**(1):155–160.

Cogswell PC, Kashatus DF, Keifer JA, Guttridge DC, Reuther JY, Bristow C, Roy S, Nicholson DW, Baldwin AS Jr. 2003. NFκB and IκBα are found in the mitochondria. Evidence for regulation of mitochondrial gene expression by NFκB. *J Biol Chem* **278**(5):2963–2968.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* **22**:1053-1066.

Collins PJ, Kobayashi Y, Nguyen L, Trinklein ND, Myers RM. 2007. The ets-related transcription factor GABP directs bidirectional transcription. *PLoS Genet* **3**(11):e208.

Comings DE. 1972. The structure and function of chromatin. *Adv Hum Genet* **3**:237–431.

Conboy CM, Spyrou C, Thorne NP, Wade EJ, Barbosa-Morais NL, Wilson MD, Bhattacharjee A, Young RA, Tavaré S, Lees JA, Odom DT. 2007. Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS One* **2**(10):e1061.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**(6121):819–823.

Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, Beresford-Smith B. 2012. Xenome–a tool for classifying reads from xenograft samples. *Bioinformatics* **28**(12):i172–178.

Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, Sidow A. 2004. Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res* **14**(4):539–548.

Cooper TA, Mattox W. 1997. The regulation of splice-site selection, and its role in human disease. *The American Journal of Human Genetics* **61**(2):259–266.

Corbo JC, Lawrence KA, Karlstetter M, Myers CA, Abdelaziz M, Dirkes W, Weigelt K, Seifert M, Benes V, Fritsche LG, Weber BH, Langmann T. 2010. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res* **20**(11):1512–1525.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**(5909):1845–1848.

Cornelison DD, Wold BJ. 1997. Single-cell analysis of regulatory gene expression in quiescent and activated mouse skeletal muscle satellite cells. *Dev Biol* **191**(2):270–283.

Corral-Debrinski M, Shoffner JM, Lott MT, Wallace DC. 1992. Association of mitochondrial DNA damage with aging and coronary atherosclerotic heart disease. *Mutat Res* **275**(3-6):169–180.

Correns C. 1900. G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. *Berichte der deutschen botanischen Gesellschaft* **18**:158-168

Coskun PE, Beal MF, Wallace DC. 2004. Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. *Proc Natl Acad Sci U S A* **101**(29):10726–10731.

Coskun P, Wyrembak J, Schriner SE, Chen HW, Marciniack C, Laferla F, Wallace DC. 2012. A mitochondrial etiology of Alzheimer and Parkinson disease. *Biochim Biophys Acta* **1820**(5):553–564.

Costessi A, Mahrour N, Tijchon E, Stunnenberg R, Stoel MA, Jansen PW, Sela D, Martin-Brown S, Washburn MP, Florens L, Conaway JW, Conaway RC, Stunnenberg HG. 2011. The tumour antigen PRAME is a subunit of a Cul2 ubiquitin ligase and associates with active NFY promoters. *EMBO J* **30**(18):3786–3798.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**(7259):1127–1131.

Covello PS, Gray MW. 1993. On the evolution of RNA editing. *Trends Genet* **9**(8):265–268.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A* **105**(51):20356–20361.

Cox D, Chao A, Baker J, Chang L, Qiao D, Lin H. 1998. A novel class of evolutionarily conserved genes defined by *piwi* are essential for stem cell self-renewal. *Genes Dev* **12**:3715-3727.

Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Brami D, Joardar VS, Johnson J, Radune D, Singh I, Badger JH, Kumar U, Saier M, Wang Y, Cai H, Gu J, Mather MW, Vaidya AB, Wilkes DE, Rajagopalan V, Asai DJ, Pearson CG, Findly RC, Dickerson HW, Wu M, Martens C, Van de Peer Y, Roos DS, Cassidy-Hanley DM, Clark TG. 2011. Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. *Genome Biol* **12**(10):R100.

Cracknell JA, Japrung D, Bayley H. 2013. Translocating kilobase RNA through the Staphylococcal α-hemolysin nanopore. *Nano Lett* **13**(6):2500–2505.

Crews S, Ojala D, Posakony J, Nishiguchi J, Attardi G. 1979. Nucleotide sequence of a region of human mitochondrial DNA containing the precisely identified origin of replication. *Nature* **277**(5693):192–198.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**(50):21931–21936.

Crick FH. 1958. On protein synthesis. *Symp Soc Exp Biol* **12**:138–163.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General nature of the genetic code for proteins. *Nature* **192**:1227–1232.

Crick FH. 1966. The Croonian lecture, 1966. The genetic code. *Proc R Soc Lond B Biol Sci* **167**(9):331–347.

Crick F. 1970. Central Dogma of Molecular Biology. *Nature* **227**:561-563.

Cristea IM, Williams R, Chait BT, Rout MP. 2005. Fluorescent proteins as proteomic probes. *Mol Cell Proteomics* **4**:1933-1941.

Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet* **24**(4):340–341.

Cross SH, Charlton JA, Nan X, Bird AP. 1994. Purification of CpG islands using a methylated DNA binding column. *Nat Genet* **6**(3):236–244.

Csuros M, Rogozin IB, Koonin EV. 2011. A Detailed History of Intron-rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes. *PLoS Comp Bio* **7**(9):e1002150

Cuccurese M, Russo G, Russo A, Pietropaolo C. 2005. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res* **33**(18):5965–5977.

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**(1):24–32.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, Hopkins JF, Kuo A, Rensing SA, Schmutz J, Symeonidi A, Elias M, Eveleigh RJ, Herman EK, Klute MJ, Nakayama T, Oborník M, Reyes-Prieto A, Armbrust EV, Aves SJ, Beiko RG, Coutinho P, Dacks JB, Durnford DG, Fast NM, Green BR, Grisdale CJ, Hempel F, Henrissat B, Hppner MP, Ishida K, Kim E, Korený L, Kroth PG, Liu Y, Malik SB, Maier UG, McRose D, Mock T, Neilson JA, Onodera NT, Poole AM, Pritham EJ, Richards TA, Rocap G, Roy SW, Sarai C, Schaack S, Shirato S, Slamovits CH, Spencer DF, Suzuki S, Worden AZ, Zauner S, Barry K, Bell C, Bharti AK, Crow JA, Grimwood J, Kramer R, Lindquist E, Lucas S, Salamov A, McFadden GI, Lane CE, Keeling PJ, Gray MW, Grigoriev IV, Archibald JM. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**(7427):59–65.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14**(5):942–950.

Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, Gugger M, Lockhart PJ, Allen JF, Brune I, Maus I, Pühler A, Martin WF. 2013. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol Evol* **5**(1):31–44.

Dai L, Zimmerly S. 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA* **9**(1):14–19.

D'Alessio JA, Ng R, Willenbring H, Tjian R. 2011. Core promoter recognition complex changes accompany liver development. *Proc Natl Acad Sci U S A* **108**(10):3906–3911.

Daley T, Smith AD. 2013. Predicting the molecular complexity of sequencing libraries. *Nat Methods* **10**(4):325–327.

Dallagiovanna B, Correa A, Probst CM, Holetz F, Smircich P, de Aguiar AM, Mansur F, da Silva CV, Mortara RA, Garat B, Buck GA, Goldenberg S, Krieger MA. 2008. Functional genomic characterization of mRNAs associated with TcPUF6, a pumilio-like protein from *Trypanosoma cruzi*. *J Biol Chem* **283**(13):8266–8273.

Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, Simpson ML, Weinberger LS. 2012. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc Natl Acad Sci U S A* **109**(43):17454–17459.

Darnell JEJ. 1978. Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science* **202**(4374):1257–1260

Darwin C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. *John Murray, London.*

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* 103:5320-5325

Davidson EH. 2006. The Regulatory Genome: Gene Regulatory Networks In Development And Evolution. *Academic Press*

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan Zj, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H. 2002a. A genomic regulatory network for development. *Science* **295**(5560):1669–1678.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Schilstra MJ, Clarke PJ, Rust AG, Pan Z, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H. 2002b. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev Biol* **246**(1):162–190.

Davis CA, Ares M Jr. 2006. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **103**(9):3262–3267.

Dawkins R. 1986. The Blind Watchmaker. *Norton, New York*.

Dawkins R. 1996. Climbing Mount Improbable. *Norton, New York*.

De Bona F, Ossowski S, Schneeberger K, Rätsch G. 2008. Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**(16):i174-180.

de Bruijn NG. 1946. A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen* **49**:758-764

de Hoon MJ, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. *Bioinformatics* **20**(9):1453–1454.

De Rasmo D, Signorile A, Roca E, Papa S. 2009. cAMP response element-binding protein (CREB) is imported into mitochondria and promotes protein synthesis. *FEBS J* **276**(16):4325–4333.

De Santa F, Narang V, Yap ZH, Tusi BK, Burgold T, Austenaa L, Bucci G, Caganova M, Notarbartolo S, Casola S, Testa G, Sung WK, Wei CL, Natoli G. 2009. Jmjd3 contributes to the control of gene expression in LPS-activated macrophages. *EMBO J* **28**(21):3341–3352.

de Souza SJ, Long M, Gilbert W. 1996 Introns and gene evolution. *Genes Cells* **1**(6):493–505.

de Vries H. 1889. Intracellulare Pangenesis. *Fisher, Jena*

de Vries H. 1900. Sur les unités des caractéres spécifiques et leur application á l'étude des hybrides. *Rev Gen Bot* **12**:257-271.

Deal RB, Henikoff S. 2010. A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Dev Cell* **18**(6):1030–1040.

Deamer DW, Akeson M. 2000. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol* **18**(4):147–151.

Deato MD, Tjian R. 2007. Switching of the core transcription machinery during myogenesis. *Genes Dev* **21**(17):2137–2149.

DeChiara TM, Brosius J. 1987. Neural BC1 RNA: cDNA clones reveal nonrepetitive sequence content. *Proc Natl Acad Sci U S A* **84**(9):2624–2628.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**(5558):1306–1311.

DeKoter RP, Singh H. 2000. Regulation of B lymphocyte and macrophage development by graded expression of PU.1. *Science* **288**:1439-1441.

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci U S A* **103**(23):8751–8756.

Dembski W. 1998. Intelligent Science and Design. *First Things* **86**:21–27

Demonacos C, Tsawdaroglou NC, Djordjevic-Markovic R, Papalopoulou M, Galanopoulos V, Papadogeorgaki S, Sekeris CE. 1993. Import of the glucocorticoid receptor into rat liver mitochondria in vivo and in vitro. *J Steroid Biochem Mol Biol* **46**(3):401–413.

Demonacos C, Djordjevic-Markovic R, Tsawdaroglou N, Sekeris CE. 1995. The mitochondrion as a primary site of action of glucocorticoids: the interaction of the glucocorticoid receptor with mitochondrial DNA sequences showing partial similarity to the nuclear glucocorticoid responsive elements. *J Steroid Biochem Mol Biol* **55**(1):43–55.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J Royal Stat Soc B* **39**(1):1-38.

Deng Q, Ramsköld D, Reinius B, Sandberg R. 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**(6167):193–196.

Deng W, Roberts SG. 2005. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev* **19**(20):2418–2423.

Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, Dennis ES. 2011. FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*. *Proc Natl Acad Sci U S A* **108**(16):6680–6685.

Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. 2008. Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9**(12):R175.

Depledge DP, Evans KJ, Ivens AC, Aziz N, Maroof A, Kaye PM, Smith DF. 2009. Comparative expression profiling of *Leishmania*: modulation in gene expression between species and in different host genetic backgrounds. *PLoS Negl Trop Dis* **3**(7):e476.

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piégu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci U S A* **103**(31):11647–11652.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**(9):1775–1789.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**(6):1173–1183.

Dervan PB, Edelson BS. 2003. Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr Opin Struct Biol* **13**(3):284–299.

Deschamps P, Lara E, Marande W, López-García P, Ekelund F, Moreira D. 2011. Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. *Mol Biol Evol* **28**(1):53–58.

Desjardins P, Frost E, Morais R. 1985. Ethidium bromide-induced loss of mitochondrial DNA from primary chicken embryo fibroblasts. *Mol Cell Biol* **5**(5):1163–1169.

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**(6):853–866.

Diaz A, Nellore A, Song JS. 2012. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biol* **13**(10):R98.

Diaz A, Park K, Lim DA, Song JS. 2012. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol Biol* **11**(3):Article 9.

DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. 2013. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res* **41**(7):4336–4343.

Dickinson DJ, Ward JD, Reiner DJ, Goldstein B. 2013. Engineering the *Caenorhabditis elegans* genome using Cas9-triggered homologous recombination. *Nat Methods* **10**(10):1028–1034.

Dimon MT, Sorber K, DeRisi JL. 2010. HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One* **5**(11):e13875.

Dinger ME, Amaral PP, Mercer TR, Mattick JS. 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief Funct Genomic Proteomic* **8**(6):407–423.

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C, Sunkin SM, Crowe ML, Grimmond SM, Perkins AC, Mattick JS. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18**(9):1433–1445.

Diribarne G, Bensaude O. 2009. 7SK RNA, a non-coding RNA regulating P-TEFb, a general transcription factor. *RNA Biol* **6**(2):122–128.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**:376-380.

Dixon RJ, Eperon IC, Hall L, Samani NJ. 2005. A genome-wide survey demonstrates widespread non-linear mRNA in expressed sequences from multiple species. *Nucleic Acids Res* **33**(18):5904–5913.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. 2012. Landscape of transcription in human cells. *Nature* **489**(7414):101–108.

Djuranovic S, Zinchenko M, Hur J, Nahvi A, Brunelle J, Rogers E, Green R. 2010. Allosteric regulation of Argonaute proteins by miRNAs. *Nat Struct Mol Biol* **17**:144-150.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1):15–21.

Dobzhansky T. 1973. Nothing in Biology Makes Sense Except in the Light of Evolution. *American Biology Teacher* **35**:125-129

Dodge JD. 1965. Chromosome structure in the dinoflagellates and the problem of the mesokaryotic cell. *Expcerpta Med Int Congr Ser* **91**:339-345

Dodge JD, Greuet C. 1987. Dinoflagellate ultrastructure and complex organelles. In Taylor FJR (Ed.) The Biology of Dinoflagellates. *Blackwell, Oxford*, pp. 93-142.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**(16):e105.

Doolittle WF. 1978. Genes in pieces: Were they ever together? *Nature* **272**:581–582

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A* **110**(14):5294–5300

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**(5757):601–603.

Doré LC, Chlon TM, Brown CD, White KP, Crispino JD. 2012. Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood* **119**(16):3724–3733.

Dorrell RG, Drew J, Nisbet RE, Howe CJ. 2014. Evolution of chloroplast transcript processing in *Plasmodium* and its chromerid algal relatives. *PLoS Genet* **10**(1):e1004008.

Douglas AGL, Wood MJA. 2011. RNA splicing: disease and therapy. *Briefings in Functional Genomics* **10**(3):151–164.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* **410**(6832):1091–1096.

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Grf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**(7):779–785.

Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de Doncker S, Maes I, Mottram JC, Quail MA, Rijal S, Sanders M, Schnian G, Stark O, Sundar S, Vanaerschot M, Hertz-Fowler C, Dujardin JC, Berriman M. 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* **21**(12):2143–2156.

Drewe P, Stegle O, Hartmann L, Kahles A, Bohnert R, Wachter A, Borgwardt K, Rätsch G. 2013. Accurate detection of differential RNA processing. *Nucleic Acids Res* **41**(10):5189–5198.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak SA, Bradley RK, Brand AD, Brent MR, Brooks AN, Brown RH, Butlin RK, Caggese C, Calvi BR, Bernardo de Carvalho A, Caspi A, Castrezana S, Celniker SE, Chang JL, Chapple C, Chatterji S, Chinwalla A, Civetta A, Clifton SW, Comeron JM, Costello JC, Coyne JA, Daub J, David RG, Delcher AL, Delehaunty K, Do CB, Ebling H, Edwards K, Eickbush T, Evans JD, Filipski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia AC, Gardiner A, Garfield DA, Garvin BE, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg AJ, Griffiths-Jones S, Gross S, Guigo R, Gustafson EA, Haerty W, Hahn MW, Halligan DL, Halpern AL, Halter GM, Han MV, Heger A, Hillier L, Hinrichs AS, Holmes I, Hoskins RA, Hubisz MJ, Hultmark D, Huntley MA, Jaffe DB, Jagadeeshan S, Jeck WR, Johnson J, Jones CD, Jordan WC, Karpen GH, Kataoka E, Keightley PD, Kheradpour P, Kirkness EF, Koerich LB, Kristiansen K, Kudrna D, Kulathinal RJ, Kumar S, Kwok R, Lander E, Langley CH, Lapoint R, Lazzaro BP, Lee SJ, Levesque L, Li R, Lin CF, Lin MF, Lindblad-Toh K, Llopart A, Long M, Low L, Lozovsky E, Lu J, Luo M, Machado CA, Makalowski W, Marzo M, Matsuda M, Matzkin L, McAllister B, McBride CS, McKernan B, McKernan K, Mendez-Lago M, Minx P, Mollenhauer MU, Montooth K, Mount SM, Mu X, Myers E, Negre B, Newfeld S, Nielsen R, Noor MA, O'Grady P, Pachter L, Papaceit M, Parisi MJ, Parisi M, Parts L, Pedersen JS, Pesole G, Phillippy AM, Ponting CP, Pop M, Porcelli D, Powell JR, Prohaska S, Pruitt K, Puig M, Quesneville H, Ram KR, Rand D, Rasmussen MD, Reed LK, Reenan R, Reily A, Remington KA, Rieger TT, Ritchie MG, Robin C, Rogers YH, Rohde C, Rozas J, Rubenfield MJ, Ruiz A, Russo S, Salzberg SL, Sanchez-Gracia A, Saranga DJ, Sato H, Schaeffer SW, Schatz MC, Schlenke T, Schwartz R, Segarra C, Singh RS, Sirot L, Sirota M, Sisneros NB, Smith CD, Smith TF, Spieth J, Stage DE, Stark A, Stephan W, Strausberg RL, Strempel S, Sturgill D, Sutton G, Sutton GG, Tao W, Teichmann S, Tobari YN, Tomimura Y, Tsolas JM, Valente VL, Venter E, Venter JC, Vicario S, Vieira FG, Vilella AJ, Villasante A, Walenz B, Wang J, Wasserman M, Watts T, Wilson D, Wilson RK, Wing RA, Wolfner MF, Wong A, Wong GK, Wu CI, Wu G, Yamamoto D, Yang HP, Yang SP, Yorke JA, Yoshida K, Zdobnov E, Zhang P, Zhang Y, Zimin AV, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley CD, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin CR, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hafez N, Hagos B, Hall J, Henson C, Hollinger A, Honan T, Huard MD, Hughes L, Hurhula B, Husby ME, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lance K, Lara M, Lee W, Lennon N, Letendre F, LeVine R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lokyitsang Y, Lubonja R, Lui A, MacDonald P, Magnisalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai O, Mihalev A, Mihova T, Mittelman R, Mlenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, Novod N, O'Neill B, Osman S, Markiewicz E, Oyono OL, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settipalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Thoulutsang Y, Thoulutsang D, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Young G, Yu Q, Zembek L, Zhong D, Zimmer A, Zwirko Z, Jaffe DB, Alvarez P, Brockman W, Butler J, Chin C, Gnerre S, Grabherr M, Kleber M, Mauceli E, MacCallum I. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**(7167):203–218.

Du J, Leng J, Habegger L, Sboner A, McDermott D, Gerstein M. 2012. IQSeq: integrated isoform quantification analysis based on next-generation sequencing. *PLoS One* **7**(1):e29175.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A, Hetzel J, Wohlschlegel JA, Pradhan S, Patel DJ, Jacobsen SE. 2012. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**(1):167–180.

du Buy H, Riley F. 1967. Hybridization between the nuclear and kinetoplast DNA's of *Leishmania enriettii* and between nuclear and mitochondrial DNA's of mouse liver. *Proc Natl Acad Sci U S A* **57**(3):790–797.

Durant L, Watford WT, Ramos HL, Laurence A, Vahedi G, Wei L, Takahashi H, Sun HW, Kanno Y, Powrie F, O'Shea JJ. 2010. Diverse targets of the transcription factor STAT3 contribute to T cell pathogenicity and homeostasis. *Immunity* **32**(5):605–615.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**:1653-1655.

Dutrow N, Nix DA, Holt D, Milash B, Dalley B, Westbroek E, Parnell TJ, Cairns BR. 2008. Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA-DNA hybrid mapping. *Nat Genet* **40**(8):977–986.

Ebert A, McManus S, Tagoh H, Medvedovic J, Salvagiotto G, Novatchkova M, Tamir I, Sommer A, Jaritz M, Busslinger M. 2011. The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity* **34**(2):175–187.

Eckner R, Ewen ME, Newsome D, Gerdes M, DeCaprio JA, Lawrence JB, Livingston DM. 1994. Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev* **8**(8):869–884.

Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. *Curr Biol* **22**(21):R898–899.

Eddy SR. 2013. The ENCODE project: missteps overshadowing a success. *Curr Biol* **23**(7):R259–261.

Eden FC, Hendrick JP, Gottlieb SS. 1978. Homology of single copy and repeated sequences in chicken, duck, Japanese quail, and ostrich DNA. *Biochemistry* **17**:5113–5121.

Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, Cheung MS, Day DS, Gadel S, Gorchakov AA, Gu T, Kharchenko PV, Kuan S, Latorre I, Linder-Basso D, Luu Y, Ngo Q, Perry M, Rechtsteiner A, Riddle NC, Schwartz YB, Shanower GA, Vielle A, Ahringer J, Elgin SC, Kuroda MI, Pirrotta V, Ren B, Strome S, Park PJ, Karpen GH, Hawkins RD, Lieb JD. 2011. An assessment of histone-modification antibody quality. *Nat Struct Mol Biol* **18**(1):91–93.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910):133–138.

Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK Jr, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol* **4**(9):e286.

Ekstrand MI, Falkenberg M, Rantanen A, Park CB, Gaspari M, Hultenby K, Rustin P, Gustafsson CM, Larsson NG. 2004. Mitochondrial transcription factor A regulates mtDNA copy number in mammals. *Hum Mol Genet* **13**(9):935–944.

Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. 2001. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**(6836):494–498.

Elmendorf HG, Singer SM, Nash TE. 2001. The abundance of sterile transcripts in *Giardia lamblia*. *Nucleic Acids Res* **29**(22):4674–4683.

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic gene expression in a single cell. *Science* **297**(5584):1183–1186.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B. 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**(5733):409–415.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* **2**(4):953–971.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* **440**(7084):623–630.

Embley TM, van der Giezen M, Horner DS, Dyal PL, Bell S, Foster PG. 2003. Hydrogenosomes, mitochondria and early eukaryotic evolution. *IUBMB Life* **55**(7):387–395.

Emerson BM, Lewis CD, Felsenfeld G. 1985. Interaction of specific nuclear factors with the nuclease-hypersensitive region of the chicken adult β-globin gene: Nature of the binding domain. *Cell* **41**(1):21–30

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**(5696):636–640.

ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146):799–816.

ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**(4):e1001046.

ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kelllis M, Khatun J, Kheradpour P, Kundaje A, Lassman T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SC, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LA, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elinitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Starnatoyannopoulos JA, Tennebaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee BK, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang

ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge EC, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry JS, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SC, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanan E, Tress ML, van Baren MJ, Walters N, Washieti S, Wilming L, Zadissa A, Zhengdong Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Raymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyenger S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Larnarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenebaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, White KP, Auer T, Centarin L, Eichenlaub M, Gruhl F, Heerman S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin TM, Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JA, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AO, Kharchenko PV, Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan KK, Yip KY, Birney E. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A; RGASP Consortium, Alioto T, Behr J, Bertone P, Bohnert R, Campagna D, Davis CA, Dobin A, Engström PG, Gingeras TR, Goldman N, Grant GR, Guigó R, Harrow J, Hubbard TJ, Jean G, Kahles A, Kosarev P, Li S, Liu J, Mason CE, Molodtsov V, Ning Z, Ponstingl H, Prins JF, Rätsch G, Ribeca P, Seledtsov I, Sipos B, Solovyev V, Steijger T, Valle G, Vitulo N, Wang K, Wu TD, Zeller G, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, Bertone P. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10**(12):1185–1191.

Enríquez JA, Fernández-Sílva P, Montoya J. 1999a. Autonomous regulation in mammalian mitochondrial DNA transcription. *Biol Chem* **380**(7-8):737–47.

Enríquez JA, Fernández-Sílva P, Garrido-Pérez N, López-Pérez MJ, Pérez-Martos A, Montoya J. 1999b, Direct regulation of mitochondrial RNA synthesis by thyroid hormone. *Mol Cell Biol* **19**(1):657–670.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**(8):817–825.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**(3):215–216.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**(7345):43–49.

Espinoza CA, Allen TA, Hieb AR, Kugel JF, Goodrich JA. 2004. *B2* RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat Struct Mol Biol* **11**(9):822–829.

Estévez AM. 2008. The RNA-binding protein TbDRBD3 regulates the stability of a specific subset of mRNAs in trypanosomes. *Nucleic Acids Res* **36**(14):4573–4586.

Esumi S, Kakazu N, Taguchi Y, Hirayama T, Sasaki A, Hirabayashi T, Koide T, Kitsukawa T, Hamada S, Yagi T. 2005. Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. *Nat Genet* **37**(2):171–176.

Eveland AL, Goldshmidt A, Pautler M, Morohashi K, Liseron-Monfils C, Lewis MW, Kumari S, Hiraga S, Yang F, Unger-Wallace E, Olson A, Hake S, Vollbrecht E, Grotewold E, Ware D, Jackson D. 2014. Regulatory modules controlling maize inflorescence architecture. *Genome Res* **24**(3):431–443.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A, Park PJ, Walsh CA. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**(3):483–496.

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ, Manning G, Lang BF, Haas B, Nusbaum C, King N. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol* **14**(2):R15.

Fan R, Bonde S, Gao P, Sotomayor B, Chen C, Mouw T, Zavazava N, Tan K. 2012. Dynamic HoxB4-regulatory network during embryonic stem cell differentiation to hematopoietic cells. *Blood* **119**(19):e139–147.

Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. 2012. Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. *Cell* **151**:1243-1255.

Fang X, Yoon JG, Li L, Yu W, Shao J, Hua D, Zheng S, Hood L, Goodlett DR, Foltz G, Lin B. 2011. The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. *BMC Genomics* **12**:11.

Falkenberg M, Gaspari M, Rantanen A, Trifunovic A, Larsson NG, Gustafsson CM. 2002. Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. *Nat Genet* **31**(3):289–294.

Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. 2012. Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. *Cell* **151**:1243-1255.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among aminal, plant, and fungal genes. *Proc Natl Acad Sci U S A.* **99:**16128–16133.

Fedorov A, Roy S, Cao X, Gilbert W. 2003. Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. *Genome Res* **13**(6A):1155–1157.

Fejes A, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones S 2008. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**:1729–1730.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, Gingeras TR; Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**(7232):1028–1032.

Felle M, Hoffmeister H, Rothammer J, Fuchs A, Exler JH, Längst G. 2011. Nucleosomes protect DNA from DNA methylation in vivo and in vitro. *Nucleic Acids Res* **39**(16):6956–6969.

Femino AM, Fay FS, Fogarty K, Singer RH. 1998. Visualization of single RNA transcripts in situ. *Science* **280**(5363):585–590.

Feng D, Liu T, Sun Z, Bugge A, Mullican SE, Alenghat T, Liu XS, Lazar MA. 2011. A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism. *Science* **331**(6022):1315–1319.

Feng J, Li W, Jiang T. 2010. Inference of Isoforms from Short Sequence Reads. In Research in Computational Molecular Biology. *Lecture Notes in Computer Science* **6044**:138-157

Feng J, Li W, Jiang T. 2011. Inference of isoforms from short sequence reads. *J Comput Biol* **18**(3):305–321.

Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**(9):1728–1740.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**(19):8689–8694.

Feng X, Grossman R, Stein L. 2011. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics* **12**:139.

Fernandez PC, Frank SR, Wang L, Schroeder M, Liu S, Greene J, Cocito A, Amati B. 2003. Genomic targets of the human c-Myc protein. *Genes Dev* **17**:1115-1129.

Fernandez-Silva P, Martinez-Azorin F, Micol V, Attardi G. 1997. The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions. *EMBO J* **16**(5):1066–1079.

Fillingham JS, Thing TA, Vythilingum N, Keuroghlian A, Bruno D, Golding GB, Pearlman RE. 2004. A non-long terminal repeat retrotransposon family is restricted to the germ line micronucleus of the ciliated protozoan *Tetrahymena thermophila*. *Eukaryot Cell* **3**(1):157–169.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. *Nature* **481**(7381):360–364.

Finster S, Eggert E, Zoschke R, Weihe A, Schmitz-Linneweber C. 2013. Light-dependent, plastome-wide association of the plastid-encoded RNA polymerase with chloroplast DNA. *Plant J* **76**(5):849–860.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**(6669):806–811.

Fish J, Raule N, Attardi G. 2004. Discovery of a major D-loop replication origin reveals two modes of human mtDNA synthesis. *Science* **306**(5704):2098–2101.

Fisher RA. 1930. The Genetical Theory of Natural Selection. *Clarendon, Oxford.*

Fisher RP, Clayton DA. 1985. A transcription factor required for promoter recognition by human mitochondrial RNA polymerase. Accurate initiation at the heavy- and light-strand promoters dissected and reconstituted in vitro. *J Biol Chem* **260**(20):11330–11338.

Fisher RP, Clayton DA. 1988. Purification and characterization of human mitochondrial transcription factor 1. *Mol Cell Biol* **8**(8):3496–3509.

Fisher RP, Lisowsky T, Parisi MA, Clayton DA. 1992. DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. *J Biol Chem* **267**(5):3358–3367.

Fisher RP, Parisi MA, Clayton DA. 1989. Flexible recognition of rapidly evolving promoter sequences by mitochondrial transcription factor 1. *Genes Dev* **3**(12b):2202–2217.

Fisher RP, Topper JN, and Clayton DA. 1987. Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements. *Cell* **50**(2):247–258.

Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, Bickel PJ, Biggin MD, Celniker SE. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proc Natl Acad Sci U S A* **109**(52):21330–21335.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**(6):461–465.

Fong AP, Yao Z, Zhong JW, Cao Y, Ruzzo WL, Gentleman RC, Tapscott SJ. 2012. Genetic and epigenetic determinants of neurogenesis and myogenesis. *Dev Cell* **22**(4):721–735.

Force A, Cresko W, Pickett FB, Proulx S, Amemiya C, Lynch M. 2005. The origin of gene subfunctions and modular gene regulation. *Genetics* **170**:433-446

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531-1545

Fornace AJ Jr, Mitchell JB. 1986. Induction of *B2* RNA polymerase III transcription by heat shock: enrichment for heat shock induced sequences in rodent cells by hybridization subtraction. *Nucleic Acids Res* **14**(14):5793–5811.

Forrester WC, Thompson C, Elder JT, Groudine M. 1986. A developmentally stable chromatin structure in the human β-globin gene cluster. *Proc Natl Acad Sci U S A* **83**:1359-1363.

Forster SC, Finkel AM, Gould JA, Hertzog PJ. 2013. RNA-eXpress annotates novel transcript features in RNA-seq data. *Bioinformatics* **29**(6):810–812.

Fortschegger K, de Graaf P, Outchkourov NS, van Schaik FM, Timmers HT, Shiekhattar R. 2010. PHF8 targets histone methylation and RNA polymerase II to activate transcription. *Mol Cell Biol* **30**(13):3286–3298.

Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, Fu XD. 2011. A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation. *Genomics* **98**(4):266-71.

Freitas MA, Sklenar AR, Parthun MR. 2004. Application of mass spectrometry to the identification and quantification of histone post-translational modifications. *J Cell Biochem* **92**(4):691–700.

Friedland AE, Tzur YB, Esvelt KM, Colaiácovo MP, Church GM, Calarco JA. 2013. Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat Methods* **10**(8):741–743.

Friedman R, Farh K, Burge C, Bartel D. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**:92-105.

Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ. 2010. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. *PLoS ONE* **5**:e15082.

Frith MC, Pheasant M, Mattick JS. 2005. The amazing complexity of the human transcriptome. *Eur J Hum Genet* **13**(8):894–897.

Fritsch EF, Lawn RM, Maniatis T. 1980. Molecular cloning and characterization of the human β-like globin gene cluster. *Cell* **19**(4):959–972.

Friz CT. 1968. The biochemical composition of the free-living amoebae *Chaos chaos*, *Amoeba dubia* and *Amoeba proteus. Comp Biochem Physiol* **26**(1):81–90.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**(5):1827–1831.

Fujii S, Toda T, Kikuchi S, Suzuki R, Yokoyama K, Tsuchida H, Yano K, Toriyama K. 2011. Transcriptome map of plant mitochondria reveals islands of unexpected transcribed regions. *BMC Genomics* **12**:279.

Fuks F. 2005. DNA methylation and histone modifications: teaming up to silence genes. *Curr Opin Genet Dev* **15**(5):490-495.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**(7269):58–64.

Fullwood MJ, Wei CL, Liu ET, Ruan Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**(4):521–532.

Fuhrmann G, Swart E, Nowacki M, Lipps HJ. 2013. RNA-dependent genome processing during nuclear differentiation: the model systems of stichotrichous ciliates. *Epigenomics* **5**(2):229–236.

Gadaleta G, D'Elia D, Capaccio L, Saccone C, Pepe G. 1996. Isolation of a 25-kDa protein binding to a curved DNA upstream the origin of the L strand replication in the rat mitochondrial genome. *J Biol Chem* **271**(23):13537–13541.

Galas DJ, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**(9):3157–3170.

Galiana-Arnoux D, Dostert C, Schneemann A, Hoffmann J, Imler J-L. 2006. Essential function in vivo for Dicer-2 in host defense against RNA viruses in *Drosophila*. *Nat Immunol* **7**:590-597.

Gall JG. 1981. Chromosome structure and the C-value paradox. *J Cell Biol* **91**:3s–14s.

Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Köhler N. 1999. Discovery of tetraploidy in a mammal. *Nature* **401**(6751):341.

Gao Z, Zhang J, Bonasio R, Strino F, Sawai A, Parisi F, Kluger Y, Reinberg D. 2012. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol Cell* **45**(3):344–356.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**(6):469–477.

Garber M, Yosef N, Goren A, Raychowdhury R, Thielke A, Guttman M, Robinson J, Minie B, Chevrier N, Itzhaki Z, Blecher-Gonen R, Bornstein C, Amann-Zalcenstein D, Weiner A, Friedrich D, Meldrim J, Ram O, Cheng C, Gnirke A, Fisher S, Friedman N, Wong B, Bernstein BE, Nusbaum C, Hacohen N, Regev A, Amit I. 2012. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol Cell* **47**(5):810–822.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**(6906):498–511.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* **196**(2):261–282.

Gaspari M, Falkenberg M, Larsson NG, Gustafsson CM. 2004. The mitochondrial RNA polymerase contributes critically to promoter specificity in mammalian cells. *EMBO J* **23**(23):4606–4614.

Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, Berney T, Montanya E, Mohlke KL, Lieb JD, Ferrer J. 2010. A map of open chromatin in human pancreatic islets. *Nat Genet* **42**(3):255–259.

Gauthier BR, Wiederkehr A, Baquié M, Dai C, Powers AC, Kerr-Conte J, Pattou F, MacDonald RJ, Ferrer J, Wollheim CB. 2009. PDX1 deficiency causes mitochondrial dysfunction and defective insulin secretion through TFAM suppression. *Cell Metabolism* **10**(2):110–118.

Ge DT, Zamore PD. 2013. Small RNA-directed silencing: the fly finds its inner fission yeast? *Curr Biol* **23**(8):R318–320.

Germain PL, Ratti E, Boem F. 2014. Junk or functional DNA? ENCODE and the function controversy. *Biology & Philosophy* March 1-25

German MA, Luo S, Schroth G, Meyers BC, Green PJ. 2009. Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc* **4**(3):356–362.

German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K, German R, De Paoli E, Lu C, Schroth G, Meyers BC, Green PJ. 2008. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* **26**(8):941–946.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414):91–100.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dosé AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM 3rd, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Rätsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X; modENCODE Consortium, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**(6012):1775–1787.

Gertz J, Reddy TE, Varley KE, Garabedian MJ, Myers RM. 2012. Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome Res* **22**(11):2153–2162.

Gertz J, Varley KE, Davis NS, Baas BJ, Goryshin IY, Vaidyanathan R, Kuersten S, Myers RM. 2012. Transposase mediated construction of RNA-seq libraries. *Genome Res* **22**(1):134–141.

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler EL, Zapp ML, Weng Z, Zamore PD. 2008. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* **320**(5879):1077–1081.

Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei CL, Ragoussis J, Natoli G. 2010. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32**(3):317–328.

Ghivizzani SC, Madsen CS, Nelen MR, Ammini CV, Hauswirth WW. 1994. In organello footprint analysis of human mitochondrial DNA: human mitochondrial transcription factor A interactions at the origin of replication. *Mol Cell Biol* **14**(12):7717–7730.

Gibbs SP. 1978. The chloroplasts of Euglena may have evolved from symbiotic green algae. *Can J Bot* **56**:2883-2889

Gilbert W. 1978. Why genes in pieces? *Nature* **271**(5645):501.

Gilbert W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **52**:901–905.

Gilbert W, de Souza SJ, Long M. 1997 Origin of genes. *Proc Natl Acad Sci U S A* **94**(15):7698–7703.

Gilbert W, Maxam A. 1973. The nucleotide sequence of the lac operator. *Proc Natl Acad Sci U S A* **70**(12):3581-3584

Gillies SD, Morrison SL, Oi VT, Tonegawa S. 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**(3):717–728.

Gilmour DS, Lis JT. 1984. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc Natl Acad Sci U S A* **81**(14):4275–4279.

Gilmour DS, Lis JT. 1985. In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Mol Cell Biol* **5**(8):2009–2018.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci U S A* **103**(25):9566–9571.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318**(5853):1136–1140.

Gingeras TR. 2009. The pervasive and interleaved transcriptome. *Nat Rev Genet* **10**(11)

Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**(7099):199–202.

Giulietti M, Piva F, D'Antonio M, D'Onorio De Meo P, Paoletti D, Castrignanó T, 'Erchia AM, Picardi E, Zambelli F, Principato G, Pavesi G, Pesole G. 2013. SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res* **41**(D1):D125–131.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**(4):1513–1518.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. *Science.* **274**(5287):546,563–567.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: a landscape takes shape. *Cell* **128**(4):635–638.

Gommers-Ampt JH, Van Leeuwen F, de Beer AL, Vliegenthart JF, Dizdaroglu M, Kowalak JA, Crain PF, Borst P. 1993. β-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan *T. brucei. Cell* **75**(6):1129–1136.

Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* **14**(7):R70.

Goodrich JA, Tjian R. 2010. Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nat Rev Genet* **11**(8):549–558.

Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr Biol* **22**(24):2303–2312.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**(5):565–577.

Gough DJ, Corlett A, Schlessinger K, Wegrzyn J, Larner AC, Levy DE. 2009. Mitochondrial STAT3 supports Ras-dependent oncogenic transformation. *Science* **324**(5935):1713–1716.

Gould SJ, Lewontin RC. 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist program. *Proc Royal Soc London B* **205**:581-598.

Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. 2013. Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci U S A* **110**(46):18584–18589.

Gowher H, Stockdale CJ, Goyal R, Ferreira H, Owen-Hughes T, Jeltsch A. 2005. De novo methylation of nucleosomal DNA by the mammalian Dnmt1 and Dnmt3A DNA methyltransferases. *Biochemistry* **44**(29):9899–9904.

Gowher H, Brick K, Camerini-Otero RD, Felsenfeld G. 2012. Vezf1 protein binding sites genome-wide are associated with pausing of elongating RNA polymerase II. *Proc Natl Acad Sci U S A* **109**(7):2370–2375.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**(7):644–652.

Grabowski P. 2011. Alternative splicing takes shape during neuronal development. *Curr Opin Genet Dev* **21**(4):388–394.

Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell* **11**:1253-1263.

Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**(18):2518–2528.

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5**(3):578–590

Graveley BR. 2002. Sex, AGility, and the regulation of alternative splicing. *Cell* **109**(4):409–412.

Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, Wildonger J, O'Connor-Giles KM. 2013. Genome engineering of *Drosophila* with the CRISPR RNA-guided Cas9 nuclease. *Genetics* **194**(4):1029–1035.

Gray MW. 2012. Mitochondrial evolution. *Cold Spring Harb Perspect Biol* **4**(9):a011403.

Gray MW, Lang BF, Burger G. 2004. Mitochondria of protists. *Annu Rev Genet* **38**:477–524.

Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. 2010. Cell Biology. Irremediable Complexity? *Science* **330**:920-921.

Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbic V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, Hernández-Crespo P, Diaz I, Martinez M, Navajas M, Sucena É, Magalhães S, Nagy L, Pace RM, Djuranovic S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi SV, Zeng J, Pires-daSilva A, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Gregory TR, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van de Peer Y. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **479**(7374):487–92.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. *Plant J* **66**(1):34–44.

Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC, Brenner SE. 2003. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* **19 Suppl 1**:i118–121.

Gregis V, Andrés F, Sessa A, Guerra RF, Simonini S, Mateos JL, Torti S, Zambelli F, Prazzoli GM, Bjerkan KN, Grini PE, Pavesi G, Colombo L, Coupland G, Kater MM. 2013. Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*. *Genome Biol* **14**(6):R56.

Gregory TR. 2007. The onion test. April 25th, 2007; `http://www.genomicron.evolverzone.com/2007/04/onion-test/`

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. *Nucleic Acids Res* **35**(Database issue):D332–8.

Greider CW, Blackburn EH. 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell* **43**(2 Pt 1):405–413.

Greider CW, Blackburn EH. 1987. The telomere terminal transferase of *Tetrahymena* is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell* **51**(6):887–898.

Greider CW, Blackburn EH. 1989. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. *Nature* **337**(6205):331–337.

Greilhuber J, Borsch T, Mller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol (Stuttg)* **8**(6):770–777.

Grewal S, Jia S. 2007. Heterochromatin revisited. *Nat Rev Genet* **8**:35-46.

Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. 2010. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat Rev Microbiol* **8**(10):743–752.

Grivna ST, Beyret E, Wang Z, Lin H. 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**(13):1709–1714.

Grossmann JG. 2013. Engineering without an engineer, or shedding new light on feats in the junkyard of DNA. A review of *Evolution: A View from the 21st Century* by James A. Shapiro. *J Creation* **27**:42–45

Grosveld F, van Assendelft GB, Greaves DR, Kollias G. 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* **51**:975-985.

Grote P, Wittler L, Hendrix D, Koch F, Währisch S, Beisaw A, Macura K, Blss G, Kellis M, Werber M, Herrmann BG. 2013. The tissue-specific lncRNA *Fendrr* is an essential regulator of heart and body wall development in the mouse. *Dev Cell* **24**(2):206–214.

Grunstein M. 1990. Histone function in transcription. *Annu Rev Cell Biol* **6**:643–678.

Gu F, Hsu HK, Hsu PY, Wu J, Ma Y, Parvin J, Huang TH, Jin VX. 2010. Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Syst Biol* **4**:170.

Gu LQ, Wang Y. 2013. Nanopore single-molecule detection of circulating microRNAs. *Methods Mol Biol* **1024**:255–268.

Gu LQ, Wanunu M, Wang MX, McReynolds L, Wang Y. 2012. Detection of miRNAs with a nanopore single-molecule counter. *Expert Rev Mol Diagn* **12**(6):573–584.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**(1):77–88.

Gunawardane L, Saito K, Nishida K, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi M. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* **315**:1587-1590.

Guo JU, Su Y, Zhong C, Ming GL, Song H. 2011. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**(3):423-434.

Guo Y, Mahony S, Gifford DK. 2012. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**(8):e1002638.

Guo Y, Papachristoudis G, Altshuler RC, Gerber GK, Jaakkola TS, Gifford DK, Mahony S. 2010. Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26**(24):3028–3034.

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY. 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**(7291):1071–1076.

Gupta SK, Carmi S, Waldman Ben-Asher H, Tkacz ID, Naboishchikov I, Michaeli S. 2013. Basal splicing factors regulate the stability of mature mRNAs in trypanosomes. *J Biol Chem* **288**(7):4991–5006.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235):223–227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**(7364):295–300.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**(5):503–510.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**(7385):339–346.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**(5):503–510.

Guzzardo PM, Muerdter F, Hannon GJ. 2013. The piRNA pathway in flies: highlights and future directions. *Curr Opin Genet Dev* **23**(1):44–52.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235):223–227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**(7364):295–300.

Ha M, Ng DW, Li WH, Chen ZJ. 2011. Coordinated histone modifications are associated with gene expression variation within and between species. *Genome Res* **21**(4):590–598.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**(8):1494–1512.

Haase A, Fenoglio S, Muerdter F, Guzzardo P, Czech B, Pappin D, Chen C, Gordon A, Hannon G. 2010. Probing the initiation and effector phases of the somatic piRNA pathway in *Drosophila*. *Genes Dev* **24**:2499-2504.

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, Morse M, Engreitz J, Lander ES, Guttman M, Lodish HF, Flavell R, Raj A, Rinn JL. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA *Firre*. *Nat Struct Mol Biol* **21**(2):198–206.

Hackett JD, Anderson DM, Erdner DL, Bhattacharya D. 2004. Dinoflagellates: a remarkable evolutionary experiment. *Am J Bot* **91**:1523-1534.

Hackett JD, Scheetz TE, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Bhattacharya D. 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics* **6**:80.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**(1):129–141.

Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, Ojo T, Luo S, Schroth G, Tuschl T. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**(9):1697–1712.

Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. 2013. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**(8):1210–1223.

Hahn MW, Wray GA. 2002. The *g*-value paradox. *Evol Dev* **4**(2):73–75.

Haldane JBS. 1932. The Causes of Evolution. *Longmans Green, New York.*

Hall SL, Padgett RA. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol* **239**(3):357–365.

Hall SL, Padgett RA. 1996. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science* **271**(5256):1716–1718.

Hall SS. 2012a Sep 18. Hidden Treasures in Junk DNA. *Scientific American*

Hall SS. 2012b. Journey to the genetic interior. *Sci Am* **307**:80-84.

Hamilton AJ, Baulcombe DC. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**(5441):950–952.

Han J, Yuan P, Yang H, Zhang J, Soh BS, Li P, Lim SL, Cao S, Tay J, Orlov YL, Lufkin T, Ng HH, Tam WL, Lim B. 2010. Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature* **463**(7284):1096–1100.

Handler D, Olivieri D, Novatchkova M, Gruber F, Meixner K, Mechtler K, Stark A, Sachidanandam R, Brennecke J. 2011. A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *EMBO J* **30**:3977-3993.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, Wong E, Sheng J, Zhang Y, Poh T, Chan CS, Kunarso G, Shahab A, Bourque G, Cacheux-Rataboul V, Sung WK, Ruan Y, Wei CL. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**(7):630–638.

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. 2013. Natural RNA circles function as efficient microRNA sponges. *Nature* **495**(7441):384–388.

Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**(1):139–144.

Hansson J, Rafiee MR, Reiland S, Polo JM, Gehring J, Okawa S, Huber W, Hochedlinger K, Krijgsveld J. 2012. Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. *Cell Rep* **2**(6):1579–1592.

Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16**(9):369–372.

Hardison RC. 2012. Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J Biol Chem* **287**(37):30932–30940.

Hardison RC, Blobel GA. 2013. Genetics. GWAS to therapy by genome edits? *Science* **342**(6155):206–207.

Hardison RC, Butler ET 3rd, Lacy E, Maniatis T, Rosenthal N, Efstratiadis A. 1979. The structure and transcription of four linked rabbit β-like globin genes. *Cell* **18**(4):1285–1297.

Hardison RC, Oeltjen J, Miller W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7**(10):959–966.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet* **4**(6):e1000106.

Harmon K. 2012 Sep 5. “Junk” DNA Holds Clues to Common Diseases. *Scientific American*

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320**(5872):106–109.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**:S4.1–9.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**(9):1760–1774.

Hashimoto H, Vertino PM, Cheng X. 2010. Molecular coupling of DNA methylation and histone methylation. *Epigenomics* **2**(5):657–669.

Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep* **2**(3):666–673.

Hawkins PG, Morris KV. 2010. Transcriptional regulation of *Oct4* by a long non-coding RNA antisense to *Oct4*-pseudogene 5. *Transcription* **1**:165-175.

Hayashi Y, Yoshida M, Yamato M, Ide T, Wu Z, Ochi-Shindou M, Kanki T, Kang D, Sunagawa K, Tsutsui H, Nakanishi H. 2008. Reverse of age-dependent memory impairment and mitochondrial DNA damage in microglia by an overexpression of human mitochondrial transcription factor a in mice. *J Neurosci* **28**(34):8624–8234.

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* **6**(2):e1000834.

He A, Kong SW, Ma Q, Pu WT. 2011. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci U S A* **108**(14):5632–5637.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat Genet* **43**:414-420.

Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. 2011. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7**:497.

Hecht A, Strahl-Bolsinger S, Grunstein M. (1996) Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature* **383**(6595):92–96.

Heikkinen S, Väisänen S, Pehkonen P, Seuter S, Benes V, Carlberg C. 2011. Nuclear hormone 1α,25-dihydroxyvitamin D$_3$ elicits a genome-wide shift in the locations of VDR chromatin occupancy. *Nucleic Acids Res* **39**(21):9181–9193.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**(7243):108–112.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**(3):311–318.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**:576-589.

Heng JC, Feng B, Han J, Jiang J, Kraus P, Ng JH, Orlov YL, Huss M, Yang L, Lufkin T, Lim B, Ng HH. 2010. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell* **6**(2):167–174.

Henikoff S. 2008. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet* **9**(1):15–26.

Henriquez FL, Richards TA, Roberts F, McLeod R, Roberts CW. 2005. The unusual mitochondrial compartment of *Cryptosporidium parvum*. *Trends Parasitol* **21**(2):68–74.

Henry GL, Davis FP, Picard S, Eddy SR. 2012. Cell type-specific genomics of *Drosophila* neurons. *Nucleic Acids Res* **40**(19):9691–9704.

Herrick G, Cartinhour SW, Williams KR, Kotter KP. 1987a. Multiple sequence versions of the *Oxytricha fallax* 81-MAC alternate processing family. *J Protozool* **34**(4):429–434.

Herrick G, Hunter D, Williams K, Kotter K. 1987b. Alternative processing during development of a macronuclear chromosome family in *Oxytricha fallax*. *Genes Dev* **1**(10):1047–1058.

Hershey AD, Chase M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**(1):39–56.

Herzberg NH, Middelkoop E, Adorf M, Dekker HL, Van Galen MJ, Van den Berg M, Bolhuis PA, Van den Bogert C. 1993. Mitochondria in cultured human muscle cells depleted of mitochondrial DNA. *Eur J Cell Biol* **61**(2):400–408.

Hess WR, Börner T. 1999. Organellar RNA polymerases of higher plants. *Int Rev Cytol* **190**:1–59.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph SJ, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA. 2009 Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, **6**:283–289.

Heyman J, Cools T, Vandenbussche F, Heyndrickx KS, Van Leene J, Vercauteren I, Vanderauwera S, Vandepoele K, De Jaeger G, Van Der Straeten D, De Veylder L. 2013. ERF115 controls root quiescent center cell division and stem cell replenishment. *Science* **342**(6160):860–863.

Heyne K, Mannebach S, Wuertz E, Knaup KX, Mahyar-Roemer M, Roemer K. 2004. Identification of a putative p53 binding sequence within the human mitochondrial genome. *FEBS Lett* **578**(1-2):198–202.

Heyse G, Jonsson F, Chang WJ, Lipps HJ. 2010. RNA-dependent control of gene amplification. *Proc Natl Acad Sci U S A* **107**:22134-22139

Hibberd DJ, Norris RE. 1984. Cytology and ultrastructure of *Chlorarachnion reptans* (*Chlorarachniophyta divisio nova, Chlorarachniophyceae classis nova*). *J Phycol* **20**:310-30

Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M. 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet* **36**(12):1255–1257.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**(23):9362–9367.

Hirakawa Y, Burki F, Keeling PJ. 2011. Nucleus- and nucleomorph-targeted histone proteins in a chlorarachniophyte alga. *Mol Microbiol* **80**(6):1439–1449.

Hirakawa Y, Suzuki S, Archibald JM, Keeling PJ, Ishida KI. 2014. Overexpression of molecular chaperone genes in nucleomorph genomes. *Mol Biol Evol* [Epub ahead of print]

Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, Ohta K. 2008. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**(7218):130–134.

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**(6935):91–96.

Ho JW, Bishop E, Karchenko PV, Négre N, White KP, Park PJ. 2011. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* **12**:134.

Ho L, Jothi R, Ronan JL, Cui K, Zhao K, Crabtree GR. 2009. An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. *Proc Natl Acad Sci U S A* **106**(13):5187–5191.

Hoeijmakers WA, Stunnenberg HG, Bártfai R. 2012. Placing the *Plasmodium falciparum* epigenome on the map. *Trends Parasitol* **28**(11):486–495.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**(5):473–476.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 41(2):827–841.

Holland CA, Mayrand S, Pederson T. 1980. Sequence complexity of nuclear and messenger RNA in HeLa cells. *J Mol Biol* **138**:755778.

Hollenhorst PC, Chandler KJ, Poulsen RL, Johnson WE, Speck NA, Graves BJ. 2009. DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet* **5**(12):e1000778.

Holley RW, Everett GA, Madison JT, Zamir A. 1965a. Nucleotide Sequences In The Yeast Alanine Transfer Ribonucleic Acid. *J Biol Chem* **240**:2122–2128.

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. 1965. Structure of A Ribonucleic Acid. *Science* **147**(3664):1462–1465.

Holliday R, Pugh JE. 1975. DNA modification mechanisms and gene activity during development. *Science* **187**(4173):226–232.

Holmstrom SR, Deering T, Swift GH, Poelwijk FJ, Mangelsdorf DJ, Kliewer SA, MacDonald RJ. 2011. LRH-1 and PTF1-L coregulate an exocrine pancreas-specific transcriptional network for digestive function. *Genes Dev* **25**(16):1674–1679.

Holt IJ, Lorimer HE, Jacobs HT. 2000. Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell* **100**(5):515–524.

Holt IJ, Reyes A. 2012. Human mitochondrial DNA replication. *Cold Spring Harb Perspect Biol* **4**(12).

Holtgrewe M. 2010. Mason a read simulator for second generation sequencing data. *Technical Report TR-B-10-06, Institut für Mathematik und Informatik, Freie Universität Berlin*

Hood L, Galas D. 2003. The digital code of DNA. *Nature* **421**(6921):444–448.

Horak CE, Snyder M. 2002. ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol* **350**:469-483.

Horiuchi S, Onodera A, Hosokawa H, Watanabe Y, Tanaka T, Sugano S, Suzuki Y, Nakayama T. 2011. Genome-wide analysis reveals unique regulation of transcription of Th2-specific genes by GATA3. *J Immunol* **186**(11):6378–6389.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, Yang L, Boley N, Andrews J, Kaufman TC, Graveley BR, Bickel PJ, Carninci P, Carlson JW, Celniker SE. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**(2):182–192.

Hotchkiss RD. 1948. The quantitative separation of purines, pyrimidines and nucleosides by paper chromatography. *J Biol Chem* **175**(1):315-332.

Hotto AM, Schmitz RJ, Fei Z, Ecker JR, Stern DB. 2011. Unexpected Diversity of Chloroplast Noncoding RNAs as Revealed by Deep Sequencing of the *Arabidopsis* Transcriptome. *G3 (Bethesda)* **1**(7):559–570.

Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D, Wu H, Ye X, Ye C, Wu R, Jian M, Chen Y, Xie W, Zhang R, Chen L, Liu X, Yao X, Zheng H, Yu C, Li Q, Gong Z, Mao M, Yang X, Yang L, Li J, Wang W, Lu Z, Gu N, Laurie G, Bolund L, Kristiansen K, Wang J, Yang H, Li Y, Zhang X, Wang J. 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**(5):873–885.

Houbaviy HB, Murray MF, Sharp PA. 2003. Embryonic stem cell-specific microRNAs. *Dev Cell* **5**:351-358.

Hough BR, Smith MJ, Britten RJ, Davidson EH. 1975. Sequence complexity of heterogeneous nuclear RNA in sea urchin embryos. *Cell* **5**:291299.

Hower V, Evans SN, Pachter L. 2011. Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics* **12**:15.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, Plasterk RH, Hannon GJ, Draper BW, Ketting RF. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**(1):69–82.

Hsu MT, Coca-Prados M. 1979. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* **280**(5720):339–340.

Hu G, Schones DE, Cui K, Ybarra R, Northrup D, Tang Q, Gattinoni L, Restifo NP, Huang S, Zhao K. 2011. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res* **21**(10):1650–1658.

Hu J, Ge H, Newman M, Liu K. 2012. OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics* **28**(14):1933–1934.

Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS. 2010. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* **38**(7):2154–2167.

Hu M, Zhu Y, Taylor JM, Liu JS, Qin ZS. 2012. Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics* **28**(1):63–68.

Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan PF, Hammond SM, Makowski L, Randell SH, Chiang DY, Hayes DN, Jones C, Liu Y, Prins JF, Liu J. 2013. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res* **41**(2):e39

Hu Y, Liu Y, Mao X, Jia C, Ferguson JF, Xue C, Reilly MP, Li H, Li M. 2014. PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. *Nucleic Acids Res* **42**(3):e20.

Hua S, Kittler R, White KP. 2009. Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell* **137**:1259-1271.

Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* **422**(6927):72–76.

Huang DW, Sherman B, Lempicki R. 2009a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**:1-13.

Huang DW, Sherman B, Lempicki R. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**:44-57.

Huang S, Zhang J, Li R, Zhang W, He Z, Lam T-W, Peng Z, Yiu S-M (2011). SOAPsplice: Genome-Wide *ab initio* Detection of Splice Junctions from RNA-Seq Data. *Front Genetics* **2**(July):46.

Huang W, Pérez-García P, Pokhilko A, Millar AJ, Antoshechkin I, Riechmann JL, Mas P. 2012. Mapping the core of the *Arabidopsis* circadian clock defines the network structure of the oscillator. *Science* **336**(6077):75–79.

Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. 2013. A major epigenetic programming mechanism guided by piRNAs. *Dev Cell* **24**(5):502–516.

Huang XY, Hirsh D. 1989. A second trans-spliced RNA leader sequence in the nematode *Caenorhabditis elegans. Proc Natl Acad Sci U S A* **86**:8640-8644

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T, Rinn JL. 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**(3):409–419.

Huff JT, Zilberman D. 2014. Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. *Cell* **156**(6)1286:1297

Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* **377**(6548):391.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**(7402):207–214.

Hunkapiller J, Shen Y, Diaz A, Cagney G, McCleary D, Ramalho-Santos M, Krogan N, Ren B, Song JS, Reiter JF. 2012. Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. *PLoS Genet* **8**(3):e1002576.

Hurtley S. 2012. No more junk DNA. *Science* **337**:1581.

Hutchins AP, Poulain S, Miranda-Saavedra D. 2012. Genome-wide analysis of STAT3 binding in vivo predicts effectors of the anti-inflammatory response in macrophages. *Blood* **119**(13):e110–119.

Hutvagner G, Simard M. 2008. Argonaute proteins: Key players in RNA silencing. *Nat Rev Mol Cell Biol* **9**:22-32.

Huxley JS. 1942. Evolution: the modern synthesis. *Allen and Unwin, London, United Kingdom*

Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31**(3):227–229.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juárez MJ, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M, Góngora-Castillo E, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, de Jesús Ortega-Estrada M, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L. 2013. Architecture and evolution of a minute plant genome. *Nature* **498**(7452):94–98.

Ide T, Tsutsui H, Hayashidani S, Kang D, Suematsu N, Nakamura K, Utsumi H, Hamasaki N, Takeshita A. 2001. Mitochondrial DNA damage and dysfunction associated with oxidative stress in failing hearts after myocardial infarction. *Circ Res* **88**(5):529–535.

Immink RG, Posé D, Ferrario S, Ott F, Kaufmann K, Valentim FL, de Folter S, van der Wal F, van Dijk AD, Schmid M, Angenent GC. 2012. Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. *Plant Physiol* **160**(1):433–449.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924):218–223.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011):931–945.

Ishizu H, Nagao A, Siomi H. 2011. Gatekeepers for Piwi-piRNA complexes to enter the nucleus. *Curr Opin Genetic Dev* **21**:484-490.

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**(7):1160–1167.

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**(2):163-136.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley SM, Bianchettin G, Borzym K, Bothe G, Bruschi CV, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson RM, Cronin A, Cruz AK, Davies RM, De Gaudenzi J, Dobson DE, Duesterhoeft A, Fazelina G, Fosker N, Frasch AC, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram JC, Mller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'neil S, Pentony M, Pohl TM, Price C, Purnelle B, Quail MA, Rabbinowitsch E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz JC, Rutter S, Saunders D, Schfer M, Schein J, Schwartz DC, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G, Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith DF, Blackwell JM, Stuart KD, Barrell B, Myler PJ. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**(5733):436–442.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**:533-538.

Izquierdo JM, Majós N, Bonnal S, Martnez C, Castelo R, Guigó R, Bilbao D, Valcárcel J. 2005. Regulation of *Fas* alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell* **19**(4):475–484.

Jackson AP, Quail MA, Berriman M. 2008. Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). *BMC Genomics* **9**:594.

Jackson CJ, Norman JE, Schnare MN, Gray MW, Keeling PJ, Waller RF. 2007. Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biol* **5**:41

Jackson IJ. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res* **19**(14):3795–3798.

Jackson B, Schnable P, Aluru S. 2009. Parallel short sequence assembly of transcriptomes. *BMC Bioinformatics* **10**(Suppl 1):S14+

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**:318-356.

Jacq C, Miller JR, Brownlee GG. 1977. A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* **12**(1):109–120.

Jaeckisch N, Yang I, Wohlrab S, Glöckner G, Kroymann J, Vogel H, Cembella A, John U. 2011. Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*. *PLoS One* **6**(12):e28012.

Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. *Annu Rev Microbiol* **56**:489-520.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**(7011):946–957.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**(6172):776–779.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**(7328):97–101.

Janzen CJ, Fernandez JP, Deng H, Diaz R, Hake SB, Cross GA. 2006. Unusual histone modifications in *Trypanosoma brucei*. *FEBS Lett* **580**(9):2306–2310.

Jao LE, Wente SR, Chen W. 2013. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc Natl Acad Sci U S A* **110**(34):13904–13909.

Jean G, Kahles A, Sreedharan VT, De Bona F, Rätsch G. 2010. RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics* Chapter 11:Unit 11.6.

Jensen BC, Sivam D, Kifer CT, Myler PJ, Parsons M. 2009. Widespread variation in transcript abundance within and across developmental stages of *Trypanosoma brucei*. *BMC Genomics* **10**:482.

Jenuwein T, Allis CD. 2001. Translating the histone code. *Science* **293**(5532):1074-1080.

Jerlström P. 2000. Pseudogenes: are they non-functional? *Creation Ex Nihilo Technical Journal* **14**:15.

Jha A. 2012 Sep 5. Breakthrough study overturns theory of 'junk DNA' in genome. *The Guardian*

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26**(11):1293–1300.

Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H, Müller-Tidow C. 2003. MALAT-1, a novel noncoding RNA, and thymosin $\beta 4$ predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**(39):8031–8041.

Jiang H, Wong WH. 2008. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**(20):2395–2396.

Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**(8):1026–1032.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21**(9):1543–1551.

Jin VX, Singer GA, Agosto-Pérez FJ, Liyanarachchi S, Davuluri RV. 2006. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics* **7**:114.

Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. 2013a. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol* **31**(3):233–239.

Jiang W, Zhou H, Bi H, Fromm M, Yang B, Weeks DP. 2013b. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice. *Nucleic Acids Res* **41**(20):e188.

Jiang Y, Schneck JL, Grimes M, Taylor AN, Hou W, Thrall SH, Sweitzer SM. 2011. Methyltransferases prefer monomer over core-trimmed nucleosomes as in vitro substrates. *Anal Biochem* **415**(1):84–86.

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. 2009. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet* **41**(8):941–945.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**(6096):816–821.

Jo K, Kwon HB, Kim S. 2014. Time-series RNA-seq analysis package (TRAP) and its application to the analysis of rice, *Oryza sativa* L. ssp. *Japonica*, upon drought stress. *Methods* S1046–2023(14)00029–2.

Johannes F, Wardenaar R, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HT, Cuppen E, Jansen RC. 2010. Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* 26(8):1000–1006.

Johannsen WL. 1909. Elemente der Exakten Erblichkeitslehre. *Fischer, Jena, F.R.G.*

Johansen K, Cai W, Deng H, Bao X, Zhang W, Girton J, Johansen J. 2009. Polytene chromosome squash methods for studying transcription and epigenetic chromatin modification in *Drosophila* using antibodies. *Methods* **48**:387-397.

John P, Whatley FR. 1975. *Paracoccus denitrificans* and the evolutionary origin of the mitochondrion. *Nature* **254**(5500):495–498.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830):1497–1502.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**(5653):2141–2144.

Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21**(2):93–102.

Johnson RF, Witzel II, Perkins ND. 2011. p53-dependent regulation of mitochondrial energy production by the RelA subunit of NF-κB. *Cancer Res* **71**(16):5588–5597.

Johnson TB, Coghill RD. 1925. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the Tubercle bacillus. *J Am Chem Soc* **47**(11):2838-2844.

Johnston JS, Ross LD, Beani L, Hughes DP, Kathirithamby J. 2004. Tiny genomes and endoreduplication in *Strepsiptera*. *Insect Mol Biol* **13**(6):581–585.

Joseph R, Orlov YL, Huss M, Sun W, Kong SL, Ukil L, Pan YF, Li G, Lim M, Thomsen JS, Ruan Y, Clarke ND, Prabhakar S, Cheung E, Liu ET. 2010. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Mol Syst Biol* **6**:456.

Josse T, Teysset L, Todeschini A-L, Sidor C, Anxolabéhére D, Ronsseray S. 2007. Telomeric trans-silencing: An epigenetic repression combining RNA silencing and heterochromatin formation. *PLoS Genetic* **3**:1633-1643.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36**(16):5221–5231.

Jung H, Lacombe J, Mazzoni EO, Liem KF Jr, Grinstein J, Mahony S, Mukhopadhyay

D, Gifford DK, Young RA, Anderson KV, Wichterle H, Dasen JS. 2010. Global control of motor neuron topography mediated by the repressive actions of a single hox gene. *Neuron* **67**(5):781–796.

Jung YL, Luquette LJ, Ho JW, Ferrari F, Tolstorukov M, Minoda A, Issner R, Epstein CB, Karpen GH, Kuroda MI, Park PJ. 2014. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.* 2014 [Epub ahead of print]

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**(7314):430–435.

Kairo A, Fairlamb AH, Gobright E, Nene V. 1994. A 7.1 kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. *EMBO J* **13**:898-905.

Kamikawa R, Nishimura H, Sako Y. 2009. Analysis of the mitochondrial genome, transcripts, and electron transport activity in the dinoflagellate *Alexandrium catenella* (Gonyaulacales, Dinophyceae). *Phycol Res* **57**:1-11.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**(3):331–342.

Kang D, Kim SH, and Hamasaki N. 2007. Mitochondrial transcription factor A (TFAM): roles in maintenance of mtDNA and cellular functions. *Mitochondrion* **7**(1-2): 39–44.

Kang D, Miyako K, Kai Y, Irie T, Takeshige K. 1997. In vivo determination of replication origins of human mitochondrial DNA by ligation-mediated polymerase chain reaction. *J Biol Chem* **272**(24):15275–15279.

Kanki T, Ohgaki K, Gaspari M, Gustafsson CM, Fukuoh A, Sasaki N, Hamasaki N, Kang D. 2004. Architectural role of mitochondrial transcription factor A in maintenance of human mitochondrial DNA. *Mol Cell Biol* **24**(22):9823–9834.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**(5569):916–919.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**(5830):1484–1488.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* **15**(7):987–997.

Kapranov P, Willingham AT, Gingeras TR. 2007. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8**(6):413–423.

Kapraun DF, Freshwater DW. 2012. Estimates of nuclear DNA content in red algal lineages. *AoB Plants* **2012**:pls005.

Karamanlidis G, Nascimben L, Couper GS, Shekar PS, del Monte F, Tian R. 2010. Defective DNA replication impairs mitochodnrial biogenesis in human failing hearts. *Circ Res* **106**(9):1541–1548.

Karamanlidis G, Bautista-Hernandez FV, Fynn-Thompson F, Del Nido P, Tian R. 2011.) Impaired mitochondrial biogenesis precedes heart failure in right ventricular hypertrophy in congenital heart disease. *Circ Heart Fail* **4**(6):707–713.

Karreth FA, Pandolfi PP. 2013. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. *Cancer Discov* **3**(10):1113–1121.

Karreth FA, Tay Y, Perna D, Ala U, Tan SM, Rust AG, DeNicola G, Webster KA, Weiss D, Perez-Mancera PA, Krauthammer M, Halaban R, Provero P, Adams DJ, Tuveson DA, Pandolfi PP. 2011. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147**(2):382–395.

Kasamatsu H, Robberson DL, Vinograd J. 1971. A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. *Proc Natl Acad Sci U S A* **68**(9):2252–2257.

Kasamatsu H, Vinograd J. 1973. Unidirectionality of replication in mouse mitochondrial DNA. *Nat New Biol* **241**(108):103–105.

Kasianowicz JJ, Brandin E, Branton D, Deamer DW. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* **93**(24):13770–13773.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M. 2010. Variation in transcription factor binding among humans. *Science* **328**(5975):232–235.

Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**(8):1064–1083.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engstrm PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309**(5740):1564–1566.

Katic I, Großlhans H. 2013. Targeted heritable mutation and gene conversion by Cas9-CRISPR in *Caenorhabditis elegans*. *Genetics* **195**(3):1173–1176.

Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarés CP. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**(6862):450–543.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**(12):1009–1015.

Kaufman BA, Durisic N, Mativetsky JM, Costantino S, Hancock MA, Grutter P, and Shoubridge EA. 2007. The mitochondrial transcription factor TFAM coordinates the assembly of multiple DNA molecules into nucleoid-like structures. *Mol Biol Cell* **18**(9):3225–3236.

Kaufmann BB, van Oudenaarden A. 2007. Stochastic gene expression: from single molecules to the proteome. *Curr Opin Genet Dev* **17**(2):107–112.

Kaufmann K, Muiño JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC. 2009. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the *Arabidopsis* flower. *PLoS Biol* **7**(4):e1000090.

Kaufmann K, Wellmer F, Muiño JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueño F, Krajewski P, Meyerowitz EM, Angenent GC, Riechmann JL. 2010. Orchestration of floral initiation by APETALA1. *Science* **328**(5974):85–89.

Kawahara Y, Oono Y, Kanamori H, Matsumoto T, Itoh T, Minami E. 2012. Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. *PLoS ONE* **7**(11):e49423.

Kedersha NL, Rome LH. 1986. Isolation and characterization of a novel ribonucleoprotein particle: large structures contain a single species of small RNA. *J Cell Biol* **103**(3):699–709.

Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. *Am J Bot* **91**(10):1481–1493.

Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukaryot Microbiol* **56**(1):1–8.

Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci* **365**(1541):729–748.

Keeling PJ. 2013. The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. *Annu Rev Plant Biol* **64**:583-607.

Keeney PM, Quigley CK, Dunham LD, Papageorge CM, Iyer S, Thomas RR, Schwarz KM, Trimmer PA, Khan SM, Portell FR, Bergquist KE, Bennett JP Jr. 2009. Mitochondrial gene therapy augments mitochondrial physiology in a Parkinson's disease cell model. *Hum Gene Ther* **20**(8):897–907.

Keightley PD. 2012. Rates and fitness consequences of new mutations in humans. *Genetics* **190**:295–304.

Keller C, Adaixo R, Stunnenberg R, Woolcock KJ, Hiller S, Buhler M. 2012. HP1(Swi6) mediates the recognition and destruction of heterochromatic RNA transcripts. *Mol Cell*

47:215-227.

Kellis M, Hardison RC, Wold BJ, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski L, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard TJP, Kent WJ, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos J, Weng Z, White KP, Members of the ENCODE Consortium. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* 111(17):6131–6138.

Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**(12):2497–2506.

Kent WJ. 2012. BLAT – the BLAST-like alignment tool. *Genome Res* **12**(4):656–664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**(6):996–1006.

Kenzelmann Broz D, Spano Mello S, Bieging KT, Jiang D, Dusek RL, Brady CA, Sidow A, Attardi LD. 2013. Global genomic profiling reveals an extensive p53-regulated autophagy program contributing to key p53 responses. *Genes Dev* **27**(9):1016–1031.

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**(7311):103–107.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. 2009. Many human large intergenic non-coding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**(28):11667–11672.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, Sandstrom R, Thurman RE, MacAlpine DM, Stamatoyannopoulos JA, Kellis M, Elgin SC, Kuroda MI, Pirrotta V, Karpen GH, Park PJ. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**(7339):480–485.

Kharchenko PV, Tolstorukov MY, and Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26:**1351–1359.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**(5):800–811.

Khorana HG, Büchi H, Ghosh H, Gupta N, Jacob TM, Kössel H, Morgan R, Narang SA, Ohtsuka E, Wells RD. 1966, Polynucleotide synthesis and the genetic code. *Cold Spring Harb Symp Quant Biol* **31**:39–49.

Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, Resch W, Baek S, Pruett N, Grøntved L, Vian L, Nelson S, Zare H, Hakim O, Reyon D, Yamane A, Nakahashi H, Kovalchuk AL, Zou J, Joung JK, Sartorelli V, Wei CL, Ruan X, Hager GL, Ruan Y, Casellas R. 2013. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**(7):1507–1520.

Kielpinski LJ, Vinther J. 2014. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic Acids Res* **42**(8):e70

Kiethega GN, Turcotte M, Burger G. 2011. Evolutionarily conserved *cox1* trans-splicing without *cis*-motifs. *Mol Biol Evol* **28**(9):2425–2428.

Kim D, Salzberg SL. 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12**(8):R72.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res* **35**(1):125–131.

Kim H, Bi Y, Pal S, Gupta R, Davuluri RV. 2011. IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-seq data. *BMC Bioinformatics* **12**:305.

Kim J, Chu J, Shen X, Wang J, Orkin SH. 2008. An extended transcriptional network for pluripotency of embryonic stem cells. Cell 21;132(6):1049–1061.

Kim K, Doi A, Wen B, Ng K, Zhao R, Cahan P, Kim J, Aryee MJ, Ji H, Ehrlich LI, Yabuuchi A, Takeuchi A, Cunniff KC, Hongguang H, McKinney-Freeman S, Naveiras O, Yoon TJ, Irizarry RA, Jung N, Seita J, Hanna J, Murakami P, Jaenisch R, Weissleder R, Orkin

SH, Weissman IL, Feinberg AP, Daley GQ. 2010. Epigenetic memory in induced pluripotent stem cells. *Nature* **467**(7313):285–290.

Kim SW, Yoon SJ, Chuong E, Oyolu C, Wills AE, Gupta R, Baker J. 2011. Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. *Dev Biol* **357**(2):492–504.

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436**(7052):876–880.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**(7295):182–187.

Kimura M. 1962. On the probability of fixation of mutant genes in populations. *Genetics* **47**:713-719

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**:624-626.

Kimura M. 1983. The neutral theory of molecular evolution. *Cambridge Univ. Press, Cambridge.*

King CR, Piatigorsky J. 1983. Alternative RNA splicing of the murine α A-crystallin gene: protein-coding information within an intron. *Cell* **32**(3):707–712.

King JL, Jukes TH. 1969. Non-Darwinian Evolution. *Science* **164**:788–797.

King MP, Attardi G. 1989. Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation. *Science* **246**(4929):500–503.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Sequencing JG, Bork P, Lim WA, Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev IV, Rokhsar D. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**(7180):783–788.

Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. 2011. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics* **27**(8):1068-75.

Kiriakidou M, Tan G, Lamprinaki S, De Planell-Saguer M, Nelson P, Mourelatos Z. 2007. An mRNA m^{7}G cap binding-like motif within human Ago2 represses translation. *Cell* **129**:1141-1151.

Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y; RIKEN GER Group; GSL Members. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res* **13**(6B):1324–1334.

Kleine T, Maier UG, Leister D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol* **60**:115-138.

Kleinjan DA, Lettice LA. 2008. Long-range gene control and genetic disease. *Adv Genet* **61**:339-388.

Klemsz MJ, McKercher SR, Celada A, Van Beveren C, Maki RA. 1990. The macrophage and B cell-specific transcription factor PU.1 is related to the *ets* oncogene. *Cell* **61**(1):113–124.

Klenov M, Sokolova O, Yakushev E, Stolyarenko A, Mikhaleva E, Lavrov S, Gvozdev V. 2011. Separation of stem cell maintenance and transposon silencing functions of Piwi protein. *Proc Natl Acad Sci* **108**:18760-18765.

Klisch TJ, Xi Y, Flora A, Wang L, Li W, Zoghbi HY. 2011. In vivo Atoh1 targetome reveals how a proneural transcription factor regulates cerebellar development. *Proc Natl Acad Sci U S A* **108**(8):3288–3293.

Klobutcher LA, Huff ME, Gonye GE. 1988. Alternative use of chromosome fragmentation sites in the ciliated protozoan *Oxytricha nova*. *Nucleic Acids Res* **16**(1):251–264.

Kloc M, Zagrodzinska B. 2001. Chromatin elimination – an oddity or a common mechanism in differentiation and development? *Differentiation* **68**(2-3):84–91.

Knoll AH. 2011. The Multiple Origins of Complex Multicellularity. *Ann Rev Earth Plan Sci* **39**:217–239.

Koch F, Jourquin F, Ferrier P, Andrau JC. 2008. Genome-wide RNA polymerase II: not genes only! *Trends Biochem Sci* **33**(6):265–273.

Koboldt DC, Ding L, Mardis ER, Wilson RK. 2010. Challenges of sequencing human genomes. *Brief Bioinform* **11**(5):484–498.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P. 2006. CAGE: cap analysis of gene expression. *Nat Methods* **3**(3):211–222.

Koeppel M, van Heeringen SJ, Kramer D, Smeenk L, Janssen-Megens E, Hartmann M, Stunnenberg HG, Lohrum M. 2011. Crosstalk between c-Jun and TAp73α/β contributes to the apoptosis-survival balance. *Nucleic Acids Res* **39**(14):6069–6085.

Kohonen T. 1982. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics* **43**(1): 59-69.

Kohonen T. 2013. Essentials of the self-organizing map. *Neural Netw* **37**:52–65.

Kolata G. 2012 Sep 5. Bits of mystery DNA, far from Junk, play crucial role. *New York Times (Science)*

Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. 2010. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog* **6**(9):e1001090.

Kolodner R, Tewari KK. 1979. Inverted repeats in chloroplast DNA from higher plants. *Proc Natl Acad Sci U S A* **76**:41–45.

Kondo S, Ueda R. 2013. Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*. *Genetics* **195**(3):715–721.

Kong SL, Li G, Loh SL, Sung WK, Liu ET. 2010. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. *Mol Syst Biol* **7**:526.

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**(7):909–915.

Koonin EV. 2004. A Non-Adaptationist Perspective on Evolution of Genomic Complexity or the Continued Detroning of Man. *Cell Cycle.* **3**(3):280–285.

Koonin EV. 2006. The Origin of Introns and Their Role in Eukaryogenesis: A Compromise Solution to the Introns-Early Versus Introns-Late Debate? *Biol Direct* **1**:22.

Koonin EV. 2011. The Logic of Chance: The Nature and Origin of Biological Evolution. *FT Press Science*

Koonin EV, Wolf YI. 2009. Is evolution Darwinian or/and Lamarckian? *Biol Direct* **4**:42.

Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**(4139):868–871.

Kornblihtt AR, Pesce CG, Alonso CR, Cramer P, Srebrow A, Werbajh S, Muro AF. 1996. The fibronectin gene as a model for splicing and transcription studies. *FASEB J* **10**(2):248–257.

Korneev S, O'Shea M. 2005. Natural antisense RNAs in the nervous system. *Rev Neurosci* **16**(3):213–222.

Korneev SA, Park JH, O'Shea M. 1999. Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J Neurosci* **19**:7711-7720.

Koufali MM, Moutsatsou P, Sekeris CE, Breen KC. 2003. The dynamic localization of the glucocorticoid receptor in rat C6 glioma cell mitochondria. *Mol Cell Endocrinol* **209**(1-2):51–60.

Kouwenhoven EN, van Heeringen SJ, Tena JJ, Oti M, Dutilh BE, Alonso ME, de la Calle-Mustienes E, Smeenk L, Rinne T, Parsaulian L, Bolat E, Jurgelenaite R, Huynen MA, Hoischen A, Veltman JA, Brunner HG, Roscioli T, Oates E, Wilson M, Manzanares M, Gómez-Skarmeta JL, Stunnenberg HG, Lohrum M, van Bokhoven H, Zhou H. 2010. Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. *PLoS Genet* **6**(8):e1001065.

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**:693-705

Koziol MJ, Rinn JL. 2010. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev* **20**(2):142–148.

Kramer S. 2012. Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. *Mol Biochem Parasitol* **181**(2):61–72.

Krause M, Hirsh D. 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* **49**(6):753–761.

Kravchenko JE, Rogozin IB, Koonin EV, Chumakov PM. 2005. Transcription of mammalian messenger RNAs by a nuclear RNA polymerase of mitochondrial origin. *Nature* **436**(7051):735–739.

Krebs AR, Demmers J, Karmodiya K, Chang NC, Chang AC, Tora L. 2010. ATAC and Mediator coactivators form a stable complex and regulate a set of non-coding RNA genes. *EMBO Rep* **11**(7):541–547.

Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J, Johnston D, Kim GE, Spitale RC, Flynn RA, Zheng GX, Aiyer S, Raj A, Rinn JL, Chang HY, Khavari PA. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**(7431):231–235.

Kriaucionis S, Heintz N. 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**(5929):929–930.

Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, Green R, Farnham PJ. 2007. Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays. *J Biol Chem* **282**:9703-9712.

Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S. 2003. Increase of functional diversity by alternative splicing. *Trends Genet* **19**(3):124–128.

Krueger F, Kreck B, Franke A, Andrews SR. 2012. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* **9**(2):145–151.

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **31**(1):147–157.

Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J. 2007. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res* **17**(8):1139–1145.

Kruse B, Narasimhan N, Attardi G. 1989. Termination of transcription in human mitochondria: identification and purification of a DNA binding protein factor that promotes termination. *Cell* **58**(2):391–397.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**(9):1639–1645.

Kucej M, Kucejova B, Subramanian R, Chen XJ, Butow RA. 2008. Mitochondrial nucleoids undergo remodeling in response to metabolic cues. *J Cell Sci* **121**(Pt 11):1861–1868.

Kudla G, Granneman S, Hahn D, Beggs JD, Tollervey D. 2011. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc Natl Acad Sci U S A* **108**(24):10010–10015.

Kuehner JN, Brow DA. 2008. Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol Cell* **31**:201-211

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M. 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* **32**(3):261-266.

Kulis M, Queirós AC, Beekman R, Martn-Subero JI. 2013. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim Biophys Acta* **1829**(11):1161–1174.

Kullman B, Tamm H, Kullman K. 2005. Fungal Genome Size Database. `http://www.zbi.ee/fungal-genomesize`

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**(7):631–634.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* **22**(9):1735–1747.

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, Hata K, Li E, Matsuda Y, Kimura T, Okabe M, Sakaki Y, Sasaki H, Nakano T. 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev* **22**(7):908–917.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40**(2):225–231.

Kwon H, Thierry-Mieg D, Thierry-Mieg J, Kim HP, Oh J, Tunyaplin C, Carotta S, Donovan CE, Goldman ML, Tailor P, Ozato K, Levy DE, Nutt SL, Calame K, Leonard WJ. 2009. Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors. *Immunity* **31**(6):941–952.

Labaj PP, Linggi BE, Wiley HS, Kreil DP. 2012. Improving RNA-Seq Precision with MapAl. *Front Genet* **3**:28.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**:853-858.

Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T. 2003. New microRNAs from mouse and human. *RNA* **9**:175-179.

Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. 1998. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* **12**(1):34–44.

Lahav T, Sivam D, Volpin H, Ronen M, Tsigankov P, Green A, Holland N, Kuzyk M, Borchers C, Zilberstein D, Myler PJ. 2011. Multiple levels of gene regulation mediate differentiation of the intracellular pathogen *Leishmania. FASEB J* **25**(2):515–525.

Lai EC, Tomancak P, Williams RW, Rubin GM. 2003. Computational identification of Drosophila microRNA genes. *Genome Biol* **4**:R42.

LaJeunesse TC, Lambert G, Andersen RA, Coffroth, MA, Galbraith DW. 2005. *Symbiodinium* (Pyrrhophyta) Genome Sizes (DNA Content) Are Smallest Among Dinoflagellates. *J Phycology* **41**(4):880–886.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci U S A* **81**(12):3786–3790.

Lam MT, Cho H, Lesch HP, Gosselin D, Heinz S, Tanaka-Oishi Y, Benner C, Kaikkonen MU, Kim AS, Kosaka M, Lee CY, Watt A, Grossman TR, Rosenfeld MG, Evans RM, Glass CK. 2013. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**(7455):511–515.

Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. *Annual Review of Genetics* **38**:1-35

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860–921.

Landry JR, Mager DL, Wilhelm BT. 2003. Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* **19**(11):640–648.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9):1813–1831.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM. 2007. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc Natl Acad Sci U S A* **104**(50):19908–19913.

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**(6632):493–497.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**:559.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.

Larsson NG, Garman JD, Oldfors A, Barsh GS, Clayton DA. 1996. A single mouse gene encodes the mitochondrial transcription factor A and a testis-specific nuclear HMG-box protein. *Nat Genet* **13**(3):296–302.

Larsson NG, Wang J, Wilhelmsson H, Oldfors A, Rustin P, Lewandoski M, Barsh GS, Clayton DA. 1998. Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. *Nat Genet* **18**(3):231–236.

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A* **94**(24):13057–13062.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**:858-862.

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* **313**(5785):363–367.

Lauer J, Shen CK, Maniatis T. 1980. The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. *Cell* **20**(1):119–130.

Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AM, Strahl BD, Patel DJ, Jacobsen SE. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**(7454):385–389.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**(3):204–220.

Law MJ, Lower KM, Voon HP, Hughes JR, Garrick D, Viprakasit V, Mitson M, De Gobbi M, Marra M, Morris A, Abbott A, Wilder SP, Taylor S, Santos GM, Cross J, Ayyub H, Jones S, Ragoussis J, Rhodes D, Dunham I, Higgs DR, Gibbons RJ. 2010. ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell* **143**(3):367–378.

Lawn RM, Heumann JM, Herrick G, Prescott DM. 1978. The gene-size DNA molecules in *Oxytricha*. *Cold Spring Harb Symp Quant Biol* **1**:483-492.

Lawrie DS, Petrov DA. 2014. Comparative population genomics: power and principles for the inference of functionality. *Trends Genet* **30**:133-139.

Le Guiner C, Gesnel MC, Breathnach R. 2003. TIA-1 or TIAR is required for DT40 cell viability. *J Biol Chem* **278**(12):10465–10476.

Lee BK, Bhinge AA, Iyer VR. 2010. Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res* **39**(9):3558–3573.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV, Park PJ; Cancer Genome Atlas Research Network. 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**(6097):967–971.

Lee HC, Gu W, Shirayama M, Youngman E, Conte D Jr, Mello CC. 2012. *C. elegans* piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell* **150**(1):78–87.

Lee J, Kim CH, Simon DK, Aminova LR, Andreyev AY, Kushnareva YE, Murphy AN, Lonze BE, Kim KS, Ginty DD, Ferrante RJ, Ryu H, Ratan RR. 2005. Mitochondrial cyclic AMP response element-binding protein (CREB) mediates mitochondrial gene expression and neuronal survival. *J Biol Chem* **280**(49):40398–40401.

Lee JS, Shilatifard A. 2007. A site to remember: H3K36 methylation a mark for histone deacetylation. *Mutat Res* **618**(1-2):130–134.

Lee JT. 2012. Epigenetic regulation by long noncoding RNAs. *Science* **338**(6113):1435–1439.

Lee JT, Davidow LS, Warshawsky D. 1999. *Tsix*, a gene antisense to *Xist* at the X-inactivation centre. *Nat Genet* **21**(4):400–404.

Lee RC, Feinbaum RL, Ambros V. 1993. The *C. elegans* heterochronic gene *lin*-4 encodes small RNAs with antisense complementarity to *lin*-14. *Cell* **75**(5):843–854.

Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee S, Lee B, Kang C, Lee S. 2011. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* **39**(2):e9.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**(5594):799–804.

Lee YL, Chiao CH, Hsu MT. 2011. Transcription of muscle actin genes by a nuclear form of mitochondrial RNA polymerase. *PLoS One* **6**(7):e22583.

Lefterova MI, Steger DJ, Zhuo D, Qatanani M, Mullican SE, Tuteja G, Manduchi E, Grant GR, Lazar MA. 2010. Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Mol Cell Biol* **30**(9):2078–2089.

LeGault LH, Dewey CN. 2013. Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs. *Bioinformatics* **29**(18):2300–2310.

Lehner B, Williams G, Campbell RD, Sanderson CM. 2002. Antisense transcripts in the human genome. *Trends Genet* **18**(2):63–65.

Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. 2007. Genomic and proteomic expression analysis of *Leishmania promastigote* and amastigote life stages: the *Leishmania* genome is constitutively expressed. *Mol Biochem Parasitol* **152**(1):35–46.

Leigh-Brown S, Enriquez JA, Odom DT. 2010. Nuclear transcription factors in mammalian mitochondria. *Genome Biol* **11**(7):215.

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**(8):1035–1043.

Lengyel P, Speyer JF, Ochoa S. 1961. Synthetic polynucleotides and the amino acid code. *Proc Natl Acad Sci U S A* **47**:1936–1942.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**(4):233–245.

Lepere G, Betermier M, Meyer E, Duharcourt S. 2008. Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia. Genes Dev* **22**:1501-1512.

Lepere G, Nowacki M, Serrano V, Gout JF, Guglielmi G, Duharcourt S, Meyer E. 2009. Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia. Nucleic Acids Res* **37**:903-915.

Lerner MR, Boyle JA, Hardin JA, Steitz JA. 1981. Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science*

211(4480):400–402.

Leroy S, Bouamer S, Morand S, Fargette M. 2007. Genome size of plant-parasitic nematodes. *Nematology* **9**:449–450.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**(14):1725–1735.

Levin JZ, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A. 2009. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* **10**(10):R115.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**(9):709–715.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424**(6945):147–151.

Levi-Setti R, Gavrilov KL, Rizzo PJ. 2008. Divalent cation distribution in dinoflagellate chromosomes imaged by high-resolution ion probe mass spectrometry. *Eur J Cell Biol* **87**:963-976.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**(1):189–192.

Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* **128**(4):707–719.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**:323.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**(4):493–500.

Li D, Qiu Z, Shao Y, Chen Y, Guan Y, Liu M, Li Y, Gao N, Wang L, Lu X, Zhao Y, Liu M. 2013. Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. *Nat Biotechnol* **31**(8):681–683.

Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi HS, Tennakoon C, Wei CL, Ruan Y, Sung WK. 2010. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol* **11**(2):R22.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei CL, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, Fullwood MJ, Cheung E, Liu E, Sung WK, Snyder M, Ruan Y. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**(1-2):84–98.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16):2078–2079.

Li J, Jiang H, Wong WH. 2010. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* **11**(5):R50.

Li JF, Norville JE, Aach J, McCormack M, Zhang D, Bush J, Church GM, Sheen J. 2013. Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9. *Nat Biotechnol* **31**(8):688–691.

Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. 2011. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci U S A* **108**(50):19867–19872.

Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, Evans MM, Scanlon MJ, Yu J, Schnable PS, Timmermans MC, Springer NM, Muehlbauer GJ. 2014. Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol* **15**(2):R40.

Li L, Jothi R, Cui K, Lee JY, Cohen T, Gorivodsky M, Tzchori I, Zhao Y, Hayes SM, Bresnick EH, Zhao K, Westphal H, Love PE. 2010. Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat Immunol* **12**(2):129–136.

Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J, Deng XW. 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* **38**(1):124–129.

Li M, He Y, Dubois W, Wu X, Shi J, Huang J. 2012. Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. *Mol Cell* **46**(1):30–42

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**(6038):53–58.

Li Q, Brown J, Huang H, Bickel P. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**:1752-1779.

Li T, Spearow J, Rubin CM, Schmid CW. 1999. Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. *Gene* **239**(2):367–372.

Li W, Feng J, Jiang T. 2011. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J Comput Biol* **8**(11):1693–1707.

Li W, Jiang T. 2012. Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformatics* **28**(22):2914–2921.

Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X, Oh S, Kim HS, Glass CK, Rosenfeld MG. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**(7455):516–520.

Li W, Teng F, Li T, Zhou Q. 2013. Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. *Nat Biotechnol* **31**(8):684–686.

Li W, Yang W, Wang XJ. 2013. Pseudogenes: pseudo or real functional elements? *J Genet Genomics* **40**(4):171–177.

Li XF, Lytton J. 1999. A circularized sodium-calcium exchanger exon 2 transcript. *J Biol Chem* **274**(12):8153–8160.

Li Y, Chien J, Smith DI, Ma J. 2011. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* **27**(12):1708–1710.

Li Y, Li-Byarlay H, Burns P, Borodovsky M, Robinson GE, Ma J. 2013. TrueSight: a new algorithm for splice junction detection using RNA-seq. *Nucleic Acids Res* **41**(4):e51

Liang K, Keleş S. 2012. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13**:199.

Liao P, Yong TF, Liang MC, Yue DT, Soong TW. 2005. Splicing for alternative structures of Cav1.2 Ca2^{+} channels in cardiac and smooth muscles. *Cardiovasc Res* **68**(2):197–203.

Liao Q, Shen J, Liu J, Sun X, Zhao G, Chang Y, Xu L, Li X, Zhao Y, Zheng H, Zhao Y, Wu Z. 2014. Genome-wide identification and functional annotation of *Plasmodium falciparum* long noncoding RNAs from RNA-seq data. *Parasitol Res* **113**(4):1269–1281.

Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**(10):e108.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**(7221):464–469.

Lidie KB, van Dolah FM. 2007. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. *J Eukaryot Microbiol* **54**(5):427–435.

Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**:327-334.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950):289–293.

Lifton RP, Goldberg ML, Karp RW, Hogness DS. 1978. The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications. *Cold Spring Harb Symp Quant Biol* **42 Pt 2**:1047–1051.

Lill R, Hoffmann B, Molik S, Pierik AJ, Rietzschel N, Stehling O, Uzarska MA, Webert H, Wilbrecht C, Mühlenhoff U. 2012. The role of mitochondria in cellular iron-sulfur protein biogenesis and iron metabolism. *Biochim Biophys Acta* **1823**(9):1491-1508.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. *Science* **299**:1540.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17**:991-1008.

Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA. 2012. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**(1):56–67.

Lin H. 2007. piRNAs in the germ line. *Science* **316**(5823):397.

Lin H, Spradling A. 1997. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* **124**:2463-2476.

Lin H, Yin H. 2008. A novel epigenetic mechanism in *Drosophila* somatic cells mediated by Piwi and piRNAs. *Cold Spring Harb Symp Quant Biol* **73**:273-281.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**(13):i275–282.

Lin S, Zhang H, Zhuang Y, Tran B, Gill J. 2010. Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proc Natl Acad Sci U S A* **107**(46):20033–20038.

Lin X, Tirichine L, Bowler C. 2012. Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. *Plant Methods* **8**(1):48.

Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, Mansson R, Sigvardsson M, Hagman J, Espinoza CA, Dutkowski J, Ideker T, Glass CK, Murre C. 2010. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat Immunol* **11**(7):635–643.

Lundblad JR, Kwok RP, Laurance ME, Harter ML, Goodman RH. 1995. Adenoviral E1A-associated protein p300 as a functional homologue of the transcriptional co-activator CBP. *Nature* **374**(6517):85–88.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC; Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK; Genome Institute at Washington University, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**(7370):476–482.

Lindmark DG, Müller M. 1973. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate *Tritrichomonas foetus*, and its role in pyruvate metabolism. *J Biol Chem* **248**(22):7724–7728.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, Cuppen E. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**(7):474–476.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**(3):523–536.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**(7271):315–322.

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM, Ecker JR. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**(7336):68–73.

Litonin D, Sologub M, Shi Y, Savkina M, Anikin M, Falkenberg M, Gustafsson CM, Temiakov

D. 2010. Human mitochondrial transcription revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro. *J Biol Chem* **285**(24):18129–18133.

Little GH, Noushmehr H, Baniwal SK, Berman BP, Coetzee GA, Frenkel B. 2011. Genome-wide Runx2 occupancy in prostate cancer cells suggests a role in regulating secretion. *Nucleic Acids Res* **40**(8):3538–3547.

Livak KJ, Wills QF, Tipping AJ, Datta K, Mittal R, Goldson AJ, Sexton DW, Holmes CC. 2013. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* **59**(1):71–79.

Liu G, Mattick JS, Taft RJ. 2013. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* **12**:127–138.

Liu G, Mercer TR, Shearwood AM, Siira SJ, Hibbs ME, Mattick JS, Rackham O, Filipovska A. 2013. Mapping of mitochondrial RNA-protein interactions by digital RNase footprinting. *Cell Rep* **5**(3):839–848.

Liu W, Tanasa B, Tyurina OV, Zhou TY, Gassmann R, Liu WT, Ohgi KA, Benner C, Garcia-Bassets I, Aggarwal AK, Desai A, Dorrestein PC, Glass CK, Rosenfeld MG. 2010. PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* **466**(7305):508–512.

Liu WM, Chu WM, Choudary PV, Schmid CW. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23**(10):1758–1765.

Liu XY, Wei B, Shi HX, Shan YF, Wang C. 2010. Tom70 mediates activation of interferon regulatory factor 3 on mitochondria. *Cell Res* **20**(9):994–1011.

Liu Z, Scannell DR, Eisen MB, Tjian R. 2011. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* **146**(5):720–731.

Lo KA, Bauchmann MK, Baumann AP, Donahue CJ, Thiede MA, Hayes LS, des Etages SA, Fraenkel E. 2011. Genome-wide profiling of H3K56 acetylation and transcription factor binding sites in human adipocytes. *PLoS ONE* **6**(6):e19778.

Lo TW, Pickle CS, Lin S, Ralston EJ, Gurling M, Schartner CM, Bian Q, Doudna JA, Meyer BJ. 2013. Precise and heritable genome editing in evolutionarily diverse nematodes using TALENs and CRISPR/Cas9 to engineer insertions and deletions. *Genetics* **195**(2):331–348.

Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**(12):1113–1117.

Löffelhardt W. 2011. The chlorarachniophyte nucleomorph is supplemented with host cell nucleus-encoded histones. *Mol Microbiol* **80**(6):1413–1416.

Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CW, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**(4):431–440.

Loh YH, Yang L, Yang JC, Li H, Collins JJ, Daley GQ. 2011. Genomic approaches to deconstruct pluripotency. *Annu Rev Genomics Hum Genet* **12**:165–185.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, Jørgensen T, Jiang T, Witte DR, Sandbæk A, Hellmann I, Lauritzen T, Hansen T, Pedersen O, Wang J, Nielsen R. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**(10):e1002326.

Lou SK, Ni B, Lo LY, Tsui SK, Chan TF, Leung KS. 2011. ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics* **27**(3):421–422.

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. 2012. Revisiting global gene expression analysis. *Cell* **151**(3):476–482.

Lu F, Tsai K, Chen HS, Wikramasinghe P, Davuluri RV, Showe L, Domsic J, Marmorstein R, Lieberman PM. 2012. Identification of host-chromosome binding sites and candidate gene targets for Kaposi's sarcoma-associated herpesvirus LANA. *J Virol* **86**(10):5752–5762.

Lubeck E, Cai L. 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* **9**(7):743–748.

Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA, Arkin AP. 2011. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A* **108**(27):11063–11068.

Luconi M, Mannelli M. 2012. Xenograft models for preclinical drug testing: implications for adrenocortical cancer. Mol Cell Endocrinol 351(1):71–77.

Ludwig M, Gibbs SP. 1985. DNA is present in the nucleomorph of cryptomonads: further evidence that the chloroplast evolved from a eukaryotic endosymbiont. *Protoplasma* **127**:9-20

Ludwig M, Gibbs SP. 1989. Evidence that the nucleomorphs of *Chlorarachnion reptans* (Chlorarachniophyta) are vestigial nuclei: morphology, division and DNA-DAPI fluorescence. *J Phycol* **25**:385-394.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a *cis*-regulatory module. *PLoS Biol* **3**(4):e93.

Lukes J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW. 2011. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63**:528–537.

Lukes J, Leander BS, Keeling PJ. 2009. Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. *Proc Natl Acad Sci U S A* **106**(Suppl 1):9963—9970.

Lun DS, Sherrid A, Weiner B, Sherman DR, Galagan JE. 2009. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol* **10**(12):R142.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**(1):18.

Lynch KW, Maniatis T. 1996. Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila doublesex* splicing enhancer. *Genes Dev* **10**(16):2089–2101.

Lynch M. 2002. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A* **99**:6118-6123.

Lynch M. 2006a. Streamlining and Simplification of Microbial Genome Architecture. *Annu Rev Microbiol* **60**:327-349.

Lynch M. 2006b. The Origins of Eukaryotic Gene Structure. *Mol Biol Evol* **23**(2):450-468

Lynch M. 2007a. The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* **8**:803–813.

Lynch M. 2007b. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* **104**(Suppl 1):8597-8604.

Lynch M. 2007c. The origins of genome architecture. *Sunderland (MA):Sinauer Associates*

Lynch M. 2010a. Evolution of the mutation rate. *Trends Genet* **26**(8):345–352.

Lynch M. 2010b. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**:961–968.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science.* **302**:1401–1404.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science* **311**(5768):1727–1730.

Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. *Mol Biol Evol* **22**:1137–1146.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**(11):1154–1159.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**:372–373.

Ma Z, Swigut T, Valouev A, Rada-Iglesias A, Wysocka J. 2010. Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat Struct Mol Biol* **18**(2):120–127.

Maassen JA, 'T Hart LM, Van Essen E, Heine RJ, Nijpels G, Jahangir Tafrechi RS, Raap AK, Janssen GM, Lemkes HH. 2004. Mitochondrial diabetes: molecular mechanisms and clinical presentation. *Diabetes* **53**Suppl1:S103–109.

MacIsaac KD, Lo KA, Gordon W, Motola S, Mazor T, Fraenkel E. 2010. A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLoS Comput Biol* **6**(4):e1000773.

Magnúsdóttir E, Dietmann S, Murakami K, Gnesdogan U, Tang F, Bao S, Diamanti E, Lao K, Gottgens B, Azim Surani M. 2013. A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nat Cell Biol* **15**(8):905–915.

Magnúsdóttir E, Gillich A, Grabole N, Surani MA. 2012. Combinatorial control of cell fate and reprogramming in the mammalian germline. *Curr Opin Genet Dev* **22**(5):466–474

Mahony S, Mazzoni EO, McCuine S, Young RA, Wichterle H, Gifford DK. 2010. Ligand-dependent dynamics of retinoic acid receptor binding during early neurogenesis. *Genome Biol* **12**(1):R2.

Maier RM, Neckermann K, Igloi GL, Kössel H. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J Mol Biol* **251**(5):614–628.

Maison C, Almouzni G. 2004. HP1 and the dynamics of heterochromatin maintenance. *Nat Rev Mol Cell Biol* **5**:296-304.

Makalowski W. 2003. Not junk after all. *Science* **300**:1246–1247.

Malceva N, Belyaeva E, King R, Zhimulev I. 1997. Nurse cell polytene chromosomes of *Drosophila melanogaster otu* mutants: Morphological changes accompanying interallelic complementation and position effect variegation. *Dev Genetic* **20**:163-174.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**(6121):823–826.

Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, Cheng J, Bell I, Ghosh S, Piccolboni A, Gingeras TR. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat Genet* **38**(10):1151–1158.

Mandava V, Janzen CJ, Cross GA. 2008. Trypanosome H2Bv replaces H2B in nucleosomes enriched for H3 K4 and K76 trimethylation. *Biochem Biophys Res Commun* **368**(4):846–851.

Mangul S, Caciula A, Glebova O, MandBoiu I, Zelikovsky A. 2012. Improved transcriptome quantification and reconstruction from RNA-Seq reads using partial annotations. *In Silico Biol* **11**(5):251–261.

Maniatis T, Ptashne M. 1973. Structure of the lambda operators. *Nature* **246**(5429):133–136.

Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH. 2012. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat Biotechnol* **30**(4):349–353.

Manrao EA, Derrington IM, Pavlenok M, Niederweis M, Gundlach JH. 2011. Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS ONE* **6**(10):e25723.

Marande W, Burger G. 2007. Mitochondrial DNA as a genomic jigsaw puzzle. *Science* **318**:415

Marande W, Lukes J, Burger G. 2005. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryot Cell* **4**(6):1137–1146.

Marban C, Su T, Ferrari R, Li B, Vatakis D, Pellegrini M, Zack JA, Rohr O, Kurdistani SK. 2011. Genome-wide binding map of the HIV-1 Tat protein to the human genome. *PLoS ONE* **6**(11):e26894.

Marchenko ND, Wolff S, Erster S, Becker K, Moll UM. 2007. Monoubiquitylation promotes mitochondrial p53 translocation. *EMBO J* **26**(4):923–934.

Marchenko ND, Zaika A, Moll UM. 2000. Death signal-induced localization of p53 protein to mitochondria. A potential role in apoptotic signaling. *J Biol Chem* **275**(21):16202–16212.

Marcker K, Sanger F. 1964. N-Formyl-Methionyl-S-RNA. *J Mol Biol* **8**:835–840.

Marco-Sola S, Sammeth M, Guigó R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**(12):1185–1188.

Mardis ER. 2010. Cancer genomics identifies determinants of tumor biology. *Genome Biol* **11**(5):211.

Margulies EH, Blanchette M; NISC Comparative Sequencing Program, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* **13**(12):2507–2518.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057):376–380.

Mariner PD, Walters RD, Espinoza CA, Drullinger LF, Wagner SD, Kugel JF, Goodrich JA. 2008. Human *Alu* RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol Cell* **29**(4):499–509.

Marinov GK, Kundaje A, Park PJ, Wold BJ. 2014. Large-Scale Quality Analysis of Published ChIP-seq Data. *G3 (Bethesda)* **4**(2):209–223.

Marinov GK, Wang YE, Chan D, Wold BJ. 2014. Evidence for site-specific occupancy of the mitochondrial genome by nuclear transcription factors. *PLoS One* **9**(1):e84713.

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. 2014. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res* **24**(3):496–510.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**(9):1509–1517.

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp PA, Bartel DP, Jaenisch R, Young RA. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**(3):521–533.

Martens PA, Clayton DA. 1979. Mechanism of mitochondrial DNA replication in mouse L-cells: localization and sequence of the light-strand origin of replication. *J Mol Biol* **135**(2):327–351.

Martens JA, Laprade L, Winston F. 2004. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429**(6991):571–574.

Martin D, Pantoja C, Fernández Miñán A, Valdes-Quezada C, Moltó E, Matesanz F, Bogdanovic O, de la Calle-Mustienes E, Domínguez O, Taher L, Furlan-Magaril M, Alcina A, Cañón S, Fedetz M, Blasco MA, Pereira PS, Ovcharenko I, Recillas-Targa F, Montoliu L, Manzanares M, Guigó R, Serrano M, Casares F, Gómez-Skarmeta JL. 2011. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol* **18**(6):708–714.

Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z. 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**:663.

Martin M, Cho J, Cesare AJ, Griffith JD, Attardi G. 2005. Termination factor-mediated DNA loop between termination and initiation sites drives mitochondrial rRNA synthesis. *Cell* **123**(7):1227–1240.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**(7080):41–45.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A* **99**(19):12246–12251.

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**:162-165.

Martinez P, Thanasoula M, Carlos AR, Gómez-López G, Tejera AM, Schoeftner S, Dominguez O, Pisano DG, Tarsounas M, Blasco MA. 2010. Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites. *Nat Cell Biol* **12**(8):768–780.

Martínez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martnez LE, Manning-Cela RG, Figueroa-Angulo EE. 2010. Gene expression in trypanosomatid parasites. *J Biomed*

Biotechnol **2010**:525241.

Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond JR, Strelets VB, Wilson RJ and the FlyBase Consortium. 2013. FlyBase: improvements to the bibliography. *Nucleic Acids Res* **41**(D1):D751–D757.

Masters BS, Stohl LL, Clayton DA. 1987. Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell* **51**(1):89–99.

Mattick JS. 1994. Introns: evolution and function. *Curr Opin Genet Dev* 4(6):823–831.

Mattick JS. 2004. RNA regulation: a new genetics? *Nat Rev Genet* **5**(4):316–323.

Mattick JS. 2007. A new paradigm for developmental biology. *J Exp Biol* **210**:1526–1547.

Mattick JS. 2009. Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. *Ann N Y Acad Sci* **1178**:29–46.

Mattick JS. 2011. The central role of RNA in human development and cognition. *FEBS Lett* **585**(11):1600–1616.

Mattick JS, Dinger ME. 2013. The extent of functionality in the human genome. *The HUGO Journal* **7**:2

Mattick JS, Gagen MJ. 2001. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol* **18**(9):1611–1630.

Mattick JS, Taft RJ, Faulkner GJ. 2010. A global view of genomic information–moving beyond the gene and the master regulator. *Trends Genet* **26**(1):21–28.

Matzke M, Aufsatz W, Kanno T, Daxinger L, Papp I, Mette M, Matzke A. 2004. Genetic analysis of RNA-mediated transcriptional gene silencing. *Biochim Biophys Acta* **1677**:129-141.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**(6099):1190–1195.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**(2):560-564

May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A. 2011. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**(1):89—93.

Mayr E. 1983. How to carry out the adaptationist program? *Am Nat* **121**:324-334.

Mazzoni EO, Mahony S, Iacovino M, Morrison CA, Mountoufaris G, Closser M, Whyte WA, Young RA, Kyba M, Gifford DK, Wichterle H. 2011. Embryonic stem cell-based mapping of developmental transcriptional programs. *Nat Methods* **8**(12):1056–1058.

McCarrey JR, Riggs AD. 1986. Determinator-inhibitor pairs as a mechanism for threshold setting in development: A possible function for pseudogenes. *Proc Natl Acad Sci U S A* **83**:679-683.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**(6):344–355.

McClintock B. 1951. Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol* **16**:1347

McClintock B. 1953. Induction of instability at selected loci in maize. *Genetics* **38**:579-599.

McClintock B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21**:197–216.

McEwan M, Humayun R, Slamovits CH, Keeling PJ. 2008. Nuclear genome sequence survey of the dinoflagellate *Heterocapsa triquetra*. *J Euk Microbiol* **55**:530-535.

McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res* **18**(2):252–260.

McGlincy NJ, Smith CW. 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci* **33**(8):385–393.

McHugh CA, Russell P, Guttman M. 2014. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol* **15**(1):203.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**(9):1527–1541.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**:816-825.

McManus S, Ebert A, Salvagiotto G, Medvedovic J, Sun Q, Tamir I, Jaritz M, Tagoh H, Busslinger M. 2011. The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. *EMBO J* **30**(12):2388–2404.

McNicoll F, Drummelsmith J, Müller M, Madore E, Boilard N, Ouellette M, Papadopoulou B. 2006. A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. *Proteomics* **6**(12):3567–3581.

McQuilton P, St. Pierre SE, Thurmond J, The FlyBase Consortium. 2012. FlyBase 101 the basics of navigating FlyBase. *Nucleic Acids Res* **40(Database issue)**:D706–714.

McPherson A, Wu C, Hajirasouliha I, Hormozdiari F, Hach F, Lapuk A, Volik S, Shah S, Collins C, Sahinalp SC. 2011. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics* **27**(11):1481–1488.

Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20**:1335–1343.

Medstrand P, Landry JR, Mager DL. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* **276**(3):1896–1903.

Mehler MF, Mattick JS. 2006. Non-coding RNAs in the nervous system. *J Physiol* **575**:333-341.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**(7205):766–770.

Melamud E, Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res* **37**(14):4873–4886.

Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL. 1997. *roX1* RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. *Cell* **88**(4):445–457.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**(3):271–277.

Melo CA, Drost J, Wijchers PJ, van de Werken H, de Wit E, Oude Vrielink JA, Elkon R, Melo SA, Léveillé N, Kalluri R, de Laat W, Agami R. 2013. eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* **49**(3):524–535.

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N. 2013. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**(7441):333–338.

Mendel G. 1866. Versuche ber Pflanzen-Hybriden. *Verh Naturforsch Vereines Abhandlungen Brünn* **4**:3-47.

Mendoza-Parra MA, Walia M, Sankar M, Gronemeyer H. 2011. Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Mol Syst Biol* **7**:538.

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**(2):716–721.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. 2011. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30**(1):99–104.

Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood AM, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS. 2011b. The human mitochondrial transcriptome. *Cell*

146(4):645–658.

Merelo P, Xie Y, Brand L, Ott F, Weigel D, Bowman JL, Heisler MG, Wenkel S. 2013. Genome-wide identification of KANADI1 target genes. *PLoS One* **8**(10):e77341.

Mereschkowsky K. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol Centralbl* **25**:593–604.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**(6114):1593–1599.

Metodiev MD, Lesko N, Park CB, Cámara Y, Shi Y, Wibom R, Hultenby K, Gustafsson CM, Larsson NG. 2009. Methylation of 12S rRNA is necessary for in vivo stability of the small subunit of the mammalian mitochondrial ribosome. *Cell Metab* **9**(4):386–397.

Mette M, Aufsatz W, van der Winden J, Matzke M, Matzke A. 2000. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J* **19**:5194-5201.

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**(7):1635–1646.

Meyer MB, Goetsch PD, Pike JW. 2012. VDR/RXR and TCF4/β-catenin cistromes in colonic cells of colorectal tumor origin: impact on c-FOS and c-MYC gene expression. *Mol Endocrinol* **26**(1):37–51.

Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, Shah S, Aparicio S, Chiang DY, Goldenberg A, Brudno M. 2013. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res* **23**(3):519–529.

Micol V, Fernández-Silva P, Attardi G. 1997. Functional analysis of in vivo and in organello footprinting of HeLa cell mitochondrial DNA in relationship to ATP and ethidium bromide effects on transcription. *J Biol Chem* **272**(30):18896–18904.

Miescher F. 1871. Ueber die chemische Zusammensetzung der Eiterzellen. *Medicinisch-chemische Untersuchungen* **4**:441-460.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153):553–560.

Miller JD, Scott EC, Okamoto S. 2006. Science communication. Public acceptance of evolution. *Science* **313**(5788):765–766.

Miller TW, Balko JM, Fox EM, Ghazoui Z, Dunbier A, Anderson H, Dowsett M, Jiang A, Smith RA, Maira SM, Manning HC, González-Angulo AM, Mills GB, Higham C, Chanthaphaychith S, Kuba MG, Miller WR, Shyr Y, Arteaga CL. 2011. ERα-dependent E2F transcription can mediate resistance to estrogen deprivation in human breast cancer. *Cancer Discov* **1**(4):338–351.

Miller WL. 2011. Role of mitochondria in steroidogenesis. *Endocr Dev* **20**:1-19.

Minning TA, Weatherly DB, Atwood J 3rd, Orlando R, Tarleton RL. 2009. The steady-state transcriptome of the four major life-cycle stages of *Trypanosoma cruzi*. *BMC Genomics* **10**:370.

Miranda TB, Jones PA. 2007. DNA methylation: the nuts and bolts of repression. *J Cell Physiol* **213**(2):384-390.

Mironov AA, Fickett JW, Gelfand MS. 1999. Frequent alternative splicing of human genes. *Genome Res* **9**(12):1288–1293.

Mirsky AE, Ris H. 1951. The desoxyribonucleic acid content of animal cells and its evolutionary significance.. *J Gen Physiol* **34**(4):451–462.

Mittal N, Zavolan M. 2014.Seq and CLIP through the miRNA world. *Genome Biol* **15**(1):202.

Miura F, Enomoto Y, Dairiki R, Ito T. 2012. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* **40**(17):e136.

Miyamoto H, Machida RJ, Nishida S. 2010. Complete mitochondrial genome sequences of the three pelagic chaetognaths *Sagitta nagae*, *Sagitta decipiens* and *Sagitta enflata*. *Comp Biochem Physiol Part D Genomics Proteomics* **5**:65-72.

Miyazaki M, Rivera RR, Miyazaki K, Lin YC, Agata Y, Murre C. 2011. The opposing roles of the transcription factor E2A and its antagonist Id3 that orchestrate and enforce the naive fate of T cells. *Nat Immunol* **12**(10):992–1001.

Mizuno T, Chou MY, Inouye M. 1984. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc Natl Acad Sci U S A* **81**(7):1966–1970.

Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. 2002. Analysis of a *piwi*-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* **110**:689-699.

Mochizuki K, Gorovsky MA. 2004. Small RNAs in genome rearrangement in *Tetrahymena*. *Curr Opin Genet Dev* **14**:181-187.

Mochizuki K, Gorovsky MA. 2005. A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev* 19:77-89.

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SC, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**(6012):1787–1797.

Monje P, Boland R. 2001. Subcellular distribution of native estrogen receptor α and β isoforms in rabbit uterus and ovary. *J Cell Biochem* **82**(3):467–479.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**(7289):773–777.

Montoya J, Christianson T, Levens D, Rabinowitz M, Attardi G. 1982. Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. *Proc Natl Acad Sci* **79**(23):7195–7199.

Montoya J, Gaines GL, Attardi G. 1983. The pattern of transcription of the human mitochondrial rRNA genes reveals two overlapping transcription units. *Cell* **34**(1):151-159.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. *Annu Rev Genet* **43**:251-264.

Moore RB, Oborník M, Janouskovec J, Chrudimský T, Vancová M, Green DH, Wright SW, Davies NW, Bolch CJ, Heimann K, Slapeta J, Hoegh-Guldberg O, Logsdon JM, Carter DA. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* **451**(7181):959–963.

Morgan TH, Sturtevant AH, Muller HJ, Bridges CB. 1915. The Mechanism of Mendelian Heredity. *Henry Holt, New York*

Morohashi K, Casas MI, Falcone Ferreyra ML, Mejía-Guerra MK, Pourcel L, Yilmaz A, Feller A, Carvalho B, Emiliani J, Rodriguez E, Pellegrinet S, McMullen M, Casati P, Grotewold E. 2012. A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. *Plant Cell* **24**(7):2745–2764.

Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon JE, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML. 2007. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**(5846):1921–1926.

Mortazavi A, Leeper Thompson EC, Garcia ST, Myers RM, Wold B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Res* **16**(10):1208–1221.

Mortazavi A, Pepke S, Jansen C, Marinov GK, Ernst J, Kellis M, Hardison RC, Myers RM, Wold BJ. 2013. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res* **23**(12):2136–2148.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7):621–628.

Moshkovich N, Lei E. 2010. HP1 recruitment in the absence of argonaute proteins in *Drosophila*. *PLoS Genetics* **6**:e1000880.

Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**(2):459–472.

Mousavi K, Zare H, Dell'orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, Hager GL, Sartorelli V. 2013. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell* **51**(5):606–617.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G, Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE, Weissman SM, Cayting P, Hariharan M, Lian J, Cheng Y, Landt SG, Ma Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C, Kumar SA, Mishra T, Jain D, Byrska-Bishop M, Blankenberg D, Lajoie1 BR, Jain G, Sanyal A, Chen KB, Denas O, Taylor J, Blobel GA, Weiss MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI, Desalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L, McCleary D, Kuan S, Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S, Xue C, Dobin A, Lin W, Fastuca M, Wang H, Guigo R, Djebali S, Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM, Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin M, Lowdon RF, Adams LB. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**(8):418.

The mouse ENCODE Consortium, Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman R, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Santos MR, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, ichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Kellis M, Mortazavi A, Weissman SM, Stamatoyannopoulos J, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B. 2014. An Integrated and Comparative Encyclopedia of DNA Elements in the Mouse Genome. *in review*

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger

JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915):520–562.

Moyroud E, Minguet EG, Ott F, Yant L, Posé D, Monniaux M, Blanchet S, Bastien O, Thévenon E, Weigel D, Schmid M, Parcy F. 2011. Prediction of regulatory interactions from genome sequences using a biophysical model for the *Arabidopsis* LEAFY transcription factor. *Plant Cell* **23**(4):1293–1306.

Mruk I, Kobayashi I. 2014. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res* **42**(1):70–86.

Muiño JM, Kaufmann K, van Ham RC, Angenent GC, Krajewski P. 2011. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **7**:11.

Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA. 2011. Master transcription factors determine cell-type-specific responses to TGFβ signaling. *Cell* **147**(3):565–576.

Müller SB, Rensing SA, Maier UG. 1994. The cryptomonad histone H4-encoding gene: structure and chromosomal localization. *Gene* **150**(2):299–302.

Mullican SE, Gaddis CA, Alenghat T, Nair MG, Giacomin PR, Everett LJ, Feng D, Steger DJ, Schug J, Artis D, Lazar MA. 2011. Histone deacetylase 3 is an epigenomic brake in macrophage alternative activation. *Genes Dev* **25**(23):2480–2488.

Muerdter F, Guzzardo PM, Gillis J, Luo Y, Yu Y, Chen C, Fekete R, Hannon GJ. 2013. A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in *Drosophila*. *Mol Cell* **50**(5):736–748.

Muerdter F, Olovnikov I, Molaro A, Rozhkov N, Czech B, Gordon A, Hannon G, Aravin A. 2012. Production of artificial piRNAs in flies and mice. *RNA* **18**:42-52.

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**(7044):903–910.

Muro EM, Mah N, Andrade-Navarro MA. 2011. Functional evidence of post-transcriptional regulation by pseudogenes. *Biochimie* **93**(11):1916–1921.

Murray BG, Leitch IJ, Bennett MD. 2012. Gymnosperm DNA C-values database (release 5.0, Dec. 2012) `http://www.kew.org/cvalues/`

Murre C. 2005. Helix-loop-helix proteins and lymphocyte development. *Nat Immunol* **6**(11):1079–1086.

Murre C, McCaw PS, Baltimore D. 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell* **56**(5):777–783.

Murre C, McCaw PS, Vaessin H, Caudy M, Jan LY, Jan YN, Cabrera CV, Buskin JN, Hauschka SD, Lassar AB, Wientraub H, Baltimore D. 1989. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* **58**(3):537–544.

Muzikar KA, Nickols NG, Dervan PB. 2009. Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression. *Proc Natl Acad Sci U S A* **106**(39):16598–16603.

Näär AM, Lemon BD, Tjian R. 2001. Transcriptional coactivator complexes. *Annu Rev*

Biochem **70**:475–501.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881):1344–1349.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**(7267):1130–1134.

Nakayama J, Rice J, Strahl B, Allis C, Grewal S. 2001. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**:110-113.

Nakayamada S, Kanno Y, Takahashi H, Jankovic D, Lu KT, Johnson TA, Sun HW, Vahedi G, Hakim O, Handon R, Schwartzberg PL, Hager GL, O'Shea JJ. 2011. Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* **35**(6):919–931.

Nash EA, Barbrook AC, Edwards-Stuart RK, Bernhardt K, Howe CJ, Nisbet RE. 2007. Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae. Mol Biol Evol* **24**(7):1528–1536.

Nass S, Nass MM, Hennix U. 1965. Deoxyribonucleic acid in isolated rat-liver mitochondria. *Biochim Biophys Acta* **95**:426-435.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martnez-Garca PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, Dejong PJ, Yorke JA, Salzberg SL, Langley CH. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* **15**(3):R59.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**(7485):635–640.

Négre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, Venken K, Bellen H, White R, Gerstein M, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. 2011. A *cis*-regulatory map of the *Drosophila* genome. *Nature* **471**(7339):527–531. doi: 10.1038/nature09990.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* **11**:265-289.

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**(7232):1038–1042.

Nekrasov V, Staskawicz B, Weigel D, Jones JD, Kamoun S. 2013. Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31**(8):691–693.

Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. 2012b. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**(6):1274–1286.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**(7414):83–90.

Ngo HB, Kaiser JT, Chan DC. 2011. The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nat Struct Mol Biol* **18**(11):1290–1296.

Nguyen TC, Deng N, Xu, G, Duan, Z, Zhu, D. 2011. iQuant: A fast yet accurate GUI tool for transcript quantification. *In BIBM Workshops* 1048-1050.

Nguyen TC, Deng N, Zhu D. 2013. SASeq: A Selective and Adaptive Shrinkage Approach to Detect and Quantify Active Transcripts using RNA-Seq. *arXiv*:1208.3619v2

Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7**(7):521–527.

Ni X, Zhang YE, Nègre N, Chen S, Long M, White KP. 2012. Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLoS Biol* **10**(11):e1001420.

Nass S, Nass MM, Hennix U. 1965. Deoxyribonucleic acid in isolated rat-liver mitochondria. *Biochim Biophys Acta* **95**:426-435.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**(4):e1000888.

Nickols NG, Dervan PB. 2007. Suppression of androgen receptor-mediated gene expression by a sequence-specific DNA-binding polyamide. *Proc Natl Acad Sci U S A* **104**(25):10418–10423.

Nicolae M, Mangul S, Mandoiu II, Zelikovsky A. 2011. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol Biol* **6**(1):9.

NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. 2009. The NIH Human Microbiome Project. *Genome Res* **19**(12):2317–2323.

Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T. 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. *PLoS Pathog* **6**(8):e1001037.

Nirenberg M, Leder P. 1964. RNA Codewords and Protein Synthesis. The Effect of Trinucleotides Upon The Binding Of sRNA to Risbosomes. *Science* **145**(3639):1399–1407.

Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* **47**:1588–1602.

Nisbet RE, Hiller RG, Barry ER, Skene P, Barbrook AC, Howe CJ. 2008. Transcript analysis of dinoflagellate plastid gene minicircles. *Protist* **159**(1):31–39.

Nishioka Y, Leder A, Leder P. 1980. Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. *Proc Natl Acad Sci U S A* **77**(5):2806–2809.

Nishiyama A, Xin L, Sharov AA, Thomas M, Mowrer G, Meyers E, Piao Y, Mehta S, Yee S, Nakatake Y, Stagg C, Sharova L, Correa-Cerro LS, Bassey U, Hoang H, Kim E, Tapnio R, Qian Y, Dudekula D, Zalzman M, Li M, Falco G, Yang HT, Lee SL, Monti M, Stanghellini I, Islam MN, Nagaraja R, Goldberg I, Wang W, Longo DL, Schlessinger D, Ko MS. 2009. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell* **5**(4):420–433.

Nitzsche A, Paszkowski-Rogacz M, Matarese F, Janssen-Megens EM, Hubner NC, Schulz H, de Vries I, Ding L, Huebner N, Mann M, Stunnenberg HG, Buchholz F. 2011. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE* **6**(5):e19470.

Niu DK, Jiang L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun* **430**(4):1340–1343.

Nix DA, Courdy SJ, Boucher KM. 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**:523.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302**(5644):413

Norman JE, Gray MW. 2001. A complex organization of the gene encoding cytochrome oxidase subunit 1 in the mitochondrial genome of the dinoflagellate, *Crypthecodinium cohnii*: homologous recombination generates two different *cox1* open reading frames. *J Mol Evol* **53**(4-5):351–363.

Norris J, Fan D, Aleman C, Marks JR, Futreal PA, Wiseman RW, Iglehart JD, Deininger PL, McDonnell DP. 1995. Identification of a new subclass of *Alu* DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J Biol Chem* **270**(39):22777–22782.

Norton L, Fourcaudot M, Abdul-Ghani MA, Winnier D, Mehta FF, Jenkinson CP, Defronzo RA. 2011. Chromatin occupancy of transcription factor 7-like 2 (TCF7L2) and its role in hepatic glucose metabolism. *Diabetologia* **54**(12):3132–3142.

Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T, Frampton GM, Drake AC, Leskov I, Nilsson B, Preffer F, Dombkowski D, Evans JW, Liefeld T, Smutko JS, Chen J, Friedman N, Young RA, Golub TR, Regev A, Ebert BL. 2011. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**(2):296–309.

Nowacki M, Haye JE, Fang W, Vijayan V, Landweber LF. 2010. RNA-mediated epigenetic regulation of DNA copy number. *Proc Natl Acad Sci U S A* **107**:22140-22144.

Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, Landweber LF. 2009. A functional role for transposases in a large eukaryotic genome. *Science* **324**(5929):935–938.

Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, Landweber LF. 2008. RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* **451**(7175):153–158.

Ntziachristos P, Tsirigos A, Van Vlierberghe P, Nedjic J, Trimarchi T, Flaherty MS, Ferres-Marco D, da Ros V, Tang Z, Siegle J, Asp P, Hadler M, Rigo I, De Keersmaecker K, Patel J, Huynh T, Utro F, Poglio S, Samon JB, Paietta E, Racevskis J, Rowe JM, Rabadan R, Levine RL, Brown S, Pflumio F, Dominguez M, Ferrando A, Aifantis I. 2012. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med* **18**(2):298–301.

Nugent PG, Karsani SA, Wait R, Tempero J, Smith DF. 2004. Proteomic analysis of *Leishmania mexicana* differentiation. *Mol Biochem Parasitol* **136**(1):51–62.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson A, Rilakovic N, Ritland C, Rosselló JA, Sena J, Svensson T, Talavera-López C, Theiβen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**(7451):579–584.

Oborník M, Janouskovec J, Chrudimský T, Lukes J. 2009. Evolution of the apicoplast and its hosts: from heterotrophy to autotrophy and back again. *Int J Parasitol* **39**(1):1–12.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**(6):730–732.

Ogita K, Okuda H, Kitano M, Fujinami Y, Ozaki K, Yoneda Y. 2002. Localization of activator protein-1 complex with DNA binding activity in mitochondria of murine brain after in vivo treatment with kainate. *J Neurosci* **22**(7):2561–2570.

Ogita K, Fujinami Y, Kitano M, Yoneda Y. 2003. Transcription factor activator protein-1 expressed by kainate treatment can bind to the non-coding region of mitochondrial genome in murine hippocampus. *J Neurosci Res* **73**(6):794–802.

Ogryzko VV, Schiltz RL, Russanova V, Howard BH, Nakatani Y. 1996. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**(5):953–959.

Ohno S. 1972. So much "junk" DNA in our genome. *Brookhaven Symp Biol* **23**:366–370.

Ohno S, Atkin NB. 1966. Comparative DNA values and chromosome complements of eight species of fishes. *Chromosoma* **18**(3):455–466.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246**:96-98

Ojala D, Montoya J, Attardi G. 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**(5806):470–474.

Okano M, Xie S, Li E. 1998. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**:219-220.

Okano M, Bell DW, Haber DA, Li E. 19999. DNA methyltransferases Dnmt3a and

Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**(3):247–257.

Oldmeadow C, Mengersen K, Mattick JS, Keith JM. 2010. Multiple evolutionary rate classes in animal genome evolution. *Mol Biol Evol* **27**:942–953.

Olins AL, Olins DE. 1974. Spheroid chromatin units (v bodies). *Science* **183**(4122):330–332.

Olivieri D, Sykora M, Sachidanandam R, Mechtler K, Brennecke J. 2010. An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in *Drosophila*. *EMBO J* **29**:3301-3317.

Olmo E. 1973. Quantitative variations in the nuclear DNA and phylogenesis of the Amphibia. *Caryologia* **26**:43–68.

Olmo E, Morescalchi A. 1978. Genome and cell size in frogs: a comparison with salamanders. *Experientia* **34**:44–46.

O'Malley J, Skylaki S, Iwabuchi KA, Chantzoura E, Ruetz T, Johnsson A, Tomlinson SR, Linnarsson S, Kaji K. 2013. High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature* **499**(7456):88–91

ÓMaoiléidigh DS, Wuest SE, Rae L, Raganelli A, Ryan PT, Kwasniewska K, Das P, Lohan AJ, Loftus B, Graciet E, Wellmer F. 2013. Control of reproductive floral organ identity specification in *Arabidopsis* by the C function regulator AGAMOUS. *Plant Cell* **25**(7):2482–2503.

Onder TT, Kara N, Cherry A, Sinha AU, Zhu N, Bernt KM, Cahan P, Marcarci BO, Unternaehrer J, Gupta PB, Lander ES, Armstrong SA, Daley GQ. 2012. Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **483**(7391):598–602.

Onodera Y, Haag J, Ream T, Costa Nunes P, Pontes O, Pikaard C. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**:613-622.

Orgel LE, Crick FH. 1980. Selfish DNA: The Ultimate Parasite. *Nature* **284**:604-607.

Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, Magnussen K, Steinmann KE, Kapranov P, Thompson JF, Zazula G, Froese D, Moltke I, Shapiro B, Hofreiter M, Al-Rasheid KA, Gilbert MT, Willerslev E. 2011. True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Res* **21**(10):1705–1719.

Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**(1):46–58.

Ouyang X, Li J, Li G, Li B, Chen B, Shen H, Huang X, Mo X, Wan X, Lin R, Li S, Wang H, Deng XW. 2011. Genome-wide binding site analysis of FAR-RED ELONGATED HYPOCOTYL3 reveals its novel function in Arabidopsis development. *Plant Cell* **23**(7):2514–2535.

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. 2002. Regulation of noise in the expression of a single gene. *Nat Genet* **31**(1):69–73.

Ozsolak F, Ting DT, Wittner BS, Brannigan BW, Paul S, Bardeesy N, Ramaswamy S, Milos PM, Haber DA. 2010. Amplification-free digital gene expression profiling from minute cell quantities. *Nat Methods* **7**(8):619–621.

Pachter L. 2011. Models for transcript quantification from RNA-Seq. *arXiv*:1104.3889v2

Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. 2008. A mitochondrial protein compendium elucidates complex I disease biology. **Cell** *134*(1):112–123.

Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, O Maoiléidigh DS, Wellmer F, Krajewski P, Riechmann JL, Angenent GC, Kaufmann K. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol* **15**(3):R41.

Palii CG, Perez-Iratxeta C, Yao Z, Cao Y, Dai F, Davison J, Atkins H, Allan D, Dilworth FJ, Gentleman R, Tapscott SJ, Brand M. 2010. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO J* **30**(3):494–509.

Palmer AC, Ahlgren-Berg A, Egan JB, Dodd IB, Shearwin KE. 2009. Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter. *Mol Cell* **34**(5):545–555.

Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev* **20**(2):153–158.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**(12):1413–1415.

Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, Frey BJ, Blencowe BJ. 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* **16**(6):929–941.

Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, Marjani SL, Euskirchen G, Ma C, Lamotte RH, Park IH, Snyder MP, Mason CE, Weissman SM. 2012. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci U S A.* **110**(2):594–509

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**(1):1–5.

Parfrey LW, Lahr DJ, Katz LA. 2008. The dynamic nature of eukaryotic genomes. *Mol Biol Evol* **25**(4):787–794.

Park E, Williams B, Wold BJ, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res* **22**(9):1626–1633.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**(10):669–680.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G. 2000. Conservation of the sequence and temporal expression of *let*-7 heterochronic regulatory RNA. *Nature* **408**(6808):86–89.

Patel AA, Steitz JA. 2005 Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol* **4**(12):960–970.

Patro R, Mount SM, Kingsford C. 2014. Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms. *arXiv*:1308.3700.

Patrushev LI, Minkevich IG. 2006. Eukaryotic noncoding DNA sequences provide genes with an additional protection against chemical mutagens. *Russ J Bioorg Chem* **32**:1068–1620.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, Ahituv N, Pennacchio LA, Shendure J. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**(3):265–270.

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**(3):577–591.

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream MA, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, Faulconbridge A, Jeffares D, Depledge DP, Oyola SO, Hilley JD, Brito LO, Tosi LR, Barrell B, Cruz AK, Mottram JC, Smith DF, Berriman M. 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet* **39**(7):839–847.

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7**(4):597–606.

Pedersen RA. 1971. DNA content, ribosomal gene multiplicity, and cell size in fish. *J Exp Zool* **177**(1):65–78.

Pehkonen P, Welter-Stahl L, Diwo J, Ryynänen J, Wienecke-Baldacchino A, Heikkinen S, Treuter E, Steffensen KR, Carlberg C. 2012. Genome-wide landscape of liver X receptor chromatin binding and gene regulation in human macrophages. *BMC Genomics* **13**:50.

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**(7447):127–131.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Bot J Linnean Society* **164**:10–15.

Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, Guo J, Dong Z, Liang Y, Bao L, Wang J. 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30**(3):253–260.

Pennisi E. 2012. Genomics. ENCODE project writes eulogy for junk DNA. *Science* **337**(6099):1159–1161.

Pepke S, Wold B, and Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6**:S22–32.

Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS. 2004. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in *Arabidopsis*. *Genes Dev* **18**(19):2368–2379.

Pervouchine DD, Knowles DG, Guigó R. 2013. Intron-Centric Estimation of Alternative Splicing from RNA-seq data. *Bioinformatics* **29**(2):273–274.

Peter IS, Davidson EH. 2011. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**(6):970–985.

Petrov DA. 2002. Mutational equilibrium model of genome size evolution. *Theor Popul Biol* **61**(4):531–544.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol* **20**:880-892.

Petruk S, Sedkov Y, Riley KM, Hodgson J, Schweisguth F, Hirose S, Jaynes JB, Brock HW, Mazo A. 2006. Transcription of *bxd* noncoding RNAs promoted by trithorax represses *Ubx* in *cis* by transcriptional interference. *Cell* **127**(6):1209–1221.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**(17):9748–9753.

Pevzner PA, Tang H. 2001. Fragment Assembly with Double-Barreled Data. *Bioinformatics/ISMB* **1**:1-9.

Pham XH, Farge G, Shi Y, Gaspari M, Gustafsson CM, Falkenberg M. 2006. Conserved sequence box II directs transcription termination and primer formation in mitochondria. *J Biol Chem* **281**(34):24647-24652.

Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. *Genome Res* **17**:1245–1253.

Philippe N, Salson M, Commes T, Rivals E. 2013. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol* **14**(3):R30

Phillips N, Kapraun DF, Gómez Garreta A, Ribera Siguan MA, Rull JL, Salvador Soler N, Lewis R, Kawai H. 2011. Estimates of nuclear DNA content in 98 species of brown algae (Phaeophyta). *AoB Plants* **2011**:plr001.

Piao Y, Kim HG, Oh MS, Pak YK. 2012. Overexpression of TFAM, NRF-1 and myr-AKT protects the MPP(+)-induced mitochondrial dysfunctions in neuronal cells. *Biochim Biophys Acta* **1820**(5):577–585.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**(11):1096–1098.

Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**(1):171–181.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**(7289):768–772.

Piehler AP, Hellum M, Wenzel JJ, Kaminski E, Haug KB, Kierulf P, Kaminski WE. 2008. The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. *BMC Genomics* **9**:165.

Pierro P, Capaccio L, Gadaleta G. 1999. The 25 kDa protein recognizing the rat curved region upstream of the origin of the L-strand replication is the rat homologue of the human mitochondrial transcription factor A. *FEBS Lett* **457**(3):307–310.

Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DR. 2011. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* **17**(5):792–798.

Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, Vitezic M, Severin J, Olivarius S, Lazarevic D, Hornig N, Orlando V, Bell I, Gao H, Dumais J, Kapranov P, Wang H, Davis CA, Gingeras TR, Kawai J, Daub CO, Hayashizaki Y, Gustincich S, Carninci P. 2010. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat Meth-*

ods **7**(7):528–534.

Pohjoismäki JL, Holmes JB, Wood SR, Yang MY, Yasukawa T, Reyes A, Bailey LJ, Cluett TJ, Goffart S, Willcox S, Rigby RE, Jackson AP, Spelbrink JN, Griffith JD, Crouch RJ, Jacobs HT, Holt IJ. 2010. Mammalian mitochondrial DNA replication intermediates are essentially duplex but contain extensive tracts of RNA/DNA hybrid. *J Mol Biol* **397**(5):1144–1155.

Polo JM, Anderssen E, Walsh RM, Schwarz BA, Nefzger CM, Lim SM, Borkent M, Apostolou E, Alaei S, Cloutier J, Bar-Nur O, Cheloufi S, Stadtfeld M, Figueroa ME, Robinton D, Natesan S, Melnick A, Zhu J, Ramaswamy S, Hochedlinger K. 2012. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell* **151**(7):1617–1632.

Polo JM, Liu S, Figueroa ME, Kulalert W, Eminli S, Tan KY, Apostolou E, Stadtfeld M, Li Y, Shioda T, Natesan S, Wagers AJ, Melnick A, Evans T, Hochedlinger K. 2010. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat Biotechnol* **28**(8):848–855.

Ponger L, Li WH. 2005. Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol Biol Evol* **22**(4):1119–1128.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res* **21**:1769-1776.

Porcel BM, Denoeud F, Opperdoes F, Noel B, Madoui MA, Hammarton TC, Field MC, Da Silva C, Couloux A, Poulain J, Katinka M, Jabbari K, Aury JM, Campbell DA, Cintron R, Dickens NJ, Docampo R, Sturm NR, Koumandou VL, Fabre S, Flegontov P, Lukes J, Michaeli S, Mottram JC, Szöor B, Zilberstein D, Bringaud F, Wincker P, Dollet M. The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS Genet* **10**(2):e1004007.

Posé D, Verhage L, Ott F, Yant L, Mathieu J, Angenent GC, Immink RG, Schmid M. 2013. Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* **503**(7476):414–417.

Poser I, Sarov M, Hutchins JR, Hériché JK, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW, Pelletier L, Kittler R, Hua S, Naumann R, Augsburg M, Sykora MM, Hofemeister H, Zhang Y, Nasmyth K, White KP, Dietzel S, Mechtler K, Durbin R, Stewart AF, Peters JM, Buchholz F, Hyman AA. 2008. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat Methods* **5**(5):409–415.

Pöyhönen M, de Vanssay A, Delmarre V, Hermant C, Todeschini A, Teysset L, Ronsseray S. 2012. Homology-dependent silencing by an exogenous sequence in the *Drosophila* germline. *G3 (Bethesda)* **2**:331-338.

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**(5909):1851–1854.

Prescott DM. 1999. The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. *Nucleic Acids Res* **27**:1243-1250.

Prescott DM, Ehrenfeucht A, Rozenberg G. 2003. Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. *J Theor Biol* **222**:323-330.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**(4825):336-341

Prokhortchouk A, Hendrich B, Jørgensen H, Ruzov A, Wilm M, Georgiev G, Bird A, Prokhortchouk E. 2001. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev* **15**(13):1613–1618.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**(Database issue):D32–36.

Psarra AM, Solakidi S, Sekeris CE. 2006. The mitochondrion as a primary site of action of steroid and thyroid hormones: presence and action of steroid and thyroid hormone receptors in mitochondria of animal cells. *Mol Cell Endocrinol* **246**(1-2):21–33.

Ptashne M. Specific binding of the lambda phage repressor to lambda DNA. *Nature* **214**(5085):232–234.

Ptasinska A, Assi SA, Mannari D, James SR, Williamson D, Dunne J, Hoogenkamp M, Wu

M, Care M, McNeill H, Cauchy P, Cullen M, Tooze RM, Tenen DG, Young BD, Cockerill PN, Westhead DR, Heidenreich O, Bonifer C. 2012. Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia* **26**(8):1829–1841.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. 2012. The Pfam protein families database. *Nucleic Acids Res* **40**(Database issue):D290–301.

Queiroz R, Benz C, Fellenberg K, Hoheisel JD, Clayton C. 2009. Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. *BMC Genomics* **10**:495.

Qi HH, Sarkissian M, Hu GQ, Wang Z, Bhattacharjee A, Gordon DB, Gonzales M, Lan F, Ongusaha PP, Huarte M, Yaghi NK, Lim H, Garcia BA, Brizuela L, Zhao K, Roberts TM, Shi Y. 2010. Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. *Nature* **466**(7305):503–507.

Qin Z, Yu J, Shen J, Maher C, Hu M, Kalyana-Sundaram S, Yu J, Chinnaiyan A. 2010. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* **11**(1):369.

Qiu S, Luo S, Evgrafov O, Li R, Schroth GP, Levitt P, Knowles JA, Wang K. 2012. Single-neuron RNA-Seq: technical feasibility and reproducibility. *Front Genet* **3**:124.

Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, Baglivo I, Pedone PV, Grimaldi G, Riccio A, Trono D. 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell* **44**(3):361–372.

Quong MW, Romanow WJ, Murre C. 2002. E protein function in lymphocyte development. *Annu Rev Immunol* **20:**301–322.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2010. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(7333):279–283.

Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M. 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci U S A* **107**:3639-3644.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc regulates transcriptional pause release. *Cell* **141**(3):432–445.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**(10):877–879.

Raj A, van Oudenaarden A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**(2):216–226.

Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A, Handunnetthi L, Handel AE, Disanto G, Orton SM, Watson CT, Morahan JM, Giovannoni G, Ponting CP, Ebers GC, Knight JC. 2010. A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res* **20**(10):1352–1360.

Ramos YF, Hestand MS, Verlaan M, Krabbendam E, Ariyurek Y, van Galen M, van Dam H, van Ommen GJ, den Dunnen JT, Zantema A, 't Hoen PA. 2010. Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. *Nucleic Acids Res* **38**(16):5396–5408.

Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**(8):777–782.

Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5**(12):e1000598.

Rao NA, McCalman MT, Moulos P, Francoijs KJ, Chatziioannou A, Kolisis FN, Alexis MN, Mitsiou DJ, Stunnenberg HG. 2011. Coactivation of GR and NFKB alters the repertoire of their binding sites and target genes. *Genome Res* **21**(9):1404–1416.

Raser JM, O'Shea EK. 2005. Noise in gene expression: origins, consequences, and control. *Science* **309**(5743):2010–2013.

Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* **12**(7):R67.

Raskatov JA, Nickols NG, Hargrove AE, Marinov GK, Wold B, Dervan PB. 2012. Gene expression changes in a tumor xenograft by a pyrrole-imidazole polyamide. *Proc Natl Acad Sci U S A* Raskatov(40):16041–16045.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:21–42.

Redi CA, Garagna S, Zuccotti M, Capanna E. 2007. Genome size: a novel genomic signature in support of Afrotheria. *J Mol Evol* **64**(4):484–487.

Reddy R, Henning D, Subrahmanyam CS, Busch H. 1984. Primary and secondary structure of 73 (K) RNA of Novikoff hepatoma. *J Biol Chem* **259**:12265–12270.

Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. 2009. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res* **19**(12):2163–2171.

Reddy TE, Gertz J, Crawford GE, Garabedian MJ, Myers RM. 2012. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. *Mol Cell Biol* **32**(18):3756–3767.

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, Crawford GE, Wold B, Willard HF, Myers RM. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22**(5):860–869.

Rees DJ, Dufresne F, Glémet H, Belzile C. 2007. Amphipod genome sizes: first estimates for Arctic species reveal genomic giants. *Genome* **50**(2):151–158.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. 2000. The 21-nucleotide *let*-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**(6772):901–906.

Remeseiro S, Cuadrado A, Gómez-López G, Pisano DG, Losada A. 2012. A unique role of cohesin-SA1 in gene regulation and development. *EMBO J* **31**(9):2090–2102.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**(5500):2306–2309.

Rey G, Cesbron F, Rougemont J, Reinke H, Brunner M, Naef F. 2011. Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biol* **9**(2):e1000595.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**(6):1408–1419.

Rhee HS, Pugh BF. 2012a. Genome-wide structure and organization of eukaryotic preinitiation complexes. *Nature* **483**(7389):295–301.

Rhee HS, Pugh BF. 2012b. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* **Chapter 21**:Unit 21.24.

Ribeiro S, Golding GB. 1998. The mosaic nature of the eukaryotic nucleus. *Mol Biol Evol* **15**(7):779–788.

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, Knox EB, Palmer JD. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* **342**(6165):1468–1473.

Riggs AD. 1975. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* **14**(1):9–25.

Rill RL, Livolant F, Aldrich HC, Davidson MW. 1989. Electron microscopy of liquid crystalline DNA: direct evidence for cholesteric-like organization of DNA in dinoflagellate chromosomes. *Chromosoma* **98**(4):280–286.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81**:145–166.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev* **17**(4):529–540.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**(7):1311–1323.

Rinn JL, Ule J. 2014. 'Oming in on RNA-protein interactions. *Genome Biol* **15**(1):401.

Rizzo PJ. 1987. Biochemistry of the dinoflagellate nucleus. In Taylor FJR (Ed.) The Biology of Dinoflagellates. *Blackwell, Oxford*, pp.143-173.

Rizzo PJ. 2003. Those amazing dinoflagellate chromosomes. *Cell Res* **13**:215-217.

Rizzo PJ, Nooden LD. 1972. Chromosomal proteins in the dinoflagellate alga *Gyrodinium cohnii*. *Science* **176**:796-797.

Roberts A, Pimentel H, Trapnell C, Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**(17):2325–2329.

Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* **10**(1):71–73.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**(3):R22.

Robertson AK, Geiman TM, Sankpal UT, Hager GL, Robertson KD. 2004. Effects of chromatin structure on the enzymatic and DNA binding functions of DNA methyltransferases DNMT1 and Dnmt3a in vitro. *Biochem Biophys Res Commun* **322**(1):110–118.

Robberson DL, Clayton DA. 1972. Replication of mitochondrial DNA in mouse L cells and their thymidine kinase - derivatives: displacement replication on a covalently-closed circular template. *Proc Natl Acad Sci U S A* **69**(12):3810–3814.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**(8):651–657.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I. 2010. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**(11):909–912.

Robine N, Lau N, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower M, Lai E. 2009. A broadly conserved pathway generates 3'UTR-directed primary piRNAs. *Curr Biol* **19**:2066-2076.

Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B. 2009. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of *Leishmania infantum*. *Mol Biochem Parasitol* **165**(1):32–47.

Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV. 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform* **6**:118-134.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable Interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**:1512–1517.

Romano LA, Wray GA. 2003. Conservation of *Endo16* expression in sea urchins despite evolutionary divergence in both *cis* and *trans*-acting components of transcriptional regulation. *Development* **130**(17):4187–4199.

Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* **103**(22):8390–8395.

Ropp PA, Copeland WC. 1996. Cloning and characterization of the human mitochondrial DNA polymerase, DNA polymerase gamma. *Genomics* **36**(3):449–458.

Rosenzweig D, Smith D, Myler PJ, Olafson RW, Zilberstein D. 2008. Post-translational modification of cellular proteins during *Leishmania donovani* differentiation. *Proteomics* **8**(9):1843–1850.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biol* **14**(5):R51.

Rossello FJ, Tothill RW, Britt K, Marini KD, Falzon J, Thomas DM, Peacock CD, Marchionni L, Li J, Bennett S, Tantoso E, Brown T, Chan P, Martelotto LG, Watkins DN. 2013. Next-generation sequence analysis of cancer xenograft models. *PLoS ONE* **8**(9):e74432.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**(7356):348–352.

Rothfels K, Sexsmith E, Heimburger M, Krause MO. 1966. Chromosome size and DNA content of species of *Anemone* and related genera (*Ranunculaceae*). *Chromosoma* **20**:54-74.

Roy J, Faktorova D, Lukes J, Burger G. 2007. Unusual mitochondrial genome structures throughout the Euglenozoa. *Protist* **158**:385-396.

Roy S, Morse D. 2012. A full suite of histone and histone modifying genes are transcribed in the dinoflagellate *Lingulodinium. PLoS One* **7**(4):e34340.

Roy SW. 2006. Intron-rich ancestors. *Trends in Genetics* **22**:468-471.

Roy SW, Gilbert W. 2005. Complex early genes. *Proceedings of the National Academy of Sciences* **102**:1986-1991.

Roy SW, Nosaka M, de Souza SJ, Gilbert W. 1999. Centripetal modules and ancient introns. *Gene* **238**(1):85–91.

Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet* **21**(8):466–475.

Rozhkov NV, Hammell M, Hannon GJ. 2013. Multiple roles for Piwi in silencing *Drosophila* transposons. *Genes Dev* **27**(4):400–412.

Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M. 2011. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**:522.

Rozowsky J, Euskirchen G, Auerbach R, Zhang Z, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein M. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**:66–75.

Ruan X, Ruan Y. 2012. Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET). *Methods Mol Biol* **809**:535-62.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**(6):1193–1207.

Ryu H, Lee J, Impey S, Ratan RR, Ferrante RJ. 2005. Antioxidants modulate mitochondrial PKA and increase CREB binding to D-loop DNA of the mitochondrial genome in neurons. *Proc Natl Acad Sci U S A* **102**(39):13915–20.

Sacomoto GA, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot MF, Peterlongo P, Lacroix V. 2012. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* **13**(Suppl 6):S5.

Sadasivam S, Duan S, DeCaprio JA. 2012. The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes Dev* **26**(5):474–89.

Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol* **14**(3):255–274.

Sahu B, Laakso M, Ovaska K, Mirtti T, Lundin J, Rannikko A, Sankila A, Turunen JP, Lundin M, Konsti J, Vesterinen T, Nordling S, Kallioniemi O, Hautaniemi S, Jänne OA. 2011. Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO J* **30**(19):3962–3976

Saito K, Nishida K, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi M. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev* **20**:2214-2222.

Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, Kotani H, Asai K, Siomi H, Siomi M. 2009. A regulatory circuit for piwi by the large Maf gene *traffic jam* in *Drosophila. Nature* **461**:1296-1299.

Sakabe NJ, Aneas I, Shen T, Shokri L, Park SY, Bulyk ML, Evans SM, Nobrega MA. 2012. Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum Mol Genet* **21**(10):2194–2204.

Sakarya O, Breu H, Radovich M, Chen Y, Wang YN, Barbacioru C, Utiramerur S, Whitley PP, Brockman JP, Vatta P, Zhang Z, Popescu L, Muller MW, Kudlingar V, Garg N, Li CY, Kong BS, Bodeau JP, Nutter RC, Gu J, Bramlett KS, Ichikawa JK, Hyland FC, Siddiqui AS. 2012. RNA-Seq Mapping and Detection of Gene Fusions with a Suffix Array Algorithm. *PLoS Comput Biol* **8**(4):e1002464.

Sakurai M, Yano T, Kawabata H, Ueda H, Suzuki T. 2010. Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nat Chem Biol* **6**(10):733–740.

Saldanha AJ. 2004. Java Treeview – extensible visualization of microarray data. *Bioinformatics* **20**(17):3246–3248.

Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. 2011. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**(3):353–358.

Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. 2013. Cell-type specific features of circular RNA expression. *PLoS Genet* **9**(9):e1003777.

Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. 2012. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* **7**(2):e30733.

Sample I. 2007 Jun 14. Study shines new light on genome. *The Guardian*

Sanchez MI, Mercer TR, Davies SM, Shearwood AM, Nygrd KK, Richman TR, Mattick JS, Rackham O, Filipovska A. 2011. RNA processing in human mitochondria. *Cell Cycle* **10**(17):2904–2916.

Sanchez-Freire V, Ebert AD, Kalisky T, Quake SR, Wu JC. 2012. Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat Protoc* **7**(5):829–838.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**(5883):1643–1647.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**:441-448.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(2):5463-5467.

Sano D, Myers JN. 2009. Xenograft models of head and neck cancers. *Head Neck Oncol* **1**:32.

Santana DM, Lukes J, Sturm NR, Campbell DA. 2001. Two sequence classes of kinetoplastid 5S ribosomal RNA gene revealed among bodonid spliced leader RNA gene arrays. *FEMS Microbiol Lett* **204**(2):233–237.

Santangelo AM, de Souza FS, Franchini LF, Bumaschny VF, Low MJ, Rubinstein M. 2007. Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet* **3**(10):1813–1826.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature* **419**:407-411.

Sarabhai AS, Strenton AO, Brenner S, Bolle A. 1964. Co-Linearity of The Gene with The Polypeptide Chain. *Nature* **201**:13–17.

Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* **14**(4):R31.

Sato S. 2011. The apicomplexan plastid and its evolution. *Cell Mol Life Sci* **68**(8):1285–1296.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana. DNA Res* **6**(5):283–290.

Satoh M, Kuroiwa T. 1991. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Exp Cell Res* **196**:137140

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, Liapis SC, Mallard W, Morse M, Swerdel MR, D'Ecclessis MF, Moore JC, Lai V, Gong G, Yancopoulos GD, Frendewey D, Kellis M, Hart RP, Valenzuela DM, Arlotta P, Rinn JL. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**:e01749.

Saxena A, Lahav T, Holland N, Aggarwal G, Anupama A, Huang Y, Volpin H, Myler PJ, Zilberstein D. 2007. Analysis of the *Leish-*

mania donovani transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. *Mol Biochem Parasitol* **152**(1):53–65.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19**(R2):R227–240.

Schatz G. 1963. The isolation of possible mitochondrial precursor structures from aerobically grown baker's yeast. *Biochem Biophys Res Commun* **12**:448–451.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**(9):1748–1759.

Schauer T, Schwalie PC, Handley A, Margulies CE, Flicek P, Ladurner AG. 2013. CAST-ChIP maps cell-type-specific chromatin states in the *Drosophila* central nervous system. *Cell Rep* **5**(1):271–282.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235):467-470.

Schiessl K, Muiño JM, Sablowski R. 2014. *Arabidopsis* JAGGED links floral organ patterning to tissue growth by repressing Kip-related cell cycle inhibitors. *Proc Natl Acad Sci U S A* **111**(7):2830-2835.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. 2012. Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**(7):676–682.

Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, Lange M, Tönjes M, Dunkel I, Sperling SR. 2010. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet* **7**(2):e1001313.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**(1-2):335–348.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**(5981):1036–1040.

Schmidt WM, Mueller MW. 1999. CapSelect: a highly sensitive method for 59 CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res* **27**(21):e31.

Schmitz SU, Albert M, Malatesta M, Morey L, Johansen JV, Bak M, Tommerup N, Abarrategui I, Helin K. 2011. Jarid1b targets genes regulating development and is involved in neural differentiation. *EMBO J* **30**(22):4586–4600.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956):1112–1115.

Schnetz MP, Handoko L, Akhtar-Zaidi B, Bartels CF, Pereira CF, Fisher AG, Adams DJ, Flicek P, Crawford GE, Laframboise T, Tesar P, Wei CL, Scacheri PC. 2010. CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. *PLoS Genet* **6**(7):e1001023.

Schödel J, Bardella C, Sciesielski LK, Brown JM, Pugh CW, Buckle V, Tomlinson IP, Ratcliffe PJ, Mole DR. 2012. Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nat Genet* **44**(4):420–425.

Schoeberl UE, Kurth HM, Noto T, Mochizuki K. 2012. Biased transcription and selective degradation of small RNAs shape the pattern of DNA elimination in *Tetrahymena*. *Genes Dev* 26:1729-1742.

Schoenherr CJ, Anderson DJ. 1995. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**(5202):1360–1363.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**(8):1086–1092.

Schwarzbauer JE, Tamkun JW, Lemischka IR, Hynes RO. 1983. Three different fibronectin mRNAs arise by alternative splicing within the coding region. *Cell* **35**:421–431.

Sebastian S, Faralli H, Yao Z, Rakopoulos P, Palii C, Cao Y, Singh K, Liu QC, Chu A, Aziz A, Brand M, Tapscott SJ, Dilworth FJ. 2013. Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. *Genes Dev* **27**(11):1247–1259.

Sehat B, Tofigh A, Lin Y, Trocmé E, Liljedahl U, Lagergren J, Larsson O. 2010. SUMOylation mediates the nuclear translocation and signaling of the IGF-1 receptor. *Sci Signal* **3**(108):ra10.

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322**(5909):1849–1851.

Seitz V, Butzhammer P, Hirsch B, Hecht J, Gütgemann I, Ehlers A, Lenze D, Oker E, Sommerfeld A, von der Wall E, König C, Zinser C, Spang R, Hummel M. 2011. Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma. *PLoS ONE* **6**(11):e26837.

Seok J, Xu W, Jiang H, Davis RW, Xiao W. 2012. Knowledge-Based Reconstruction of mRNA Transcripts with Short Sequencing Reads for Transcriptome Research. *PLoS ONE* **7**(2): e31440.

Shadel GS, Clayton DA. 1997. Mitochondrial DNA maintenance in vertebrates. *Annu Rev Biochem* **66**:409–435.

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**(7453):236–240.

Shan Q, Wang Y, Li J, Zhang Y, Chen K, Liang Z, Zhang K, Liu J, Xi JJ, Qiu JL, Gao C. 2013. Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat Biotechnol* **31**(8):686–688.

Shang J, Clayton DA. 1994. Human mitochondrial transcription termination exhibits RNA polymerase independence and biased bipolarity in vitro. *J Biol Chem* **269**(46):29112–29120.

Shankaranarayanan P, Mendoza-Parra MA, Walia M, Wang L, Li N, Trindade LM, Gronemeyer H. 2011. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat Methods* **8**(7):565–567.

Shao R, Kirkness EF, Barker SC. 2009. The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome Res* **19**(5):904–912.

Shao R, Zhu XQ, Barker SC, Herd K. 2012. Evolution of extensively fragmented mitochondrial genomes in the lice of humans. *Genome Biol Evol* **4**(11):1088–1101.

Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. 2012. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol* **13**(3):R16.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**(9):618–630.

Shapiro JA. 1999. Transposable elements as the key to a 21st century view of evolution. *Genetica* **107**(1-3):171–179.

Shapiro JA. 2002. A 21(st) Century View of Evolution. *J Biol Phys* **28**(4):745–764.

Shapiro JA. 2005. Retrotransposons and regulatory suites. *Bioessays* **27**(2):122–125.

Shapiro JA. 2009. Revisiting the central dogma in the 21st century. *Ann N Y Acad Sci* **1178**:6–28.

Shapiro JA. 2013. Rethinking the (im)possible in evolution. *Prog Biophys Mol Biol* **111**(2-3):92–96.

Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. *Biol Rev Camb Philos Soc* **80**(2):227–250.

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**(11):1009–1014.

Sharp PA, Burge CB. 1997. Classification of introns: U2-type or U12-type. *Cell* **91**(7):875–879.

Shaw JM, Feagin JE, Stuart K, Simpson L. 1988. Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. *Cell* **53**(3):401–411.

She H, Yang Q, Shepherd K, Smith Y, Miller G, Testa C, Mao Z. 2011. Direct regulation of complex I by mitochondrial MEF2D is disrupted in a mouse model of Parkinson disease and in human patients. *J Clin Invest* **121**(3):930–940.

Shearer TL, Van Oppen MJ, Romano SL, Wörheide G. 2002. Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Mol Ecol* **11**(12):2475–2487.

Shearwin KE, Callen BP, Egan JB. 2005. Transcriptional interference–a crash course. *Trends Genet* **21**(6):339–345.

Shen L, Gao G, Zhang Y, Zhang H, Ye Z, Huang S, Huang J, Kang J. 2010. A single amino acid substitution confers enhanced methylation activity of mammalian Dnmt3b on chromatin DNA. *Nucleic Acids Res* **38**(18):6054–6064.

Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, Carstens RP, Xing Y. 2012. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* **40**(8):e61.

Shen T, Aneas I, Sakabe N, Dirschinger RJ, Wang G, Smemo S, Westlund JM, Cheng H, Dalton N, Gu Y, Boogerd CJ, Cai CL, Peterson K, Chen J, Nobrega MA, Evans SM. 2011. Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. *J Clin Invest* **121**(12):4640–4654.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**(5741):1728–1732.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**(4):761–772.

Shirayama M, Seth M, Lee H-C, Gu W, Ishidate T, Conte D, Mello C. 2012. piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell* **150**:65-77.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**(7371):74–79.

Shulga N, Pastorino JG. 2012. GRIM-19-mediated translocation of STAT3 to mitochondria is necessary for TNF-induced necroptosis. *J Cell Sci* **125**(Pt 12):2995–3003.

Shuter BJ, Thomas JE, Taylor WD, Zimmerman AM. Phenotypic Correlates of Genomic DNA Content in Unicellular Eukaryotes and Other Cells. *The American Naturalist* **122**(1):26–44

Shutt TE, Bestwick M, Shadel GS. 2011. The core human mitochondrial transcription initiation complex: It only takes two to tango. *Transcription* **2**(2):55–59.

Shutt TE, Gray MW. 2006. Bacteriophage origins of mitochondrial replication and transcription proteins. *Trends Genet* **22**(2):90–95.

Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA. 2010. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res* **38**(15):4946–4957.

Sienski G, Dönertas D, Brennecke J. 2012. Transcriptional silencing of transposons by piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* **151**:964-980.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**(8):1034–1050.

Siersbæk R, Nielsen R, John S, Sung MH, Baek S, Loft A, Hager GL, Mandrup S. 2011. Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *EMBO J* **30**(8):1459–1472.

Silver TD, Koike S, Yabuki A, Kofuji R, Archibald JM, Ishida K. 2007. Phylogeny and nucleomorph karyotype diversity of chlorarachniophyte algae. *J Eukaryot Microbiol* **54**:403-410

Simon JA, Kingston RE. 2013. Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol Cell* **49**(5):808–824.

Simon MD, Wang CI, Kharchenko PV, West JA, Chapman BA, Alekseyenko AA, Borowsky ML, Kuroda MI, Kingston RE. 2011. The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci U S A* **108**(51):20497–20502.

Simmons RA, Suponitsky-Kroyter I, Selak MA. 2005. Progressive accumulation of mitochondrial DNA mutations and decline in mitochondrial function lead to beta-cell failure. *J Biol Chem* **280**(31):28785–28791.

Simpson AG, Stevens JR, Lukes J. 2006. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol* **22**(4):168–174.

Simpson AM, Suyama Y, Dewes H, Campbell DA, Simpson L. 1989. Kinetoplastid mitochondria contain functional tRNAs which are encoded in nuclear DNA and also contain small minicircle and maxicircle transcripts of unknown function. *Nucleic Acids Res* **17**:5427-5445.

Simpson L. 1997. The genomic organization of guide RNA genes in kinetoplastid protozoa: several conundrums and their solutions. *Mol Biochem Parasitol* **86**:133-141.

Simpson L, Thiemann OH. 1995. Sense from nonsense: RNA editing in mitochondria of kinetoplastid protozoa and slime molds. *Cell* **81**:837-840.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**(6):1117–1123.

Singh K, Carey M, Saragosti S, Botchan M. 1985. Expression of enhanced levels of small RNA polymerase III transcripts encoded by the *B2* repeats in simian virus 40-transformed mouse cells. *Nature* **314**(6011):553–556.

Siomi M, Sato K, Pezic D, Aravin A. 2011. PIWI-interacting small RNAs: The vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**:246-258.

Slack FJ. 2006. Regulatory RNAs and the demise of junk DNA. *Genome Biol* **7**:328.

Slamovits CH, Saldarriaga JF, Larocque A, Keeling PJ. 2007. The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *J Mol Biol* **372**:356-368.

Sleutels F, Zwart R, Barlow DP. 2002. The noncoding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415**:810-813

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol* **10**(1):e1001241.

Smeenk L, van Heeringen SJ, Koeppel M, Gilbert B, Janssen-Megens E, Stunnenberg HG, Lohrum M. 2011. Role of p53 serine 46 in p53 target gene regulation. *PLoS ONE* **6**(3):e17574.

Smith DR. 2013. RNA-Seq data: a goldmine for organelle research. *Brief Funct Genomics* **12**(5):454–456.

Smith DR, Kayal E, Yanagihara AA, Collins AG, Pirro S, Keeling PJ. 2012. First complete mitochondrial genome sequence from a box jellyfish reveals a highly fragmented linear architecture and insights into telomere evolution. *Genome Biol Evol* **4**(1):52–58.

Smith ER, Lin C, Garrett AS, Thornton J, Mohaghegh N, Hu D, Jackson J, Saraf A, Swanson SK, Seidel C, Florens L, Washburn MP, Eissenberg JC, Shilatifard A. 2011. The little elongation complex regulates small nuclear RNA transcription. *Mol Cell* **44**(6):954–965.

Smith JD, Gregory TR. 2009. The genome sizes of megabats (Chiroptera: Pteropodidae) are remarkably constrained. *Biol Lett* **5**(3):347–351.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**(9):1021–1028.

Soccio RE, Tuteja G, Everett LJ, Li Z, Lazar MA, Kaestner KH. 2011. Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol Endocrinol* **25**(4):694–706.

Sologub M, Litonin D, Anikin M, Mustaev A, Temiakov D. 2009. TFB2 is a transient component of the catalytic site of the human mitochondrial RNA polymerase. *Cell* **139**(5):934–944.

Solomon MJ, Larsen PL, Varshavsky A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**(6):937–947.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Gräf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Zhang Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21**(10):1757–1767.

Soppe WJ, Jasencakova Z, Houben A, Kakutani T, Meister A, Huang M, Jacobsen S, Schubert I, Fransz P. 2002. DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *EMBO J* **21**:6549-6559.

Sorek R, Shamir R, Ast G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet* **20**(2):68–71.

Southall TD, Gold KS, Egger B, Davidson CM, Caygill EE, Marshall OJ, Brand AH. 2013. Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells. *Dev Cell* **26**(1):101–112.

Speijer D. 2006. Is kinetoplastid pan-editing the result of an evolutionary balancing act? *IUBMB Life* **58**(2):91–96.

Speijer D. 2011. Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity. *Bioessays* **33**:344-349.

Squazzo SL, OGeen H, Komashko VM, Krig SR, Jin VX, Jang S, Margueron R, Reinberg D, Green R, Farnham PJ. 2006. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* **16**:890-900.

Sreedharan VT, Schultheiss SJ, Jean G, Kahles A, Bohnert R, Drewe P, Mudrakarta P, Görnitz N, Zeller G, Rätsch G. 2014. Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis. *Bioinformatics* **30**(9):1300–1301.

Srividya G, Duncan R, Sharma P, Raju BV, Nakhasi HL, Salotra P. 2007. Transcriptome analysis during the process of in vitro differentiation of *Leishmania donovani* using genomic microarrays. *Parasitology* **134**(Pt 11):1527–1539.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**(7378):490–495.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* **344**:1–20.

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, Flicek P, Odom DT. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**(3):530–540.

Steger DJ, Grant GR, Schupp M, Tomaru T, Lefterova MI, Schug J, Manduchi E, Stoeckert CJ Jr, Lazar MA. 2010. Propagation of adipogenic signals through an epigenomic transition state. *Genes Dev* **24**(10):1035–1044.

Steijger T, Abril JF, Engstrm PG, Kokocinski F; RGASP Consortium, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P, Bohnert R, Bucher P, Cloonan N, Derrien T, Djebali S, Du J, Dudoit S, Engstrm PG, Gerstein M, Gingeras TR, Gonzalez D, Grimmond SM, Guigó R, Habegger L, Harrow J, Hubbard TJ, Iseli C, Jean G, Kahles A, Kokocinski F, Lagarde J, Leng J, Lefebvre G, Lewis S, Mortazavi A, Niermann P, Rtsch G, Reymond A, Ribeca P, Richard H, Rougemont J, Rozowsky J, Sammeth M, Sboner A, Schulz MH, Searle SM, Solorzano ND, Solovyev V, Stanke M, Steijger T, Stevenson BJ, Stockinger H, Valsesia A, Weese D, White S, Wold BJ, Wu J, Wu TD, Zeller G, Zerbino D, Zhang MQ, Hubbard TJ, Guigó R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**(12):1177–1184.

Steiner FA, Talbert PB, Kasinathan S, Deal RB, Henikoff S. 2012. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Res* **22**(4):766–777.

Stoebe B, Maier UG. 2002. One, two, three: nature's tool box for building plastids. *Protoplasma* **219**(3-4):123–130.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**(5696):655–660.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J Mol Evol* **49**:169-181

Stoltzfus A. 2012. Constructive neutral evolution: exploring evolutionary theory's curious disconnect. *Biology Direct* **7**:35

Stroud H, Otero S, Desvoyes B, Ramírez-Parra E, Jacobsen SE, Gutierrez C. 2012. Genome-wide analysis of histone H3.1 and H3.3 variants in *Arabidopsis thaliana. Proc Natl Acad Sci U S A* **109**(14):5370–5375.

Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**(2):103–105.

Stumpf JD, Saneto RP, Copeland RC. 2013. Clinical and Molecular Features of POLG-Related Mitochondrial Disease. *Cold Spring Harb Perspect Biol* **4**(5):a011395.

Suarez J, Hu Y, Makino A, Fricovsky E, Wang H, Dillmann WH. 2008. Alterations in mitochondrial function and cytosolic calcium induced by hyperglycemia are restored by mitochondrial transcription factor A in cardiomyocytes. *Am J Physiol Cell Physiol* **295**(6):1561–1568.

Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sánchez-Pons N, Torruella G, Derelle R, Manning G, Lang BF, Russ C, Haas BJ, Roger AJ, Nusbaum C, Ruiz-Trillo I. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat Commun* **4**:2325.

Sugiyama S, Hattori K, Hayakawa M, Ozawa T. 1991. Quantitative analysis of age-associated accumulation of mitochondrial DNA with deletion in human age-associated accumulation of mitochondrial DNA with deletion in human hearts. *Biochem Biophys Res Commun* **180**(2):894–899.

Sugiyama T, Cam H, Verdel A, Moazed D, Grewal S. 2005. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proc Natl Acad Sci* **102**:152-157.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**(5891):956–960.

Sun J, Pan H, Lei C, Yuan B, Nair SJ, April C, Parameswaran B, Klotzle B, Fan JB, Ruan J, Li R. 2011. Genetic and genomic analyses of RNA polymerase II-pausing factor in regulation of mammalian transcription and cell growth. *J Biol Chem* **286**(42):36248–36257.

Sun L, Goff LA, Trapnell C, Alexander R, Lo KA, Hacisuleyman E, Sauvageau M, Tazon-Vega B, Kelley DR, Hendrickson DG, Yuan B, Kellis M, Lodish HF, Rinn JL. 2013. Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* **110**(9):3387–3392.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A* **109**(45):18488–18492.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. 2009. MEN ϵ/β nuclear-retained non-coding RNAs are upregulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19**(3):347–359.

Suomalainen A, Isohanni P. 2010. Mitochondrial DNA depletion syndromes - many genes, common mechanisms. *Neuromuscul Disord* **20**(7)429–437.

Sureau A, Gattoni R, Dooghe Y, Stévenin J, Soret J. 2001. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J* **20**(7):1785–1796.

Surono A, Takeshima Y, Wibawa T, Ikezawa M, Nonaka I, Matsuo M. 1999. Circular dystrophin RNAs consisting of exons that were skipped by alternative splicing. *Hum Mol Genet* **8**(3):493–500.

Sutton WS 1902. On the morphology of the chromosome group in Brachystola magna. *Biol Bull* **4**:24–39.

Sutton WS 1903. The chromosomes in heredity. *Biol Bull* **4**:231–251.

Suzuki K, Miyagishima SY. 2010. Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. *Mol Biol Evol* **27**:581-590.

Swanton MT, Greslin AF, Prescott DM. 1980. Arrangement of coding and non-coding sequences in the DNA molecules coding for rRNAs in *Oxytricha sp.* DNA of ciliated protozoa. VII. *Chromosoma* **77**:203-215.

Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, Khurana JS, Goldman AD, Nowacki M, Schotanus K, Jung S, Fulton RS, Ly A, McGrath S, Haub K, Wiggins JL, Storton D, Matese JC, Parsons L, Chang WJ, Bowen MS, Stover NA, Jones TA, Eddy SR, Herrick GA, Doak TG, Wilson RK, Mardis ER, Landweber LF. 2013. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol* **11**(1):e1001473.

Szczepanek K, Chen Q, Derecka M, Salloum FN, Zhang Q, Szelag M, Cichy J, Kukreja RC, Dulak J, Lesnefsky EJ, Larner AC. 2011. Mitochondrial-targeted Signal transducer and activator of transcription 3 (STAT3) protects against ischemia-induced changes in the electron transport chain and the generation of reactive oxygen species. *J Biol Chem* **286**(34):29610–29620.

Szczepanek K, Chen Q, Larner AC, Lesnefsky EJ. 2012. Cytoprotection by the modulation of mitochondrial electron transport chain: the emerging role of mitochondrial STAT3. *Mitochondrion* **12**(2):180–189.

Szczepanek K, Lesnefsky EJ, Larner AC. 2012. Multi-tasking: nuclear transcription factors with novel roles in the mitochondria. *Trends Cell Biol* **22**(8):429–437.

Szklarczyk R, Huynen MA. 2010. Mosaic origin of the mitochondrial proteome. *Proteomics* **10**(22):4012–4024.

Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, Daub CO, Hume DA, Suzuki H, Orlando V, Carninci P, Hayashizaki Y, Mattick JS. 2009a. Tiny RNAs associated with transcription start sites in animals. *Nat Genet* **41**(5):572–578.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**:288–299

Taft RJ, Kaplan CD, Simons C, Mattick JS. 2009b. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* **8**(15):2332–2338.

Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ, Rasko JE, Rokhsar DS, Degnan BM, Mattick JS. 2010. Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol* **17**(8):1030–1034.

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**(5929):930–935.

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**(4):663–676.

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**(5):861–872.

Takeshima H, Suetake I, Shimahara H, Ura K, Tate S, Tajima S. 2006. Distinct DNA methylation activity of Dnmt3a and Dnmt3b towards naked and nucleosomal DNA. *J Biochem* **139**(3):503–515.

Takeshima H, Suetake I, Tajima S. 2008. Mouse Dnmt3a preferentially methylates linker DNA and is inhibited by histone H1. *J Mol Biol* **383**:810-821.

Tallack MR, Whitington T, Yuen WS, Wainwright EN, Keys JR, Gardiner BB, Nourbakhsh E, Cloonan N, Grimmond SM, Bailey TL, Perkins AC. 2010. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res*

20(8):1052–1063.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**(7194):534–538.

Tan PY, Chang CW, Chng KR, Wansa KD, Sung WK, Cheung E. 2011a. Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Mol Cell Biol* **32**(2):399–414.

Tan SK, Lin ZH, Chang CW, Varang V, Chng KR, Pan YF, Yong EL, Sung WK, Cheung E. 2011b. AP-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *EMBO J* **30**(13):2569–2581.

Tang C, Shi X, Wang W, Zhou D, Tu J, Xie X, Ge Q, Xiao PF, Sun X, Lu Z. 2010. Global analysis of in vivo EGR1-binding sites in erythroleukemia cell using chromatin immunoprecipitation and massively parallel sequencing. *Electrophoresis* **31**(17):2936–2943.

Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, Lao K, Surani MA. 2010. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc.* **5**(3):516–535.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**(5):377–382.

Tang F, Lao K, Surani MA. 2011. Development and applications of single-cell transcriptome analysis. *Nat Methods* **8**(4 Suppl):S6–11.

Tang S, Riva A. 2013. PASTA: splice junction identification from RNA-Sequencing data. *BMC Bioinformatics* **14**(1):116.

Tanifuji G, Onodera NT, Moore CE, Archibald JM. 2014. Reduced nuclear genomes maintain high gene transcription levels. *Mol Biol Evol* **31**(3):625–635.

Tanifuji G, Onodera NT, Wheeler TJ, Dlutek M, Donaher N, Archibald JM. 2011. Complete nucleomorph genome sequence of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set. *Genome Biol Evol* **3**:44–54.

Tarn WY, Steitz JA. 1996. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273**(5283):1824–1832.

Tarn WY, Steitz JA. 1996 A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84**(5):801–811.

Tasic B, Nabholz CE, Baldwin KK, Kim Y, Rueckert EH, Ribich SA, Cramer P, Wu Q, Axel R, Maniatis T. 2002. Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell* **10**(1):21–33.

Taslim C, Wu J, Yan P, Singer G, Parvin J, Huang T, Lin S, Huang K. 2009. Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics* **25**(18):2334–2340.

Tay Y, Karreth FA, Pandolfi PP. 2014a. Aberrant ceRNA activity drives lung cancer. *Cell Res* **24**(3):259–260.

Tay Y, Rinn J, Pandolfi PP. 2014b. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**(7483):344–352.

Tay Y, Kats L, Salmena L, Weiss D, Tan SM, Ala U, Karreth F, Poliseno L, Provero P, Di Cunto F, Lieberman J, Rigoutsos I, Pandolfi PP. 2011. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* **147**(2):344–357.

Tchieu J, Kuoy E, Chin MH, Trinh H, Patterson M, Sherman SP, Aimiuwu O, Lindgren A, Hakimian S, Zack JA, Clark AT, Pyle AD, Lowry WE, Plath K. 2010. Female human iPSCs retain an inactive X chromosome. *Cell Stem Cell* **7**(3):329–342.

Tentler JJ, Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, Arcaroli JJ, Messersmith WA, Eckhardt SG. 2012. Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol* **9**(6):338–350.

Teo AK, Arnold SJ, Trotter MW, Brown S, Ang LT, Chng Z, Robertson EJ, Dunn NR, Vallier L. 2011. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev* **25**(3):238–250.

Teodorovic S, Walls CD, Elmendorf HG. 2007. Bidirectional transcription is an inherent feature of *Giardia lamblia* promoters and contributes to an abundance of sterile antisense transcripts throughout the genome. *Nucleic Acids Res* **35**(8):2544–2553.

ter Schegget J, Flavell RA, Borst P. 1971. DNA synthesis by isolated mitochondria. 3. Characterization of D-loop DNA, a novel intermediate in mtDNA synthesis. *Biochim Biophys*

Acta **254**(1):1–114.

Terzioglu M, Ruzzenente B, Harmel J, Mourier A, Jemt E, López MD, Kukat C, Stewart JB, Wibom R, Meharg C, Habermann B, Falkenberg M, Gustafsson CM, Park CB, Larsson NG. 2013. MTERF1 binds mtDNA to prevent transcriptional interference at the light-strand promoter but is dispensable for rRNA gene transcription regulation. *Cell Metab* **17**(4):618-626.

Teytelman L, Ozaydin B, Zill O, Lefrançois P, Snyder M, Rine J, Eisen MB. 2009. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One* **4**(8):e6700.

Thévenaz P, Ruttimann U, Unser M. 1998. A pyramid approach to subpixel registration based on intensity. *IEEE Trans Image Process* **7**:27-41.

Thiebaut M, Colin J, Neil H, Jacquier A, Seraphin B, Lacroute F, Libri D. 2008. Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in S. cerevisiae. *Mol Cell* **31**:671-682

Thiebaut M, Kisseleva-Romanova E, Rougemaille M, Boulay J, Libri D. 2006. Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Mol Cell* **23**(6):853–864.

Thomas CA Jr. 1971. The genetic organization of chromosomes. *Annu Rev Genet* **5**:237–256.

Thompson CC, Brown TA, McKnight SL. 1991. Convergence of Ets- and notch-related structural motifs in a heteromeric DNA binding complex. *Science* **253**(5021):762–768.

Thompson D, Parker R. 2007. Cytoplasmic Decay of Intergenic Transcripts in *Saccharomyces cerevisiae. Mol Cell Biol* **27**:92-101.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414):75–82.

Tian D, Sun S, Lee JT. 2010. The long non-coding RNA, *Jpx*, is a molecular switch for X chromosome inactivation. *Cell* **143**:390-403.

Tiersch TR, Wachtel SS. 1991. On the evolution of genome size of birds. *J Hered* **82**(5):363–368.

Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK, Wang X, Ottersbach K, Stemple DL, Green AR, Ouwehand WH, Göttgens B. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20**(5):597–609.

Tiwari VK, Burger L, Nikoletopoulou V, Deogracias R, Thakurela S, Wirbelauer C, Kaut J, Terranova R, Hoerner L, Mielke C, Boege F, Murr R, Peters AH, Barde YA, Schübeler D. 2011a. Target genes of Topoisomerase IIβ regulate neuronal survival and are defined by their chromatin state. *Proc Natl Acad Sci U S A* **109**(16):E934–943.

Tiwari VK, Stadler MB, Wirbelauer C, Paro R, Schübeler D, Beisel C. 2011b. A chromatin-modifying function of JNK during stem cell differentiation. *Nat Genet* **44**(1):94–100.

Toedling J, Servant N, Ciaudo C, Farinelli L, Voinnet O, Heard E, Barillot E. 2012. Deep-sequencing protocols influence the results obtained in small-RNA sequencing. *PLoS ONE* **7**(2):e32724.

Tovar J, Fischer A, Clark CG. 1999. The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica. Mol Microbiol* **32**(5):1013–1021.

Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, van der Giezen M, Hernández M, Mller M, Lucocq JM. 2003. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* **426**(6963):172–176.

Tragante V, Moore JH, Asselbergs FW. 2014. The ENCODE Project and Perspectives on Pathways. *Genet Epidemiol* doi: 10.1002/gepi.21802. [Epub ahead of print]

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**(1):46–53.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9):1105–1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**(3):562–578.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5):511–515.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2012a. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* **31**(1):46–53.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012b. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**(3):562–578.

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* **38**(15):e159.

Trifunovic A, Larsson NG. 2008. Mitochondrial dysfunction as a cause of ageing. *J Intern Med* **263**(2):167–178.

Trompouki E, Bowman TV, Lawton LN, Fan ZP, Wu DC, DiBiase A, Martin CS, Cech JN, Sessa AK, Leblanc JL, Li P, Durand EM, Mosimann C, Heffner GC, Daley GQ, Paulson RF, Young RA, Zon LI. 2011. Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* **147**(3):577–589.

Trowbridge JJ, Sinha AU, Zhu N, Li M, Armstrong SA, Orkin SH. 2012. Haploinsufficiency of Dnmt1 impairs leukemia stem cell function through derepression of bivalent chromatin domains. *Genes Dev* **26**(4):344–349.

Turro E, Su SY, Gonçalves Â, Coin LJ, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12**(2):R13.

Tuteja G, White P, Schug J, Kaestner KH. 2009. Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res* **37**(17):e113.

Tzur YB, Friedland AE, Nadarajan S, Church GM, Calarco JA, Colaiácovo MP. 2013. Heritable custom genomic modifications in *Caenorhabditis elegans* via a CRISPR-Cas9 system. *Genetics* **195**(3):1181–1185.

Udvardy A, Maine E, Schedl P. 1985. The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *J Mol Biol* **185**(2):341–358.

Uhler JP, Hertel C, Svejstrup JQ. 2007. A role for noncoding transcription in activation of the yeast *PHO5* gene. *Proc Natl Acad Sci U S A* **104**:8011-8016

Umbarger MA, Toro E, Wright MA, Porreca GJ, Baù D, Hong SH, Fero MJ, Zhu LJ, Marti-Renom MA, McAdams HH, Shapiro L, Dekker J, Church GM. 2011. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol Cell* **44**(2):252–264.

Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, Haussler D. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* **7**(12):995-1001.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nat Genet* **15**(1):57–61.

Vaidya AB, Akella R, Suplick K. 1989. Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malarial parasite. *Mol Biochem Parasitol* **35**(2):97–107.

Vaidya AB, Mather MW. 2009. Mitochondrial evolution and functions in malaria parasites. *Annu Rev Microbiol* **3**:249–267.

Vagin V, Sigova A, Li C, Seitz H, Gvozdev V, Zamore P. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**:320-324.

Valdes C, Seo P, Tsinoremas N, Clarke J. 2013. Characteristics of cross-hybridization and cross-alignment of expression in pseudo-xenograft samples by RNA-Seq and microarrays. *J Clin Bioinforma* **3**(1):8.

Valouev A, Johnson D, Sundquist A, Medina C, Anton E, Batzoglou S, Myers R, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**:829–834.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**(5):e1000371.

van der Giezen M. 2009. Hydrogenosomes and mitosomes: conservation and evolution of functions. *J Eukaryot Microbiol* **56**:221-231

van Heeringen SJ, Akhtar W, Jacobi UG, Akkers RC, Suzuki Y, Veenstra GJ. 2011. Nucleotide composition-linked divergence of vertebrate core promoter architecture. *Genome Res* **21**(3):410–421.

van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A, Ramasamy G, Vainio S, Heidebrecht T, Perrakis A, Pagie L, van Steensel B, Myler PJ, Borst P. 2012. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania. Cell* **150**(5):909–921.

van Oppen MJ, Catmull J, McDonald BJ, Hislop NR, Hagerman PJ, Miller DJ. 2002. The mitochondrial genome of *Acropora tenuis* (Cnidaria; Scleractinia) contains a large group I intron and a candidate control region. *J Mol Evol* **55**(1):1–13

van Weerden WM, Bangma C, de Wit R. 2009. Human xenograft models as useful tools to assess the potential of novel therapeutics in prostate cancer. *Br J Cancer* **100**(1):13–18.

Vanhée-Brossollet C, Vaquero C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211**(1):1–9.

Vanin EF, Goldberg GI, Tucker 1980. A mouse α-globin-related pseudogene lacking intervening sequences. *Nature* **286**(5770):222–226.

Vansant G, Reynolds W 1995. The consensus sequence of a major *Alu* subfamily contains a functional retinoic acid response element. *Proc Natl Acad Sci U S A* **92**(18):8229–8233.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**(4):252–263.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE, Absher DM, Wold BJ, Myers RM. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **23**(3):555–567.

Vaseva AV, Moll UM. 2009. The mitochondrial p53 pathway. *Biochim Biophys Acta* **1787**(5):414–420.

Vasquez JJ, Hon CC, Vanselow JT, Schlosser A, Siegel TN. 2014. Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res* **42**(6):3623–3637.

Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crété P. 2004. Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol Cell* **16**(1):69–79.

Veitch NJ, Johnson PC, Trivedi U, Terry S, Wildridge D, MacLeod A. 2010. Digital gene expression analysis of two life cycle stages of the human-infective parasite, *Trypanosoma brucei gambiense* reveals differentially expressed clusters of co-regulated genes. *BMC Genomics* **11**:124.

Veldhuis MJW, Cucci TL, Sieracki ME. 1997. Cellular DNA content of marine phytoplankton using two new fluorochromes: taxonomic and ecological implications. *J Phycol* **33**:527-541.

Venables JP, Tazi J, Juge F. 2012. Regulated functional alternative splicing in *Drosophila. Nucleic Acids Res* **40**(1):1–10.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu

F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. *Science* **291**(5507):1304–1351.

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. 1998. Shotgun sequencing of the human genome. *Science* **280**(5369):1540-1542

Venters BJ, Pugh BF. 2013. Genomic organization of human transcription initiation complexes. *Nature* **502**(7469):53–58.

Vermeulen M, Eberl HC, Matarese F, Marks H, Denissov S, Butter F, Lee KK, Olsen JV, Hyman AA, Stunnenberg HG, Mann M. 2010. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell 142*(6):967–980.

Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM. 2012. Personal and population genomics of human regulatory variation. *Genome Res* **22**(9):1689–1697.

Verzi MP, Shin H, He HH, Sulahian R, Meyer CA, Montgomery RK, Fleet JC, Brown M, Liu XS, Shivdasani RA. 2010. Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. *Dev Cell* **19**(5):713–726.

Verzi MP, Shin H, Ho LL, Liu XS, Shivdasani RA. 2011. Essential and redundant functions of caudal family proteins in activating adult intestinal genes. *Mol Cell Biol* **31**(10):2026–2039.

Vierstra J, Wang H, John S, Sandstrom R, Stamatoyannopoulos JA. 2014. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat Methods* **11**(1):66–72.

Vilagos B, Hoffmann M, Souabni A, Sun Q, Werner B, Medvedovic J, Bilic I, Minnich M, Axelsson E, Jaritz M, Busslinger M. 2012. Essential role of EBF1 in the generation and function of distinct mature B cell types. *J Exp Med* **209**(4):775–792.

Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet* **15**(4):221–233.

Vinogradov AE. 1998. Buffering: a possible passive-homeostasis role for redundant DNA. *J Theor Biol* **193**:197-199.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231):854–858.

Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ, Hansen DV, Nord AS, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry

M, Afzal V, Kaplan T, Kriegstein AR, Rubin EM, Ovcharenko I, Pennacchio LA, Rubenstein JL. 2013. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**(4):895–908.

Vivar OI, Zhao X, Saunier EF, Griffin C, Mayba OS, Tagliaferri M, Cohen I, Speed TP, Leitman DC. 2010. Estrogen receptor beta binds to and regulates three distinct classes of target genes. *J Biol Chem* **285**(29):22059–22066.

Vivarès CP, Méténier G. 2000. Towards the minimal eukaryotic parasitic genome. *Curr Opin Microbiol* **3**(5):463–467.

Vlcek C, Marande W, Teijeiro S, Lukes J, Burger G. 2011. Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Res* **39**(3):979-988

Voigt O, Erpenbeck D, Worheide G. 2008. A fragmented metazoan organellar genome: the two mitochondrial chromosomes of *Hydra magnipapillata*. *BMC Genomics* 9:350

von Sternberg R. 2002. On the Roles of Repetitive DNA Elements in the Context of a Unified GenomicEpigenetic System. *Ann N Y Acad Sci* **981**:154–188.

von Sternberg R, Shapiro JA. How Repeated Retroelements format genome function. *Cytogenet Genome Res* **110**(1-4):108–116.

Waaijers S, Portegijs V, Kerver J, Lemmens BB, Tijsterman M, van den Heuvel S, Boxem M. 2013. CRISPR/Cas9-targeted mutagenesis in *Caenorhabditis elegans*. *Genetics* **195**(3):1187–1191.

Wallace EV, Stoddart D, Heron AJ, Mikhailova E, Maglia G, Donohoe TJ, Bayley H. 2010. Identification of epigenetic DNA modifications with a protein nanopore. *Chem Commun (Camb)* **46**(43):8195–8197.

Walkup LK. 2000. Junk DNA: evolutionary discards or God's tools? *Creation Ex Nihilo Technical Journal* **14**:18–30.

Wallace EV, Stoddart D, Heron AJ, Mikhailova E, Maglia G, Donohoe TJ, Bayley H. 2010. Identification of epigenetic DNA modifications with a protein nanopore. *Chem Commun (Camb)* **46**(43):8195–8197.

Waller RF, Jackson CJ. 2009. Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *Bioessays* **31**:237-245.

Walter P, Blobel G. 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**(5885):691–698.

Wang C, Xu J, Zhang D, Wilson ZA, Zhang D. 2010. An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. *BMC Bioinformatics* **11**:81.

Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG, Fu XD. 2011a. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**(7351):390–394.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221):470–476.

Wang H, Iacoangeli A, Popp S, Muslimov IA, Imataka H, Sonenberg N, Lomakin IB, Tiedge H. 2002. Dendritic BC1 RNA: functional role in regulation of translation initiation. *J Neurosci* **22**(23):10232–10241.

Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. 2013. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**(4):910–918.

Wang H, Zou J, Zhao B, Johannsen E, Ashworth T, Wong H, Pear WS, Schug J, Blacklow SC, Arnett KL, Bernstein BE, Kieff E, Aster JC. 2011b. Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. *Proc Natl Acad Sci U S A* **108**(36):14908–14913.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**(9):1798–1812.

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. 2010. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**(18):e178.

Wang K, Wu Y, Zhang W, Dawe RK, Jiang J. 2014. Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res* **24**(1):107–116.

Wang L, Feng Z, Wang X, Wang X, Zhang X. 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**(1):136–138.

Wang L, Xi Y, Yu J, Dong L, Yen L, Li W. 2010. A statistical method for the detection of alternative splicing using RNA-seq. *PLoS One* **5**(1):e8529

Wang PL, Bao Y, Yee MC, Barrett SP, Hogan GJ, Olsen MN, Dinneny JR, Brown PO, Salzman J. 2014. Circular RNA Is Expressed across the Eukaryotic Tree of Life. *PLoS One* **9**(3):e90859.

Wang S, Elgin S. 2011. *Drosophila* Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. *Proc Natl Acad Sci* **108**:21164-21169.

Wang X, Su H, Bradley A. 2002. Molecular mechanisms governing *Pcdh*-γ gene expression: evidence for a multiple promoter and *cis*-alternative splicing model. *Genes Dev* **16**(15):1890–1905.

Wang X, Wu Z, Zhang X. 2010. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *J Bioinform Comput Biol* **8** Suppl 1:177–192.

Wang X-H, Aliyari R, Li W-X, Li H-W, Kim K, Carthew R, Atkinson P, Ding S-W. 2006. RNA interference directs innate immunity against viruses in adult *Drosophila*. *Science* **312**:452-454.

Wang Y, Zheng D, Tan Q, Wang MX, Gu LQ. 2011. Nanopore-based detection of circulating microRNAs in lung cancer patients. *Nat Nanotechnol* **6**(10):668-674.

Wang YE, Marinov GK, Wold BJ, Chan DC. 2013. Genome-wide analysis reveals coating of the mitochondrial genome by TFAM. *PLoS ONE* **8**(8):e74513.

Wang XJ, Gaasterland T, Chua NH. 2005. Genome-wide prediction and identification of *cis*-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biol* **6**(4):R30.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**(7):897–903.

Wanrooij S, Falkenberg M. 2010. The human mitochondrial replication fork in health and disease. *Biochim Biophys Acta* **1797**(8):1378–1388.

Ward BL, Anderson RS, Bendich AJ. 1981. The mitochondrial genome is large and variable in a family of plants (*Cucurbitaceae*). *Cell* **25**:793-803.

Ward LD, Kellis M. 2012a. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**:1675–1678.

Ward LD, Kellis M. 2012b. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**(Database issue):D930–934.

Warrior R, Gall J. 1985. The mitochondrial DNA of *Hydra attenuata* and *Hydra littoralis* consists of two linear molecules. *Arch Sci Geneva* **38**:439-445.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24**(4):616–628.

Watanabe A, Arai M, Koitabashi N, Niwano K, Ohyama Y, Yamada Y, Kato N, Kurabayashi M. 2011. Mitochondrial transcription factors TFAM and TFB2M regulate Serca2 gene transcription. *Cardiovasc Res* **90**(1):57–67.

Watanabe H, Wada T, Handa H. 1990. Transcription factor E4TF1 contains two subunits with different functions. *EMBO J* **9**(3):841–847.

Watanabe KI, Bessho Y, Kawasaki M, Hori H. 1999. Mitochondrial genes are found on minicircle DNA molecules in the mesozoan animal *Dicyema*. *J Mol Biol* **286**:645-650.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**(7194):539–543.

Watson JD, Crick FH. 1953a. Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature* **171**:964-967.

Watson JD, Crick FH. 1953b. Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid. *Nature* **171**:737-738.

Watts PC, Lundholm N, Ribeiro S, Ellegaard M. 2013. A century-long genetic record reveals that protist effective population sizes are comparable to those of macroscopic species. *Biol Letters* **9**(6):20130849.

Weatherby K, Carter D. 2013. *Chromera velia*: The Missing Link in the Evolution of Parasitism. *Adv Appl Microbiol* **85**:119–144.

Weber, M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**:853862

Wegrzyn J, Potla R, Chwae YJ, Sepuri NB, Zhang Q, Koeck T, Derecka M, Szczepanek K, Szelag M, Gornicka A, Moh A, Moghaddas S, Chen Q, Bobbili S, Cichy J, Dulak J, Baker DP, Wolfman A, Stuehr D, Hassan MO, Fu XY, Avadhani N, Drake JI, Fawcett P, Lesnefsky EJ, Larner AC. 2009. Function of mitochondrial Stat3 in cellular respiration. *Science* **323**(5915):793–797.

Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martnez-Garca PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, Dejong PJ, Mockaitis K, Main D, Langley CH, Neale DB. 2014. Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. *Genetics* **196**(3):891–909.

Wei G, Abraham BJ, Yagi R, Jothi R, Cui K, Sharma S, Narlikar L, Northrup DL, Tang Q, Paul WE, Zhu J, Zhao K. 2011. Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* **35**(2):299–311.

Wei L, Vahedi G, Sun HW, Watford WT, Takatori H, Ramos HL, Takahashi H, Liang J, Gutierrez-Cruz G, Zang C, Peng W, O'Shea JJ, Kanno Y. 2010. Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. *Immunity* **32**(6):840–851.

Weigert MG, Garen A. 1965. Base composition of nonsense codons in E. coli. Evidence from amino-acid substitutions at a tryptophan site in alkaline phosphatase. *Nature* **206**(988):992–994.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**:235-244.

Weiss R. 2007. Jun 14. Intricate Toiling Found In Nooks of DNA Once Believed to Stand Idle. *The Washington Post*

Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG. 2009. ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J* **28**(10):1418–1428.

Wells J. 2011. The Myth of Junk DNA. *Discovery Institute Press, Seattle.*

Wells J. 2013. Not junk after all: non-protein-coding DNA carries extensive biological information. *Biological InformationNew Perspectives... World* 210–231.

Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R. 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**(7151):318–324.

Westermann AJ, Gorski SA, Vogel J. 2012. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol* **10**(9):618–630.

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, Hansen CL. 2011. High-throughput microfluidic single-cell RT-qPCR. *Proc Natl Acad Sci U S A.* **108**(34):13999–14004.

Whiteford N, Haslam N, Weber G, Prgel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C. 2005. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* **33**(19):e171.

Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA. 2011. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**(7384):221–225.

Wieland C. 1994. Junk moves up in the world. *Creation Ex Nihilo Technical Journal* **8**:125.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**(7199):1239–1243.

Will CL, Luhrmann R. 2005. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem* **386**(8):713–724.

Williams BA, Hirt RP, Lucocq JM, Embley TM. 2002. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature* **418**(6900):865–869.

Williams BAP, Slamovits CH, Patron NJ, Fast NM, Keeling PJ. 2005. A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc Natl Acad Sci U S A* **102**:10936-10941.

Williams GS, Boyman L, Chikando AC, Khairallah RJ, Lederer WJ. 2012. Mitochondrial calcium uptake. *Proc Natl Acad Sci U S A* **110**(26):10479-10486.

Williams RS. 1986. Mitochondrial gene expression in mammalian striated muscle. Evidence that variation in gene dosage is the major regulatory event. *J Biol Chem* **261**26:12390–12394.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**(7479):231–236.

Wilson NK, Foster SD, Wang X, Knezevic K, Schtte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, Pimanda JE, de Bruijn MF, Göttgens B. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**(4):532–544.

Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, Dawson MA, Donaldson IJ, Dumon S, Frampton J, Janky R, Sun XH, Teichmann SA, Bannister AJ, Göttgens B. 2009. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood* **113**(22):5456–5465.

Wold B, Myers RM. 2008. Sequence census methods for functional genomics. *Nat Methods* **5**(1):19–21.

Wolf YI, Koonin EV. 2013. Genome reduction as the dominant mode of evolution. *Bioessays* **35**(9):829–837.

Wolfe J. 1967. Structural aspects of amitosis: a light and electron microscope study of the isolated macronuclei of *Paramecium aurelia* and *Tetrahymena pyriformis*. *Chromosoma* **23**:59-79.

Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CW. 2004. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell* **13**(1):91–100.

Wollmann H, Holec S, Alden K, Clarke ND, Jacques PÉ, Berger F. 2012. Dynamic deposition of histone variant H3.3 accompanies developmental remodeling of the *Arabidopsis* transcriptome. *PLoS Genet* **8**(5):e1002658.

Wood DL, Xu Q, Pearson JV, Cloonan N, Grimmond SM. 2011. X-MATE: a flexible system for mapping short read data. *Bioinformatics* **27**(4):580–581.

Woodfield GW, Chen Y, Bair TB, Domann FE, Weigel RJ. 2010. Identification of primary gene targets of TFAP2C in hormone responsive breast carcinoma cells. *Genes Chromosomes Cancer* **49**(10):948–962.

Woodmorappe J. 2000. Are pseudogenes 'shared mistakes' between primate genomes? *Creation Ex Nihilo Technical Journal* **14**:55–71.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**(1):e7.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* **16**:97-159.

Wrutniak C, Cassar-Malek I, Marchal S, Rascle A, Heusser S, Keller JM, Fléchon J, Dauça M, Samarut J, Ghysdael J, Cabello, G. 1995. A 43-kDa protein related to c-Erb A α1 is located in the mitochondrial matrix of rat liver. *J Biol Chem* **270**(27):16347–16354.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, Quake SR. 2014. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11**(1):41–46.

Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, Sun YE, Zhang Y. 2011a. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* **25**(7):679–684.

Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, Zhao K, Sun YE, Zhang Y. 2011b. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**(7347):389–393.

Wu J, Akerman M, Sun S, McCombie WR, Krainer AR, Zhang MQ. 2011. SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics* **27**:3010-3016.

Wu J, Anczuków O, Krainer AR, Zhang MQ, Zhang C. 2013. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* **41**(10):5149–5163.

Wu JQ, Seay M, Schulz VP, Hariharan M, Tuck D, Lian J, Du J, Shi M, Ye Z, Gerstein M, Snyder MP, Weissman S. 2012. Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS Genet* **8**(3):e1002565.

Wu Q, Maniatis T. 1999. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**(6):779–790.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7):873–881.

Wu Z, Wang X, Zhang X. 2011. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics* **27**(4):502–508.

Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, Lohan AJ, Loftus B, Graciet E, Wellmer F. 2012. Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. *Proc Natl Acad Sci U S A* **109**(33):13452–13457.

Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM. 2002. A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. *Nucleic Acids Res* **30**(11):2524–2237.

Wyatt GR, Cohen SS. 1952. A new pyrimidine base from bacteriophage nucleic acids. *Nature* **170**(4338):1072–1073.

Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Régnault B, Devaux F, Namane A, Séraphin B, Libri D, Jacquier A. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**(5):725–737.

Xia Z, Wen J, Chang CC, Zhou X. 2011. NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics* **12**:162.

Xiao S, Xie D, Cao X, Yu P, Xing X, Chen CC, Musselman M, Xie M, West FD, Lewin HA, Wang T, Zhong S. 2012. Comparative epigenomic annotation of regulatory DNA. *Cell* **149**(6):1381–1392.

Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M. 2013. Dynamic trans-acting factor colocalization in human cells. *Cell* **155**(3):713-724.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GK, Wang J. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 2014 [Epub ahead of print].

Xing D, Wang Y, Xu R, Ye X, Yang D, Li QQ. 2013. The regulatory role of Pcf11-similar-4 (PCFS4) in *Arabidopsis* development by genome-wide physical interactions with target loci. *BMC Genomics* **14**:598.

Xing Y, Lee CJ. 2005. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet* **1**(3):e34.

Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. 2006. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res* **34**(10):3150–3160.

Xu C, Fan ZP, Müller P, Fogley R, DiBiase A, Trompouki E, Unternaehrer J, Xiong F, Torregroza I, Evans T, Megason SG, Daley GQ, Schier AF, Young RA, Zon LI. 2011. Nanog-like regulates endoderm formation through the Mxtx2-Nodal pathway. *Dev Cell* **22**(3):625–638.

Xu K, Doak TG, Lipps HJ, Wang J, Swart EC, Chang WJ. 2012. Copy number variations of 11 macronuclear chromosomes and their gene expression in *Oxytricha trifallax*. *Gene* **505**(1):75–80.

Xu Q, Modrek B, Lee C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30**(17):3754–3766.

Xu S, Zhong M, Zhang L, Wang Y, Zhou Z, Hao Y, Zhang W, Yang X, Wei A, Pei L, Yu Z. 2009. Overexpression of Tfam protects mitochondria against beta-amyloid-induced oxidative damage in SH-SY5Y cells. *FEBS J* **276**(14):3800–3809.

Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H, He W, Zeng L, Xing M, Wu R, Jiang H, Liu X, Cao D, Guo G, Hu X, Gui Y, Li Z, Xie W, Sun X, Shi M, Cai Z, Wang B, Zhong M, Li J, Lu Z, Gu N, Zhang X, Goodman L, Bolund L, Wang J, Yang H, Kristiansen K, Dean M, Li Y, Wang J. 2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**(5):886–895.

Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu JY, Horvath S, Fan G. 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**(7464):593–597.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Mnster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**(7232):1033–1037.

Yaffe D, Saxel O. 1977. Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle. *Nature* **270**(5639):725–727.

Yagi Y, Shiina T. 2014. Recent advances in the study of chloroplast gene expression and its evolution. *Front Plant Sci* **5**:61.

Yakubovskaya E, Chen Z, Carrodeguas JA, Kisker C, Bogenhagen DF. 2006. Functional human mitochondrial DNA polymerase gamma forms a heterotrimer. *J Biol Chem* **281**(1):374–382.

Yamanaka S. 2010. Elite and stochastic models for induced pluripotent stem cell generation. *Nature* **460**(7251):49–52.

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. 1985. Mitochondrial origins. *Proc Natl Acad Sci U S A* **82**(13):4443–4447.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20**(5):614–622.

Yang F, Nickols NG, Li BC, Marinov GK, Said JW, Dervan PB. 2013. Antitumor activity of a pyrrole-imidazole polyamide. *Proc Natl Acad Sci U S A*. **110**(5):1863–1868.

Yang XP, Ghoreschi K, Steward-Tharp SM, Rodriguez-Canales J, Zhu J, Grainger JR, Hirahara K, Sun HW, Wei L, Vahedi G, Kanno Y, O'Shea JJ, Laurence A. 2011. Opposing regulation of the locus encoding IL-17 through direct, reciprocal actions of STAT3 and STAT5. *Nat Immunol* **12**(3):247–254.

Yang Y, Lu Y, Espejo A, Wu J, Xu W, Liang S, Bedford MT. 2010. TDRD3 is an effector molecule for arginine-methylated histone marks. *Mol Cell* **40**(6):1016–1023.

Yanofsky C, Carlton BC, Guest JR, Helinski DR, Henning U. 1964. On The Colineairty of Gene Structure and Protein Structure. *Proc Natl Acad Sci U S A* **51**:266–272.

Yant L, Mathieu J, Dinh TT, Ott F, Lanz C, Wollmann H, Chen X, Schmid M. 2010. Orchestration of the floral transition and floral development in *Arabidopsis* by the bifunctional transcription factor APETALA2. *Plant Cell* **22**(7):2156–2170.

Yao H, Brick K, Evrard Y, Xiao T, Camerini-Otero RD, Felsenfeld G. 2010. Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev* **24**(22):2543–2555.

Yasukawa T, Reyes A, Cluett TJ, Yang MY, Bowmaker M, Jacobs HT, Holt IJ. 2006. Replication of vertebrate mitochondrial DNA entails transient ribonucleotide incorporation throughout the lagging strand. *EMBO J* **25**(22):5358–5371.

Yasukawa T, Yang MY, Jacobs HT, Holt IJ. 2005. A bidirectional origin of replication maps to the major noncoding region of human mitochondrial DNA. *Mol Cell* **18**(6):651–662.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, Nemzer S, Pinner E, Walach S, Bernstein J, Savitsky K, Rotman G. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21**(4):379–386.

Yildirim O, Li R, Hung JH, Chen PB, Dong X, Ee LS, Weng Z, Rando OJ, Fazzio TG. 2011. Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**(7):1498–1510.

Yoon OK, Brem RB. 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA* **16**(6):1256–1267.

Yoon SJ, Wills AE, Chuong E, Gupta R, Baker JC. 2011. HEB and E2A function as SMAD/FOXH1 cofactors. *Genes Dev* **25**(15):1654–1661.

Yoshida Y, Izumi H, Torigoe T, Ishiguchi H, Itoh H, Kang D, Kohno K. 2003. P53 physically interacts with mitochondrial transcription factor A and differentially regulates binding to damaged DNA. *Cancer Res* **63**(13):3729–3734.

Yu GL, Bradley JD, Attardi LD, Blackburn EH. 1990. In vivo alteration of telomere sequences and senescence caused by mutated *Tetrahymena* telomerase RNAs. *Nature* **344**(6262):126–132.

Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min JH, Jin P, Ren B, He C. 2012. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**(6):1368–1380.

Yu M, Mazor T, Huang H, Huang HT, Kathrein KL, Woo AJ, Chouinard CR, Labadorf A, Akie TE, Moran TB, Xie H, Zacharek S, Taniuchi I, Roeder RG, Kim CF, Zon LI, Fraenkel E, Cantor AB. 2012. Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. *Mol Cell* **45**(3):3303–3343.

Yu M, Riva L, Xie H, Schindler Y, Moran TB, Cheng Y, Yu D, Hardison R, Weiss MJ, Orkin SH, Bernstein BE, Fraenkel E, Cantor AB. 2009. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36**(4):682–695.

Yu S, Cui K, Jothi R, Zhao DM, Jing X, Zhao K, Xue HH. 2010. GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells. *Blood* **117**(7):2166–2178.

Yuan P, Han J, Guo G, Orlov YL, Huss M, Loh YH, Yaw LP, Robson P, Lim B, Ng HH. 2009. Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev* **23**(21):2507–2520.

Yun K, Wold B. 1996. Skeletal muscle determination and differentiation: story of a core regulatory network and its context. *Curr Opin Cell Biol* **8**(6):877–889.

Yunis JJ, Yasmineh WG. 1971. Heterochromatin, satellite DNA, and cell function. *Science* **174**:1200-1209.

Zahler AM, Neeb ZT, Lin A, Katzman S. 2012. Mating of the stichotrichous ciliate *Oxytricha trifallax* induces production of a class of 27 nt small RNAs derived from the parental macronucleus. *PLoS One* **7**:e42371.

Zahn K, Blattner FR. 1987. Direct evidence for DNA bending at the lambda replication origin. *Science* **236**(4800):416–422.

Zalzman M, Falco G, Sharova LV, Nishiyama A, Thomas M, Lee SL, Stagg CA, Hoang HG, Yang HT, Indig FE, Wersto RP, Ko MS. 2010. Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature* **464**(7290):858–863.

Zambon RA, Vakharia VN, Wu LP. 2006. RNAi is an antiviral immune response against a dsRNA virus in *Drosophila melanogaster*. *Cell Microbiol* **8**:880-889.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**(15):1952–1958.

Zaphiropoulos PG. 1996. Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping. *Proc Natl Acad Sci U S A* **93**(13):6536–6541.

Zaphiropoulos PG. 1997. Exon skipping and circular RNA formation in transcripts of the human cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol Cell Biol* **17**(6):2985–2993.

Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**(21):2227–2241.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**(5980):916-919.

Zeng PY, Vakoc CR, Chen ZC, Blobel GA, Berger SL. 2006. In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. *Biotechniques* **41**(6):694, 696, 698.

Zenklusen D, Larson DR, Singer RH. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* **15**(12):1263–1271.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**(5):821-829.

Zhang B, Horvath S. 2005. A General Framework for Weighted Gene Co-expression Network Analysis. *Stat Appl Genet Mol Biol 2005*, **4**:Article 17.

Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res* **20**(5):646–654.

Zhang H, Dungan CF, Lin S. 2011. Introns, alternative splicing, spliced leader trans-splicing and differential expression of *pcna* and *cyclin* in *Perkinsus marinus*. *Protist* **162**(1):154–167.

Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci U S A* **104**(11):4618–4623.

Zhang X, Borevitz JO. 2009. Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics* **182**:943-954.

Zhang X, Robertson G, Krzywinski M, Ning K, Droit A, Jones S, Gottardo R. 2011. PICS: probabilistic inference for ChIP-seq. *Biometrics* **67**(1):151–163.

Zhang Y, Lameijer EW, 't Hoen PA, Ning Z, Slagboom PE, Ye K. 2012. PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics* **28**(4):479–486.

Zhang Y, Laz EV, Waxman DJ. 2011. Dynamic, sex-differential STAT5 and BCL6 binding to sex-biased, growth hormone-regulated genes in adult mouse liver. *Mol Cell Biol* **32**(4):880–896.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9):R137.

Zhang Y, Mayba O, Pfeiffer A, Shi H, Tepperman JM, Speed TP, Quail PH. 2013. A quartet of PIF bHLH factors provides a transcriptionally centered signaling hub that regulates seedling morphogenesis through differential expression-patterning of shared target genes in *Arabidopsis*. *PLoS Genet* **9**(1):e1003244.

Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J. 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**(5):908-921.

Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature* **400**:155-159.

Zhang Z, Green BR, Cavalier-Smith T. 2000. Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: a possible common origin for sporozoan and dinoflagellate plastids. *J Mol Evol* **51**:26-40.

Zhao B, Zou J, Wang H, Johannsen E, Peng CW, Quackenbush J, Mar JC, Morton CC, Freedman ML, Blacklow SC, Aster JC, Bernstein BE, Kieff E. 2011a. Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. *Proc Natl Acad Sci U S A* **108**(36):14902–14907.

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. 2010. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**(6):939–953.

Zhao L, Glazov EA, Pattabiraman DR, Al-Owaidi F, Zhang P, Brown MA, Leo PJ, Gonda TJ. 2011b. Integrated genome-wide chromatin occupancy and expression analyses identify key myeloid pro-differentiation transcription factors repressed by Myb. *Nucleic Acids Res* **39**(11):4664–4679.

Zheng B, Chen X. 2011. Dynamics of histone H3 lysine 27 trimethylation in plant development. *Curr Opin Plant Biol* **14**(2):123–129.

Zheng Q, Rowley MJ, Böhmdorfer G, Sandhu D, Gregory BD, Wierzbicki AT. 2012. RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *Plant J* doi: 10.1111/tpj.12034. [Epub ahead of print]

Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, Macneil DJ, Weingarth DT, Zhang B, Greenawalt D, Dobrin R, Hao K, Woo S, Fabre-Suver C, Qian S, Tota MR, Keller MP, Kendziorski CM, Yandell BS, Castro V, Attie AD, Kaplan LM, Schadt EE. 2010. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* **6**(5):e1000932.

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HY, Preston E, Slightham C, Hillier LW, Brock T, Agarwal A, Auerbach R, Hyman AA, Gerstein M, Mango SE, Kim SK, Waterston RH, Reinke V, Snyder M. 2010. Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet* **6**(2):e1000848.

Zhong X, Hale CJ, Law JA, Johnson LM, Feng S, Tu A, Jacobsen SE. 2012. DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nat Struct Mol Biol* **19**(9):870–875.

Zhou ZX, Zhang MJ, Peng X, Takayama Y, Xu XY, Huang LZ, Du LL. 2013. Mapping genomic hotspots of DNA damage by a single-strand-DNA-compatible and strand-specific ChIP-seq method. *Genome Res* **23**(4):705–715.

Zimin AV, Marais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**(21):2669–2677.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB, Salzberg SL, Yorke JA, Langley CH. 2014. Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* **196**(3):875–890.

Zipursky SL, Sanes JR. 2010. Chemoaffinity Revisited: Dscams, Protocadherins, and Neural Circuit Assembly. *Cell* **143**(3):343–353.

Zonneveld BJM. 2009. The systematic value of nuclear genome size for "all" species of *Tulipa L.* (Liliaceae). *Plant Systematics Evolution* **281**:217–245.

Zonneveld BJM. 2010. New record holders for maximum genome size in eudicots and monocots. *J Botany* **2010**:527357

Zuckerkandl E. 1976. Gene control in eukaryotes and the C-value paradox: "Excess" DNA as an impediment to transcription of coding sequences. *J Mol Evol* **9**:73–104.

Zuckerkandl E. 1997. Junk DNA and sectorial gene expression. *Gene* **205**:323–343.

Zylber E, Vesco C, Penman S. 1969. Selective inhibition of the synthesis of mitochondria-associated RNA by ethidium bromide. *J Mol Biol* **44**(1):195–204.