STRUCTURE AND EVOLUTION OF MAMMALIAN GENE NETWORKS

Thesis by

Ali Mortazavi

In Partial Fulfillment of the Requirements for the

degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2008

(Defended May 15th, 2008)

ACKNOWLEDGEMENTS

I had great doubts upon graduating with a B.S. from Caltech in 1993 that I would ever be able to pursue a Ph.D. in science given my undistinguished undergraduate academic record. Hence this thesis is nothing short of a small miracle and most of the credit for it belongs to my co-workers, friends and family who cajoled me over the years to go back to graduate school and urged me to finish promptly.

My advisor, Barbara Wold, boldly pushed me routinely beyond what I thought was capable of. I learned a tremendous amount from our multitude of conversations and arguments about the analyses of our data and science in general, which is a gift of wisdom that too few of my fellow graduate students seem to get from their graduate advisors. No analysis or figure was ever good enough for Barbara before the tenth version or so, but her perfectionism elevated the quality and significance of our results tremendously. Her total dedication to Science, with a big S, has set a high standard for me to emulate in my own career and I am forever grateful that she gave me a chance to redeem myself at Caltech as a graduate student.

I would also like to thank my thesis committee. I have tremendously enjoyed learning genetics and working on *C. elegans* genomics with Paul Sternberg, and Ellen Rothenberg's love of science, and T-cells in particular, was always infectious. As for Eric Davidson, his books and exquisite body of work have been the organizing framework and driving force of my interests in biology, even while we agree to disagree on how to get there from time to time.

I want to thank Rick Myers of Stanford University and now at the HudsonAlpha Institute for Biotechnology for the tremendous amount of help and support that he gave me when visiting and collaborating with them on my first two papers. I would probably be still struggling to publish my first paper if it weren't for all of the help of Rick and his lab members gave me, in particular from Evonne Chen Leeper as well as David Johnson and his Stanford ENCODE team.

I want to thank all of the members of the Wold and Sternberg group with which I had tremendous interactions with over the years. In particular, Brian Williams and Steven Kuntz helped me tremendously with discussions and advice.

Stepping back a little bit, I am grateful to the professors at Cal State LA that helped me launch me on my graduate adventure before returning to Caltech. Dr Momand and Dr Sharp ran the exciting Souhern California Bioinformatics summer institute that gave me my first foray into bioinformatics and sent several bright interns my way every summer for help with our projects, such as Sarah Aerni and Qing Yuan. And Dr Vellanoweth was a wonderful Master's advisor who taught me to enjoy biochemistry and to think critically when reading the scientific literature!

My family has provided me with tremendous, unconditional support over all of these years and enabled me to return to graduate school. Gorgy (Fig 3.14) is the reason why I became interested in evolution and returned in biology. My sisters Azar and Arezou have always been wonderful and supportive. My mother-in-law, Chandra Maharaj, prayed for my success constantly and has selflessly put up with my insane schedule. My mother and father, Manijeh and Seyed, have nurtured me with a love of learning from my earliest childhood, when they were both pursuing their PhD's; It is because of my childhood during their graduate studies that I knew that I wouldn't be happy until I returned to academia and earned mine. They too have put up gamely with my relentless schedule.

And last, but not least, my lovely wife Shantika is the person most responsible for this thesis. My partner in all things, she went with me through every emotional up and down of every paper and has sacrificed a tremendous amount of time and effort for me to do well in graduate school. She has always expected the best of me, and I would not have started, nor finished, a PhD. in biology at Caltech without her love and dedication. I am proud and consider myself blessed to be married to her. Shantika is my better half, and I dedicate this thesis to her.

ABSTRACT

Accurate measurements of protein:DNA and RNA expression levels are critical to building meaningful models of gene regulatory networks. We develop here two new techniques doing such measurements using ultra-high-throughput DNA sequencing combined with extensive computational analyses, which we call respectively ChIP-seq and RNA-seq. To show the power and versatility of these techniques, we apply them to the study of two model problems that are representative of the research agenda of regulatory biology. We use ChIP-seq to study the conservation and evolution of the binding repertoire of the transcription factor NRSF/REST in boreoeutherian mammals, whereas we use ChIP-seq of RNA Polymerase II phosphoisoforms and RNA-seq to study a developmental time course of myogenesis in the C2C12 mouse cell line. Together, ChIP-seq and RNA-seq show the promise of ultra-high-throughout sequencing in mapping and studying gene regulatory networks which will likely supplant the previous generation of microarray-based technologies as the new generations of sequencers mature and become more generally available.

TABLE OF CONTENTS

NOMENCLATURE

Cohort. A set of sites or genes that are analyzed as a group

ChIP. Chromatin Immunoprecipitation. A common technique for enrichment of DNA fragments bound by a protein using antibodies

ChIP-seq. ChIP experiment assayed with ultra-high throughput-sequencing

Cistematic. Software library written in Python for analyzing genes and motifs on a genome-scale

CTD. C-Terminal Domain of RNA polymerase II that is highly conserved in Eukaryotes and consisting of the repeated heptad YSPTSPS. It can be phosphorylated on every Serine

ERANGE. Enhanced Read Analysis of Gene Expression. A set of programs written on Python to analyze both ChIP-seq and RNA-seq data (dual-use)

GO. Gene Ontology. A set of structured vocabulary terms describing the function of a gene

Motif. A representation of a set of sequences with a common pattern

Multiread. Read that maps equally well to more than one position onto the genome

NRSE. Neuron Restrictive Silencer Element. The binding site of NRSF

NRSF. Neuron Restrictive Silencer Factor

Polymerase stalling. Distinctive accumulation of RNA polymerase II at the promoter of genes with little or no productive poly-A RNA output

PSFM. Position Specific Frequency Matrix. A matrix representation of a motif that is closely related to Position Weight Matrices (PWM)

Read. Short 25–35 bp long sequence of DNA

RNA-seq. Ultra-high-throughput sequencing of RNA

Ultra-high-throughput sequencing. Sequencing of millions of short reads of DNA in parallel

Uniquely mappable read. Read that maps best to only a single position on the genome

Chapter 1

INTRODUCTION

The regulation of gene expression is the central, foundational problem of regulatory biology. While post-transcriptional and post-translational processes play a critical role in modulating the expression level of any RNA and of its derived protein, their absence or presence in a cell is ultimately dictated by transcription factor proteins that interact with DNA to recruit or repel the eukaryotic transcriptional machinery. The last 40 years have revealed a stunningly elegant and combinatorial use of multiple transcription factors that respond to the current regulatory state as well as extra-cellular signaling cues to control transcription of any gene with the necessary precision in time and space according to both its lineage and its environment (Davidson, 2006). While much of our current model of transcriptional regulation is built on a multitude of studies of particular aspects of this process at a few genes at a time, the availability of sequenced genomes and of ultra-high-throughput methods affords us also an approach to the problem from a genome-wide, top-down perspective of: what constitutes the repertoire of all genes regulated by a single transcription factor, as well as how cells turn on gene expression in a coordinated fashion upon differentiation. But any successful top-down approach is dependent on both measurements that are more accurate than those used to date and new computational methods to analyze and integrate the data.

This thesis uses a new generation of ultra-high-throughput DNA sequencing technologies to analyze with unmatched accuracy both the binding repertoire of a multi-faceted, vertebrate transcription factor — the Neuron Restrictive Silencer Factor (NRSF) — and the transcriptome of the C2C12 mouse myogenic cell line through differentiation. These analyses shed light respectively on the evolution of the binding repertoire of a transcription factor over 100 million years of evolution, as well as some of the key general transcriptional changes that accompany the differentiation of muscle myoblasts into myotubes.

NRSF as a tractable model of the evolution of gene regulatory networks

The Neuron Restrictive Silencer Factor, also known as the RE-1 Silencing Transcription factor (NRSF/REST), was discovered simultaneously by the Anderson and Mandel labs (Schoenherr, 1995; Chong, 1995) and has accumulated an impressive and rapidly expanding literature that makes it one of the most studied vertebrate-specific transcriptional repressors. The ever-expanding roles of NRSF beyond neurogenesis as a tumor suppressor (Westbrook, 2008; Westbrook, 2005), guardian of genome stability (Guardavaccaro, 2008), and necessary factor in maintaining the pluripotency of embryonic stem cells (Singh, 2008) have significantly increased its profile across different subfields of biology.

Several features of NRSF make it particularly interesting for us to study in the context of the evolution of gene regulatory networks. NRSF has a large binding site (NRSE), which makes it practical for us to identify its target binding site in all sequenced genomes. While the co-repressors of NRSF are present in invertebrates, there is no convincing evidence to date that NRSF is present in invertebrates (Dallman, 2004), and the canonical NRSE is missing from their genomes (Mortazavi, 2006). However, all available vertebrate genomes have a copy of both NRSF and an enrichment of NRSEs in genes associated with a neurosecretory phenotype. There appears to be no functional paralogs of NRSF in any of the genomes and the DNA binding site has remained extremely conserved in all vertebrates. NRSF is thus a tractable example of a vertebrate-specific evolution in gene regulation which affords us the opportunity to quantify the change in its target repertoire using both computational and in vivo experiments.

The discovery of the Neuron Restrictive Silencer Element (NRSE) and of its cognate binding factor NRSF

Early independent studies of the transcriptional regulation of neuronal genes such as Superior Cervical Ganglion 10 (SCG10, now known as Stathmin2) and the sodium channel Nav1.2 (now Scn2a) revealed the presence of a repressive element proximal to their promoters which restricted the expression of these genes outside of neurons. Unlike most cis-regulatory elements that are 6–10 bp long, the repressive element, which was called the Neuron Restrictive Silencer Element (NRSE, also RE-1 for Repressive Element 1) was determined to be nearly 21–23 bp long (Mori, 1992; Kraner, 1992). This element was used to identify NRSF as a zinc finger transcription factor that bound the canonical NRSE/RE-1, that repressed constructs with the NRSE, and whose expression pattern was predominantly non-neuronal. The full-length NRSF is thought to bind its cognate site without the combinatorial assist of other transcription factors. While this makes it unusual in the context of the current transcriptional literature, NRSF and its fellow canonical zinc finger CTCF (Lobanenkov, 1990; Filipova, 2008) may be representative of the large family of zinc finger repressors that have been expanding rapidly in the mammalian and primate lineages (Huntley, 2006).

An early survey of available genomic sequences revealed that the NRSE was embedded in a host of frankly neuronal genes such as BDNF, but was also found in the proximity of other genes such as skeletal muscle actin and the hormone somatostatin (Schoenherr, 1996). Nevertheless, NRSF was proposed as master regulator of neuronal fate through a control of neuronal gene batteries (Schoenherr, 1996). However, the early embryonic lethal knockout failed to reveal any ectopic neurogenesis or even mis-expression of the several known targets at the time (Chen, 1998), while the neurogenic ability of several bHLH factors with similar functional homologs in Drosophila, such as NeuroD, were demonstrated conclusively (Lee, 1995). As the primacy of the activators in neurogenesis took hold, NRSF was demoted to being an eccentric repressor of some neuronal genes.

The interaction of NRSF with its co-repressors

The next decade of NRSF research focused on the biochemistry of NRSF and of its repression, which is accomplished by the recruitment of 3 different complexes. The N-terminus recruits the ubiquitous eukaryotic repressor mSin3a (Huang, 1999). The C-terminus recruits another repressor named CoREST, which has attracted much attention because it also shares a similar expression pattern that is non-neuronal (Andres, 1999). CoREST works in conjunction with HDAC2 and, while recruited by NRSF, it is thought to stay behind after the NRSF protein is degraded (see below) and to continue repression of its targets until a further de-repression event that is CoREST specific (Ballas, 2005). Small CTD Phosphatases (SCP) are a third family of NRSF co-repressors that have recently been identified that also play a role in the repression of NRSF targets (Yeo, 2005). SCP family members work by dephosphorylating the 5th serine of the heptad repeats of C-terminal domain of RNA polymerase II, which would either prevent initiation or promote the early termination of transcription. All 3 modes of repression (mSin3A, CoREST, and SCP mediated) are known to repress different genes (Ballas, 2005; Lunyak, 2001; Yeo, 2005). It is not known what determines the recruitment and/or the selectivity of the different co-repressors to their appropriate targets.

NRSF splice isoforms affect its DNA binding domain and likely function

A parallel series of efforts determined that NRSF can be found in a variety of splice isoforms that includes three different 5' UTRs, the N-terminal (sin3A-interacting) domain, and different fractions of the DNA-binding domain and C-terminal (CoREST interacting) domains. In particular, some neurons and cancers express a splice isoform, called REST4, that include zinc fingers 1 through 5 (Shimojo, 2001). Mutagenesis studies of the different fingers of NRSF showed that fingers 5, and 6 through 8, are most important for binding to the full NRSE, and that REST4 binds much more weakly to the right half of the canonical NRSE (Shimojo, 2001; Lee, 2000). NRSF was shown early on to be first degraded at the protein level before being transcriptionally shutdown. Recent work shows that NRSF is degraded by the ubiquitin conjugating enzyme βTRCP, which recognizes a degron present in the C-terminal domain of NRSF and which is absent in REST4 (Westbrook, 2008). While the role of REST4 as an activator has been debated over the years, given conflicting results (Magin, 2002), a recent report shows that whereas full-length NRSF is a repressor of glutamine synthetase, REST4 interacts with the glucocorticoid receptor (GR) to activate the transcription of the same gene (Abramovitz, 2008). It is unknown to what extent REST4 would activate the remainder of the target repertoire of NRSF and whether REST4 or full-length NRSF can act as activators in conjunction with other transcription factors besides GR.

The NRSF gene target repertoire

The availability of whole genome sequences rejuvenated the search for NRSF targets through computational means. Genomic scans with the NRSE consensus (Schoenherr, 1996; Bruce, 2004) or more sensitive position-specific-frequency-matrix-like methods (Mortazavi 2006; Zhang, 2006) revealed that NRSE was associated with a large, specific subset of genes highly enriched for neuronal expression, including several neuronal transcription factors and RNA splicing factors. Another insight from computational studies that was quickly confirmed in vivo related NRSF-mediated repression to the expression of neuron-specific microRNAs, such as miR-124 and mir-9 (Conaco, 2006; Johnson, 2008). Thus NRSF-derepression is necessary for the coordinate expression of neuronal-specific genes, microRNAs, and splicing factors, which together change dramatically the regulatory state of cells commited to neurogenesis (Lim, 2005). Whereas NRSF knockouts or dominant negatives (without the N- and C-terminal domains) cannot transform non-neuronal cells into neurons, an NRSF-VP16 activator does transform C2C12 myoblasts into neuron-like cells (Watanabe, 2004). Hence NRSF de-repression is a necessary step in neurogenesis, although it is not sufficient without the concomitant action of neurogenic activators such as those of the NeuroD1 activator.

The availability of a large list of true positive and of several good chromatin immunoprecipitation (ChIP) "grade" antibodies made NRSF an ideal candidate for a genome-wide survey of its in vivo binding across the genome. Whereas we surveyed the NRSF binding repertoire using a novel, high-resolution technique called ChIPSeq (also ChIP-seq) as described in Chapter 3, the Mandel lab and their collaborators independently surveyed the same repertoire using a related method called SACO (Otto, 2007). Both surveys revealed that NRSF had a wider set of targets than expected based on the computational surveys alone, due to the fact that NRSF had an expanded family of split sites, where the two halves of the canonical NRSE are separated by a spacer that is preferentially between 5–8 bp and the repressive ability of these sites was soon verified (Otto, 2007; Patel, 2007). In addition to additional members of known gene family targets, these genome-wide surveys revealed novel NRSEs, such as the one in the exon NeuroD1, which provides a clear link between NRSF-mediated repression and the neurogenic phenotype of NRSF-VP16. The new targets also pointed to a potential role for NRSF in the specification of pancreatic islet cells.

NRSF as a regulator of the neurosecretory phenotype

While NRSF is found associated with a variety of neuronal gene ontology terms, much attention has been given to the particular enrichment in genes that are important in neurosecretion (Bruce, 2006). In particular, NRSF expression in rat PC12 cells blocks secretion (Pance, 2006; D'alessandro, 2008) and beta-cell specific expression of NRSF (see below) reduces insulin secretion (Martin, 2008). Combined with the lack of expression of full-length NRSF in some cancers with neurosecretory phenotypes such as small-cell lung cancer (Coulson, 2000), this suggests that NRSF may play a critical role in allowing the co-option of the neuroendocrine phenotype in novel, non-neuronal settings. Several studies have demonstrated a role for NRSF in the cardiac program as well as a potential role in the specification of pancreatic islet cells (Kuwahara, 2003; Atouf, 1997; Kemp, 2003), neither of which are part of the ectodermal lineage. However, several NRSF targets such as somatostatin are expressed in endocrine cells besides the nervous system and islet cells such as gastric and intestinal endocrine cells. We would thus expect that NRSF de-repression to turn out as a prerequisite for the development of any endocrine cells, such as somtatin-producing cells, or indeed any cell found to express neurosecretory gene markers such as SNAP-25. NRSF would thus permit non-ectodermal lineages to acquire this neurosecretory gene battery, given the appropriate expression of activators that would only turn on parts of the neurogenesis program. Since NRSF must be absent in order to have a neuroendocrine phenotype, and therefore cannot play its other role as a tumor surpressor, these cells will likely turn out to be more susceptible to uncontrolled proliferation that ultimately could turn into neuroendocrine tumors. Thus future studies will likely return to the topic of re-expressing NRSF and maintaining the expression of full-length NRSF to control proliferation (Fuller, 2005).

Theme of Thesis

This thesis develops the techniques and principles to analyze gene regulatory networks in mammalian genomes. In Chapter 2, I show how to search for transcription factor binding sites with and without restrictions of conservation, and how to analyze the resulting set of genes using gene ontology analysis and microarray gene expression data. I also show that the canonical NRSEs are absent in invertebrates and that NRSF is predicted to regulate both neuronal microRNAs, as well as alternative splicing factors. Evonne Chen Leeper and

Sarah T. Garcia of the Rick Myers lab at Stanford did the ChIP validation of a subset of our predicted sites. Chapter 2 was published in *Genome Research* in 2006 (Mortazavi, 2006). Our collaboration with the Myers lab continued in Chapter 3 with the development of ultra-high-throughput sequencing of ChIP (ChIP-seq) as a replacement for previous microarray methods (ChIP-chip) and in its application in identifying the binding repertoire of NRSF in the human jurkat cell line. While Dave Johnson of the Myers lab and his team of technicians developed and performed all of the original ChIP-seq protocols in human, I performed mouse and dog ChIP-seq experiments and did all of the data analysis in Chapter 3 (which is an expanded version of our human-centric 2007 publication in *Science* (Johnson, 2007) that includes massive updates to account for the comparative analysis). Chapter 4 switches gear to the application of ultra-high-throughput sequencing to the analysis of the transcriptome of multiple tissues and of the C2C12 myogenic model, by sequencing of the polyA-selected RNA (RNA-seq) and ChIP-seq of the RNA polymerase II phosphoisoforms to correlate Pol II occupancy to the observed mRNA. Brian Williams did all of the wet-bench development of RNA-seq, Lorian Schaeffer built and sequenced the bulk of the libraries, and Ken McCue developed the RNA-seq read uniformity metrics. I developed the entire RNA-seq analysis pipeline, designed and did the actual polymerase ChIP-seqs, and did the integrated RNA-seq and polymerase analysis. Chapter 4 will be published as two separate publications, one focusing on RNA-seq in a methods-oriented paper (Mortazavi, 2008), and another focused on the C2C12 differentiation story and the polymerase stalling changes. Chapter 5 concludes with a brief discussion of the future of studies of gene-regulation using the methods described herein.

Chapter 2

COMPARATIVE GENOMICS MODELING OF THE NRSF/REST REPRESSOR NETWORK: FROM SINGLE CONSERVED SITES TO GENOME-WIDE REPERTOIRE

Abstract

We constructed and applied an open source informatic framework called Cistematic in an effort to predict the target gene repertoire for transcription factors with large binding sites. Cistematic uses two different evolutionary conservation-filtering algorithms in conjunction with several analysis modules. Beginning with a single conserved and biologically tested site for the neuronal repressor NRSF/REST, Cistematic generated a refined PSFM (position specific frequency matrix) based on conserved site occurrences in mouse, human, and dog genomes. Predictions from this model were validated by chromatin immunoprecipitation (ChIP) followed by quantitative PCR. The combination of transfection assays and ChIP enrichment data provided an objective basis for setting a threshold for membership and rank-ordering a final gene cohort model consisting of 842 high-confidence sites in the human genome associated with 733 genes. Statistically significant enrichment of NRSE-associated genes was found for neuron-specific Gene Ontology (GO) terms and neuronal mRNA expression profiles. A more extensive evolutionary survey showed that NRSE sites matching the PSFM model exist in roughly similar numbers in all fully sequenced vertebrate genomes but are notably absent from invertebrate and protochordate genomes, as

is NRSF itself. Some NRSF/REST sites reside in repeats, which suggests a mechanism for both ancient and modern dispersal of NRSEs through vertebrate genomes. Multiple predicted sites are located near neuronal microRNA and splicing factor genes, and these tested positive for NRSF/REST occupancy in vivo. The resulting network model integrates post-transcriptional and translational controllers, including candidate feedback loops on NRSF and its co-repressor, CoREST.

The Cistematic source code and associated databases are available at http://cistematic.caltech.edu. All data in this paper, as well as the scripts used to generate them, can be found at http://cistematic.caltech.edu/~alim/cispaper.

Introduction

Specific repressors, such as canonical zinc finger transcription factors, stand out in vertebrate genomes because of their large number, significant expansion in mammals, and the diversity of cellular and organismic functions they affect (Hamilton, 2003). The Krab family of zinc finger sequence specific DNA-binding repressors, for example, numbers over 400 in rodent and human genomes (Dehal, 2001; Shannon, 2003). For the vast majority of these, nothing is known about their target gene repertoire or binding motif. A few, studied in more detail, play important roles in diverse cellular and organismic functions ranging from regulation of rodent male specific genes by the Rsl (regulator of sex limitation) Krab repressors (Krebs, 2005), to lipid metabolism and possible predisposition to hypoalphalipoproteinemia by znf202 (Wagner, 2000). Much more is known about NRSF/REST, a zinc finger repressor famous for negative regulation of neuronal genes in non-neuronal cell types, and in neuronal stem cells and progenitors prior to differentiation (Schoenherr, 1995; Chong, 1995; Chen, 1998). The main isoform of NRSF represses transcription by recruiting cofactors such as CoREST (Andres, 1999), CTD phosphatases (Yeo, 2005), mSin3A, and histone deacetylases (Huang, 1999). Another isoform, REST4, is thought to act in a dominant negative fashion (Hersh, 2003). In addition to neuronal development, NRSF/REST may have other roles in cardiac development (Kuwahara, 2003), pancreatic islet development (Atouf, 1997; Abderrahmani, 2001), and perhaps B- or T-cell lineages (Scholl, 1996). Little is known about which genes affecting these non-neuronal lineages are direct NRSF/REST targets, or how many overlap with the neuronal set.

A first step toward understanding how a regulator fits into the design logic and function of a gene network is to define its genome-wide target gene set. In multicellular animals and plants, this is not easily done by direct experimental measurements, because the matrix of all possible target DNA sites, across many tissues and developmental states, is so vast. An alternate starting point is to use comparative genomics, constrained by some smaller sets of functional data, to generate a computational genome-wide model that can then be tested directly and interrogated to develop new focused hypotheses.

Two considerations make the NRSF/REST repressor a superior candidate for this analysis. First, factors with tandem arrays of zinc fingers can recognize relatively long and specific target motifs, and this makes computational approaches for finding target genes more feasible. Specifically, NRSF has a 21 bp binding site (NRSE or RE-1), and much is known about where and how NRSEs function. They can direct repression from positions within 5'-UTRs, in introns and at intron/exon junctions, as well as upstream of the transcription start and downstream of the coding stop (Schoenherr, 1996; Thiel et al., 1998). One study also reported that repression can extend to neighboring genes at one locus, although it is not clear whether this is general or not (Lunyak, 2002). NRSF transcriptional repression also appears to be tuned in vivo for strength and timing at different target genes during the progression from pluripotent stem cell to differentiated neuron or glial cell (Kuwabara, 2004; Ballas, 2005). It is not known whether these distinctions, so far studied for only a few genes, reflect differences in the sequence, number, or organization of NRSE sites.

The second virtue of NRSF/REST for genome-wide target prediction is that a collection of NRSF sites has been quantitatively assayed for activity in vivo (Schoenherr, 1996; Bruce, 2004). These assays, which include sequences that resemble the consensus binding site but lack function, are invaluable for calibrating and interpreting any model of NRSF binding derived by other criteria, including evolutionary conservation of NRSE occurrences.

In addition to direct transcriptional regulation, post-transcriptional and translational mechanisms mediated by microRNAs are implicated in neurogenesis. Cells undergoing terminal differentiation express tissue-specific microRNAs that are currently thought to modulate translation and/or degradation of large networks of target mRNAs (reviewed in Kosik, 2005). miR-124a, for example, is neuron-specific and can target hundreds of genes when expressed in HeLa cells (Lim, 2005). A broad survey of microRNA expression in brain and neuronal cell culture (Sempere, 2004) suggests there are at least a dozen different microRNAs that are predominantly expressed in the brain. While prediction of likely target sites in 3' UTRs of known mRNAs has been very active (John, 2004; Krek, 2005; Lewis, 2005), little is known about how microRNAs are themselves transcriptionally regulated, except that microRNAs located within introns of protein coding genes tend to be expressed along with their “host” gene (reviewed in Ying, 2004). This emerging picture raises the question of how transcriptional regulators are connected to and coordinated with the post-transcriptional ones.

In the first part of this work, we use NRSF/REST as an amenable test case to build a comprehensive genome-wide model for the corresponding gene cohort. To do this, we develop a set of generally applicable algorithms and open source software tools (Cistematic) to make and refine site predictions and enumerate the target gene cohort. We show it is possible to begin with a single biologically defined, evolutionarily conserved NRSF/REST site, then use conservation among mouse, human and dog genomes to develop a refined model for NRSF sites. The resulting model is compared and contrasted with prior ones (Schoenherr, 1996; Bruce et al., 2004), and we show that the major known functions of NRSF can be deduced computationally by using RNA expression and GO analysis modules in Cistematic. We test our model by experimentally measuring in vivo binding at 113 loci by chromatin immunoprecipitation followed by Q-PCR. In the second part of the study, we use the PSFM model to investigate evolution of the NRSF network over much greater evolutionary distances, and to develop and test specific hypotheses about links between NRSF/REST and post-transcriptional regulatory pathways. High confidence candidate sites near neuronal microRNAs and splicing factors are identified, and in vivo interaction of NRSF at these loci is experimentally verified.

Experimental outline

The availability of multiple whole-genome DNA sequences raises the possibility of building strong predictive models for the entire binding site repertoire of a sequence-specific DNA binding factor by leveraging preferential conservation of functionally important sites. We used a two-part strategy that begins by deriving and refining a PSFM model for the binding site. The starting point is one or more functionally tested and conserved instances to seed a multi-genome search for additional conserved instances. The cisMatcher algorithm used to do this is designed to focus on site instances that are embedded in somewhat larger conserved domains. The reasoning is that functional sites are often located within larger conserved cis-regulatory modules. At later times in the process, one can exercise an option to recover other instances of the site that do not require conservation beyond the boundaries of the site model. The second process develops a model for the genome-wide cohort of genes associated with sites defined by the fact that they match the PSFM at or above a specified score. At this stage, various conservation and gene geography criteria are selected and applied. They can require, for example, that PSFM match sites occur near orthologs in multiple genomes and that candidate cohort genes be located within a specified distance of a PSFM match site. Thus the refined PSFM from the first part of the process is used to interrogate the genome(s) to find which genes are located near site instances. The PSFM match score, coupled with archival and new experimental data, is then used to help establish an appropriate threshold for inclusion in target gene cohort.

Deriving and refining a conservation-based PSFM site model.

The Cistematic pipeline is outlined in Figure 2.1 and summarized here, with details in Methods. In one experiment, the derivation pipeline was initiated with orthologs from a single gene, SCG10 (STMN2) in human, mouse, and dog genomes (Mori, 1992; Schoenherr, 1995). This seed PSFM was used to run a genome-wide search that used the cisMatcher algorithm. It collected additional similar instances that occur in domains of conservation (here set for 87.5% PSFM match and 85% similarity in a 25–65bp window) shared by at least two of the three participating genomes (Fig 2.2 and below). These conserved occurrences (81) of the motif were then used to derive a refined SCG10 PSFM, which we call NRSE2. In a second experiment, by contrast, we began with a collection of

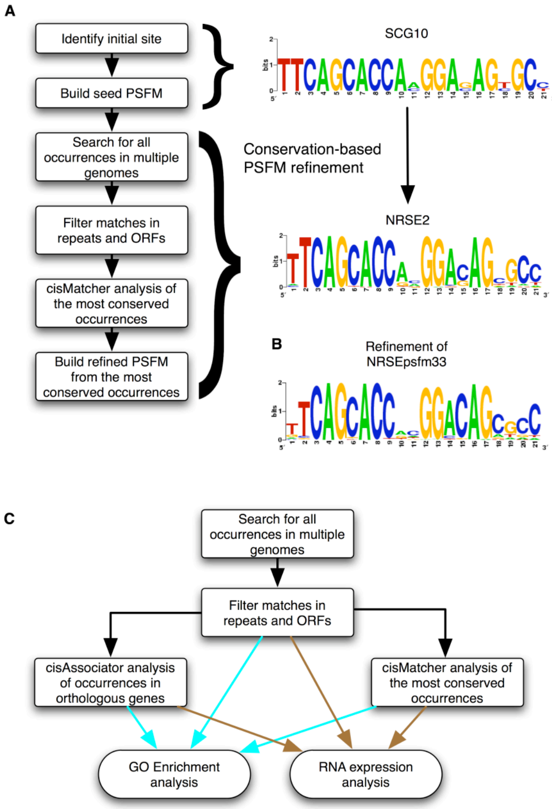

A
Identify initial site
Build seed PSFM
Search for all occurrences in multiple genomes
Filter matches in repeats and ORFs
cisMatcher analysis of the most conserved occurrences
Build refined PSFM from the most conserved occurrences
SCG10
Conservation-based PSFM refinement
NRSE2
B
Refinement of NRSEpsfm33
C
Search for all occurrences in multiple genomes
Filter matches in repeats and ORFs
cisAssociator analysis of occurrences in orthologous genes
cisMatcher analysis of the most conserved occurrences
GO Enrichment analysis
RNA expression analysis

Figure 2.1. Experimental approach — **A.** Matches from genome-wide matches to the initial NRSE PSFM (SCG10) were analyzed with cisMatcher and used to create a refined NRSE PSFM (NRSE2). **B.** A refinement starting with a PSFM of 33 known sites (Table S2) produces a result very similar to NRSE2. **C.** NRSE1 (consensus: NTYAGMRCCNNRGMSAGNNNN; Bruce, 2004) and NRSE2 were searched for, genome-wide, with either its consensus or with its position-specific frequency matrix (PSFM). Their respective gene cohorts were then analyzed for Gene Ontology (GO) enrichment and expression analysis (**I**). The NRSE2 PSFM was further processed and analyzed for GO enrichment and expression analysis of two subsets: (**II**) human genes with matches that co-occur in mouse and/or dog, and (**III**) human genes that are nearest to the "most conserved" matches, as identified by cisMatcher.

33 different known NRSEs, used them to develop the seed PSFM (nrsePWM33). In a third experiment, we ran the PSFM pipeline on several other individual NRSE instances.

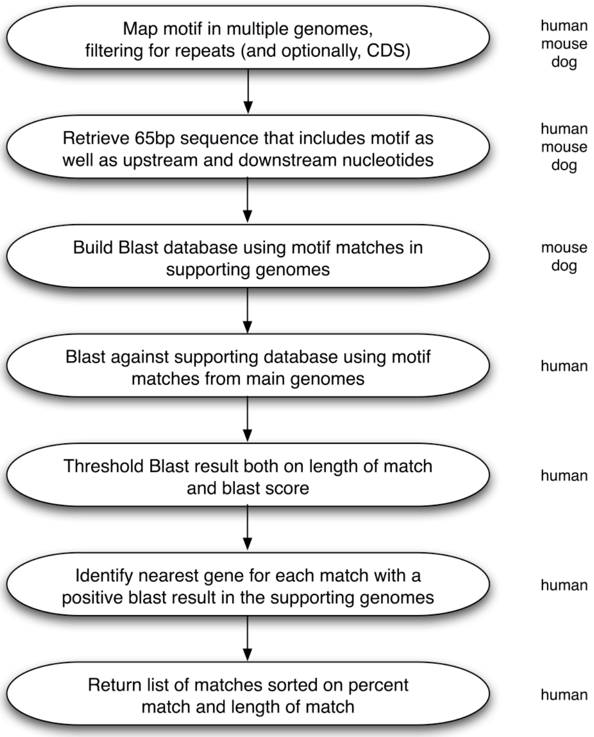

Map motif in multiple genomes, filtering for repeats (and optionally, CDS)
human mouse dog
Retrieve 65bp sequence that includes motif as well as upstream and downstream nucleotides
human mouse dog
Build Blast database using motif matches in supporting genomes
mouse dog
Blast against supporting database using motif matches from main genomes
human
Threshold Blast result both on length of match and blast score
human
Identify nearest gene for each match with a positive blast result in the supporting genomes
human
Return list of matches sorted on percent match and length of match
human

Figure 2.2. cisMatcher algorithm. After mapping motif instances across multiple genomes, each of these sequences are compared to one another in order to match conserved motifs across genomes regardless of their proximity to gene annotations. Example genome inputs and outputs are given on the right-hand side for a human versus mouse and dog comparison.

The resulting site models were remarkably similar to each other (Fig 2.3). We conclude that cisMatcher, operating over this set of genomes, derives a set of convergent PSFM models for NRSE sites. This means that our refinement process, which draws into the model many additional conserved instances, is robust to the identity of the specific initiating NRSE.

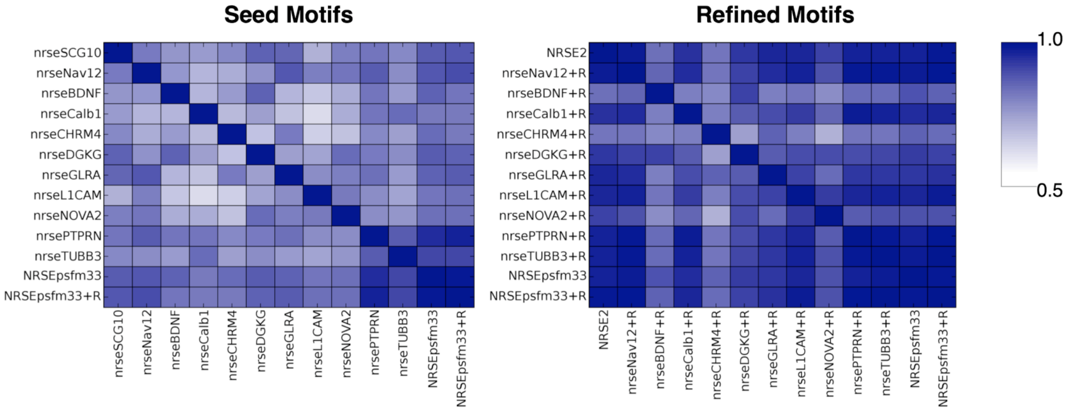


Figure 2.3 Different seed motifs converge following motif refinement. A. 10 initial seed motifs from known or predicted sites are compared using the motif similarity score (see methods) to our starting motif (SCG10), as well as a PSFM of 33 known instances (NRSEpsfm33) and the refined version (NRSEpsfm33+R). The correlation median is 0.80. **B**. Motif refinement of SCG10 (called NRSE2) and of the 10 initial motifs (denoted with a +R) are markedly more similar, with a motif correlation median of 0.91, and several inter-motif correlations rising above 0.95.

Estimating a membership threshold

How similar to the site model does a sequence need to be to function in vivo? We used multiple kinds of experimental data to iterate toward an informed and increasingly objective membership threshold. Setting a threshold is, at this stage, a useful and necessary simplification, but there is no biochemical or biological reason to expect a crisp boundary between sites that do and do not bind the factor. Fig 2.4A displays archival data for known NRSE sites, plus a few previously tested negative sites that resemble the NRSE, plotted as a function of PSFM match score. These data suggested starting with an estimated 84% match score threshold. We also asked if the PSFM match score correlated with the bioactivity of individual instances in a reporter transfection assay, drawing on data from Schoenherr (1996). Remarkably, there was a significant correlation of PSFM match score with repression strength (R^2 = 0.82, Fig 2.4B). The repression activity data are in general agreement with panel 2.4A, and support a threshold value in the low 80s. The relationship of PSFM match score with repression efficiency in the transfection assay may also indicate that both reflect binding affinity.

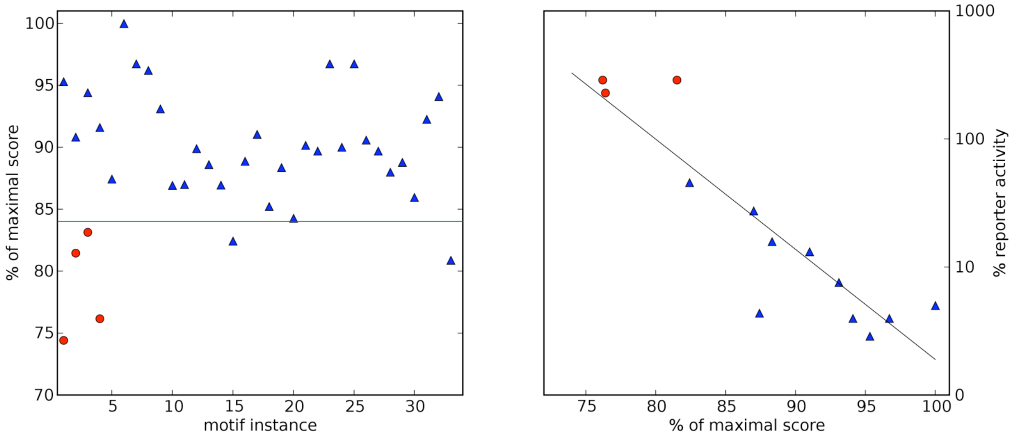


Fig. 2.4. Selection of a threshold for NRSE2 and correlation of score with repression activity. A. 33 known instances (filled triangles) and 4 false positives (filled ovals) listed in Table S1 were scored with the NRSE2 PSFM using a consensus score, as described in the text and methods. A threshold of 84% of the best possible score (match #5) was selected conservatively to exclude the known false positives. The PSFMs exclude about 6% of known instances at this relatively high threshold. **B**. The NRSE2 PSFM score of 10 known instances and 3 false positives were plotted against their relative repression in a transient transfection of a reporter from Schoenherr et al. (1996), where 100% and above reporter activity represents no repression. The regression shows a marked correlation between PSFM match score and repression (R^2= 0.82).

Assembling and testing target gene cohort models.

The three mammalian genomes were then searched for every match to NRSE2 above a predetermined threshold, and genes within a 10 kb radius were grouped into cis-regulatory cohorts of genes. This cohort of human NRSE2-associated genes was filtered for evolutionary conservation by requiring that matches also exist within 10 kb of an ortholog in mouse and/or dog genomes using Cistematic's cisAssociator algorithm. Because some known NRSF sites can apparently act in isolation without surrounding conserved elements, cisAssociator deliberately does not require alignment or additional conservation outside the site. Note that when a match is within 10 kb of more than one gene, cisAssociator includes all genes into the cohort. This choice is based on the report that single NRSE instances can apparently silence multiple nearby genes (Lunyak, 2002). However, it also means that, even if the definition of the NRSE2 PSFM is optimal, some genes included in the cohort model will be false positives. We also wanted to collect additional sites that might function from distances greater than 10 kb, but without greatly increasing false-positives. The cohort was therefore expanded by using the Cistematic cisMatcher algorithm to identify genes with conserved NRSE2 matches that are distal.

We then used the cohort model to revisit the threshold issue, evaluating it experimatally by sampling 113 candidate NRSE sites that spanned a range of high scoring and low scoring PSFM scores. Chromatin immunoprecipitation (ChIP) was performed and and assayed by quantitative PCR (QPCR) (Fig 2.5). These in vivo protein:DNA interaction data generally validate the PSFM model (see below) and the Probit model in Fig 2.5B suggests that a threshold around 84 is reasonable, but also indicates that there is no sharp PSFM boundary. This means that users of this and related models will select membership thresholds, or ranges for thresholds, to best serve different specific uses of the model for which pressure on sensitivity versus selectivity are different.

How do previously identified NRSE cohorts, based on conventional consensus sites, compare with the new PSFM? We compared the NRSE2 PSFM matches with instances found using the original (NRSE0) consensus of Schoenherr et al. (1996) and the recent

(NRSE1) consensus used in the genome survey of Bruce et al. (2004). Cistematic recovered the respective gene cohorts corresponding to NRSE0 and NRSE1 instances.

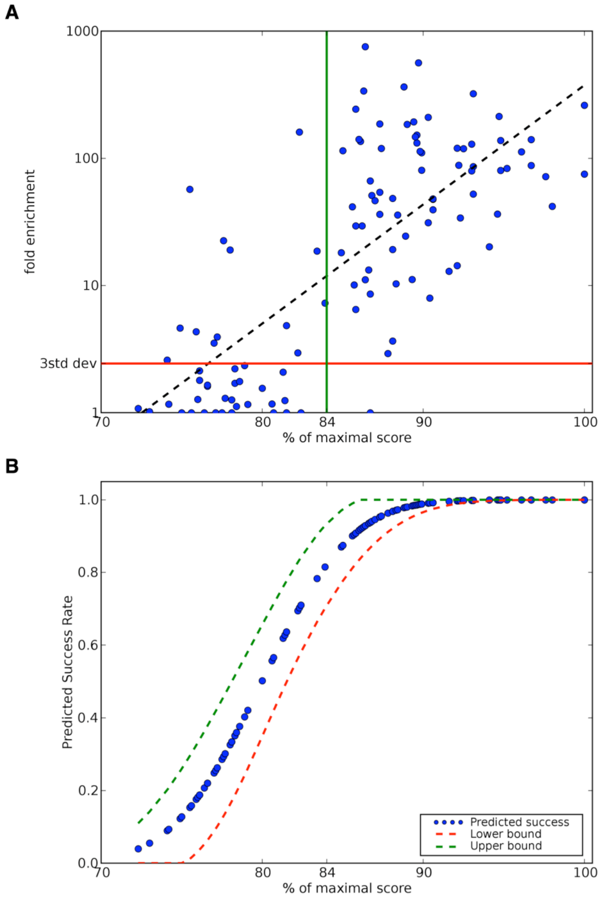


Figure 2.5. Quantitative Analysis of Chromatin Immunoprecipitation of NRSF. A. 113 potential NRSE2 matches, 42 of whom fell below our threshold of 84% (green vertical line), were assayed using chromatin IP followed by quantitative PCR. Fold enrichments were calculated by dividing the absolute number of genomic equivalents of each NRSE by the mean of the recovered amounts of 5 random non-genic, non-conserved regions. Fold enrichments that were above 3 standard deviations from the mean of the 5 random non-genic amounts (red line, 2.44x enrichment), were considered to be occupied sites. An exponential regression (black line in this semilog plot), which would correspond to the regression in Fig 2b, accounts for about half of the data's variation (R^2 = 0.56). 13 of the 83 occupied sites (16%) fell below our 84% threshold. **B**. Cumulative normal distribution function of probit coefficient versus score with 95% confidence levels shown by dashes. The estimated chance of a success match goes up by nearly half between 80 and 84%.

Fig 2.6 shows that NRSE1 contains a significant fraction of matches that score poorly with the PSFM model (< 80%), with many low-scoring NRSE1 matches occurring in complex repeats. Matches within repeats were excluded from subsequent analyses for both NRSE1 and NRSE2, although we note that individual instances embedded within repeats might be functional.

We then asked how cisMatcher positive sites are distributed relative to gene anatomy. Many known instances of the NRSE that have been studied in detail are either intragenic or are located near the promoter, but it is not known how great a role ascertainment bias based on proximity has played in selecting them for study. We mapped the genome-wide cisMatcher set, which is not biased by the method of selection for its position relative to adjacent genes. There is an obvious enrichment of NRSF motifs within 5 kb of gene model start sites (40%), although a full quarter of the conserved matches are more than 10 kb from either the 5' or 3' boundary of the nearest gene model, and 3'UTRs have substantial numbers.

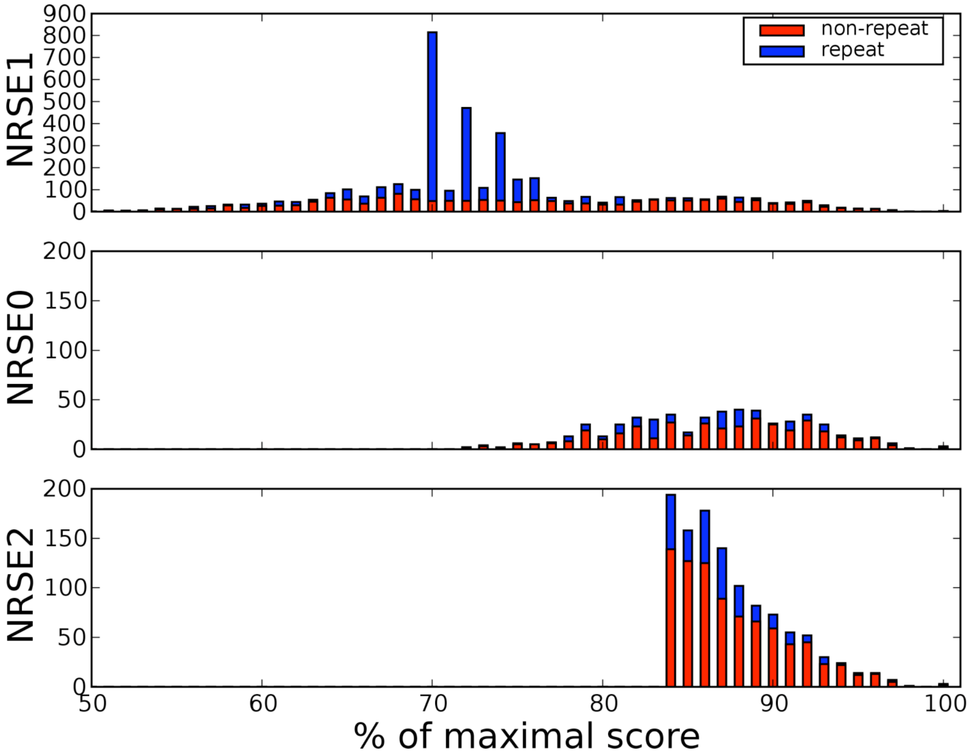


Figure 2.6. NRSE matches in repeats. Genome-wide human matches for the NRSE1 (Bruce) consensus (top), the NRSE0 (Schoenherr) consensus, and the NRSE2 PSFM (threshold of 84%) are scored using the NRSE2 PSFM. Whereas all three NRSE motifs show matches within repeats (black), the NRSE1 motif disproportionally matches within repeats at scores between 70 and 76%.

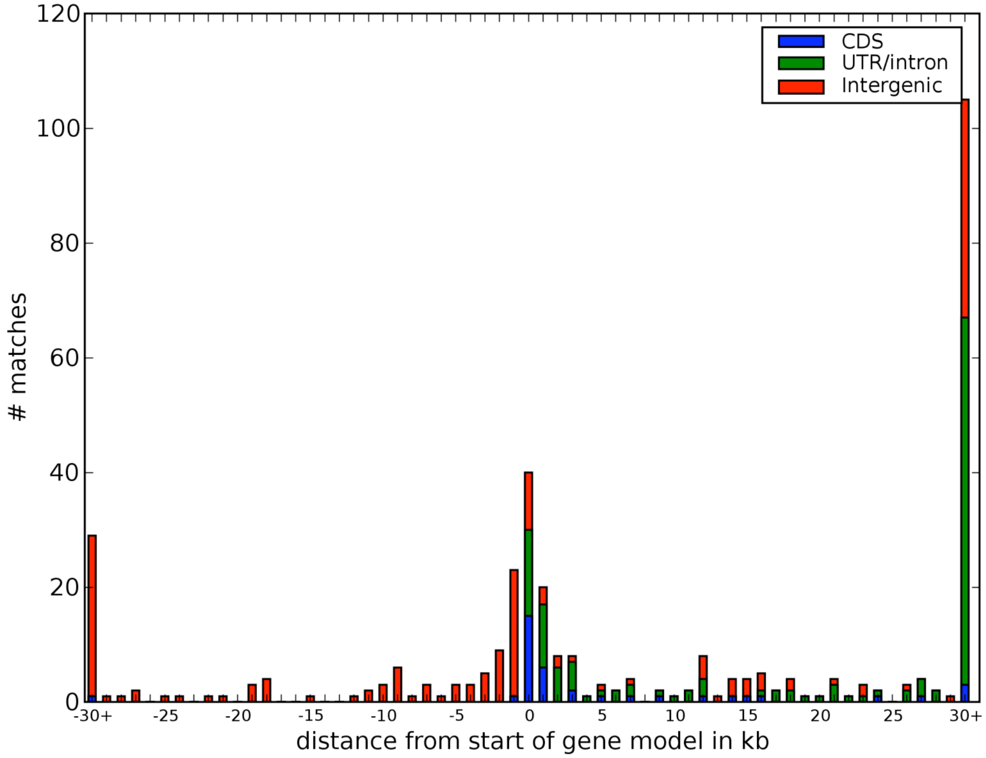


Figure 2.7. Spatial distribution of cisMatcher-identified NRSEs. cisMatcher matches were binned based on their distance from the start of the gene model (either transcription or translation, depending on the model), which show that while there is a clear enrichment of NRSE around the start of the model, more than half of the matches are further than 5 kb away.

Chromatin Immunoprecipitation (ChIP) analysis of predicted NRSEs

We tested the NRSE2 cohort experimentally at 113 sites (Fig 2.5), 42 of which fell below our 84% threshold, by using chromatin IP coupled-with Q-PCR in Jurkat cells (see Methods). Of 71 candidate sites ranking above the 84% threshold, 70 were ChIP positive. In contrast, at slightly lower PSFM match scores, 29 of 42 sites were negative for NRSF ChIP. Thus, predicted sites could be quite effectively partitioned by PSFM score into those that will certainly be ChIP positive and those that are likely to be negative (p-value = $8.6*10^{-16}$, Fisher's exact test). The associated Probit analysis allows one to select other thresholds and to consider the confidence limits at any selected threshold.

The 84% value is a conservative membership threshold designed to minimize false-positives as much as possible, at the cost of accepting some false negative predictions (13/83, or 16%). Cistematic provides the option of sliding the threshold to provide cohort models that correspond to differing stringencies for false positive or false negative members.

Gene Ontology (GO) analysis of the NRSE2 cis-regulatory cohort.

We next asked if functions of the NRSF-regulated cohort could be inferred based on enrichment of Gene Ontology (GO) terms. Statistically significant enrichment, subject to Bonferroni correction for multiple hypothesis testing, was observed for each cohort model but not for a large set of randomly scrambled version of the PSFM (Methods). The NRSE2 PSFM identified a larger cohort (660 human genes within 10 kb of an NRSE2) than the original NRSE0 consensus (362 human genes) or the seed SCG10-based PSFM (192 human genes), with significant enrichments in functional GO categories such as “synaptic transmission”, “neurogenesis”, and “transporter activity”. These functions nicely recapitulate much of the NRSF literature. In contrast, several GO categories significantly enriched in the larger NRSE1 cohort (1270 genes), such as “synaptogenesis” or “calcium-dependent cell-cell adhesion”, are conspicuously absent from the other NRSE cohorts. On detailed inspection, the latter results are mainly due to NRSE1 matches within the paralogous protocadherin β cluster. This calls attention to a specific interpretation issue in GO enrichment analysis, which is the power of very similar paralogs in gene families to drive an entire term to significance. Similar paralogy issues do not appear to dominate most significant other terms for any of the NRSE models.

Cistematic's orthology matching function was next used to develop a conserved cohort. NRSE2 instances in human, mouse, and dog were collected and subjected to both cisMatcher and cisAssociator conservation criteria. 505 human genes met at least one of these criteria. GO analysis of the resulting conserved cohort (Fig 2.8) shows further enrichment of several GO terms such as "transporter activity", "synapse", and "synaptic vesicle" when compared to the larger NRSE2 cohort, but these effects were not substantial.

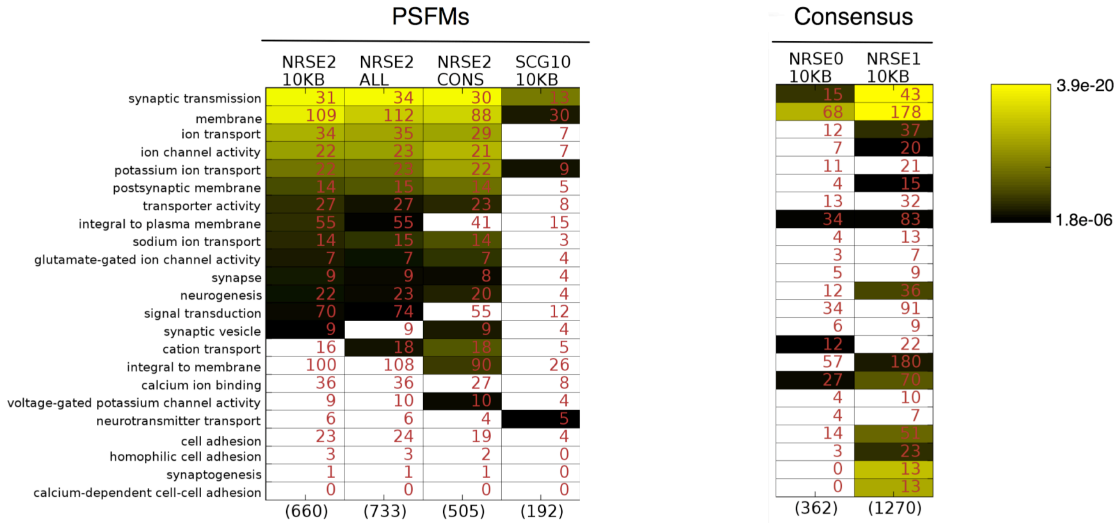


Figure 2.8. Gene ontology enrichment comparison of different NRSE cis-regulatory cohorts. Cohorts of human genes within 10 kb of a candidate NRSE0 (Schoenherr, 1996), NRSE1 (Bruce, 2004), SCG10 (the original seed motif), NRSE2, All NRSE2 matches, and conserved NRSE2 matches were filtered of repeat matches and were analyzed for GO term over-representation. Significantly enriched GO terms in at least one of the cohorts (out of 4576 possible GO terms) are shown. Numbers in cells represent the genes with the term in the cohort while numbers in parentheses represent the cohort size. Cells shown in color pass the threshold of significance, as determined by a Bonferroni correction. GO terms are sorted in decreasing order by p-values of the leftmost column. Note that GO enrichments are in term of decrease in p-values, which are directly correlated to the size of the cohorts, the number of genes in the shared association cohort with a particular GO term may go down or stay the same, while its significance increases. The NRSE1 motif behaves differently from the other definitions, as seen in the enrichment of synaptogenesis, which is the result of weak matches within the paralogous protocadherin β family.

To test the robustness of the GO analysis, the columns of the NRSE2 PSFM were scrambled repeatedly and the entire analysis pipeline was repeated 100 times (data not shown). Only two scrambled motifs recovered any significantly enriched GO term, and they found just one each. No scrambled motif recovered significant GO terms when either of our conservation criteria was applied. These results argue that enrichment of specific GO terms for NRSE2 is statistically sound.

Comparative Expression Analysis of NRSE2

We asked if the NRSE2 cohort is enriched in genes with a specific RNA expression pattern. One prediction from prior studies of NRSF is that genes expressed predominantly in neurons will be enriched among true biological targets of NRSF (Ballas, 2005; Chen, 1998). The GNF gene atlas (http://symatlas.gnf.org, Su, 2004) of mRNA expression across 79 human tissues was used to investigate the expression profile of our gene cohorts. CompClust (Hart, 2005) was used to cluster the NRSE2 cohort with k-means, k-medians, and DiagEM for k=5, 10, and 15. While all three algorithms returned similar pan-neuronal clusters, k-medians with a Pearson correlation metric and k=5 performed best qualitatively and was used for all subsequent analyses. The NRSE1 cohort was also clustered for comparison and produced similar clusters (Fig 2.9). The NRSE2 clustering is shown in Fig 2.10A. In every clustering, one or more clusters had a distinctly brain-specific expression pattern, whose medoid weights are shown in Figure 2.10B. The percentage of each cohort falling within these brain-specific clusters ranged from 21% for NRSE1 to 40% for NRSE2. These reactions are significantly higher than the percentage of genes in GNF that have a Pearson correlation coefficient > 0.4 with our pan-neuronal medoid vector (1,482 out of 16,054 genes with current NCBI Gene IDs, or about 9%), which gives a p-value of

$8.0 * 10^{-71}$ ($\chi^2 = 316.58$, χ^2 test for equality of distributions) for the neuronal enrichment of the NRSE2 cohort. Nevertheless, the majority of neuronal genes are not associated with a recognizable NRSE. As would be predicted if many NRSE2 genes are regulated by NRSF in a neuronal context, there is a large (> 4-fold) enrichment for brain expression to 40% of all NRSE2-associated genes (Fig 2.10C, 2.9).

Fig 2.10D gives match score distributions for the subset of genes that display a predominantly brain-specific RNA expression pattern. Genes within the brain-specific expression clusters share a similar scoring distribution pattern to the entire population of matches for both NRSE0 and NRSE2, whereas NRSE1 pan-neuronal matches show a bimodal distribution with a local minimum at 77%, which is below our predicted cut-off for repression activity (Fig 2.4B). Based on the PSFM score and its relation to functional assays, these NRSE1 instances are unlikely to be biologically active on their own.

Confusion matrices (Hart, 2005) were used as a generalized Venn diagram to compare the overlap of the genes and expression pattern of the different cohorts. Fig 2.9 shows the confusion matrix for NRSE1 versus NRSE2; while both motifs agree on about 323 genes, both cohorts have large sets of non-overlapping genes (also known as outersects or relative complements; see Methods). The outersect of NRSE1 is comprised of 615 additional genes not present in the NRSE2 cohort, whereas the corresponding NRSE2 outersect includes 172 genes. Neuronal genes comprise 34% of the NRSE2 outersect, but only 14% of the NRSE1 outersect (p-value = $5*10^{-9}$, $\chi^2 = 34.07$, χ^2 test for equality of distributions)

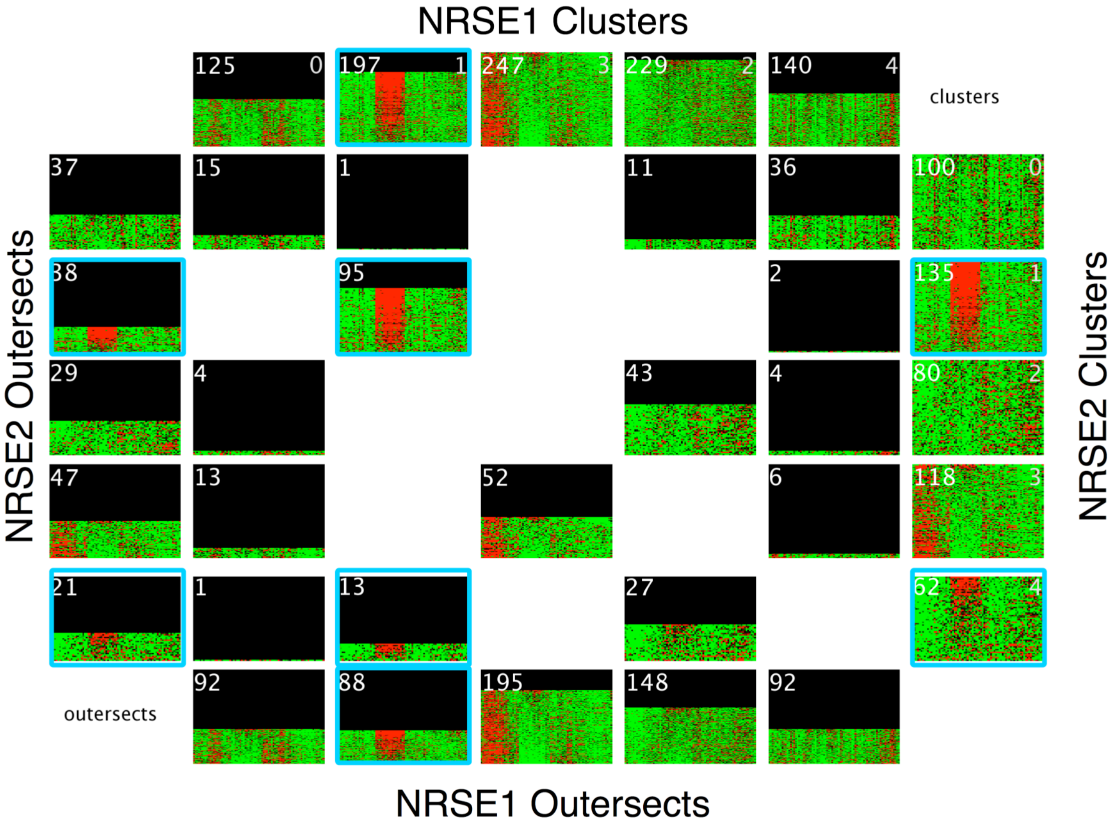


Figure 2.9. Confusion matrix comparison of tissue expression pattern of NRSE1 and NRSE2 matching genes. Human genes with an NRSE1 (top row) or NRSE2 (right-most column) with an expression pattern in the GNF survey of 79 human tissues, were clustered using the k-medians algorithm as described in the methods, with the cluster number in the upper-right-hand corner, and the cluster size in the upper left hand. Genes that were unique to each dataset are shown in the bottom row / rightmost column, whereas genes that are in common between the two datasets are shown at the intersections of their respective clusters. In both datasets, clusters with a blue border represent those genes with a high expression pattern in neuronal tissues and low expression pattern elsewhere. These highly brain-enriched expressed genes make up a greater percentage of the NRSE2 cohort (40%) and of its outersect (34 %) than of the NRSE1 cohort (21 %) or of its outersect (14%), the latter containing low-scoring matches by our PSFM.

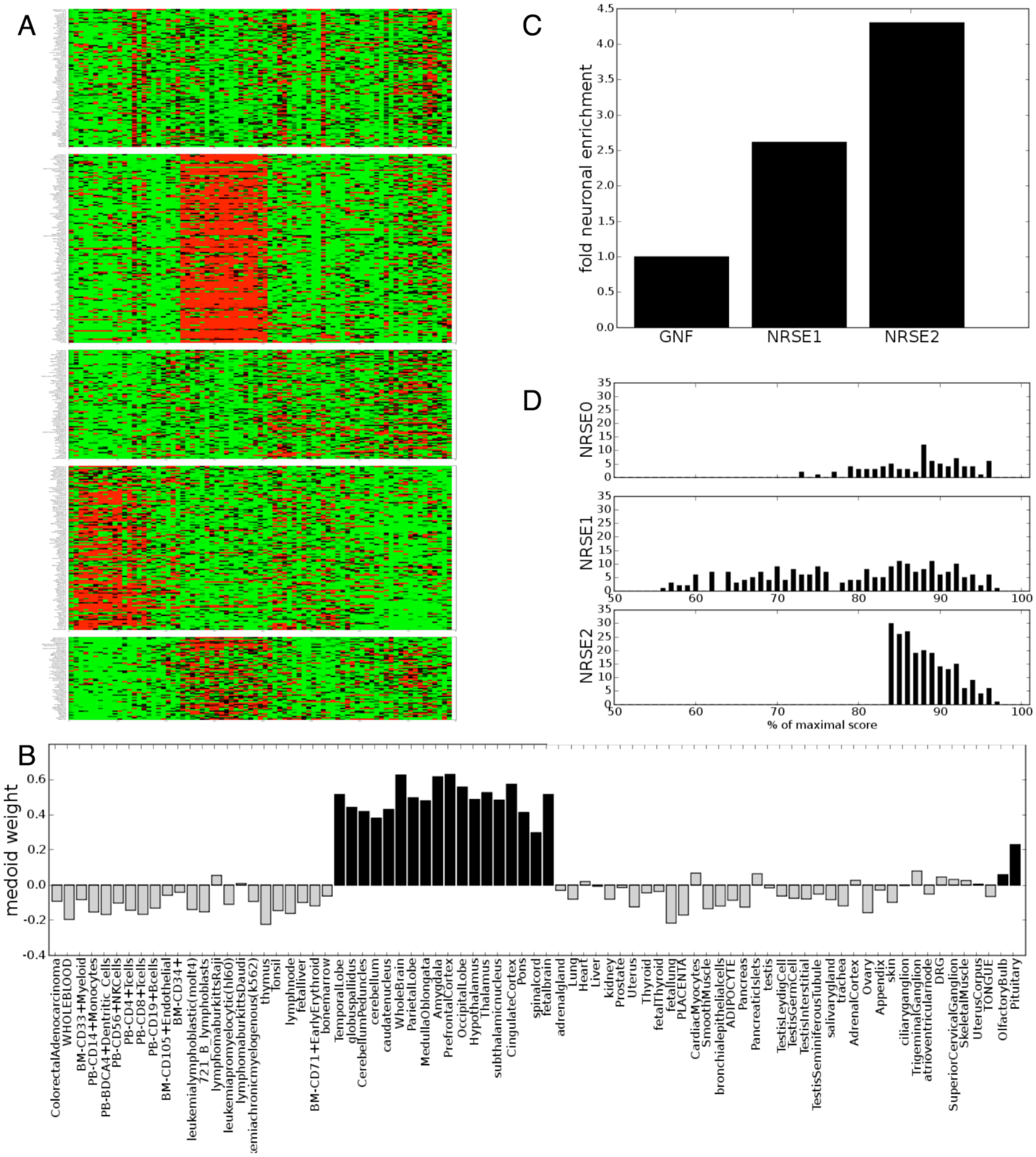


Figure 2.10. Tissue expression pattern of NRSE associated-genes shows brain-specific expression enrichment. **A**. Human genes with an NRSE2 with an expression pattern in the GNF survey of 79 human tissues, were clustered using the k-medians algorithm as described in the methods. The second and fifth clusters, which encompass 40% of the NRSE2-associated genes shows a clear, brain-specific expression pattern. **B**. Weights of the k-medoid for cluster 2, with brain tissues highlighted in black. Note that cardiac myocytes and pancreatic islet cells also have positive weights. **C**. NRSE2 shows a 3.5 fold enrichment of "brain specific" genes (as defined by the medoid in B) compared to the GNF datasets and show greater enrichment than NRSE1. **D**. NRSE0 (top), NRSE1, and NRSE2 matches associated with genes than have a greater than 0.4 correlation with the medoid vector in B. NRSE1 shows a double-humped distribution of matches, with matches weaker than 77% accounting for half of its matches; these low scoring matches are likely false-positives.

suggesting that consensus-based approaches like NRSE1 likely miss neuronal, NRSE-associated genes (Figure 2.11) as also suggested by Zhang et al (2006).

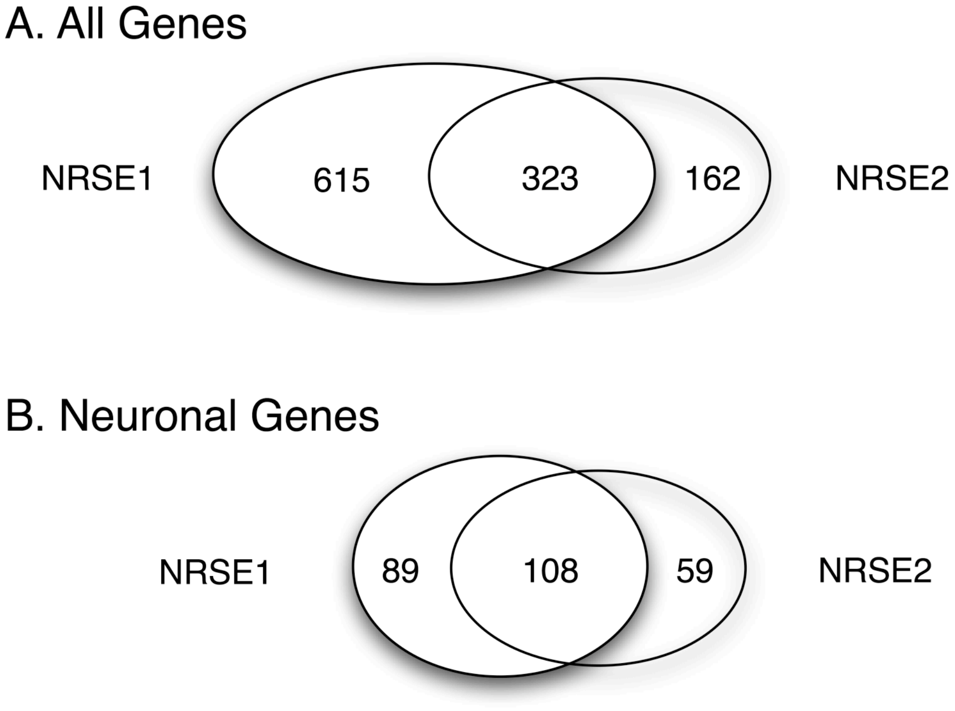


Figure 2.11. Venn diagram of NRSE1 and NRSE2 matching genes. **A**. Human genes with an NRSE1 or NRSE2 with an expression pattern in the GNF survey of 79 human tissues. **B**. Highly brain-enriched expressed genes make up a greater percentage of the NRSE2 cohort (40%) and of its outersect (34 %) than of the NRSE1 cohort (21 %) or of its outersect (14%), the latter containing low-scoring matches by the NRSE2 PSFM.

NRSEs are only found in vertebrate genomes

NRSE2 matches were sought in genomes representing four invertebrate phyla (arthropod, nematode, echinoderm, and urochordate), together with seven additional vertebrate species. The remarkable result is that there are essentially no matches in invertebrate genomes, while all vertebrate genomes have the same order of magnitude of matches, regardless of genome size, with the pufferfish genome being especially informative (Fig 2.12).

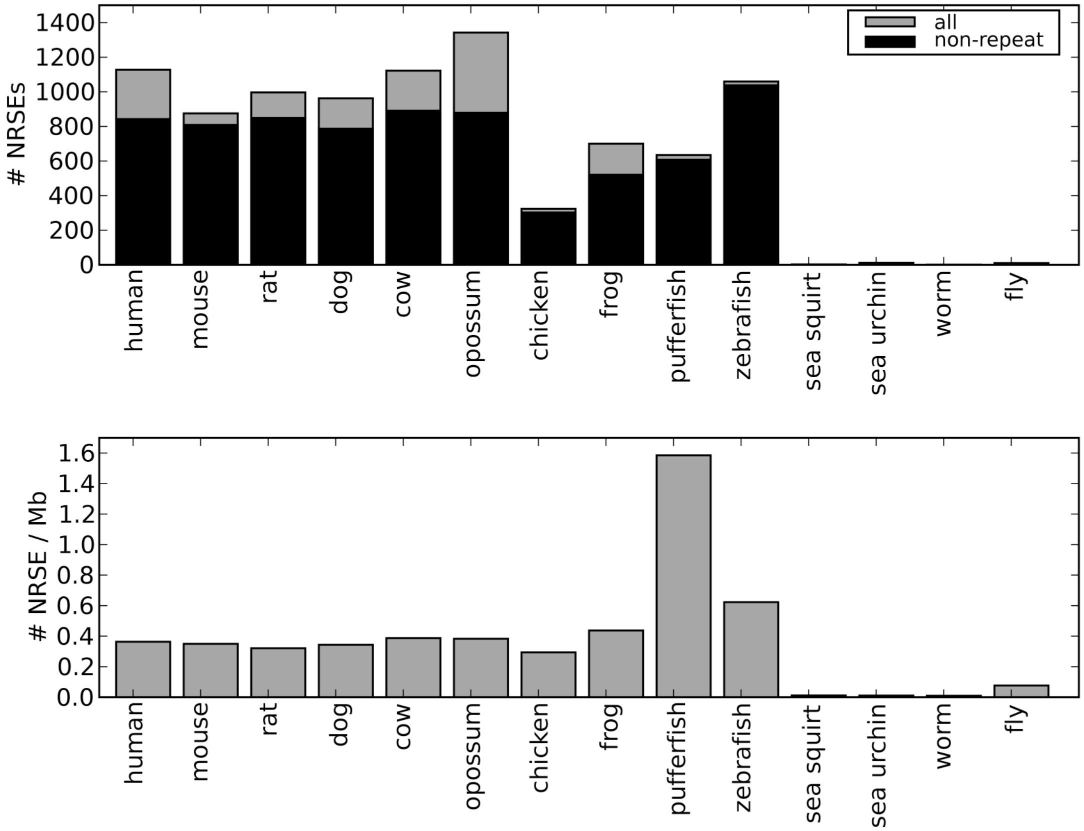


Figure 2.12. NRSE distribution in vertebrate and invertebrate genomes. A. The number of NRSE2 matches in mammalian genomes is relatively constant and include a significant number of matches within repeats when compared to other vertebrates, compared to the virtual absence of NRSE2 matches in invertebrates. **B**. The higher density of all NRSE matches / Mb of genomic sequences in pufferfish and zebrafish when compared to chicken suggest that fish and mammalian NRSE matches may have been expanding independently.

Tetraodon has a highly compressed genome that retains functional sequences such as ORFs at 3–5-fold elevated density. A similar enrichment is seen for NRSE2 occurrences, which suggests that many of them are functional. The notable paucity of NRSE2 sites from the sea urchin, Drosophila, and Ciona (a urochordate) genomes argues that this repression network is absent up into protochordata, and it calls into question a previous tentative assignment of NRSF orthology to CoREST-interacting zinc fingers in *C. elegans*

(Lakowski, 2003). We also found that there is only one NRSE2 instance in the entire *C. elegans* genome, and it is not conserved in related worm genomes (*C. briggsae* and *C. remanei*).

NRSE2 PSFM matches in the Tetraodon genome were related to matches in the human genome using cisAssociator to identify genes that remain associated with a high scoring NRSE in both fish and mammals. There were only 33 matches that pass our criteria for best reciprocal match of the corresponding gene models. Occurrences of NRSE1 and NRSE2 in human repeats were analyzed using the UCSC repeatMasker annotations (http://genome.ucsc.edu, Kent, 2002; Karolchik, 2005) to address whether NRSE instances were found preferentially within the same repeat families. While most NRSE2 matches (285 instances that meet or exceed the 84% match score threshold of Figure 2.4) reside mainly in the old vertebrate LINE2 family (226 matches, 79%), the overwhelming majority of NRSE1 consensus matches are in the ERV1 SINE family (1,858 of 2,339 matches, 79%), which score between 70 and 74%. This dichotomy is particularly striking because there are no NRSE2 matches in the ERV1 family. With two or three strategic chance mutations, many of these repeats could achieve a low functional match score upon which selection could operate to favor further optimization.

NRSE2 PSFM matches associated with microRNAs

We proceeded to identify microRNAs in the human genome located within a 25 kb radius of a non-repeatmasked NRSEs. The search radius was increased from the cisAssociator 10 kb used for the NRSE2 cohort to respond to the observation that some microRNAs are embedded in, and expressed as part of, primary transcripts from protein coding genes

(Ying, 2004). The sites were mapped against the UCSC entries of the microRNA registry (Griffiths-Jones, 2004; Weber, 2005). Twenty-one microRNAs were identified (out of 326 in the annotations) that represent sixteen distinct families. All but one of these microRNAs had been previously characterized in the context of mammalian neuronal differentiation (Sempere, 2004). MiR-375 was shown separately to be pancreatic β-cell line specific (Poy, 2004). It has been shown to target at least one gene (myotrophin) in the murine pancreatic cell line MIN6 in coordination with miR-124a (Krek, 2005). Six NRSE-associated miR families also assayed in Sempere belong to 14 families (out of 100 surveyed) categorized in Sempere et al (2004) as "brain specific" or "brain enriched". This pattern of coherent tissue specificity in expression is significant by the criterion of p-value of 0.02 (Fisher's exact test). Seven of these microRNAs are located in introns of genes in the NRSE2 cohort, i.e., miR-153 in PTPRN, miR-139 in PDE2A; miR-9-1 in CROC4; miR-7-3 in C19orf30); and miR-24-1, miR-27b, as well as miR-23b in C9orf3. In the case of miR-153, miR-139, and miR-9-1, the RNA expression pattern of the "host" gene falls in the brain-specific cluster (Fig 2.10A). We assayed NRSEs from 11 of these by ChIP, and 10 scored positive for NRSF/REST occupancy. Our results for miR-124a and miR-9 agree with those reported in by Conoco et al. (2006).

By inspecting lists of predicted target RNAs for NRSE-associated MicroRNAs (Lewis, 2005) we found that CoREST (GenBank D31888) is a candidate target for three of our sixteen microRNA families (miR-29b, miR-124a, miR-153), and that NRSF itself (GenBank U22680) is a prospective target of miR-153, which has recently been shown to be brain-specific in the zebrafish embryo (Kloosterman, 2006). These postulated interactions create a potential feedforward loop that might have the effect of more quickly

Name	**NRSE2 PSFM (%)**	**Distance (bp)**	**Human Brain**	**Mouse Brain**	**P19 + RA**	**NT2 + RA**	**ChIP Fold Enrichment**
miR-153-1	97	14,208	Low	Low	Low		87.9
miR-135b	93	10,826	Low	Low	Medium	Low	79.6
miR-124a-2 (*)	92	934	Medium	Medium	Low	Low	
miR-9-1 (*)	91	5,681	High	Medium	High	Low	7.97
miR-29a (clust 1)	91	11,106	Medium	Medium		Low	48.03
miR-29b-1(clust 1)	91	11,818	Medium	Medium		Low	48.03
miR-212 (clust 2)	88	111				Low	
miR-132 (clust 2)	88	252	Medium	High			
miR-133a-2	88	23,034	Low	Low		Low	10.32
miR-124a-3 (*)	87	487	Medium	Medium	Low	Low	1.00
miR-375	87	9,768	-	-	-	-	8.56
miR-7-3	86	1,097	Medium	Medium	Low	Low	
miR-139	86	2,255	Medium	Medium			29.37
miR-9-3 (*)	86	3,050	High	Medium	High	Low	11.12
miR-124a-1 (*)	86	21,763	Medium	Medium		Low	10.09
miR-124a-3 (*)	86	2,394	Medium	Medium	Low	Low	
mirR-24 (clust 3)	85	1,743					
miR-27b (clust 3)	85	2,319	Medium	Low	Low	Low	
miR-23b (clust 3)	85	2,556	High	Low	Medium	Medium	
miR-203	85	15,684	Low	Low			

Table 2.1. microRNAs with associated NRSE2 matches in the human genome have a neuronal expression pattern. MicroRNAs with an NRSE2 match with PSFM score greater than 84% within 25 kb are shown along with their expression pattern from Sempere et al. (2004) in human and mouse brain as well as in mouse P19 and human NT2 cell lines undergoing retinoic-acid induced neuronal differentiation and where several miRs (bold) were categorized as "brain specific" or "brain enriched". Multiple microRNAs that are near the same NRSE are labeled with the same "clust" ID. Entries with asterisks mark members of the same microRNA family that only have one entry in Sempere et al. (2004), and are hence shown with the same expression pattern. miR-375 was found separately to be expressed specifically in pancreatic β cells by Poy et al. (2005). Chromatin Immunoprecipitation fold enrichments for those microRNA-associated NRSE2 matches that were part of our 113 sites tested that are higher than 2.44 are considered positives.

or definitively down-regulating NRSF mRNA, as NRSF activity begins to fall (Figure 2.14). This is given additional impetus by the observations that miR-153 is the microRNA with the best-scoring NRSE site (Table 2.1), and that its NRSE is embedded in PTPRN, a gene expressed strongly and widely in the nervous system.

Discussion

Our effort to model the conserved NRSF binding site and its target gene cohort differs substantially in design, tools, and outcome from prior attempts (Lunyak et al., 2002, Bruce et al., 2004). We show that a successful PSFM site model can be derived from a single starting conserved NRSE by using iterations of motif refinement that incorporate additional site instances based on their conservation in multiple mammalian genomes. Prior designs started from collections of multiple genes and produced conventional consensus sites. The NRSF PSFM model, unlike standard consensus motif, captures more information about site structure and affords a way to rank score matches, according to how well they match the model site. We then tested the model experimentally across a range of PSFM match scores, including below-threshold borderline values, by ChIP/QPCR experiments. This allowed us to assess the predictive qualities of the model relative to PSFM score. These results encourage us to think that other relatively large and well-specified motifs could be usefully modeled in the same manner. However it is important to recognize that shorter or less well-specified motifs — those with lower information content — will be difficult or even impossible to treat in this manner without additional algorithms to help discriminate functional occurrences from chance occurrences.

The PSFM site model captures more information about site preferences at each position than does a basic consensus. We showed that the PSFM score correlated well with repression activity in transient transfection assays, arguing that it is a good first-order predictor of function. Our ChIP independently showed that a high PSFM match score is predictive of in vivo NRSF occupancy at a given locus. In most prior attempts to develop genome-wide target site models, including NRSF/REST studies, thresholds for membership were set arbitrarily. Based on NRSF/REST results, we think that integration of functional data in this manner is a natural way to bound computational models, establish confidence limits, and then further refine them. However, the apparent intensity of the ChIP interaction differed greatly from one positive locus to another, and we do not yet know what modulates levels of ChIP signal. Obvious biological possibilities include chromatin structure, the presence or absence of various collaborating factors, and contributions from weaker NRSE sites near strong ones.

Cistematic permitted us to efficiently generate and compare families of related models by varying parameters for conservation, position of sites relative gene anatomy, PSFM match stringency, and initiating seed sites. The ability to do this in an automated manner is useful for finding out if a model is vulnerable to changes in input parameters. In one pertinent example, we ran the pipeline beginning with different individual starting site instances, as well as a starting site pool, and found the results are robust to these variations in the initial seed site.

The NRSE2 matches were analyzed for statistically significant functional covariates, from GO and from RNA expression data, using Cistematic modules designed for these purposes.

The software architecture (Fig 2.2, 2.13) and Open Source license are meant to encourage users to add other analytical modules at will. A key conclusion from these experiments is that the principle function of NRSF could have been inferred solely from analysis of the final NRSE2 cohort model. The enrichment relationships for neuronally expressed RNAs and neuronal GO functions within the NRSE2 cohort model were statistically far above background, despite incompleteness of GO annotations and imperfections in large-scale expression databases. RNA analysis of the NRSE cohort model benefited from strong sampling of brain tissues in the GNF data, and application of this approach to other motifs will be effective as global RNA datasets and GO annotations become more extensive. Had we not already known that NRSF acts as a repressor, this also could have been inferred *de novo* from the NRSE2 cohort, together with expression data for NRSF/REST itself. In mouse and human, the RNA profile for NRSF/REST is in frank opposition to the expression of its direct target repertoire. These inferences show that PSFMs based on evolutionary conservation, and the target gene cohort models derived from them, can successfully predict organismic and molecular functions. The model generates hypotheses at the level of the entire network and also at the level of individual genes (Fig 2.14 and below).

We think the approach taken here will be applicable to many transcriptional regulators in vertebrates that meet several criteria. In practical terms, the cardinal requirement is a long

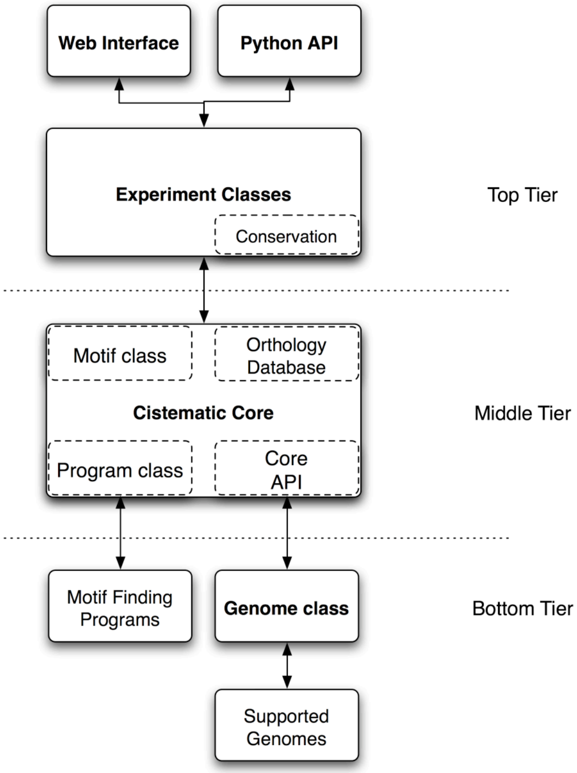


Figure 2.13. Cistematic architecture — Cistematic's three-tiered architecture consists of the top-level Experiment classes, which provide a framework for accessing the data and results produced by the middle-tier classes that form the Cistematic core. The Cistematic core itself relies on the Genome class and external programs to retrieve sequence data, annotations, and candidate motifs.

and specific binding motif. The length of the NRSE2 PSFM was critical for evading the most dire consequences of Wasserman and Sandelin's "futility theorem", namely that the vast majority of binding site instances predicted based on motif knowledge will have no functional significance (Wasserman, 2004). Large families of factors whose members are likely to be eligible for Cistematic PSFM models include multifinger zinc finger class regulators that have been expanding rapidly in mammals (Shannon, 2003). The second criterion is evolutionary conservation. If a site/factor pair is very new, it will not be possible to leverage conservation, although the addition of increasing numbers of genomes will provide more branch length and resolution within clades such as the mammals (Boffelli, 2004). Finally, whether the data are obtained before the initial PSFM model building or after, quantitative functional analysis of a sample of true positive and true negative sites makes a powerful contribution that can be used to bound model membership and, in the best cases, to predict which instances are likely to be most active in vivo.

The NRSF/REST network is a chordate invention.

All currently available data argue that the neuronal NRSF repression network is a chordate invention. Extending the analysis of NRSE2 matches to an additional eleven available genomes (Fig 2.13) revealed that while NRSE2 is not only absent in Drosophila as previously noted (Bruce 2004, Dahlman, 2005; Yeo, 2005), but also is essentially absent from all invertebrate genomes. In sharp contrast, all vertebrate genomes we surveyed have between 302–1047 non-repeat matches, with an average of 750. Within mammals the average number is modestly higher (842). Furthermore, preliminary surveys of amphioxus (a cephalochordate) and lamprey (a basal vertebrate) whole genome shotgun traces found that NRSE2 matches are present in both at high densities, while the motif is entirely absent from the urochordate, *Ciona intestinalis*. This, along with the absence of any gene models that are convincingly similar to NRSF in Ciona or invertebrate genomes, suggests that NRSF emerged after the time of the last common ancestor shared by vertebrate and urochordates. Paralleling this, NRSF/REST itself is present and highly conserved in all vertebrate genomes but absent from *Ciona* and multiple invertebrate genomes. We did not detect NRSF in searches of sea urchin or *C. elegans*, and others have reported it absent from Drosophila, even though its principal co-repressors are present there (Dallman, 2004; Yeo, 2005). We did not detect NRSF in amphioxus trace coverage either, which could be a simple technical issue, but also raises the possibility that the target motif might have emerged ahead of the factor itself.

These data, combined with the existence of high-scoring sites within old LINE2 elements in the human genome, suggest that NRSEs may have first been distributed across vertebrate genomes via repeats at roughly the same time the NRSF DNA binding factor first appeared. In such a scenario, NRSEs that land near or in genes and also confer some advantage when repressed by NRSF, are starting points to expand an NRSF network. The much larger reservoir of weak, probably inert, NRSE1 (Bruce consensus) sites present in other repeat families might provide new NRSF/target gene pairs, given one or two key mutations.

A subset of neuronal genes belong to the NRSE cohort

RNA expression and GO term analyses showed that, under the NRSE2 model, NRSF does not directly act on a majority of genes with broad brain expression or with distinctly neuronal GO classifications. There are roughly 1,400 genes preferentially and broadly expressed in adult brain, but only 11% of these have a high confidence NRSE2 motif. Some of the non-NRSE brain genes are probably glial, while another subset might be explained by weaker NRSEs, functioning individually or multiply. NeuroD1/Beta2, for example, is an attractive candidate target based on its expression pattern and function in neurogenesis and pancreatic islet cell genesis (Lee 1995, Huang 2000). It has one NRSE ~ 4.5 kb upstream that scores above our threshold in mouse and dog, but slips below threshold in human. However, closer inspection shows that NeuroD1, like the related factors, NeuroG1 and NeuroG2, has additional low scoring NRSE matches embedded in its open reading frame. Learning the rules governing use of weaker sites awaits a fully comprehensive experimental mapping of NRSF/REST in vivo interactions, but many neuronal genes probably depend on other factors for their neuronal expression. A corollary is that substantial numbers of additional pan-brain genes present in relaxed-stringency models, including the NRSE1 cohort, are likely neuronal due to other regulatory factors, rather than by the action of a functional NRSE.

The converse is also true. Significant (~ 4-fold) enrichment of the NRSE2 cohort for a brain expression profile leaves 60% unaccounted for. Some reasons for this include incomplete gene annotations, genes restricted to specific kinds of neurons, mRNAs present at levels below microarray threshold, and inclusion of some extra NRSE2 neighborhood genes into the model by the cisAssociator algorithm. For example, several of the NRSE2 associated transcription factors are well known for important functions in specific neuronal populations (Neurogenin-3, POU4F1, POU4F3, LHX3, and LHX5), but none are in the pan-brain cluster, nor is their expression utterly specific to brain. It is also unclear how many genes in this model cohort might be targets of NRSF regulation relevant to its cardiac, pancreatic, or other functions.

NRSF/REST interactions at neuronal transcription factor, microRNA and RNA splicing factor loci

The NRSE2 model target gene cohort included other transcription factors, microRNAs and splicing regulatory factors, all of which could extend the regulatory effects of NRSF/REST. Multiple NRSE instances are associated with transcription factors. In addition to an expected complement of channels and synaptic proteins, highly conserved NRSE instances shared between human and fish are associated with transcription factors of interest. LHX5 and LHX3 are LIM homeobox factors important for specification and function of distinct neuronal populations. LHX5 also controls regulation of neuronal precursor exit from the cell cycle in the hippocampus (Zhao, 1999). Among NRSE2 instances conserved among mammals, there are at least 25 other transcription factors, including NeuroD2 (McCormick, 1996), a known mediator of neuronal differentiation; its conserved NRSE is located ~ 13 kb downstream in mammalian genomes and was validated by the ChIP experiments. Another pro-neural transcription factor with an NRSE is Neurogenin-3, which marks both a subset of neuronal precursors and the early precursors of pancreatic islet cells (Sommer, 1996; Gradwohl, 2000). In addition, several genes encoding RNA-binding proteins involved in RNA splicing and editing have NRSEs. Among these, NOVA2 is especially interesting because it regulates brain specific RNA splicing for a substantial group of synaptic proteins (Ule, 2005). Both of NOVA2's NRSEs (one in the third intron, the other one downstream in a LINE2 repeat) were occupied by NRSF/REST according to the ChIP data.

NRSE2 matches are also associated with multiple neuronal microRNAs, several of which (miR-9-1, miR-9-3, miR-29a/miR-29b, miR-124a-1, miR-133, miR-135b, miR-139, miR-153, miR-375), were validated by ChIP. This suggests the circuit model in Fig 2.14: In stem cells and progenitors of Fig 2.14A, NRSF acts by repressing hundreds of protein coding genes and a handful of microRNA genes. Upon developmental progression to the differentiated state (Fig 2.14B), NRSF is downregulated, first at the protein level and then transcriptionally (Ballas, 2005). Thus, its targets are freed — perhaps sequentially according to NRSE strength and number — for induction by various transcription activators. In this model, feedforward connections of microRNAs onto CoREST and

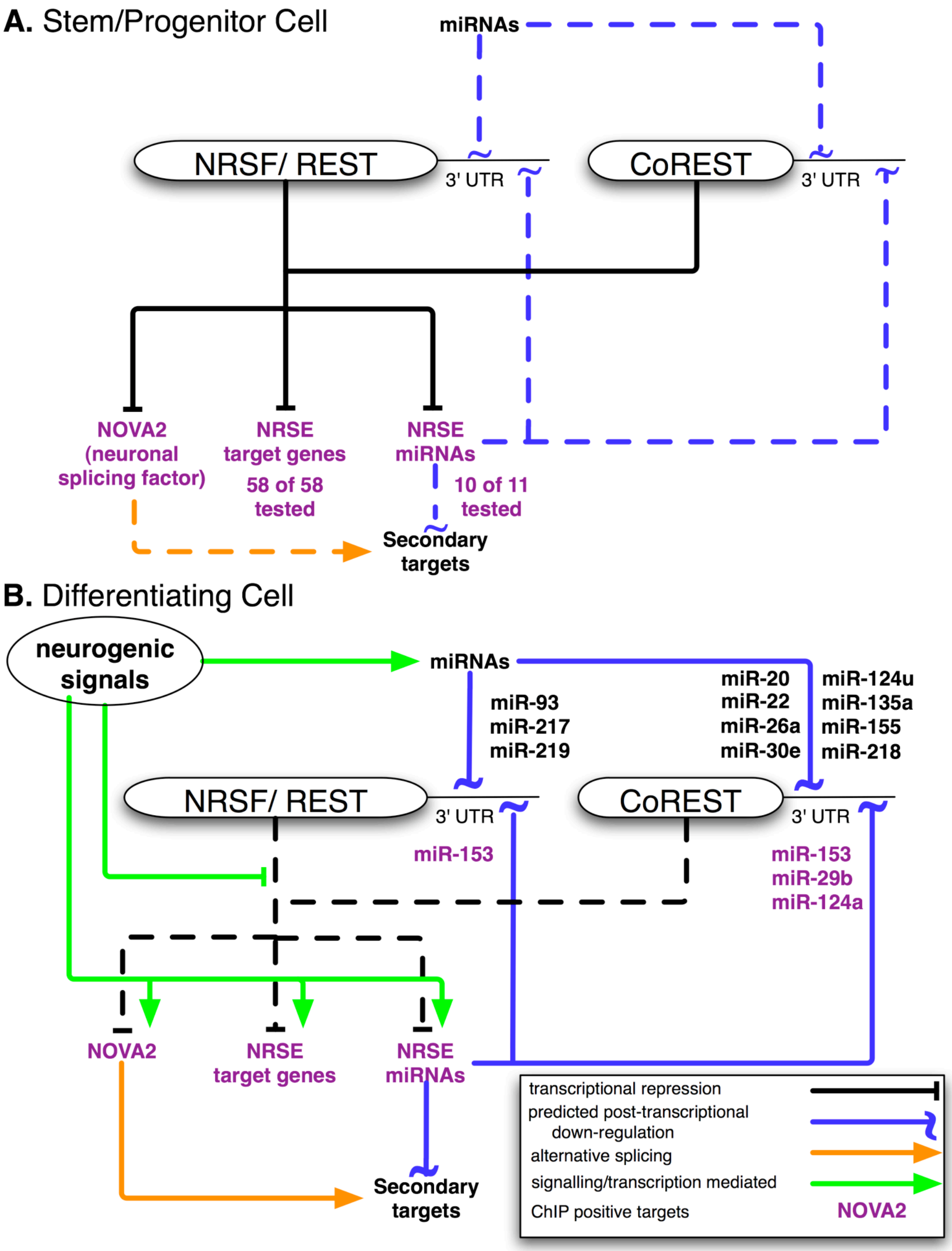


Fig. 2.14. NRSF gene regulatory network model. A. NRSF in conjuction with CoREST and other co-repressors prevents the transcription of several hundred targets, including neuronal splicing factors, transcription factors, and microRNAs, as well as many terminal differentiation genes in a stem cell. **B**. Upon receiving neurogenic signals to terminally differentiate, the NRSF protein is degraded, which leads to derepression of its targets, which are now available to activators. In particular the NRSE-associated miR-153, which is embedded in the pan-neuronal gene PTPRN that has a NRSE in one of its introns, is predicted to down-regulate both NRSF and CoREST mRNAs (which is also the predicted target of the NRSE-associated miR-29b and miR-124a), thus maintaining the derepression.

NRSF may modulate or accelerate the change from precursor cell to neuron. MicroRNAs and splicing factors can go on to down-regulate other target genes not wanted in differentiating neurons. This extended reach of NRSF from direct negative regulation to indirect positive regulation may also explain why only a fraction of neuronal genes are direct NRSF targets. Embryonic lethality of NRSF null mice at day E10.5, before the onset of neurogenesis (Chen, 1998), might therefore result from mis-expression of neuronal microRNAs or splicing factors.

METHODS

Cistematic

Cistematic is a Python package for automated motif identification in eukaryotic genomes. Cistematic has a 3-tiered architecture of objects written in the Python scripting language, which encapsulate the concepts of motifs, genome sequences and annotations, as well as motif-finding programs (Fig 2.13). The sequences and annotations that Cistematic uses for vertebrate genomes are derived from the UCSC Genome database. The primary objectives of Cistematic are to identify, refine, and/or map candidate motifs by determining their genome-wide distribution, their association with potentially co-expressed or co-regulated genes, and their GO enrichment.

A typical Cistematic script consists of Python commands that perform a set of operations on certain Cistematic objects. A set of Experiment objects provides ready-made logic to do much of the work for the user. Most of these Experiment objects are designed to handle various aspects of phylogenetic footprinting across multiple metazoan and fungal genomes. Cistematic stores all of its information and results in SQL-queryable databases, using the

Sqlite 3.0 database library and the pysqlite 2.0 Python library. Cistematic can also generate tab-delimited files that can be imported into Excel for browsing. Cistematic currently runs on Mac OS X, Linux, and Solaris with Python 2.4 and sqlite installed and is available at http://cistematic.caltech.edu, along with the scripts used to generate the data in this paper, which are available at http://cistematic.caltech.edu/~alim/cispaper .

Motif Similarity Score

We define the motif similarity score of two PSFMs A and B as:

$$\mathrm{MSS}(A, B) = \mathrm{Max}(\Sigma\ \mathrm{PearsonCorr}(A_i,B_i),\Sigma\ \mathrm{PearsonCorr}(A_i,\mathrm{rev}B_i))/\mathrm{length}(A)$$

where the index i represents the corresponding columns in the PSFMs, revB is the reverse complement PSFM of B and PearsonCorr is the Pearson Correlation. The MSS of two motifs ranges between 0 and 1.0.

Genome-wide Cis-Regulatory Cohort Identification

We used the Cistematic Locate experiment object class to map every instance of our motifs in human, mouse, and dog with either the consensus or the PSFM. The consensus score for a candidate window m of length L was calculated as:

$$\Sigma\ f_{i_{i}}(m_{i})$$

where f_i is the frequency of the nucleotide at position m_i in the i^{th} column of the PSFM.

The best possible score for each PSFM was calculated and all matches that scored higher than the best score times a predetermined threshold (see Results and Fig 2.4) were accepted as matches. We have found that this particular scoring function performs as well as the traditional log-likelihood scoring (data not shown), allows us to use PSFMs without resorting to pseudo-counts or Dirchlet distributions to account for unseen valid nucleotides, and that the threshold can be intuitively related to the number of mismatches of the site to the consensus of the PSFM (about 5% per major mismatch in the case of NRSE2).

One or more genes were identified for every match as members of the cis-regulatory cohort using the criteria that the match instance is (a) within the gene model or (b) within a 10 kb radius of either the 3’ or 5’ gene model boundaries. The relative location of the motif to each neighboring gene was noted as upstream, 5’-UTR, coding sequence, intron, 3’-UTR, or downstream. Results from each genome were saved to a separate file to serve as inputs for the ensuing steps of the analysis.

We used the following annotations from NCBI or UCSC along with the corresponding genomic sequences from UCSC: human (NCBI Build 35), mouse (NCBI Build 35), dog (NCBI Build 2), and Tetraodon (geneid, UCSC tetNig1).

Orthology Matching

Genes from each genome that were flagged as neighbors in our genome-wide search were cross-matched using the Cistematic orthology database, which is built from a combination of NCBI’s Homologene (version 41.2) supplemented with pre-computed best reciprocal Blast searches for additional genomes that are not yet included into Homologene. Cistematic considers a motif occurrence in genes in a genome (human) conserved if the orthologous gene was present in the genome-wide search results for one or more of the other genomes (here mouse and dog) or if it was present in another paralog in the original genome.

cisMatcher and Motif Refinement

Cistematic can identify motif conservation arbitrarily far from a gene using the cisMatcher algorithm, as outlined in Fig 2.2 and described below. The objective is to be able to specify gene proximity with flexibility that will bring in all flanking sequence to the next gene, for example, whether that distance is several hundred kilobases in a gene desert or only one kb in a gene-dense neighborhood. Cistematic genome-wide results were purged of matches that were marked as occurring within repeats; operationally, these are any of the partially or completely lower-case matches in the genomic sequences from UCSC, which are soft-repeatmasked. Remaining matches were used to retrieve 65 bp sequences with 22 bp upstream from the motif, the motif itself, and the remainder downstream of the motif, which were saved in one file in fasta format per genome. The resulting sequence files from mouse and dog were used to build a Blast database, which was then searched using the human sequences. For each human match, the best match with an e-value less than 0.01 with length longer than 25 bp and similarity greater than 85% in each of mouse and dog were imported into a custom sqlite database. A query was used to retrieve the best mouse or dog match for each human sequence that was available. For each human match with a conserved match in another genome, the nearest human gene within 200-kb was mapped using a radius that was expanded in 1-kb increments; in cases where more than one gene are within the same radius, the one with the lower starting numerical coordinate on the pseudomolecule was picked. Matches were annotated as occurring upstream, in the 5’-UTR or 3’-UTR, in the coding sequence, introns, or downstream relative to their nearest gene.

Matches within coding sequences were optionally filtered where indicated. Matching sequences, and their corresponding gene and relative locations are sorted in decreasing order of similarity and length.

Human matches that were picked up by cisMatcher as well as their corresponding mouse and/or dog matches are then used to calculate the refined PSFM, which can then be used again by Cistematic to repeat the analysis.

Gene Ontology Analysis

Cistematic can flag particular Gene Ontology terms as enriched or depleted, at a statistically significant level, for any set of genes. Cistematic tabulates gene ontology (GO) terms associated with a gene cohort using its own GO annotations, provided for mammalian genomes from NCBI's loc2go dataset. P-values are calculated for every GO term using the hypergeometric. We apply a stringent protocol for significance in which the Bonferroni correction is applied to account for multiple hypotheses testing, where each GO term in the genome represents a hypothesis. We report as significantly enriched or depleted GO terms that (a) are still significant following the Bonferroni correction, and that (b) contain more than 15 genes in the genome. Note that we only show the GO terms (rows) in our GO summary figures that have at least one statistically significant enrichment or depletion in one cohort (column) included in each figure.

To test for the robustness of our analysis, we also generated 100 motifs where we scrambled the order of the columns in NRSF and repeated the entire analysis pipeline in Fig 2.1 and asked whether we recovered any enriched GO terms.

Expression Analysis

The GNF expression dataset was pre-processed by discarding all entries with NCBI gene ID's that are missing or that are not found in the latest NCBI human annotations. If more than one expression pattern for the same gene ID was available, only the first one was kept. For the remaining genes, tissue replicates were averaged and each gene was median-centered.

Confusion matrices were done as by Hart el al (2005), with the following modifications to accommodate genes present in only one cohort. Outersects were defined as the relative complement of each cluster i of set A with respect to set B, i.e.,

$$A_i \setminus B = \{x \mid x \in A_i, x \notin B\}$$

Cell Culture Conditions

Culture conditions were as follows: Jurkat cells (Schneider et al., 1977) were grown in Advanced RPMI 1640 (GIBCO Invitrogen Cell Culture, Carlsbad, CA) supplemented with 15% fetal bovine serum, 100 U/ml of penicillin-streptomycin, and 1x Glutamax (GIBCO Invitrogen Cell Culture, Carlsbad, CA) at 37°C with 5% CO2.

Chromatin Immunoprecipitation

This protocol was adapted from the laboratory of Peggy Farnham (http://mcardle.oncology.wisc.edu/farnham/protocols). We cross-linked the Jurkat cells by adding formaldehyde to a final concentration of 1% for 10 minutes. Cross-linking was stopped by adding glycine to a final concentration of 0.125 M. Then, we collected 2 x 10^7 cells per IP and washed once with 1x phosphate-buffered saline (PBS). We resuspended the cells in lysis buffer (5 mM 1,4-piperazine-bis-(ethanesulphonic acid), pH 8.0, 85 mM KCl, 0.5% NP-40, Protease Inhibitor Cocktail [Roche, Indianapolis, IN]) and centrifuged to collect the crude nuclear preparation. We resuspended the crude nuclear preparation in RIPA buffer (1x PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate [SDS], Protease Inhibitor Cocktail) and sonicated at power output 5–6 with the Sonics Vibra-Cell VC130 (Sonics, Newtown, CT) 4 times for 30 seconds each on ice to produce an average DNA fragment size of 500 base pairs. We centrifuged the chromatin solution at 4°C for 15 minutes at 20,000 rcf. Sonicated chromatin was incubated with NRSF mouse monoclonal antibody (12C11; Chen et al., 1998) coupled to sheep anti-mouse IgG magnetic beads (Dynabeads M-280, Invitrogen, Carlsbad, CA). After bead pelleting, the supernatent was retained as mock IP DNA for use in quantitative PCR. The magnetic beads were washed five times with wash buffer (100 mM Tris, 500 mM LiCl, 1% NP-40, 1% Deoxycholate), and washed once with TE (10 mM Tris at pH 8.0, 1 mM EDTA). After washing, the bound DNA was eluted by heating the beads to 65°C in elution buffer (0.1 M $NaHCO_3$ and 1% SDS). The eluted DNA and mockIP DNA were incubated at 65°C for 12 h more to reverse the cross-links. Then, we extracted with phenol-chloroform and back extracted the organic phase once. We concentrated the DNA in the aqueous phase using the QIAquick PCR Purification Kit (QIAGEN Inc., Valencia, CA), substituting 3 volumes of Qiagen Buffer PM for 5 volumes of Qiagen Buffer PB.

Quantitative PCR

We used Primer3 software to design primers by inputting 500 bp of upstream genomic sequence and 500 bp downstream of each predicted NRSE. Each primer pair was required to flank the NRSE. We performed real-time PCR to quantitate the absolute amount of enriched DNA for each NRSE (amplicon size range between 60–217 bp, average size of 79 bp). Each reaction contained 3.5 mM $MgCl_2$, 0.125 mM dNTPs, 0.5 uM forward primer, 0.5 uM reverse primer, 0.1X Sybr Green (Molecular Probes Invitrogen Detection Technologies, Carlsbad, CA), 1U Stoffel fragment (Applied Biosystems, Foster City, CA), and template DNA in a final volume of 20 uL. For each amplicon, we measured a standard curve of 50 ng, 5 ng, 500 pg, and 50 pg mock IP DNA in addition to our replicate ChIP DNA samples. We measured product accumulation for 40 cycles on the Bio-Rad Icycler and calculated the threshold cycle for each dilution of the standard curve. We then performed a linear regression to fit the threshold cycle from our ChIP DNA sample to this standard curve and divided that result by the amplicon size to measure the absolute number of genomic equivalents of that NRSE in the pool of ChIP DNA. We measured the levels of five random non-genic, non-conserved regions in each ChIP DNA preparation to normalize for any variation in absolute quantities of DNA in each prep.

Acknowledgements

The authors thank Drs. Erich Schwarz and Brian Williams for thoughtful comments on the manuscript; Dr. Ken McCue for advice on statistical analysis; members of the Wold and Myers groups, Sarah Aerni, and Profs. David Bartel and Paul Sternberg for helpful discussions of the project. This work was supported by a NIH/NSF Southern California Bioinformatics Summer Institute Fellowship and a NIH Training Grant Fellowship to AM, a Hubert Shaw and Sandra Lui Stanford Graduate Fellowship to ECL, a Stanford Genome Training Program traineeship to SG (NHGRI grant T32 HG00044), NIH grant U01 HG003162 to RMM, and grants from DOE BER and NASA to BJW.

Chapter 3

A COMPARATIVE ANALYSIS OF NRSF BINDING SITES IN BOREOEUTHERIAN MAMMALS USING CHIP-SEQ

Abstract

In vivo transcription factor-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these interactions comprehensively across multiple mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIP-seq) based on direct ultra-high-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuronal restrictive silencing factor NRSF/REST, to 2155 regions in the human, 2520 regions in the mouse, and 2104 regions in the dog genomes. The data display sharp resolution of binding position (+/-20 bp), which facilitated motif finding and allowed us to identify noncanonical NRSF binding motifs. These ChIP-seq data also have high sensitivity and specificity (ROC area ≥ 0.96), properties that were important for inferring new candidate interactions that include key transcription factors in the gene network that regulates pancreatic islet cell development. Analysis of the evolution and turn-over of NRSF binding sites in the three boreoeutherian genomes shows that while there is a core subset of conserved sites and of conserved target-genes, two-third of the sites and target genes are specific to each species, and also show a preferential evolution of non-canonical sites into canonical sites over time.

Introduction

Although much is known about transcription factor binding and action and the conservation of their binding sites at specific genes, far less is known about the composition, function, and conservation of entire factor:DNA interactomes, especially for organisms with genomes larger than yeast (Harbison, 2004), beyond several computational analyses of conservation of non-coding elements in fly (Stark, 2007), and mammals (Xie, 2005). Now that the human, mouse, and dog genomes have been sequenced, it is possible in principle to measure how any transcription factor is deployed across each genome for a given cell type and physiological condition and to assess the evolution of the target repertoire. Such measurements are important for systems level studies because they provide a global map of candidate gene network input connections and provide us with a test case for the value of transcription factor binding site conservation in identifying functional sites. These direct physical interactions between transcription factors or cofactors and the chromosome can be detected by chromatin immunoprecipitation (ChIP; Kim, 2006). In ChIP experiments, an immune reagent specific for a DNA binding factor (an antibody) is used to enrich target DNA fragments to which the factor was bound in the living cell. The enriched DNA fragments are then identified and quantified.

For the gigabase-size genomes of vertebrates, it has been difficult to make ChIP measurements that combine high accuracy, whole-genome completeness, and high binding site resolution (less than 200 bp). These data quality and depth issues dictate whether primary gene network structure can be inferred with reasonable certainty and comprehensivity, and how effectively the data can be used to discover binding site motifs by computational methods. For these purposes, statistical robustness, sampling depth across the genome, absolute signal, and signal-to-noise ratio must be good enough to detect nearly all in vivo binding locations for a transcription factor with minimal inclusion of false positives. A further challenge in genomes large or small is to map factor binding sites with high positional resolution. To address this issues, we turned to ultra-high-throughput DNA sequencing to gain sampling power and applied size selection on immuno-enriched DNA to enhance positional resolution.

The ChIPSeq assay shown here differs from other large-scale ChIP methods such as ChIP-chip (Kim, 2006), ChIP-SAGE (SACO; Impey, 2004) or ChIP-Pet (Wei, 2006) in design, data produced, and cost. The design is simple and robust as, unlike SACO or ChIP-Pet, it involves no plasmid library construction. Unlike microarray assays, 75% of the genome sequence that is single-copy are accessible for ChIP-seq assay (Fig 4.1D), rather than the ~ 50% that are selectable to be array features in custom or tiling arrays. In addition, the sequence counting nature of ChIP-seq avoids constraints imposed by array hybridization chemistry, such as Tm-related base composition constraints, cross-hybridization, and secondary structure interference. Finally, ChIP-seq is feasible for any sequenced genome, rather than being restricted to species for which whole-genome tiling arrays have been produced and hence makes genomes such as dog as practical to work with as human and mouse.

ChIP-seq illustrates the power of new sequencing platforms, such as those from Solexa/Illumina and ABI/SoLID, Helicos (Harris, 2008), and Polonies (Kim, 2007) to perform sequence census counting assays. The generic task in these applications is to identify and quantify the molecular contents of a nucleic acid sample whose genome of origin has been sequenced. The very large numbers of short of individual sequence reads produced by these instruments (currently reaching up to ~ 80 million reads of 25 nt, per instrument run, depending on the platform used) are extremely well suited for making direct digital measurements of the sequence content of a nucleic acid sample. By determining a short sequence read from each of many randomly selected molecules from the sample that was incorporated into the library, then informatically mapping each sequence read onto the reference genome, the identity of each starting molecule is learned and its frequency in the sample calculated. Desired levels of sensitivity and statistical certainty needed to detect rare molecular species can be achieved, in principle, by sequencing the library deeper. Sequence census assays do not require knowing in advance that a sequence is of interest as a promoter, enhancer or RNA-coding domain as most current microarray designs do. The Solexa/Illumina platform used in this chapter provides the best combination of price effectiveness and robustness at this time.

We used ChIP-seq to build a high-resolution interactome map for human mouse and dog Neuron Restrictive Silencer Factor (also known as RE-1 Silencing Transcription Factor). This vertebrate-specific zinc finger repressor negatively regulates many neuronal genes in stem and progenitor cells and in non-neuronal cell types, such as the Jurkat T-cell line studied here using a variety of cofactors such as Sin3a, CoREST, and Small CTD-phosphatases (Ballas, 2005). A primary reason for selecting NRSF as a test case is that prior studies provide a rich set of known "gold standard" target genes, including more than 80 in vivo binding sites defined by ChIP-QPCR (Mortazavi, 2006). A subset of these have also been tested for regulatory function by transfection assays (Schoenherr, 1996). In addition, the DNA motif bound by NRSF, called NRSE (also called Repressor Element-1), is long (21 bp) and well specified (Mori, 1992), and was in fact used to identify NRSF. This has led to a rich group of computational models for the site and for all its occurrences in several mammalian genomes (Mortazavi, 2006; Schoenherr, 1996; Bruce, 2004; Zhang, 2006; Wu, 2006). These sites provide a framework of explicit predictions that can now be tested by measuring repressor binding globally. Finally, there is a high quality monoclonal antibody (Chen, 1998) that recognizes the N-terminus of NRSF efficiently in ChIP experiments in human (Mortazavi, 2006), mouse, and dog as well as a polyclonal antibody to the C-terminus that is also used in ChIP-studies (Ballas, 2005).

An earlier version of this work focusing on identifying NRSEs in human Jurkats using only the monoclonal antibody was published in *Science* (Johnson, 2007) using an early version of the peak-finder. The results and figures below were rederived using the current enhanced version of the peak finder and also include unpublished comparisons to the polyclonal antibody as well as additional sites that are only found once, including low-prevalence multireads (reads that map to more than one, but less than ten locations in the genome). The mouse and dog ChIP-seqs are also novel and, when combined with the original data, allow for new insights that we could only speculate on at the time of the original paper.

ChIP-seq with N-terminus monoclonal antibody in human Jurkat cells

We first prepared two DNA samples for each ChIP-seq experiment: an NRSF/REST-enriched ChIP sample and a companion control sample of the same fixed chromatin, but without immuno-enrichment. In an effort to increase positional precision and to provide optimal substrate for the Solexa/Illumina sequencing platform, we introduced a size selection step after crosslink reversal (Materials and Methods), which likely greatly contributes to our increased positional resolution. DNA sequencing of each sample was performed by the standard Solexa protocol. 2 to 5 million 25 nt sequence reads were produced per sample, of which approximately half mapped to single sites in their respective genome (Table 3.1). Sequence reads that map to more than 1 but less than 10 location in the genome were analyzed separately to gage their effect on the enrichment predictions, while reads mapping to more than 10 sites were removed from subsequent analyses. This eliminates sequences in simple repeats and in some recent complex repeats. The location of each remaining unique sequence read in the genome was recorded. To accommodate polymorphisms relative to the reference genome, up to 2 mismatches were allowed. The resulting sequence read distribution was processed with a ChIP-seq peak locator algorithm developed for this purpose (now part of dual use ERANGE package). The algorithm finds a local concentration of sequence hits (a location cluster), and within that location calls a peak. We then required of these a minimum 5-fold enrichment of sequence reads in the ChIP sample relative to the corresponding location in the control. Five-fold enrichment is a conservative choice among enrichment thresholds commonly used in contemporary large-scale ChIP studies. A location that passed these criteria and also had 8 or more sequence reads per million reads (8 RPM, a threshold value selected based on the sensitivity and specificity analysis described below) was called as an NRSF-positive binding event.

An example of primary ChIP-seq data from two independent experiments is shown in Fig 3.1 for the NeuroD1 locus, a transcription factor involved in neurogenesis (Lee, 1995).

This positive signal, which has intermediate signal intensity identifies a novel NRSF binding target that is recognizably, but weakly related to the NRSF consensus. The NRSF/REST sequence tag distribution centers directly over the only canonical NRSE motif in a 4 kb region, which is located in the open reading frame of the NeuroD1 gene. The

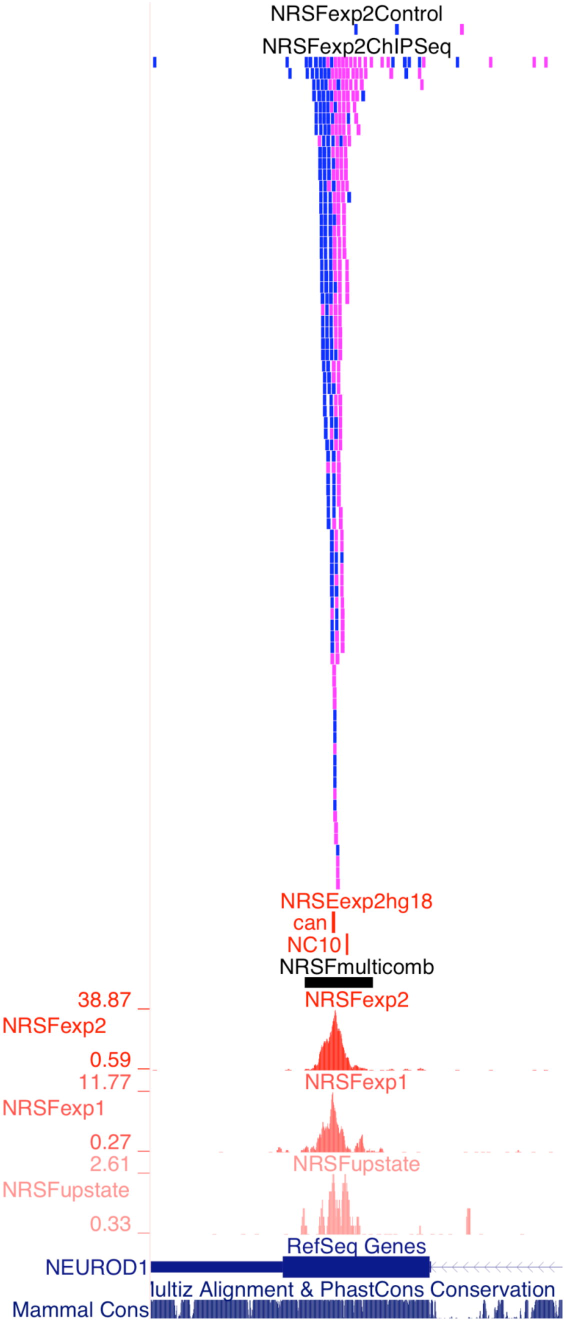


Figure 3.1. ChIP-seq reveals a canonical NRSE within the coding sequence of NeuroD1. 25 bp reads from multiple ChIP-seq experiments in the Jukat cell line were mapped onto the human genome (UCSC hg18) in a 3 kb region of the neurogenic NeuroD1 locus. Individual reads for the control and ChIP and control for experiment 2 using the monoclonal anti-NRSF N-terminus are shown in blue or purple, depending on whether they map to the forward or reverse strand, respectively. A 500 bp region was called by the ERANGE peak finder as enriched (black), within which 2 candidate sites were found. The canonical NRSE (CAN) is within 3 bp of the peak of the signal. The normalized UCSC genome browser wigglegram representations of experiment 2, experiment 1, and an experiment using the upstate polyclonal antibody against the C-terminus are also shown. Note the that while all three signal peaks agree, experiment 2 show a substantial signal enrichment versus both experiment 1 and the upstate experiment, even though the upstate experiment was done following the exact same protocol as experiment 2. The upstate (anti-C-terminus) experiment shows an equally strong peak at the non-canonical site (NC10), whereas the signal from the canonical site predominates in the anti-N-terminus experiments.

reads further show the typical ChIP-seq directionality transition at the binding site. This site was called as a ChIP-seq peak by the locator algorithm (now part of ERANGE). A previous study had implicated NRSF in repression of NeuroD1, but had failed to find a local site computationally and hence theorized long-distance repression to explain the effect (Lunyak, 2002), but our ChIP-seq results suggest a simpler explanation of a degenerate site within the NeuroD1 coding sequence. Over the entire primary dataset, the distribution of sequence-tag number per location ranged from the threshold value of 8 RPM (Reads Per Million, threshold determined from the ROC curve analysis discussed below) to a maximum of 3174 RPM at the highest signal. The two ChIP-seq experiments produced similar results (Fig 3.2), with our experiment 2 mapping 2155 enriched regions, most of

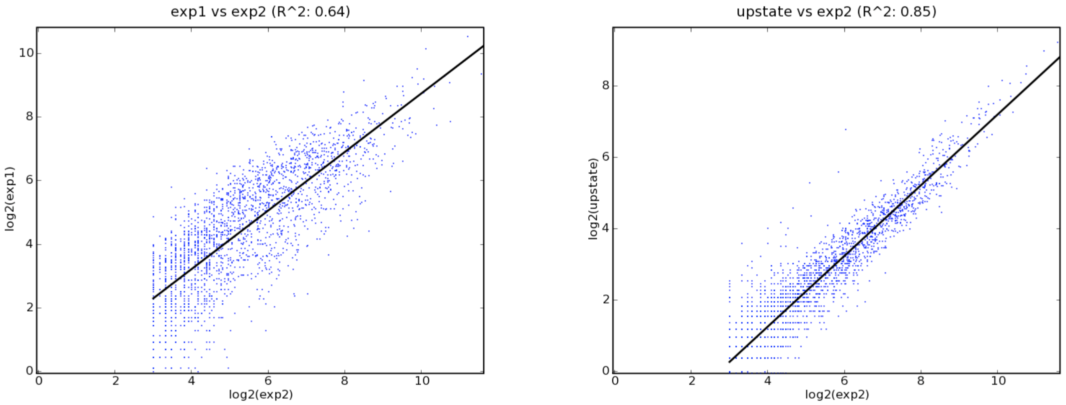


Figure 3.2. Comparison of ChIP-seq normalized counts between different experiments in enriched regions. We compared the results obtained from our different experiments done in Jurkats in the regions identified by our algorithm as enriched in experiment 2 above background. We first compare experiment 1 (60 IPs pooled, unamplified) versus experiment 2 (4 IPs pooled, LM-PCR amplified before gel selection) and find a reasonable correlation (R^2 = 0.64). The correlation of signals between experiment 2 and the anti-C-terminus (upstate) experiment is quite good (R^2 = 0.85), suggesting that both antibodies see the same set of sites.

which occur in or near 1272 genes. We can recover 90% of these regions in Experiment 1, even though our enrichment is substantially lower. Since all subsequent experiments are done following the protocol for experiment 2, we will only use the results for experiment 2 in all further comparisons.

Assessing ChIP-seq efficiency with the C-terminus NRSF antibody in Jurkats

ChIP-seq is only as good as the immunoprecipitation that is performed before the library building. Major sources of ChIP variability include experimental design (how many cells, how much cross-linking), epitope availability, as well as biological variability (expression level of a factor). In addition, the epitope might be species-specific, such as in the case of NRSF, where the N-terminus antibody was raised to specifically not recognize the chicken ortholog (Chen, 1998) A less-efficient ChIP would result in wanted bound-DNA fragments representing a smaller fraction of the total pool of recovered DNA; the resulting higher background would hence give a lower signal-to-noise ratio. Hence we will use the fraction of reads that fall within enriched regions as a measure of IP efficiency (called fIP) that allows us to compare the performance of the same antibody across different conditions or of different antibodies. The fIP for the two human monoclonal experiments are 0.12 and 0.2 and are listed for all experiments in this chapter in Table 3.1.

The availability of a second ChIP-grade commercial polyclonal antibody to the C-terminus of NRSF from Upstate allows us to compare our enriched regions independently in the same Jurkat cell line and assess the strength of the two antibodies relative to one another.

Sample	*IP reads*	*Mock reads*	*#regions (8RPM+)*	*IP efficiency (fIP)*
Jurkat 1	3.7M	3.8M	1991	0.12
Jurkat 2	1.7M	2.3M	2155	0.2
Jurkat Upstate	3.1M	2.3M	1033	0.03
EL4	6.5M	7.2M	1320	0.1
C2C12	3.4M	7.5M	2520	0.1
MDCK	5.9M	4.3M	2104	0.1

Table 3.1. Summary statistics of our ChIP-seq experiments

Our C-terminal NRSF ChIP-seq analyzed with the same parameters as experiment 2 only returned 1033 regions with a fIP of 0.03, which matches our expectation for lower enrichment with the polyclonal antibody based on ChIP-QPCR (data not shown). A scatter plot of the upstate polyclonal ChIP-seq regions to the monoclonal antibody shows that the signal from the two antibodies are highly correlated (Fig 3.2), but that the C-terminus ChIP-seq simply does not capture the weaker regions seen by the monoclonal antibody above threshold. While there exist neuronal settings where we expect the two antibodies to give us biologically significant differences in signal from recognizing different NRSF splice isoforms (see REST4 below), we do not believe this to be the case in Jurkats.

NRSF ChIP-seq in mouse and dog

In order to study the extent of conservation of NRSE in mammals, we performed NRSF ChIP-seq using the monoclonal N-terminus antibody in the mouse lymphoblastoid EL-4 cell line, the undifferentiated mouse myoblast C2C12 line, and the dog MDCK cell lines. While the EL-4 cell line can be thought of as roughly equivalent to Jurkats, C2C12 and MDCK cells are adherent cell lines. Using the same threshold as used previously, we found 1320 enriched regions associated with 945 genes in EL-4s, 2520 enriched regions associated with 1927 genes in C2C12s, and 2104 enriched regions associated with 1300 genes in MDCK. All three ChIP-seq experiments have an fIP of 0.1. However only 994 regions and 716 genes were in common to both EL4 and C2C12, suggesting that the NRSF binding repertoire is not static, but might change to a certain extent depending on the cell state including variables such as chromatin accessibility.

Assessing ChIP-seq resolution by comparison to known sites

Human NRSF binding sites previously identified by QPCR or transfection assays (Mortazavi, 2006) plus a set of known negatives derived from custom Agilent tiling array experiments (data not shown) were used to measure sensitivity (successful detection of true positives) and specificity (successful rejection of true negatives) of the ChIP-seq assay in human. A ROC (receiver operator characteristic) analysis provides a way of measuring and graphically portraying sensitivity (fraction true positive on the Y axis) versus specificity (1- the fraction false positives, displayed on the X axis) (Fig 3.3). The observed ROC areas for experiment 2 is high at 0.96. The selected threshold of 8 sequence reads per region per million reads required for inclusion in the ChIP-seq interactome corresponds to a sensitivity of 87% and a specificity of 98%. We conclude that the ChIP-seq NRSF measurements are accurate and, as suggested by P-values (not shown), statistically robust.

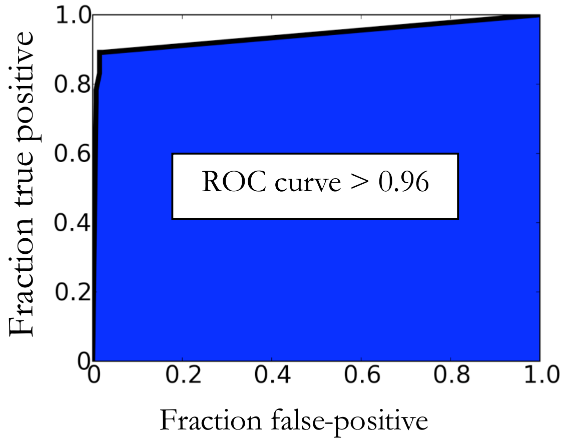


Figure 3.3. ROC curve for experiment 2 (area under curve > 0.96) showing the performance of ChIP-seq in detecting previously validated true positives (y-axis, 83) and true negatives (x-axis, 130). Our sensitivity at the thresholds used for our analyses are 87% and our specificity is 98%. All of the ChIP-seq false negatives are at marginal sites that barely pass the threshold in the validation assay.

We next assessed the precision of ChIPSeq site location relative to 771 computationally high-scoring NRSE motifs in the genome that also have positive ChIPSeq signals by measuring the distance from the experimental ChIPSeq peak to the center of the computational NRSE sequence motif. 754 sites in this group were ChIP-Seq positive in the two human monoclonal experiments, and the center of a 21 bp NRSE motif was within +/-

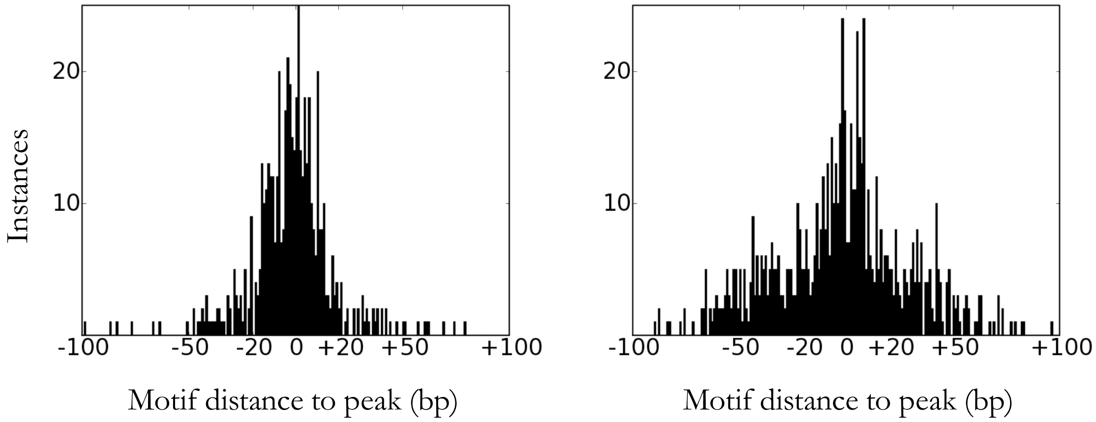


Figure 3.4 Distribution of the distance of the center of 771 canonical human NRSEs (84%+ of optimal score) in ChIPSeq enriched regions to the called ChIPSeq peaks in experiment 1. 46% of these peaks fall within the boundaries of the NRSE (here, +/- 10 bp), and 94% of the canonical NRSEs fall within 50 bp of the peak. Increasing the number of reads, as in the case of the NRSF EL4 experiment, improves our resolution to within 20 bp of 751 canonical sites.

50 bp of the called ChIP-Seq peak for 94% of these (Fig 3.4). Increasing read numbers as in the case of our mouse and dog NRSF ChIP-seqs improves our resolution to +/- 20 bp (Fig 3.4), while a new generation of ChIP-seq peak finders may achieve base-pair resolution level for solo sites. The resolution, which depends in part on size selection of sheared chromatin after immunoenrichment, is much higher than is typical for ChIP-chip or ChIP-SAGE (+/- 500 to 1,000 bp; Impey, 2004; Cawley, 2004).

How comprehensive are the NRSF ChIP-Seq measurements? Several lines of evidence address this question. First, as shown in Fig 3.5, virtually all strong canonical NRSF motifs instances across the human genome were detectably occupied. We defined strong sites as those having $\geq$ 90% match to a previously developed motif model (a position specified frequency matrix), which is based on evolutionarily conserved site instances across multiple mammalian genomes *(*Mortazavi, 2006*)*. This high representation of detectable binding suggests that no strong sites were missed by undersampling. It also implies that all sites are accessible for NRSF/REST binding in Jurkat cells, at least part time in some individual cells, although the degree of accessibility might vary and may account for wide differences in the number of tags per site (Fig 3.2). Second, we observed ChIP-seq positive signals for sites previously studied in detail by transfection analysis (Schoenherr, 1996), and they correspond to a wide range of ChIP-seq signals, with all but one scoring positive in both ChIP-seq experiments. Taken together with the sensitivity results (Fig 3.3), these observations suggest that the NRSF/REST interactome measurements are genome-comprehensive and have been sampled deeply enough to include most sites known by any other criteria to be biologically active, even if relatively weakly. This level of genome completeness is attributable to the depth of Solexa/Illumina sequence sampling, and is substantially greater than in prior studies of the CREB interactome measured by SACO (Impey, 2004) and the p53 interactome measured by ChIP-pet (Wei, 2006).

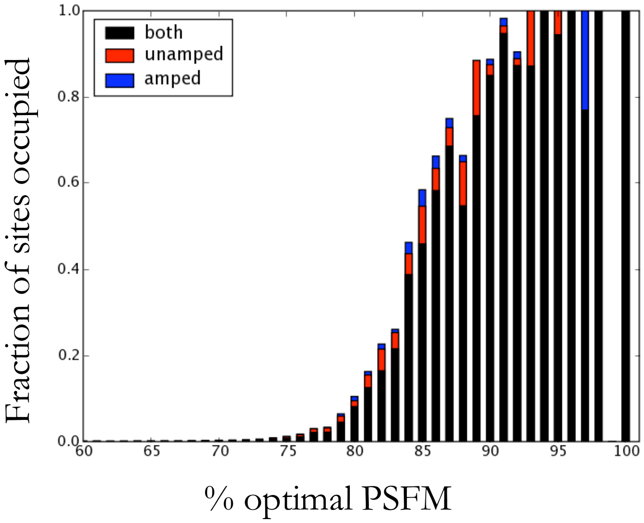


Figure 3.5. Fractional site occupancy of NRSEs (y-axis) in Jurkat cells as a function of PSFM score (x-axis) for experiment 1 (unamped) and experiment 2 (amped). Nearly all high-scoring sites scoring 90% or better are detected.

The impact of multireads on region and site prediction

Two of the previously validated NRSEs in Jurkats (Mortazavi, 2006) that failed to pass the threshold using unique reads are associated with HBA1 and HBA2, which are the result of a relatively recent gene duplication in the human lineage. We therefore evaluated the effect of including multireads (reads that can map up to 10 places in the genome) on our predictions. We reran dual-use ERANGE on the human Jurkat experiment 2 dataset supplemented with 0.2M ChIP Multireads mapping to 0.8M locations and 0.3M Control Multireads mapping to 1.4M locations. This resulted in 2276 enriched regions associated with 1,358 genes with an fIP of 0.19 with a 29 k additional enriched reads coming from the multireads (8%), which is in proportion with multiread prevalence in the genome. Many of these reads come from sites that we already called as enriched based on unique reads alone, but we indeed recover 117 new genes that are part of gene families that are part of segmental duplications, including both HBA genes (Fig 3.6); while it is formally possible that only one of the gene family members would be associated with NRSF, it is difficult to justify that given the similarity of the binding region. However the increase in total number of reads means that some regions that barely passed our threshold with unique reads alone would fail to pass the same threshold once we included additional reads unless these regions also accumulated additional multireads; 31 genes that made the threshold

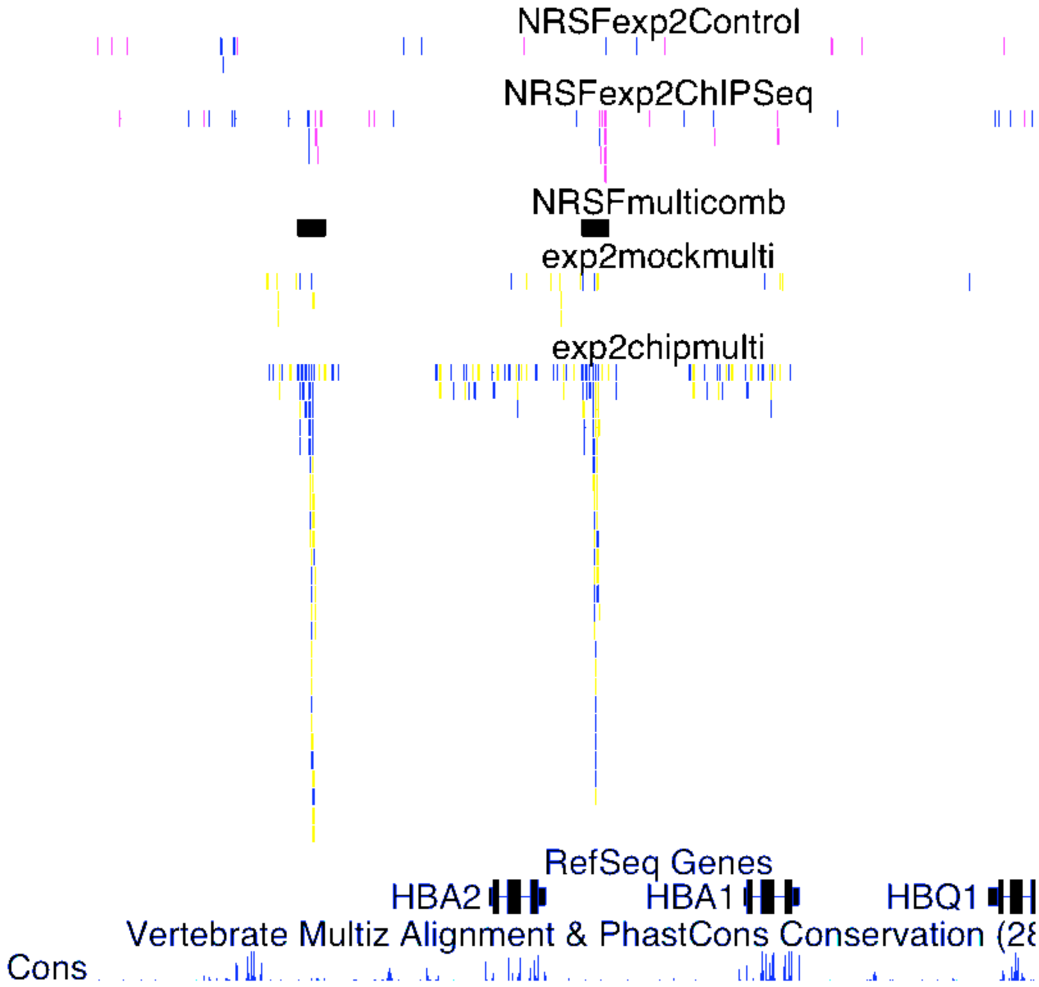


Figure 3.6. Multireads can also be used to define NRSEs. The HBA1 and HBA2 NRSEs, which had been validated in Jurkats (Mortazavi, 2006), did not make our threshold with unique reads. The inclusion of multireads that map between 2 to 10 locations in the genome does indeed reveal both sites.

with unique reads alone fail to make our threshold once including multireads. While the rest of the analysis below is done without further inclusion of multireads, their inclusion may be ultimately necessary in order to correctly track the amount of site turnover in the evolutionary analysis.

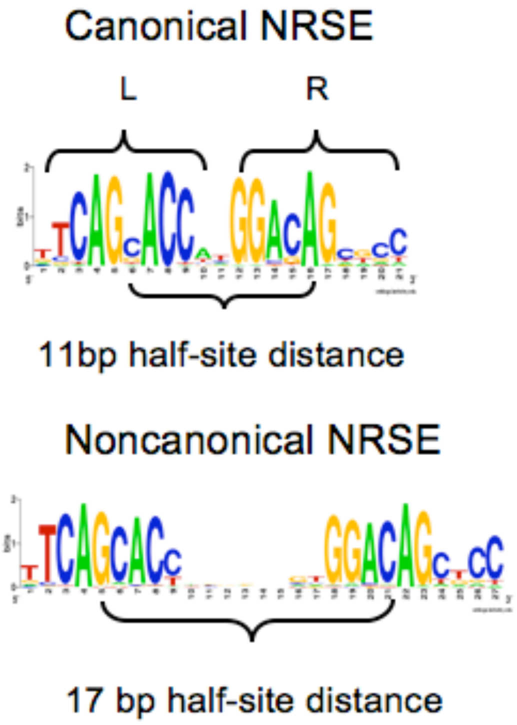


Figure 3.7. Canonical NRSE weblogo (top) showing the two half-sites ("L" and "R") and the canonical 11bp distance between the center two half sites of 11bp compared to weblogo of noncanonical NRSE (bottom) with half-site distance of 17, showing the lack of conservation in the spacer nucleotides.

Identification of a novel family of non-canonical NRSE

The positional resolution and low number of false positives in these experiments can greatly facilitate motif finding algorithms by significantly restricting the search space to a smaller region around the peak, which both improves signal-to-noise and also greatly reduces the run times for many algorithms. The canonical NRSF binding site (NRSE) has been studied extensively (Mortazavi, 2006; Bruce, 2004), and this allowed us to ask whether we recover it using the motif finding algorithm MEME (Bailey, 1995) when a sample of the experimental interactome peak domains are used. When given all sites in the top 10% of signal intensity (100 bp segments from 198 regions having 500 reads or more), MEME returned the full previously known known motif as its top motif. Single or multiple matches to this canonical motif, using a 70% match threshold, account for 75% of all ChIP-seq regions mapped in this study.

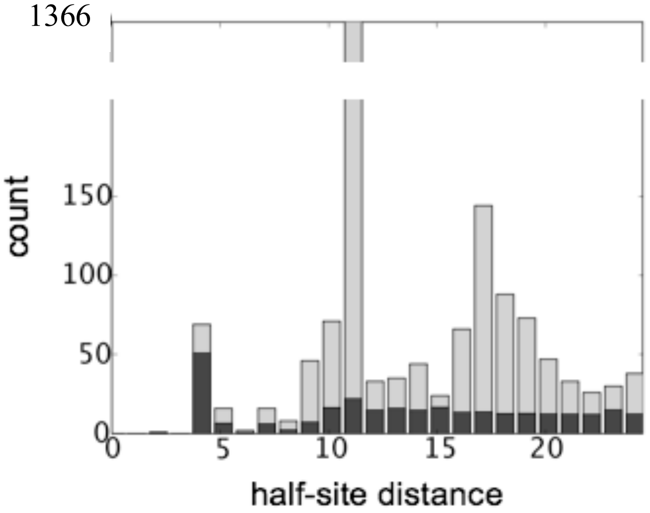


Figure 3.8. Histogram of half-site distances in base pairs in the ChIP-seq-enriched regions, showing the observed (grey) and expected (black, based on frequency in genome) counts. In addition to the expected canonical peak at distance 11, there is also a enrichment of half-sites occurring within non-canonical distances of 16–19 bp.

We next focused attention on those remaining ChIP-seq positive regions that have 175+ RPM, yet have no canonical motif match. There are 22 such locations, and when they were run in MEME, only two candidate motifs stood out. By inspection, the large canonical NRSF binding motif of 21 bp is naturally subdivided into two prominent, non-identical, non-pallindromic half-sites (Fig 3.7). The two motifs from the MEME search correspond directly to the separate left and right sides of the canonical motif. We next asked if these motifs occur at other ChIP-seq binding locations, and if they are organized in any discernable pattern. A distinctive pattern was discovered within 50 bp of many ChIP-seq peaks, in which left and right half site motifs are separated by additional “spacer” sequence that increases the center-to-center distance from the canonical 11 bp to 16–19 bp, or decreases it by one base pair to 10 bp (Fig 3.8). Thus, the canonical site has two central positions that have no sequence specificity, and the noncanonical group is similarly oriented but has increased the separation distance by an additional 5–9 bp (Fig 3.7). These linked half sites, oriented with respect to each other in the same way as in the canonical site, occur in NRSF ChIPSeq binding domains in a statistically significant manner relative to random sequence windows in the genome (χ^2= 1,309 for half-site distance of 17, P-value of 0) and account for 197 regions lacking a canonical motif (Fig 3B and S4). We also found that some binding locations have multiple clustered occurrences of noncanonical motif(s) along with a canonical one.

There are no structural data available for NRSF, so we cannot relate this new family of binding site motifs to a known DNA binding structure. However, the protein has eight zinc fingers in its DNA binding domain, and other C2H2 zinc finger proteins such as Zif268/Egr-1 bind DNA with three fingers per 10 bp turn, but they show considerable strain when binding with six fingers (Peisach, 2003). This makes simultaneous binding of one molecule of NRSF to these non-canonical half-site configurations plausible, but it is also possible that the protein is bound to only one half-site at a time by using a subset of its fingers in these cases. Mutagenesis studies show that Fingers 3–8 are necessary for binding the full NRSE (Shimojo, 2001.) However, the REST4 isoform of NRSF, which only has the first 5 of the zinc finger and is missing the C-terminal, CoREST-interacting domain, can act as an activator when binding cooperatively with the glucocorticoid repressor to turn on glutamine synthetase (Abramovitz, 2008), which is normally repressed by full-length REST. The REST4 binding site does indeed resemble the right half-site (Lee, 2000). Interestingly, REST4 would be recognized by our monoclonal N-terminus antibody, but not by the Upstate C-terminus antibody, which would allow us to verify its binding in vivo in the right setting. It will be interesting to learn if there are other functional and molecular characteristics that set the different classes of sites apart. For example, do the different NRSF co-repressors differ in their interactions at non-canonical sites compared with canonical ones? (Ballas, 2005; Yeo, 2005).

We also asked whether half-sites are significantly enriched in our ChIP-seq neighborhoods, without regard to orientation or spacing, relative to expectations based on their occurrence in the genomes, and found that these regions are greatly enriched for left half-sites (χ^2= 3,070) and right half-sites (χ^2= 11,674). This range of configurations, from concentrated half-sites to the noncanonical 16–19 bp spaced left and right sites, to the canonical 11 bp spaced full site, is quite striking. Significant NRSF binding occurs in vivo, according to our data, at all three kinds of loci. An analysis of the genes with NRSEs (Fig 3.9) reveals that while most of the genes with canonical (75–77%) and non-canonical sites (57–71%) are detected in both C2C12 and EL4, few of the genes with half-sites (6–21%) are in common. In particular, we found far more enriched half-sites in both C2C12 and MDCK when compared to EL4 and Jurkats, which could be a reflection of the more active metabolic state and hence open chromatin state of the two adherent cell lines. We discuss site conservation in the “transcription factor binding site evolution” section.

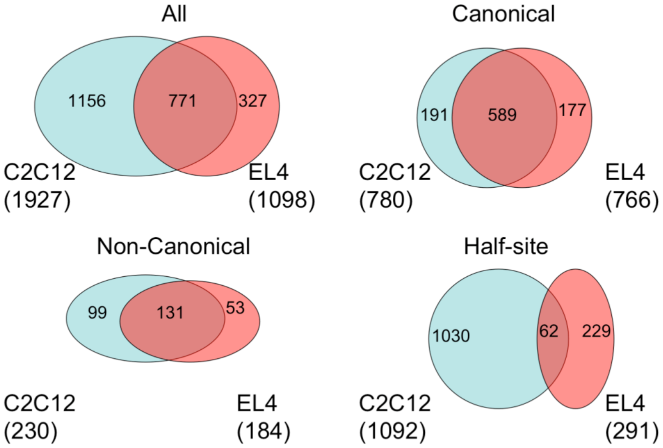
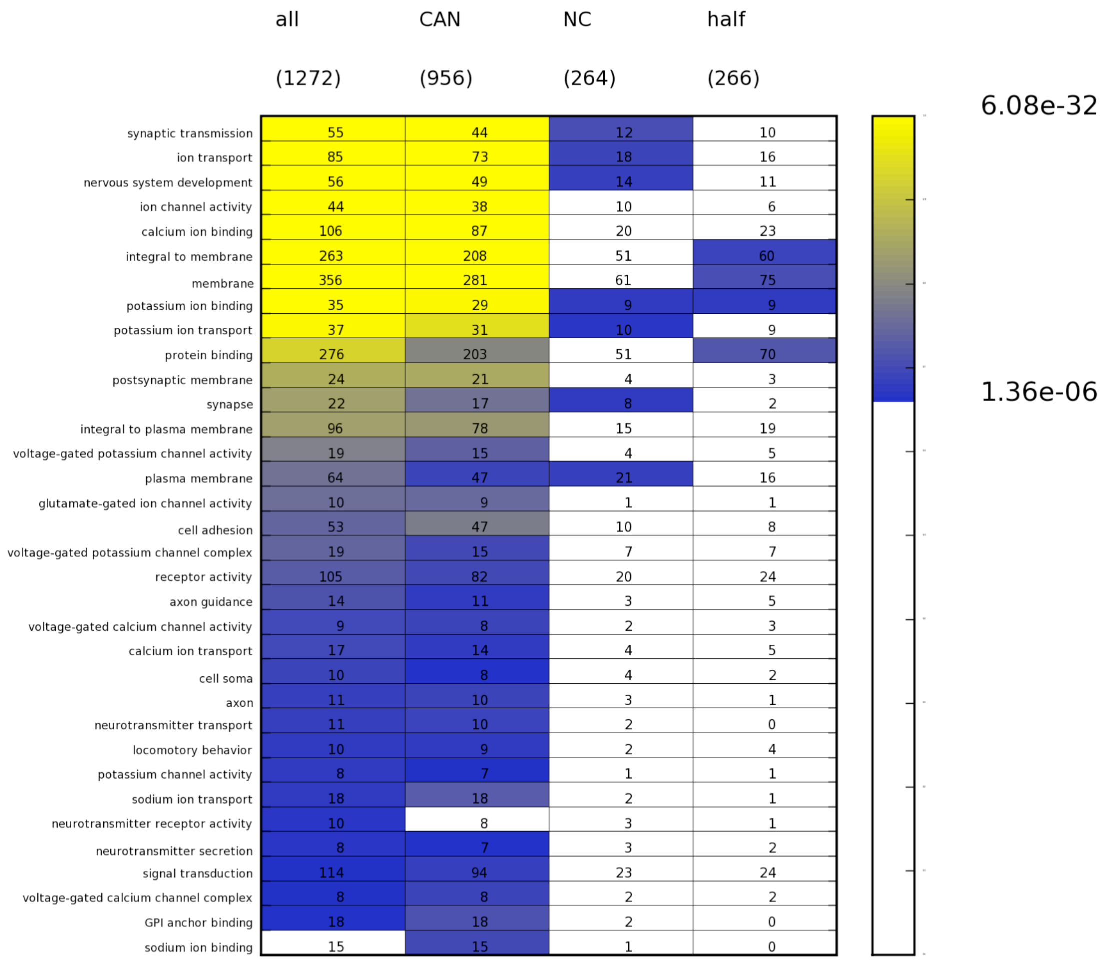


Figure 3.9. Comparison of the genes in common between the ChIP-seq experiments between C2C12 and EL4. While genes with canonical NRSEs were most likely to be in common between the two datasets, genes with half-sites showed little overlap and may represent binding that is dependent on chromatin state.

The conserved core network

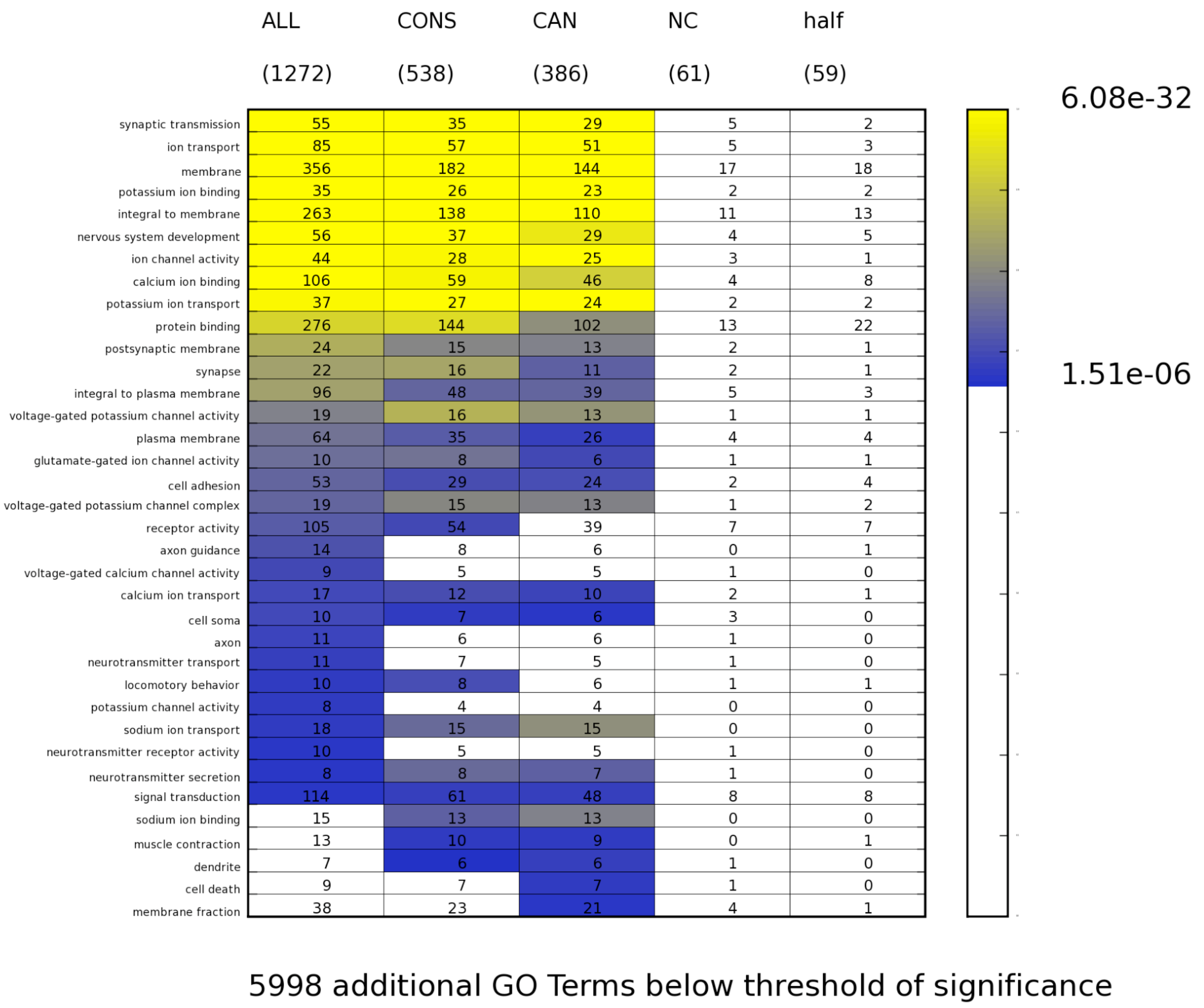
101 of the genes in the NRSF gene cohort in Jurkats are involved in transcriptional regulation (GO:0006355), and we found NRSEs associated with 22 microRNAs, and five splicing regulators. NRSEs occur prominently in introns, including a non-canonical site located about 500 bp downstream of the transcription start site of the NRSF gene itself, which suggests the possibility of negative autoregulatory feedback. We also found, as expected, that NRSF-bound loci are highly enriched in GO terms related to neurons and their development (Fig 3.10). The enrichment for the experimentally determined sites exceeded that achieved for any computationally predicted target gene cohort (Fig 2.9). Synaptic transmission and nervous system development rank in the top three GO terms among 6,000, with P-values for over-representation of the NRSF target genes of 10^{-24} and 10^{-17} (Fig 3.10).



	all (1272)	CAN (956)	NC (264)	half (266)
synaptic transmission	55	44	12	10
ion transport	85	73	18	16
nervous system development	56	49	14	11
ion channel activity	44	38	10	6
calcium ion binding	106	87	20	23
integral to membrane	263	208	51	60
membrane	356	281	61	75
potassium ion binding	35	29	9	9
potassium ion transport	37	31	10	9
protein binding	276	203	51	70
postsynaptic membrane	24	21	4	3
synapse	22	17	8	2
integral to plasma membrane	96	78	15	19
voltage-gated potassium channel activity	19	15	4	5
plasma membrane	64	47	21	16
glutamate-gated ion channel activity	10	9	1	1
cell adhesion	53	47	10	8
voltage-gated potassium channel complex	19	15	7	7
receptor activity	105	82	20	24
axon guidance	14	11	3	5
voltage-gated calcium channel activity	9	8	2	3
calcium ion transport	17	14	4	5
cell soma	10	8	4	2
axon	11	10	3	1
neurotransmitter transport	11	10	2	0
locomotory behavior	10	9	2	4
potassium channel activity	8	7	1	1
sodium ion transport	18	18	2	1
neurotransmitter receptor activity	10	8	3	1
neurotransmitter secretion	8	7	3	2
signal transduction	114	94	23	24
voltage-gated calcium channel complex	8	8	2	2
GPI anchor binding	18	18	2	0
sodium ion binding	15	15	1	0

Figure 3.10. Gene Ontology analysis of Jurkat NRSE gene cohort. The bulk of enrichment is coming from the canonical gene cohort (CAN). While the non-canonical gene cohort (NC) also shows some characteristic neuronal enrichment, the half sites (half) are only enriched in non-specific terms.

How much of the NRSF gene cohort is shared across all three genomes? We find that 538 genes have an NRSE in human and/or dog (Fig 3.11). While this number is quite close to what we predicted computationally, the number of genes falling into particular GO groups are quite different, in particular for genes related to ion transport. Furthermore, the genes analyzed in this section are being identified as conserved without multireads, whereas our computational predictions do include genes that are impacted by multireads. All observed enriched GO categories come from conservation of the canonical gene cohort, as few genes with non-canonical or half sites show up as conserved (Fig 3.11). Interestingly, not all GO



	ALL (1272)	CONS (538)	CAN (386)	NC (61)	half (59)
synaptic transmission	55	35	29	5	2
ion transport	85	57	51	5	3
membrane	356	182	144	17	18
potassium ion binding	35	26	23	2	2
integral to membrane	263	138	110	11	13
nervous system development	56	37	29	4	5
ion channel activity	44	28	25	3	1
calcium ion binding	106	59	46	4	8
potassium ion transport	37	27	24	2	2
protein binding	276	144	102	13	22
postsynaptic membrane	24	15	13	2	1
synapse	22	16	11	2	1
integral to plasma membrane	96	48	39	5	3
voltage-gated potassium channel activity	19	16	13	1	1
plasma membrane	64	35	26	4	4
glutamate-gated ion channel activity	10	8	6	1	1
cell adhesion	53	29	24	2	4
voltage-gated potassium channel complex	19	15	13	1	2
receptor activity	105	54	39	7	7
axon guidance	14	8	6	0	1
voltage-gated calcium channel activity	9	5	5	1	0
calcium ion transport	17	12	10	2	1
cell soma	10	7	6	3	0
axon	11	6	6	1	0
neurotransmitter transport	11	7	5	1	0
locomotory behavior	10	8	6	1	1
potassium channel activity	8	4	4	0	0
sodium ion transport	18	15	15	0	0
neurotransmitter receptor activity	10	5	5	1	0
neurotransmitter secretion	8	8	7	1	0
signal transduction	114	61	48	8	8
sodium ion binding	15	13	13	0	0
muscle contraction	13	10	9	0	1
dendrite	7	6	6	1	0
cell death	9	7	7	1	0
membrane fraction	38	23	21	4	1

Figure 3.11. Gene Ontology analysis of Conserved NRSE gene cohorts. All of the GO enrichment is coming from the conserved canonical gene cohort. Neither the non-canonical gene cohort nor the half sites show any enrichment.

terms are equally conserved. While two thirds of the genes annotated as involved in synaptic transmission show conservation, all of the genes annotated as neurotransmitter secretion are conserved, as well as most genes that have voltage-gated potassium channel activity or sodium transport annotations.

The Pancreatic islet cell network

Amongst the transcription factors identified is a set that have not previously been suggested as NRSF targets, but that are known to be critical in the gene network that drives islet cell development in the pancreas (Fig 3.12). The transcription factors NeuroD1/Beta2, HNF4a, HNF6/Onecut1, and Hes1 were all detected here for the first time as in vivo binding targets of NRSF, and — together with Neurogenin3, which is a previously identified target (Mortazavi, 2006) — they are positioned critically in the regulatory network that controls pancreatic β-cell development (Davidson, 2006). We further find that Pax4, which has previously been described as a target of NRSF (Kemp, 2003) is only bound in mouse; this also extends to Ptf1a. Although in vivo binding does not ensure NRSF repression activity, these regulators are known to function as positive drivers of pancreatic neuroendocrine development. If NRSF repression is active at all these sites, as might be the case in progenitor cells, the circuit repression would be very effectively blocked. In this hypothesis, NRSF acts as a permissivity factor, gating entry into and progress through the developmental pathway. These pancreatic network sites are among the more modest ChIP-seq signals, ranging from 33 RPM for HNF6 to 119 RPM for NeuroD1 in Jurkats, values that are comfortably above the significance threshold of 8 RPM, yet they fall in the bottom quartile. Thus these ChIP-seq data were statistically robust enough to map parts of this gene network that might otherwise have gone undetected or been highly uncertain.

There are precedents in other systems that show that relatively weak sites are biologically important, specifically because they are, in the biochemical binding sense, suboptimal. For example, in *C. elegans*, the Pha4/FoxA factor is the key activator of a large interactome,

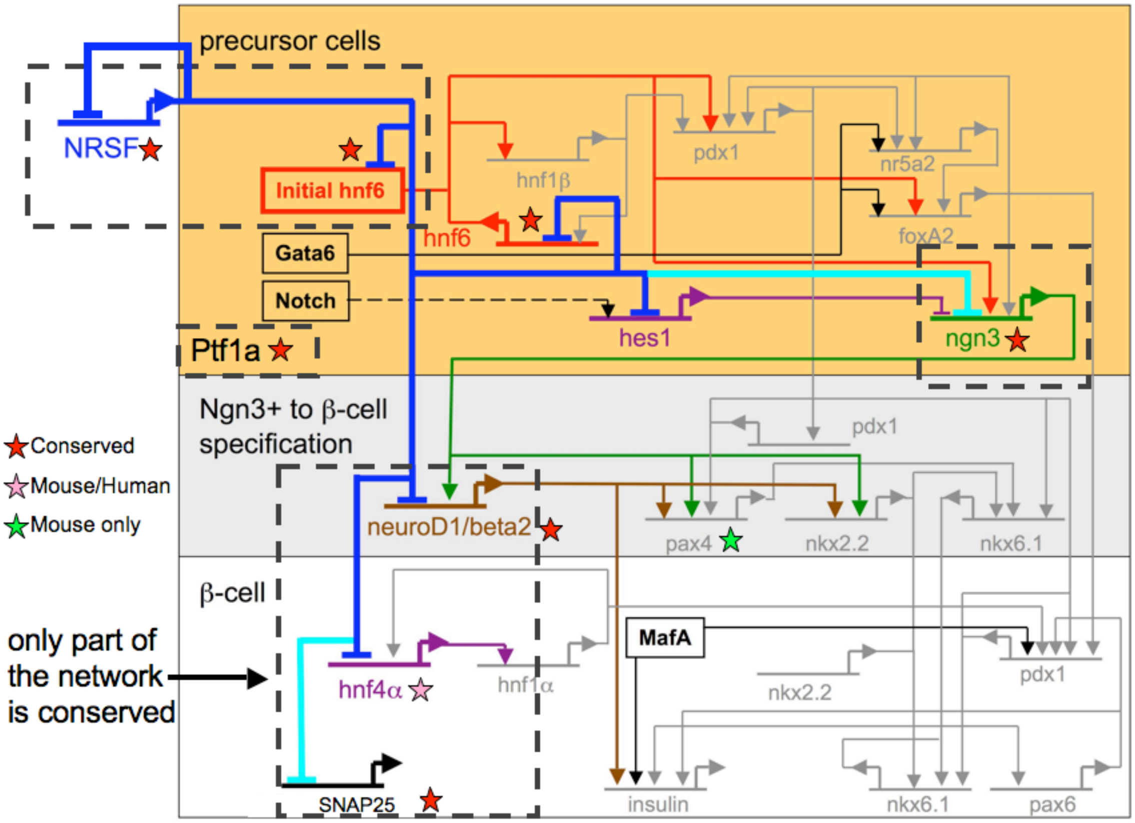


Figure 3.12. A role for NRSF in pancreatic islet specification. Current ChIP-seq-based implied scaffolding of NRSF regulatory interactions on top of gene regulatory network of pancreatic islet β-cell specification (adapted from Davidson, 2006) shows that NRSF represses several key transcription factors that sit on top of the GRN hierarchy, including the key gatekeepers ngn3 and NeuroD1, in addition to its already known terminal differentiation gene battery targets such as SNAP25. These targets represent some of the main drivers of β-cells' neuroendocrine-like behavior. However, only part of the NRSF sub-network is conserved in all three genomes, which is most likely the set of genes most directly related to turning on the neurosecretory network. Note that NRSF is also found bound to one of its own promoters, presumably in order to autoregulate itself.

and a subset of target genes have suboptimal sequences and numbers of sites (Gaudet, 2002). In that system binding suboptimality is believed to help program the temporal order of action during development, with poor binders turning on at later times in the developmental progression, when Pha4 levels are highest. By analogy, the regulators that govern the pancreatic network may be released from NRSF repression relatively early in downregulation of the repressor to create a permissive state that must be established before launching the neuroendocrine development program. Also following this logic, the critical neurosecrotory gene SNAP25 is a classic NRSF target that is expressed later in development in differentiated islet cells, and it displayed relatively higher ChIP-seq tag scores than most of the transcription factors that are positioned higher and earlier in the network, thanks to its double-site. Independent evidence suggests that SNAP25 expression depends on relief from NRSF-mediated repression in islet cells (Martin, 2008). Targets of the regulatory class highlighted here (Fig 3.12) can also participate in positive autoregulatory and cross-regulatory interactions that we expect would stabilize and push forward the circuit once it begins (Davidson, 2006). This makes a "protective" repressor, active in nonpancreatic cell types or progenitor cells, an attractive piece of regulatory logic.

Transcription factor binding site evolution

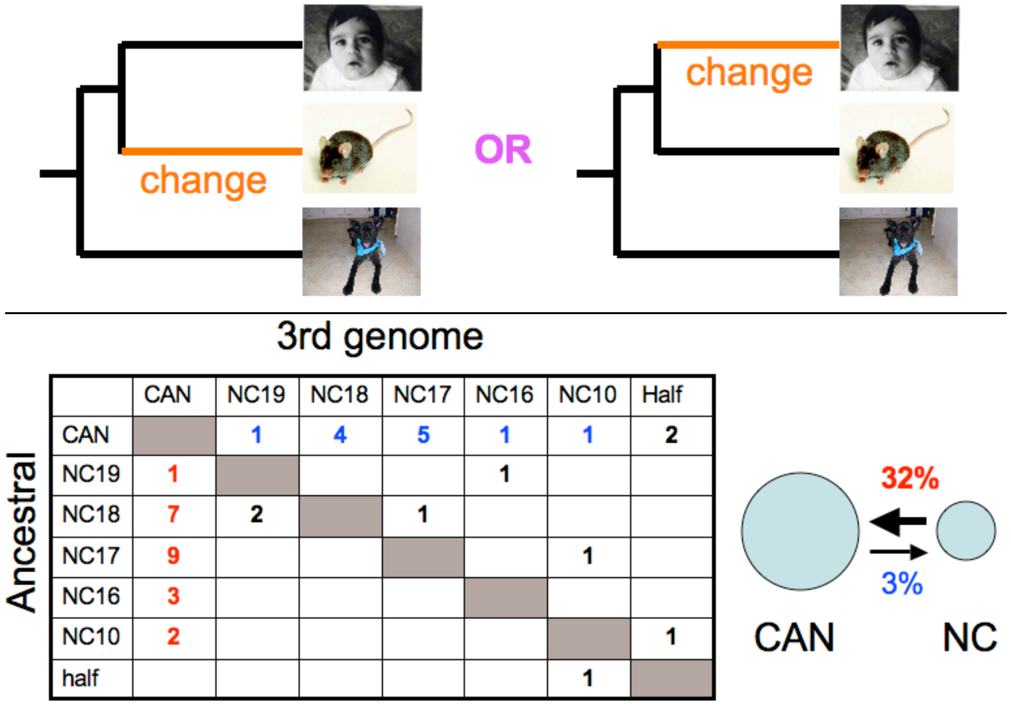
Our gene level analysis showed that only about half of the NRSF gene cohorts in any one genome were present in the conserved core of the network, suggesting either a large turnover of sites or sites that are only bound in some conditions. We further wanted to find out whether there was any conversion between different types of NRSEs. Because the half-sites are much shorter than the full 21 bp NRSE motif, they also occur widely over the genome, presumably mainly by chance. This would mean that there is a rich pool of possible binding sites from which higher affinity canonical sites could be gradually made and tested in evolution, as suggested previously (Zhang, 2006). However, these sites were considered unlikely to interact with NRSF specifically (Zhang, 2006), whereas the non-canonical motif family we define here show binding on their own, especially when clustered. We mapped human jurkats onto mouse EL4 binding sites and asked what fraction of these sites mapped onto sites in these two lymphoblastoid cell lines.

We found that only 677 of the 1774 sites (38%) that were mappable between the two genomes actually overlapped between the two genomes (Fig 3.13). These numbers are similar when comparing jurkats to C2C12 and MDCK (data not shown), which suggest that this is not related to cell-type specificity. We then asked whether we could determine the state of the ancestral NRSE and see whether there is a preferred directionality of changes between the different classes by asking for sites that change classes in human when compared to mouse/dog and in mouse when compared to human/dog. We find a small, but definite preference for converting a non-canonical site into a canonical site rather than the other way around (Fig 3.14); this is in fact more pronounced, given that there is a larger starting pool of canonicals than non-canonical ones in our set of sites. We see almost no conversion of half sites into any of the other classes, but do see a couple of canonical sites decay into half sites.

Mouse EL4 sites

Human Jurkat sites	CAN	NC19	NC18	NC17	NC16	NC10	Half	Not Enriched	Not Mappable
CAN	532	1	2	6	3	3	6	702	487
NC19	1	5	6	0	2	0	0	28	13
NC18	6	0	18	1	0	0	2	30	12
NC17	8	0	0	26	3	0	4	55	26
NC16	4	0	0	1	13	1	0	29	11
NC10	2	0	0	0	0	7	1	28	14
half	4	0	0	0	0	0	10	225	114
Total	557	6	26	34	20	11	23	1097	677

Figure 3.13. Confusion matrix for human Jurkat NRSEs (rows) mapped onto mouse EL4 NRSEs (columns) using UCSC's liftOver annotation between genome builds, which are based on global alignments. Of the three quarters of the sites that could be mapped onto the mouse genome, 38% were also enriched in our ChIP-seq experiment in mouse. 82% of these sites were of the same class as in human (red).



3rd genome

Ancestral	CAN	NC19	NC18	NC17	NC16	NC10	Half
CAN		1	4	5	1	1	2
NC19	1				1		
NC18	7	2		1			
NC17	9					1	
NC16	3						
NC10	2						1
half						1	

Figure 3.14. Motif evolution for NRSF. Comparing sites that change class in one genome (mouse or human) but stay the same in dog and the other genome were pooled and show that there is a preference for going from a non-canonical site to a canonial one (red), rather than the other way around (blue).

Discussion

From a technical perspective, the unparalled resolution of ChIP-seq in identifying the global map of NRSF binding sites in mammalian cell lines compared to previous technologies such as tiling arrays highlights the benefits of global assays. While non-canonical sites were also enriched in our Agilent and Affymetrix tiling arrays (data not shown), we could not locate or derive a motif to account for their reproducible binding without the 21 bp NRSE. The combination of ChIP-seq with motif finding delivered the unexpected surprise of the non-canonical NRSEs, which have been independently confirmed and also tested for function (Otto, 2007; Patel, 2007). As a well-documented test case of a repressor that was already known to be bound to many of its functional sites in multiple cell types, it was relatively straightforward for us to develop a strategy to identify the likely bound sites. However, our own data in this chapter, as well as ChIP-seq experiments for other factors that have been done since our original science paper, show that there can be a wide variability in IP efficiency and that using the same threshold can cause us to miss valid sites if using a single threshold. Although lowering our threshold requirements will recover additional regions, it would be difficult to know off-hand what the adequate threshold should be without the evidence from the monoclonal data. We could also justify a lower threshold by simply sequencing the upstate and control libraries deeper such that even the lower threshold and fold enrichment become unlikely to be seen by chance alone. For factors and antibodies where we don’t have existing calibration data, current rule of thumb for ChIP-seq sequencing used by the ENCODE community is to sequence a library to 10 million mappable reads, assuming that library complexity (the number of unique molecules actually captured in the library) is not a limiting factor.

What is the significance of the non-canonical NRSEs and of the half-sites? Our original Jurkat, as well as our mouse and dog data, make it clear that the canonical sites form the bulk of the sites associated with the neuronal targets that are typical of NRSF and are also the ones most likely to be conserved. While the gene ontology profile of the non-canonical genes is similar to that of the canonical set, our motif evolution analysis implies that they are more likely to become canonical than the other way around. Meanwhile, the half-sites remain something of a mystery, as they not only do not fit the gene ontology profile, but seem to change dramatically even between cell types. These half sites might only be bound opportunistically (or maybe more appropriately, incidentally) when the chromatin state is permissive, as opposed to the full sites (canonical or non-canonical) which more likely control/dictate chromatin state by recruiting its co-repressors.

NRSF has long been thought to control the neurosecretory phenotype (D'Allessandro, 2008; Pance, 2006; Bruce, 2006) and our results tie beautifully with the neurosecretory phenotype of the islet cells. However, the most enlightening pancreatic islet gene is not even in Fig 3.12 and has been known since the very first survey done over a decade ago by Schoenherr et al. (1996). Somatostatin has two closely spaced NRSE in its intron. While somatosatin-producing δ-cells are a minor component of pancreatic islets, somatostatin is highly expressed in the brain and several other endocrine cell types through the rest of the body such as the gastric D-cells. It is probable that any cell expressing somatostatin has shut off NRSF. It also suggests that the δ-cells may have been the ancestral cell types from which the other islet cells were derived.

The direct cis-regulatory connection of NeuroD1 and NRSF provides an exciting connection between the most famous repressor of neurogenesis and the best known neurogenic factor. It is worth noting that other members of the NeuroD family are also NRSE associated, although not in their coding sequence. This brings the question of whether the NeuroD family became neurogenic because of NRSF or whether the fact that these factors were neurogenic permitted the rise of NRSEs. Together with NRSF's regulation of neuronal microRNAs such as miR-124 and of neuronal splicing factors, this connection with NeuroD brings NRSF back into the role of an organizing regulation of neurogenesis. The fact that NRSF simply doesn't seem to exist in non-vertebrates makes this fundamental role in neurogenesis even more remarkable. It will be fascinating to understand how NRSF arose and acquired such a critical role.

Methods

Cell culture and chromatin immunoprecipitation

The Jurkat human T lymphoblast cell line was cultured according to standard protocols. NRSF/REST chromatin immunoprecipitation was performed as previously (Mortazavi, 2006) with a custom monoclonal antibody (Chen, 1998). For experiment 1, 70 separate chromatin immunoprecipitations, corresponding to a single batch of chromatin, were pooled for a single Solexa library preparation. For experiment 2, four chromatin immunoprecipitations, corresponding to one batch of chromatin, were pooled for a single Solexa library preparation. For both experiments, the controls used chromatin that was reverse crosslinked, phenol extracted, and purified on a QIAQuick PCR cleanup column (Qiagen). The control chromatin matched the chromatin preps of the ChIPs used for each experiment. EL4, C2C12, and MDCK experiments followed the experiment 2 protocol, except that a single IP was used for C2C12 and MDCK each.

Library preparation for Solexa

The Solexa library was prepared as per Ilumna's instructions. The size selection of this library was performed by gel electrophoresis and subsequent excision and purification of DNA (QIAex II, Qiagen) in the ~ 200–700 bp range. The control library for experiment 1 was constructed in an identical manner, using ~ 2 μg input DNA. For experiment 2, we modified the Solexa library construction protocol to include a PCR preamplification (30 sec at 98^{o}C; [10 sec at 98^{o}C, 30 sec at 65^{o}C, 30 sec at 72^{o}C] x 25 cycles; 5 min at 72^{o}C) following linker ligation and preceding gel electrophoresis. Size selection was performed by gel electrophoresis and subsequent excision and purification of DNA in the ~ 150–300 bp range. The control library for Experiment 2 was prepared in an identical manner, using ~ 50 ng input control DNA. We used QPCR (Mortazavi, 2006) to estimate the enrichment of five loci in these two libraries. QPCR loci are (in genome build hg17, NCBI v35): chr1:151353339-15153415 (QPCR1), chr1:158498975-158499043 (QPCR2), chr16:88520383-88520482 (QPCR3), chr17:3247959-3248037 (QPCR4), and chr2:165920458-165920534 (QPCR5). To calculate fold enrichment, each primer pair was normalized against two putative "negative" primer pairs, chr7:115817618-115817717 (NEG1), and chr7:115712789-115712882 (NEG2).

Enriched Region Identification

Solexa ChIP and control reads were analyzed jointly for each experiment, to identify regions that have an over-representation of reads in the ChIP sample versus the control sample, using a set of python scripts (available at http://woldlab.caltech.edu/ChIPSeq), which are now part of dual use ERANGE (Chapter 4). Candidate enriched regions were identified as aggregations of 8+ RPM not separated by more than 75 bp and were assigned the normalized number of reads as a score. The threshold of 8 RPM was selected on the basis of the ROC analysis described in the text and in Fig 3.3; this threshold will need to be selected in future studies based on the structure of each data set, and with consideration of the false discovery rate that will be tolerated in a given study. Regions (a) with at least 20% or more control reads within the same boundaries (these regions corresponded typically either to satellite repeats or to the mitochondrial genome), (b) with peaks having less than five partly overlapping reads, or (c) with more than 90% of the reads in a single direction were filtered out and did not participate in subsequent analyses.

Site Analysis

We performed all motif-oriented analyses by using Cistematic (Mortazavi, 2006). The NRSE2 PSFM (position specific frequency matrix) derived in that work was used to identify and local canonical NRSE sites (match score thresholds specified in the text and figure legends), across human genome hg18 (NCBI v36). These site locations were then used to compare and call the distances from peaks of ChIPSeq read-tag distributions at each location. The analysis of NRSE2 locations relative to called peak site locations was done on the shared set of NRSF ChIPSeq positive regions. Analyses in mouse and dog were done against mouse mm9 and dog cf2 genomic builds, respectively.

We merged enriched shared regions within 500 bp of one another and combined reads within these regions. We then applied a triangular 5-point smooth to reads within these consolidated regions to identify the coordinate(s) with the greatest number of overlapping reads as the peak(s). If there was more than one coordinate with the maximum score in the region, we selected the first one as the peak.

Motif Searches by Meme and NRSE half-sites

Enriched shared regions with > 500 reads (a subset that was selected to keep compute resources modest but to focus on quantitatively robust signals) were analyzed with MEME (Bailey, 1995) from within Cistematic using the zoops model for 10 motifs of 8–28 bp in length. We repeated the same analysis with enriched shared regions with 300 or more reads but no 70% NRSE2 match within them to identify non-canonical motifs.

Following the results of the MEME analysis of the strongly enriched non-canonical regions, we divided NRSE2 into two half sites, i.e., NRSE2-left (position 1 through 10) and NRSE2-right (position 12–21). Position 11 serves as motif-center, allowing us to compare distances between motifs in an orientation-independent manner. We also defined the distance between positions 5 of NRSE2-left and NRSE2-right (corresponding to position 16 of NRSE2). This gives a canonical distance of 11 bp between the two half sites within NRSE2 and allowed us to consider shorter as well as longer distances. We tabulated the occurrences in experiment 1 enriched regions of half-sites with distances of 1 up to 25 bp in various orientations, in addition to the canonical orientation (NRSE2-left followed by NRSE2-right) and compare their observed occurrences to their expected occurrences based on their genome-wide occurrences adjusted for the fractional size of the enriched regions (0.06% of the genome, for experiment 1). We also analyzed other motif arrangements, such as right-left (opposite of canonical), left-left, and right-right.

P-value estimation for NRSEs

P-values for sites were estimated by counting the number of 25 nt sequence reads in a 400 bp window centered on the motif and comparing them to the observed frequency of that window count in the control. The P-values presented are from an extremely conservative calculation, in which the entire control dataset was used, unfiltered for sequence read pileups in repeat DNA sequences. We know the latter to be artifactual because they also occur at the same location in the ChIP experiments. Using instead a control from which these reads have been filtered eliminates any and all 400 bp windows with more than 11 reads. This results in effective P-values of 0 for all ChIP-seq positive sites.

Final Site Analysis

Based on the enrichment of particular distances in the canonical arrangement, we used the following procedure to identify NRSEs in each region of the common enriched region set:

1. Analyze each region with NRSE2left and NRSE2right PSFMs using a threshold of 70%.

2. Accept any canonical half-sites with distances of 10, 11, 16, 17, 18, 19.

3. Accept any 70% or higher canonical NRSE2 site that has not already been picked up by 2.

4. If there is still no site assigned, pick the nearest half site to the peak of the region.

5. Filter sites to only retain those with P-values lower than 10^{-4}.

Associated Gene and Gene Ontology Analysis

We used Cistematic to identify the nearest NCBI gene model in the human, mouse, and dog genomes within 20 kb of each enriched regions in the shared set of ChIP-seq positive regions. We then analyzed this gene cohort for Gene Ontology enrichment using Cistematic, as previously described (Mortazavi, 2006). Briefly, we tabulated the count of each GO term in our gene cohort, calculated a P-value for that occurrence by chance in a gene cohort of that size, and applied a Bonferroni correction for multiple hypotheses testing.

Cistematic was also used for the orthology matching between the genes in the multiple genomes.

Acknowledgements

We thank Dr. Gary Schroth and his group at Solexa/Illumina, Inc., for access to the sequencing platform, without which this work would not have been possible. We thank Betsy Anton, Loan Nguyen, Cathy Medina, and Dr. Larisa Tsavaler for outstanding experimental work and Dr. Kenneth F. McCue for invaluable counsel on statistical analyses. A.M. also thanks Lorian Schaeffer for her assitance in building and running the Solexa 1G sequencing platform for the mouse and dog libraries. We are grateful to Prof. David Anderson of Caltech for the gift of monoclonal antibody against NRSF. This work was supported by NIH Grant U01 HG003162 to R.M.M. with a supplement to B.W. and by a grant from the Caltech Beckman Institute. A.M. was supported by NIH/NRSA 5T32GM07616; B.W. by a Bren Chair endowment at Caltech, and R.M.M. by the Stanford W. Ascherman Chair endowment at Stanford University.

Chapter 4

QUANTITATION OF MOUSE RNA BY RNA-SEQ AND RNA POLYMERASE II CTD PHOSPHOISOFORMS BY CHIP-SEQ REVEAL A NOVEL FORM OF PROMOTER STALLING

Abstract

We map and quantify mammalian transcriptomes by sequencing them deeply and recording how frequently each gene is represented in the sequence sample (RNA-seq). This application of ultra-high-throughput DNA sequencing provides a digital measure of the presence and prevalence of transcripts from both known and novel genes. Reference measurements consisting of 20 to 40 million mapped 25 bp reads are reported for polyA selected RNA from three adult mouse tissues and the C2C12 muscle cell line. Exogenous RNA standards established linear quantification spanning five orders of magnitude in prevalence that allow for the calculation of transcripts/cell for all genes. RNA splice events were detected for thousands of genes in each transcriptome by direct mapping of splice-crossing sequence reads. An independent and complementary genome-wide transcriptome map was made by measuring RNA polymerase II occupancy across the genome C2C12s, using phosphoisoform specific antibodies that mark different Pol II activity states. We identify two distinct stalled pol II signatures at a subset promoters with no detectable RNA-seq expression based on the level of Ser 2 phosphorylation. We finally show the change in these polymerase stalling states at genes upon myogenic differentiation of C2C12 cells correlates with specific biological processes.

Introduction

Transcriptome analysis has become an important general phenotyping method, with microarrays of several kinds now in routine use. However some significant limitations remain for expression array methods, including hybridization and cross-hybridization artifacts, dye-based detection issues, and design constraints that preclude or limit detection of RNA splice patterns and previously unmapped genes. A different approach to large scale RNA analysis has been SAGE (serial amplification of gene expression) and related methods that use DNA sequencing of previously cloned 17–25 bp long tags from the transcripts' 3' (or 5') ends (Harbers, 2005). These sequence tags are then identified by informatic mapping to mRNA reference databases or, for longer tag lengths, to the source genome. A strength of SAGE methods is that they produce digital counts of transcript abundance in contrast to the analog style signals from fluorescent dye-based microarrays. However, SAGE family methods rarely provide information about splice isoforms, promoter choice, or new gene discovery. Past SAGE studies required bacterial cloning intermediates and have generally been limited to < 200,000 tags, so that fully comprehensive measurements across all abundance classes have been beyond reach. Recently, very dense whole-genome tiling microarrays have been developed and applied to transcriptome mapping (Kapranov et al., 2007). In contrast to SAGE or conventional expression arrays, they have the capacity to discover new exons for known genes or entirely novel genes, and many new RNAs of different kinds have been reported. But the amount of input RNA needed for these multi-slide tiling formats is high, and the other limitations of microarray sensitivity and specificity remain.

RNA-seq

A simpler and potentially more comprehensive way to measure transcriptome output is to deeply sequence cDNA made from source RNA by ultra-high-throughput sequencing (Wold, 2008). Resulting DNA sequence reads are then informatically mapped back to the source genome and counted so that the number and density of reads corresponding to any gene, exon, splice event, or previously unknown candidate gene can be calculated and compared with any other sequenced sample (Fig 4.1A). If enough reads (> 20-40 M) are collected from a sample, it should be possible to unambiguously detect and quantify RNAs from all biologically relevant abundance classes, to map RNA splice choices in transcripts of moderate and high abundance classes, to discover previously unknown RNAs or exons, and to detect at least some characteristic transcripts from minority cell populations in mixed tissues such as brain.

We tested these expectations by performing RNA-seq on polyA selected RNA from mouse liver, muscle, and brain tissues, and from the myogenic C2C12 cell line, using the Solexa 1G sequencing system. Among ultra-high-throughput sequencing platforms in wide use in 2008, it maximizes total read number produced per machine run and per dollar. High read number is critical for this assay because it optimizes the number of independent pieces of evidence (sequence reads) for the presence and abundance of transcripts from all genes. In preliminary experiments we found that controlled hydrolysis of RNA samples prior to cDNA synthesis steps significantly improved uniformity of sequence coverage across transcripts (data not shown), which translates into greater sensitivity of detection, accuracy of quantification, and the completeness of splice and exon maps. Randomly primed cDNA

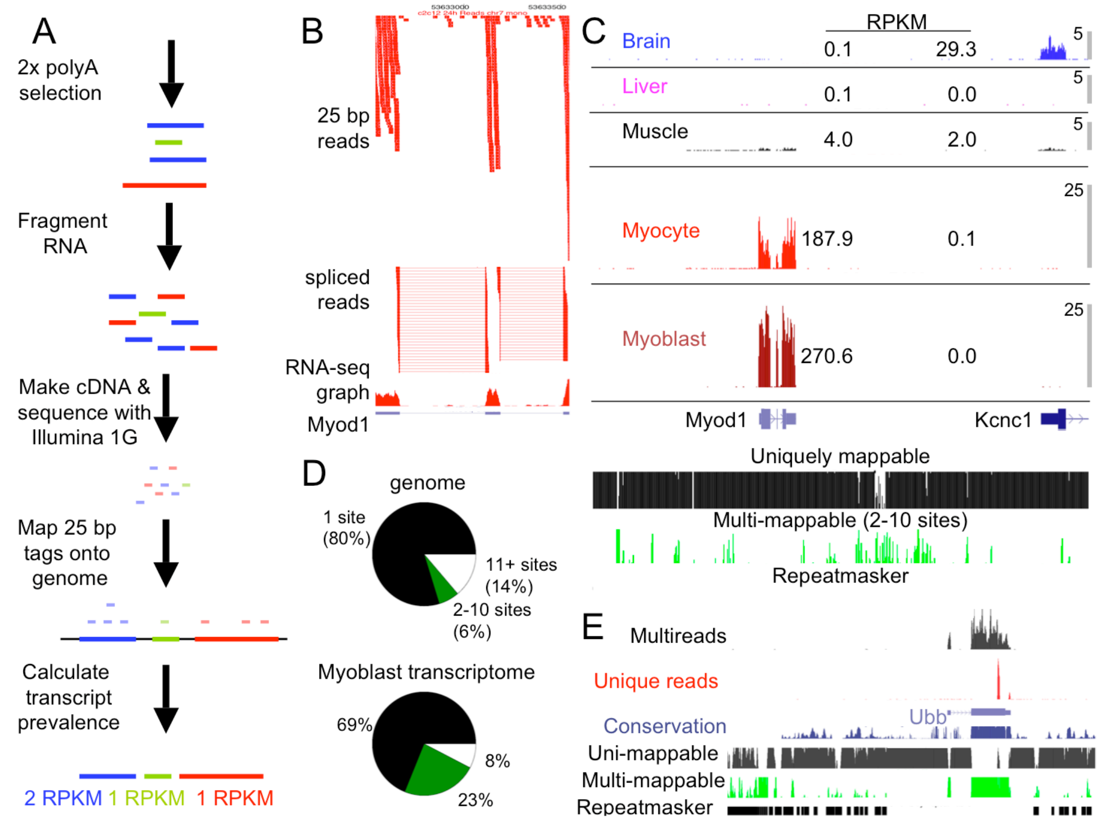


Figure 4.1. RNA-Seq outline. A Following 2 rounds of polyA selection, RNA is fragmented to an average length of 200 bp, which is then converted into cDNA via random priming. The resulting cDNA is then converted into a library for Solexa 1G sequencing. The resulting 25 bp reads are then mapped onto the genome. Normalized transcript prevalence (in RPKM — Reads Per Kb per Million reads) is then calculated with an ERANGE package algorithm. **B.** Primary data from C2C12 myocyte RNA that map uniquely to a 1 kb region of the MyoD locus, including reads that span to introns. The RNASeq graph above the gene model summarizes the quantity of reads, where each point represents the number of reads covering each nucleotide per million reads (normalized scale of 0 to 4.5). **C.** Detecting and quantifying differential expression. Mouse polyA-selected RNAs from brain, liver, skeletal muscle, and C2C12 myocytes and myoblasts are shown for a 35 kb region of chr7 containing Myod, which is muscle specific, and its neighboring gene Kcnc1, which is adult muscle and brain-specific. **D.** Multireads form a significantly larger fraction of the sequenced transcriptome than their fraction of the genome. **E.** A 10 kb region of the ubiquitin Ubb locus is shown. An ERANGE algorithm allocates multireads (2–10 occurances in the genome) using a weighting function based on the density of uniquely-mapping reads at each paralog (shown in red).

was made from 100 ng of 200–300 nt hydrolyzed polyA RNA per sample, and a Solexa molecular library was then constructed according to the manfacturer's standard protocol (Fig 4.1A). 10–30 million mappable reads were obtained from each library, with two or more independent libraries made for each tissue.

Enhanced Read Analysis of Gene Expression (ERANGE)

To analyze these data, we developed ERANGE package (Enhanced Read Analysis of Gene Expression, Fig 4.2) in order to 1) assign reads that map uniquely in the genome to their site of origin (Fig 4.1B, 4.1D), or, for reads that match equally well to several sites in the genome ("multireads"), allocate them to their most likely site(s) of origin (Fig 4.1D, 4.1E); 2) detect splice-crossing reads and assign them to their gene of origin (Fig 4.1B); 3) organize reads that cluster together, but do not map to an already known exon, into candidate exons or parts of exons; and 4) calculate the prevalence of transcripts from each known or newly proposed RNA, based on sums of weighted unique reads, spliced reads, and multireads that mapped onto exons (Fig 4.2; Materials and Methods). The new candidate RNA exons can be thought of as ESTs and, like ESTs, some are provisionally appended to existing gene models, if they meet several additional criteria (Materials and Methods). Remaining unassigned candidate transcribed regions (labeled RNAFAR) can then be used, in conjunction with polymerase location data presented below, to support new or revised gene annotations.

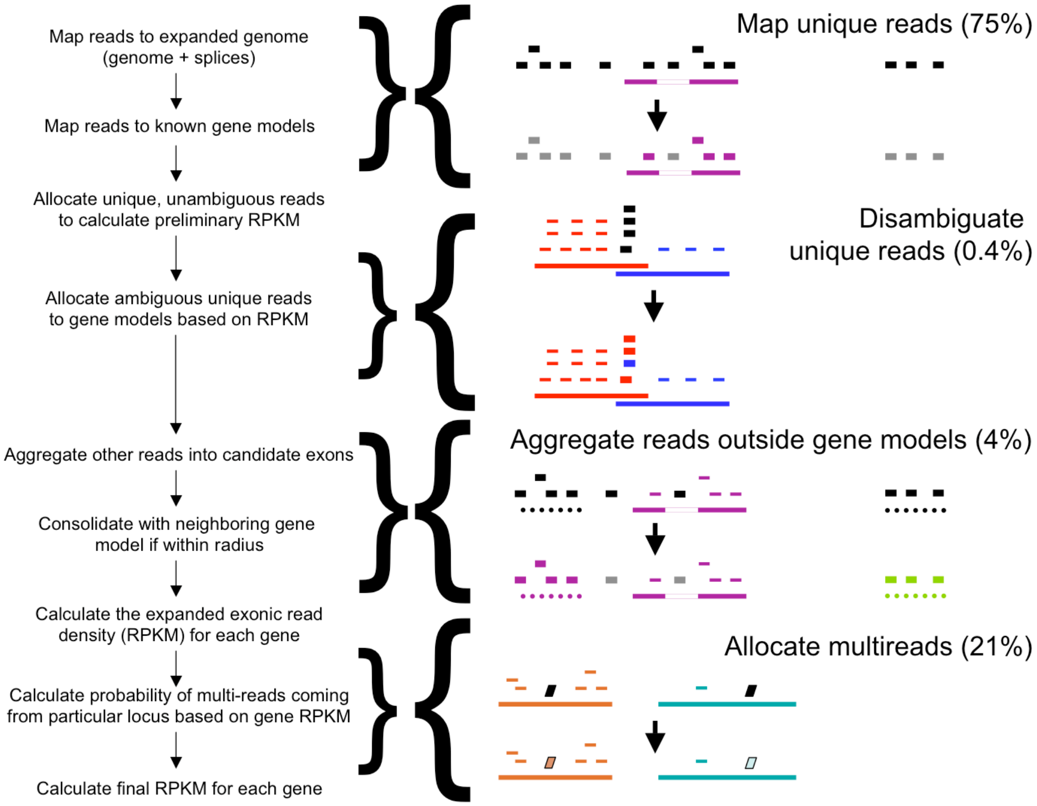

Map reads to expanded genome (genome + splices)
Map reads to known gene models
Allocate unique, unambiguous reads to calculate preliminary RPKM
Allocate ambiguous unique reads to gene models based on RPKM
Aggregate other reads into candidate exons
Consolidate with neighboring gene model if within radius
Calculate the expanded exonic read density (RPKM) for each gene
Calculate probability of multi-reads coming from particular locus based on gene RPKM
Calculate final RPKM for each gene
Map unique reads (75%)
Disambiguate unique reads (0.4%)
Aggregate reads outside gene models (4%)
Allocate multireads (21%)

Figure 4.2. Enhanced Read Analysis of Gene Expression (ERANGE). The main steps of the computational pipeline are outlined to the left, with different aspects of read assignment and weighing diagrammed to the right, with the corresponding number of gene model reads in muscle shown in parenthesis. Reads under consideration are shown as black rectangle with their assignment to gene models shown in color, while reads falling outside of known or predicted (RNAFAR) exons are shown in grey. RNAFAR regions are shown as dotted lines and can either be assigned to neighboring gene models if close enough (purple) or assigned their own predicted model otherwise (green). Multireads (shown as trapezoids) are assigned fractionally to their different possible locations based on the expression levels of their respective gene models.

Primary data from a 21 million-read transcriptome measurement of C2 myogenic cells (Fig 4.1B,C) show key characteristics. Prior studies (reviewed in Berkes, 2005) have established that the MyoD gene is expressed at a modest level in myoblast, at a lower level in adult muscle tissue, and is silent in liver and brain. RNA-seq data matched these expectations, with 5,002 25 bp reads mapping uniquely to MyoD exons out of 21 M reads from a C2 myocyte RNA-seq library, but only 1 and 2 reads from individual liver and brain RNA-seq libraries, respectively. Reads mapping to exons are prominent, while reads in MyoD introns are at much lower density. Another 255 reads that did not map to the mouse genome mapped uniquely across MyoD's splice junctions, while 62 reads that can map from 2 to 10 places in the genome ("multireads") fell on MyoD exons. However multireads are present at much higher percentages in our sequenced transcriptome than in the genome as a whole (Fig 4.1D) and can dominate over uniquely mappable reads for paralogous genes or genes with highly conserved domains, such as Ubiquitin B (Fig 4.1E), which has 1,946 unique reads, but also 65,466 multireads. Across the transcriptome, the read density in introns for all expressed genes was < 1% of levels in exons, as expected for

Sample	*Uniques*	*Splices*	*Multi (2-10)*	*Multi (11+)*	*Spikes*	*No Match*	*Used*	*Total*
Brain 1	50.3%	3.1%	9.7%	1.7%	0	34.7%	63.5%	28.8M
Brain 2	54.3%	3.3%	10.6%	1.6%	0.1%	25%	73.4%	48.8M
Liver 1	44.3%	3.4%	12.8%	2.3%	0	37.2%	60.5%	29.6M
Liver 2	42.7%	3.4%	13%	2.4%	1%	37.6%	60%	41.5M
Muscle 1	46.2%	4%	14%	1.6%	0	34.2%	64.1%	30.1M
Muscle 2	44.6%	3.5%	15.3%	1.6%	1.6%	33.3%	65.1%	37.2M
C2C12 Undiff 1	51.5%	3.8%	18.7%	6.1%	0	19.9%	74.0%	22.8M
C2C12 Undiff 2	49.6%	3.6%	18.5%	7.8%	2.2%	18.3%	73.9%	16.4M
C2C12 24h R	36.1%	2.6%	16.8%	2.8%	0	41.8%	55.4%	35.2M
C2C12 24h T 1	34%	2%	16.2%	2.8%	0	44.9%	52.3%	39.4M
C2C12 24h T 2	44.8%	3%	20.5%	9.3%	1%	21.4%	69.3%	10.9M

Table 4.1. RNA-seq summary read statistics

partially processed precursors present in whole-cell polyA RNA. Data for technical and biological replicate determinations showed high reproducibility, with R^2 of 0.96 (Fig 4.3A) and 0.95 respectively. Primary sequence read data for the mouse brain, muscle and liver replicates are available at http://woldlab.caltech.edu/RNA-Seq, and submitted to the Short Read Archive (SRA10030). The summary data about the datasets are in Table 4.1.

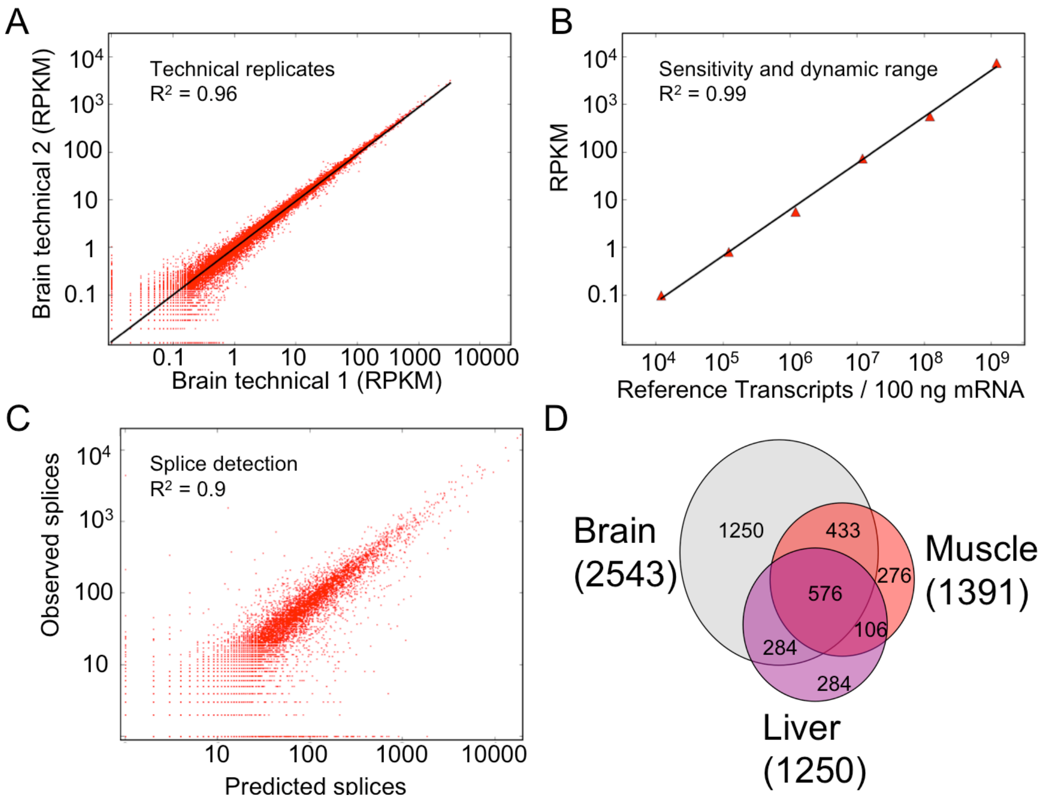


Figure 4.3. Reproducibility, linearity, sensitivity, and splice detectability. A. Comparison of two brain technical replicate RNASeq determinations for all mouse gene models (UCSC), measured in RPKM (reads per kilobase of exon per million mapped sequence reads) which is a normalized measure of exonic read density (R^2 = 0.96). **B.** 6 reference transcripts of lengths 0.3 to 10 kb were added to the liver RNA sample (1.2x10^4 to 1.2x10^9 transcripts per sample) (R^2 > 0.99). **C.** The number of expected spliced reads for each gene model was predicted based on the number of introns and the exonic read density is compared with the observed number of splices (R^2 = 0.90). **D.** Genes with two splice isoforms in the same tissue.

Mapping spliced reads and alternative splicing

Splice-crossing reads, as shown for MyoD (Fig 4.1B), were identified by mapping otherwise unassigned sequence reads across all known splice events in all UCSC mm9

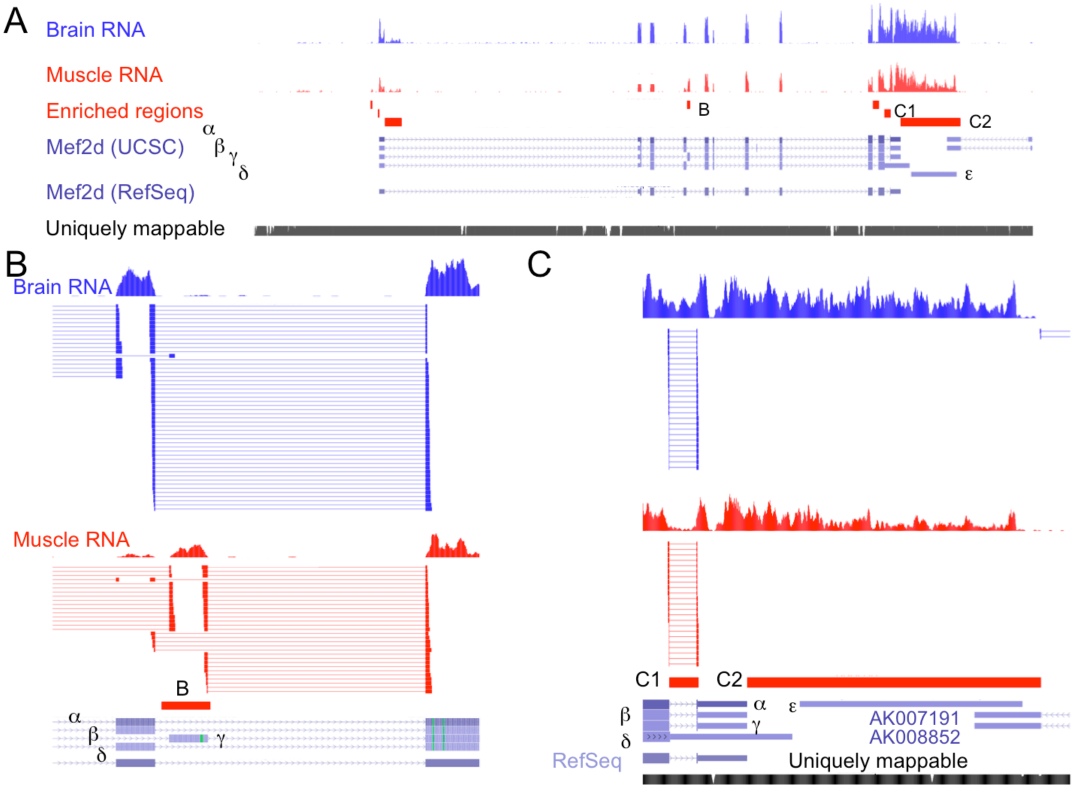


Figure 4.4. Alternative splicing in Mef2D. A. 40 kb region encompassing mef2D, which is expressed in the myoblast and adult muscle (28 RPKM in muscle and 45 RPKM in brain) and its neighboring gene, which is expressed at a much lower level in brain. Whereas RefSeq has only a single annotation for Mef2d, UCSC has five (labeled α-ε).Whereas α corresponds to the RefSeq model, γ is a muscle-specific isoform (Martin, 1994). Our RNAFAR algorithm identified 7 regions enriched with reads that fell outside of the NCBI gene annotations (red) that were assigned by the algorithm to the Mef2d locus. **B.** 1.5 kb close-up of muscle specific alternative splicing at RNAFAR region "B". The prevalence of splicing switches from the canonical exon to the RNAFAR exon upon muscle differentiation, as seen both in the ratio of spliced reads and in the number of reads falling on the two diagnostic exons. **C.** 5 kb closeup of 3' end of Mef2d. RNAFAR region C1 has a greater density of reads in the brain than in adult muscle. Both the brain and muscle datasets show reads across region C2, which is consistent with ε being the 3'UTR for mef2d in muscle. Spliced reads also show that the partially overlapping transcript AK007191/AK00852 is present in brain.

gene model splices. Over our entire dataset, splice-spanning reads comprised ~ 3% of all mapped reads (Table 4.1), which is consistent with expected splice frequency from gene models across the genome (see Methods). To assess the efficiency of splice detection, we computationally predicted all reads expected to cross known splices in a transcriptome, by considering all gene models and their respective levels of expression based on exon reads. Observed splice-crossing reads were in good agreement with predictions (Fig 4.3C). We conclude that splice events can be readily detected and quantified for abundant and moderately abundant transcripts. Alternative splice isoforms were detectable in proportion to their relative prevalence within and across tissues. For example, Mef2D (Fig 4.4) has a muscle-specific exon that is only prominent in adult muscle (Fig 4.4B), whereas Mef2D in the brain has a bleeding exon that does not go through a canonical splice (Fig 4.4C). In both cases, these features were not part of the NCBI or RefSeq annotations for that gene model and were thus picked up by our RNAFAR algorithm. However, our calculations of RNA prevalence were at the locus level and we explicitly did not attempt to quantitate transcript isoform prevalence, which would be greatly helped by the use of paired reads given that we typically detect more than 1 splice isoform for an alternatively spliced gene within the same tissue (Fig 4.3D) Detection of splice events is primarily a function of transcript prevalence (Fig 4.3C).

Transcript quantitation

To assess the dynamic range, linearity, sensitivity, and sequence coverage, we introduced into each experimental RNA sample a set of known RNA standards that were transcribed in vitro from Arabidopsis and phage lambda templates (Fig 4.3B). The RNA standards were designed to test for possible effects of sequence length and sequence composition on the observed transcript abundance, and were added to the sample RNA at concentrations spanning the full range of abundances observed in natural transcriptomes. RNA-seq data for the standards were linear across a dynamic range of five orders of magnitude in RNA concentration (R^2 = .99) (Fig 4.3B).

The RNA standards showed that the sensitivity of RNA-seq is, as expected, a function of molar concentration and transcript length. We therefore quantified transcript levels in RPKM units (Reads Per Kilobase of exon model per Million mapped reads, Fig 4.1A), where reads counted include weighted unique, splice, and assigned multireads (Fig 4.2). This measure of read density transparently reflects the molar concentration of a transcript in the starting sample. The RPKM unit also normalizes for the total read number per measurement, which allows direct comparison of expression levels across multiple transcriptomes. Finally, if used in conjunction with the RNA standard data, it allows calculation of absolute transcript amounts (supplementary materials and methods). In the case of our C2C12 cell lines, we recovered 100 ng of polyA RNA per 2.8 * 10^6 myoblast cells, whereas we needed only 9*10^5 nuclei (to account for any early fusion of myocytes into myotubes) of the 24 hour differentiated myocytes to recover the same amount of polyA RNA. While the RPKM of MyoD goes down from 270.6 RPKM in the myoblast down to 188 RPKM in the myocyte, its abudance actually increases during that period from 35.4 transcripts per cell to 156.1 transcripts per nuclei, assuming uniform recovery of RNA.

At current practical sequencing capacity and cost (~ 20–40M mapped reads from 4–8 lanes of one Solexa 1G flowcell), transcript detection was robust down to 1.1 RPKM, which corresponds to 1 transcript per nuclei in our myocyte samples. The RNA standards also showed that sequence coverage throughout a long transcript is highly reproducible and quite uniform. Whereas uniquely placed 25-mers provide surprisingly good coverage of single copy and many distant paralogs, such as the four transcription factors of the MyoD family, few of these reads fall on genes such as ubiquitin (Fig 4.1E) or Gapdh that have either very recent duplications or pseudogenes in the genome. We found that 76% of the genome is uniquely mappable with 25-mers, a further 6% are accessible when considering reads that could map from 2 to 10 positions on the genome (Fig 4.1D). Given that these mappable multireads contribute between 14 to 30 percent of the reads from RNA-seq depending on the tissue suggests that these multireads are most likely contributed by highly expressed genes. In order to avoid underestimating the expression of genes that would be sensitive to multireads, we calculate the probability that a multiread comes from a particular gene based on the expanded exon read density (unique plus splice-spanning reads) already calculated for that gene, versus the expanded exon read density for the other possible source locations (see materials and methods). Genes so similar that they lack uniquely assigned reads will receive a symmetric distribution of multireads. But for older and more diversified gene families, ERANGE distributes multireads asymmetrically in proportion to the unique and splice reads recorded. The impact on the analysis pipeline from identifying and allocating multireads in this manner was to change RNA quantitation for ~ 32% of genes above 1 transcript / cell in the myocyte transcriptome by more than 30%.

RNA polymerase II CTD phosphoisoform ChIP-seq

RNA polymerase II catalyzes transcription of protein-coding genes and many small RNA precursors. Pol II is recruited to promoter regions by sequence specific transcription factors where, according to current models, the enzyme might pause, then initiate transcription and finally traverse the body of the gene (Core, 2008). Elongation can occur with or without pausing and, and in most genes continues on considerably beyond the 3'polyA processing site before finally disengaging somewhere downstream at a location that is well mapped for only a few metazoan genes (Gromak, 2006, Tantravahi, 1993), although increasing evidence links factors involved in polyadenylation to termination (Buratowski, 2005). Measuring Pol II occupancy across the genome is an independent way to map a transcriptome. This can be done by chromatin immunoprecipitation experiments (ChIP) in which an immune reagent against Pol II is used to retrieve small DNA segments occupied by Pol II in the cell, followed by deep sequencing (Barski, 2007; Schones, 2008) or dense tiling microarray assays (Zeitlinger, 2007; Guenther, 2007), which have shown that a large fraction of genes, including genes with little or no expression, have a distinct polymerase signal at the promoter

Pol II initiation and elongation activities are associated with phosphorylation of specific serine residues in its C-terminal domain (CTD) domain. The CTD contains 52 repeats of a heptad consensus YSPTSPS motif which is highly conserved in eukaryotes and can be phosphorylated at every serine, though not necessarily uniformly on each of the 52 repeats (Chapman et al., 2007; Egloff et al., 2007). The CTD apparently serves as an organizing scaffold for RNA processing enzymes involved in 5' capping, splicing, and polyadenylation, among others (Phatnani and Greenleaf, 2006). These enzymes have

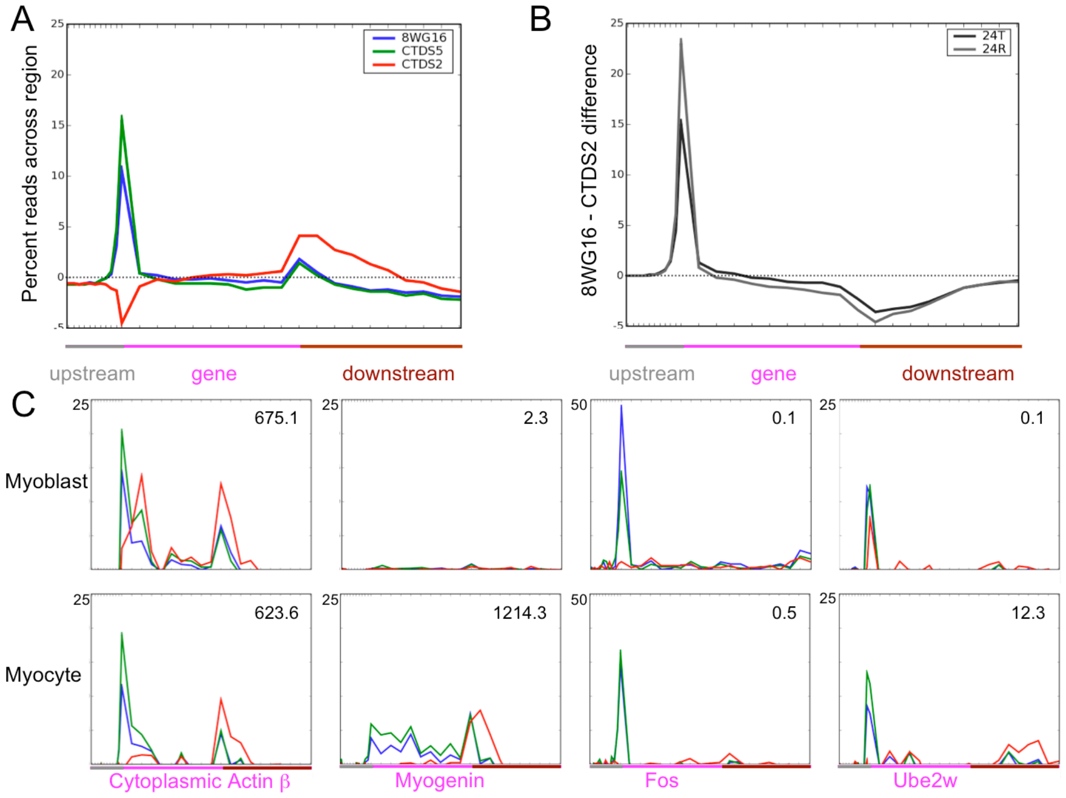


Figure 4.5 CTD. polymerase ChIP-seq defines several classes of polymerase signatures. A. ChIPSeq profile of 3 different antibodies to RNA polymerase II that recognize the unphosphorylated (8WG16) and Pol II C-terminal domain S5 and S2 phosphorylation states for the 500 top genes expressed in c2c12 myocytes 24 hours after differentiation on a normalized gene model of 3 kb upstream in green, a normalized gene locus in red and 10 kb downstream in blue. While 8WG16 and CTD S5 show very similar enrichment patterns at the TSS and throughout the gene locus, The CTD S2 pol II mark rises throughout the gene model and peaks past the polyadenylation site that defines the final, mature mRNA. **B.** The direction and extent of a gene model can be mapped using the difference in the 8WG16 and CTD S2 signals. Genes, and scales are the same as in A. **C.** Pol II phosphorylation marks define 5 classes of genes: highly expressed genes with significant S2 in the vicinity of the TSS (actb in Myoblast), highly expressed genes with little CTD S2 signal in the vicinity of the TSS (Actb and Myog in Myocytes), genes with almost no polymerase signal above background (Myog in myoblasts), unexpressed genes with high 8WG16 and S5 but no S2 signal at the TSS (Fos), and genes with no expression but high S2 signal at the TSS (Ube2w in myoblast), which are lowly expressed when the S2 signal goes down at the TSS but come up at the 3' end of the gene (Ube2w in myocyte). Ube2w is "stalled" in myoblasts in spite of having the S2 phosphorylation, suggesting an additional process regulating productive elongation in at least some stalled genes. Estimated transcripts/cell in upper right corner of every panel.

domains that recognize and dock to CTD repeats with specific phosphorylation patterns (Meinhart et al., 2005), suggesting that there is a CTD phosphorylation code to coordinate Pol II activity with RNA processing machinery as the enzyme traverses the length of the gene (Corden, 2007). Two key indicators of Pol II status are phosphorylation of CTD-Ser5 by TFIIH, which marks initiation, and phosphorylation of Ser2 by PTEFb (also known as CTDK-I), which marks elongation (Phatnani and Greenleaf, 2006).

For ChIP analysis, we used immune reagents that specifically recognize Pol II with accessible Ser2P only or with Ser5P only CTD residues, as well as an immune reagent (8WG16) which recognizes unphosphorylated repeats. Each polymerase molecule can contain unmodified, singly modified (S2P or S5P), or doubly modified repeat motifs within its 52-motif CTD, which means that reactivity with the three immune reagents is expected to reflect a multiple modification code. We performed ChIPSeq on undifferentiated, actively proliferating C2 myoblasts and on actively differentiating C2s in the process of becoming myocytes. Pol II occupancy maps were generated for the three phophoisoforms of the CTD repeat motif. We first asked whether polymerase signal density was predictive of expression, and found it to be so only weakly ($R^2 = 0.37$). We next asked if the polymerase isoforms give distinct ChIP-seq profiles. Fig 4.5A shows that ChIP-seq results for Ser5 and nonphosphorylated (8WG16) forms are indistinguishable from each other, but the occupancy map for Ser2 is distinct in an informative way. The relative amount of Ser2 ChIP-seq signal increases markedly, compared to the other two isoforms, after the polyA addition site. It is common to see Ser2 signals dominate the profile for kilobases downstream, if there is no other active gene in the region (Fig 4.5B). When this

Pol II isoform difference pattern is observed, it can be used to postulate the orientation of transcription observed from the locus, such as at the locus of myocyte specific mir-206 and mir-133b (Fig 4.6). These microRNAs occur in an otherwise unannotated region of mouse and human genomes. When the Pol II isoform data are integrated with RNA data, the picture is consistent with a novel ~15kb transcription locus, with a defined TSS, candidate conserved proximal promoter, polyA addition site and downstream transcription termination domain ~ 2 kb from the polyA addition site. This primary transcript would produce the two microRNAs by processing, plus an RNA with two proposed larger exons,

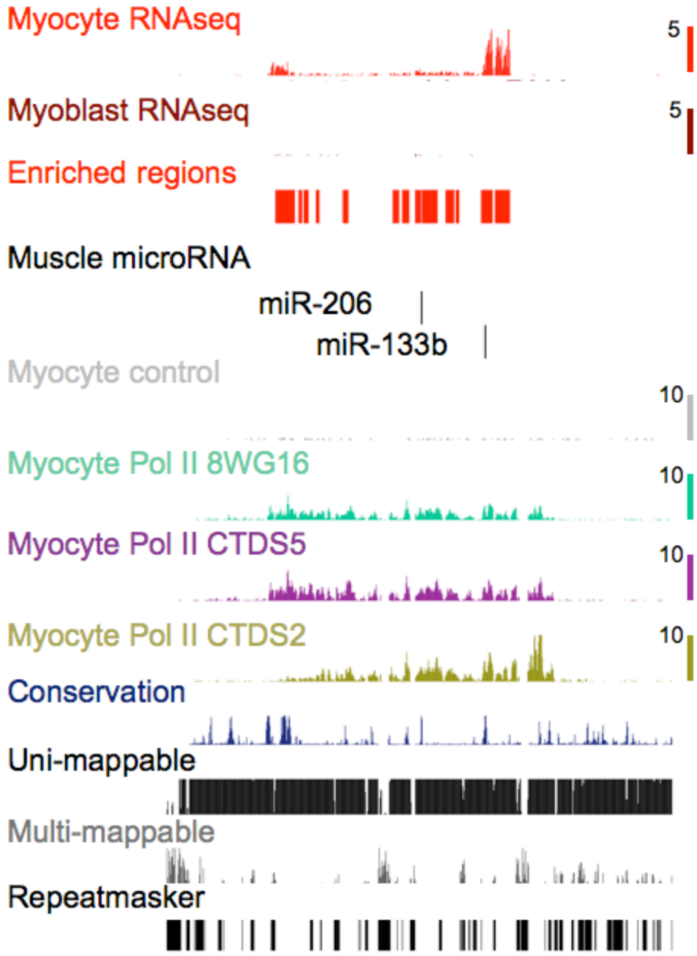


Figure 4.6. miR-206 and miR-133b transcript defined using RNA-seq and polymerase ChIP-seq. 30 kb region of chromosome 1 centered on mir-206 and mir-133b, which are expressed upon the myoblast to myocyte transition that are embedded in a novel transcript that is not in the NCBI or UCSC gene models and is also expressed upon differentiation. Our algorithm flagged the enriched regions as candidate exons, whereas the Pol II ChIP-seq marks show that the transcript is on the same strand (+) as the microRNAs.

one from each end of the proposed domain. 43 RNAFAR candidates were specific to muscle and myocytes.

RNA polymerase II stalling and its change upon differentiaton

We quantified polymerase stalling using a modified version of the polymerase stalling index that has previously been used for ChIP-chip (Zeitlinger, 2007) and ChIP-seq (Shones, 2008) with a higher threshold of 15 times more signal density around the TSS than along the rest of the gene for each phosphoisoform. While the 8WG16 and CTD S5 signals correlate very well, we found that stalled promoters showed two distinct patterns of CTDS2 signal. Whereas 58% of the stalled promoters with expression of less than 1 transcript / cell in the Myoblast, such as Fos and the heat shock factors, showed 8WG16 stalling but no CTD S2 signal, the remainder showed stalling with both the 8WG16 and the CTDS2 signal — such as in the case of the ubiquitinating enzyme Ube2w (Fig 4.5C). In particular, we noticed that the CTDS2, but not 8WG16 stalling signal disappears from the promoter of Ube2w upon differentiation. We therefore decided to classify genes at both endpoints by their expression, 8WG16 stalling status, and CTD S2 stalling status to quantitate the size and magnitude of these transitions (Fig 4.7A). We found that 1034 genes followed the pattern of no-expression with high 8WG16, high CTD S2 stalling signal in myoblasts transitioning to expression with high 8WG16, low CTD S2 signal in the myocytes and that these genes tend to be enriched for gene ontology terms such as ubiquitin cycle, protein kinase activity, and mitochondrion (Fig 4.7B). These genes are presumably primed to go on upon differentiation, which presumably releases them from

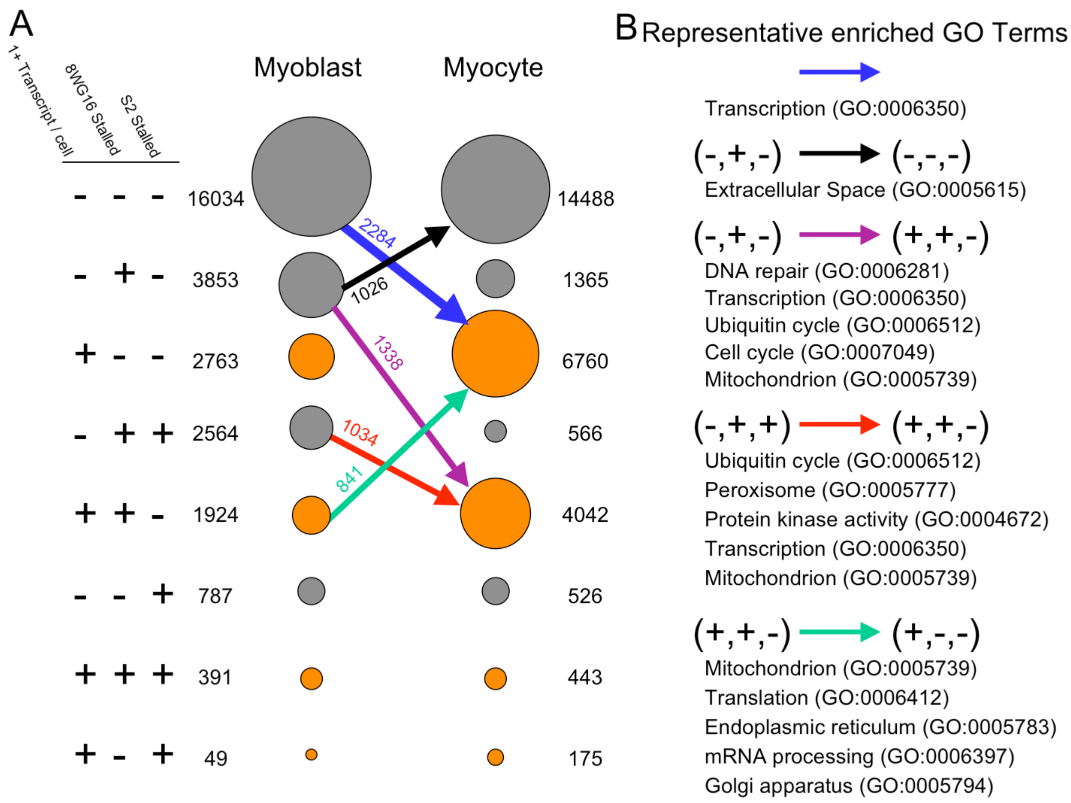

A
1+ Transcript / cell
8WG16 Stalled
S2 Stalled
Myoblast
Myocyte
- - - 16034
14488
- + - 3853
1365
+ - - 2763
6760
- + + 2564
566
+ + - 1924
4042
- - + 787
526
+ + + 391
443
+ - + 49
175
2284
1026
1338
1034
841
B Representative enriched GO Terms
Transcription (GO:0006350)
(-,+,-) → (-,-,-)
Extracellular Space (GO:0005615)
(-,+,-) → (+,+,-)
DNA repair (GO:0006281)
Transcription (GO:0006350)
Ubiquitin cycle (GO:0006512)
Cell cycle (GO:0007049)
Mitochondrion (GO:0005739)
(-,+,+) → (+,+,-)
Ubiquitin cycle (GO:0006512)
Peroxisome (GO:0005777)
Protein kinase activity (GO:0004672)
Transcription (GO:0006350)
Mitochondrion (GO:0005739)
(+,+,-) → (+,-,-)
Mitochondrion (GO:0005739)
Translation (GO:0006412)
Endoplasmic reticulum (GO:0005783)
mRNA processing (GO:0006397)
Golgi apparatus (GO:0005794)

Figure 4.7. Analysis of Polymerase Stalling Transitions. Genes with genomic loci longer than 1.2 kb were categorized in terms of their expression level, 8WG16 stalling index, and S2 stalling index in the myoblast and myocyte samples, where RNA Polymerase was considered stalled at the promoter if its stalling index is greater than 15. **A.** The top five transitions between expression and polymerase classes are shown. In addition to genes turning on upon differentiation (blue, purple, and red arrows), we have also unexpressed genes that change polymerase status (black arrow), and genes that are already expressed that lose their 8WG16 stalling signature (green arrow). **B.** representative enriched Gene Ontology terms for each of the 5 transitions. Many of the GO terms enriched for the genes with stalling in the myoblast that are expressed in the Myocyte are related to the upregulation of metabolic activity in the differentiated cell. Note that for some classes of genes with general GO terms such as Transcription (GO:0006350), a variety of strategies are used to regulate promoter loading.

one or more common repressor of elongation, whereas muscle specific genes show no preferential polymerase pre-loading at the myoblast or myocyte stage (data not shown). We finally note that RNA-seq is far more sensitive than ChIP-seq, as we are able to detect transcripts that are only present in a fraction of the cells in our sample, such as in the case of Myogenin which, while present at over 1461 RPKM in our myocyte sample (~ 1214.3

transcript/nuclei), is detected at 17 RPKM (~ 2.3 transcripts/cell) in our mybolasts. Given that less than 1% of our myoblasts cells stain as Myogenin-positive and assuming that the expression of the few exponential cells that are myogenin positive is closer to 1214 transcript/cell, then we calculate that as few as 1,600 cells in the myoblast sample (out of 2.8 million, or 0.06%) are myogenin positive. Hence, we observe nearly no signal in our polymerase ChIP-seq from $2*10^7$ cells at the Myogenin locus, while we can measure the mRNA.

Discussion

RNA sequencing revealed ~ 18 k previously unannotated regions of transcription which might identify new exons of existing genes or new genes in 3 tissues and one cell line. RNA splicing has been problematic for microarray methods and inaccessible for SAGE. Here, the shear number of reads produced by the Solexa platform RNA made it possible to identify splice events very effectively for high and moderate abundance RNAs, and to sporadically detect splice events for rarer transcripts. To map splice isoforms comprehensively for RNAs of all prevalence classes and for rarer alternative splice isoforms, requires a different approach, such as building prevalence-normalized input cDNA samples to distribute the sequence sampling power evenly across all transcript species. This would require no novel technology. The long-range contiguity of splice choices cannot be extracted from our data, but the application of rapidly improving "paired-end" variations of ultra-high throughput sequencing can address this issue ("paired" sequences are determined for both ends of single segment of DNA, and starting DNA is of a known length class). Coupled with additional bioinformatics tools, this could allow mapping of long-range splice contiguity by RNA-seq. Combining prevalence normalization of input RNA with paired-end sequencing would presumably give the most complete splicing pattern map, and this analysis might benefit from longer reads, such as those from the 454 platform, to gain more contiguity.

Our data were very reproducible. There was strong evidence for the presence and quantitfication of mRNA down to 1 transcript / cell. These measurements reach the lowest mRNA levels that are likely to be biologically relevant in animal cells, although we caution that the lower limit of biological significance may well be function dependent, i.e. half a transcript / cell might still be significant for a transcription factor, whereas 10 transcript / cell might not be significant for genes that would need to be expressed at a high level such as actins or tubulins. For mixed cell tissues like our differentiating C2C12 cells, RNA-seq at the sampling levels used was also able to detect transcripts that were prevalent in a minority cell population. However, current technology and costs mean that low prevalence RNAs from minority cells will not be accessible. Instead, it should be possible to map the transcriptomes of rare cell types and states by greatly reducing the amount of input mRNA — aiming for the single-cell level. Anticipated improvements in sequencing platforms also suggest that RNA-seq will increase in accuracy because of longer reads and increase in sensitivity because of larger read numbers.

Although RNA-seq has no background from cross-hybridization as microarrays do, it is not free of ambiguities. 14 to 30% of 25 bp reads from each mouse (or human) RNA-seq sample matched equally well at 2–10 different locations on the mouse genome, rather than mapping best to a single site (Table 4.1). Such "multireads" are expected, given the prominent role of gene and segmental duplications in genome evolution; they accounted for

83–90% of all multireads in our data. Discarding all multireads (as default settings in current Solexa software does), will report that duplicated genes are expressed at very low levels or not expressed at all, as illustrated for the Ubb Ubiquitin gene family member (Fig 4.1E). In less closely related gene families, multireads will have variable impact on RNA quantification, depending on the extent of sequence similarity within each gene and across the extended gene family. The strength of evidence for detection of any given rare transcript by RNA-seq, especially if it has garnered multiple unique sequence reads, may be stronger than for microarrays, because array fluorescence signals from a low abundance true positive can be very difficult to distinguish, numerically and statistically, from fluctuations in background hybridization and dye-labeling. However, short sequence reads that contain one or more errors (wrong base calls) might coincidentally match — in mutated form — to an existing sequence in the genome. This creates a different kind of background in RNA-seq. Different sequences will be more vulnerable to this than others, and future improvements in machine platforms, lengths of high quality sequence reads, better algorithms for base calling and for — especially — development of more sophisticated probabilistic error models for each transcriptome, should all contribute to improving certainty for the rarest RNAs. Like RNA-seq, high density tiling arrays can discover previously unknown RNAs (Krapanov, 2007). However, the data are not directly comparable, because our input RNA was doubly polyA selected, and because they used mixed sources of RNA. This focuses the attention on protein coding mRNAs, and to a much lesser extent, on their partly processed precursors. Future applications of RNA-seq to those other RNA preparations and also to very newly synthesized RNAs will allow more direct comparisons to non-polyA selected data as well as to RNA polymerase II occupancy maps.

The sensitivity of RNA-seq combined with the ChIP-seq of the different RNA polymerase II CTD phosphoisoforms revealed an unexpected dichotomy at stalled promoters depending on the presence or absence of a CTD S2 Pol II signal. We note that our CTD S2 signal seems to be specific to heptads phosphorylated at S2 only, as opposed to repeats phosphorylated at both S2 and S5, which suggests that the S2 signal at these promoters is due to either S2 overphosphorylation or CTD S5 dephosphorylation of CTD S2–S5 repeats. In either case, this suggests that there is an additional process beyond the classically described control of elongation by phosphorylation of S2 by PTEFb. This could be a signature of known repressors of elongation such as NELF (Lee, 2008) or represent the presence of recruited CTD S5 phosphatases such as the SCP family that are recruited by repressors such as NRSF/REST (Yeo, 2005). Histone modification data under the same differentiation conditions should highlight whether these different stalled promoters have distinct histone modification profiles.

Methods

Cell culture and immunocytochemistry

The mouse C2C12 skeletal muscle cell line was grown in 15 cm dishes in a 37˚C incubator at 5% CO_2 in DMEM (Gibco) with 20% fetal bovine serum (Hyclone), penicillin at 50 units/mL, and streptomycin at 50 ug/ml (Gibco). At confluence, the cells were rinsed twice with phosphate-buffered saline (PBS), and the medium was changed to low serum medium (DMEM with 2% horse serum, 1 μM insulin, and penicillin/streptomycin). Differentiated cultures were fed every 24 hours until harvest. At harvest, parallel plates were set aside for fixation in 4% paraformaldehyde or 70% ethanol for immunocytochemistry. Primary antibodies were from Novocastra (NCL anti-myoD1), Cortex Biochem (CR2031R rabbit anti-myosin heavy chain), or from a laboratory supply of ascites fluid (myogenin — F5D). Cells were initially permeabilized in 0.15% Triton X-100 in PBS for 15 minutes, then rinsed twice in PBS and blocked for 1 hour in 10% normal goat serum. The blocking serum was removed and an aliquot of primary antibody (1:500 anti-myogenin, 1:100 anti-myoD, 1:300 anti-MHC) diluted in 1.5% goat or donkey serum in PBS was added and incubated at room temperature for 1 hour. The primary was then removed, the dish was rinsed in 0.1% Triton in PBS 3 times for 5 minutes each at room temperature, and fluorescent secondary antibody was applied (goat anti-mouse Alexa 488, donkey anti-rabbit Alexa 564, Molecular Probes) for 1 hour at room temperature. The dishes were rinsed again 3 times for 5 minutes in 0.1% Triton in PBS, rinsed once for two minutes in PBS, and then equilibrated in equilibration buffer (Anti-fade kit, Molecular Probes) before mounting in glycerol with DAPI. Cells were visualized on a Zeiss Axiophot and > 1000 nuclei were counted in both channels for percentage of myogenin, myoD, and MHC-positive nuclei. Images were collected using OpenLab software.

Estimate of nuclear counts

Parallel cell preparations fixed in 1% paraformaldehyde were used for nuclear count estimates. The cell pellets were lysed in 5mM PIPES, pH 8.0, 85 mM KCl, 0.5% NP-40, and the remaining nuclear fraction was subjected to overnight incubation at 65˚C in Tissue and Cell lysis buffer with Proteinase K (EpiCentre MasterPure DNA extraction kit). Genomic DNA was then extracted using the EpiCentre protocol, and quantified on the

NanoDrop spectrophotometer. Genomic DNA yields were then converted to nuclei counts using the value of 6.4 pgs of DNA per nucleus.

RNA preparation

Cells were removed from the incubator, rinsed twice quickly with room temperature PBS, and then lysed with 2.25 ml Trizol per 15 cm plate (InVitrogen). The lysate was sheared 10 times through a 21 gauge needle and then frozen on pulverized dry ice until extraction. Lysates were thawed at room temperature, spun at 12,000 X g for 10 minutes at 4˚C, and the supernatant transferred to a new tube. 200 μLs of chloroform:isoamyl alcohol (24:1) were added, the tube was agitated vigorously by hand for 15 seconds, and then spun for 10 minutes at 12,000 X g at 4˚C. The supernatant was transferred to a new tube, and re-extracted with an equal volume of phenol:chloroform:isoamyl alcohol equilibrated to pH 4.5. After spinning at 12,000 X g for 15 minutes at 4˚C, the aqueous phase was mixed with an equal volume of isopropanol, and placed on ice for 10 minutes to precipitate RNA. The sample was spun for 10 minutes at 12,000 X g at 4˚C, the pellet was rinsed once with 75% ethanol, and then dried on the benchtop for 10 minutes, and resuspended in 20 μLs of ddH_2O for 10 minutes at 37˚C. The sample was then treated with 10 μLs Baseline Zero DNAse (EpiCentre Biotechnology) for 20 minutes, and respiked with an additional 5 μLs of DNAse for another 20 minutes. After the addition of stop buffer, the sample was phenol:chloroform extracted and precipitated overnight in 0.3M sodium acetate in 70% ethanol. The sample was spun at 14,000 X g at 4˚C for 25 minutes, the pellet was rinsed once with 70% ethanol, spun for 5 minutes at 8,000 X g at 4˚C, and the supernatant removed. The pellet was dried at room temperature, and then resuspended in 20 μLs of ddH_2O. Concentration was determined on a Nanodrop spectrophotometer, and small aliquots reserved for evaluation of RNA integrity on the Agilent 6000 BioAnalyzer. Oligo dT selection was performed twice on 75 ugs of input total RNA using Dynal magnetic beads (InVitrogen) according to the manufacturer's protocol. After selection, a single 100 ng aliquot of mRNA was reserved for evaluation on the BioAnalyzer.

cDNA preparation

100 ngs of double-selected mRNA was used as template for cDNA synthesis. The mRNA was fragmented by addition of 5X fragmentation buffer (200 mM Tris Acetate, pH 8.2, 500 mM KOAc, and 150 mM MgOAc) and heating at 94˚C for 2 minutes 30 seconds in a thermocycler. The sample was immediately transferred to ice, and then run over a G50 sephadex column (USA Scientific) to remove the fragmentation ions. The sample was reduced to 10.5 µls in a speedvac, 3 µgs of random hexamers were added and the sample was reheated to 65˚C for 5 minutes in a thermocycler. The sample was then placed on ice, and reagents for first-strand reverse transcription were added according to the manufacturer's protocol (InVitrogen cDNA synthesis kit, catalog # 48190-011). After the first strand was synthesized, a custom second strand synthesis buffer was added (Illumina), and dNTPs, RNAseH, and E. coli polymerase I were added to nick translate the second strand synthesis for 2.5 hours at 16˚C. The reaction was then cleaned up on a QiaQuick PCR column (Qiagen) and eluted in 30 uls of EB buffer.

Adapter ligation, size selection and amplification

cDNA ends were subjected to an end repair protocol (Illumina Genomic DNA sequencing kit), followed by addition of an A base for ligation, and ligation of custom amplification adapters (Illumina). After the final QiaQuick PCR cleanup, an additional cleanup was performed over G50 Sephadex (USA Scientific), to prevent sample dispersal when loading into the agarose gel for size selection. In order to size select the sample for a narrow distribution of cDNA around 200 bps, a 2% low melt agarose gel (NuSieve GTG Agarose, Cambrex) in TAE was prepared in a 10 cm mold. The sample was loaded in the gel with a 100 bp ladder (Invitrogen) located 3 lanes to the side to prevent contamination of the sample with ladder bands. The gel was run at 80 volts constant voltage until the dye front was about 1.5 cm from the bottom of the gel. The gel was post-stained in ethidium bromide for 20 minutes, destained for 20 minutes in ddH2O, and the bands were visualized and excised on a UV illumination stand. Bands were excised in a narrow distribution of 200 +/- 25 bps with a disposable scalpel. The gel slices were extracted using the QiaExII kit (Qiagen) according to the manufacturer's protocol, and 1/6 of the eluate volume was used in an amplification protocol with Phusion polymerase (Finnzymes) and custom amplification primers (Illumina). The template was amplified for 15 cycles using the following protocol: 98°C for 30 seconds, 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds, and 72°C for 5 minutes. The amplification product was cleaned up on a QiaQuick PCR column, reduced to 25 μLs volume in a speed vac, and then passed over a G50 Sephadex column. The DNA concentration was determined on a Nanodrop spectrophotometer, and an aliquot of the sample was diluted to 10 nanomolar concentration with EB buffer to be used as input for the Illumina Cluster generation protocol. We regularly use a 4 pmole mass to seed the flow cell to an approximate cluster density of 30,000 clusters per tile.

Spike controls derivation and validation

The MM9 version of the RefSeq database was used as a source of mRNA lengths for the known mouse transcriptome. After plotting the length profile for the mouse transcriptome, we chose several lengths to cover the distribution of mouse mRNA lengths (~ 300 nts, ~ 1500 nts, and ~ 10,000 nts). *Arabidopsis* total RNA was reverse transcribed into double stranded DNA, and amplification primers were designed to 3 expressed sequences of about 1400 nt length each and 3 sequences between 300 and 400 nts each. After PCR amplification and gel electrophoresis, specific bands were excised, eluted, and cloned into a modified version of pBluescript KS II (-) (Stratagene). Additionally, three fragments of the lambda genome were also recovered from electrophoresis after an Sph I digest of lambda genomic DNA, and cloned into the same expression vector. The expression vectors were then digested to allow sense strand transcription for the expressed sequences, and *in vitro* transcription was performed using the AmpliScribe T3 or T7 kits (EpiCentre). The reaction products were split in two, passed first over a G50 Sephadex column (USA Scientific), and then further cleaned up over RNEasy Minicolumns (Qiagen). First pass quantitation was performed on the Nanodrop UV spectrophotometer. To corroborate the Nanodrop estimates, further quantitation of RNA mass was performed using the RiboGreen reagent (Molecular Probes) in a BioRad Fluorometer. The full length integrity of the RNA was inspected first on a 1.5% agarose gel stained with ethidium bromide, and also using the Agilent BioAnalyzer. After determination of the concentration of the RNA samples, serial dilutions were made to the concentrations indicated.

Sequencing and read mapping on the genome and across splices

Libraries were sequenced as 36-mers using the standard Solexa pipeline (version 0.2.6), but raw reads were truncated as 25-mers and remapped using version 0.3 of Eland (David Cox, under preparation) using the –multi option. Mouse ChIPSeq reads were mapped against the standard mm9 mouse build from UCSC (NCBI v37). Mouse RNAseq reads were similarly mapped against an expanded genome consisting of the standard mm9 genome and 42-mers representing the last 21 bp of the upstream exon and the first 21 bp of the corresponding downstream exon of each mRNA splice documented in the knownGene table for mm9.

Normalized gene locus expression level analysis and multi-read probability assignment

Unique reads in the expanded genome that landed within any exons of NCBI gene models (v37.1) were counted. Reads from all samples that did not fall within known exons where aggregated into candidate exons by requiring regions with at least 15 reads whose starts are not separated by more than 30 bp. These candidate exons were then rerun against the mappable, unmatched reads from each respective experiment to obtain their counts for that experiment. Reads that fell onto exons and candidate exons as well as splices were summed up for each locus and normalized by the locus length into the expanded exonic read density, i.e., reads per KB per million reads (RPKM), using the formula:

$$R = \frac{10^9 C}{NL}$$

where C is the number of mappable reads that fell onto the gene's exons, N is the total number of mappable reads in the experiment, and L is the sum of the exons in base pairs. In particular, candidate exons are consolidated with neighboring NCBI gene models if they (a) have an RPKM that is within the same order of magnitude or higher than that of the gene model and (b) meet either of 2 criteria:

— candidate is within an intron of the gene model,

— candidate is within 20 kb of the 3’ or 5’ end of the nearest, known exon.

The requirement for (a) prevents counting of reads within introns of highly expressed genes that are likely from unspliced introns or downstream of the polyadenylation signal and that would be from partially unspliced hnRNA rather than processed mRNA. Candidate exons that fell within 20 kb of one another but further than 2 0kb from any other gene were aggregated into predicted “FAR” loci.

The expanded RPKM were then used to calculate the probability that a multi-read came from a particular known or candidate exon, and the resulting fractional counts were added to the total count for the gene locus, which was renormalized into a multi RPKM. The expanded and multi RPKM for each locus were combined to produce a final RPKM.

Conversion of RPKM into absolute transcript numbers

Assuming uniform distribution of the mappable reads across the transcriptome, the probability of observing C reads on a transcript of length L in N tries corresponds to the fraction of the transcriptome composed of the transcript:

$$\frac{C}{N} = \frac{XL}{T}$$

where X is the copy number of the transcript and T is the length of the transcriptome in base pairs. We can substitute final RPKMs to get:

$$X = \frac{C}{NL}T = \frac{R}{10^9}T$$

where we can either derive T from the starting amount of mRNA (assuming that we had 100% efficiency in cDNA synthesis), or fit T from spike-in data, which should be more accurate.

Chromatin Immunoprecipitations

Chromatin immunoprecipitations and sequencing libraries were prepared as described in (Mortazavi et al., 2006) with the following modifications: 4 x 10^7 cells were used per IP using 100 uL of dynal anti-igG beads with 12.5 ug of antibodies. The following antibodies from Abcam were used: the monoclonal antibody 8WG16 (unphosphorylated CTD), rabbit polyclonal ab-5131 (Serine 5 CTD phosphorylated, ChIP-grade), and the rabbit polyclonal ab-5095 (Serine 2 CTD phosphorylated, ChIP-grade). Libraries for each phosphorylation state and control libraries were prepared as previously described (Johnson, 2007) from a single IP.

Polymerase Stalling Index Calculations

Stalling index calculations were done as previously described (Zeitlinger, 2007; Schones, 2008) with modifications to deal with ChIP-seq background and identification of 5' and 3' boundaries. Briefly, RNAFAR annotations were used to find gene boundaries, including the predicted TSS. Normalized (i.e., per million) read counts in all regions for each ChIP were corrected by subtracting the corresponding normalized reads from the control libraries. Stalling ratios were calculated by dividing the IP density on 600 bp region centered on the TSS by the read density over the gene model minus the first 600 bp. Genes with ratios over 15 were considered stalled.

Chapter 5

CONCLUSIONS

I have spent the last four years of my life living and breathing NRSF, ChIP-seq, and RNA-seq. It is difficult to think of the way things used to be without access to an ultra-high - throughput sequencer, which is destined to be the biologist's new best friend, alongside the old trusty QPCR thermocyclers. I don't think this thesis would be complete without a couple of pages discussing my opinions on the outlook for the techniques introduced herein as well as where I see the field headed or, at least, where I would like it to go.

ChIP-seq has been an unqualified success. Our publication of ChIP-seq for NRSF in *Science*, along with near simultaneous publication of two ChIP-seq publications primarily focused on ChIP-seq of histone modifications in Cell (Barski, 2007) , Nature (Mikkelsen, 2007) and of the transcription factor STAT1 (Robertson, 2007) marked a quantum leap in the field of mammalian chromatin immunoprecipitation. The resolution and quality of the output were an order of magnitude above and beyond what anybody, including ourselves, thought was achievable at the time on the genomic scale. While not part of my thesis, the work that we have done in the Wold lab on ChIP-seq with the myogenic transcription factors MyoD and Myogenin makes it clear that the technique works spectacularly well even for the smallest binding site, and that there can be no turning back. For all practical purposes, ChIP-chip and other derivatives are now obsolete for mammalian genomes as long as one has access to Solexa-class sequencers, and the field is rapidly switching over as the technology has caught on. This does not mean that ChIP-seq can save an otherwise lousy IP — if anything, we are more sensitive to the quality of the input than ever. However, the digital nature of the technology allows us to develop new metrics, such as our IP efficiency, to assess and compare different IPs and antibodies objectively. As the pool of ChIP-grade antibodies continues to expand, we can expect efforts such as ENCODE to produce ChIP-seq results for hundred of factors, thus giving us a real chance to look comprehensively at transcription factor-DNA binding in at least a few human cell types in the near future. I am also looking forward to seeing ChIP-seq applied to organisms other than mouse or human, as ChIP-seq, particularly with the Pol II CTD antibodies, can accommodate any of the sequenced eukaryotic genomes such as sea urchins or ascidians.

RNA-seq is younger, yet seems also destined for a bright future. While finishing up my thesis, three papers have come out in April and May 2008 describing RNA-seq in *Arabidopsis thaliana* (Lister, 2008), *S. cerevisiae* (Nagalakshmi, 2008), and S. pombe (Wilhelm, 2008). The major benefits of RNA-seq are both transcript/exon discovery and the first comprehensive and quantitative assay of alternative splicing. While ChIP-seq effectively made mammalian tiling arrays obsolete, RNA-seq is threatening the primacy of the venerable microarray for expression analysis. While microarrays will continue to have a cost advantage in the foreseeable future, experiments that can benefit from exon and splice discovery and quantification will benefit greatly from RNA-seq. In particular, scaling down the amount of starting material so that we can sequence the transcriptome of a single cell, such as a nemtode or human neuron, is the next quantum step in the technology and is being worked on as we speak.

Stepping back to take in the state of (regulatory) biology in 2008, the sequencing of the reference human genome in 2001 represented a watershed in the history of biology, as it gave us access to the first mammalian scale genomes and forever transformed biology from a relatively data-poor field to a data-rich one, which has reshaped every aspect of being a biologist and doing biology. While the reference sequence for human and the other model organisms gave us much to work on in terms of annotating genes and regulatory elements, which continues to be the driving force behind projects such as the NHGRI ENCODE (the ENCyclopedia Of Dna Elements) initiative, the human genome project revolutionized high-throughput DNA sequencing technology, which had three major benefits, two of which were predictable and one of which has come as a surprise. All three are worth discussing, as together they will continue to influence the future of biology and of its allied fields.

The first and most obvious benefit is that many mammalian and deuterostome genomes have been sequenced in the last decade, giving us the raw material necessary for the analysis of the evolution of vertebrates in general and of our own lineage in particular. As our analyses of the role of conservation and the evolution of gene regulation will extend beyond solo and possibly exceptional examples such as that of NRSF, we shall be able to understand and decode the gene regulatory networks underlying each species, and to identify those changes that have been selected for in the life history of each species. A logical extension of this is that we will be able to reconstruct our ancestral genomes, such as the ancestral primate and probably the ancestral mammal and vertebrate, as well as understand the evolution of these gene regulatory networks at the molecular and informational level. While this will require much work and major advances in our understanding and prediction of protein-protein interaction, as well as protein-DNA binding, it is conceivable that we will have the requisite knowledge at hand in the coming decades. A predictive decoding and understanding of all gene regulatory relationships from primary sequence is the holy grail and would take its place alongside the Central Dogma of Biology and the Evolutionary Synthesis as one of the great foundational, intellectual achievements of the entire field. This is one of the grand, organizing projects of regulatory biology and may ultimately be our most significant achievement as a species. Whether it takes a decade or a century to achieve, it is worth pursuing regardless of any positive societal benefits or not.

The second benefit is that the availability of the reference sequence has allowed us to ask how each individual member of our species differs from the reference and from one another as well as to quantify it. This has been a long time goal of human and population geneticists and offers the alluring prospects of identifying the underlying genetic determinants controlling phenotypic traits that are so evident when comparing a child to his parents. It also offers the hope to tailor personalized medical treatment to people based on their genotype. These twin promises have driven the recent progresses that has given us the current generation of ultra-high-throughput sequencing, where we can get 40–80 million reads of 25–32 bp per run, i.e., nearly 2.5 Gbp of sequence or nearly 1x coverage of the human genome. While still too expensive to do routinely at the 20x coverage necessary for these short reads to achieve a reliable resequencing of a human genome, we are within grasp of the “$1000 genome” that is considered the threshold for “routine” medical resequencing. Of course, obtaining the genoptype simply to use it for the strictly correlational studies that are still the bread-and-butter of population genetics is not nearly as meaningful as one would hope. However, when individual genotypes are decoded down to the base-bp level changes to underlying, reference gene regulatory networks coming from the studies discussed as the first benefit, we may achieve the holy grail of understanding and acting upon our health predisposition and achieve even longer, more productive lives.

The third, unexpected consequence of the rise of and rapid advances in ultra-high-throughput sequencing is that any assay that can be converted into counting DNA fragment becomes immediately practical. There is today an effervescence of new techniques and protocols applying the sequence census assays to problems such as DNAse hypersensitivity, DNA methylation, small RNA discovery, etc., which will ultimately affect nearly all fields of biology. While counting assays such as SAGE or low-coverage mRNA sequencing (for transcript discovery) have been around for over a decade, they were expensive, technically challenging and not particularly comprehensive. The fortuitous availability of robust, cloning-free sequencing developed for resequencing repurposed for counting assays is one of these now-obvious ideas that nobody at the time (circa 2005), including I, thought was anything more than fanciful. It took the truly visionary and dedicated single-mindedness of my advisor to make it happen and her embrace of ENCODE made it all possible. I was fortunate that NRSF was the project that lead to the wonderful collaboration between Barbara's and Rick Myers' groups and which allowed me to witness the birth of a transformational technology that will be a highlight of both of their distinguished careers.

BIBLIOGRAPHY

Abderrahmani, A., M. Steinmann, V. Plaisance, G. Niederhauser, J.A. Haefliger, V. Mooser, C. Bonny, P. Nicod, and G. Waeber. 2001. The transcriptional repressor REST determines the cell-specific expression of the human MAPK8IP1 gene encoding IB1 (JIP-1). Molecular and Cellular Biology 21: 7256–67.

Abramovitz L., T. Shapira, I. Ben-Dror, V. Dror, L. Granot, T. Rousso, E. Landoy, L. Blau, G. Thiel, and L .Vardimon. 2008. Dual role of NRSF/REST in activation and repression of the glucocorticoid response. Journal of Biological Chemistry 283(1):110–9

Andres, M.E., C. Burger, M.J. Peral-Rubio, E. Battaglioli, M.E. Anderson, J. Grimes, J. Dallman, N. Ballas, and G. Mandel. 1999. CoREST: a functional co-repressor required for regulation of neural-specific gene expression. Proceedings of the National Academy of Sciences of the United States of America 96: 9873–78.

Atouf, F., P. Czernichow, and R. Scharfmann. 1997. Expression of neuronal traits in pancreatic beta cells — Implication of neuron-restrictive silencing factor/repressor element silencing transcription factor, a neuron-restrictive silencer. Journal of Biological Chemistry 272: 1929–34.

Ballas, N., C. Grunseich, D.D. Lu, J.C. Speh, and G. Mandel. 2005. REST and its co-repressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. Cell 121: 645–57.

Bailey, T.L., and C. Elkan. 1995. Unsupervised Learning of Multiple Motifs In Biopolymers Using EM. Machine Learning 21: 51–80.

Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129(4):823–37.

Bejerano, G., C.B. Lowe, N. Ahituv, B. King, A. Siepel, S.R. Salama, E.M. Rubin, W.J. Kent, and D. Haussler. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90.

Berkes, C.A., and S.J. Tapscott. 2005. MyoD and the transcriptional control of myogenesis. Semin Cell Dev Biol. 16(4–5):585–95

Boffelli, D., C.V. Weer, L. Weng, K.D. Lewis, M.I. Shoukry, L. Pachter, D.N. Keys, and E.M. Rubin. 2004. Intraspecies sequence comparisons for annotating genomes. Genome Research 14: 2406–11.

Bruce, A.W., I.J. Donaldson, I.C. Wood, S.A. Yerbury, M.I. Sadowski, M. Chapman, B. Gottgens, and N.J. Buckley. 2004. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. Proceedings of the National Academy of Sciences of the United States of America 101: 10458–63.

Bruce, A.W., A. Krejcí, L. Ooi, J. Deuchars, I.C. Wood, V. Dolezal, and N.J. Buckley. 2006. The transcriptional repressor REST is a critical regulator of the neurosecretory phenotype. Journal of Neurochemistry 98(6):1828–40.

Buratowski, S. 2005. Connections between mRNA 3' end processing and transcription termination. Curr Opin Cell Biol. (3):257–61.

Cawley, S., S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A.J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T.R. Gingeras. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499–509.

Chapman, R.D., M. Heidemann, T.K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. 2007. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science 318(5857):1780-2

Chen, Z.F., A.J. Paquette, and D.J. Anderson. 1998. NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. Nature Genetics 20: 136-142.

Chong, J.A., J. Tapia-Ramirez, S. Kim, J.J. Toledo-Aral, Y. Zheng, M.C. Boutros, Y.M. Altshuller, M.A. Frohman, S.D. Kraner, and G. Mandel. 1995. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. Cell 80:949–57.

Corden JL. 2007. Transcription. Seven ups the code. Science 318(5857):1735–6.

Coulson, J.M., J.L. Edgson, P.J. Woll, and J.P. Quinn. 2000. A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: a potential role in derepression of neuroendocrine genes and a useful clinical marker. Cancer Research 60(7):1840-4.

Core, L.J., and J.T. Lis. 2008. Transcription Regulation Through Promoter-Proximal Pausing of RNA Polymerase II. Science 319 (5871):1791–2.

D'Alessandro, R., A. Klajn, L. Stucchi, P. Podini, M.L. Malosio, and J. Meldolesi. 2008. Expression of the neurosecretory process in pc12 cells is governed by rest. Journal of Neurochemistry 105(4):1369–83.

Dallman, J.E., J. Allopenna, A. Bassett, A. Travers, and G. Mandel. 2004. A conserved role but different partners for the transcriptional co-repressor CoREST in fly and mammalian nervous system formation. Journal of Neuroscience 24: 7186–93.

Davidson, E.H. 2006. The regulatory genome: gene regulatory networks in development and evolution. Academic Press/Elsevier, San Diego, CA.

Dehal, P., P. Predki, A.S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C.L.E. Zhou, S. Rash, Q. Zhang, L. Gordon, J. Kim, C. Elkin, M.J. Pollard, P. Richardson, D. Rokhsar, E. Uberbacher, T. Hawkins, E. Branscomb, and L. Stubbs. 2001. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. Science 293: 104–11.

Egloff S, D. O'Reilly, R.D. Chapman, A. Taylor, K. Tanzhaus, L. Pitts, D. Eick, and S. Murphy. 2007. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science 318(5857):1777–9.

Fickett, J.W. 1996. Quantitative discrimination of MEF2 sites. Molecular and Cellular Biology 16:437–41.

Filippova, G.N. 2008. Genetics and epigenetics of the multifunctional protein CTCF. Curr Top Dev Biol. 80:337–60.

Fuller, G.N., X. Su, R.E. Price, Z.R. Cohen, F.F. Lang, R. Sawaya, and S. Majumder. 2005. Many human medulloblastoma tumors overexpress repressor element-1 silencing transcription (REST)/neuron-restrictive silencer factor, which can be functionally countered by REST-VP16. Mol Cancer Ther. 4(3):343–9.

Gaudet, J., and S.E. Mango. 2002. Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. Science 295(5556):821–5.

Guardavaccaro, D., D. Frescas, N.V. Dorrello, A. Peschiaroli, A.S. Multani, T. Cardozo, A. Lasorella, A. Iavarone, S. Chang, E. Hernando, and M. Pagano. 2008. Control of chromosome stability by the beta-TrCP-REST-Mad2 axis. Nature 452(7185):365–9.

Gradwohl, G., A. Dierich, M. LeMeur, and F. Guillemot. 2000. Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. Proceedings of the National Academy of Sciences of the United States of America 97:1607–11.

Griffiths-Jones, S. 2004. The microRNA Registry. Nucleic Acids Research 32:D109–111.

Gromak, N., S. West, and N.J. Proudfoot. 2006. Pause sites promote transcriptional termination of mammalian RNA polymerase II. Molecular and Cellular Biology (10):3986–96.

Guenther, M.G., S.S. Levine, L.A. Boyer, R. Jaenisch, and R.A. Young. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130(1):77–88.

Hamilton, A.T., S. Huntley, J. Kim, E. Branscomb, and L. Stubbs. 2003. Lineage-specific expansion of KRAB zinc-finger transcription factor genes: Implications for the evolution of vertebrate regulatory networks. Cold Spring Harbor Symposia on Quantitative Biology 68:131–140.

Harbers, M., and P. Carninci. 2005. Tag-based approaches for transcriptome research and genome annotation. Nature Methods. (7):495-502

Harbison, C.T., D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature. 431(7004):99–104.

Harris, T.D., P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J.W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S.R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss, and Z. Xie. 2008. Single-molecule DNA sequencing of a viral genome. Science. 320(5872):106–9

Hart, C.E., L. Sharenbroich, B.J. Bornstein, D. Trout, B. King, E. Mjolsness, and B.J. Wold. 2005. A mathematical and computational framework for quantitative comparison and integration of large-scale gene expression data. Nucleic Acids Research 33: 2580–94.

Hentsch, B., A. Mouzaki, I. Pfeuffer, D. Rungger, and E. Serfling. 1992. The Weak, Fine-Tuned Binding of Ubiquitous Transcription Factors to the Il-2 Enhancer Contributes to Its T-Cell-Restricted Activity. Nucleic Acids Research 20: 2657–65.

Hersh, L.B. and M. Shimojo. 2003. Regulation of cholinergic gene expression by the neuron restrictive silencer factor/repressor element-1 silencing transcription factor. Life Sciences 72: 2021–28.

Huang, H.P., Liu, M., El-Hodiri H.M., Chu K., Jamrich M., and Tsai M.J. 2000. Regulation of the pancreatic islet-specific gene BETA2 (neuroD) by neurogenin 3. Molecular and Cellular Biology 20: 3292–307.

Huang, Y.F., S.J. Myers, and R. Dingledine. 1999. Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. Nature Neuroscience 2: 867–72.

Impey, S., S.R. McCorkle, H. Cha-Molstad, J.M. Dwyer, G.S. Yochum, J.M. Boss, S. McWeeney, J.J. Dunn, G. Mandel, and R.H. Goodman. 2004. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. Cell 119(7):1041–54.

John, B., A.J. Enright, A. Aravin, T. Tuschl, C. Sander, and D.S. Marks. 2004. Human MicroRNA targets. Public Library of Science Biology 2:1862–79.

Johnson DS, A. Mortazavi, R.M. Myers, and B. Wold. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830):1497–502.

Johnson, R., C. Zuccato, N.D. Belyaev, D.J. Guest, E. Cattaneo, and N.J. Buckley. 2008. A microRNA-based gene dysregulation pathway in Huntington's disease. Neurobiol Dis. 3:438–45.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. Nucleic Acids Research 31:51–54.

Kemp, D.M., J.C. Lin, and J.F. Habener. 2003. Regulation of Pax4 paired homeodomain gene by neuron-restrictive silencer factor. Journal of Biological Chemistry 278(37):35057–62.

Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. Genome Research 12: 996–1006.

Kim, J.B., G.J. Porreca, L. Song, S.C. Greenway, J.M. Gorham, G.M. Church, C.E. Seidman, and J.G. Seidman. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316(5830):1481–4.

Kim, T.H., and B. Ren. 2006. Genome-wide analysis of protein-DNA interactions. Annual Review of Genomics and Human Genetics 7:81–102.

Kloosterman W.P., E. Wienholds, E. de Bruijn, S. Kauppinen, and R.H. Plasterk. 2006. In Situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. Nature Methods 3:27–29.

Kapranov, P., et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 316(5830):1484–8.

Kraner, S.D., J.A. Chong, H.J. Tsay, and G. Mandel. 1992. Silencing the type II sodium channel gene: a model for neural-specific gene regulation. Neuron. 9(1):37–44.

Krek, A., D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, P. MacMenamin, I. daPiedade, K.C. Gunsalus, M. Stoffel, and N. Rajewsky. 2002. Combinatorial microRNA target predictions. Nature Genetics 37: 495-500.

Kosik, K.S., and A.M. Krichevsky. 2005. The elegance of the microRNAs: A neuronal perspective. Neuron 47: 779–782.

Krebs, C.J., L.K. Larskins, S.M. Khan, and D.M. Robins. 2005. Expansion and Diversification of KRAB zinc-finger genes within a cluster including regulation of sex-limitation 1 and 2. Genomics 6: 752–61.

Kuwabara, T., J. Hsieh, K. Nakashima, K. Taira, and F.H. Gage. 2003. A small modulatory dsRNA specifies the fate of adult neural stem cells. Cell 116: 779–793.

Kuwahara, K., Y. Saito, M. Takano, Y. Arai, S. Yasuno, Y. Nakagawa, N. Takahashi, Y. Adachi, G. Takemura, M. Horie, Y. Miyamoto, T. Morisaki, S. Kuratomi, A. Noma, H. Fujiwara, Y. Yoshimasa, H. Kinoshita, R. Kawakami, I. Kishimoto, M. Nakanishi, S. Usami, M. Harada, and K. Nakao. 2003. NRSF regulates the fetal cardiac gene program and maintains normal cardiac structure and function. EMBO Journal 22:6310–21.

Lakowski, B., S. Eimer, C. Gobel, A. Bottcher, B. Wagler, and R. Baumeister. 2003. Two suppressors of sel-12 encode C2H2 zinc-finger proteins that regulate presenilin transcription in Caenorhabditis elegans. Development 130:2117-2128.

Lee C, Li X, Hechmer A, Eisen M, Biggin MD, Venters BJ, Jiang C, Li J, Pugh BF, Gilmour DS (2008) NELF and GAGA factor are linked to promoter proximal pausing at many genes in Drosophila. Molecular and Cellular Biology 28(10):3290–300.

Lee, J.E., S.M. Hollenberg, L. Snider, D.L. Turner, N. Lipnick, and H. Weintraub. 1995. Conversion of Xenopus ectoderm into neurons by NeuroD, a basic helix-loop-helix protein. Science 268:836–44.

Lee, J.H., M. Shimojo, Y.G. Chai, L.B. Hersh. 2000. Studies on the interaction of REST4 with the cholinergic repressor element-1/neuron restrictive silencer element. Brain Res. Mol. Brain Res. 80:88–98

Lewis, B.P., C.B. Burge, and D.P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20.

Lim, L.P., N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433:769–773.

Lister, R., R.C. O'Malley, J. Tonti-Filippini, B.D. Gregory, C.C. Berry, A.H. Millar, and J.R. Ecker. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(3):523-36

Lobanenkov, V.V., R.H. Nicolas, V.V. Adler, H. Paterson, E.M. Klenova, A.V. Polotskaja, and G.H. Goodwin. 1990. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene 12:1743-53

Lunyak, V.V., R. Burgess, G.G. Prefontaine, C. Nelson, S.H. Sze, J. Chenoweth, P. Schwartz, P.A. Pevzner, C. Glass, G. Mandel, and M.G. Rosenfeld. 2002. Co-repressor-dependent silencing of chromosomal regions encoding neuronal genes. Science 298:1747–52.

Magin, A., M. Lietz, G. Cibelli, and G. Thiel. 2002. RE-1 silencing transcription factor-4 (REST4) is neither a transcriptional repressor nor a de-repressor. Neurochem International (3):195–202.

Martin, D., F. Allagnat, G. Chaffard, D. Caille, M. Fukuda, R. Regazzi, A. Abderrahmani, G. Waeber, P. Meda, P. Maechler, and J.A. Haefliger. 2008. Functional significance of repressor element 1 silencing transcription factor (REST) target genes in pancreatic beta cells. Diabetologia, Apr 3 (epub).

Martone, R., G. Euskirchen, P. Bertone, S. Hartman, T.E. Royce, N.M. Luscombe, J.L. Rinn, F.K. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. 2003. Distribution of NF-kappa B-binding sites across human chromosome 22. Proceedings of the National Academy of Sciences of the United States of America 100:12247–52.

Meinhart, A., T. Kamenski, S. Hoeppner, S. Baumli, and P. Cramer. 2005. A structural perspective of CTD function.Genes and Development 19(12):1401-15.

Mikkelsen, T.S., M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E.S. Lander, and B.E. Bernstein. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153):553–60.

Monaco, C., S. Otto, J. Han, and G. Mandel. 2006. Reciprocal actions of REST and a microRNA promote neuronal identity. Proceedings of the National Academy of Sciences of the United States of America 103: 2422–27.

Mori, N., C. Schoenherr, D.J. Vanderbergh, D.J. Anderson. 1992. A common silencer element in the SCG10 and type II Na+ channel gene binds a factor present in non-neuronal cells but not in neuronal cells. Neuron 9: 45–54.

Mortazavi, A., E.C. Leeper Thompson, S.T. Garcia, R.M. Myers, B. Wold. 2006. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. Genome Research (10):1208–21.

Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods, May 30 (epub).

McCormick, M.B., R.M. Tamimi, L. Snider, A. Asakura, D. Bergstrom, and S.J. Tapscott. 1996. neuroD2 and neuroD3: Distinct expression patterns and transcriptional activation potentials within the neuroD gene family. Molecular and Cellular Biology 16: 5792–800.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science, May 1 (epub).

Otto, S.J., S.R. McCorkle, J. Hover, C. Conaco, J.J. Han, S. Impey, G.S. Yochum, J.J. Dunn, R.H. Goodman, and G. Mandel. 2007. A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. Journal of Neuroscience 27(25):6729–39.

Pance, A., F.J. Livesey, and A.P. Jackson. 2006. A role for the transcriptional repressor REST in maintaining the phenotype of neurosecretory-deficient PC12 cells. Journal of Neurochemistry. 99(5):1435–44.

Patel, P.D., D.A. Bochar, D.L. Turner, F. Meng, H.M. Mueller, and C.G. Pontrello. 2007. Regulation of tryptophan hydroxylase-2 gene expression by a bipartite RE-1 silencer of transcription/neuron restrictive silencing factor (REST/NRSF) binding motif. Journal of Biological Chemistry 282(37):26717–24.

Peisach, E., and C.O. Pabo. 2003. Constraints for zinc finger linker design as inferred from X-ray crystal structure of tandem Zif268-DNA complexes. Journal of Molecular Biology 330(1):1–7.

Phatnani, H.P., and A.L. Greenleaf. 2006. Phosphorylation and functions of the RNA polymerase II CTD. Genes and Development 20(21):2922–36

Poy, M.N., L. Eliasson, J. Krutzfeldt, S. Kuwajima, X.S. Ma, P.E. MacDonald, B. Pfeffer, T. Tuschl, N. Rajewsky, P. Rorsman, and M. Stoffel. 2004. A pancreatic islet-specific microRNA regulates insulin secretion. Nature 432:226–30.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, S. Jones. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods. (8):651–7.

Schoenherr, C.J., and D.J. Anderson. 1995. The Neuron-Restrictive Silencer Factor (NRSF) — a Coordinate Repressor of Multiple Neuron-Specific Genes. Science 267:1360–63.

Schoenherr, C.J., A.J. Paquette, and D.J. Anderson. 1996. Identification of potential target genes for the neuron-restrictive silencer factor. Proceedings of the National Academy of Sciences of the United States of America 93:9881–86.

Scholl, T., M.B. Stevens, S. Mahanta, and J.L. Strominger. 1996. A zinc finger protein that represses transcription of the human MHC class II gene, DPA(1,2). Journal of Immunology 156:1448–57.

Schones, D.E., K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132(5):887-98.

Sempere, L.F., S. Freemantle, I. Pitha-Rowe, E. Moss, E. Dmitrovsky, and V. Ambros. 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biology 5(3):R13.

Shannon, M., A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Research 13:1097–110.

Shimojo, M., J.H. Lee, and L.B. Hersh. 2001. Role of zinc finger domains of the transcription factor neuron-restrictive silencer factor/repressor element-1 silencing transcription factor in DNA binding and nuclear localization. Journal of Biological Chemistry 276(16):13121–6.

Singh SK, Kagalwala MN, Parker-Thornburg J, Adams H, Majumder S. 2008. REST maintains self-renewal and pluripotency of embryonic stem cells. Nature 453(7192):223–7.

Sommer, L., Q. Ma, and D.J. Anderson. 1996. Neurogenins, a novel family of atonal-related bHLH transcription factors, are putative mammalian neuronal determination genes that reveal progenitor heterogeneity in the developing CNS and PNS. Molecular and Cellular Neuroscience. 8: 221–41.

Stark, A., M.F. Lin, P. Kheradpour, J.S. Pedersen, L. Parts, J.W. Carlson, M.A. Crosby, M.D. Rasmussen, S. Roy, A.N. Deoras, J.G. Ruby, J. Brennecke; Harvard FlyBase curators; Berkeley Drosophila Genome Project, E. Hodges, A.S. Hinrichs, A. Caspi, B. Paten, S.W. Park, M.V. Han, M.L. Maeder, B.J. Polansky, B.E. Robson, S. Aerts, J. van Helden, B. Hassan, D.G. Gilbert, D.A. Eastman, M. Rice, M. Weir, M.W. Hahn, Y.

Park, C.N. Dewey, L. Pachter, W.J. Kent, D. Haussler, E.C. Lai, D.P. Bartel, G.J. Hannon, T.C. Kaufman, M.B. Eisen, A.G. Clark, D. Smith, S.E. Celniker, W.M. Gelbart, and M. Kellis. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450(7167):219–32.

Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America 101: 6062–67.

Tantravahi, J., M. Alvira, and E. Falck-Pedersen. 1993. Characterization of the mouse beta maj globin transcription termination region: a spacing sequence is required between the poly(A) signal sequence and multiple downstream termination elements. Molecular and Cellular Biology. (1):578-87.

Thiel, G., M. Lietz, and M. Cramer. 1998. Biological Activity and Modular Structure of RE-1-silencing Transcription Factor (REST), a Repressor of Neuronal Genes. Journal of Biological Chemistry 273: 26891-26899.

Ule, J., A. Ule, J. Spencer, A. Williams, J.S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B.R. Zeeberg, D. Kane, J.N. Weinstein, J. Blume, and R.B. Darnell. 2005. Nova regulates brain-specific splicing to shape the synapse. Nature Genetics 37:844–52.

Wagner, S., M.A. Hess, P. Ormonde-Hanson, J. Malandro, H.P. Hu, M. Chen, R. Kehrer, M. Frodsham, C. Schumacher, M. Beluch, C. Honer, M. Skolnick, D. Ballinger, and B.R. Bowen. 2000. A broad role for the zinc finger protein ZNF202 in human lipid metabolism. Journal of Biological Chemistry 275:15685–90.

Wasserman, W.W., and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics 5: 276–87.

Watanabe, Y., S. Kameoka, V. Gopalakrishnan, K.D. Aldape, Z.Z. Pan, F.F. Lang, and S. Majumder. 2004. Conversion of myoblasts to physiologically active neuronal phenotype. Genes and Development 8(8):889–900.

Weber, M.J. 2005. New human and mouse microRNA genes found by homology search. FEBS Journal 272:59–73.

Wei, C.L.,Q. Wu, V.B. Vega, K.P. Chiu, P. Ng, T. Zhang, A. Shahab, H.C. Yong, Y. Fu, Z. Weng, J. Liu, X.D. Zhao, J.L. Chew, Y.L. Lee, V.A. Kuznetsov, W.K. Sung, L.D. Miller, B. Lim, E.T. Liu, Q. Yu, H.H. Ng, and Y. Ruan. 2006. A global map of p53 transcription-factor binding sites in the human genome. Cell 124(1):207–19.

Westbrook, T.F., E.S. Martin, M.R. Schlabach, Y.M. Leng, A.C. Liang, B. Feng, J.J. Zhao, T.M. Roberts, G. Mandel, G.J. Hannon, R.A. DePinho, L. Chin, and S.J. Elledge. 2005. A genetic screen for candidate tumor suppressors identifies REST. Cell 121:837–848.

Westbrook ,T.F., G. Hu, X.L. Ang, P. Mulligan, N.N. Pavlova, A. Liang, Y. Leng, R. Maehr, Y. Shi, J.W. Harper, and S.J. Elledge. 2008. SCFbeta-TRCP controls oncogenic transformation and neural differentiation through REST degradation. Nature. 452(7185):370–4.

Wilhelm B.T., S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C.J. Penkett, J. Rogers, and J. Bähler. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, (epub).

Wold, B., and R.M. Myers. 2008. Sequence census methods for functional genomics. Nature Methods. (1):19–21.

Xie, X., J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434(7031):338–45.

Yeo, M., S.K. Lee, B. Lee, E.C. Ruiz, S.L. Pfaff, and G.N. Gill. 2005. Small CTD phosphatases function in silencing neuronal gene expression. Science 307:596–600.

Ying, S.Y., and S.L. Lin. 2004. Intron-derived microRNAs - fine tuning of gene functions. Gene 342:25–28.

Zhao, Y.U., H.Z. Sheng, R. Amini, A. Grinberg, E. Lee, S.P. Huang, M. Taira, and H. Westphal. 1999. Control of hippocampal morphogenesis and neuronal differentiation by the LIM homeobox gene Lhx5. Science 284:1155–58.

Zhang, C., Z. Xuan , S. Otto, J.R. Hover, S.R. McCorkle, G. Mandel, and M.Q. Zhang. 2006. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Research 34: 2238–46.

Zeitlinger, J., A. Stark, M. Kellis, J.W. Hong, S. Nechaev, K. Adelman, M. Levine, and R.A. Young. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nature Genetics 39(12):1512-6.