Statistical Methods for Gene Differential Expression Analysis of RNA-Sequencing

Thesis by

Lynn Yi

In Partial Fulfillment of the Requirements for
the degree of
Ph.D. in Biology

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2019
(Defended April 15, 2019)

Lynn Yi
ORCID: 0000-0003-4575-0158

ACKNOWLEDGEMENTS

The Pachter Lab has been my work family for the past four years. Lior Pachter's passion for scientific research and dedication to use his work for humanistic good served as a reminder for why and how science should be pursued. Members who introduced me to the Pachter Lab at Berkeley and provided wisdom throughout my PhD include Harold Pimentel, Lorian Schaeffer, Shannon McCurdy, Vasilis Ntranos, Nicolas Bray and Páll Melsted. At Caltech, it was a pleasure to work and watch the lab grow with Jase Gehring, Lisa Sledd and Valentine Svensson. My PhD committee members, David Chan, Matt Thomson and Venkat Chandrasekaran, have dedicated many hours of personal time advising me through my years at Caltech, without which I would be adrift.

Completing my PhD would not have been possible without Elizabeth Ayala and Tess Legaspi, who have taken care of not only me but also hundreds of other graduate students at Caltech. Bonnie Sturner helped me develop the mental fortitude necessary for the pursuit of this work. The Broads of Broad were sisters with whom the struggles of research could be shared openly. Knowing I could always call Amy Li, Mandy Xi, Jeff Ning and Daniel Ichwan meant I could get through the toughest of times.

My parents Jizu Yi and Lihua Chen have been my first role models of hard work, perseverance, and love. There was not a sacrifice that was too much for them if it could bring me additional opportunity. Watching David and Joycena Yi grow into upright young adults is a source of joy and inspiration. Over the past several years, Wei and Ming Chi have become my second set of parents, and Emily and Eric Chi, another pair of siblings, thereby doubling the excitement and warmth in my life. My husband Brian Chi's love and care, coupled with his level-headedness and silly sense of humor, keep me up and running every day. These people are my raison d'être.

ABSTRACT

RNA-Sequencing ("RNA-Seq") is performed to measure gene expression, often to ask the question of what genes are differentially expressed across various biological conditions. Statistical methods have been used to model RNA-Seq quantifications in order to determine differential expression, and have traditionally be divided into gene-level methods and transcript-level methods. There has been little attempt to connect the statistical divide, although transcript expression and gene expression are biologically inextricably linked. In this thesis, we provide a case study of a comparative differential expression analysis, demonstrating that many differential expression events happen on the isoform-level, and that performing an analysis using only summarized gene quantifications would fail to capture these events. Furthermore, we develop statistical methods that unify the transcript-level and gene-level analysis. In bulk RNA-Seq, by using p-value aggregation methods, we are able to translate transcript-level results into gene-level results under a unified framework. For single cell RNA-Seq, we propose using multiple logistic regression, leveraging the high dimensionality of the data in order to determine if the transcript quantifications pertaining to a gene are able to constitute a linear discriminant for cell type. This method combines differential transcript expression analysis and differential gene expression analysis into a unified framework which we call "gene differential expression." Lastly, we demonstrate that our methods could be used on transcript compatibility counts instead of transcript quantifications in order to bypass ambiguous read assignment and improve accuracy. We show that transcript compatibility counts obtained via transcriptome pseudoalignment are comparable in quantification accuracy to quantifications from genome alignment methods.

PUBLISHED CONTENT AND CONTRIBUTIONS

Yi L, Pimentel H, Pachter L. Zika infection of neural progenitor cells perturbs transcription in neurodevelopmental pathways. *PLoS One*. 2017;12(4). doi:10.1371/journal.pone.0175744

L.Y. performed the RNA-Seq data reanalysis and participated in the writing of this manuscript.

Yi L, Pimentel H, Bray NL, Pachter L. Gene-level differential analysis at transcript-level resolution. *Genome Biol*. 2018;19(1):53. doi:10.1186/s13059-018-1419-z

L.Y. performed the proof-of-concept analysis for methods development, performed the simulations, and participated in the writing of this manuscript.

Ntranos V, Yi L, Melsted P, Pachter L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat Methods*. 2019;16(2):163-166. doi:10.1038/s41592-018-0303-9

L.Y. joined in method development, developed a single cell RNA-Seq simulation pipeline for benchmarking, and participated in the writing of this manuscript.

Yi L, Liu L, Melsted P, Pachter L. A direct comparison of genome alignment and transcriptome pseudoalignment. *bioRxiv*. October 2018:444620. doi:10.1101/444620

L.Y. managed the software development, performed computational analysis, and participated in the writing of this manuscript.

FUNDING SOURCES

I have been fortunate to have received financial support from the following grants and fellowships for the scientific research documented by this thesis: NIH NIGMS GM008042 through the UCLA-Caltech MSTP, NIH T32 GM07616 through the Caltech Biology Department, NIH R012017-0569 through the Brain Initiative, the Lee Ramo Fund, and the Walter and Sylvia Treadway Fund.

TABLE OF CONTENTS

NOMENCLATURE

Differential Gene Expression: gene quantifications are modeled in order to determine differential expression

Differential Transcript Expression: transcript quantifications are modeled in order to determine differential expression

Differential Transcript Usage: transcript quantifications are modeled to determine whether transcript allocation across the gene is differential

Equivalence Class (EC): the set of transcripts that a read is compatible with

Pseudoalignment: the process by which the set of transcripts a read is compatible with is determined

Transcript compatibility counts (TCCs): the counts associated with equivalence classes for a RNA-seq sample

Bulk RNA-Seq: method where the RNA from many cells are sequenced in aggregate

Single cell RNA-Seq (scRNA-Seq): method by which single cells are isolated and its mRNA sequenced, and the cell of origin of a read can be identified

Chapter I
Introduction

This page intentionally left blank.

Why study gene expression

Gene expression is the master regulator of biology. In humans, for example, most tissues share the same genome, and yet a diversity of tissues and processes can be achieved and regulated due to differential gene expression. Measuring gene expression is vital for studying tissue differentiation, attributing disease risk to gene expression, and studying the genomic landscape of diseases, drugs, or other perturbations.

Methods for Measuring Gene Expression

There exists a myriad of molecular biology techniques to measure gene expression. Reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) first reverse transcribes RNA into cDNA, and then uses fluorescently-coupled polymerase-based amplification with specifically designed primers to detect the presence of a target cDNA. The number of cycles needed to produce a fluorescent signal provides a quantitative measure of cDNA concentration. qPCR is a targeted, low throughput, yet sensitive method for measuring gene expression.

In the 1980's, microarrays were developed as a high-throughput method that can quantify the expression of many genes in one experiment. While it was initially developed to study DNA structural variation, its most common use has been to measure RNA expression. In this technology, DNA oligonucleotide probes are printed and bound to distinct regions of a chip, not unlike the printing of circuits onto a silicon chip. Hybridization of a sample on the chip allows target cDNA to be captured by the bound oligonucleotide probes. A fluorescent or chemiluminescent signal is coupled to hybridization, allowing for readout by photodetection, usually by a specialized machine. Signal from different regions of the chip correspond to hybridization of different probes, and quantification of luminescence allows for quantification of target RNA in the sample. Popular commercial suppliers of microarray technology include Affymetrix and Illumina, i.e., Affymetrix GeneChip and Illumina BeadChip.[1,2]

Next Generation Sequencing

In the early 2000's, next-generation sequencing (NGS) came into the arena. It would not be an understatement to call the advent of NGS a revolution for molecular biology. Prior to NGS, Sanger sequencing was possible at low-throughput and high cost. Sequencing the first human genome, completed in 2001 with Sanger sequencing, was estimated to have cost of up to 1 billion dollars. The concept of quantifying gene expression is changed with NGS. Compared to microarray, which specifies a set of sequences and asks whether cDNA containing these subsequences is expressed, NGS allows one to sample a pool of cDNA. Each read from NGS now corresponds to a sequenced molecule of DNA, which can literally be counted up for each gene. [3]

There are several platforms for sequencing. Oxford Nanopore is a single-molecule method that identifies bases from measuring current while a motor protein threads a string of bases through a synthetic biologic nanopore. It allows for long length (median of 6kb, maximum of 60kb) but with high error rates (5-10%) and low throughput (on the order of 10K reads), limiting its use in genome-wide studies. Ion Torrent uses clonal amplification of DNA, measuring pH changes due to DNA extension in order to identify the base being added. [4] [5]

Illumina sequencing by far the most popular method for genome-wide sequencing studies. It uses clonal amplification of DNA with fluorescently labeled dNTPs. Fluorescent imaging after each round of base extension reveals the base being that has been added. Illumina is popular due to its low error rates (<1% across its sequencing platforms) with errors being primarily substitutions, as well as its speed and high throughput. In contrast to Oxford Nanopore sequencing, Illumina sequencing has a shorter maximum read length of a few hundred base pairs. [6] [7] Library preparation for Illumina sequencing involves reverse transcribing RNA to cDNA as well as adaptor ligation of pre-specified Illumina adaptors for recognition by the sequencer. Ultimately, the choice in sequencing platform will depend on a combination of factors, including cost, throughput, read length, error rate, and speed. For the purpose of this thesis, we will be discussing computational methods for analyzing the short reads such as those produced by Illumina sequencers.

The cost per read is becoming cheaper, the time to sequence is shorter, the maximum number of base pairs that could be sequenced per experiment is growing, and empirically, the number of reads per experiment is growing exponentially. [8] The decreasing cost of sequencing (biological Moore's law) is also coupled to the decreasing cost of computation (siliconic Moore's law). Growing sequencing datasets motivate the need for algorithmic and statistical methods that scale with the number of reads.

Single cell RNA-Seq is a technique that was developed in the last decade and has been quickly growing in popularity. In bulk RNA-Seq, the lysate generated from many cells are sequenced. A bulk RNA-Seq sample of brain tissue, for example, would include sequenced mRNA from neuron, glia, and connective tissue, and gene expression could only be suggested as a mixed aggregate unless the cell types are sorted and purified prior to RNA-Seq. In contrast, single cell RNA-Seq allows one to identify the mRNA corresponding to individual cells, making a scRNA-Seq a nature method to assay gene expression of heterogeneous biological systems, such as cancer, developmental biology and differentiation, and complex tissues. scRNA-Seq is currently being used as the primary method for developing a human brain atlas, where the composition of cell types in various regions of the brain could be identified and mapped.[9]

There are several methods for performing scRNA-Seq, including isolating single cells into wells to be individually processed for sequencing (SMART-seq) or generating single cell suspensions using oil droplets where cell barcoding is performed prior to pooled library preparation (10x). In addition to cell barcodes, methods may also employ oligonucleotides as unique molecular identifiers (UMI) to identify sequences that are PCR amplified from the same original molecule. When choosing the specific scRNA-Seq protocol, common considerations include the desired number of cells, the sequencing depth per cell, and cost of experiment. Like those of other sequencing methods, scRNA-Seq datasets are growing in size, but in the number of cells as well as the total number of reads, motivating the need for computational methods to analyze the high dimensional datasets produced.

Computational Analysis of RNA-Seq

Here, we present an overview of the conventional computational pipeline for performing differential expression analysis of RNA-Seq. In the first section, we use a discussion of gene-level analysis of bulk RNA-Seq to introduce commonly-used methods for quantification and differential expression. The second section discusses the unique challenges and several solutions for analyzing bulk RNA-Seq on a transcript isoform level.

For each of the first two sections (gene level analysis and transcript level analysis), first we discuss the process for obtaining quantifications from a sequencing experiment. This involves taking a set of sequences generated from the sequencer and obtaining quantifications for genomic features, such as gene counts. In order to make conclusions about gene expression, such as whether a gene is up or downregulated between conditions, these quantifications must be modeled to in order to account for inherent variability. So, we first discuss methods for quantification and then methods for differential expression analysis.

In the last section, we present a brief discussion of methods for the analysis of single cell RNA-Seq (scRNA-Seq). While the quantification of single cell RNA-Seq may also involve similar alignment or pseudoalignment of reads as in bulk RNA-Seq, the unique challenges of barcode and unique molecular identification (UMI) mapping, as well as those of analyzing a high dimensional data matrix, makes scRNA-Seq a rich topic of its own. Indeed, methods such as imputation for dropped-out expression, dimensionality reduction of the high dimensional data matrix, and clustering to identify cell types have been developed for scRNA-Seq uniquely.

While the primary purpose is to introduce the current statistical methods for RNA-Seq analysis, we will also discuss limitations of existing methods and, where appropriate, where the work proposed in this thesis improves upon these limitations.

RNA-Seq Analysis on the Gene Level

Alignment to Reference Genome and Quantification

In the conventional pipeline for analyzing RNA-Seq, reads are first be aligned to the reference genome corresponding to the species of interest. The quality of the reference genome is highly species dependent. Common model organisms, including human, mouse, C. elegans and Drosophila, have easily-accessible, accurate reference genomes, whereas obtaining reference genome for an uncommonly studied organism could be challenging. However, even these generally accurate references genomes could be unreliable or uninformative at loci with high amounts of genetic diversity and/or structural variation, such as the *HLA* and *ABO* loci in humans. Special attention should be paid if these regions are of interest to the scientific investigation. In addition to the genomic sequence itself, an annotation that specifies the gene and/or transcript tracks on this genomic sequence is necessary for RNA-Seq analysis. Common annotations used include those provided by Ensembl, RefSeq and UCSC.

Bowtie/Bowtie 2 is the classic aligner for short reads such as those produced by Illumina sequencing. It uses the Burrows-Wheeler transform, a reversible string compression algorithm, to build a space-efficient index of the genome on which subsequences (i.e. reads) could be found efficiently.

While DNA sequencing alignment could be simply formulated as an error-tolerant substring matching question, the problem of RNA-Sequencing alignment has an additional layer which is to account for splice junctions. Because of the splicing of mRNAs, a contiguous mRNA subsequence may not correspond to contiguous subsequence in the genome, and aligners for RNA-Seq analysis must specifically allow for gaps. Bowtie2[10] was extended from Bowtie to support gapped alignment. TopHat/TopHat2[11] identifies splice junctions using alignments from Bowtie. Tophat/TopHat2 was then deprecated by its authors for HISAT/HISAT2[12], a splice-aware aligner that provides further speed improvements through an extension of Burrows Wheeler transform to graphs. STAR[13] is another splice-aware aligner, independent of the Burrows Wheeler transform, that was built using the concept of a maximal mappable prefix, and is extendable to long reads such as those produced by Oxford Nanopore.

After alignment to the genome, reads are then assigned to genes in order to determine gene expression. Methods including HTSeq[14] and featureCounts[15] sum up the reads aligning to genomic intervals defined by an annotation, i.e. all exons corresponding to a gene, to produce gene counts. When these methods are used to aggregated counts across multiple exons to a single gene count, they lose information about distinct transcript isoforms, which are unique in sequence and could have different biological functions.

Differential expression analysis

The experimenter performing RNA-Seq often has conditions he/she is wishing to measure and compare gene expression under. For example, in Chapter 1, we analyze a dataset from an experiment where the goal was to discover genes whose expressions were affected by Zika infection. The experimenters infected human neuroprogenitor cells with the Zika virus and with a mock control and performed a *differential gene expression analysis* on the dataset.[16] Such experimental setups for RNA-Seq are common and motivates the need for a systematic and statistically principled approach to performing genome-wide differential expression analysis.

RNA-Seq is not a deterministic process and is inherently noisy. Biologically, different handling and preparation of the sample may result in expression differences in the samples. Furthermore, there is stochasticity in gene expression, even at "steady states."[17] [18]Technically, the process of library preparation and sequencing, where a subsample of the biologically expressed RNA molecules are selected to be reverse transcribed, adaptor ligated, amplified and eventually sequenced, is a multi-step procedure where variance is introduced at every step. Lastly, the computational procedure of read alignment and quantification introduces inferential variance, for example as a result of alignment error or as a result of short, ambiguous reads that could map to more than one gene. Statistical methods for differential expression analysis must therefore model variance accurately in order to determine the statistical significance of an observed difference in gene expression between conditions.

The procedure can be outlined as follows:

1. After alignment and read counting, we arrive at a gene count or abundance matrix of dimension *samples* by *genes*. The expression matrix is first normalized to account for the differences in sequencing depth across samples. In addition to the expression table, another table of dimension *samples* by *conditions* describes the covariates for each sample.

2. A model on gene expression is fit for each gene. Two common parameterizations for RNA-Seq include a linear model on log-transformed counts and a negative binomial model on discrete counts. limma[19] and sleuth[20] use a linear model; edgeR[21] and DESeq[22]/DESeq2[23] use the negative binomial model. The negative binomial model can be motivated as a Poisson with an additional parameter to account for overdispersion.

This model, linear or negative binomial, is fit for each gene g:

$$Y_{ig} \sim \sum_{j \in covariates} \beta_j * X_{ij} \quad ,$$

where Y*ig* is the expression of gene *g* in sample *i*, and *Xi* is the (often indicator) variable corresponding to covariate *j* for sample *i*. Covariates may be included in the model to account for confounders and batch effects, even if they are not tested for effect on expression. To test the significance of differential expression for a specific covariate *j*, a likelihood ratio test could be performed comparing this alternate model to the null model where covariate j does not have an effect on gene expression, i.e. Bj is excluded from model. Another option is the Wald test, which tests the null hypothesis that $Bj = 0$.

3. After obtaining p-values for each gene, multiple testing correction should be performed. The Benjamini-Hochberg correction to control false discovery rate is standard. Quantile-quantile plots (QQ plots) can also be used to examine whether the p-values are distributed as expected, i.e. uniform between 0 - 1 as under the null hypothesis.

One known issue of RNA-Seq differential expression is that sample sizes are often limited. With only two or three biological replicates, the sample variance is unreliable and could lead to false positives. The solution that most methods (including limma, edgeR, DESeq2, and sleuth) have adopted is to perform shrinkage on the genome-wide mean variance relationship, thereby sharing information across genes, i.e. the variances of genes with similar mean expressions are shrunk towards their mean variance. The shrinkage estimator of variance is then used for statistical testing, instead of the sample variance. sleuth is unique in that it distinguishes technical from biological variance and performs shrinkage only on the biological component of variance, while using an alternative approach to compute technical variance (discussed below).

RNA-Seq Analysis on the Transcript Level

One feature of molecular biology grossly ignored in the gene-level analysis is that of transcript isoforms. Transcript isoforms are different mRNAs that are transcribed from the same genomic locus, but have different sequences due to differential splicing leading to inclusion/exclusion of different exons. A gene-level analysis fails to capture the amount of regulation that is happening through biological processes such as alternative 5'/3'UTRs, differential splicing, and up/down regulation of specific transcripts. In Chapter II, we showcase a differential transcript expression analysis of RNA-Seq of Zika infection, demonstrating that many genes have individual transcripts that are differentially expressed with opposing effect sizes. The same genes are likely to be missed in the differential gene expression analysis, which we compare to our differential transcript expression analysis.

Nonetheless, performing analysis on the transcript-level has its challenges. First, transcript quantification is not trivial. There may be many transcripts per gene that share exons and overlap in sequence. Reads that may map unambiguously to a genomic locus could correspond to multiple transcript isoforms. Furthermore, differential transcript expression comes with its own set of challenges, including the fact transcript quantifications contain greater inferential variance due to ambiguous read assignment. In this section, we address some methods that have been developed to handle the challenges of obtaining transcript quantifications and performing differential transcript expression analysis.

Transcript Quantification from Genome Alignments

Methods have been developed specifically to obtain transcript isoform quantifications after genome alignment. In general, a probabilistic model for RNA sequencing is required for assigning genomic alignments to transcripts. For example, this model may assume uniform sequencing across mRNAs and uniform coverage along each mRNA, or it may consider sequencing bias (non-uniform sequences at the start and end of fragments) and positional bias (non-uniform coverage along the mRNA). An assumption about the distribution of RNA-Seq reads, even if that is of uniform, unbiased sequencing along the expressed mRNA molecule, is necessary for choosing how to distribute read alignments to transcripts. RSEM[24] is a method that uses the expectation maximization algorithm to assign ambiguous counts to transcripts under a probabilistic model of counts. Cufflinks performs isoform discovery alongside isoform quantification by solving for the minimum number of isoforms required to explain the observed reads.

Pseudoalignment: An Alternative to Genome Alignment

Aligning RNA-Seq reads to the genome is computationally expensive, as the index of the reference genome must be stored in memory and algorithmically scales with the length of the reference genome and the read.[25] Instead of alignment, a class of "pseudoalignment" methods were developed in as a light-weight alternative. Instead of seeking to locate the genomic site of origin of an RNA-Seq read, it looks for the set of transcripts a read may have originated from. This set of transcripts is the "equivalence class" of the read, in that two reads are defined to be in equivalence if they are compatible with the same set of transcripts.

The pseudoaligner Sailfish[26] segments each read into its constituent kmers and hashes each kmer to obtain that kmer's equivalence class. Following pseudoalignment, kmer counts are assigned to transcripts with the EM algorithm. kallisto[27] also uses kmer hashing, but employs a de Brjuin graph to map each read (instead of each kmer) to an equivalence classes, thereby improving the accuracy of Sailfish while retaining the value of its simplistic approach.

There is an enormous speed-up provided by and less computational demand required for performing pseudoalignment instead of alignment, allowing RNA-Seq to be quantified in several minutes on a personal laptop instead of several hours on a dedicated server. In Chapter V, we show that the quantification accuracy of pseudoalignment methods is comparable to that of alignment methods.

Transcript Compatibility Counts (TCCs)

While originally an intermediate product of pseudoalignment alignment methods generated prior to read assignment, transcript compatibility counts ("TCCs"), the counts associated with each equivalence class, have been explored as a useful data matrix of its own. TCCs are free of any inferential variance and constitute summary statistics for RNA-Seq quantification. In the differential expression method sleuth (further discussed below in "Differential transcript expression analysis"), TCCs allow for efficient bootstrap estimation of inferential variance that would otherwise be computationally infeasible. In Ntranos et al.[28], TCCs are used as the data matrix instead of gene quantifications to cluster single cell RNA-Seq datasets. In Chapters III and IV, we show using TCCs instead of transcript counts in performing differential expression analysis improves accuracy.

Differential transcript expression analysis

After transcript quantification, a differential transcript expression analysis ("DTE") is performed to investigate which transcripts have statistically different abundances across conditions. Traditionally, transcript quantifications are used as input to differential expression methods used for differential gene expression ("DGE"), and there has been little thought in modeling DTE differently from DGE.

However, there are substantial differences between performing DGE and performing DTE that have substantial statistical implications. Given that eukaryotes often have more transcripts than genes (~170,000 transcripts compared to 30,000 genes in human genome), there is a greater burden to pass multiple testing correction when performing DTE. Furthermore, because transcript quantification compared to gene quantification involves more specific read assignment to transcripts with some overlap in sequence, there is greater inferential variance. The number of transcripts in a gene and the extent of shared sequence across transcripts, for instance, have a large effect on the quantification certainty of transcripts within a gene, and would be difficult to determine *a priori* for each gene.

Cuffdiff2[29] performs differential expression of transcripts by modeling transcript counts with a beta negative binomial distribution. The beta distribution is used to capture additional inferential variance from ambiguous transcript quantification in addition to the over-dispersion modeled by the negative binomial. Cuffdiff2 performs model fitting by calculating the empirical covariance matrix from read assignment to transcripts.

sleuth[30] uses another strategy for calculating inferential variance. Instead of parametric modeling, sleuth performs bootstrapping to estimate inferential variance resulting from ambiguous read assignment. sleuth then performs shrinkage on the variance component that is not captured by the estimated inferential variance. What would otherwise be a computationally infeasible task of bootstrapping, aligning and quantifying millions of RNA-Seq reads was made feasible by leveraging the concept of TCCs developed by kallisto. Bootstrapping TCCs is equivalent to bootstrapping reads, and quantification of bootstrapped TCCs with the EM algorithm is fast and requires no additional alignment on the bootstrapped samples. In simulations, sleuth has much more conservative and accurate false discovery rates compared to other differential expression methods, suggesting that accurate estimation of inferential variance is an important component of differential transcript expression.

Differential Expression for Single Cell RNA-Seq

Single cell RNA-Seq (scRNA-Seq) represents a new frontier in experimental and computational methods. On the analysis side, although naively it could be thought of as an extension of bulk RNA-Seq with more samples via more cells, closer examination reveals unique challenges associated with scRNA-Seq.

The maximum of cells that can be assayed in one scRNA-Seq experiment is exponentially and is now on the order of millions, leading to a high dimensional data matrix.[31] Compared to bulk RNA-Seq where there may only be a few samples that are usually deeply sequenced, scRNA-Seq datasets often have many cells that are usually shallowly sequenced. Furthermore, due to the lower amounts of starting RNA in single cells, scRNA-Seq is thought of as a "noisier" experiment compared to bulk RNA-Seq. For example, a gene may have no mapped reads in a subset cells of a cell type that are known to express that gene, a concept referred to as "dropout."

Single cell methods commonly model dropout in order to perform differential expression. Instead of a negative binomial distribution on counts, many methods use a zero-inflated negative binomial to better model the additional zero expression values. In SCDE[32], a Bayesian approach is used to fit an error model to each cell in order to model per-cell in addition to per-gene probability of dropout. In MAST[33], each cell is now parameterized by its detection rate, the proportion of genes that are expressed in the cell. In Monocle[34], instead of the negative binomial, a Tobit model is used to account for zero-inflation. However, whether dropout exists as an independent phenomenon from merely shallow sequencing is disputed.[35]

While the measurement of a specific gene in a specific cell may be unreliable, the dataset provides information in aggregate. For example, common scRNA-Seq pipelines perform smoothing and dimensionality reduction[36], identify distinct cell types through unsupervised clustering, and then perform differential expression between clusters to identify markers of these cell types.[37] Furthermore, as the scRNA-Seq datasets increase in number of cells, machine learning methods that leverage this increasing dimensionality are being adopted.[38] In Chapter IV, we show that logistic regression can be used for performing gene differential expression, an example of leveraging the number of cells to fit higher dimensional models.

Outline of Chapters

This thesis discusses novel methods we developed in order to perform RNA-Seq gene differential expression, which unifies differential gene expression and differential transcript expression. By using transcript quantifications, our methods are sensitive to transcript-level differential events, while being able to summarize our statistical results to the gene-level. However, not only is our method more sensitive and accurate for detecting differential transcript expression, it can also detect differential gene expression. We present two methods for performing gene differential expression: one for bulk RNA-Seq and one for scRNA-Seq. Furthermore, we show that in both methods, we can use transcript compatibility counts (TCCs) instead of transcript counts, from which we can obtain even more accurate results from reducing inferential variance due to ambiguous read assignment to isoforms.

Chapter II is a case-study that compares a gene-level analysis with a transcript-level analysis, showcasing that there are many isoform-level events that are missed when RNA-Seq quantifications are summarized to the gene level. It motivates the significance of detecting differential transcript-level events in biological systems.

Chapter III provides a more formal discussion of the statistical drawbacks of performing differential gene expression analysis using gene quantifications. In addition to missing differential transcript expression events, a gene-level analysis is fraught on its premise of constructing gene counts, which when calculated from transcript counts leads to a distortion of variance. We provide a p-value aggregation method that allows one to perform differential transcript expression and then obtain gene-level results that are statistically unified and coherent. We show that performing differential expression on TCCs instead of transcript counts and aggregating to the gene-level increases accuracy.

Chapter IV discusses a differential expression method for single cell RNA-Seq. It leverages the large number of cells sequenced in RNA-Seq in a higher dimensional model corresponding to the dimension of the number of transcript isoforms. While it is a distinct method from the method discussed in Chapter III for bulk RNA-Seq, it conceptually achieves the same results: gene-level results that are consistent with the transcript-level results and that is sensitive to transcript-level events. We again show that using TCCs instead of transcript counts with this approach can lead to more accurate results. Furthermore, in this chapter's supplement, we present a more formal argument for how our method unifies differential gene expression ("DGE") and differential transcript expression ("DTE") into a common test.

Chapter V discusses a comparison of pseudoalignment and alignment. It demonstrates that the two suites of methods are comparable in accuracy in their transcript and gene quantifications. Furthermore, we discuss the new tools we developed new tools to convert between the quantifications produced by the two approaches, i.e. producing TCCs from genome alignments and producing alignment-like visualization from pseudoalignment.

References

[1] Bumgarner R. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol*. 2013;Chapter 22:Unit 22.1. doi:10.1002/0471142727.mb2201s101

[2] Pirrung MC, Southern EM. The genesis of microarrays. *Biochem Mol Biol Educ*. 2014;42(2):106-113. doi:10.1002/bmb.20756

[3] Reuter JA, Spacek D V, Snyder MP. High-throughput sequencing technologies. *Mol Cell*. 2015;58(4):586-597. doi:10.1016/j.molcel.2015.05.004

[4] Reuter JA, Spacek D V, Snyder MP. High-throughput sequencing technologies. *Mol Cell*. 2015;58(4):586-597. doi:10.1016/j.molcel.2015.05.004

[5] Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J Microbiol*. 2016;56(4):394-404. doi:10.1007/s12088-016-0606-4

[6] Reuter JA, Spacek D V, Snyder MP. High-throughput sequencing technologies. *Mol Cell*. 2015;58(4):586-597. doi:10.1016/j.molcel.2015.05.004

[7] Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J Microbiol*. 2016;56(4):394-404. doi:10.1007/s12088-016-0606-4

[8] NIH NHGRI. DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI). https://www.genome.gov/27541954/dna-sequencing-costs-data/. Accessed March 31, 2019.

[9] Ecker JR, Geschwind DH, Kriegstein AR, et al. The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron*. 2017;96(3):542-557. doi:10.1016/j.neuron.2017.10.007

[10] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923

[11] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120

[12] Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511-515. doi:10.1038/nbt.1621

[13] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

[14] Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169. doi:10.1093/bioinformatics/btu638

[15] Liao Y, Smyth GK, Shi W. Sequence analysis featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

[16] Tang H, Hammack C, Ogden SC, et al. Zika Virus Infects Human Cortical Neural Progenitors and Attenuates Their Growth. *Cell Stem Cell*. 2016;18(5):587-590. doi:10.1016/j.stem.2016.02.016

[17] Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science*. 2002;297(5584):1183-1186. doi:10.1126/science.1070919

[18] Singer ZS, Yong J, Tischler J, et al. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol Cell*. 2014;55(2):319-331. doi:10.1016/j.molcel.2014.06.029

[19] Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29. doi:10.1186/gb-2014-15-2-r29

[20] Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods*. 2017;14(7):687-690. doi:10.1038/nmeth.4324

[21] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

[22] Anders S, Huber W, Nagalakshmi U, et al. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106

[23] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

[24] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. doi:10.1186/1471-2105-12-323

[25] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;48(3):443-453. doi:10.1016/0022-2836(70)90057-4

[26] Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*. 2014;32(5):462-464. doi:10.1038/nbt.2862

[27] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-Seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

[28] Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol*. 2016;17(1):112. doi:10.1186/s13059-016-0970-8

[29] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2013;31(1):46-53. doi:10.1038/nbt.2450

[30] Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods*. 2017;14(7):687-690. doi:10.1038/nmeth.4324

[31] Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc*. 2018;13(4):599-604. doi:10.10

[32] Kharchenko P V, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods*. 2014;11(7):740-742. doi:10.1038/nmeth.2967

[33] Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015;16(1):278. doi:10.1186/s13059-015-0844-5

[34] Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381-386. doi:10.1038/nbt.2859

[35] Svensson V. Droplet scRNA-seq is not zero-inflated. *bioRxiv*. March 2019:582064. doi:10.1101/582064

[36] Moon KR, Van Dijk D, Wang Z, et al. PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data. doi:10.1101/120378

[37] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. April 2018. doi:10.1038/nbt.4096

[38] Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun*. 2019;10(1):390. doi:10.1038/s41467-018-07931-2

This page intentionally left blank.

Chapter II

Zika infection of neural progenitor cells perturbs transcription in neurodevelopmental pathways

Lynn Yi[1,2], Harold Pimentel[3] and Lior Pachter[1,4]

1. Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA
2. David Geffen School of Medicine, University of California, Los Angeles, CA, USA
3. Department of Genetics, Stanford University, Palo Alto, CA, USA
4. Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA

Corresponding author: lpachter@caltech.edu

This page intentionally left blank.

Abstract

Background

A recent study of the gene expression patterns of Zika virus (ZIKV) infected human neural progenitor cells (hNPCs) revealed transcriptional dysregulation and identified cell cycle-related pathways that are affected by infection. However, deeper exploration of the information present in the RNA-Seq data can be used to further elucidate the manner in which Zika infection of hNPCs affects the transcriptome, refining pathway predictions and revealing isoform-specific dynamics.

Methodology/Principal Findings

We analyzed data published by Tang *et al.* using state-of-the-art tools for transcriptome analysis. By accounting for the experimental design and estimation of technical and inferential variance we were able to pinpoint Zika infection affected pathways that highlight Zika's neural tropism. The examination of differential genes reveals cases of isoform divergence.

Conclusions

Transcriptome analysis of Zika infected hNPCs has the potential to identify the molecular signatures of Zika infected neural cells. These signatures may be useful for diagnostics and for the resolution of infection pathways that can be used to harvest specific targets for further study.

Introduction

As infection with Zika virus (ZIKV) is associated with increasing cases of congenital microcephaly and adult Guillain-Barre Syndrome, a characterization of its pathophysiology becomes crucial. A characterization of the molecular effects of infection may help in the development of fetal diagnostics and can accelerate the identification of genes and pathways critical in disease progression. RNA-Sequencing (RNA-Seq) is an effective technology for probing the transcriptome and has been applied to study the effects of ZIKV infection of human neuroprogenitor cells (hNPCs) [1].

While initial analyses of the data conducted a general survey of transcriptome changes upon infection [1-3], they [1,2] used a method, Cufflinks/Cuffdiff [4], that failed to take advantage of the experimental design used in Tang et. al [1]. They [1-3] also did not examine transcriptome dynamics at the isoform level.

We applied the recently-developed kallisto [5] and sleuth [6] programs to improve the accuracy of quantification and to extract information from the data that was previously inaccessible. We found that sleuth's improved control of false discovery rate [6] resulted in differential transcript and gene lists that are much more specific and that are significantly enriched in neurodevelopmental pathways. They reveal ZIKV's neural tropism and the host's response to viral infection. Furthermore, we demonstrate that the combination of accurate kallisto quantification, assessment of inferential variance and the sleuth response error model allows for the detection of post infection isoform-specific changes that were missed in previous analyses.

The sleuth Shiny app drives a freely available website that allows for reproducibility of our analyses, and provides tools for interacting with the data. This makes the dataset useful for analysis by infectious disease experts who may not have bioinformatics expertise.

Methods

We ran kallisto and sleuth on a total of eight RNA-seq samples of ZIKV-infected and mock-infected hNPCs (GEO: Series GSE78711) (See Table 1 for experimental design and description of samples). We used kallisto to pseudoalign the RNA-seq reads and perform bootstraps, using an index based on the ENSEMBL GRC38 *Homo sapiens* release 85 transcriptome. For single-end read quantification, we used default parameters (kmer size = 31, fragment length = 187 and sd = 70). For each of the eight samples, kallisto quantified transcript abundances and performed 100 bootstraps.

The response error model of sleuth was then used to identify differentially expressed transcripts. Sleuth used the bootstraps performed by kallisto to estimate the inferential variance of each transcript, and an adjusted variance was used to determine differential expression for that transcript. This data set had a unique experimental design, however. For each sequencing library corresponding to a biological sample, Tang et al. performed both paired-end and single-end sequencing. To take advantage of the technical replicates performed by Tang et al., we modified sleuth to perform a weighted average of the inferential variance with the number of fragments sequenced (Table 1).

Principle component analysis of the transcript abundances provided a quick verification of the accuracy of our methods, as the first principle component separated the samples by infection status and the second principle component separated the samples by sequencing method (Fig. 1).

The data analysis pipeline was performed on a laptop and can be repeated using the provided scripts at http://www.github.com/pachterlab/zika/. The kallisto quantifications, the modified version of sleuth, as well as a script for the pipeline, are available on the github. One can use the script to start the Shiny app, which recreates the statistics and figures referenced throughout this paper, along with interactive data visualization tools. Alternatively, the preloaded sleuth Shiny app can be found via http://128.32.142.223/tang16/.

Results

Using a false discovery rate of 0.05 as the threshold for differential expression, we detected 4610 transcripts across 3646 genes that are differentially expressed between ZIKV-and mock-infected samples. (Fig. 2, S1 and S2 Tables) For the 3969 genes that Cuffdiff found differentially expressed but sleuth did not, sleuth reported an average false discovery rate of 0.55.

It was not surprising that the many differentially expressed genes discovered by Cuffdiff were considered false positives by sleuth. In simulations by Pimentel et al [6], sleuth provided the most accurate false discovery rates, whereas other methods including

DESeq2, edgeR, and Cuffdiff2 underestimated their false discovery rates. In other words, these methods provided differential gene lists that had many more false positives than what was suggested by their p-values. The fundamental idea underlying sleuth is that, by using bootstraps to estimate inferential variance, it does not assume a parametric distribution to account for uncertainty in isoform mapping.

Furthermore, we found a few hundred genes with differentially expressed transcripts not identified by Cuffdiff. We ascribe these to the accounting of experimental design and the isoform-level analysis.

Zika induced isoform divergence

Differentially regulated genes may be missed in gene-level analysis for several reasons. Noise in the measurement of highly expressed transcripts can mask expression changes in lowly expressed transcripts. In the case of isoform switching, upregulation in one isoform and downregulation in another may "cancel out." We identified 108 genes that contain transcript(s) that are significantly upregulated and other transcript(s) that are significantly downregulated, a phenomenon we coin "isoform divergence" (S3 Table). Of these 108 isoform diverging genes, 57 were not considered differentially expressed by Cuffdiff analysis.

We performed a pathway analysis on the 108 genes using Reactome [7]. Enrichment was identified in neuronal system (specifically transmission across chemical synapses and protein-protein interactions at the synapses), developmental biology (specifically axon guidance), immune system, DNA repair, chromatin modifying enzymes, gene expression (rRNA and transcriptional regulation), metabolism, signal transduction, transmembrane transport and vesicle-mediated transport.

One of these 57 isoform diverging genes not picked up by Cufflink is NRCAM, neuronal cell adhesion molecule, which is putatively involved in neuron-neuron adhesion and axonal cone growth. Another is CHRNA7, cholinergic receptor nicotinic alpha 7 subunit. [8] Fig. 3 shows transcript abundances in NRCAM and CHRNA7 across different samples, highlighting isoform-specific changes.

A gene ontology (GO) analysis of sleuth-discovered genes showcase neural and head development networks

We performed a side-by-side gene ontology (GO) analysis with the differential genes identified by sleuth and Cuffdiff, using ClueGO [9, 10] over the Biological Processes ontology network, using GO Term Fusion. We set the network specificity to global (GO tree interval: 1-4), using pathways with a minimum of 50 genes and kappa score of 0.5. We highlighted enriched nodes of particular interest and their enrichments in Fig. 5. Provided in the supplementary materials are the side-by-side GO analysis results tables (S4 and S5 Tables).

Discussion

RNA-Seq can provide rapid and high resolution probing of infection response, and initial studies of Zika infection highlight isoforms, genes and pathways that may play an important role in disease etiology. However, the simplicity of RNA-Seq library prep and cDNA sequencing belies the complexity of analysis. We have shown that a careful analysis of previously published data can reveal novel targets with higher confidence, and in the process rendering a valuable dataset usable by the community of Zika researchers.

The kallisto and sleuth tools we have used in our analysis are particularly powerful when coupled with the interactive sleuth Shiny application, and our publicly available server provides access to numerous interactive plots and figures that cannot be reproduced in a static publication. This highlights the utility and importance of data sharing [11], and we hope that our analysis, aside from its usefulness for the Zika scientific community, can also serve as a blueprint for future data sharing efforts.

sleuth is a fast and accurate pipeline for analyzing RNA-Seq data that allows for testing at the isoform level. The alignment and quantification pipeline is feasible and compatible with a standard desktop computer. The interactive Sleuth application, made publically available, allows for informative data visualization, including those of library prep fragment size distributions, principle component analysis, and gene and transcript expression changes. We invite the scientific community studying Zika to utilize this toolkit.

Figures

Sample	Accession Number	Condition	Sequencing method	Sequencing machine	Reads	No. Fragments / weights
Mock1-1	SRR3191542	mock	paired-end	MiSeq	15855554	7927777
Mock2-1	SRR3191543	mock	paired-end	MiSeq	14782152	7391076
ZIKV1-1	SRR3191544	zika	paired-end	MiSeq	14723054	7361527
ZIKV2-1	SRR3191545	zika	paired-end	MiSeq	15242694	7621347
Mock1-2	SRR3194428	mock	single-end	NextSeq	72983243	72983243
Mock2-2	SRR3194429	mock	single-end	NextSeq	94729809	94729809
ZIKV1-2	SRR3194430	zika	single-end	NextSeq	71055823	71055823
ZIKV2-2	SRR3194431	zika	single-end	NextSeq	66528035	66528035

Table 1. Experimental design. Tang et al. infected two samples with ZIKV and two with a mock infection. Library preparation was performed for each sample to make four cDNA libraries. Each library was then sequenced with MiSeq using paired-end reads and NextSeq using single-end reads.

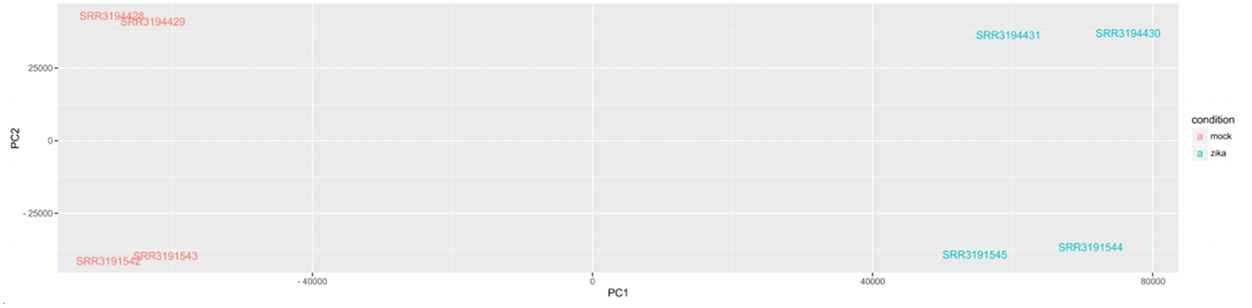


Figure 1: Principle component analysis. PCA of the eight samples shows that the primary contributor to variance is ZIKV infection status (ZIKV vs mock), while the secondary component is sequencing method (paired-end vs single-end).

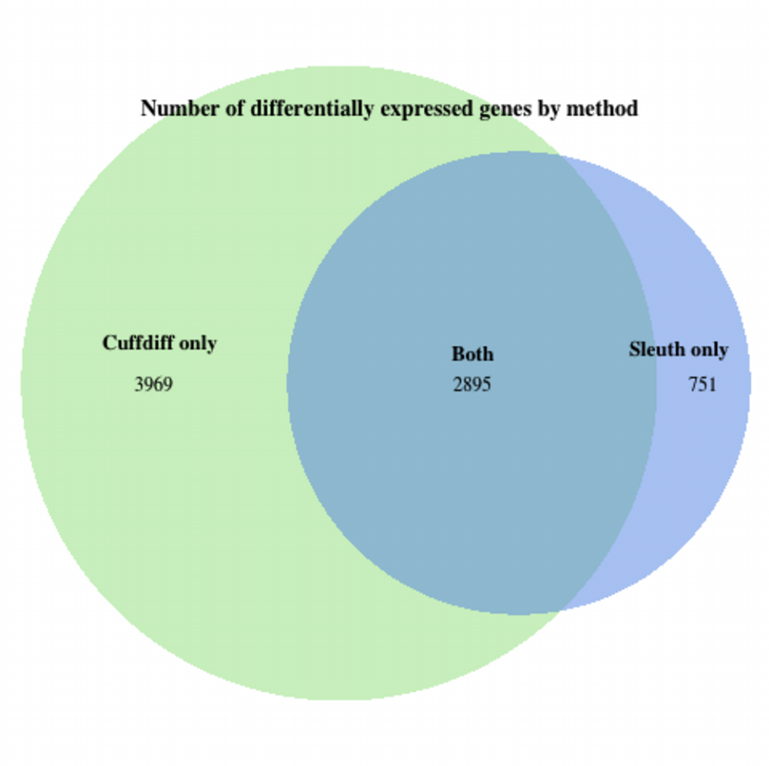
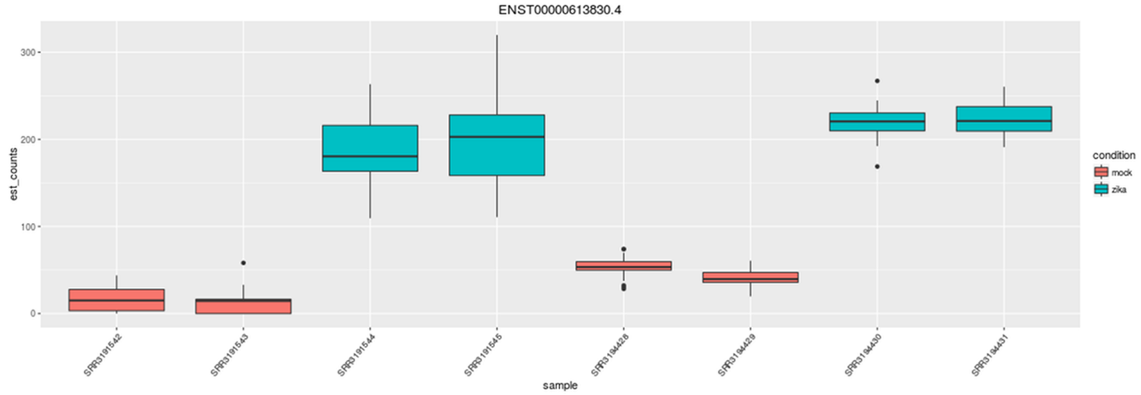
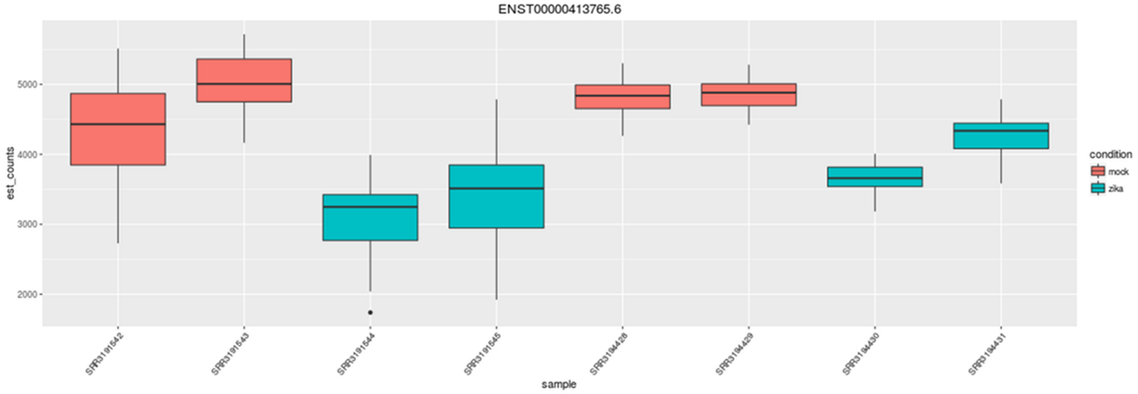
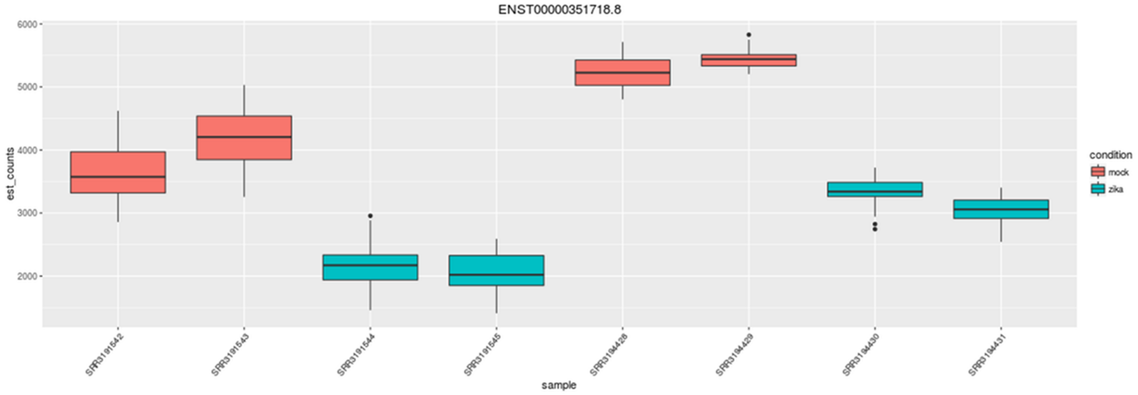


Figure 2: Venn diagram of differential expression analysis. Sleuth identified 3646 differentially expressed genes. Cuffdiff identified 6864 differentially expressed genes. 2895 of the 3646 differentially expressed genes were also reported in Tang et. al [1], but they reported an additional 3969 genes that we failed to identify. Furthermore, we found 751 differentially expressed genes corresponding to 5426 transcripts not detected by Cuffdiff.


ENST00000613830.4
est_counts
300
200
100
0
SRR3191542
SRR3191543
SRR3191544
SRR3191545
SRR3194428
SRR3194429
SRR3194430
SRR3194431
sample
condition
mock
zika


ENST00000413765.6
est_counts
5000
4000
3000
2000
SRR3191542
SRR3191543
SRR3191544
SRR3191545
SRR3194428
SRR3194429
SRR3194430
SRR3194431
sample
condition
mock
zika


ENST00000351718.8
est_counts
6000
5000
4000
3000
2000
SRR3191542
SRR3191543
SRR3191544
SRR3191545
SRR3194428
SRR3194429
SRR3194430
SRR3194431
sample
condition
mock
zika

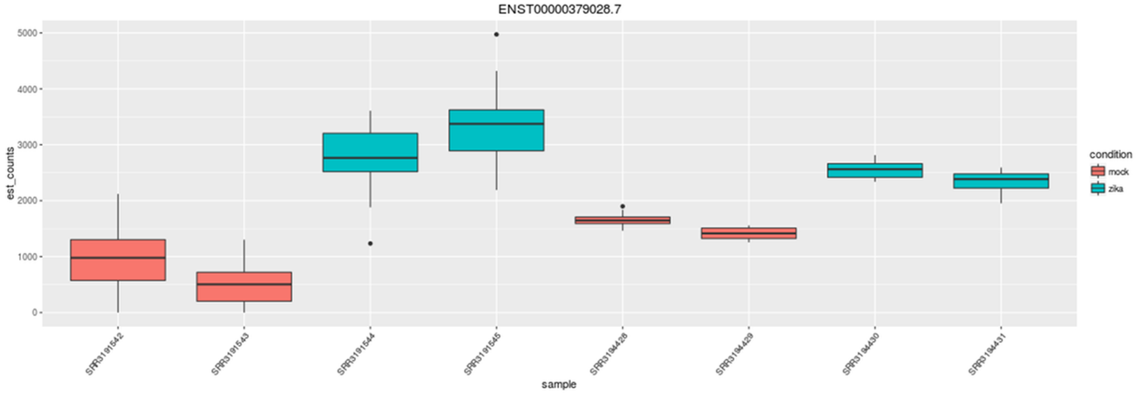

ENST00000379028.7
est_counts
5000
4000
3000
2000
1000
0
SRR3191542
SRR3191543
SRR3191544
SRR3191545
SRR3194428
SRR3194429
SRR3194430
SRR3194431
sample
condition
mock
zika

Figure 3: NRCAM is an example of an isoform divergent gene involved in neuron-neuron adhesion. For a specific gene, the sleuth Shiny app plots the counts corresponding to each transcript and sample. Visualized here are counts for four transcripts of NRCAM across the eight samples, colored by infection status.

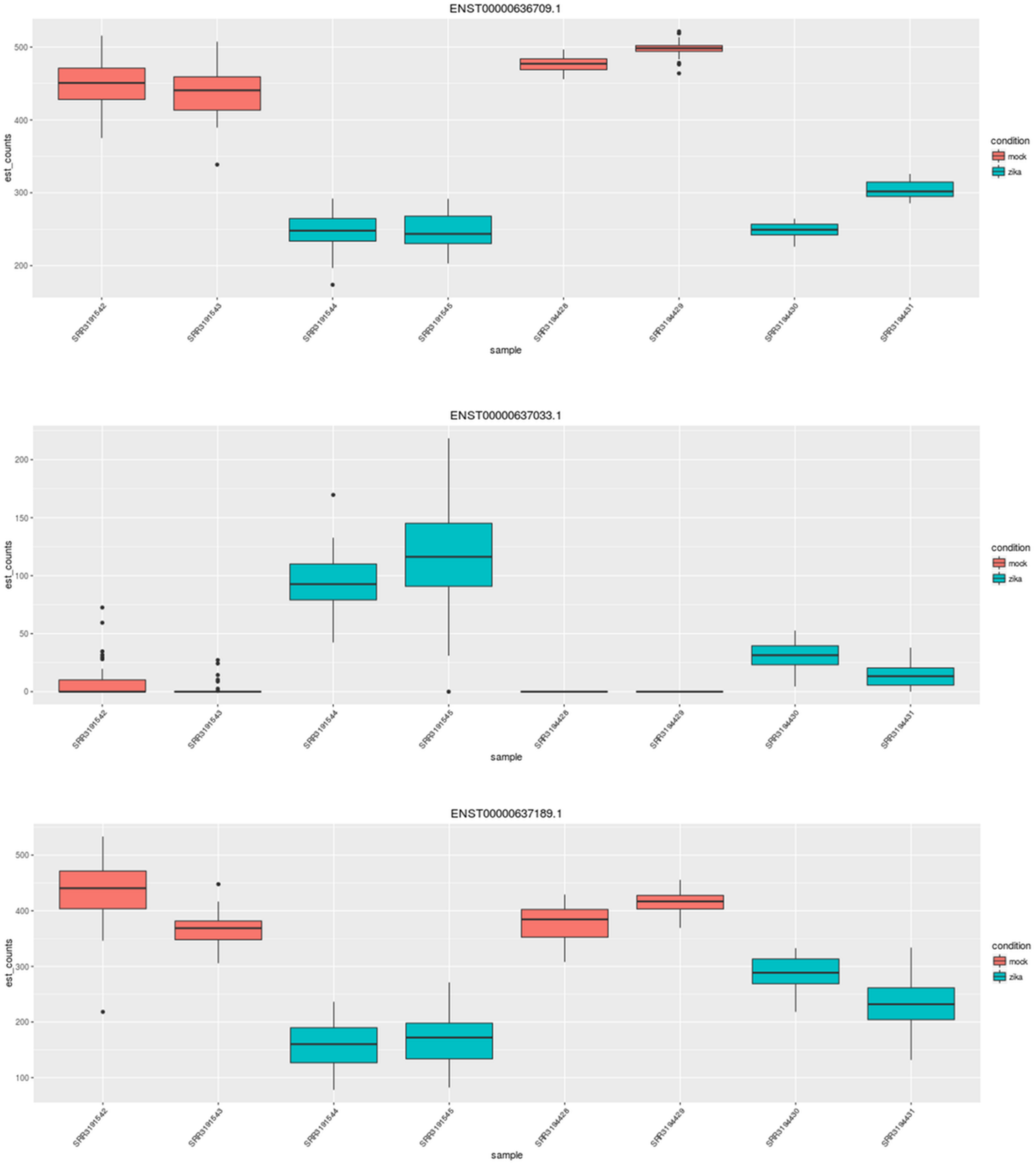


Figure 4: The counts of CHRNA7, another isoform diverging gene, plotted by the sleuth Shiny app. Visualized here are counts for three transcripts of CHRNA7 across eight samples, colored by infection status.

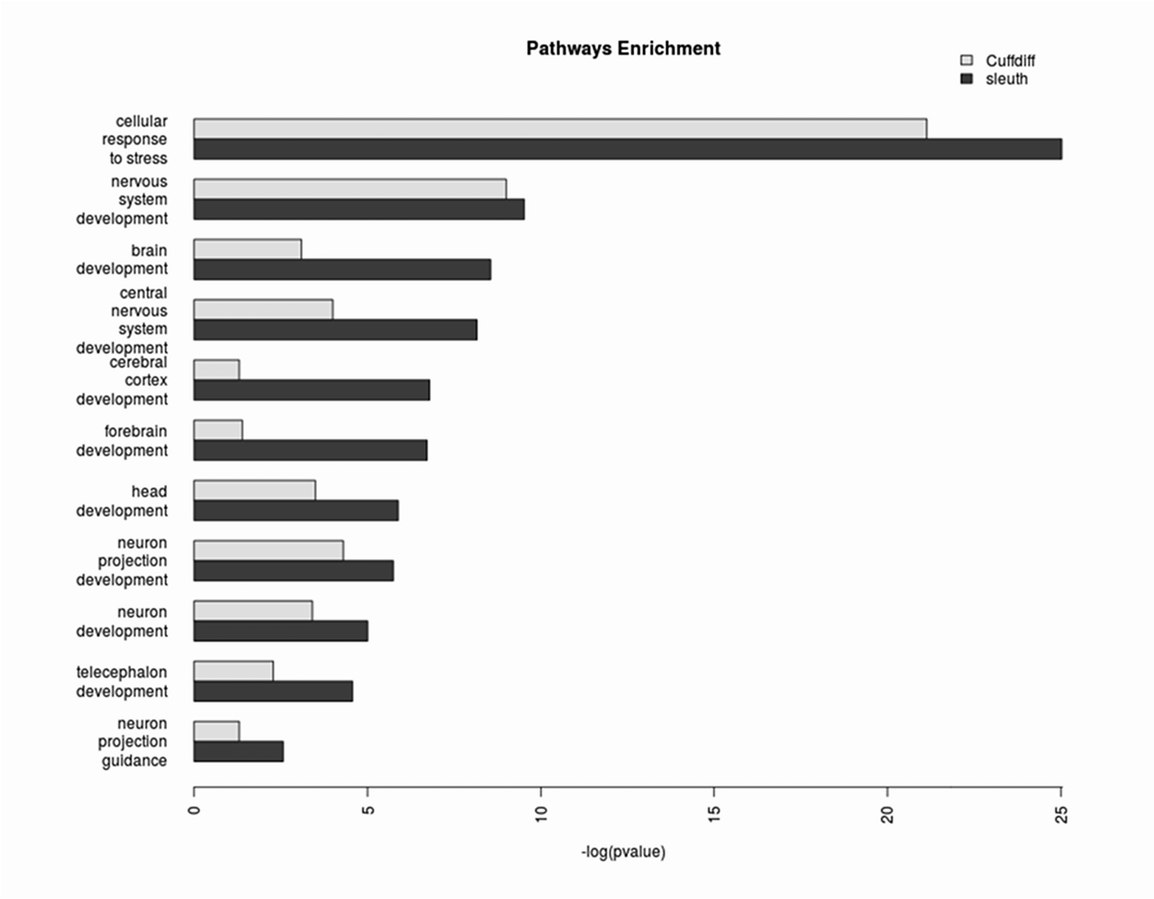


Figure 5: GO pathways enrichment. The enriched nodes of particular interest include neuron projection guidance (p-value = 2.7E-3 vs >0.05 with Cuffdiff), cerebral cortex development (1.6E-7 vs >0.05), neuron development (9.9E-6 vs 3.9E-4), neuron projection development (1.8E-6 vs 5.0E-5), nervous system development (3.0E-10 vs 1.0E-9), central nervous system development (6.9E-9 vs 1.0E-4), brain development (2.8E-9 vs 8.0E-4), forebrain development (1.9E-7 vs 4.1E-2), telecephalon development (2.7E-5 vs 5.2E-3), head development (1.3E-6 vs 3.2E-4), and cellular response to stress (9.4E-26 vs 7.3E-22).

References

[1] Tang H, Hammack C, Ogden SC, Wen Z, Qian X, Li Y, et al. Zika Virus Infects Human Cortial Neuro Progenitors and Attenuate Their Growth. Cell Stem Cell. 2016 May 5;18(5): 587-90.

[2] Rolfe AJ, Bosco DB, Wang J, Nowakowski RS, Fan J, Ren. Bioinformatic analysis reveals the expression of unique transcriptomic signatures in Zika virus infected human neural stem cells. Cell Biosci. 2016; 6:42, doi: 10.1186/s13578-016-0110-x.

[3] Wang Z, Ma'ayan A. An open RNA-seq data analysis pipeline tutorial with an example of reprocessing data from a recent Zika virus study. F1000Research. 2016;5: 1574, doi:10.5256/f1000research.9804.r14924.

[4] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nature Protocols. 2012;7: 562-578.

[5] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016; **34**:525-527, doi:10.1038/nbt.3519.

[6] Pimentel HJ, Bray N, Puente S, Melsted P, Pachter L. Differential analysis of RNA-Seq incorporating quantification uncertainty.
bioRxiv 058164, doi: http://dx.doi.org/10.1101/058164.

[7] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014 Jan;42 (Database issue):D472-7, doi: 10.1093/nar/gkt1102.

[8] Stelzer G, Rosen R, Plaschkes I, Lieder I, Zimmerman S, Twik M, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis. Curr. Protoc. Bioinform. *2016;* 54:1.30.1-1.30.33. *doi: 10.1002/cpbi.5.*

[9] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009, 25(8):1091-3.

[10] Bindea G, Galon J, Mlecnik B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics. 2013, 29(5):661-3.

[11] Longo D, Drazen J. Data Sharing. N Engl J Med 2016; 374:276-277, doi: 10.1056/NEJMe1516564.

Chapter III

Gene-level differential analysis at transcript-level resolution

Lynn Yi[1,2], Harold Pimentel[3], Nicolas L. Bray[4] and Lior Pachter[2,5]

1. UCLA-Caltech Medical Science Training Program, Los Angeles, CA, USA
2. Division of Biology and Biological Engineering, Caltech, Pasadena, CA, USA
3. Department of Genetics, Stanford University, Palo Alto, CA, USA
4. Innovative Genomics Institute, Berkeley, Berkeley, CA, USA
5. Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA, USA

Corresponding authors: Nicolas Bray (nicolas.bray@gmail.com),
Lior Pachter (lpachter@caltech.edu)

This page intentionally left blank.

Abstract

Compared to RNA-Seq transcript differential analysis, gene-level differential expression analysis is more robust and experimentally actionable. However, the use of gene counts for statistical analysis can mask transcript-level dynamics. We demonstrate that *'analysis first, aggregation second,'* where the *p*-values derived from transcript analysis are aggregated to obtain gene-level results, increase sensitivity and accuracy. The method we propose can also be applied to transcript compatibility counts obtained from pseudoalignment of reads, which circumvents the need for quantification, and is fast, accurate, and model-free. The method generalizes to various levels of biology, and we showcase an application to gene ontologies.

Keywords

RNA-Seq, differential expression, metaanalysis, *p*-value aggregation, Lancaster method, Fisher's method, Šidák correction, RNA-Seq quantification, RNA-Seq alignment, pseudoalignment, transcript compatibility counts, gene ontology

Background

Direct analysis of RNA abundance by sequencing cDNAs using RNA-Sequencing (RNA-Seq) offers the possibility to analyze expression at the resolution of individual transcripts (Wang *et al.* 2009). Nevertheless, RNA-Seq continues to be mostly studied at the gene-level, partly because such analyses appear to be more robust (Soneson *et al.* 2016), and also because gene-level discoveries are more experimentally actionable than transcript-level discoveries due to the difficulty of knocking down single isoforms (Kisielow *et al.* 2002).

Gene-level RNA-Seq differential analysis is, at first glance, similar to transcript-level analysis, with the caveat that transcript counts are first summed to obtain gene counts (Anders and Huber 2010, Anders et al. 2015). However, despite such superficial simplicity, there is considerable complexity involved in transitioning from transcripts to genes. In (Trapnell *et al.* 2013), it was shown that naïve approach of summing transcript counts to gene counts lead to inaccurate estimates of fold-change between conditions when transcripts have different lengths. Because transcript counts are proportional to transcript lengths, summing transcript counts is not equivalent to summing transcript abundances.

A remedy to this problem is to estimate gene abundances (e.g. in transcript-per-million units) by summing transcript abundances (Trapnell *et al.* 2010), but regularization methods for variance estimation of gene counts (Robinson *et al.* 2010) cannot be directly applied to abundances. For this reason, recent workflows for gene-level differential analysis rely on converting gene abundance estimates to gene counts (Soneson *et al.* 2016, Pimentel *et al.* 2017). Such methods have two major drawbacks. First, even though the resulting gene counts can be used to accurately estimate fold changes, the associated variance estimates can be distorted (see Figure 1 and Additional file 1: Section 1). Second, the assignment of a single numerical value to a gene can mask dynamic effects among its multiple constituent transcripts (Figure 2). In the case of "cancellation" (Figure 2a), the abundance of transcripts changing in opposite directions cancels out upon conversion to gene abundance. In "domination" (Figure 2b), an abundant transcript that is not changing can mask substantial change in abundance of a minor transcript. Finally, in the case of "collapsing" (Figure 2c), due to overdispersion in variance, multiple isoforms of a gene with small effect sizes in the same direction do not lead to a significant change when observed in aggregate, but their independent changes constitute substantial evidence for differential expression. As shown in Figure 2, these scenarios are not only hypothetical scenarios in a thought experiment, but events that occur in biological data.

Rather than aggregating quantifications prior to differential analysis, one approach is to first perform a transcript-level differential analysis followed by a gene-level meta-analysis. Such a method is implemented in the DEXSeq program (Anders *et al.* 2012), although it is not effective at recovering differential events lost due to collapsing, and is suboptimal even for cancellation or domination events (see Results and Additional file 1: Section 2). Meta-analysis has been suggested for microarray studies to aggregate probe-level *p*-values (Hess *et al.*, 2007) and is performed in genome-wide association studies to aggregate single nucleotide polymorphism *p*-values to make gene-level (Chen *et al.*, 2014. Dai *et al.*, 2011., Lamparter *et al.*, 2016) and pathway-level inferences (Li *et al.*, 2011, Lamparter *et al.*, 2016), but such approaches do not appear to have been extensively explored for RNA-Seq.

We present a new framework for gene-level differential analysis that utilizes the Lancaster method (Lancaster, 1961). In this framework, differential expression is performed on transcripts as usual, but then transcript-level *p*-values are aggregated to obtain gene-level *p*-values. (See Methods for details about the Lancaster method. See Additional file 1 for applicability of the Lancaster method to RNA-Seq.)

Our approach can be based on *p*-values derived from transcript-level differential analysis, but can also be applied to *p*-values derived from comparisons of transcript compatibility counts (TCCs), a concept introduced by the pseudoalignment method in kallisto (Bray *et al.*, 2016). Transcript compatibility counts are the number of reads that are compatible with a set of transcripts, i.e. an equivalence class. In default RNA-Seq quantification mode, kallisto matches each read with its equivalence class, thus generating TCCs, and then applies the expectation-maximization (EM) algorithm on TCCs to obtain transcript quantifications. Differential analysis performed on directly TCCs has the advantage of being fast and model-free, and we show that it is particularly useful for positionally biased RNA-Seq data.

Finally, we highlight the generality of our approach at varying levels of biological resolution by extending it to gene ontology analysis. In contrast to classical gene ontology (GO) tests that identify enrichment of GO terms with respect to gene lists, our approach identifies GO terms in which there is significant perturbation among the associated genes. We combine this idea with TCC-based differential analysis to illustrate how GO analysis can be performed on RNA-Seq data without transcript quantification.

Results

We first examined the performance of aggregation in comparison to standard gene-level differential expression methods using three simulated scenarios from (Pimentel *et al.* 2017). In these simulations, transcripts are perturbed independently, in a correlated fashion with other transcripts of the gene, or according to effect sizes observed in a biological experiment. In the first scenario of independent effects, random transcripts in the transcriptome are independently chosen to be perturbed, and the effect size for each transcript is chosen independently. In the second scenario of correlated effects, genes are independently chosen to be differentially expressed, and all transcripts of the same gene are perturbed in the same direction. In the third scenario of experimentally based effects, effect sizes are learned from an experimental data set and applied to the simulation. (See Methods for more details.) Each of the three scenarios was simulated twenty times.

We evaluated the performance of various aggregation methods on these simulations with two differential expression methods, sleuth and DESeq2. These differential expression methods were chosen for their superior performance in previously published simulations (Pimentel *et al,* 2017). sleuth utilizes bootstraps on reads to estimate inferential variance due to read-mapping and quantification uncertainty, which is then used in a linear model to perform differential expression analysis. DESeq2 utilizes a negative binomial model on counts *(Love et al. 2014)*. We evaluated every aggregation method using each differential expression method in each of the three simulation scenarios.

Figure 3 shows the results of performing aggregation using sleuth in the simulation scenario that is modeled after experimental effect sizes, plotted as a false discovery rate (FDR)-sensitivity tradeoff curve. (Additional File 1: Figs S1 and S2 show results with other two simulation scenarios using sleuth. Additional File 1: Fig S3 shows results with the three simulation scenarios using DESeq2.) Aggregation of transcript *p*-values using the Lancaster method (Lancaster, 1961) outperforms standard gene-level analysis; it provides greater power at lower FDR. Furthermore, Lancaster-based aggregation outperforms the Šidák method of DEXSeq, which utilizes the minimum transcript *p*-value to make the gene-level determination (method corrected, Additional file 1: Section 2). While the Šidák method performs well when transcripts are perturbed independently (Additional file 1: Fig. S1), it performs very poorly in the more common case of correlated effect (Additional file 1: Fig. S2). In addition to providing more power at lower FDR than the other methods, the Lancaster method is also better at controlling and accurately reporting FDR (See Figure 3b for reported FDRs). Additional file 1: Fig. S3 shows similar improvements when aggregation is performed using *p*-values that are derived from DESeq2 (Love *et al.* 2014) instead of sleuth. Regardless of the differential expression method used to compute *p*-values, the Lancaster method of aggregation outperforms the other methods, showing that improvements in performance are due to the aggregation method and not the differential expression software.

Transcript-level *p*-values are computed from transcript quantifications, a process that introduces uncertainty from multiple-mapping RNA-Seq reads. (Pimentel et al, 2017) showed that propagating uncertainty from the transcript quantification to differential expression analysis increases accuracy of the differential expression analysis. In kallisto (Bray *et al.* 2016), pseudoalignment was performed to generate transcript compatibility counts (TCCs), which are the number of reads that are compatible with sets of transcripts and therefore do not contain any quantification uncertainty. Given the improved results observed with performing Lancaster aggregation, we asked whether it is possible to perform differential expression analysis directly on TCCs and aggregate on TCC *p*-values to obtain gene *p*-values, thereby bypassing transcript quantification and the uncertainty it entails altogether. Figure 3 shows that aggregating TCC *p*-values outperforms other methods, including that of aggregating transcript *p*-values. Furthermore, aggregating TCC *p*-values reported FDRs that are as or more accurate than those reported by other methods. In this instance, we used only TCCs that mapped solely to the transcripts of a single gene, which accounts for 88% of the RNA-Seq reads. It may be possible to continue to improve performance by accounting for intergenic TCCs.

Aggregation of TCCs is useful when quantification is complicated due to non-uniformity of reads coverage across transcript spans. While non-uniformity in coverage is prevalent in RNA-Seq (Hayer *et al.*, 2015), it is particularly extreme in variants of RNA-Seq that enrich for 5’ or 3’ sequences. We used TCC aggregation to perform differential expression on QuantSeq data (Moll *et al.,* 2014), where an experiment involved mechanically stretching rat primary type I like alveolar epithelial cells and then performing QuantSeq 3' mRNA sequencing to detect changes in 3' untranslated region (UTR) expression (Dolinay *et al.,* 2017, GEO Series GSE89024). Figure 4a shows that overall results with TCC-based aggregation are similar to standard analysis based on gene counts obtained by summing the number of reads that map to any constituent isoforms. However, TCC-based aggregation allows for the discovery of events that are masked in standard count-based analysis. Figure 4b shows an example where we discovered 3' UTR isoform switching, an event which could not be identified with a gene counts-based analysis. While *p*-value aggregation works well for gene-level differential expression analysis, aggregation can be extended to other natural groupings. To demonstrate the generality of the approach, we applied *p*-value aggregation to gene ontologies (Ashburner, 2000). Classic gene ontology (GO) analysis of a RNA-Seq experiment involves first performing gene differential expression analysis to obtain either a list of statistically differential genes (i.e. all genes with q-value < 0.05) or a rank order list of genes (i.e. ordered by p-value) and then identifying GOs that are statistically enriched in this gene list. Common statistical tests for enrichment include Fisher's exact test and Wilcoxon rank-sum test (Huang *et al., 2009,* Mi *et al., 2013*). Instead of testing for enrichment of GOs, we examined the complementary question of "perturbation analysis," namely, whether the GO is significantly perturbed. To test for perturbation, we aggregated *p*-values based on transcript quantifications or TCCs for all genes in each GO term to obtain *p*-values for each GO term, which are then Bonferroni corrected. Unlike standard GO enrichment analysis, this perturbation analysis utilizes the information derived from all genes and reveals information not only about membership, but also about the significance of perturbation.

We performed differential expression and GO analysis on recently published RNA-Seq data that examined the effect of dexamethasone treatment on primary neural progenitor cells of embryonic mice (Frahm *et al.*, 2017, GEO Series GSE95363). First, we performed differential expression using each of the four previously discussed aggregation methods to obtain differential gene lists (FDR < 0.05). (Additional file 1: Fig. S4 compares differential expression with sleuth standard gene mode vs. Lancaster aggregating TCC *p*-values.) Then, we applied classical GO enrichment analysis to each method's differential gene list. The Lancaster method applied to TCC derived *p*-values produced the differential gene list that is enriched for the most "immune"-containing GO terms (Figure 5a). To apply the GO perturbation test, we performed further aggregation on the gene *p*-values resulting from differential expression analysis to generate GO *p*-values, resulting in a total of four GO perturbation tests. Each GO perturbation test resulted in a perturbed GO list (FWER < 0.05) that was more enriched for "immune"-containing GO terms than the corresponding enrichment test (FWER < 0.05) (Additional file 1: Fig. S5).

To highlight some specific results, in the GO perturbation test based on aggregating TCC *p*-values, we found 6396 GO terms (<0.05 FWER) perturbed by dexamethasone treatment. Example terms at the top of the perturbed list included: system process (GO:0003008), response to stress (GO:0006950), metabolic process

(GO:0008152), immune system process (GO:0002376), inflammatory response (GO:0006954), and response to hormone (GO:0009725). As a comparison, the corresponding classical enrichment analysis using Fisher's exact test revealed 2123 enriched GO terms (<0.05 FWER). Many of the perturbed GOs mentioned above were also enriched, but system process and inflammatory response were not (FWER = 0.27 and 1.00). In other words, an enriched ontology is likely perturbed, but not vice versa, and indeed, many "immune"-containing GO terms were perturbed but not enriched (Figure 5b). These results suggest that perturbation analysis can be a useful and powerful complementary analysis to standard GO enrichment analysis.

Discussion

We have shown that aggregating *p*-values to obtain gene-level *p*-values is a powerful and tractable method that provides biologically interpretable results. By using only the resulting *p*-values from a differential expression analysis, aggregation bypasses issues of different variances and directions of change across constituent transcripts, allowing it to capture cancellation, domination and collapsing events. All the examples of failure modes of traditional gene differential analysis showcased in Figure 2 were successfully identified with the Lancaster method. Furthermore, performing the Lancaster method on TCC *p*-values leverages the idea of pseudoalignment for RNA-Seq, enabling a fast and model-free approach to differential analysis that circumvents numerous drawbacks of previous methods.

The method of *p*-value aggregation is also extendable to testing other features of biological interest. We have demonstrated its utility for GO analysis to test for perturbation of gene ontologies, a complementary analysis that can be used in addition to existing GO enrichment tests. Aggregation can be performed hierarchically to maintain resolution at all levels including transcripts, genes and gene ontology terms. Further applications can include testing for intron retention, differential transcript start site (TSS) usage, and other use cases where aggregation of features is of interest. Finally, gene-level testing directly from TCC counts is particularly well-suited for single-cell RNA-Seq analysis, where many technologies produce read distributions that are non-uniform across transcripts.

While this paper has focused on higher-order differential analysis, the complementary problem of differential analysis of individual transcripts can also benefit from some of the aggregation ideas described here. The stageR method, recently described in (Van den Berge *et al.*, 2017), incorporates a two-step testing procedure in which an initial meta-analysis at the gene-level (using DEXSeq) is used to identify differential transcripts without losing power due to testing of all transcripts. The use of the Šidák method for aggregation of *p*-values makes sense in that context, as it is desirable to identify genes with at least one differential isoform. However, it is possible that some of the methods we have introduced, including testing of TCCs and weighting, could be applied during the screening stage.

Conclusions

Transcript differential analysis and gene differential analysis for RNA-Seq have been two independent procedures up until now. Aggregating transcript p-values with the Lancaster method to call gene differential expression not only outperforms other gene-level methods, it also retains information about transcript dynamics and produces one coherent analysis between transcripts and genes. This framework can be leveraged to study multiple resolutions of biology, such as performing a hierarchical analysis of transcripts, genes and gene ontologies, or to bypass artifacts introduced at a particular resolution, such as obtaining gene-level results without transcript quantification by aggregating on transcript compatibility counts.

Methods

Aggregation of p-values

Fisher's method aggregates K p-values $p_1,\dots, p_{K,}$ which, under the null hypothesis, are independent and uniformly distributed between 0 and 1. Under the null hypothesis, the test statistic $T = \sum_{i=1}^{K} -2\log(p_i)$ is chi-squared distributed with degrees of freedom $(df) = 2K$. The aggregated p-value is therefore $1 - \phi(\sum_{i=1}^{K} -2\log(p_i))$, where ϕ is the cumulative distribution function (CDF) of a chi-squared distribution with $df = 2K$. (Fisher, 1932)

The Lancaster method (Lancaster, 1961) generalizes Fisher's method for aggregating p-values by introducing the possibility of weighting the p-values with weights

$w_1,\ldots,w_K$. According to the Lancaster method, under the null hypothesis where all studies have zero effect, the test statistic $T = \sum_{i=1}^{K} \phi_{wi}^{-1}(p_i)$, where ϕ_{wi}^{-1} is the inverse CDF of the chi-squared distribution with $df = w_i$, follows a chi-squared distribution with $df = \sum_{i=1}^{K} w_i$. Fisher's method is a specific instance of the Lancaster method where all p-values are uniformly weighted by 2, and we found that the Lancaster method applied with a weighting scheme based on transcript counts outperformed Fisher's method (Additional file 1: Fig. S6).

We investigated whether the assumptions of Fisher's and the Lancaster method, namely that p-values are independent and uniformly distributed under the null hypothesis, apply to RNA-Seq. Additional file 1: Fig. S7 shows a distribution of the transcript p-values for the dexamethasone RNA-Seq data we examined. Aside from a peak close to 0, presumably corresponding to the differential transcripts, the p-values appear to be uniformly distributed. Furthermore, the Additional file 1: Section 3 contains a walkthrough of the experiments we performed to test the independence between transcripts under the null hypothesis, showing that while transcripts of the same are not independent in general, the dependence is weak and does not lead to exaggerated p-values or inflated false discovery rates (Additional file 1: Figs. S8, S9).

The Šidák method (Šidák, 1967) utilizes a test based on the minimum p-value $m = min(p_1,\ldots, p_K)$, namely the adjustment $\theta = 1 - (1 - \mathrm{m})^K$. In the context of K isoforms with p-values $p_1,\ldots, p_K$, θ is the gene-level p-value based on adjusting for the number of isoforms in the gene. If there are M genes, the adjustments will generate p-values $\theta_{1,\ldots,}\theta_M$, which can be corrected for multiple testing. This method is similar to the perGeneQvalue result from DEXSeq (Anders *et al.*, 2012), and while both methods control the false discovery rate, the gene ranking is different between the two methods (Additional file 1: Section 2).

Transcript differential analysis and aggregation

RNA-Seq reads were quantified with kallisto v.0.43.1 to obtain transcript counts and abundances. These transcript counts were used as inputs in differential expression methods sleuth and DESeq2 in order to obtain transcript *p*-values, which were then aggregated with the Lancaster method to obtain gene *p*-values. sleuth and DESeq2 were run with their respective default filters and the Wald test. sleuth was run with 30 bootstraps. Transcripts filtered out from the differential expression analysis due to low counts were also filtered out from the *p*-value aggregation. To obtain *p*-value weights for the Lancaster method, we used as weights the mean expression level for the transcript extracted by the differential expression analysis (i.e. the mean_obs parameter in sleuth, the baseMean parameter in DESeq2). FDRs were calculated for the gene-specific *p*-values using the Benjamini-Hochberg method. While we used the Wald test in this manuscript for obtaining transcript and gene differential expression analysis, we also tested the likelihood ratio test, which showed similar improvements with Lancaster aggregation and whose performance is comparable to the Wald test (Additional file 1: Fig. S10).

Transcript compatibility count differential analysis and aggregation

Transcript compatibility counts (TCCs) of RNA-Seq reads were obtained with the kallisto *pseudo* option, which outputs a TCC matrix whose two dimensions are the number of samples and number of equivalence classes. Each TCC represents the RNA-Seq counts corresponding to an equivalence class of transcripts. All TCCs corresponding to transcripts from more than one gene were filtered out from the analysis; 88% of reads were retained after applying this filter. The remaining TCCs were used to perform differential expression with sleuth (Pimentel *et al.* 2017) and DESeq2 (Love *et al.* 2015) by using TCCs in lieu of transcript/gene counts. In order to use sleuth, we performed 30 bootstraps on TCCs, whose results were inputted into sleuth to estimate inferential variance. Non-expressed TCCs were filtered from the sleuth analyses and the default filter in DESeq2 was used. Both methods were performed with the likelihood ratio test because we found that the Wald test applied to TCCs reported overly liberal FDRs. The resulting TCC *p*-values from the differential expression analysis were aggregated using the Lancaster method, with *p*-value weights equal to the log-transformed mean counts normalized to 1. In other words, given *K* TCCs of the same gene with mean counts $t_1, \ldots, t_K$, the weight for the *i*th TCC is $w_i = \frac{\log(t_i+1)}{\sum_{j=1}^{K} \log(t_j+1)}$.

Gene differential analysis

The aggregation methods were compared to standard gene-level differential analysis performed with sleuth and DESeq2. sleuth was run in gene mode with 30 bootstraps. DESeq2 was run on gene counts obtained using tximport (Soneson *et al.* 2015) to aggregate transcript quantifications, except the case of 3' QuantSeq data set, where gene counts were obtained by summing reads that uniquely map to a gene. Both sleuth and DESeq2 were run with the Wald test and their respective default filters.

Simulations

The simulations used to benchmark the method followed the approach of (Pimentel *et al.* 2017). A null distribution consisting of the negative binomial model for transcript counts was learned from the Finnish female lymphoblastic cell lines subset of GEUVADIS (Lappalainen *et al.*, 2013). A distribution of fold changes to the mean was learned from an experimental data set from (Trapnell *et al.*, 2013), and 20% of genes were chosen randomly to be differentially expressed, with fold changes of the transcripts assigned by rank-matching transcript abundances. Twenty simulations were performed, each with different randomly chosen sets of differentially expressed genes. (For further details on the simulation structure see (Pimentel *et al.* 2017).)

The simulations were quantified with kallisto v0.43.1 using an index constructed from Ensembl *Homo sapiens* GRCh38 cDNA release 79. Differential expression analyses were performed with sleuth and DESeq2 and then aggregated with various methods described above. Sensitivities and corresponding FDRs were calculated and then averaged across the twenty simulations. The average sensitivity at each average FDR was plotted with the mamabear package (Pimentel *et al.,* 2017, https://github.com/pimentel/mamabear).

Rat Alveolar Epithelial Cell Stretching Data Set Analysis

We used a 3' QuantSeq data set (GEO Series GSE89024) of stretched and unstretched rat primary type I like alveolar epithelial cells. Five replicates for each condition were performed by the original experimenters, resulting in a total of 10 single-end RNA-Seq samples (Dolinay *et al.,* 2017). Reads were trimmed to remove poly-A tails with fqtrim-0.9.5 using the default parameters (Johns Hopkins Center for Computational Biology, 2015). As discussed above in Methods section 'Transcript compatibility count differential analysis and aggregation,' TCCs were obtained with the kallisto *pseudo* option, differential expression of TCCs was performed in sleuth, and TCC *p*-values were aggregated with the Lancaster method. Because kallisto quantification is invalid for this non-uniform sequencing dataset and it cannot be used to provide bootstrap estimates of inferential variance required for sleuth, we used DESeq2's default pipeline to perform gene differential analysis, summing all reads mapping uniquely to a gene to obtain gene counts.

Dexamethasone Data Set Analysis

We analyzed a data set (GEO Series GSE95363) consisting of reads derived from RNA-Seq on primary mouse neural progenitor cells extracted from two regions of the brain, from female and male embryonic mice, and with and without dexamethasone treatment. Three replicates were performed for each of the eight combinatorial conditions, resulting in a total of 24 single-end RNA-Seq samples (Frahm *et al.*, 2017). As detailed above in 'Transcript differential analysis and aggregation', samples were quantified with kallisto v0.43.1 (default kmer length 31, with 30 bootstraps per sample), using an index constructed from Ensembl *Mus musculus* GRCm38 cDNA release 88. Within sleuth, a linear model with three parameters (gender, brain region, and treatment) was constructed, a Wald test was performed to test for effect of treatment on transcript expression, and the resulting *p*-values were aggregated. As detailed above in 'Transcript compatibility count differential analysis and aggregation,' TCCs were obtained with kallisto v0.43.1 using the *pseudo* option, differential expression of TCCs was performed in sleuth, and the resulting *p*-values aggregated. On this data set, we also performed the sleuth's standard gene pipeline (detailed in 'Gene differential analysis') and the Sidak aggregation method, resulting in a total of four different aggregation methods.

Each method's significant gene list, thresholded at FDR < 0.05, was inputted into a classical GO analysis to test for GO enrichment. topGO_2.26.0 (Alexa *et al.*, 2016) was invoked to perform Fisher's exact test, using gene ontologies drawn from GO.db_3.4.0 and mouse gene annotations drawn from org.Mm.eg.db_3.4.0 (The Gene Ontology Consortium, 2015). Furthermore, the gene *p*-values from each aggregation method were used in a GO perturbation test. In the GO perturbation test, gene *p*-values are weighted by the counts mapping uniquely to the gene and aggregated with the Lancaster method, using the ontology-to-gene mappings provided by topGO. The GO *p*-values were Bonferroni corrected to obtain FWER.

Software Versions

DESeq2 1.14.1 and sleuth 0.29.0 were used in R version 3.4.1 to perform differential analyses. tximport 1.2.0 was used to sum transcript counts within genes to perform gene-level differential expression with DESeq2. We implemented Fisher's method and Lancaster method with the chisq and gamma functions in the R Stats Package. A lightweight R package containing the functionality for performing *p*-value aggregation with Fisher's, Lancaster and Šidák methods, which is applicable generally to outside the domain of RNA-Seq, is available on CRAN as "aggregation" (https://cran.r-project.org/web/packages/aggregation/index.html). Our method to perform gene-level differential analysis via Lancaster aggregation of transcript *p*-values has been implemented in sleuth. Scripts to reproduce the figures and results of the paper are available at http://github.com/pachterlab/aggregationDE/.

Figures

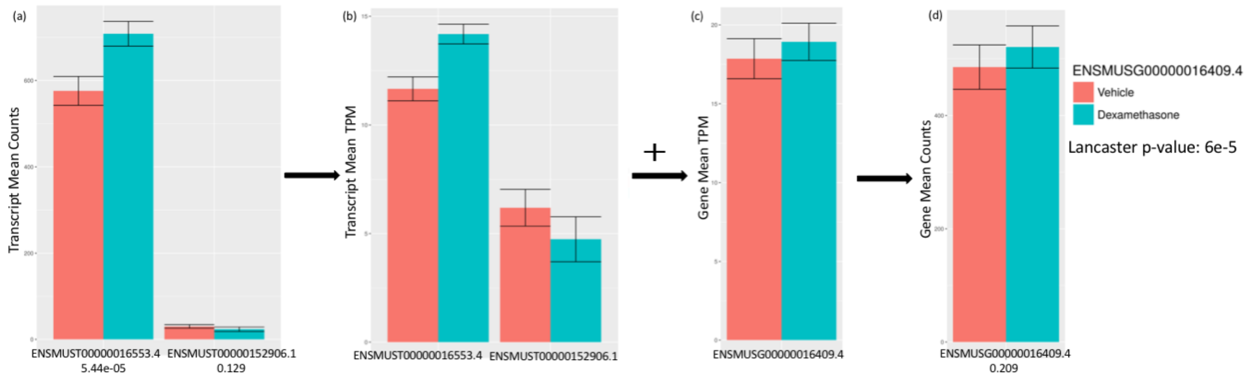


Figure 1. Conversion of transcript counts to gene counts for the *Nkap* gene in the dexamethasone dataset under two conditions (dexamethasone and vehicle treatment). The *x*-axis is labeled with the Ensembl gene and transcript IDs, along with *p*-values obtained by performing sleuth on transcripts and genes. In this process, the transcript counts (a) are converted into transcript abundances (b) by normalization according to transcript lengths. Transcript abundances are then summed to obtain gene abundances (c), and then converted to gene counts (d) using the median or mean transcript length as a proxy for the gene length. The converted gene counts mask significant changes among the constituent transcripts, and the gene count variance does not directly reflect the combined variance in transcript counts. In this example *Nkap* is not differential when examined using the converted gene counts, but can be identified as differential when the *p*-values of the constituent transcripts are aggregated using the Lancaster method.

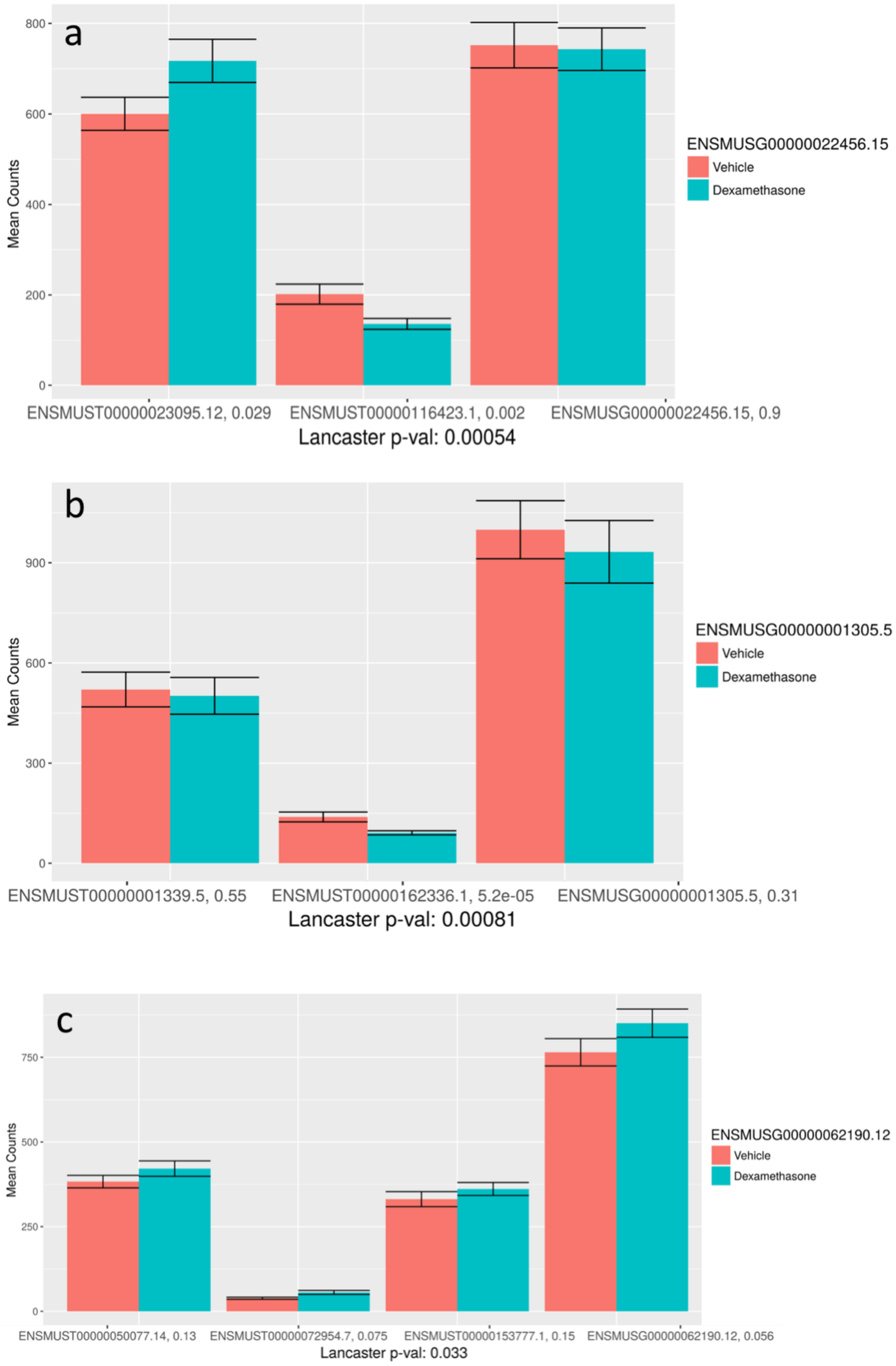

a
Mean Counts
0
200
400
600
800
ENSMUST00000023095.12, 0.029
ENSMUST00000116423.1, 0.002
ENSMUSG00000022456.15, 0.9
Lancaster p-val: 0.00054
ENSMUSG00000022456.15
Vehicle
Dexamethasone
b
Mean Counts
0
300
600
900
ENSMUST00000001339.5, 0.55
ENSMUST00000162336.1, 5.2e-05
ENSMUSG00000001305.5, 0.31
Lancaster p-val: 0.00081
ENSMUSG00000001305.5
Vehicle
Dexamethasone
c
Mean Counts
0
250
500
750
ENSMUST00000050077.14, 0.13
ENSMUST00000072954.7, 0.075
ENSMUST00000153777.1, 0.15
ENSMUSG00000062190.12, 0.056
Lancaster p-val: 0.033
ENSMUSG00000062190.12
Vehicle
Dexamethasone

Figure 2: Differential transcript masking. Dynamics among transcripts may not be detected with gene-level analyses due to cancellation (a), domination (b) and collapsing (c). Gene counts and constituent transcript counts are plotted between conditions (dexamethasone versus vehicle treatment) and annotated with Ensembl ID and sleuth-derived *p*-values. In the case of cancellation (a), the abundance of transcripts changing in opposite directions cancels out upon conversion to gene abundance. In domination (b), an abundant transcript that is not changing can mask substantial change in abundance of a minor transcript. In the case of collapsing (c), multiple isoforms of a gene with small effect sizes in the same direction do not lead to a significant change when observed after summation, but their independent changes constitute substantial evidence for differential expression. In all these examples, gene-level differential analysis with sleuth failed to identify the genes as differential (*p*-values listed on x-axis), whereas Lancaster aggregation of transcript *p*-values resulted in detection of the genes as differential.

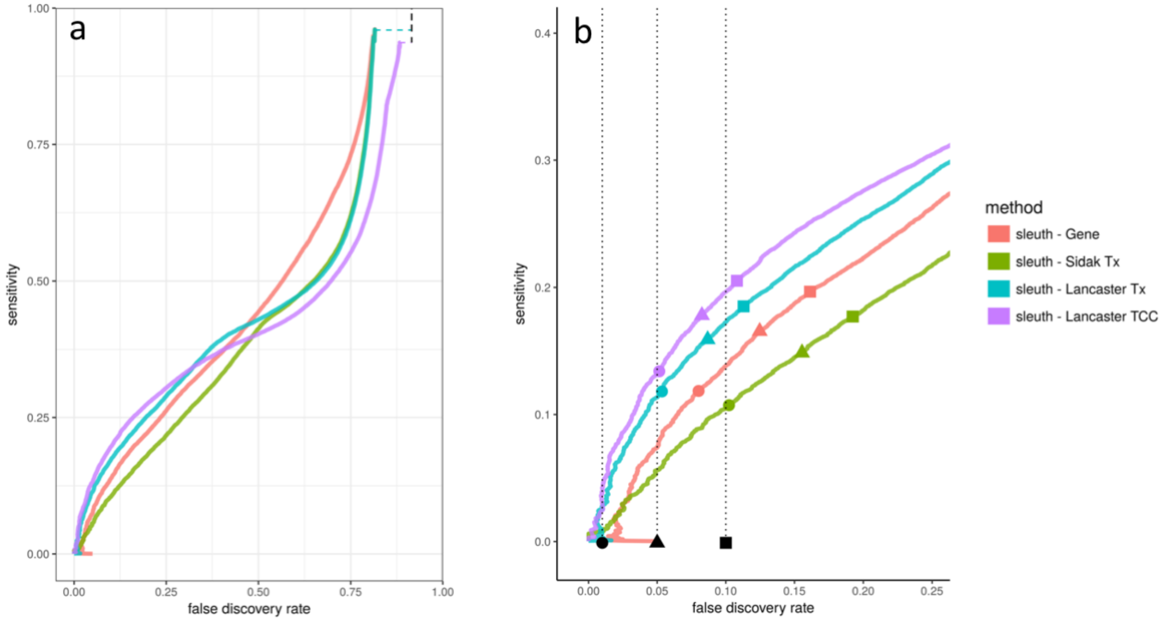
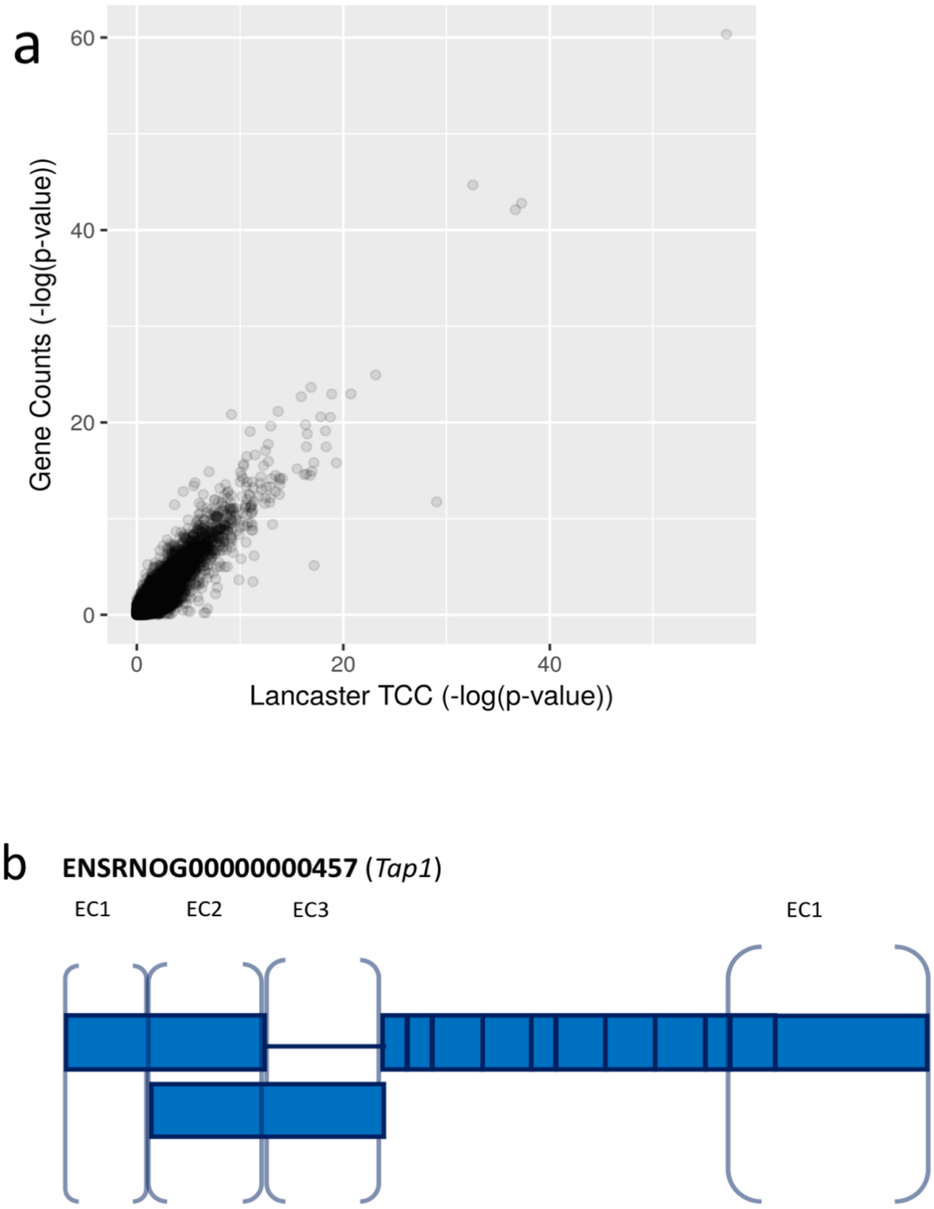


Figure 3: Sensitivity and false discovery trade-off curves of aggregation methods. Twenty simulated experiments based on parameters estimated from biological data were analyzed with different aggregation methods and averaged producing (a), and zoomed in (b). sleuth in gene mode ('sleuth-Gene') is a standard gene-level differential analysis method. Aggregation results based on transcript *p*-values are shown using two approaches: sleuth transcript *p*-values aggregated by the Lancaster method ('sleuth-Lancaster Tx') and sleuth transcript *p*-values aggregated by the Šidák-adjusted minimum method ('sleuth – Sidak Tx'). Finally, sleuth TCC *p*-values obtained by running sleuth on TCC counts were aggregated with the Lancaster method ('sleuth-Lancaster TCC'). Dashed lines indicate true FDR at 0.01, 0.05, and 0.1. The shapes (circle, triangle, square)

on each sensitivity-FDR curve indicate the true FDR and sensitivity at each method's reported FDRs of 0.01, 0.05, and 0.1.



Equivalence Class	p-value	Effect Size of Stretching (log fold change)	Mean Counts
EC1	0.0037	- 0.992	15.5
EC2	0.34	- 0.229	0.1
EC3	0.14	0.649	6.3

Lancaster TCC p-value: 0.0056
Gene count p-value: 0.169

Figure 4: Analysis of positionally-biased RNA-Seq data using TCC aggregation. A log-log plot of *p*-values comparing aggregated sleuth-derived TCC *p*-values using the Lancaster method (x-axis) to *p*-values obtained by differential analysis in DESeq2 with gene counts (y-axis) shows overall agreement (a). DESeq2 applied on gene counts discovered 460 DE genes (FDR < 0.05); Lancaster aggregation on TCCs discovered 243 genes (FDR < 0.05). TCC aggregated analysis can detect differential 3' UTR usage that is masked in gene count analyses (b). An example is shown from the rat gene *Tap1*, with rectangular blocks representing individual exons (blank = noncoding, solid = coding), and distinct equivalence classes (EC's) labeled with brackets. Two other transcripts and their corresponding (zero count) equivalence classes are not shown. Significance levels for *Tap1* under effects of alveolar stretching were calculated using the Lancaster method (*p*-value = 0.0056) and compared to *p*-values derived from gene counts (*p*-value = 0.169).

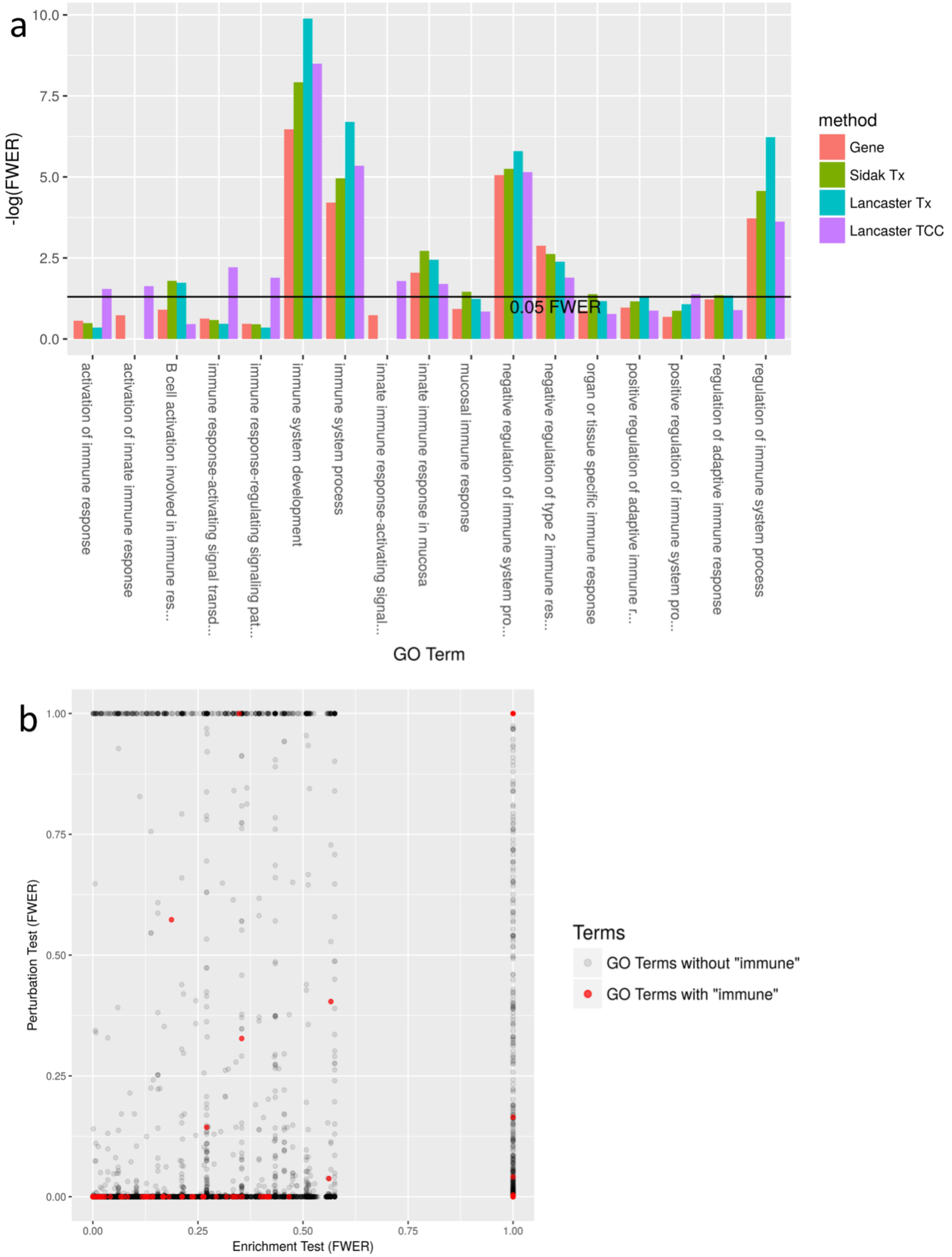

a
-log(FWER)
10.0
7.5
5.0
2.5
0.0
0.05 FWER
method
Gene
Sidak Tx
Lancaster Tx
Lancaster TCC
activation of immune response
activation of innate immune response
B cell activation involved in immune res...
immune response-activating signal transd...
immune response-regulating signaling pat...
immune system development
immune system process
innate immune response-activating signal...
innate immune response in mucosa
mucosal immune response
negative regulation of immune system pro...
negative regulation of type 2 immune res...
organ or tissue specific immune response
positive regulation of adaptive immune r...
positive regulation of immune system pro...
regulation of adaptive immune response
regulation of immune system process
GO Term
b
Perturbation Test (FWER)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Enrichment Test (FWER)
Terms
GO Terms without "immune"
GO Terms with "immune"

Figure 5: GO analysis based on *p*-value aggregation. (a) Four aggregation methods ('Lancaster TCC', 'Lancaster Tx', 'Sidak Tx' and 'Gene') were performed with sleuth to obtain gene-level differential expression analysis on response to dexamethasone treatment. The significant genes (FDR<0.05) from each differential expression analysis were tested for GO enrichment (Fisher's exact test) and Bonferroni-corrected. GO terms containing the word 'immune,' for which at least one differential expression analysis provided a significant enrichment (FWER < 0.05), are shown with corresponding FWERs. Aggregation methods ('Lancaster TCC', Lancaster Tx' and 'Sidak Tx') are better at detecting "immune" enrichment than *p*-values derived from standard gene-level analysis ('Gene'). (b) TCC *p*-values aggregated by GO term ('Perturbation Test') reveal complementary information to classical GO enrichment ('Enrichment Test').

Declarations

Ethics approval and consent to participate

No data from humans were used in this manuscript.

Availability of Data and Materials

Scripts to reproduce the figures and results of the paper are available at http://github.com/pachterlab/aggregationDE/, which is under GNU General Public License v3.0. [34] The RNA-Seq datasets used in the analysis can be found at GEO GSE89024 [35] and GEO GSE95363. [36]

Competing Interests

The authors declare that they have no competing interests.

Funding

LY was partially funded by the UCLA-Caltech Medical Science Training Program, NIH T32 GM07616 and the Lee Ramo Fund. Harold Pimentel was partially funded by NIH R01 HG008140.

Authors' Contributions

LY, NLB and LP devised the methods. LY analyzed the biological data. LY and LP performed computational experiments. HP developed and implemented the simulation framework. LY and LP wrote the paper. NLB and LP supervised the research.

Acknowledgments

We thank Jase Gehring, Páll Melsted and Vasilis Ntranos for discussion and feedback during development of the methods. Conversations with Cole Trapnell regarding the challenges of functional characterization of individual isoforms were instrumental in launching the project.

References

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq:
a revolutionary tool for transcriptomics.
Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484.

[2] Soneson C, Love MI and Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015, 4:1521. doi: 10.12688/f1000research.7563.1.

[3] Kisielow M, Kleiner S, Nagasawa M, Faisal A, Nagamine Y. Isoform-specific knockdown and expression of adaptor protein ShcA using small interfering RNA. Biochem J. 2002 Apr 1; 363(Pt 1):1-5. doi: 10.1042/bj3630001.

[4] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10):R106. doi: 10.1186/gb-2010-11-10-r106.

[5] Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data, Bioinformatics. 2015; 31 (2), 166-169. doi: 10.1093/bioinformatics/btu638.

[6] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013 Jan;31(1):46-53. doi: 10.1038/nbt.2450.

[7] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May; 28(5):511-5. doi: 10.1038/nbt.1621.

[8] Robinson M., McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1; 26(1): 139–140. Published online 2009 Nov
11. doi: 10.1093/bioinformatics/btp616.

[9] Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017 Jul;14(7):687-690. doi: 10.1038/nmeth.4324.

[10] Anders S, Reyes A, Huber W.
Detecting differential usage of exons from RNA-seq data. Genome Res. 2012 Oct;22(10):2008-17. doi: 10.1101/gr.133744.111.

[11] Hess A, Iyer H. Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. Genome Biology. 2007; 8:96. doi: 10.1186/1471-2164-8-96.

[12] Chen Z, Yang W, Liu Q, Yang JY, Li J, Yang MQ. A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. BMC Bioinformatics. 2014;15(Suppl 17):S3. doi:10.1186/1471-2105-15-S17-S3.

[13] Dai H, Charnigo R, Srivastava T, Talebizadeh Z, Ye S.Q., Integrating P-values for Genetic and Genomic Data Analysis. J Biom Biostat. 2012; 3:e117. doi:10.4172/2155-6180.1000e117

[14] Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergman S, Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLOS Computational Biology. 2016. doi: 10.1371/journal.pcbi.1004714.

[15] Li S, Williams BL, Cui Y, A combined p-value approach to infer pathway regulations in eQTL mapping. Stat. Interface. 2011, 4, 389–402. doi: 10.4310/SII.2011.v4.n3.a13

[16] Lancaster, HO. The Combination of Probabilities: An Application of Orthonormal Functions. Australian and New Zealand Journal of Statistics. 1961 Apr. doi: 10.1111/j.1467-842X.1961.tb00058.x

[17] Bray N, Pimentel H, Melsted H, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016; 34, 525–527. doi:10.1038/nbt.3519.

[18] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15, pp. 550. doi: 10.1186/s13059-014-0550-8.

[19] Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant G.R. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015; 31(24), 3938-3945. doi: 10.1093/bioinformatics/btv488.

[20] Moll P, Ante M, Seitz A, Reda T. QuantSeq 3 [prime] mRNA sequencing for RNA quantification. Nature Methods. 2014 Dec 1;11(12).

[21] Dolinay T, Himes BE, Shumyatcher M, Lawrence GG, Margulies SS. Integrated Stress Response Mediates Epithelial Injury in Mechanical Ventilation. Am J Respir Cell Mol Biol. 2017 Aug;57(2):193-203. doi: 10.1165/rcmb.2016-0404OC.

[22] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al., Gene ontology: tool for the unification of biology. Nat Genet. 2000; 25(1):25-9. doi: 10.1038/75556.

[23] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13. doi: 10.1093/nar/gkn923.

[24] Mi H, Muruganujan A, Casagrande JT, Thomas PD, Large-scale gene function analysis with the PANTHER classification system. Nature Protocols. 2013; 8, 1551–1566. doi:10.1038/nprot.2013.092.

[25] Frahm KA, Waldman JK, Luthra S, Rudine AC, Monaghan-Nichols AP, Chandran UR. A comparison of the sexually dimorphic dexamethasone transcriptome in mouse cerebral cortical and hypothalamic embryonic neural stem cells. Mol Cell Endocrinol. 2017 May 26. pii: S0303-7207(17)30295-2. doi: 10.1016/j.mce.2017.05.026.

[26] Van den Berge K, Soneson C, Robinson MD, Clement L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. Genome Biol. 2017 Aug 7;18(1):151. doi: 10.1186/s13059-017-1277-0.

[27] Fisher RA. Statistical Methods for Research Workers. 4th edition. London: Oliver and Boyd; 1932.

[28] Šidàk, Z. Rectangular confidence region for the means of multivariate normal distributions. Journal of the American Statistical Association. 1967. 62, 626–633. doi:10.1080/01621459.1967.10482935.

[29] Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, *et al*. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501, 506–511. doi: 10.1038/nature12531.

[30] Johns Hopkins Center for Computational Biology. fqtrim. 2015; July 16. doi: 10.5281/zenodo.20552. https://github.com/gpertea/fqtrim/tree/v0.9.4

[32] Alexa A and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.28.0. 2016.

[33] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucl Acids Res. 2015; 43 Database issue D1049–D1056. doi: 10.1093/nar/gku1179.

[34] Yi L, Pimentel H, Bray NL, Pachter L. aggregationDE. Github. 2016; Feb 19. doi: 10.5281/zenodo.1179317. https://github.com/pachterlab/aggregationDE

[35] Dolinay T, Himes BE, Shumyatcher M, Lawrence GG, Margulies SS. Integrated Stress Response Mediates Epithelial Injury in Mechanical Ventilation. Quant Seq analysis of primary stretched rat alveolar type I-like epithelial (AEC-I) cells. 2016; Oct 22. GEO GSE89024.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89024

[36] Frahm KA, Waldman JK, Luthra S, Rudine AC, Monaghan-Nichols AP, Chandran UR. A comparison of the sexually dimorphic dexamethasone transcriptome in mouse cerebral cortical and hypothalamic embryonic neural stem cells. 2017, May 26. GEO GSE95363.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95363

Chapter IV

A discriminative learning approach to differential expression analysis for single-cell RNA-seq

Vasilis Ntranos[1,2^], Lynn Yi[3,4^], Páll Melsted[5] and Lior Pachter[4,6]

1. Department of Electrical Engineering & Computer Science, UC Berkeley, Berkeley, CA, USA
2. Department of Electrical Engineering, Stanford University, CA, USA
3. UCLA-Caltech Medical Science Training Program, Los Angeles, CA, USA
4. Division of Biology and Biological Engineering, Caltech, Pasadena, CA, USA
5. Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland
6. Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA, USA

^ Authors contributed equally

Corresponding author: Lior Pachter (lpachter@caltech.edu)

This page intentionally left blank.

Abstract

Single-cell RNA-seq makes it possible to characterize the transcriptomes of cell types across different conditions and to identify their transcriptional signatures via differential analysis. Our method detects changes in transcript dynamics as well as changes in overall gene abundance in large numbers of cells to determine differential expression. When applied to transcript compatibility counts obtained via pseudoalignment, our approach provides a quantification-free analysis of 3’ single-cell RNA-seq that can identify previously undetectable marker genes.

Single-cell RNA-seq (scRNA-seq) technology provides transcriptomic measurements at single-cell resolution, making possible the identification and characterization of cell types in heterogeneous tissue. The problem of identifying transcripts or genes that are differential between cell groups is analogous to the differential expression problem in bulk RNA-seq. Bulk RNA-seq differential expression methods can be applied directly to test transcripts or genes for differences between groups of cells[1], and methods that account for technical artifacts in scRNA-seq experiments such as by modeling dropout seem to offer some advantages.[2,3] However, one aspect of scRNA-seq that current methods do not take advantage of is the large number of cells sampled in single-cell experiments. Furthermore, current scRNA-seq methods are mostly based on quantifications of gene counts, thus precluding analysis of individual isoforms. In contrast, bulk RNA-seq is often performed to study dynamics of isoform expression, which have been shown to be important both in cell development [4] and pathology [5]. In the case of single-cell RNA-seq, isoform analysis is more complicated than in bulk RNA-seq[6] but just as important[7]. To investigate these transcript dynamics from bulk RNA-seq, methods have been developed to test for differential transcript usage (DTU) but such methods rely on sampling reads from across isoforms. One of the challenges with scRNA-seq is that many methods produce data only from the 3' ends of transcripts.

We show how prediction methods that take advantage of large numbers of cells and that fully exploit all the transcript information that can be extracted from reads, can greatly improve results both for differential gene expression and for differential transcript usage. We make use of logistic regression, which was considered when microarray gene expression assays were developed[8,9], but not pursued due to limited sample sizes. Now, scRNA-seq provides the large number of samples required to accurately fit a logistic regression model. Instead of the traditional approach of using the cell labels as covariates for gene expression, we perform logistic regression for each gene to predict cell labels from the quantifications of constituent transcripts when transcript quantifications can be accurately obtained. This is possible with certain technologies, such as SMART-Seq. Fitting the logistic regression model provides a linear combination of transcript quantifications that distinguishes cell groups, providing information about effect sizes of constituent transcripts, i.e. the "direction of change" (**Figure 1, Supplementary Figure 1**). Unlike traditional methods that test either for changes in overall gene abundance or for changes in transcript allocation, our method has the power to detect a change in any linear combination of transcript quantificiations and provides a unified testing framework that eliminates the need for a dichotomy between differential gene expression (DGE) and differential transcript usage (DTU) methods (**Supplementary Figure 2**).

In a simulation based on experimental effect sizes (see Methods), logistic regression outperforms other existing scRNA-seq differential expression methods, even with different normalizations (**Supplementary Figures 3, 4b**). In the case that isoforms move in concert, naïve gene quantification by summing of isoform counts performs similarly to logistic regression (**Figure 1a-c, Supplementary Figures 1a, 3e,f**), but logistic regression can also detect isoform switching (**Figure 1d-f,**

Supplementary Figures 1b, 3c-d, 5a). When applied to a dataset of differentiating myoblasts from Trapnell et al.,[10] the method reveals diverse transcript dynamics across multiple genes known to be important for myogenesis (**Figure 1h**). In addition to the Trapnell et al. dataset, we applied our method to a Smart-Seq2 dataset of embryonic cells in order to find genes that are differential between day 3 and day 4 post-fertilization, and compared the results with those of other methods. We showcase several genes undergoing isoform switching that are only found by our method (**Supplementary Figure 6)**. These results suggest that methods that test only for changes in overall gene expression are likely to miss a significant proportion of differential genes **(Supplementary Figure 2b).**

While transcript quantifications are biologically meaningful, in some cases they may be infeasible to obtain, such as in cases where only the 3' ends of transcripts are sequenced.[11,12] The reason is that transcripts of the same gene often share 3'UTRs and therefore cannot be differentiated solely from 3' end sequences. We therefore examined the possibility of performing logistic regression directly on the transcript compatibility counts (TCCs) obtained via pseudoalignment. This is a procedure that for each read, finds a set of transcripts from which the read could have originated.[13] The sets of compatible transcripts are called "equivalence classes", and the TCCs correspond to the number of reads that map to each equivalence class.[14] TCCs were used by Ntranos et al.[15] as a more accurate, technology-independent transcriptomic signature for single cell clustering, since, unlike transcript quantifications, TCCs do not depend on a specific coverage model. In the case of 3' sequencing where transcript quantification is infeasible, TCCs can be readily obtained via pseudoalignment, maintaining the isoform-level information that is available in the data.[15,16]

On simulated data, the performance of logistic regression with TCCs is comparable with that of other methods (**Supplementary Figure 7a-b**). Furthermore, it has more power to detect isoform-switching (**Supplementary Figure 7c-d**). To investigate whether logistic regression with TCCs confers an advantage over gene-count based differential analysis from such data, we examined 10X Chromium scRNA-seq from three human T-cell populations that were purified using antibodies specific to different isoforms of PTPRC (CD45).[11] Using TCCs, we performed pairwise differential analyses of purified CD45RO+ memory helper T-cells, CD45RA+ naïve helper T-cells, and CD45RA+ naïve cytotoxic T-cells, providing two positive controls (CD45RA+ vs CD45RO+) and a negative control (CD45RA+ vs CD45RA+) for the method. Logistic regression was able to detect differential expression of *CD45* in the purified CD45RO+ memory and CD45RA+ naïve T-cell populations (**Figure 2a,b**). This result was deemed impossible in Peterson et al.,[17] where it was noted that 3' mRNA sequencing alone could not resolve these markers. We confirmed that gene counts alone could not identify *CD45* as differential, and furthermore we found that independent testing of TCCs reduced statistical power. In contrast, when testing CD45RA+ naïve helper T-cells and CD45RA+ naïve cytotoxic T-cells (**Figure 2c**), *CD45* was not found significant with any method for this subsample, and there was little difference in overall p-value distribution between independently testing equivalence classes and performing multiple logistic regresion. A power analysis showed that logistic regression with TCCs finds *CD45* to be differentially expressed after multiple testing correction (FDR < 0.01). At any cell number, performing logistic regression on gene counts failed to find *CD45*. A distribution of the p-values obtained by each method suggests that while both methods find genes with the largest change in overall gene expression, logistic regression using TCCs detects isoform switching **(Supplementary Figure 8).**

Further examination of the transcripts corresponding to the ECs identified by our method pointed to transcripts that were differentially expressed **(Supplementary Figure 9).** Visual inspection of the differential ECs identified for *CD45* revealed that the corresponding isoforms were being distinguished by virtue of alternative unannotated 3' untranslated regions (UTRs) **(Supplementary Figure 9a-b).** To quantify the extent of isoform accessibility by 3'-end sequencing, we estimated the distribution of read pseudoalignments with respect to the annotated 3'UTRs **(Supplementary Figure 9d).** We found a substantial number of reads farther from 3'ends than expected, which points to a large number of unannotated 3' UTRs. To mitigate the effect of unannotated 3' UTRs on our analysis, we updated the transcriptome with novel 3' UTRs for *CD45* and redid the analysis. *CD45* remained differential (**Supplementary Methods, Supplementary Figures 10, 11).** Our results are concordant with previous work on lymphocytic surface receptor isoform diversity.[18] Equivalence classes provide access to isoforms in genes other than *CD45*, and we found multiple other genes that also exhibited isoform switching between memory and naïve T cells **(Supplementary Figure 12).**

To examine whether TCCs are informative in a *de novo* scRNA-seq experiment, we analyzed a PBMC dataset that consists of 68579 cells sequenced at an average of 20491 reads per cell.[11] After clustering and using known cell type markers to annotate the clusters **(Supplementary Figure 13)**, we were able to recapitulate our previous *CD45* differential analysis: *CD45* was identified as differential between memory and naïve T-cells **(Supplementary Figure 14),** showing that TCC-based logistic regression can be applied to cell groups generated by unsupervised clustering as in standard, *de novo* scRNA-seq workflows.

Logistic regression is especially powerful for scRNA-seq since it leverages the large number of cells available in scRNA-seq experiments and incorporates isoform information for gene-level testing. It reveals the contribution of individual isoforms to the gene-level differential analysis, aiding in interpretability of results. While we have demonstrated the power of logistic regression for performing gene-level differential expression between two cell types, the method extends to more general groupings. Furthermore, logistic regression can be performed on all genes simultaneously to discover gene markers characterizing cell types. Finally, our method scales effectively with both the number of reads and cells, which is critical for processing increasingly large scRNA-seq datasets **(Supplementary Figure 5b-c).**

Methods

Model

We model cell membership as a function of gene expression. Let $X_{\mathrm{t,i}}$ be the transcript abundance for transcript t in cell i. Let y_i be the indicator variable for the membership of cell i. Then for a gene g, the transcript abundances are linear predictors of cell membership

$$y_i \mid X_{t,i} \sim Bernoulli(\sigma(c + \textstyle\sum_{t \in g} B_t X_{t,i})),$$

where $\sigma(t)$ is the logistic function defined as $\sigma(t) = 1/(1 + e^{-t})$. This framework is multiple logistic regression, where for each gene, the number of predictors is equal to the number of transcripts in the gene. To obtain significance scores for each gene, we perform a likelihood ratio test. The null model used for the likelihood ratio test is that cell membership does not depend on gene expression:

$$y_i \sim Bernoulli(\sigma(c)).$$

For each gene, the difference in the degrees of freedom between the alternate model and the null model is therefore equal to the number of transcripts contained in the gene.

In the case where transcript quantifications are not available, TCCs may be used instead. Let $T_{\mathrm{e,i}}$ be the TCC for equivalence class e in cell i. The TCC models are

$$y_i \mid T_{e,i} \sim Bernoulli(\sigma(c + \textstyle\sum_{e \in g} B_e T_{e,i})).$$

Our null models remain $y_i \sim Bernoulli(\sigma(c))$ and the differences in the degrees of the freedom are equal to the number of TCCs associated with each gene g.

Trapnell et al. 2014 analysis

We downloaded the preprocessed Trapnell et al. 2014 data from the conquer[1] database, which included the quantified transcript-per-million (TPM) values and cell labels for 222 serum-induced primary myoblasts over a time course of 0, 24 and 48 hours. We selected the 85 myogenic precursor cells and the 97 differentiating myoblast cells for differential expression analysis. We used Ensembl Homo_sapiens.GRCh38.rel84.cdna.all.fa to group 176241 transcripts into 38694 genes and tested each gene for differential expression between myogenic precursors and differentiating myoblasts using our method. Logistic regression was run using *sklearn.linear_model.LogisticRegression()*. After Benjamini-Hochberg correction, we obtained 1308 significant differential genes (< 0.01 FDR). We visualized these genes in a circle plot by performing logistic regression on the primary and secondary isoforms, which are defined as the isoforms with the largest and second largest average expression over all cells.

Zheng et al. 2017 analysis

We obtained the raw reads for the three human PBMC purified cell sub-type datasets described in Zheng et al., 2017, CD4+/CD45RA+/CD25- naïve T-cells, CD4+/CD45RO+ memory T-cells and CD8+/CD45RA+ naïve cytotoxic T-cells, from https://support.10xgenomics.com/single-cell-gene-expression/datasets. The reads were preprocessed (barcode detection, error-correction and pseudoalignment) with the scRNA-Seq-TCC-prep kallisto wrapper (SC3Pv1 chemistry) to obtain the single cell transcript compatibility counts (TCC) matrix (https://github.com/pachterlab/scRNA-Seq-TCC-prep). After filtering out cells with total UMI counts outside the interval [1K-30K], we obtained 31831 cells (9923 CD4+/CD45RA+/CD25- naïve T-cells, 9994 CD4+/CD45RO+ memory T-cells and 11914 CD8+/CD45RA+ naïve cytotoxic T-cells respectively). We selected all the equivalence classes that contained at least one isoform associated with the *CD45* gene (also known as *PTPRC, ENSG00000081237, ENSG00000262418*) and filtered out the ones with total UMI counts less than 0.25% of the total number of cells, i.e. equivalence classes with fewer than ~79 UMI counts across all cells. This resulted in seven equivalence classes uniquely associated with subsets of the annotated isoforms of *CD45*. The gene counts for each cell were obtained by summing the TCCs. We performed all three pairwise tests for differential expression between the purified cell sub-types using a multiple logistic regression model on the seven TCCs, a logistic regression model on the aggregated gene counts, and a logistic regression model independently on each equivalence class. Logistic regression was run using *sklearn.linear_model.LogisticRegression()*, and the likelihood ratio test was used to obtain p-values for all three tests, as described in the "Model" section. For each pairwise test, we randomly subsampled 3000 cells per group across 200 independent subsamples to generate *p*-value distributions for each method.

The raw reads for the 68k PBMC dataset were preprocessed with the scRNA-Seq-TCC-prep kallisto wrapper to obtain the TCC matrix. Equivalence classes that mapped to multiple Ensembl gene names and cells with total UMI counts outside the interval [2K-20K] were filtered out. The resulting 65444 cell by 95426 EC matrix was subsequently used for post-processing and clustering with scanpy 0.2.6.[19] We used the same steps outlined in the "Zheng et al. recipe" that is provided by scanpy, except we selected the 5000 most variable equivalence classes in lieu of the 1000 most variable genes. To verify the clustering structure, we plotted the cells with t-SNE using specific marker gene abundances obtained by summing all the constituent TCCs. Supplementary Figure 14 focuses on the clusters that most likely correspond to populations of naïve cytotoxic T-cells (Cluster A, CD8A+/CD4-/CCR7+, 5226 cells), naïve helper T-cells (Cluster B, CD4+/CCR7+, 12424 cells) and memory helper T-cells (Cluster C, CD4+/S100A4+/CCR10+, 4173 cells). Clusters A and C corresponded to clusters 3 and 6 in Supplementary Figure 13, whereas cluster B was obtained by manually merging clusters 1 and 2. We performed pairwise differential expression tests between these three clusters using multiple logistic regression on TCCs and the likelihood ratio test (see 'Model' section for construction of the likelihood ratio test). To compare to our method, we also performed logistic regression on gene counts, and independent logistic regressions on each TCC followed by Bonferroni correction. P-value distributions were obtained from performing these three differential expression tests across 200 subsamples, each time subsampling 2000 cells per cluster.

In order to estimate the distribution of read distances to the 3' end, we pseudoaligned the reads from the three purified T-cell populations to the transcriptome using the pseudobam option of kallisto 0.44.0. In the case of read multiple alignment, the weight of the read was split evenly across all reported transcripts. The distance to the 3' end was inferred from the transcriptome coordinates reported in the BAM file.

To detect novel 3' UTR ends for *CD45*, we identified reads whose alignment extends past the UTR into the polyA-tail and that kallisto pseudoaligned to the *CD45* gene. These reads were clustered according to their pseudoaligned genomic coordinates. After discarding clusters corresponding to known 3'UTR ends, 3 clusters remained corresponding to unannotated 3' UTR ends, containing 69, 71 and 97 reads respectively. For each of these clusters, we removed the polyA-tail and generated a consensus sequence via multiple alignment using FSA[20]. The consensus sequence was aligned to the genome to determine the genomic coordinates of the novel 3' UTR endpoint. The reference transcriptome was modified by creating a new version of each transcript belonging to *CD45* that overlapped the new 3' UTR endpoint, resulting in 13 novel transcripts added. For visualization purposes, we also ran kalliso pseudobam with the updated the GTF file (Supplementary Figure 10).

Petropoulos et al., 2016 analysis

We downloaded the Petropoulos et al., 2016 dataset[21] from the conquer[1] database, which contains quantifications for 1529 human preimplantation embryonic cells. We used the provided Ensembl transcript and gene quantifications (counts and TPMs) to perform differential analysis between the day three embryonic cells and the day four embryonic cells (271 total cells). The differential expression methods were run with the same normalization and filters as with the simulations (see below). The method *glm* from R's native *stats* library was used to perform logistic regression, by using the parameter *family='binomial'* with its default logit link function. UpSetR 1.3.3[22] was used to plot the size of the intersection sets of the 3000 most significantly differential genes.

Read simulation framework

We developed a scRNA-seq simulation framework that can simulate reads (https://github.com/pachterlab/NYMP_2018/tree/master/simulations). Parameters for the simulator were estimated using data from Trapnell et al., 2014. In each simulation, cells were simulated from two different cell groups: a null group and a perturbed group, each with 105 cells. The null type was modeled after the cluster of proliferating myoblasts from the Trapnell et al., 2014 dataset. Specifically, after quantification of the dataset using kallisto and clustering on TCCs, the cluster containing 105 cells with *MYOG* expression was identified and used as the basis of our simulations.

The nonzero TPMs from the myoblast cluster were used to estimate the parameters of a lognormal distribution for each transcript. To simulate the null cell type, TPMs for each transcript were drawn from a lognormal distribution. Then, for each transcript, a subset of cells were chosen at random in which the transcript abundance was set to 0 (‘dropout’). The percentage of dropout for each transcript was matched to the experimental dataset. Mathematically, given d_t as the dropout rate for transcript t, and μ_t and σ_t parameterizing the mean and variance of the expressed component, the transcript expression is modeled as:

$$x_t | \, \mu_t, \sigma_t, d_t = \begin{cases} 0 & with\, probabiltiy\; d_t \\ lognormal(\mu_t, \sigma_t) & with\, probability\; (1 - d_t) \end{cases}.$$

The use of lognormal distribution on TPMs was motivated by the Tobit model in Monocle[2], the mixture of dropout and expression components was motivated by SCDE, and the masking of random cells with 0's to reach sufficient dropout was modeled after Splatter. The mean variance plots and the distribution of 0's are greatly concordant between the simulations and the experimental data after which they are modeled (Supplementary Figure 15).

Three different types of simulated data were prepared to reflect distinct perturbation scenarios and effect sizes. Transcripts expressed in fewer than 5 of the 105 cells were deemed too lowly expressed and filtered out from the perturbation. In the independent effects simulation, 30% of the transcripts that passed the filter (20456 out of 68179 expressed transcripts) were chosen at random to be perturbed. For each transcript, a minimum effect size of 2-fold was drawn from a truncated lognormal distribution. The direction of each perturbation was chosen uniformly at random (50% upregulated, 50% downregulated). In the correlated effect simulations, genes with all transcripts passing the filter also passed the filter. 30% of remaining genes (~5220 of 17390 genes) were chosen at random to be perturbed and expressed transcripts (defined as expressed in >= 5 cells) of that gene were perturbed with the same effect size drawn from a truncated log normal distribution at a minimum of 2. In the experiment-based simulations, the effect sizes were learned from Trapnell et al., 2014 from the set of transcripts that DESeq2[23] found to be differentially expressed (p-value < 0.05). The same transcripts are perturbed with their DESeq2-derived effect sizes in the simulation.

The effect sizes were applied to the mean expression, and abundances per cell were generated by sampling from lognormal distribution truncated at zero. Given these cell-by-cell abundances, RSEM[24] was used to generated paired-end reads uniformly distributed across transcripts using a model learned from a proliferating myoblast cell from the Trapnell et al., 2014 data set and a background noise read percentage (parameter theta) of 20%. The number of reads per cell was learned from the myoblast cluster by fitting a lognormal distribution of reads per cell ($\mu = 14.42$, $\sigma = 0.336$), corresponding to a mean of 193,000 paired-end reads per cell.

Splatter Simulation Framework

We also used Splatter[25], which simulates transcript counts directly instead of reads. The same 105 myoblasts from Trapnell et al., 2014 used to model the simulations above were used to fit Splatter simulation parameters. Transcripts with more than 90% zeros were filtered from the simulation, leaving 47606 transcripts to be simulated. We used Splatter's default parameters to simulate two groups, i.e. 0.1 chance of perturbation in each group, resulting in 9095 perturbed transcripts and corresponding to 19% perturbation rate across the two groups. The 47606 transcripts were randomly assigned to 15420 genes according the transcriptomic structure and transcript counts were summed to provide gene counts. These transcripts correspond to 6393 perturbed genes across 15420 total genes.

Simulation Analyses and Benchmarking

Logistic regression, Monocle's Tobit model,[10] DESeq2 1.16.11,[23] MAST 1.2.1,[3] and SCDE 1.99.4[2] were used to benchmark the simulations in R. Monocle's Tobit model method, DESEq2, and MAST were invoked using Seurat's wrapper functionality through *Seurat::FindMarkers*.[26] The method *glm* from R's native *stats* library was used to perform logistic regression, by using the parameter *family='binomial'* with its default logit link function.

The FASTQ files output from the RSEM simulations were quantified using kallisto v0.44.0. tximport[27] was used to aggregate transcript-level counts and abundances into gene-level counts and abundances prior to inputting into the various methods. In contrast, the Splatter[25] simulation did not require read quantification as transcripts counts were directly simulated. In order to afford each method its optimal input, normalizations native to each method were used. For SCDE, and DESeq2, the gene counts were used as input. For Monocle and MAST, the TPM abundances were used as input. For our method, we used DESeq2's library size method of normalization on transcript counts prior to performing logistic regression. To apply DESeq2's method, size factors were calculated based on the transcript counts using *DESeq2::calculateSizeFactors*, and the normalized counts were obtained by dividing by the cell's size factor. For all methods, we filtered out genes/transcripts with zero expression in >90% of cells from the analysis with logistic regression.

To perform logistic regression using TCCs, we filtered out ECs that contained transcripts from multiple genes and ECs with >90% zeros. Additionally, genes with fewer than 4 cells per TCC were filtered from analysis. TCCs were normalized with DESeq2's size factor method.

We benchmarked the accuracy of the methods by evaluating their tradeoff between sensitivity and false discovery rate (FDR). FDR is defined as the number false positives divided by the number of total declared positives. We ranked the genes by significance, i.e. lowest to highest p-value, and then calculated and plotted the FDR and sensitivity at each level of significance.

In addition to benchmarking the methods by accuracy, we evaluated the runtimes of the methods on the Splatter simulation. Every method was run in series three times on the same dataset on a machine with 40 cores and 350 GB. Their runtimes were benchmarked with R's *system.time()*. All methods were run using a single core, except SCDE which was run with its default 20 cores. The real elapsed time and the total processing time, calculated as the sum of the user time and the system time, were plotted in Supplementary Figure 4.

Figures

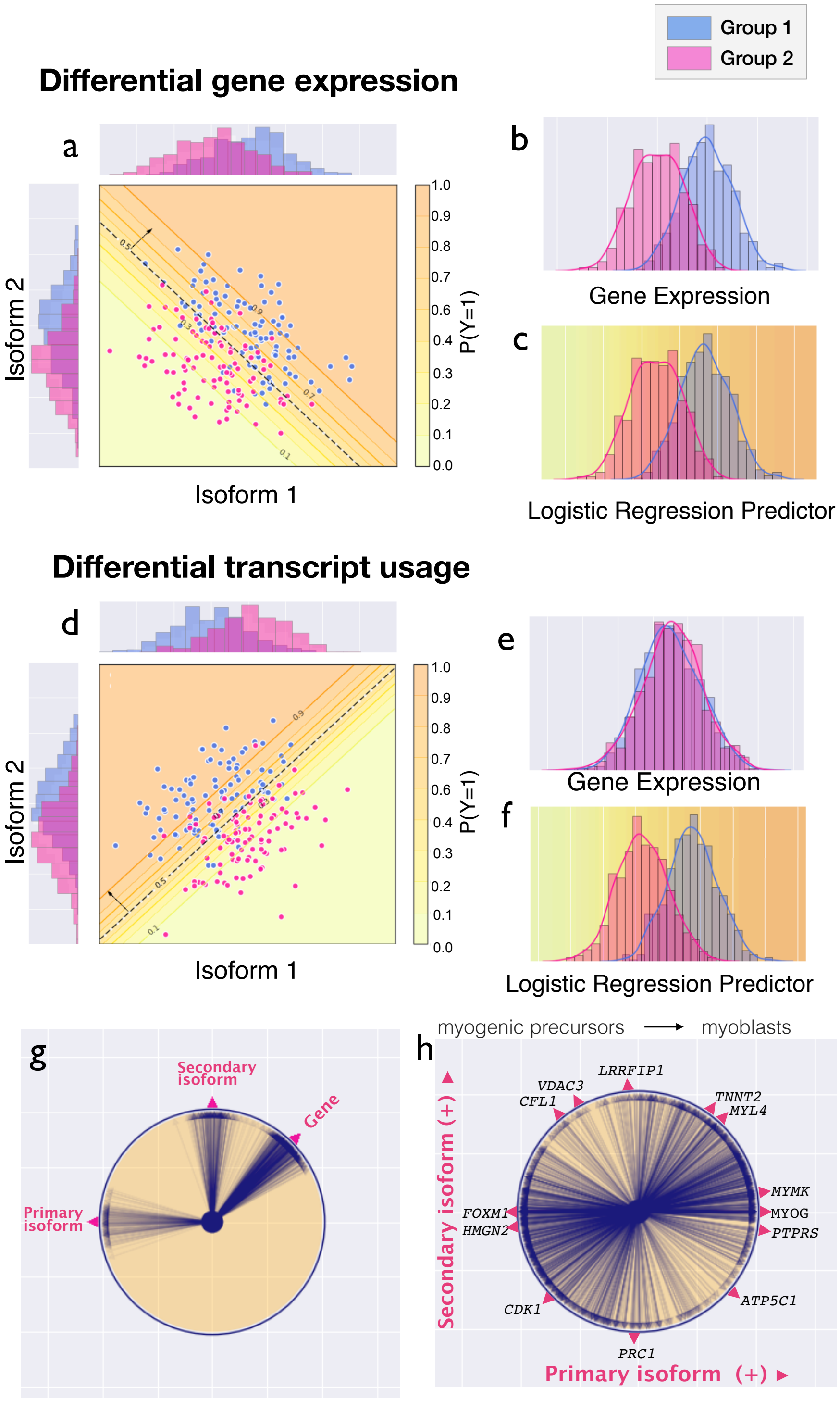

Group 1
Group 2
Differential gene expression
a
Isoform 2
Isoform 1
P(Y=1)
1.0
0.9
0.8
0.7
0.6
0.4
0.3
0.2
0.1
0.0
b
Gene Expression
c
Logistic Regression Predictor
Differential transcript usage
d
Isoform 2
Isoform 1
P(Y=1)
1.0
0.9
0.8
0.7
0.6
0.4
0.3
0.2
0.1
0.0
e
Gene Expression
f
Logistic Regression Predictor
g
Secondary isoform
Gene
Primary isoform
h
myogenic precursors → myoblasts
LRRFIP1
VDAC3
CFL1
TNNT2
MYL4
MYMK
MYOG
PTPRS
FOXM1
HMGN2
ATP5C1
CDK1
PRC1
Secondary isoform (+)
Primary isoform (+)

Figure 1: Logistic regression applied to scRNA-seq. Logistic regression can be used to detect gene differential expression at isoform level resolution. Panels **(a-c)** show a hypothetical scenario with two cell groups, ‘Group 1 and ‘Group 2’, colored blue and pink respectively, where both isoforms of a gene change with the same effect size. In panel (**a**), each cell is plotted by its two isoforms’ abundances. The dashed line along the x=y indicates the direction of change inferred by logistic regression. The orange shading corresponds to the probability of being in Group 1 under the logistic regression model; cells farther above the dashed line are more likely to be in Group 1 (darker orange) and cells farther below the dashed line are more likely to be in Group 2 (lighter orange). Panel **(b)** shows the histogram of gene abundances. Panel (**c**) shows the histogram of the linear combination of transcript abundances learned by logistic regression along with the same probability gradient as in panel (**a**). In this scenario, the linear combination found by logistic regression is the same as the summed gene abundances. Panels **(d-f)** depict another scenario where two isoforms have effect sizes in opposite directions, i.e. isoform switching. Panel (**e**) is a histogram of gene abundance and panel (**f**) is a histogram of the linear combination of transcripts from logistic regression. In (**g**), transcript dynamics for 1000 genes are visualized via a circle plot, in which the directions of arrow correspond to the direction of the change for each gene. The x-axis corresponds to primary isoform and the y-axis corresponds to the secondary isoform. Panel **(h)** shows the directions of change of 1308 genes from the Trapnell et al. data set that were identified by logistic regression as differentially

expressed between myogenic precursors and differentiating myoblasts. Pink arrows corresponding to known myogenic genes are marked along the circle.

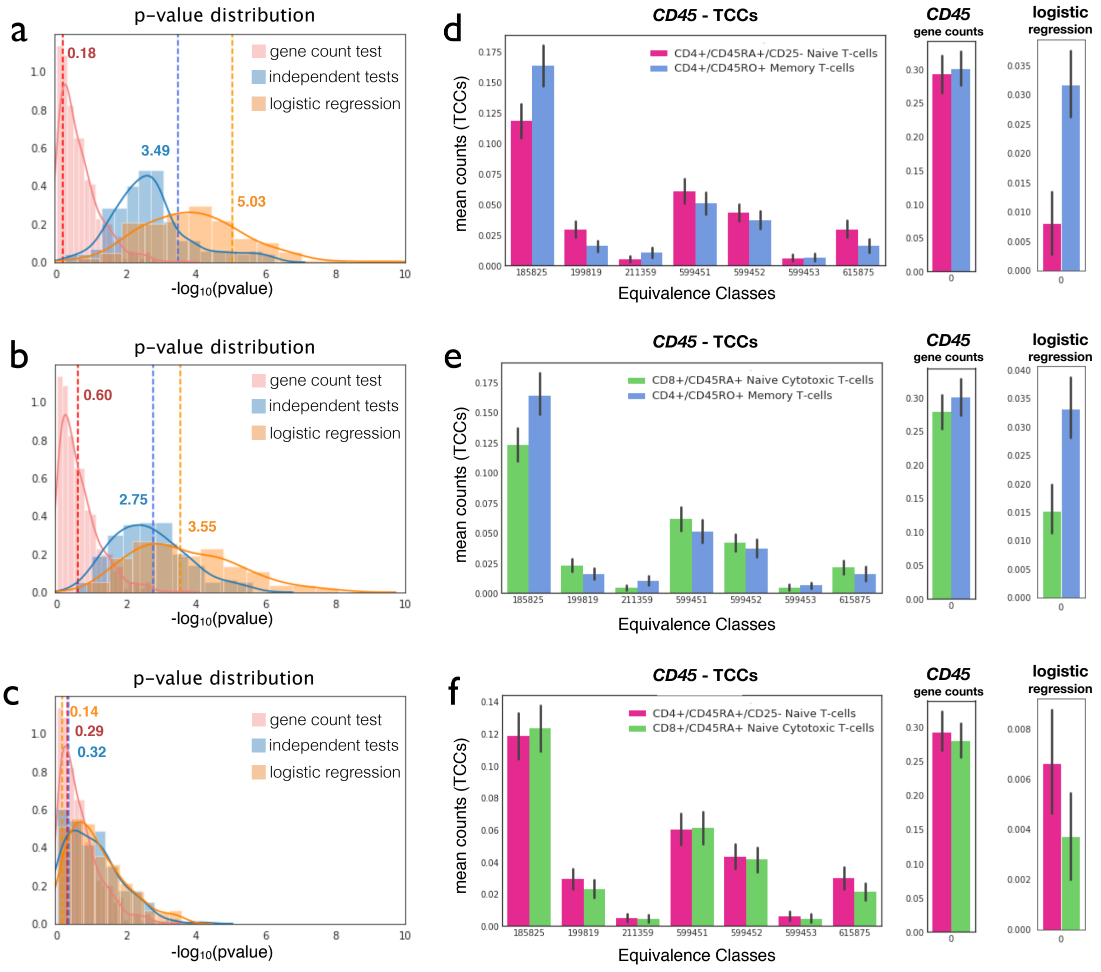


Figure 2: Logistic regression discovers *CD45* in purified T-cell types. Pairwise differential expression analysis was performed between purified memory helper T-cells, naïve helper T-cells, and naïve cytotoxic T-cells that were sequenced with 10X. In panels **(a, b, c)**, p-value distributions corresponding to three different differential expression methods were generated from 200 subsamples. Each subsample contained 3000 cells of each cell type from the full dataset of 9923 naïve helper T-cells, 9994 memory helper T cells, and 11914 naïve cytotoxic T cells. The three methods included

our method of multiple logistic regression on TCCs (‘logistic regression’), logistic regression on gene counts (‘gene count test’) and logistic regression on each equivalence class followed by Bonferonni correction (‘independent tests’). Bar plots corresponding to the expression profiles of a specific subsample are shown in **(d, e, f)**, where the whiskers correspond to the 95% confidence interval of the mean expression across 3000 cells. The p-values corresponding to this particular subsample are marked on the p-value distribution.

Declarations

Code Availability

The code required to perform the simulations and reproduce the analyses are available at https://github.com/pachterlab/NYMP_2018. We also provide the Github repository that was zipped at the time of manuscript acceptance as Supplementary Software.

Data Availability

The myogenesis dataset (Trapnell et al., 2014) is available on conquer database and on GEO as Series GSE52529. The dataset on embryogenesis is available on the conquer database (Petropoulos et al., 2016). The 10x dataset on PBMCs is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets.

Acknowledgments

We thank N. Bray, J. Gehring and V. Svensson for discussion and comments on the manuscript, and H. Pimentel for assisting with the simulations. We thank A. Butler and R. Satija for implementing this method in Seurat.

Author Contributions

V.N. developed the model during discussions with L.Y. and L.P. V.N. performed analysis of the 10x PBMC dataset. L.Y. performed the simulations and analysis of the embryo SMARTseq dataset. P.M. developed kallisto genomebam and assisted with analysis. All authors contributed extensively in the interpretation of the results and wrote the manuscript.

Competing Financial Interests Statement

The authors declare no competing interests.

References

[1] Soneson C, Robinson MD, Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data. Nat Methods. 2018 Apr;15(4):255-261. doi: 10.1038/nmeth.4612

[2] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014 Jul;11(7):740-2. doi:10.1038/nmeth.2967.

[3] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015; 16: 278.

[4] Yamazaki T, Liu L, Lazarev D, Al-Zain A, Fomin V, Yeung PL et al. TCF3 alternative splicing controlled by hnRNP H/F regulates E-cadherin expression and hESC pluripotency. Genes Dev. 2018 Sep 1;32(17-18):1161-1174. doi: 10.1101/gad.316984.118.

[5] Vitting-Seerup K, Sandelin A. The Landscape of Isoform Switches in Human Cancers.
Mol Cancer Res. 2017 Sep;15(9):1206-1220. doi: 10.1158/1541-7786.MCR-16-0459.

[6] Arzalluz-Luque A, Conesa A. Single-cell RNAseq for the study of isoforms—how is that possible? Genome Biol. 2018;19: 110. doi: 10.1186/s13059-018-1496-z.

[7] Gupta I, Collier PG, Haase G, Mahfouz A, Joglekar A, Floyd T et al. Single-cell isoform RNA sequencing (ScISOr-Seq) across thousands of cells reveals isoforms of cerebellar cell types. bioRxiv. 2018 Jul; doi: 10.1101/364950.

[8] Xing E, Jordan MI, Karp RM. Feature Selection for High-Dimensional Genomic Microarray Data. ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning. 2001 June.

[9] Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics, Volume 19, Issue 17, 22 November 2003, Pages 2246–2253, doi:10.1093/bioinformatics/btg308.

[10] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381-386. doi: 10.1038/nbt.2859.

[11] Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Jan 16;8:14049. doi:10.1038/ncomms14049.

[12] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M et al., Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202-1214. doi: 10.1016/j.cell.2015.05.002.

[13] Bray N, Pimentel H, Melsted H, Pachter L. Near-optimal probabilistic RNA-Seq quantification. Nat Biotechnol. 2016; 34, 525–527. doi:10.1038/nbt.3519.

[14] Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms Mol Biol. 2011 Apr 19;6(1):9. doi: 10.1186/1748-7188-6-9.

[15] Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome Biol. 2016 May 26;17(1):112. doi:10.1186/s13059-016-0970-8.

[16] Yi L, Pimentel H, Bray NL, Pachter L. Gene-level differential analysis at transcript-level resolution. Genome Biol. 2018 Apr 12;19(1):53. doi: 10.1186/s13059-018-1419-z.

[17] Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA. Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnol. 2017 Oct;35(10):936-939. doi:10.1038/nbt.3973.

[18] Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun. 2017 Jul 19;8:16027. doi:10.1038/ncomms16027.

[19] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome biology. 2018 Dec;19(1):15.

[20] Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. Fast statistical alignment. PLoS Comput Biol. 2009 May;5(5):e1000392. doi:10.1371/journal.pcbi.1000392.

[21] Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, Plaza Reyes A, Linnarsson S, Sandberg R, Lanner F. Single-Cell RNA-Seq Reveals Lineage

and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016 May 5;165(4):1012-26. doi:10.1016/j.cell.2016.03.023.

[22] Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017 Sep 15;33(18):2938-2940. doi:10.1093/bioinformatics/btx364.

[23] Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

[24] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12:323. doi:10.1186/1471-2105-12-323.

[25] Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017 Sep 12;18(1):174. doi: 10.1186/s13059-017-1305-0.

[26] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015 May;33(5):495-502. doi:10.1038/nbt.3192.

[27] Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015 Dec 30;4:1521. doi:10.12688/f1000research.7563.2.

This page intentionally left blank.

Chapter V

A direct comparison of genome alignment and transcriptome pseudoalignment

Lynn Yi[1,2], Lauren Liu[1], Páll Melsted[3] and Lior Pachter[1]

1. Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA
2. UCLA-Caltech Medical Scientist Training Program, Los Angeles, CA, USA
3. Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavik, Iceland

Corresponding author: Lior Pachter (lpachter@caltech.edu)

This page intentionally left blank.

Abstract

Motivation

Genome alignment of reads is the first step of most genome analysis workflows. In the case of RNA-Seq, transcriptome pseudoalignment of reads is a fast alternative to genome alignment, but the different "coordinate systems" of the genome and transcriptome have made it difficult to perform direct comparisons between the approaches.

Results

We have developed tools for converting genome alignments to transcriptome pseudoalignments, and conversely, for projecting transcriptome pseudoalignments to genome alignments. Using these tools, we performed a direct comparison of genome alignment with transcriptome pseudoalignment. We find that both approaches produce similar quantifications. This means that for many applications genome alignment and transcriptome pseudoalignment are interchangeable.

Availability and Implementation

bam2tcc is a C++14 software for converting alignments in SAM/BAM format to transcript compatibility counts (TCCs) and is available at https://github.com/pachterlab/bam2tcc. kallisto genomebam is a user option of kallisto that outputs a sorted BAM file in genome coordinates as part of transcriptome pseudoalignment. The feature has been released with kallisto v0.44.0, and is available at https://pachterlab.github.io/kallisto/.

Introduction

Read alignment programs are used to locate the genome coordinates from which a sequenced read could originate (e.g., Langmead and Salzberg, 2012). In the case of RNA-Seq, the sequenced reads correspond to cDNA that have been reverse transcribed from mRNA. Because splicing occurs as part of post-transcriptional processing, the alignments to the genome may span multiple exons and skip the introns between the exon-exon junctions. The task of aligning RNA-Seq reads to the genome in a way that is robust to splicing is known as "genome spliced alignment." There are several programs that perform this task, such as TopHat/TopHat2 (Trapnell *et al.*, 2009; Kim *et al.*, 2013), STAR (Dobin *et al.*, 2013), and HISAT/HISAT2 (Kim *et al.*, 2015). When spliced alignment is used for RNA-Seq, subsequent analysis is required to assign reads to genes (e.g., Liao *et al.*, 2014) or to transcripts (e.g., Trapnell *et al.*, 2010) as part of quantification.

An alternative to align reads to the genome is to align reads directly to the transcriptome. The transcriptome is defined as the set of sequences corresponding to mature mRNA after post-transcriptional processing. Methods such as eXpress (Roberts and Pachter, 2013) and RSEM (Li and Dewey, 2011) use transcriptome alignments for read assignment under a quantification model. In previous benchmarks of RNA-Seq quantification methods, it has been unclear whether performance improvements from transcriptome alignment are due to the mode of alignment or to a different quantification model (Teng *et al.*, 2016).

In 2016, Bray *et al*. introduced the concept of pseudoalignment to the transcriptome, which, rather than a performing a full alignment, records information about the set of transcripts a read is compatible with. Specifically, in pseudoalignment, reads are assigned to these sets of transcripts, i.e. “equivalence classes” of transcripts (Nicolae *et al.*, 2011). Transcript compatibility counts (TCCs) constitute the number of reads within the equivalence classes and serve as sufficient statistics for transcript quantification. (See Figure 1a for workflow of pseudoalignment and quantification.) Transcriptome pseudoalignment is orders-of-magnitude faster than traditional genome alignment (Bray *et al.*, 2016; Patro *et al.*, 2017); however while pseudoalignment has increased in popularity since its introduction two years ago (Vivian *et al.*, 2017; Lachmann *et al.*, 2018), genome alignment programs are still widely used.

While there have been comparisons of genome alignment and pseudoalignment methods (Teng *et al.*, 2016; Bray *et al.*, 2016; Patro *et al.*, 2017), the benchmarks have examined only the final quantifications and have not teased apart the algorithmic components. The output of peudoalignment, in the form of TCCs, is conceptually different from genome alignments. To compare genome alignments directly to pseudoalignments, one must perform a conversion of the underlying data models, a task that is considerably more complicated than converting file formats. Furthermore, procedures for quantification after genome alignments are fundamentally different than those that are used with pseudoalignments, making a direct comparison challenging.

Results

To compare alignment to pseudoalignment methods, we created a tool, bam2tcc, that converts genome alignments in the format of a BAM or SAM file to transcript compatibility counts, the primary output of transcriptome pseudoalignment. We then quantified genome alignments and transcriptome pseudoalignments using the exact same model and method (Figure 1b), thus separating the effects of alignment from those of quantification. We used bam2tcc to convert HISAT2 and STAR genome alignments into transcript compatibility counts, which were then quantified using the expectation maximization (EM) algorithm for a uniform coverage model (Bray *et al.*, 2016). We chose HISAT2 and STAR because of their popularity as well as their accuracy in previous benchmarks (Baruzzo *et al.*, 2017).

We compared the accuracy of the genome spliced alignment programs HISAT2 and STAR and transcriptome pseudoalignment programs kallisto and Salmon on simulations where the true abundances are known. We ran the methods with default parameters and with minimal parameterization. Performance of aligners and pseudoaligners on the simulations were comparable, as demonstrated by their mean absolute relative differences (MARDs) on expressed transcripts. We separately benchmarked accuracy on transcripts that were not expressed in the simulation by examining their distances from zero and taking the mean of this distance across all transcripts. On both measures, across 10 simulated samples, the results of methods were highly concordant with each other (Figure 1b-e), with the exception of STAR.

kallisto, Salmon and HISAT2 were more accurate than STAR (Tables 1-4), demonstrating that transcriptome pseudoalignment methods can outperform genome alignment methods.

We also compared the results of the four methods on experimental RNA-Seq data from Zika-infected human neuroprogenitor cells (Tang *et al.*, 2016; Yi *et al.*, 2017). Since a simulation cannot capture all sources of variance in an experiment, the inter-method correlations on experimental data were lower than on simulated data. Nonetheless, the cross-method correlations show that the methods still produced concordant quantifications (Tables 5-6).

Examining the TCCs derived by each method on this experimental dataset also shows concordance on the level of the TCCs, even prior to quantification. We examined the number of transcripts in each equivalence classes as proxy for the uncertainty in read assignment. The distribution of equivalence class size, defined as the number of transcripts per equivalence class, weighted by the counts for the equivalence class, was similar for all methods (Figure 2a). The weighted mean equivalence class size was similar across all methods and showed that each read on average is compatible with 3.8 ambiguous transcripts (Figure 2b). An examination of the intersections of identical equivalence classes across methods showed that the majority of equivalence class were shared amongst all methods, and that almost all of the reads were in equivalence classes that were common to all methods (Figure 2c).

One feature of genome alignment methods is that the output can be used to produce a visualization of the reads along the genome. Such visualizations are important for quality control and interpretation. To enable the feature with transcriptome pseudoalignment, we developed a tool, *kallisto quant --genomebam*, that generates a BAM file that can be used for visualization, alongside kallisto's usual quantification. This will allow users to benefit from the speed of pseudoalignment while still being able to visualize the pseudoalignments.

Discussion

Our analysis is the first direct comparison that specifically examines the differences in alignment compared to pseudoalignment. Whereas previous comparisons confounded alignment/pseudoalignment with quantification, we have controlled for quantification by developing a new tool to convert genome alignments to TCCs. One application of our tool is to convert preexisting alignments into pseudoalignments. Previous work has shown that using TCCs directly for clustering and differential expression is as good as or better than using transcript quantifications (Ntranos *et al.*, 2016; Yi *et al.*, 2018; Ntranos *et al.*, 2018). Direct analysis of TCCs can be advantageous since it does not introduce inferential ambiguity through transcript assignment and is feasible even in the absence of full length sequencing (Ntranos *et al.*, 2016; Yi *et al.*, 2018; Ntranos *et al.*, 2018). Furthermore, TCCs can be used as a light-weight format to share patient RNA-seq data that maintains individual privacy. Unlike alignments in the form of a BAM file, TCCs are devoid of sequence-specific information and is thus anonymized, while still maintaining all information necessary for transcript quantification.

In choosing the benchmarking metrics for our analyses, we separated analysis of expressed transcripts from non-expressed transcripts. This is not typically done but we found that such a separation is important as metrics such as mean absolute difference (MARD) can be biased by zeroes. Because relative differences are more meaningful on expressed transcripts and absolute differences are more meaningful on non-expressed transcripts, we propose that subsequent benchmarks should always separately evaluate the two.

One advantage of performing genome spliced alignment with RNA-Seq reads is that alignments can be readily visualized on browsers (e.g., Robinson *et al.*, 2011). We provide, for the first time, a tool for visualizing pseudoalignments as projections to the genome. Previously, the pseudoalignment programs RapMap (Srivastava *et al.*, 2016) and kallisto could output SAM formatted alignments, but only with respect to the transcriptome, and were therefore not directly useful for visualization.

Finally, our results demonstrate a practical point for bioinformaticians: for the purpose of transcript quantification, transcriptomic pseudoaligners perform as accurately as aligners. One key advantage of pseudoaligners is speed, and with our new feature, we can support visualization of the pseudoalignments in genomic coordinates. Aside from cases where alignment to noncoding regions is valued (e.g. when transcriptome annotations are incomplete) or where alignments are important for the biology of interest (e.g. for the discovery of novel splice junctions), pseudoalignment should suffice.

Conclusion

In a first direct comparison between aligners and pseudoaligners, we showed that pseudoaligners are as accurate as genome aligners. We created a tool that converts genome alignment in the form of a SAM/BAM into TCCs that can be quantified with kallisto. Furthermore, we implemented a new feature in kallisto for projecting pseudoalignments to the genome, which is output as a BAM file and can be visualized like genome alignments. Our tools place genome alignment and transcriptome pseudoalignment on an equal footing.

Methods

bam2tcc

bam2tcc is written in C++14 and uses the SeqAn software library (Reinert *et al.*, 2017) for efficient parsing of BAM and GTF files. bam2tcc requires as inputs a GTF/GFF file for the annotation and a sorted BAM or SAM file of alignments. The output of bam2tcc is a vector of TCCs and a map of ECs to transcripts.

Briefly, bam2tcc first combines the transcript coordinates and the sorted read alignments. For each alignment, and every transcript, it considers whether the alignment is compatible with the transcript based on the exon coordinates of the transcript. An alignment is compatible with a transcript if it starts within an exon of the transcript, ends within an exon of the transcript, and its gaps coincide within the start and end coordinates of all the exons between the start and end exon. For each alignment, the set of transcripts that are compatible with the alignment is its equivalence class. In the case of reads with multiple genome alignments, bam2tcc computes the union of the alignments' equivalence classes to obtain the equivalence class of the read. For paired-end sequencing, bam2tcc takes the intersection of the equivalence classes corresponding to the two reads to obtain the equivalence class of the pair.

GenomeBam

kallisto v0.44.0 adds a new option of projecting pseudoalignments of reads to genomic coordinates, where alignments are annotated with the posterior probability of the alignment. To this end, using a user-provided GTF file, kallisto constructs a model of the transcriptome consisting of genes, transcripts and exon coordinates. The reporting of the alignment uses a two-stage process. In the first stage, kallisto performs pseudoalignment and the equivalence class of each read is recorded on disk in a temporary file. Following pseudoalignment, the EM algorithm is run to obtain transcript quantifications. This is the usual workflow of kallisto quantification. In the second stage, with quantification results available, kallisto then loads the temporary file of equivalence classes in conjunction with the reads. For each read, kallisto identifies the first k-mer in the read that maps to the transcripts of the equivalence class with non-zero abundances. Using coordinates of the k-mer within the transcript, kallisto then projects the transcript coordinates to genome coordinates, accounting for exon structure. This subsequent projection is done without additional sequence information beyond the first matching k-mer. The set of genome projections are collapsed, such that a read mapping to multiple transcripts but to a single genomic position has a single alignment record in the BAM file. All multiple genome alignments are reported, but the alignment supported by the highest transcript abundance is reported as the primary alignment. The genome is divided into fixed intervals and each alignment is written to a temporary BAM file on disk corresponding to the interval. After all reads have been processed, each temporary BAM file is sorted and concatenated to a final sorted BAM file. Finally, the sorted BAM file is indexed for fast random access.

Datasets

We used RSEM v1.3.0 to simulate paired end RNA-Seq samples with uniform coverage. The RSEM model was built using data from single cell RNA-Seq (SMART-Seq) performed on differentiating myoblasts (Trapnell *et al.*, 2014). With this model, we simulated 10 samples with an average of 2 million paired end reads per sample, and used the isoform counts that RSEM reported to have simulated (RSEM's '*sim.isoform.results*' file) as ground truth. Isoform counts were summed to gene counts to obtain ground truth gene counts.

The Zika-infected human neuroprogenitor cell (hNPC) dataset is available at GEO database (GEO Series GSE78711). For summary statistics, we performed the analyses on all four paired end samples in the dataset and reported the mean and standard deviations across all four samples. For figures showcasing one sample, we used SRR3191542, although we performed the analysis on all four samples and found similar results across them.

Genome and Transcriptome

We performed quantification and analysis using Ensembl *Homo sapiens* genome GRCh38 release 92 (ftp://ftp.ensembl.org/pub/release-92/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna_sm.toplevel.fa) and its corresponding annotation (GRCh38 release 92, ftp://ftp.ensembl.org/pub/release-92/gtf/homo_sapiens). The transcriptome was extracted from the annotation using *tophat -G*. This generation of the transcriptome file puts the genomic and pseudoaligners on an equal footing, as transcripts originating from alternate loci are not included in the transcriptome FASTA file.

Generating TCCs

We used Salmon v0.11.2 (labeled "Salmon" or "Salmon_0.11.2" in figures) and kallisto v0.44.0 (labeled 'kallisto' in figures). We also included Salmon v0.8.2 in several benchmarks, which would be labeled explicitly as "Salmon_0.8.2." Salmon and kallisto indices were built using k-mer length equal to 31. kallisto TCCs were obtained by running *kallisto pseudo*. Salmon TCCs were obtained with Salmon's quasimapping mode by running *Salmon --dumpEQ* and reformatting Salmon's output to match the format of TCCs in kallisto.

We used HISAT2 v2.1.0 and STAR version 2.4.2a to perform genome alignment. We used samtools v.1.2 (Li *et al.*, 2009) to sort the alignments by genomic coordinates. We then ran bam2tcc on the STAR and HISAT2 alignments to generate TCCs.

Quantification

The TCCs generated by all four methods were quantified using kallisto's EM algorithm, which is built on a uniform sequencing model. kallisto's EM algorithm was run by using a branch of kallisto written specifically for this analysis (https://github.com/pachterlab/kallisto/tree/pseudoquant) and invoking *kallisto pseudoquant -l 187 -s 70* on the TCCs generated from all four methods. The *-l* and *-s* parameters correspond to the fragment size distribution (mean length and standard deviation), which are required for quantification with the EM algorithm.

Benchmarking

In comparing the quantifications across the methods, we use the mean absolute relative distance (MARD) and the mean absolute distance. We defined mean absolute relative distance (MARD) as:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{|\hat{x}_t - x_t|}{x_t},$$

where *T* is the number of transcripts/genes considered, $\hat{x}_t$ is the estimated quantification for transcript/gene *t*, and x_t is the ground truth quantification for transcript/gene *t*. We define mean absolute distance as:

$$\frac{1}{T}\sum_{t=1}^{T}|\hat{x}_t - x_t|.$$

Because we use mean absolute distance on only the set of unexpressed transcripts/genes, the mean absolute distance simplifies to

$$\frac{1}{T}\sum_{t=1}^{T}|\hat{x}_t|.$$

We use transcript and gene counts to calculate MARDs and mean absolute differences, obtaining gene counts from summing counts of the corresponding transcripts. We perform the Pearson and Spearman correlations on the log-transformed counts.

Declarations

Availability and Implementation

bam2tcc is available at https://github.com/pachterlab/bam2tcc. kallisto v0.44.0 containing the novel genomebam feature is available at https://pachterlab.github.io/kallisto/. The scripts and code used to regenerate our analysis are available at https://github.com/pachterlab/YLMP_2018.

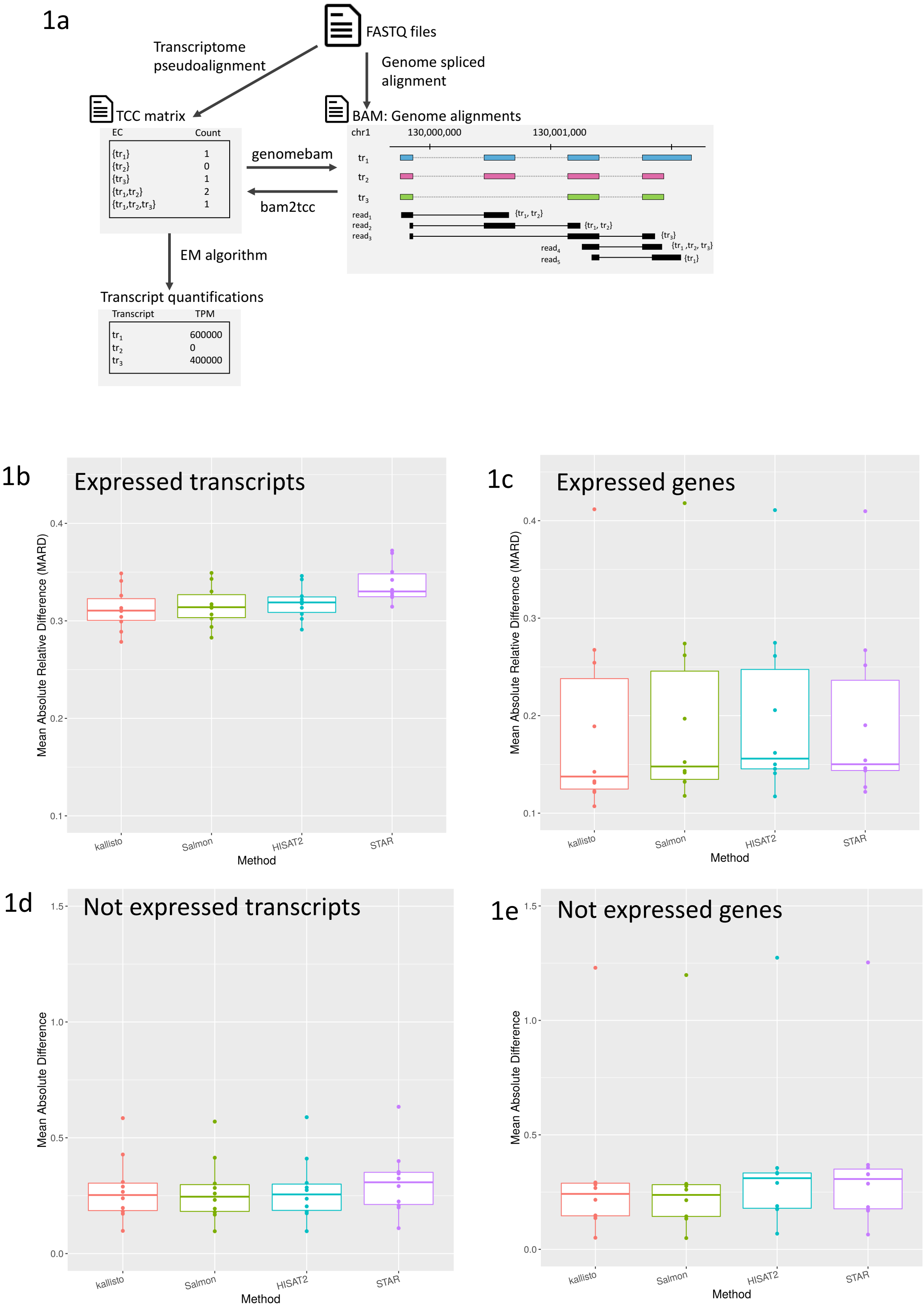
Acknowledgement

We thank Valentine Svensson for helpful feedback during the implementation of bam2tcc.

Funding

LY was funded by the UCLA/Caltech MSTP, NIH T32 GM007616, NIH U19MH114830, the Lee Ramo Endowment, the Treadway Endowment, and the Hearst Endowment. LP was partly funded by U19MH114830.

Figures



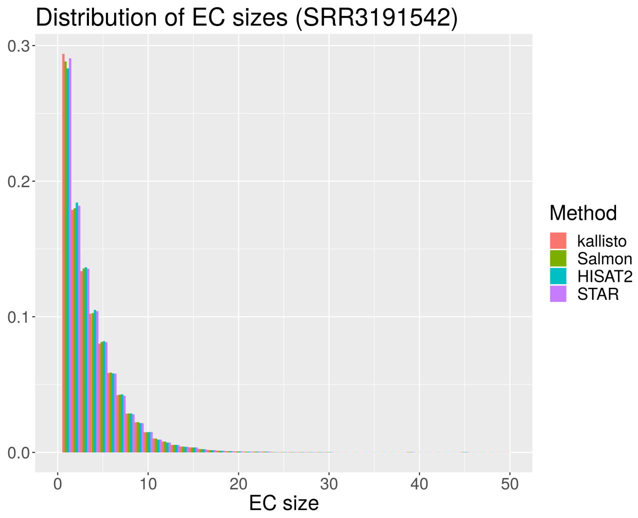
EC	Count
$\{tr_1\}$	1
$\{tr_2\}$	0
$\{tr_3\}$	1
$\{tr_1,tr_2\}$	2
$\{tr_1,tr_2,tr_3\}$	1

Transcript	TPM
tr_1	600000
tr_2	0
tr_3	400000

Figure 1

(a) We compared genome aligners and pseudoaligners by obtaining transcript compatibility counts from all methods and using kallisto to perform the same EM quantification. bam2tcc was used to convert genome alignments from HISAT and STAR to transcript compatibility counts prior to quantification. We then plotted the mean absolute relative distances (MARDs) across ten simulations for transcripts and genes where the true expression is nonzero **(b-c)** and the mean absolute distance for transcripts and genes where the true expression is zero **(d-e)**.

2a



2b

Mean EC Size (SD)

kallisto	3.837 (0.025)
Salmon	3.842 (0.024)
HISAT2	3.828 (0.024)
STAR	3.799 (0.025)

2c

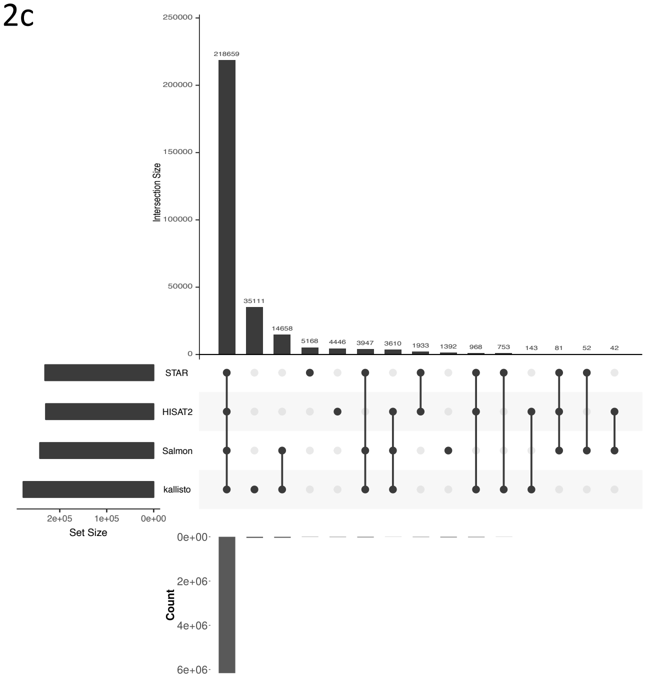


Figure 2

Distribution of equivalence class sizes in a dataset of paired-end RNA-Seq of human neuroprogenitor cells (SRR3191542). The size of an equivalence class is measured as the number of transcripts, weighted by the number of counts in the equivalence class. All methods have similar distributions of equivalence class sizes **(a)**, and furthermore, the methods have comparable mean equivalence class size across the four samples in this dataset **(b)**. The other three samples in the dataset also had similar distributions of equivalence class sizes (data not shown). **(c)** shows the equivalence classes that are shared across methods using an upset plot. The number of shared equivalence classes across the methods are plotted in the top bar graph. The read density in these equivalence classes are plotted in the bottom bar graph, which was calculated as the sum of the counts of the ECs within that intersection.

References

Baruzzo,G. *et al.* (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods*, **14**, 135–139.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–60.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Lachmann,A. *et al.* (2018) Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.*, **9**, 1366.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Liao,Y. *et al.* (2014) Sequence analysis featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. **30**, 923–930.

Nicolae,M. *et al.* (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.*, **6**, 9.

Ntranos,V. *et al.* (2016) Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.*, **17**, 112.

Ntranos,V. *et al.* (2018) Identification of transcriptional signatures for cell types from single-cell RNA-Seq. *bioRxiv*, 258566.

Patro,R. *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.

Reinert,K. *et al.* (2017) The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. *J. Biotechnol.*, **261**, 157–168.

Roberts,A. and Pachter,L. (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*, **10**, 71–3.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–6.

Srivastava,A. *et al.* (2016) RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, **32**, i192–i200.

Tang,H. *et al.* (2016) Zika Virus Infects Human Cortical Neural Progenitors and Attenuates Their Growth. *Cell Stem Cell*, **18**, 587–590.

Teng,M. *et al.* (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biol.*, **17**, 74.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–11.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Vivian,J. *et al.* (2017) Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.*, **35**, 314–316.

Yi,L. *et al.* (2018) Gene-level differential analysis at transcript-level resolution. *Genome Biol.*, **19**, 53.

Yi,L. *et al.* (2017) Zika infection of neural progenitor cells perturbs transcription in neurodevelopmental pathways. *PLoS One*, **12**.