UNIVERSITY OF CALIFORNIA

Los Angeles

Statistical Methods for Bulk and Single-cell RNA Sequencing Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Wei Li

2019

ABSTRACT OF THE DISSERTATION

Statistical Methods for Bulk and Single-cell RNA Sequencing Data

by

Wei Li
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2019
Professor Jingyi Li, Chair

Since the invention of next-generation RNA sequencing (RNA-seq) technologies, they have become a powerful tool to study the presence and quantity of RNA molecules in biological samples and have revolutionized transcriptomic studies on bulk tissues. Recently, the emerging single-cell RNA sequencing (scRNA-seq) technologies enable the investigation of transcriptomic landscapes at a single-cell resolution, providing a chance to characterize stochastic heterogeneity within a cell population. The analysis of bulk and single-cell RNA-seq data at four different levels (samples, genes, transcripts, and exons) involves multiple statistical and computational questions, some of which remain challenging up to date.

The first part of this dissertation focuses on the statistical challenges in the transcript-level analysis of bulk RNA-seq data. The next-generation RNA-seq technologies have been widely used to assess full-length RNA isoform structure and abundance in a high-throughput manner, enabling us to better understand the alternative splicing process and transcriptional regulation mechanism. However, accurate isoform identification and quantification from RNA-seq data are challenging due to the information loss in sequencing experiments. In Chapter 2, given the fast accumulation of multiple RNA-seq datasets from the same biological condition, we develop a statistical method, MSIQ, to achieve more accurate isoform quantification by integrating multiple RNA-seq samples under a Bayesian framework. The MSIQ method aims to (1) identify a consistent group of samples with homogeneous quality and (2) improve isoform quantification accuracy by jointly modeling multiple RNA-seq samples and allowing for higher weights on the consistent group. We show that MSIQ provides a consistent estimator of isoform abundance, and we demonstrate the accuracy of MSIQ compared with alternative methods through both simulation and real data studies. In Chapter 3, we introduce a novel method, AIDE, the first approach that directly controls false isoform discoveries by implementing the statistical model selection principle. Solving the isoform discovery problem in a stepwise manner, AIDE prioritizes the annotated isoforms and precisely identifies novel isoforms whose addition significantly improves the explanation of observed RNA-seq reads. Our results demonstrate that AIDE has the highest precision compared to the state-of-the-art methods, and it is able to identify isoforms with biological functions in pathological conditions.

The second part of this dissertation discusses two statistical methods to improve scRNA-seq data analysis, which is complicated by the excess missing values, the so-called dropouts due to low amounts of mRNA sequenced within individual cells. In Chapter 5, we introduce scImpute, a statistical method to accurately and robustly impute the dropouts in scRNA-seq data. The scImpute method automatically identifies likely dropouts, and only performs imputation on these values by borrowing information across similar cells. Evaluation based on both simulated and real scRNA-seq data suggests that scImpute is an effective tool to recover transcriptome dynamics masked by dropouts, enhance the clustering of cell subpopulations, and improve the accuracy of differential expression analysis. In Chapter 6, we propose a flexible and robust simulator, scDesign, to optimize the choices of sequencing depth and cell number in designing scRNA-seq experiments, so as to balance the exploration of the depth and breadth of transcriptome information. It is the first statistical framework for researchers to quantitatively assess practical scRNA-seq experimental design in the context of differential gene expression analysis. In addition to experimental design, scDesign also assists computational method development by generating high-quality synthetic scRNA-seq datasets under customized experimental settings.

The dissertation of Wei Li is approved.

Alexander Hoffmann

Yingnian Wu

Qing Zhou

Jingyi Li, Committee Chair

University of California, Los Angeles

2019

To my dearest parents,

Hongwei Li and Yuexia Wu,

for their unconditional love, acceptance, and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Throughout my doctoral studies and the writing of this dissertation, I have received a great deal of support and assistance from my mentors, collaborators, family and friends.

I would like to express my deepest gratitude to my advisor, Prof. Jingyi Jessica Li, who has been extremely supportive of my study, research, and professional development. Jessica has led me to the field of statistical genomics, and it has always been inspiring to observe her scientific rigor and enthusiasm for this interdisciplinary field. I am grateful for Jessica' constant encouragement, patient guidance, and the tremendous support she provided throughout my doctoral studies at UCLA.

I would like to thank Prof. Alexander Hoffmann, Prof. Yingnian Wu, and Prof. Qing Zhou for serving on my doctoral committee and providing insightful discussions and feedbacks on my research. I am deeply indebted to Prof. Mark Handcock, Prof. Hoffmann, and Prof. Xin Tong for their invaluable advice on my academic career. In addition, I am very grateful to Prof. Hubing Shi, Prof. Shihua Zhang, and Dr. Anqi Zhao for collaborating with me on transcriptomic studies, which contribute to the first part of this dissertation.

I would also like to extend my appreciation to Dr. Zahra Razaee, Dr. Medha Uppala, Dr. Yucheng Yang, and Dr. Esther Hsiao, for their advice and assistance while I was a junior student. I also appreciate the help and cooperation I received from other previous and current members in Jessica's group, in particular Yidan Sun, Xinzhou Ge, Haowen Zhang, and Ruochen Jiang. I thank Prof. Nicolas Christou and Prof. Robert Gould for their advice on my teaching assistantships. I also thank other faculty members in the Department of Statistics at UCLA for the numerous help I received from them.

My sincere thanks also go to my family and friends for their love and support. I thank my friends at UCLA, Yidan Sun, Yvonne Xiao, Yu Gao, Qiaoling Ye, and Kun Zhou, for their generous help and friendship. I also thank my longtime friends, Yingyao Bai, Junlin Li, Jie Liao, and Zhichao Wu, for always being there for me. Most importantly, I wish to thank my dearest parents, Hongwei Li and Yuexia Wu, and my boyfriend, Jingdong Sun. This dissertation would not have been possible without their warm love and endless support.

VITA

2010–2014	B.S. in Statistics, School of Mathematics and Statistics, Huazhong University of Science & Technology
2015–2016	Teaching Assistant, Department of Statistics, University of California, Los Angeles
2016-2018	Graduate Student Researcher, Department of Statistics, University of California, Los Angeles
2018-2019	Teaching Assistant Consultant, Department of Statistics, University of California, Los Angeles

HONORS AND AWARDS

2011	National Fellowship, Ministry of Education, China
2012	National Fellowship, Ministry of Education, China
2013	Student Award of Excellence, HUST
2013	National Fellowship, Ministry of Education, China
2015	The Most Promising Computational Statistician Award, UCLA
2016	Doctoral Student Travel Grant, UCLA
2016	Diversity Scholarship, UseR 2016 Conference
2018	Pearl Cohen Poster Award, UCLA Biomedical & Science Innovation Day

2018	Dissertation Year Fellowship, UCLA

PUBLICATIONS

Li, W. V., Razaee, Z. S., & Li, J. J. (2016). Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states. *BMC Genomics*, 17(1), S10.

Li, W. V., Chen, Y., & Li, J. J. (2017). TROM: A testing-based method for finding transcriptomic similarity of biological samples. *Statistics in Biosciences*, 9(1), 105-136.

Li, W. V., Zhao, A., Zhang, S., & Li, J. J. (2018). MSIQ: joint modeling of multiple RNA-seq samples for accurate isoform quantification. *Annals of Applied Statistics*, 12(1), 510-539.

Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9, 997.

Li, W. V., & Li, J. J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology*, 6(3), 195-209.

Li, W. V., & Li, J. J. (2019). A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics*, in press.

Ge, X., Zhang, H., Xie, L., **Li, W. V.**, Kwon, S.B., & Li, J. J. (2019). EpiAlign: an alignment-based bioinformatic tool for comparing chromatin state sequences. *Nucleic Acids Research*, in press.

CHAPTER 1

Introduction to Bulk RNA Sequencing Analysis

1.1 Background

Transcriptomes are the complete sets of RNA molecules in biological samples. Unlike the genome, which is largely invariant in different tissues and cells of the same individual, transcriptomes can vary greatly and cause different tissue and cell phenotypes. Understanding transcriptomes is essential for interpreting genome function and investigating molecular bases for various disease phenomena. RNA sequencing (RNA-seq) uses the next generation sequencing (NGS) technologies to reveal the presence and quantity of RNA molecules in biological samples. Since its invention, RNA-seq has revolutionized transcriptome analysis in biological research. RNA-seq does not require any prior knowledge on RNA sequences, and its high-throughput manner allows for genome-wide profiling of transcriptome landscapes [1, 2]. Researchers have been using RNA-seq to catalog all transcript species, such as messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs), to determine the transcriptional structure of genes, and to quantify the dynamic expression patterns of every transcript under different biological conditions [1]. Depending on whether RNA molecules are sequenced for cells in bulk or for individual cells, the RNA-seq technologies are divided into bulk RNA-seq and single-cell RNA-seq protocols. In Chapters 1-3, we discuss how to model the bulk RNA-seq data generated by the Illumina sequencers (Figure 1.1), which are used to analyze the transcriptomes from large populations of cells. In Chapters 4-6, we dicuss the statistical modeling and analysis of single-cell RNA-seq data.

Due to the popularity of bulk RNA-seq technologies and the increasing need to analyze large-scale RNA-seq datasets, more than two thousand computational tools have been devel-

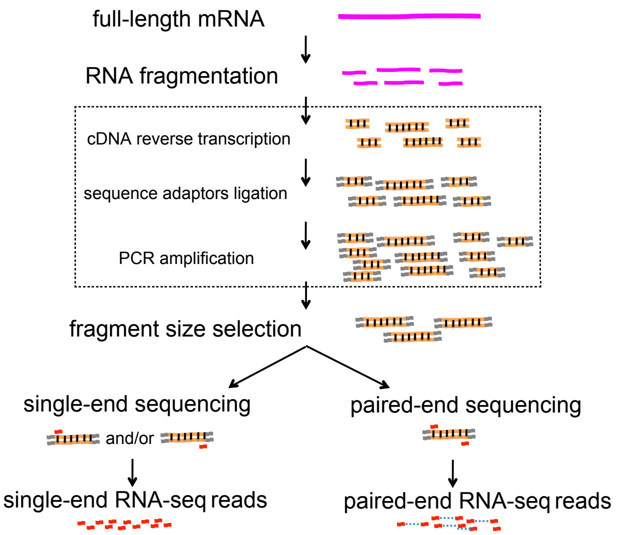


Figure 1.1: Workflow of an RNA-seq experiment. The first step is to break full-length mRNA transcripts into short fragments, because the current state-of-the-art sequencing machines have various length limits on their input nucleotide sequences. To stabilize the resulting short single-stranded RNA fragments, they are reversely transcribed into double-stranded complementary DNAs (cDNAs). Then adapters are added to both ends to ease the sequencing step. Since some cDNA fragments are rare and might not be captured in sequencing, the polymerase chain reaction (PCR) technique is used to amplify the copies of each cDNA fragment to achieve stronger sequencing signals. After this amplification step, a fragment size selection step is used to filter out cDNA fragments that are too short or too long to be accurately sequenced. Finally, in the sequencing step, short sequences of fixed length, starting from the ligated adapters and extending into the actual fragment sequences, will be captured by the sequencing machine from the two ends of double-stranded cDNA fragments.

oped in the past ten years to assist the visualization, processing, analysis, and interpretation of RNA-seq data. The two most computationally intensive steps are data processing and analysis. In data processing, for organisms with reference genomes available, short RNA-seq reads (fragments) are aligned (or mapped) to the reference genome and converted into genomic positions; for organisms without reference genomes, *de novo* transcriptome assembly is needed. Regarding the reference-based alignment, the RNA-seq Genome Annotation Assessment Project (RGASP) Consortium has conducted a systematic evaluation of mainstream spliced alignment programs for RNA-seq data [3]. In this chapter, we focus on the statistical questions engaged in RNA-seq data analyses, assuming reads are already aligned to the reference genome. Depending on the biological questions to be answered from RNA-seq data, we categorize RNA-seq analyses at four levels, which require three different ways of RNA-seq data summary. Sample-level analyses (e.g., sample clustering) and gene-level analyses (e.g., identifying differentially expressed genes [4]) mostly require gene read counts, i.e., how many RNA-seq reads are mapped to each gene. Transcript-level analyses, such as RNA transcript assembly and quantification [5], often need read counts of genomic regions within a gene, i.e., how many RNA-seq reads are mapped to each region and each region-region junction, or even the exact position of each read. Exon-level analyses, such as identifying differential exon usage [6], usually require read counts of exons and exon-exon junctions. As these four levels of analysis use different statistical and computational methods, we introduce the key statistical models used at each level of RNA-seq analysis (Figure 1.2), with an emphasis on the identification of differential expression and alternative splicing patterns, two of the most common goals of bulk RNA-seq experiments.

1.2 Sample-level analysis: transcriptome similarity

The availability of numerous public RNA-seq datasets has created an unprecedented opportunity for researchers to compare multi-species transcriptomes under various biological conditions. Comparing transcriptomes of the same or different species can reveal molecular mechanisms behind important biological processes, and help one understand the conservation and differentiation of these molecular mechanisms in evolution. Researchers need similarity measures to directly evaluate the similarities of different samples (i.e., transcriptomes) based on their genome-wide gene expression data summarized from RNA-seq experiments. Such similarity measures are useful for outlier sample detection, sample classification, and sample clustering analysis. In addition to gene expression, it is also possible to evaluate transcriptome similarity based on alternative splicing events [7]. Correlation analysis is a

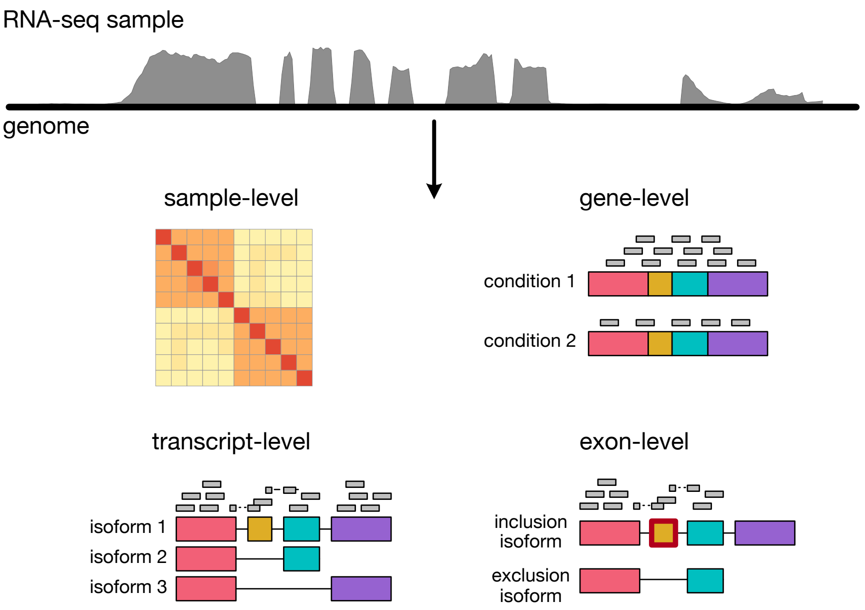


Figure 1.2: RNA-seq analyses at four different levels: sample-level, gene-level, transcript-level, and exon-level. In the sample-level analysis, the RNA-seq reads are usually summarized into a similarity matrix. Taking a 4-exon gene as an example, the gene-level analysis summarizes the counts of RNA-seq reads mapped to the gene in samples of different conditions, and it subsequently compares the gene's expression levels calculated based on read counts; the transcript-level analysis focuses on reads mapped to different isoforms; the exon-level analysis mostly considers the reads mapped to or skipping the exon of interest (the yellow exon marked by a red box in this example).

classical approach to measure transcriptome similarity of biological samples [8,9]. The most commonly used measures are Pearson and Spearman correlation coefficients. The analysis starts by calculating pairwise correlation coefficients of normalized gene expression between any two biological samples, resulting in a correlation matrix. Users can visualize the correlation matrix (usually as a heatmap) to interpret the pairwise transcriptome similarity of biological samples, or they may use the correlation matrix in downstream analysis such as sample clustering.

However, a caveat of using correlation analysis to infer transcriptome similarity is that the existence of housekeeping genes would inflate correlation coefficients. Moreover, correlation measures rely heavily on the accuracy of gene expression measurements and are not robust when the signal-to-noise ratios are relatively low. Therefore, we have developed an alternative transcriptome overlap measure TROM [10] to find sparse correspondence of transcriptomes in the same or different species. The TROM method compares biological samples based on their *associated genes* instead of the whole gene population, thus leading to a more robust and sparse transcriptome similarity result than that of the correlation analysis. TROM defines the associated genes of a sample as the genes that have z-scores (normalized expression levels across samples) greater than or equal to a systematically selected threshold. Pairwise TROM scores are then calculated by an overlap test to measure the similarity of associated genes for every pair of samples. The resulting TROM score matrix has the same dimensions as the correlation matrix, and the TROM score matrix can be easily visualized or incorporated into downstream analysis.

Aside from the correlation coefficients and the TROM scores, there are other statistical measures useful for measuring transcriptome similarity in various scenarios. First, partial correlation can be used to measure sample similarity after eliminating the part of the sample correlation attributable to other variables such as batch effects or experimental conditions [11]. Second, with evidence of a non-linear association between RNA-seq samples, it is suggested to use measures that can capture non-linear dependences, such as the mutual information. Similarly, one may consider using the conditional mutual information [12] or partial mutual information [13] to remove the effects of other confounding variables. In addition to the direct calculation of the sample similarity matrix by applying a similarity measure to the high-dimensional gene expression data, sometimes it is helpful to visualize the gene expression data and investigate the sample similarities after dimensionality reduction. Popular dimension reduction methods include principal component analysis (PCA), t-stochastic neighbor embedding (t-SNE) [14], and multidimensional scaling (MDS) [15].

1.3 Gene-level analysis: gene expression dynamics

RNA-seq technologies enable the measurement and comparison of genome-wide gene expression patterns across different samples. The profiling of gene expression patterns is the key to investigating new biological processes in various tissues and cells of different organisms. A common but important question in a large cohort of biological studies is how to compare gene expression levels across different experimental conditions, time points, tissue and cell types, or even species.

1.3.1 Differential gene expression (DGE) analysis

The main approach to comparing two biological conditions is to find *differentially expressed* (DE) genes. A gene is defined as DE if it is transcribed into different amounts of mRNA molecules per cell under the two conditions [16]. However, since we do not observe the true amounts of mRNA molecules, statistical tests are principled approaches that help biologists understand to what extent a gene is DE.

It is commonly acknowledged that normalization is a crucial step prior to DGE analysis due to the existence of batch effects, which could arise from different sequencing depths or various protocol-specific biases in RNA-seq experiments [17]. The reads per kilobase per million mapped reads (RPKM) [18], the fragments per kilobase per million mapped reads (FPKM) [19], and the transcripts per million mapped reads (TPM) [20] are the three most frequently used units for gene expression measurements from RNA-seq data, and they remove the effects of total sequencing depths and gene lengths. Even though in these units, gene expression data may still contain protocol-specific biases [21], and further normalization is often needed. There are two main categories of normalization methods: distribution-based and gene-based. Distribution-based normalization methods aim to make the distribution of all or most gene expression levels similar across different samples, and such methods include the quantile normalization [22], DESeq [23], and TMM [24]. Gene-based normalization methods aim to make non-DE genes or housekeeping genes have the same expression levels in different samples, and such methods include a method by Bullard et al. [17] and PoissonSeq [25]. For a comprehensive comparison of the assumptions and performance of these normalization methods, we refer readers to [16,17,26].

How to form a proper statistical hypothesis test is the core question in the development of a DGE method. Most existing methods use the Poisson distribution [27] or the Negative Binomial (NB) distribution [23,28,29] to model the read counts of an individual gene in different samples. In our discussion here, we focus on the NB distribution because it is commonly used to account for the observed over-dispersion of RNA-seq read counts. We consider two biological conditions $k = 1, 2$, each with J_k samples. $Y_{k,ij}$ denotes the read count of gene i in the jth sample of condition k. The basic assumption is that

$$Y_{k,ij} \sim \text{NB}(\text{mean} = s_{kj}\theta_{ki},\ \text{dispersion} = \phi_i), \tag{1.1}$$

where s_{kj} is the size factor of the jth sample of condition k, θ_{ki} is the true expression level of gene i under condition k, and ϕ_i is the dispersion of gene i. It is necessary to consider the size factor s_{kj} because it accounts for the fact that different samples usually have different numbers of sequenced reads. The dispersion parameter ϕ_i controls the variability of the expression levels of gene i across biological samples. The estimation of the parameters s_{kj}, θ_{ki}, and ϕ_i is the key step to investigating the differential expression of gene i between the two conditions. Bayesian modeling is often used, and prior distributions and relationships of s_{kj}, θ_{ki}, and ϕ_i are often assumed. Note that assuming s_{kj} being independent of gene i simplifies the problem, but it can be advantageous to calculate gene-specific factors $s_{k,ij}$ to account for technical biases dependent on gene-specific GC contents or gene lengths [30]. The DGE analysis is carried out by testing

$$H_0 : \theta_{1i} = \theta_{2i} \text{ vs. } H_1 : \theta_{1i} \neq \theta_{2i} \tag{1.2}$$

for each gene i.

Starting from model (1.1), most methods include six steps. First, they estimate θ_{ki} and ϕ_i for each gene. The dispersion parameter characterizes the mean-variance relationship, consistent with the observation that genes with similar true expression levels exhibit similar variances [28,30]. When the sample sizes are small, one may consider using shrinkage estimation of ϕ_i's to borrow information across genes or to incorporate prior knowledge, for the purpose of obtaining more robust results [31]. Second, they construct a test statistic based on the estimators to reflect the mean difference between the two conditions. Third, they derive the null distribution of the test statistic under H_0. Fourth, they calculate the observed value of the test statistic for each gene. Fifth, they convert the observed values of test statistics into p-values based on the null distribution. Sixth, they perform multiple testing correction on the p-values to determine a reasonable threshold, and the genes with p-values under that threshold would be called as DE.

For example, edgeR [28] first estimates the dispersion parameters using a conditional maximum likelihood and then develops a test analogous to the Fisher's exact test. DESeq2 [30] adds a layer to the model by estimating $(\theta_{2i} - \theta_{1i})$ using a generalized linear model with a logarithmic link function, $Y_{k,ij}$ as the response variable, and the condition as a binary predictor. This generalized linear model setup can easily incorporate the information on experimental design as additional predictors. In the testing step, DESeq2 transforms the problem into testing if the condition predictor has a significant effect on the logarithmic fold change of gene expression, which is equivalent to testing whether $\theta_{2i} - \theta_{1i} = 0$. EBSeq [32] and ShrinkSeq [33] are also based on model (1.1), but under a Bayesian framework they use hyper-parameters to borrow information across genes, and they calculate the posterior probability of a gene being differentially expressed.

There are other DGE methods that do not assume the NB distribution as in model (1.1) but take a different approach by assuming that $\log(Y_{k,ij})$ follows a Normal distribution, which has more tractable mathematical theory than count distributions have. For example, the voom method [34] estimates the mean-variance relationship of $\log(Y_{k,ij})$ and generates a precision weight for each observation. Then voom inputs $\log(Y_{k,ij})$ and precision weights into the limma empirical Bayes analysis pipeline [35], which was initially designed for microarray data and has multiple modeling advantages: using linear modeling to analyze complex experiments with multiple treatment factors, using quantitative weights to account for variation in the precision of different observations, and using empirical Bayes methods to borrow information across genes. Another method sleuth [36] is applicable to finding both differentially expressed genes and transcripts between two conditions.

Remark 1.1. A common scenario is that a study only includes a small number of RNA-seq replicates [37]. Even though most methods introduced in this section are technically applicable to data with as few as two replicates per condition, there is no guarantee of good performance for these methods with a small number of replicates. In fact, it was observed that many methods did not have a good control on false discovery rates (FDR) under this scenario [37]. We suggest that users carefully check or consult a statistician if the assumptions of a method are reasonable for their study before using the method, as a way to reduce the chance of misusing statistics.

Remark 1.2. Comparisons of DGE methods show that none of the methods is optimal in all circumstances, and methods can produce very different results regarding both the ranking and number of DE genes on the same dataset [4, 26]. In some applications, users are more concerned about the ranking of DE genes than the p-values of genes, especially when setting a reasonable threshold on the p-values is difficult. In other applications where thresholding on p-values is required to control the FDR, users need to address the multiple-testing issue. Common approaches to addressing the multiple-testing issue include the Bonferroni correction [38], the Holm-Bonferroni method [39], and the Benjamini-Hochberg FDR correction [40], with a decreasing level of conservatism. The first two methods aim to control the family-wise error rate (the probability of making one or more false discoveries), while the third method aims to control the expected proportion of false discoveries among the discoveries.

1.3.2 Gene co-expression network analysis

A gene co-expression network (GCN) is an undirected graph, where nodes correspond to different genes, and edges connecting the nodes denote the co-expression relationships between genes. GCNs can help people learn the functional relationships between genes and infer and annotate the functions of unknown genes. To the best of our knowledge, the first GCN analysis on a genome-wide scale across multiple organisms was completed in 2003, enabled by the availability of high-throughput microarray data [41]. One of the most commonly used GCN analysis methods, WGCNA, is popularly applied to gene expression datasets to detect gene clusters and modules and investigate gene connectivity by analyzing correlation networks [42]. Here we introduce the GCN methods based on the framework proposed in [43]. We denote the gene expression matrix as $\boldsymbol{X}_{N\times J}$, where the N rows represent genes, and the J columns represent samples. The N genes are considered as N nodes in the co-expression network. The first step is to construct a symmetric adjacency matrix $\boldsymbol{A}_{N\times N}$, where A_{ij} is a similarity score in the range from 0 to 1 between genes i and j. A_{ij} measures the level of concordance between gene expression vectors $\boldsymbol{X}_{i\cdot}$ and $\boldsymbol{X}_{j\cdot}$. As discussed in Chapter 1.2, the similarity measure can be calculated based on the correlation coefficients or the mutual information measures, depending on the type of gene co-expression relationships to be studied. The elements in the adjacency matrix only consider each pair of genes when evaluating their similarity in expression profiles. However, it is important to consider the relative connectedness of gene pairs with respect to the entire network in order to detect co-expression gene modules. Therefore, one needs to calculate the topological overlap matrix $\boldsymbol{T}_{N\times N}$, where T_{ij} is the topological overlap between node i and j. One such example used in previous studies is $T_{ij} = \frac{\sum_{k=1}^{N} A_{ik}A_{kj}+A_{ij}}{\min\{\sum_{k=1}^{N} A_{ik},\sum_{k=1}^{N} A_{jk}\}+1-A_{ij}}$ [44]. The final distance between nodes i and j is defined as $d_{ij} = 1 - T_{ij}$. Clustering methods can then be applied to search for gene modules based on the resulting distance matrix. The identified gene modules are of great biological interest in many applications. For example, the modules can serve as a prioritizer to evaluate functional relationships between known disease genes and candidate genes [45]. Gene modules can also be used to detect regulatory genes and study the regulatory mechanisms in various organisms [46].

1.4 Transcript-level analysis: isoform discovery and quantification

During the transcription process from genes to mRNA transcripts, one gene may give rise to multiple mRNA transcripts with different nucleotide sequences, thus contributing to the diversity of transcriptomes. RNA transcripts from the same gene are often referred to as *isoforms*, which are different combinations of whole or partial exons. An important use of RNA-seq data is to recover full-length mRNA transcript structures and expression levels based on short RNA-seq reads. This application involves two major tasks. The first task, to identify novel transcripts in RNA-seq samples, is commonly referred to as transcript/isoform reconstruction, discovery, assembly, or identification. This is one of the most challenging problems in this area due to the large searching space of candidate isoforms and limited information contained in short reads (Figure B.1a). The second task, to estimate the expression of known or newly discovered transcripts, is usually referred to as transcript/isoform quantification or abundance estimation. It is a common practice to combine the two tasks into one step, and many popular computational tools simultaneously perform transcript reconstruction and quantification [47]. This is usually achieved by estimating the expression levels of all the candidate isoforms with penalty or regularity constraints, and the resulting isoforms with non-zero estimated expression are treated as identified isoforms. We introduce these two tasks together, as they can be tackled by the same statistical framework in many existing tools. We focus on the basic models that are commonly used by multiple methods, and these models are generally annotation-based and assume that a reference genome is available for the organism of interest.

The transcript reconstruction and quantification are performed separately for individual genes, so the following discussion applies to one gene. Throughout this section, we index the isoforms of a gene as $\{1, 2, \ldots, J\}$. In the reconstruction setting, J is the total number of candidate isoforms; in the quantification setting, J is the number of annotated (or newly discovered) isoforms to be quantified. We index the exons of the gene as $\{1, 2, \ldots, I\}$. Suppose that a total of n reads are mapped to the gene, and they are denoted as $\boldsymbol{R} = \{r_1, r_2, \ldots, r_n\}$. The goal of most methods is to estimate $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_J)^T$, where $\theta_j =$ fraction of isoform $j = \mathbb{P}$(a random read is from isoform j).

1.4.1 Likelihood-based methods

The first type of transcript quantification methods estimates transcript abundance by maximizing the likelihood or the posterior based on a statistical model. These methods are flexible and can be easily modified to incorporate prior biological information. The statistical models are further divided into three categories: region-based, read-based, and fragment-based models.

Region-based models summarize the read counts based on the genomic regions of interest, such as exons and exon-exon junctions. Suppose that S is the index set that denotes all the genomic regions of interest. Read counts can be summarized as $\boldsymbol{X} = \{X_s \mid s \in S\}$, where X_s is the total number of reads mapped to region s. The basic model assumes that X_s follows a Poisson distribution with parameter λ_s. Given the structures of isoforms and their compatibility with the regions, it is reasonable to assume λ_s as a linear function of the θ_j's: $\lambda_s = \sum_{j=1}^{J} a_{sj}\theta_j$. The likelihood function can then be derived, and the task of estimating $\boldsymbol{\Theta}$ reduces to a maximum likelihood estimation (MLE) problem:

$$\begin{aligned} L(\boldsymbol{\Theta}|\boldsymbol{X}) &= \prod_{s\in S} \frac{e^{-\lambda_s}\lambda_s^{X_s}}{X_s!} = \prod_{s\in S} \frac{\exp\left(-\sum_{j=1}^{J} a_{sj}\theta_j\right)\left(\sum_{j=1}^{J} a_{sj}\theta_j\right)^{X_s}}{X_s!}\,, \\ \hat{\boldsymbol{\Theta}} &= \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_J\right)^T = \arg\max_{\boldsymbol{\Theta}} \sum_{s\in S} \log L(\boldsymbol{\Theta}|\boldsymbol{X})\,. \end{aligned} \tag{1.3}$$

The first isoform quantification method [48] uses a region-based model.

In contrast to region-based models, read-based methods directly use the likelihood as a product of the probability densities of individual reads instead of first summarizing reads into region counts. The likelihood function is written as

$$L(\boldsymbol{\Theta}|\boldsymbol{R}) = \prod_{i=1}^{n}\sum_{j=1}^{J} \mathbb{P}(r_i|\text{isoform } j)\ \theta_j = \prod_{i=1}^{n}\sum_{j=1}^{J} \mathbb{P}(s_i|\text{isoform } j)\ \mathbb{P}(\ell_{ij}|\text{isoform } j)\ \theta_j\,, \tag{1.4}$$

where $\boldsymbol{R} = \{r_1, \ldots, r_n\}$, s_i is the starting position, and ℓ_{ij} is the read length (for single-end reads) or fragment length (for paired-end reads) of read r_i if it belongs to isoform j (Figure B.1b). While many methods do not explicitly state it, they assume that s_i and ℓ_{ij} are independent in the above model. If the two ends of read i are mapped to the same exon or two neighboring exons, its corresponding fragment length can be determined and fixed. Otherwise, the corresponding fragment length l_{ij} of read i could vary for each different compatible isoform j (Figure B.1b). Even though each read has the same weight in the likelihood model, the reads that are mapped to two non-neighboring exons play a critical role in the detection of splicing junctions and the reconstruction of full-length transcripts. Cufflinks [49], eXpress [50], RSEM [20], and Kallisto [51] all adapted the above model in their quantification step, and they mainly differ in how they model $\mathbb{P}(s_i|\text{isoform } j)$ and $\mathbb{P}(l_i|\text{isoform } j)$ to incorporate sequencing bias adjustment. To estimate $\boldsymbol{\Theta}$ by maximizing the likelihood in (1.4), the Expectation-Maximization (EM) algorithm [52] is the standard optimization algorithm.

Some other methods, including WemIQ [53], Salmon [54], iReckon [55], and MSIQ [56], introduce hidden variables to denote the isoform origins of reads and use these variables to simplify the form of the likelihood function. Suppose that the isoform origins of reads $\boldsymbol{R} = \{r_1, \ldots, r_n\}$ are denoted as $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^T$, where $Z_i = j$ if read r_i comes from isoform j. Then the joint probability of $\boldsymbol{R}$ and $\boldsymbol{Z}$ can be written as

$$\mathbb{P}(\boldsymbol{R}, \boldsymbol{Z}|\boldsymbol{\Theta}) = \prod_{i=1}^{n} \mathbb{P}(r_i, Z_i|\boldsymbol{\Theta}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[\mathbb{P}(r_i|\text{isoform } j)\ \theta_j\right]^{\mathbb{1}\{Z_i=j\}} . \tag{1.5}$$

Such model formulation is especially useful when one would like to estimate $\boldsymbol{\Theta}$ under the Bayesian framework (Figure B.1c), as what has been done in MISO [57], Salmon [54], and MSIQ [56]. Prior knowledge on $\boldsymbol{\Theta}$ can be incorporated via modeling the prior distribution of $\boldsymbol{\Theta}$, and $\boldsymbol{\Theta}$ would be estimated with the maximum-a-posteriori (MAP) estimator. As shown in Figure B.1c, another advantage of the Bayesian framework is that the model can be easily extended to incorporate multiple RNA-seq samples and borrow isoform abundance information across samples [56].

A more recent isoform quantification method alpine [58] belongs to the third fragment-based category. Alpine is specifically designed to adjust for multiple sources of sequencing biases in isoform quantification. It models the fragment counts using a Poisson generalized linear model, whose predictors are bias features including the length, the relative position, the read sequence bias, the guanine-cytosine (GC) content and the presence of long GC stretches within every fragment. Alpine estimates the read start sequence biases using the variable length Markov model, which was proposed by Roberts et al. [59] and implemented in Cufflinks [49]. After estimating bias parameters, alpine outputs bias-corrected isoform abundance estimates. The Poisson parameter λ_s for a potential fragment s is assumed to be $\lambda_s = \sum_{j=1}^{J} a_{sj}\theta_j$, similar to the assumptions in region-based models. Hence, θ_j's are estimated based on the bias-corrected estimates $\hat{\lambda}_s$'s.

The above approaches, however, would not lead to accurate isoform reconstruction results when directly used to discover new isoforms, because the number of candidate isoforms can be huge when the number of exons is large. A common practice is to add penalty terms before maximizing the objective function, i.e., the likelihood or the posterior. The regularization aims to enforce sparsity on the estimated $\hat{\mathbf{\Theta}}$, whose nonzero entries indicate the discovered isoforms. Two such reconstruction methods are iReckon [55] and NSMAP [60].

1.4.2 Regression-based methods

The second type of statistical methods for isoform quantification is regression-based. These methods formulate the problem as a linear or generalized linear model and treat the region-based read count (or proportion) as the response variable, candidate isoforms as predictor variables, and isoform abundances as coefficients to be estimated. Regression-based methods include rQuant [61], SLIDE [62], IsoLasso [63], and CIDANE [47].

The basic model is a linear model with region-based read proportions as the responses. As for the design matrix, IsoLasso considers a binary matrix to denote the compatibility between the isoforms and genomic regions, while the other three methods consider a conditional probability matrix, for which the read proportions are modeled as:

$$\begin{aligned} \frac{X_s}{n} &= \sum_{j=1}^{J} \mathbb{P}(\text{a random read falls into region } s|\text{isoform } j)\ \mathbb{P}(\text{isoform } j) + \epsilon_s \\ &\triangleq \sum_{j=1}^{J} F_{sj}\ \theta_j + \epsilon_s \qquad (s \in S)\,, \end{aligned} \tag{1.6}$$

where ϵ_s represents independent random noise with mean 0. As in the likelihood-based methods, the probability F_{sj} depends on the structure of region s and the length of isoform

j. Especially when region s spans alternative splicing junctions (e.g., region s skips the middle exon but includes the two end exons), the estimation accuracy of F_{sj} is critical in the modeling. Then the estimation task reduces to a penalized least-squares problem

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta} \geq \mathbf{0}} \sum_{s=1}^{S} \left(\frac{X_s}{n} - \sum_{j=1}^{J} F_{sj}\, \theta_j \right)^2 + \text{penalty}, \tag{1.7}$$

where the penalty term is only needed for isoform discovery and often excluded for isoform quantification. For example, IsoLasso sets the penalty term as $\lambda \sum_{j=1}^{J} \frac{n\theta_j}{L_j}$, where L_j is the length of isoform j, while SLIDE uses $\lambda \sum_{j=1}^{J} \frac{\theta_j}{m_j}$, where m_j is the number of exons in isoform j. In both methods, λ is a tuning parameter to control the level of regularization. IsoLasso selects λ based on the resulting number of isoforms with non-zero estimated expression, while SLIDE uses a stability criterion [64].

Remark 1.3. There are isoform discovery methods that reconstruct RNA transcripts based on deterministic graph methods. Examples include a *de novo* approach Trinity [65], and reference genome-based approaches Scripture [66], Cufflinks [19], and Stringtie [67], which all construct splice graphs based on aligned reads and then use various criteria to parse the graph into transcripts in a deterministic way, without resorting to statistical models.

Remark 1.4. Despite many methods developed for isoform quantification, not all of them discuss the estimation uncertainty of isoform abundance levels. Even though the point estimates of expression levels have led to new scientific discoveries in many biological studies, it is important to consider estimation uncertainty, especially when the differential expression analysis is of interest, or when some candidate isoforms are highly similar in structures. One way to evaluate the uncertainty in Bayesian methods is to construct posterior or credible intervals of the estimated abundance levels [56,68]. In regression-based methods, it is possible to calculate the standard errors of the abundance estimates.

Remark 1.5. There have been multiple efforts to quantify transcripts for better accuracy based on multiple RNA-seq samples (especially biological replicates), thanks to reduced sequencing costs and the rapid accumulation of publicly available RNA-seq samples. Model-based methods include CLIIQ [69], MITIE [70], FlipFlop [71], and MSIQ [56]. These methods generalize the models designed for isoform quantification based on a single sample (Figure

B.1c), and their results show that aggregating the information from multiple samples can achieve better accuracy in isoform abundance estimation. We discuss the statistical modeling of multiple RNA-seq samples for isoform quantification in Chapter 2.

Remark 1.6. Current statistical methods differ in their perspectives to formulate the isoform quantification problem, the trade-off between the complexity and flexibility of models, and the methods to adjust for various sources of sequencing biases and errors. Because of the complexity of transcript-level analysis and the noise and biases in RNA-seq samples, it is impossible to identify a superior method for all real datasets. We suggest that users consider their preferences on the precision and recall rates in discovery problems, and to evaluate the assumptions of different methods for RNA-seq read generation and bias correction, before selecting the tool to apply on their data [5, 72].

1.5 Exon-level analysis: exon inclusion rates

Since transcript-level analysis of complex genes in eukaryotic organisms remains a great challenge [72], there are approaches focusing on exon-level signals, seeking to study alternative splicing based on exons and exon-exon junctions instead of full-length transcripts. When transcriptomic studies focus on the exon-level, a primary step is usually to estimate the *percentage spliced in* (PSI or Ψ, [57]) of an exon of interest. Our discussion below applies to an individual exon. Considering two isoforms, one includes the exon and the other skips the exon, the goal of model-based methods is to estimate

$$\begin{aligned}\Psi &= \text{exon's inclusion rate}\\ &= \frac{\mathbb{P}(\text{a read is from the inclusion isoform})}{\mathbb{P}(\text{a read is from the inclusion isoform}) + \mathbb{P}(\text{a read is from the exclusion isoform})}.\end{aligned} \tag{1.8}$$

A direct estimator of PSI is $\hat{\Psi} = \frac{\frac{C_I}{L_I}}{\frac{C_I}{L_I} + \frac{C_E}{L_E}}$, where C_I denotes the number of reads supporting the inclusion isoform (e.g., reads spanning the upstream splicing junction, the exon of interest, and the downstream splicing junction), and C_E denotes the number of reads supporting the exclusion isoform (e.g., reads spanning parts of the upstream and downstream exons but skipping the exon of interest). L_I and L_E denote the lengths or the adjusted lengths (after accounting for constraints on read and isoform lengths) of the inclusion and exclusion isoforms.

To evaluate the estimation uncertainty, methods including MISO [57], SpliceTrap [73], and rMATS [74] use different statistical models. Both MISO and SpliceTrap construct models similar to model (1.5) under the Bayesian framework, with Ψ as the parameter of interest. The Bayesian confidence interval of Ψ can then be obtained based on its posterior distribution. The rMATS method accounts for the information from multiple replicates through the following hierarchical model

$$\begin{aligned} C_{Ik}|\Psi_k &\sim \text{Binomial}(n = C_{Ik} + C_{Ek}, p = f(\Psi)), \\ \text{logit}(\Psi_k) &\sim \text{Normal}(\mu = \text{logit}(\Psi), \sigma^2), \end{aligned} \tag{1.9}$$

where C_{Ik} (C_{Ek}) is the number of reads supporting inclusion (exclusion) in replicate k ($k = 1, 2, \ldots, K$); Ψ_k is the PSI of the exon of interest in replicate k; Ψ and σ^2 are the mean and variance of PSI in the biological condition of interest; f is a function to normalize Ψ based on the effective length of the exon. Since MISO and rMATS can estimate Ψ and the uncertainty of $\hat{\Psi}$, they can be used to detect differential exon usage between two biological conditions through statistical testing.

Remark 1.7. There is a trade-off in alternative splicing studies concerning whether to use transcript-level or exon-level information. Full-length transcripts provide global information on splicing patterns which directly lead to knowledge on protein isoforms, but accurate quantification of transcripts suffer from the limited information in short RNA-seq reads. On the other hand, exon-level analysis results in more accurate quantification of individual splicing events, but limits the scope of studies to local genomic regions. The accumulation of multiple RNA-seq samples and the increasingly large databases of annotated transcripts [75] might provide a solution to this dilemma: combining information from multiple samples with prior knowledge on transcripts may assist the reconstruction and quantification of full-length isoforms from short RNA-seq reads.

1.6 Discussion

RNA-seq has become the standard experimental method for transcriptome profiling, and its application to numerous biological studies have led to new scientific discoveries in various biomedical fields. We have summarized the key statistical considerations and methods involved in gene-level, transcript-level, and exon-level RNA-seq analyses. Despite the fact that continuous efforts on the development of new tools improve the accuracy of analyses on all levels, challenges posted by relatively short RNA-seq reads remain in studying full-length transcripts, making it difficult to fully understand the dynamics of mRNA isoforms and their protein products. In complex transcriptomes, probabilistic models have limited power in distinguishing different but highly similar transcripts. It has been noted that identification of all constituent exons of a gene is not always successful, and in cases where these exons are correctly reported, it is challenging to assemble them into complete transcripts with high accuracy [72]. Given the current read lengths in NGS, we emphasize the importance of jointly using multiple samples (i.e., technical or biological replicates) to aggregating information on alternative splicing and sequencing noise. Naïve pooling or averaging methods are shown to be inadequate in the multiple-sample analysis [56], and statistical discussion on this topic is still insufficient. On the other hand, new sequencing technologies such as PacBio [76] and Nanopore [77,78] sequencing technologies can produce longer reads with average lengths of $2-3$ kbp [79]. A primary barrier of the current long-read sequencing technologies is their relatively high error rates and sequencing costs [80]. One current approach to take advantage of these technologies is to combine the information in next-generation short reads and third-generation long reads in isoform analysis [79].

To demonstrate the efficiency of statistical methods developed for RNA-seq data, method developers must show the reproducibility and interpretability of these methods. As we have discussed in Remarks 1.2 and 1.6, there is hardly a method that is superior in every application. However, a useful method should at least demonstrate its advantage under specific assumptions or on a particular type of datasets. Meanwhile, no matter how complicated a statistical model is, its general framework and logical reasoning should be interpretable to the users. Also, comparison of different methods over benchmark data can be beneficial for the development of new methods. However, experimentally validated benchmark data for RNA-seq experiments are still limited on the genome-wide scale.

Aside from the analysis tasks introduced in this chapter, bulk RNA-seq is also widely applied to other areas like RNA-editing analysis [81,82], non-coding RNA discovery and characterization [83, 84], expression quantitative trait loci (eQTLs) mapping [85], and prediction of disease progression [86], with interesting statistical questions involved. Transcriptomic data can also be integrated with genomic and epigenomic data to advance our understanding of gene regulation and other biological processes [87]. In recent years, the emerging single-cell RNA sequencing (scRNA-seq) technologies enable the investigation of transcriptomic landscapes at the single-cell resolution, bringing RNA-seq analyses to a new stage [88]. We discuss the modeling of scRNA-seq data in more detail in Chapters 4-6.

1.7 Acknowledgments

This chapter is partially based on my previous work [89]. I would like to thank the co-author Dr. Jingyi Jessica Li of this work. We also thank the insightful feedbacks from Dr. Lior Pachter at California Institute of Technology and Dr. Michael I. Love at University of North Carolina at Chapel Hill.

CHAPTER 2

Transcript Quantification with Multiple Bulk RNA Sequencing Samples

2.1 Introduction

One primary goal of the transcriptomics field is to quantify the dynamic expression levels of mRNA isoforms under different biological conditions. For common species, such as *H. sapiens* (humans), *M. musculus* (mice), and *D. melanogaster* (fruit flies), existing transcript annotations record a large number of mRNA isoforms reported in previous literature. For example, the UCSC Genome Browser [90], GENCODE [75] and RefSeq [91] databases contain known mRNA isoform structures in transcriptomes of humans and several other species. However, the annotations lack gold standard abundance information of these isoforms. In many biological studies, it is important to identify and catalog expression levels of novel or alternative transcripts [92] in order to perform downstream analyses such as identification of differentially expressed transcripts and construction of transcript co-expression networks. Hence, how to accurately estimate isoform abundance is a key question.

Over the past decade, the next-generation RNA-seq technologies have generated numerous datasets with unprecedented nucleotide-level information on transcriptomes, providing new opportunities to study the dynamic expression of known and novel mRNAs in a high-throughput manner [1, 19, 93]. The ideal data would include the sequences of full-length mRNA transcripts; however, the most widely used Illumina sequencers generate millions of short reads, typically shorter than 300 base pairs (bp), from the two ends of mRNA transcript fragments [1]. In this chapter, our discussion focuses on paired-end RNA-seq data generated by Illumina sequencers (Figure 1.1).

Due to the presence of numerous isoforms in existing annotations, inference of their abundance from RNA-seq reads has been an active field of research since 2009 [48,49,53,62]. A necessary step is to first map reads to reference genomes so that researchers know the number of reads generated from each exon. Then, a common approach to summarizing RNA-seq reads is to categorize the reads by the genomic regions to which they map so that the number of reads in different genomic regions can be used to distinguish the abundance of various isoforms. As different isoforms may consist of overlapping but not identical exons, many methods divide exons into *subexons* (Figure 2.1), which are defined as transcribed regions between every two adjacent splicing sites in annotations [53,62,94]. By this definition, every gene is composed of non-overlapping subexons and introns. Since combinations of subexons form a superset of all the annotated isoforms, it is reasonable to categorize RNA-seq reads based on the sets of subexons to which they map. For the ease of terminology, we will refer to subexons as exons for the remainder of this dissertation.

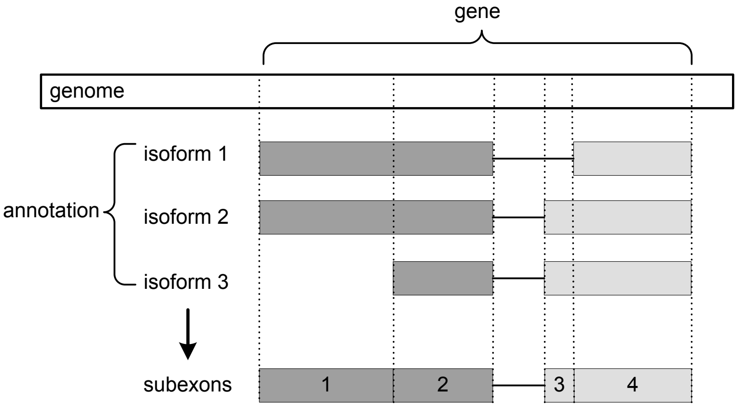


Figure 2.1: Illustration of subexons. The example gene has two exons, represented by light and dark gray boxes, and three mRNA isoforms. The solid lines between exons represent introns in the gene that have been spliced out in isoforms. Adjacent splicing sites in these isoforms define four non-overlapping subexons: the first exon is divided into subexon 1 and 2, and the second exon is divided into subexon 3 and 4.

How to infer isoform abundance from observed RNA-seq reads is a statistical problem, as reads are generated by a mixture of isoforms. We illustrate this using a hypothetical gene, which is composed of four non-overlapping exons (Figure 2.2). Suppose that the gene is transcribed into two mRNA isoforms: 60% of the transcripts are isoform 1, which consists of exons 1, 2 and 4, and 40% of the transcripts are isoform 2, which consists of all four exons. In reality, the isoform proportions, though of great interest to biologists, remain unobservable under the current experimental settings. Our aim is to estimate the relative abundance of annotated isoforms based on reads generated in RNA-seq experiments. Suppose that n paired-end reads are generated and mapped to the reference genome. Some of the mapped reads have obvious isoform origins. For example, read 3 is compatible only with isoform 2, and thus must have isoform 2 as its origin. On the other hand, many mapped reads can have ambiguous origins. For example, read 1 is compatible with both isoforms 1 and 2, and thus we cannot determine its origin isoform. The much more complex structures of real genes complicate the situation even further. For instance, human genes have nine exons on average [95], and a large proportion of human genes have more than ten annotated isoforms (Figure 2.3). Therefore, this problem requires powerful statistical methods to provide good estimates of isoform proportions.

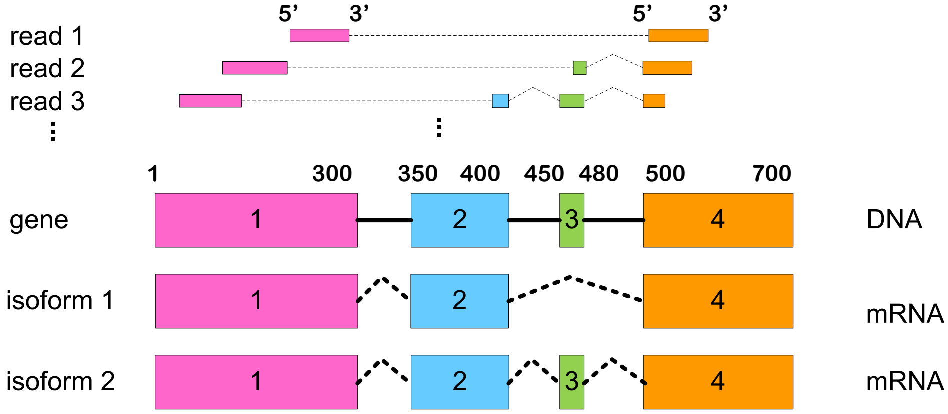


Figure 2.2: The four exons of this gene are represented as boxes of different lengths and colors. The starting and ending positions of the four exons are marked on top of the gene. In an RNA-seq experiment, multiple reads are generated and the number of reads coming from each isoform is proportional to the isoform's abundance. Each read has a 5'-end and a 3'-end, as shown in read 1. These reads are mapped to the reference genome and their overlapping exons are key information for estimating isoform abundance.

A number of isoform quantification methods have been developed to estimate the abun-

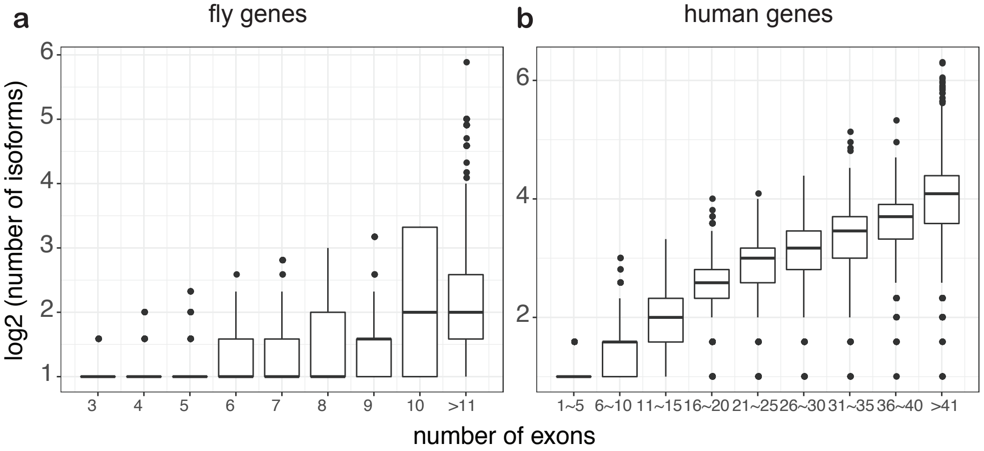


Figure 2.3: Relationship of isoform numbers and exon numbers in fly and human genes. **a**: $\log 2$(isoform numbers) versus exon numbers in 3421 fly genes, whose exon numbers range from 3 to 98. **b**: $\log 2$(isoform numbers) versus exon numbers in 15,268 human genes that have at least two annotated isoforms. The exon numbers in human genes range from 2 to 380.

dance of specific isoforms. These methods perform isoform quantification using either direct computation or model-based approaches [1, 5, 72]. Direct computation approaches count the number of reads compatible with each isoform and then normalize the counts by isoform lengths and the total number of reads to generate estimates of isoform abundance. The most commonly used unit is RPKM [18]. However, for complex gene structures, counts of RNA-seq reads compatible with isoforms may not be proportional to isoform abundance, as some reads cannot be assigned unequivocally to only one isoform. To address this issue, model-based approaches are needed to assess the likelihood of a read coming from different isoforms. In the first model-based isoform quantification method [48], read counts in genomic regions are modeled as Poisson variables with isoform abundance as the mean parameter, under the assumption that reads are uniformly sampled within each isoform. Isoform abundance is estimated by maximum likelihood estimates. Cufflinks [49], the most widely used method for discovering novel isoforms from RNA-seq data, also has the functionality to estimate isoform abundance. Its approach is similar to the likelihood-based approach in [48], and it proposed a new unit for isoform abundance based on paired-end RNA-seq data, FPKM, which accounts for the dependency between paired-end reads. MISO [57] is another model-based method constructed under a Bayesian framework, and it provides maximum-*a-posteriori* estimates and confidence intervals of isoform abundance. There are other isoform quantification methods with different features [96]. For example, iReckon [55] utilizes a regularized EM algorithm; WemIQ [53] replaces the Poisson distribution with a more general and realistic generalized Poisson distribution; Sailfish [97] is a fast alignment-free method that saves the read mapping step. Please see Chapter 1.4 for more detailed discussion of existing transcript quantification methods.

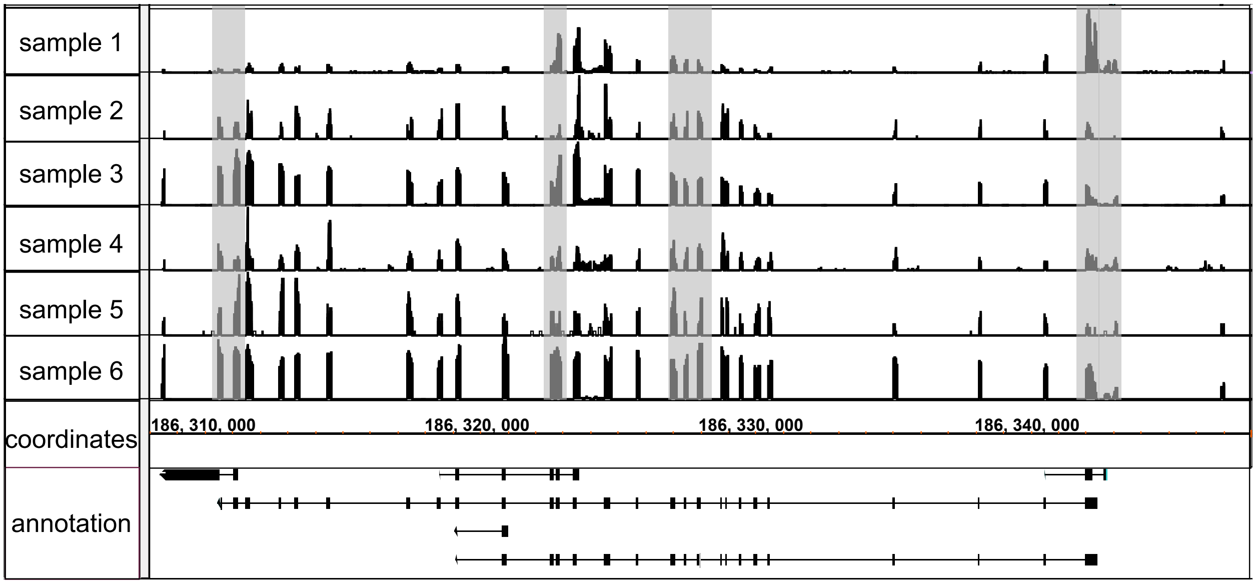


Figure 2.4: Reads in six human ESC RNA-seq samples mapped to the gene *TPR*. Detailed information on these samples is listed in Table C.1. The counts of RNA-seq reads are summarized in the histograms. The annotated isoform structures of this gene were shown in the bottom row. We mark four example sites where the six samples are obviously inconsistent with gray shades.

However, there remains much space to improve the accuracy of isoform quantification due to the noises and biases in RNA-seq data. Because of the accumulation of RNA-seq samples in public databases, multiple RNA-seq data sets are often available for the same biological condition (e.g., the same cell or tissue type), and they provide more information than a single RNA-seq dataset. For example, the GTEx (Genotype-Tissue Expression) study comprises

$9,662$ samples from 54 tissues, and the Cancer Genome Atlas (TCGA) study comprises $11,350$ samples from 33 cancer types [98]. Here, the concept of *multiple samples* includes both *technical replicates* (different aliquots of the same sample measured multiple times [92]) and *biological replicates* (replicates obtained from multiple samples of the same type of cells or tissues). The availability of multiple RNA-seq samples from the same biological condition in public databases (e.g., NIH Gene Expression Omnibus [99]) motivated us to develop a new statistical method for better isoform quantification by taking advantage of the common and thus more reliable information provided by multiple samples. The necessity of such a method is twofold. First, the number of RNA-seq samples produced by a single lab is limited since experimental costs increase each time an additional replicate is added. A statistical method that allows for multiple samples enables researchers to combine their own data with public data to obtain more accurate and robust isoform abundance estimates. Second, such a method supports better reuse of public data for both biological discovery and method development.

Several methods have been developed to use multiple RNA-seq samples from the same biological condition for isoform quantification. For example, CLIIQ [69] uses integer linear programming to jointly model RNA-seq data from multiple samples. MITIE [70] assumes that the same isoforms are expressed in all samples but may have different abundances, and it then reduces the problem to solving systems of linear equations. FlipFlop [71] uses a convex formulation and introduces the group-lasso penalty to ensure sparsity in estimation. However, none of these methods considers the quality variation of different RNA-seq samples and how such variation might affect the inference of isoform abundance. It is commonly recognized that RNA-seq samples generated by different protocols or different labs can vary greatly with respect to the signal-to-noise ratios, biases, etc. For example, Figure 2.4 shows RNA-seq read coverage of the *TPR* gene in six human ESC samples. There is obvious variation in the read coverage profiles of these six samples. For example, the second sample has little signal in the last exon while the other samples have obviously stronger signals in the last exon. Thus, it is inappropriate to treat all the samples equally during isoform quantification by assuming that they come from the same population (i.e., the same tissue or cell type). Hence, results from these methods may be sensitive to the heterogeneity of samples or even, in some cases, be dominated by biased samples, which do not accurately reflect the transcriptome information of the biological condition of interest.

In this chapter, we propose a quantification method, MSIQ (**M**ultiple RNA-seq **S**amples for accurate **I**soform **Q**uantification), for accurately estimating isoform expression. MSIQ is a model-based method for estimating isoform abundance by discerning and using multiple RNA-seq samples that share similar transcriptome information, which we define as the *consistent group*. Our modeling consists of two components: (1) estimating the probability of each sample being in the consistent group via evaluating the sample similarities, and (2) estimating isoform abundance from reweighted samples, with greater weights given to the samples that are more likely to be consistent. These two components enable the method to distinguish between the large variation stemming from experimental factors and the reasonable biological variation. In Chapter 2.2, we describe the Bayesian hierarchical model used in MSIQ to bridge unknown isoform proportions and observed read counts mapped to a gene in multiple RNA-seq samples. Our model allows for different isoform proportions of RNA-seq samples inside and outside the consistent group. This approach reduces the probability that the estimated isoform abundance is biased by samples of poor quality. We conduct parameter estimation by Gibbs sampling and prove the consistency of the MSIQ estimator. We show that the isoform proportions estimated by MSIQ are consistent with the unknown isoform proportions in the consistent group, while the estimates based on the assumption that all samples have equal weights are not. In Chapter 2.3, we apply MSIQ to both simulated and real data to illustrate the efficiency and robustness of MSIQ under various parameter settings. We also compare MSIQ with the oracle estimators and other widely used estimation methods. In Chapter 2.6, we discuss the advantages and limitations of MSIQ and its possible extensions.

2.2 Methods

For a given gene, the MSIQ method aims to achieve two goals with respect to isoform expression quantification. First, we want to identify the samples that represent the tissue or cell type of interest. We refer to these samples as the *consistent group* and assume that the group contains at least one sample. We identify samples in the consistent group under the assumption that these samples share the most similar read distributions among all the samples. Second, we would like to estimate the proportion of reads coming from each RNA isoform in the given tissue or cell type, with larger weights given to samples in the consistent group. We focus our efforts on RNA-seq data with paired-end reads, but the model can easily be extended for single-end reads.

2.2.1 Observed data and parameters of interest

Suppose we are studying a gene with N exons, J annotated RNA isoforms, and D RNA-seq samples. Ideally, we are interested in the true proportion of each isoform

$$p_j = \mathbb{P}(\text{an mRNA transcript is of isoform } j),\ \ j = 1, 2, ..., J. \tag{2.1}$$

However, these hidden parameters are not directly observable in RNA-seq experiments. Instead of directly estimating p_j, we aim to estimate the practical parameters

$$\alpha_j = \mathbb{P}(\text{an RNA-seq read is from isoform } j),\ \ j = 1, 2, ..., J, \tag{2.2}$$

which we refer to as isoform proportions.

We denote the observed data, D independent samples of reads mapped to the given gene, by $\boldsymbol{R}^{(d)} = \{r_1^{(d)}, r_2^{(d)}, ..., r_{n_d}^{(d)}\} \quad (d = 1\,, 2\,, ..., D)$, where n_d denotes the total number of reads in sample d and $r_i^{(d)}$ denotes the ith read ($i = 1, 2, ..., n_d$) in sample d. To use the read information, an efficient data summary is needed to preserve the most relevant information for isoform quantification while limiting the computational complexity to a manageable level [100]. We write each read as $r_i^{(d)} = \left\{\boldsymbol{s}_{1i}^{(d)}, \boldsymbol{s}_{2i}^{(d)}, \left\{y_{i1}^{(d)}, y_{ic^{(d)}}^{(d)}, y_{i(c^{(d)}+1)}^{(d)}, y_{i(2c^{(d)})}^{(d)}\right\}\right\}$, where $\boldsymbol{s}_{1i}^{(d)}$ and $\boldsymbol{s}_{2i}^{(d)}$ respectively denote the index sets of exons overlapping with the read's left end and right end; $y_{ik}^{(d)}$ denotes the kth genomic position of read i; $c^{(d)}$ is read length in sample d. Please refer to Appendix A.1 for the advantages of this summarizing approach over alternative approaches.

2.2.2 Assumptions and prior

In addition to the observed data, we consider the hidden data, which are the isoform origins of the reads $\boldsymbol{Z}^{(d)} = (Z_1^{(d)}, Z_2^{(d)}, \ldots, Z_{n_d}^{(d)})'$, where $Z_i^{(d)} \in \{1, 2, \ldots, J\}$ indicates the isoform origin of read i, and $Z_i^{(d)} = j$ if read i actually comes from isoform j. The differences between RNA-seq samples are reflected in their isoform proportions $\boldsymbol{\tau}^{(d)}$, $d = 1, 2, \ldots, D$. In sample d, we denote the true probability of reads from isoform j as $\tau_j^{(d)} = P(Z_i^{(d)} = j)$ and the isoform proportion vector as $\boldsymbol{\tau}^{(d)} = (\tau_1^{(d)}, \tau_2^{(d)}, \ldots, \tau_J^{(d)})'$, with $\sum_{j=1}^{J} \tau_j^{(d)} = 1$. We define a hidden state variable E_d for each sample such that

$$E_d = \mathbb{1}\{\text{sample } d \text{ belongs to the consistent group}\}. \tag{2.3}$$

We assume samples in the consistent group all have the same proportion vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_J)'$ with $\sum_{j=1}^{J} \alpha_j = 1$, while samples outside the consistent group can each have different isoform proportions $\boldsymbol{\beta}^{(d)} = (\beta_1^{(d)}, \beta_2^{(d)}, \ldots, \beta_J^{(d)})'$ with $\sum_{j=1}^{J} \beta_j^{(d)} = 1$. Thus the isoform proportions can be expressed as

$$\boldsymbol{\tau}^{(d)} = E_d \cdot \boldsymbol{\alpha} + (1 - E_d) \cdot \boldsymbol{\beta}^{(d)} = \begin{cases} \boldsymbol{\alpha}, & \text{if } E_d = 1\,, \\ \boldsymbol{\beta}^{(d)}, & \text{if } E_d = 0\,. \end{cases} \tag{2.4}$$

The isoform proportion vector $\boldsymbol{\alpha}$ of the consistent group is our key parameter of interest.

We assume $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}^{(d)}$ are *a priori* Dirichlet($\boldsymbol{\lambda}$), and E_d is *a priori* Bernoulli(γ): $E_d|\gamma \sim$ Bernoulli(γ), where $\gamma \sim$ Beta(a, b). Intuitively, $\boldsymbol{\lambda}$ controls the distance between the isoform proportions of samples inside and outside the consistent group, while γ controls the tendency of assigning a sample to the consistent group. We describe the relationship between observed RNA-seq reads and hidden isoform proportions in multiple samples under a Bayesian framework (Figure 2.5).

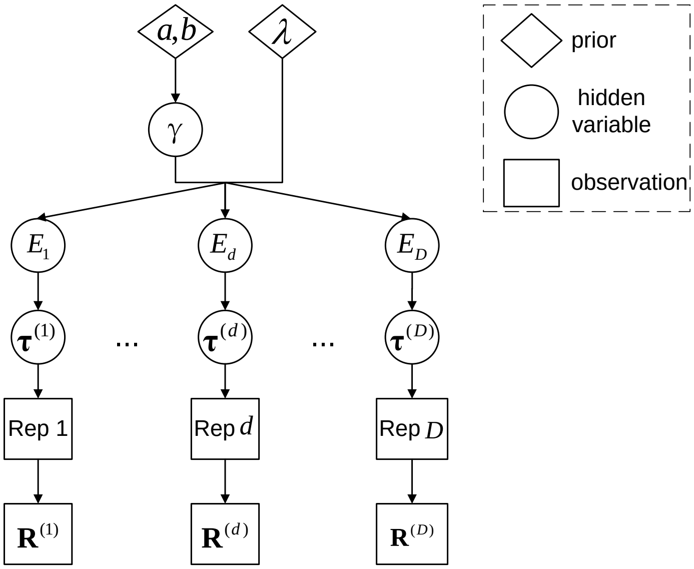


Figure 2.5: Joint modeling of multiple RNA-seq samples with MSIQ. In this framework, E_d ($d = 1, 2, \ldots, D$) is a binary hidden state variable indicating whether RNA-seq sample d is in the consistent group, while a, b, and γ are hyper-parameters (priors) in E_d's distribution. Depending on E_d, the isoform proportion vector $\boldsymbol{\tau}^{(d)}$ takes either the consistent group's isoform proportion vector $\boldsymbol{\alpha}$ or sample-specific $\boldsymbol{\beta}^{(d)}$. Given the isoform proportions, RNA-seq reads are generated in each sample, and the observed data are summarized as $\boldsymbol{R}^{(d)}$.

2.2.3 The probabilistic model of MSIQ

We introduce $I_{i,j}^{(d)}$ as a short notation of the binary variable $\mathbb{1}\{Z_i^{(d)} = j\}$. Given a sample with isoform proportion $\boldsymbol{\tau}^{(d)}$, the probability of read $r_i^{(d)}$ and origin $Z_i^{(d)}$ can be written as:

$$\begin{aligned}\mathbb{P}\left(r_i^{(d)}, Z_i^{(d)} | \boldsymbol{\tau}^{(d)}\right) &= \prod_{j=1}^{J} \mathbb{P}\left(r_i^{(d)}, Z_i^{(d)} = j | \boldsymbol{\tau}^{(d)}\right)^{\mathbb{1}\{Z_i^{(d)}=j\}} \\ &= \prod_{j=1}^{J} \left[\mathbb{P}\left(r_i^{(d)} | Z_i^{(d)} = j\right) \tau_j^{(d)}\right]^{I_{ij}^{(d)}} \triangleq \prod_{j=1}^{J} \left(h_{i,j}^{(d)} \tau_j^{(d)}\right)^{I_{i,j}^{(d)}},\end{aligned} \tag{2.5}$$

where $\mathbb{P}\left(r_i^{(d)}, Z_i^{(d)} | \boldsymbol{\tau}^{(d)}\right)$ refers to the joint density of read $r_i^{(d)}$ and its isoform origin $Z_i^{(d)}$ given the model parameters, and $h_{i,j}^{(d)}$ is the generating probability of read $r_i^{(d)}$ given isoform j. If read $r_i^{(d)}$ and isoform j are incompatible (e.g., read 2 in Figure 2.2 cannot come from isoform 1), $h_{i,j}^{(d)} = 0$. Otherwise, $h_{i,j}^{(d)}$ depends on the model of the read generation mechanism.

We adopt the following model from [53]:

$$h_{i,j}^{(d)} = \frac{1}{\ell_j'} \times \mathbb{P}\left(L_{i,j}^{(d)}\right), \tag{2.6}$$

where ℓ_j' is the effective length (i.e., the number of possible starting positions on the fragment) of isoform j and can be calculated as $\ell_j' = \ell_j - L^{(d)}$: ℓ_j is the length of isoform j and $L^{(d)}$ is the mean fragment length in sample d. $L_{i,j}^{(d)}$ denotes the fragment length of $r_i^{(d)}$ if it comes from isoform j. $L_{i,j}^{(d)}$ is assumed to be a Gaussian random variable and its mean $L^{(d)} = \mathbb{E}(L_{i,j}^{(d)})$ and variance $\text{var}(L_{i,j}^{(d)})$ can be estimated from single-isoform genes, whose mapped reads directly specify fragment lengths.

Let $\boldsymbol{E} = (E_1, E_2, \ldots, E_D)'$ be the hidden state vector indicating whether each sample is among the consistent group or not, and let $\boldsymbol{R} = \{\boldsymbol{R}^{(d)}\}_{d=1}^D$, $\boldsymbol{Z} = \{\boldsymbol{Z}^{(d)}\}_{d=1}^D$, and $\boldsymbol{\tau} = \{\boldsymbol{\tau}^{(d)}\}_{d=1}^D$ represent the reads, origins of reads, and isoform proportions in all the samples, respectively. To simplify the notation, we also introduce $n_j^{(d)} = \sum_{i=1}^{n_d} I_{ij}^{(d)}$ to represent the total number of reads coming from isoform j in sample d. Based on equation (2.5), the joint probability of all reads is given by the MSIQ model as follows:

$$\mathbb{P}\left(\boldsymbol{R}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{E}, \gamma | \boldsymbol{\lambda}, a, b\right) = \mathbb{P}\left(\boldsymbol{R}, \boldsymbol{Z} | \boldsymbol{\tau}, \boldsymbol{E}\right) \mathbb{P}\left(\boldsymbol{\tau} | \boldsymbol{\lambda}, \boldsymbol{E}\right) \mathbb{P}\left(\boldsymbol{E} | \gamma\right) \mathbb{P}(\gamma | a, b), \tag{2.7}$$

where

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{R}, \boldsymbol{Z} | \boldsymbol{\tau}, \boldsymbol{E}\right) &= \prod_{d=1}^{D} \left\{ \left[\prod_{i=1}^{n_d} \prod_{j=1}^{J} \left(h_{i,j}^{(d)} \alpha_j \right)^{I_{i,j}^{(d)}} \right]^{E_d} \left[\prod_{i=1}^{n_d} \prod_{j=1}^{J} \left(h_{i,j}^{(d)} \beta_j^{(d)} \right)^{I_{i,j}^{(d)}} \right]^{1-E_d} \right\}, \\
\mathbb{P}\left(\boldsymbol{\tau} | \boldsymbol{\lambda}, \boldsymbol{E}\right) &\propto \prod_{j=1}^{J} \alpha_j^{\lambda_j - 1} \prod_{d=1}^{D} \left[\prod_{j=1}^{J} \left(\beta_j^{(d)} \right)^{\lambda_j - 1} \right]^{1-E_d}, \\
\mathbb{P}\left(\boldsymbol{E} | \gamma\right) &\propto \gamma^{\sum_{d=1}^D E_d} (1-\gamma)^{D - \sum_{d=1}^D E_d}, \\
\mathbb{P}(\gamma | a, b) &\propto \gamma^{a-1} (1-\gamma)^{b-1}.
\end{aligned} \tag{2.8}$$

As a result, the joint probability can be simplified as

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{R}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{E}, \gamma | \boldsymbol{\lambda}, a, b\right) \propto & \left[\prod_{j=1}^{J} \alpha_j^{\lambda_j - 1 + \sum_{d=1}^D E_d n_j^{(d)}} \right] \left[\prod_{d=1}^{D} \prod_{j=1}^{J} \left(\beta_j^{(d)} \right)^{(1-E_d)\left(\lambda_j - 1 + n_j^{(d)}\right)} \right] \\
& \times \left[\prod_{d=1}^{D} \prod_{j=1}^{J} \prod_{i=1}^{n_d} \left(h_{i,j}^{(d)} \right)^{I_{i,j}^{(d)}} \right] \gamma^{\sum_{d=1}^D E_d + a - 1} (1-\gamma)^{D - \sum_{d=1}^D E_d + b - 1}.
\end{aligned} \tag{2.9}$$

2.2.4 Posterior sampling

In the MSIQ model (2.9), the reads $\boldsymbol{R}$ are the observed data, the isoform origins $\boldsymbol{Z}$ and the consistent group indicator $\boldsymbol{E}$ are the hidden data, while isoform proportions $\boldsymbol{\alpha}, \{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$, and consistent group proportion γ are the parameters. To estimate the parameters, a useful approach is to implement a Gibbs sampler to iteratively draw posterior samples of hidden data and parameters from their conditional distributions. Since our ultimate parameter of interest is $\boldsymbol{\alpha}$, whose inference becomes obvious given $\boldsymbol{Z}$ and $\boldsymbol{E}$, we integrate out $\boldsymbol{\tau}$ (i.e., $\boldsymbol{\alpha}$ and $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$) in model (2.9) to achieve better computational efficiency. This step is based on a property of the Dirichlet distribution:

$$\int \cdots \int_{\{(\tau_1,\ldots,\tau_J):0\leq\tau_j\leq 1,\sum_j \tau_j=1\}} \prod_{j=1}^{J} \tau_j^{\lambda_j-1} d\tau_1 \cdots d\tau_J = B(\boldsymbol{\lambda}), \quad \forall \lambda_j > 0\,, \tag{2.10}$$

where $B(\boldsymbol{\lambda}) = \frac{\Pi_{j=1}^{J}\Gamma(\lambda_j)}{\Gamma(\sum_{j=1}^{J}\lambda_j)}$. Hence,

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{R},\boldsymbol{Z},\boldsymbol{E},\gamma\,|\,\boldsymbol{\lambda},a,b\right) \propto B_1(\boldsymbol{Z},\boldsymbol{E}) &\left[\prod_{d=1}^{D} B_0^{(d)}(\boldsymbol{Z}^{(d)},E_d)\right] \left[\prod_{d=1}^{D}\prod_{i=1}^{n_d}\prod_{j=1}^{J}\left(h_{i,j}^{(d)}\right)^{I_{i,j}^{(d)}}\right] \\ &\times \gamma^{\sum_{d=1}^{D}E_d+a-1}(1-\gamma)^{D-\sum_{d=1}^{D}E_d+b-1}\,,\end{aligned} \tag{2.11}$$

where

$$\begin{aligned}&B_0^{(d)}(\boldsymbol{Z}^{(d)},E_d=1) = 1,\\ &B_0^{(d)}(\boldsymbol{Z}^{(d)},E_d=0) = \Pi_{j=1}^{J}\Gamma\left(\lambda_j+n_j^{(d)}\right)/\Gamma\left(\textstyle\sum_{j=1}^{J}\lambda_j+n_d\right),\\ &B_1(\boldsymbol{Z},\boldsymbol{E}) = \Pi_{j=1}^{J}\Gamma\left(\lambda_j+\textstyle\sum_{d=1}^{D}E_d\cdot n_j^{(d)}\right)/\Gamma\left(\textstyle\sum_{j=1}^{J}\lambda_j+\sum_{d=1}^{D}E_d\cdot n_d\right).\end{aligned} \tag{2.12}$$

We denote $\boldsymbol{\Theta} \triangleq \{\boldsymbol{Z},\boldsymbol{E},\gamma\}$. The distribution of each parameter or hidden variable conditional on everything else can thus be estimated by Gibbs sampling as follows.

(1) E_d follows a Bernoulli distribution:

$$E_d|\Theta/\{E_d\} \quad \sim \quad \text{Bern}\left(\frac{\text{odds}(E_d;\boldsymbol{\lambda},\tau)}{1+\text{odds}(E_d;\boldsymbol{\lambda},\tau)}\right)\,, \tag{2.13}$$

where

$$\begin{aligned}\text{odds}(E_d;\boldsymbol{\lambda},\tau) &= \frac{\mathbb{P}(E_d=1|\boldsymbol{\Theta}/\{E_d\})}{\mathbb{P}(E_d=0|\boldsymbol{\Theta}/\{E_d\})}\\ &= \frac{B_1(\boldsymbol{Z},\boldsymbol{E}_{-d},E_d=1)}{B_1(\boldsymbol{Z},\boldsymbol{E}_{-d},E_d=0)}\cdot\frac{B_0^{(d)}(\boldsymbol{Z}^{(d)},E_d=1)}{B_0^{(d)}(\boldsymbol{Z}^{(d)},E_d=0)}\cdot\frac{\gamma}{1-\gamma}\,.\end{aligned} \tag{2.14}$$

(2) $Z_i^{(d)}$ follows a multinomial distribution:

$$Z_i^{(d)}|\boldsymbol{\Theta}/\{Z_i^{(d)}\} \quad \sim \quad \text{Multinomial}\left(q_{i1}^{(d)}, q_{i2}^{(d)}, \ldots, q_{iJ}^{(d)}\right) , \tag{2.15}$$

where $q_{ij}^{(d)} = {\scriptstyle \mathbb{P}\left(\boldsymbol{R},\boldsymbol{Z}_{-i}^{(-d)},Z_i^{(d)}=j,\boldsymbol{E},\gamma\middle|\boldsymbol{\lambda},a,b\right)/\sum_{j'=1}^{J}\mathbb{P}\left(\boldsymbol{R},\boldsymbol{Z}_{-i}^{(-d)},Z_i^{(d)}=j',\boldsymbol{E},\gamma\middle|\boldsymbol{\lambda},a,b\right)}$.

(3) γ follows a Beta distribution:

$$\gamma|\Theta/\{\gamma\} \sim \text{Beta}\left(\sum_{d=1}^{D} E_d + a, D - \sum_{d=1}^{D} E_d + b\right) . \tag{2.16}$$

2.2.5 Estimators of isoform proportions

With the above posterior distribution of the hidden variables and parameters, we can draw samples iteratively to estimate the hidden state of each RNA-seq sample and the true isoform proportions in the consistent group. We suppose that there are T iterations after discarding the burn-in period of Gibbs sampling. In each iteration, we denote the sampled hidden state vector as $\boldsymbol{E}^{(t)} = (E_1^{(t)}, E_2^{(t)}, \ldots, E_D^{(t)})'$ and the hidden origin vector in sample d as $(Z_1^{(d,t)}, \ldots, Z_{n_d}^{(d,t)})'$.

To estimate isoform proportions in each iteration, we pool the reads from sample d whose estimated state varibale $E_d^{(t)} = 1$ to calculate $\boldsymbol{\alpha}^{(t)}$, where

$$\alpha_j^{(t)} = \frac{\lambda_j + \sum_{d=1}^{D}\left(E_d^{(t)}\sum_{i=1}^{n_d}\mathbb{1}\{Z_i^{(d,t)} = j\}\right)}{\sum_{j=1}^{J}\lambda_j + \sum_{d=1}^{D}E_d^{(t)}n_d}. \tag{2.17}$$

Overall, the MSIQ estimator of the isoform proportion vector is $\hat{\boldsymbol{\alpha}}^{\text{MSIQ}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}^{(t)}$, and the relative estimation error (REE) is calculated as

$$\text{REE}(\hat{\boldsymbol{\alpha}}^{\text{MSIQ}}) = \sum_{j=1}^{m}|\alpha_j - \hat{\alpha}_j^{\text{MSIQ}}|/\alpha_j. \tag{2.18}$$

We show the consistency property of the MSIQ estimator $\hat{\boldsymbol{\alpha}}^{\text{MSIQ}}$ in the following lemma (Appendix A.2).

Lemma 2.1. $\hat{\boldsymbol{\alpha}}^{\text{MSIQ}}$ converges to the posterior mean of isoform proportion $\mathbb{E}(\boldsymbol{\alpha}|\boldsymbol{R}, \boldsymbol{\lambda}, a, b)$:

$$\lim_{T\to\infty}\hat{\boldsymbol{\alpha}}^{\text{MSIQ}} = \mathbb{E}(\boldsymbol{\alpha}|\boldsymbol{R}, \lambda, a, b). \tag{2.19}$$

We can also estimate the posterior probability of each sample belonging to the consistent group: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)'$, where $\theta_d = P(E_d = 1|\boldsymbol{R}, \boldsymbol{\lambda}, a, b)$, and the estimator is $\hat{\theta}_d^{\text{MSIQ}} = \frac{1}{T}\sum_{t=1}^{T} E_d^{(t)}$. Based on this posterior probability, we can predict the state variable of each sample: $\hat{E}_d = \mathbb{1}\{\hat{\theta}_d^{\text{MSIQ}} > 1/2\}$.

To further evaluate the biological variation within the consistent group, we estimate the standard error of the MSIQ estimator given the posterior samples. For isoform j, the standard error of the respective entry in $\hat{\boldsymbol{\alpha}}^{\text{MSIQ}}$ is estimated as:

$$\hat{\sigma}_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\alpha_j^{(t)} - \hat{\alpha}_j^{\text{MSIQ}})^2}\,. \tag{2.20}$$

Even though the consistent group is assumed to have a consensus isoform proportion, it is useful to account for the biological variation, especially when the overall heterogeneity is non-negligible.

We also propose six competing estimators to demonstrate the effectiveness of MSIQ in accurate isoform quantification. We know from Chapter 2.2.3 that the log likelihood of all reads in sample d is

$$\log\left(\mathbb{P}(\boldsymbol{R}^{(d)}, \boldsymbol{Z}^{(d)}|\boldsymbol{\tau}^{(d)})\right) = \sum_{i=1}^{n_d}\sum_{j=1}^{m} I_{ij}^{(d)} \log\left(h_{ij}^{(d)}\tau_j^{(d)}\right). \tag{2.21}$$

Then the EM algorithm can be implemented to estimate $\boldsymbol{\tau}^{(d)}$. The six competing estimators are calculated using the EM algorithm based on different sets of samples:

AVG (averaging): We calculate the isoform proportion in each sample and take the average of them as the estimator of isoform proportion: $\hat{\boldsymbol{\alpha}}^{\text{AVG}} = \frac{1}{D}\sum_{d=1}^{D}\hat{\boldsymbol{\tau}}^{(d)}$.

AVG* (oracle averaging): We calculate the isoform proportion in each sample in the consistent group (truth) and take the average of them as the estimator of isoform proportion: $\hat{\boldsymbol{\alpha}}^{\text{AVG*}} = {\sum_{d=1}^{D}\hat{\boldsymbol{\tau}}^{(d)}\mathbb{1}\{E_d=1\}}\big/{\sum_{d=1}^{D}\mathbb{1}\{E_d=1\}}$.

POOL (pooling): We pool the reads in all samples together, then we use the EM algorithm to estimate the isoform proportion as $\hat{\boldsymbol{\alpha}}^{\text{POOL}}$.

POOL* (oracle pooling): We pool the reads in samples in the consistent group (truth) together, then we use the EM algorithm to estimate the isoform proportion as $\hat{\boldsymbol{\alpha}}^{\text{POOL*}}$.

MSIQa (MSIQ averaging): We calculate the isoform proportion in each sample in the consistent group (identified by MSIQ) and take the average of them as the estimator of isoform proportion: $\hat{\boldsymbol{\alpha}}^{\text{MSIQa}} = \sum_{d=1}^{D} \hat{\boldsymbol{\tau}}^{(d)} \mathbb{1}\{\hat{\theta}_d^{\text{MSIQ}} > 1/2\} \big/ \sum_{d=1}^{D} \mathbb{1}\{\hat{\theta}_d^{\text{MSIQ}} > 1/2\}$.

MSIQp (MSIQ pooling): We pool the reads of the given gene in the samples from the identified consistent group together, then we use the EM algorithm to estimate the isoform proportion as $\hat{\boldsymbol{\alpha}}^{\text{MSIQp}}$.

Among these estimators, $\hat{\boldsymbol{\alpha}}^{\text{AVG*}}$ and $\hat{\boldsymbol{\alpha}}^{\text{POOL*}}$ are oracle estimators that can be used as gold standards in simulations but are unknown in real data; $\hat{\boldsymbol{\alpha}}^{\text{MSIQa}}$ and $\hat{\boldsymbol{\alpha}}^{\text{MSIQp}}$ are MSIQ-dependent and rely on $\hat{\boldsymbol{\theta}}$ estimated by MSIQ.

2.3 Results

2.3.1 MSIQ achieves the lowest error rates in simulations

To show that MSIQ provides more accurate estimation of isoform expression than the current averaging or pooling method, we compare the REE rates of $\hat{\boldsymbol{\alpha}}^{\text{MSIQ}}$ with those of the six competing estimators: $\hat{\boldsymbol{\alpha}}^{\text{AVG*}}$, $\hat{\boldsymbol{\alpha}}^{\text{MSIQa}}$, $\hat{\boldsymbol{\alpha}}^{\text{AVG}}$, $\hat{\boldsymbol{\alpha}}^{\text{POOL*}}$, $\hat{\boldsymbol{\alpha}}^{\text{MSIQp}}$, and $\hat{\boldsymbol{\alpha}}^{\text{POOL}}$. It is difficult to compare these methods on real data because true isoform abundances in samples are unknown. Therefore, we used simulated data to compare the performance of these estimators under various scenarios.

We simulated RNA-seq reads from $3,421$ *D.melanogaster* (fly) genes that have multiple isoforms available in the UCSC Genome Browser (September 2010). Among these genes, 221 have 3 exons, 330 have 4 exons, 365 have 5 exons, 370 have 6 exons, 320 have 7 exons, 311 have 8 exons, 256 have 9 exons, 292 have 10 exons, and 956 genes have more than 10 exons. The isoform numbers increase at a roughly exponential rate as the exon numbers increase (Figure 2.3). We simulated ten samples and 500 paired-end reads from each gene

Table 2.1: Four parameter settings in the simulation study.

setting	average fragment length (bp)	read length (bp)
1	150	50
2	250	50
3	150	100
4	250	100

in every sample. To fully evaluate the performance of the seven estimators, we considered five different scenarios with different numbers of samples in the consistent group.

For each gene, we first independently generated the isoform proportion vector $\boldsymbol{\alpha}$ of samples in the consistent group and the isoform proportion vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_5$ for the other five samples. The five scenarios were designed as follows. In scenario 1, all ten samples are in the consistent group. In scenario 2, five samples are in the consistent group, and the other five samples have individual isoform proportions $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_5$. In scenario 3, seven samples are in the consistent group, and the other three samples have individual isoform proportions $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$. In scenario 4, seven samples are in the consistent group, and the other three samples have the same isoform proportion vector $\boldsymbol{\beta}_6 = \underset{\boldsymbol{\beta}_i, i=1,\ldots,5}{\arg\max} ||\boldsymbol{\beta}_i - \boldsymbol{\alpha}||_2^2$, which is the proportion vector most different from $\boldsymbol{\alpha}$. In scenario 5, seven samples are in the consistent group, and the other three samples have the same isoform proportion vector $\boldsymbol{\beta}_7 = \underset{\boldsymbol{\beta}_i, i=1,\ldots,5}{\arg\min} ||\boldsymbol{\beta}_i - \boldsymbol{\alpha}||_2^2$, which is the proportion vector most similar to $\boldsymbol{\alpha}$.

We also considered four settings of fragment and read length (Table 2.1) to examine how these parameters affect the performance of the seven estimators on isoform quantification. Under each setting, we first determined the origin of a fragment according to the designated isoform proportion, and then the starting position and the fragment length were simulated from a uniform distribution and a normal distribution, respectively. Once the starting and ending positions of the fragments were determined, the corresponding paired-end reads were also obtained.

For each scenario and parameter setting, we calculated the seven estimators, and then

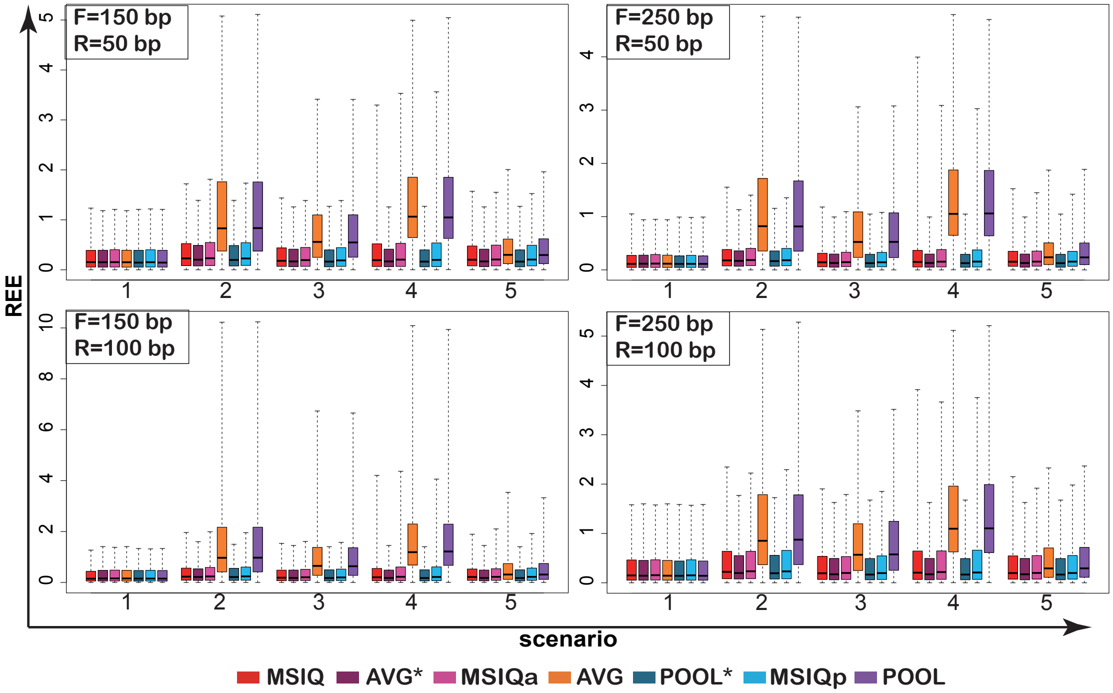
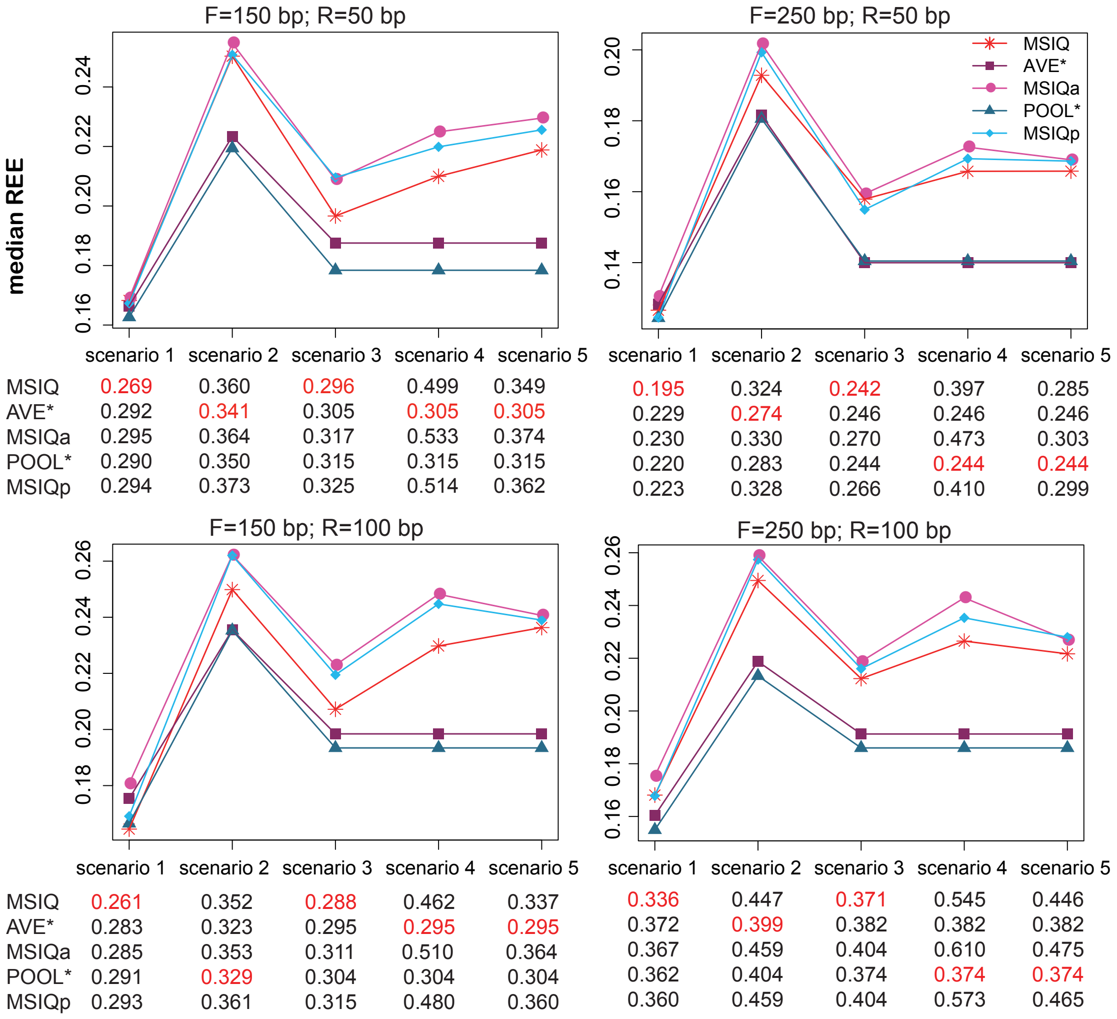


Figure 2.6: REE rates of the seven estimators in scenarios 1-5. REE rates are calculated for 2465 fly genes with 3-10 exons. In each boxplot, the REE rates of MSIQ, AVG* (oracle averaging), MSIQa, AVG (averaging), POOL* (oracle pooling), MSIQp and POOL (pooling) are plotted side by side under each scenario and the whiskers extend to the most extreme REE rates. The top-right legend of each plot displays the parameter setting: the mean fragment length (F) and the read length (R).

evaluated their estimation accuracy by calculating the REE of these estimators against the true isoform proportions (Figure 2.6). When calculating $\hat{\boldsymbol{\alpha}}^{\text{MSIQ}}$, we set the hyper-parameters in model (2.9) as $a = 7$ and $b = 2$. We have also included a sensitivity analysis of the MSIQ method on these two parameters in Appendix A.3. The results suggest that given the samples not in the consistent group (scenarios 2-5), especially when these samples constitute a large proportion or are vastly different from the consistent group, MSIQ and MSIQ-based methods achieve much smaller error rates than the averaging or pooling methods. Compared with MSIQ, AVG results in a 17.3-fold increase in the REE rates on average, and POOL results in a 17.6-fold increase. These results suggest that, compared with the direct averaging or pooling method, the MSIQ methods, which take the quality of samples into consideration,



F=150 bp; R=50 bp

	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
MSIQ	0.269	0.360	0.296	0.499	0.349
AVE*	0.292	0.341	0.305	0.305	0.305
MSIQa	0.295	0.364	0.317	0.533	0.374
POOL*	0.290	0.350	0.315	0.315	0.315
MSIQp	0.294	0.373	0.325	0.514	0.362

F=250 bp; R=50 bp

	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
MSIQ	0.195	0.324	0.242	0.397	0.285
AVE*	0.229	0.274	0.246	0.246	0.246
MSIQa	0.230	0.330	0.270	0.473	0.303
POOL*	0.220	0.283	0.244	0.244	0.244
MSIQp	0.223	0.328	0.266	0.410	0.299

F=150 bp; R=100 bp

	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
MSIQ	0.261	0.352	0.288	0.462	0.337
AVE*	0.283	0.323	0.295	0.295	0.295
MSIQa	0.285	0.353	0.311	0.510	0.364
POOL*	0.291	0.329	0.304	0.304	0.304
MSIQp	0.293	0.361	0.315	0.480	0.360

F=250 bp; R=100 bp

	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
MSIQ	0.336	0.447	0.371	0.545	0.446
AVE*	0.372	0.399	0.382	0.382	0.382
MSIQa	0.367	0.459	0.404	0.610	0.475
POOL*	0.362	0.404	0.374	0.374	0.374
MSIQp	0.360	0.459	0.404	0.573	0.465

Figure 2.7: Median REE of the MSIQ-based and oracle estimators. MSIQ outperforms MSIQa and MSIQp and gives error rates close to those of the oracle estimators. The parameters, the mean fragment length (F) and the read length (R), are listed on the top of each panel. The standard errors of the REE rates are given under each scenario. The smallest standard error in each scenario is marked in red.

can lead to more accurate isoform quantification. MSIQ can also constrain the estimation error to a much narrower range compared with direct averaging and pooling (Figure 2.6). For example, MSIQ is able to control the REE rate below 1.33 for 90% of the $2,465$ genes, while direct averaging and pooling give rise to REE rates larger than 2.00 for more than 15% of these genes. We conclude that MSIQ is a more robust method than direct averaging and pooling.

We also summarized the median REE rates under different scenarios (Figure 2.7). Since AVG and POOL are observed to have much poorer accuracy than the other five estimation methods, we exclude them from this comparison. The results show that MSIQ not only outperforms direct averaging and pooling, but also achieves more accurate abundance estimation than MSIQa and MSIQp. Compared with MSIQ's median REE rate, MSIQa and MSIQp have average REE rates that are greater by 0.009 and 0.007, respectively. We also conclude that the estimation results of MSIQ are similar to those of AVE* and POOL*, the two oracle estimators that are impossible to calculate on real data. On average, the REE rate of MSIQ is only 0.019 larger than AVE* and 0.058 larger than POOL*.

From Figure 2.7, it is obvious that the proportion of samples in the consistent group and the difference between the consistent group and other samples have large effects on the performance of all five estimating methods: MSIQ, AVG*, MISQa, POOL*, and MISQp. In scenario 1 when all the samples are in the consistent group, the five methods exhibit their lowest median REE rates. In scenario 2, which has the smallest proportion of samples in the consistent group, all five methods have the largest median REE rates among all scenarios. This phenomenon can be explained by the fact that having fewer samples in the consistent group leads to more error-prone identification of these samples and less accurate estimates of the isoform proportions. In scenarios 3-5, in which 70% of the samples are in the consistent group, the REE rates of the five methods lie between those of scenarios 1 and 2. Among all three non-oracle estimation methods, MSIQ has the best performance in all five scenarios. Unlike MSIQa and MSIQp, which discard the samples outside of the identified consistent group, MSIQ partially borrows information from these samples through the Bayesian hierarchical framework.

2.4 MSIQ has the highest estimation accuracy in real data studies

Although the true isoform proportions are mostly unknown in real data, we are still able to evaluate multi-sample isoform abundance estimation methods by creating a set of samples with the majority from one tissue of interest (the consistent group) and other samples from

Table 2.2: RNA-seq samples and consistent groups in the five sets.

set ID	consistent group	other samples	sample IDs
1	ESC	/	1-6
2	ESC	brain	1-9
3	ESC	Flux Simulator	1-6, 10-14
4	ESC	Flux Simulator	1-6, 15-19
5	ESC	Flux Simulator	1-6, 20-24

a different tissue. Even though this setup is not a realistic scenario in biological studies, it provides a good opportunity to evaluate different estimation methods. In this setup, we know the true states of the hidden state variables, i.e., which samples belong to the consistent group. We used six public RNA-seq data sets of human ESCs and consider these samples to be the consistent group (Table C.1). We mixed these samples with three samples of human brain tissues or three samples simulated by Flux Simulator [101].

We obtained five sets of RNA-seq samples by mixing the six human ESC samples in the consistent group with other samples in different combinations (Table 2.2). Since MSIQ has the best performance among all the three non-oracle MSIQ-based estimation methods in the simulation studies in Chapter 2.3.1, we only considered MSIQ and not MSIQa or MSIQp in the real data studies. We compared MSIQ with direct averaging (AVG) and pooling (POOL) on these five sets of real RNA-seq samples to estimate the isoform proportions in the consistent group (human ESC). We also evaluated three previously developed methods for single RNA-seq samples, Cufflinks [49], MISO [57], and iReckon [55]), in this comparison. For Cufflinks, we used both the averaging (Cuffa) and the pooling (Cuffp) approach to calculate the isoform proportions. For MISO and iReckon, pooling is not a feasible approach due to the large memory requirements when analyzing a merged RNA-seq sample, so we only implemented the averaging approach. When evaluating the above seven methods, we used each method's estimates on set 1 (i.e., the consistent group) as the standards. The estimation results of MSIQ, AVG, POOL, Cuffa, Cuffp, MISO, and iReckon on sets 2-5 were

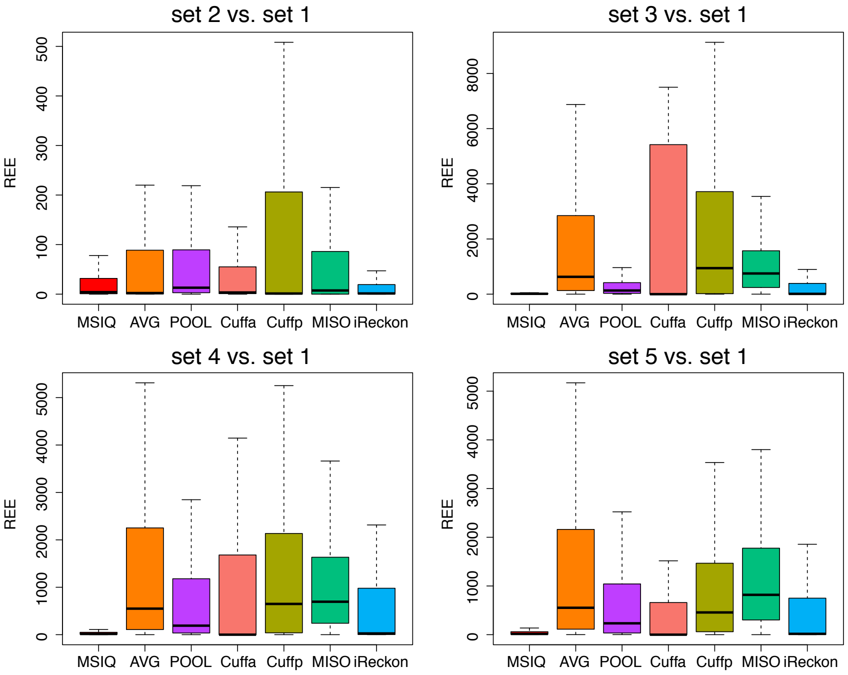


Figure 2.8: REE rates of MSIQ, AVG (averaging), POOL (pooling), Cuffa (Cufflinks averaging), Cuffp (Cufflinks pooling), MISO, and iReckon on sets 2 to 5. We used these seven estimators to perform isoform quantification on sets 2 to 5 and calculated the REE rates by treating their correpsonding estimates on set 1 as the standards.

compared with their own standards on set 1, and REE rates were calculated accordingly.

In this study, the true RNA isoform structures were extracted from the *Homo sapiens* annotation of the UCSC Genome Browser (February 2009) [102]. According to the annotation, there are 15, 268 human genes with multiple isoforms (Figure 2.3b). For each sample set, we only performed estimation for genes that have reads in all the samples. As a result, isoform proportions were calculated for 11, 091 genes in set 1, 9753 genes in set 2, 460 genes in set 3, 404 genes in set 4, and 497 genes in set 5. Comparing the REE rates of MSIQ and the other six methods (Figure 2.8), we clearly see that MSIQ achieves the lowest median error rates and the smallest inter-quantile ranges in the four comparison cases. This result is strong evidence supporting the effectiveness of MSIQ in identifying the consistent group and estimating its isoform proportions. Note that even though iReckon also leads to relatively accurate results, especially when comparing set 2 and set 1, the number of genes for which iReckon can provide estimation is much smaller compared with other methods. In the four cases, iReckon obtains estimates only for 1065, 255, 377 and 374 genes. This comparison also suggests that pooling is not an ideal approach when the depths of sequencing coverage in multiple RNA-seq samples vary greatly.

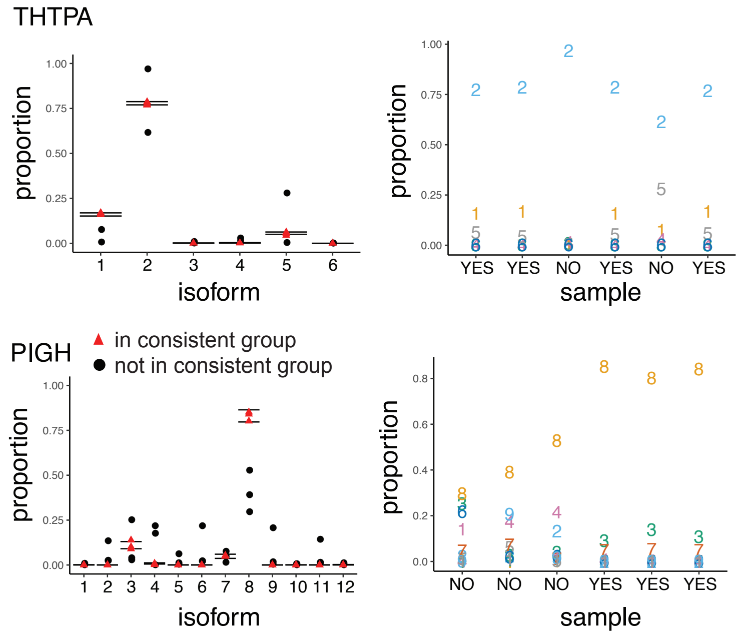


Figure 2.9: MSIQ's estimated isoform proportions and standard errors for gene *THTPA* (6 isoforms) and gene *PIGH* (12 isoforms). The left plots give the estimated proportions by isoform. The intervals denote the respective MSIQ estimator ± one standard error: $\hat{\alpha}_j^{\text{MSIQ}} \pm \hat{\sigma}_j$. The right plots give the estimated isoform proportions by sample. The numbers denote the isoform indices and the horizontal axis denotes whether the corresponding sample is identified as being within the consistent group or not.

We use set 1 (i.e., the 6 human ESC samples) to illustrate why the consistent group represents more reliable transcriptome landscapes and how the standard deviation defined in (2.20) can be used to assess the biological variation within the consistent group. Shown in Figure 2.9 are two example genes *THTPA* (6 isoforms) and *PIGH* (12 isoforms). We use these two examples to illustrate that (1) MSIQ is able to identify consistent groups that have comparably more consistent isoform abundances, and (2) the biological variation within the consistent group is much smaller compared to the overall variation among all the samples, and this variation is well captured by the estimated standard errors.

2.5 MSIQ leads to the highest correlation with NanoString counts

We present a second real data study to evaluate different methods by comparing their reported isoform abundances (in FPKM values) with NanoString counts on the same datasets. The NanoString nCounter technology is considered to be a highly reproducible and robust method for detecting gene and isoform expression [103], so the NanoString measurements can be used as a benchmark for isoform expression [72,104]. We compared our MSIQ method with three other estimation methods, Cufflinks, iReckon, and MISO, based on their performance on six samples of the human liver hepatocellular carcinoma (HepG2) cell line (Table C.1).

Even though genome-wide isoform abundances are not available for these HepG2 data, the NanoString counts are available for a small set of genes [72]. These NanoString measurements include 140 probes that correspond to 470 isoforms of 107 genes. We applied MSIQ, Cufflinks, iReckon, and MISO on the six HepG2 samples and used each method to estimate isoform abundances for this set of genes. Cufflinks and iReckon directly report the FPKM values of the relevant isoforms. MSIQ and MISO estimate isoform proportions, and the FPKM values can be calculated accordingly. For each sample, we calculated the Pearson correlation coefficient between each method's estimated isoform expression and the benchmark NanoString counts. Since the NanoString probe counts do not have a one-to-one correspondence with isoform expression, for each NanoString probe we either used the isoform with the largest expression (Figure 2.10a) or added up the expression of all the isoforms corresponding to that probe (Figure 2.10b). Overall, the estimated expression of MSIQ has the highest correlation with the NanoString counts and achieves the best consistency with this benchmark measurement. This result again suggests that MSIQ leads to more accurate isoform quantification by incorporating the information in multiple RNA-seq samples.

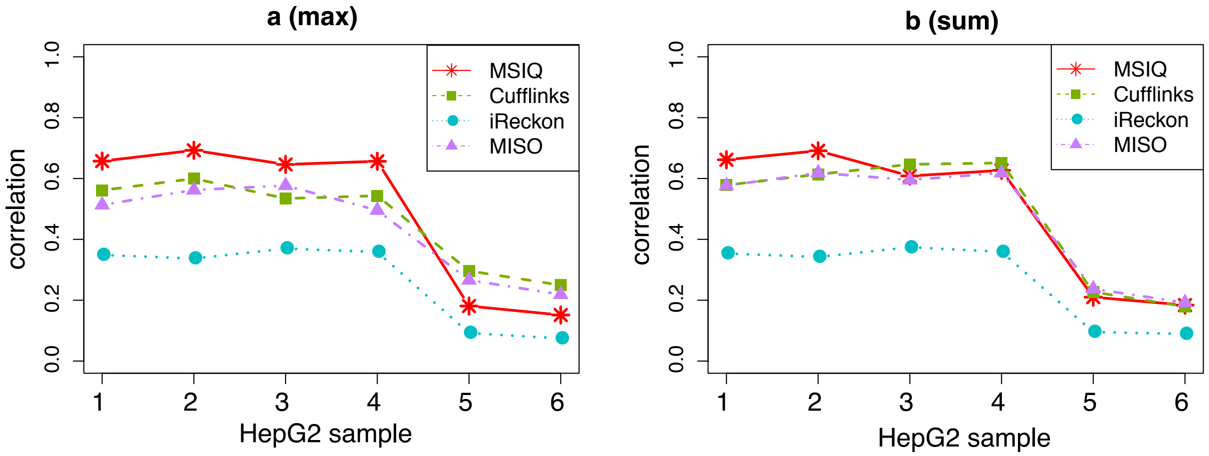


Figure 2.10: Correlation between NanoString counts and the estimated isoform expression. **a**: For each NanoString probe, the corresponding isoform with the largest estimated FPKM value was used to calculate the correlation. The standard error of the calculated correlation coefficients is between 0.069 and 0.099. **b**: For each NanoString probe, the sum of all the corresponding isoforms' estimated FPKM values was used to calculate the correlation. The standard error of the calculated correlation coefficients is between 0.065 and 0.085.

2.6 Discussion

In this chapter, we propose a new method, MSIQ, to more accurately estimate isoform expression levels associated with biological conditions of interest using multiple RNA-seq datasets. Accurate isoform quantification from RNA-seq data has long been a challenge because the existence of multiple isoforms makes it impossible to uniquely assign most reads and determine the reads' isoform origins. MSIQ tackles this challenge by utilizing data from multiple RNA-seq samples derived from the same biological condition; we reason that aggregating more information can improve accuracy in isoform abundance estimation. Unlike previous work that treats all the samples equally, MSIQ identifies a consistent group of samples that are most representative of the biological condition and estimates isoform proportions of the consistent group.

Applications of MSIQ to both simulated and real data demonstrate that MSIQ yields more accurate isoform quantification than direct averaging or pooling methods given the existence of poor quality or mislabeled samples. These results suggest MSIQ's potential as a powerful and robust transcriptomic tool for isoform quantification. MSIQ's estimation results provide accurate transcriptome profiles, which can be used to construct co-expression networks, investigate cell-type-specific isoform expression, and identify differentially expressed transcripts between two biological conditions. The MSIQ method also provides standard error estimates to measure the variability of isoform abundance within the consistent group. This information can be especially useful when users need to compare multiple tissues or cell types. We currently estimate the standard errors using the posterior samples of isoform proportions, and our method can be extended to directly model the variation at the cost of increased complexity in the model and computations. In addition to isoform abundance estimation, MSIQ can also be applied to evaluate the quality of multiple RNA-seq samples of the same tissue or cell type. This application can help researchers evaluate the reproducibility of RNA-seq samples and determine which samples to include in downstream analyses.

An important step in the MSIQ method is the identification of the consistent group, which depends on posterior draws of the hidden state variables. We currently use a Beta-Bernoulli model to describe the probability of each sample belonging to the consistent group. However, it is possible to improve the model once gold standard data for the biological condition of interest become available [20, 105]. We can extend our MSIQ model to account for the heterogeneous quality of multiple RNA-seq samples based on the similarity of the isoform abundance estimates between each sample and the gold standard. Such quality assessment can be integrated with inter-sample similarity to better identify the consistent group. As a result, the samples that have higher agreement with gold standards and high similarity with each other will be more likely to be considered as in the consistent group. This procedure is supposed to identify more reliable samples and can potentially increase the re-use of public RNA-seq data as it will provide an interpretable measure of the quality of multiple RNA-seq datasets. We would also like to point out that biological knowledge can be incorporated into MSIQ modeling to further improve abundance estimation. For example, mRNA fragments are actually not uniformly distributed within the isoforms [53], and a high correlation was observed between read coverage and genome GC content [62]. Our proposed hierarchical model can be considered an umbrella framework that can be easily extended to incorporate more detailed modeling procedures as long as these procedures use likelihoods to describe read generating processes. Such extension might help MSIQ achieve better performance on complex genes.

Another interesting extension of our MSIQ method is to model single-cell RNA-seq (scRNA-seq) data, which contain information on the technical and biological noise of isoform abundance at the single-cell level [106, 107]. ScRNA-seq data are needed for the analysis of (1) subpopulations of cells from a large heterogeneous population and (2) rare cell types, for which sufficient molecules cannot be obtained for conventional RNA-seq experiments [18]. Given scRNA-seq data, MSIQ can be iteratively utilized to evaluate the transcriptional heterogeneity and detect subpopulations (i.e., consistent groups) in the set of samples. Meanwhile, MSIQ may also reveal the principal isoform expression pattern in a given cell population. An alternative approach is to allow for multiple consistent groups as subpopulations of single cells in the modeling.

2.7 Acknowledgments

This chapter is based on my joint work with Dr. Anqi Zhao, Dr. Shihua Zhang, and Dr. Jingyi Jessica Li [56]. I would like to thank Dr. Anqi Zhao and Dr. Jessica Li for obtaining preliminary results that led to this project. I also thank Dr. Shihua Zhang and Dr. Li for their advice on the project. In addition, we thank Dr. Yucheng Yang for his help in the data processing and Dr. Katherine R. McLaughlin for commenting on our work.

CHAPTER 3

Precise Transcript Reconstruction with Bulk RNA Sequencing Data

3.1 Introduction

Alternative splicing, a post-transcriptional process during which particular exons of a gene may be included into or excluded from a mature RNA isoform transcribed from that gene, is a key contributor to the diversity of eukaryotic transcriptomes [108]. Alternative splicing is a prevalent phenomenon in multicellular organisms, and it affects approximately 90%-95% of genes in mammals [109]. Understanding the diversity of eukaryotic transcriptomes is essential to interpreting gene functions and activities under different biological conditions. In transcriptome analysis, a key task is to accurately identify the set of existing isoforms and estimate their abundance levels under a specific biological condition, because the information on isoform composition is critical to understanding the isoform-level dynamics of RNA contents in different cells, tissues, and developmental stages. Abnormal splicing events have been known to cause many genetic disorders [110], such as retinitis pigmentosa [111] and spinal muscular atrophy [112]. Accurate isoform identification and quantification will shed light on gene regulatory mechanisms of genetic diseases, thus assisting biomedical researchers in designing targeted therapies for diseases.

The identification of truly expressed isoforms is an indispensable step preceding accurate isoform quantification. However, compared with the quantification task, isoform discovery is a more challenging problem both theoretically and computationally. The reasons behind this challenge are threefold. First, next-generation RNA-seq reads are too short compared with full-length RNA isoforms. RNA-seq reads are typically no longer than 300 bp in Illumina sequencing [113] (Figure 1.1), while more than 95% of human isoforms are longer than 300 bp, with a mean length of 1712 bp (GENCODE annotation, Release 24) [75]. Due to the fact that most isoforms of the same gene share some overlapping regions, many RNA-seq reads do not unequivocally map to a unique isoform (Figure B.1a). As a result, isoform origins of those reads are ambiguous and need to be inferred from a huge pool of candidate isoforms. Another consequence of short reads is that *junction reads* spanning more than one exon-exon junctions are underrepresented in RNA-seq data, due to the difficulty of mapping junction reads (every read needs to be split into at least two segments and has all the segments mapped to different exons in the reference genome). The underrepresentation of those junction reads further increases the difficulty of accurately discovering full-length RNA isoforms. Second, the number of candidate isoforms increases exponentially with the number of exons. Hence, computational efficiency becomes an inevitable factor that every method must account for, and an effective isoform screening step is often needed to achieve accurate isoform discovery [94]. Third, it is a known biological phenomenon that often only a small number of isoforms are truly expressed under one biological condition. Given the huge number of candidate isoforms, how isoform discovery methods balance the parsimony and the accuracy of their discovered isoforms becomes a critical and meanwhile difficult issue [47, 55].

Over the past decade, researchers have developed multiple state-of-the-art isoform discovery methods to tackle one or more of the challenges mentioned above. The two earliest annotation-free methods are Cufflinks [49] and Scripture [66], which can assemble RNA isoforms solely from RNA-seq data without using annotations of known isoforms. Both methods use graph-based approaches, but they differ in how they construct graphs and then parse a graph into isoforms. Scripture first constructs a connectivity graph with nodes as genomic positions and edges determined by junction reads. It then scans the graph with fixed-sized windows, scores each path for significance, connects the significant paths into candidate isoforms, and finally refines the isoforms. Cufflinks constructs an overlap graph of mapped reads, and it puts a directed edge based on the genome orientation between two compatible reads that could arise from the same isoform. It then finds a minimal set of paths that cover all the fragments in the overlap graph. A more recent method StringTie also uses the graph idea. It first creates a splice graph with read clusters as nodes to identify isoforms and then constructs a flow network to estimate the expression levels of isoforms [67]. Another suite of methods utilize different statistical tools and regularization methods to tackle the problems of isoform discovery. For example, IsoLasso [63], SLIDE [62], and CIDANE [47] all build linear models, where read counts are summarized as the response variable, and isoform abundances are treated as parameters to be estimated. The candidate isoforms with non-zero estimated abundance are then considered as discovered. For more comprehensive discussion and comparison of existing isoform discovery methods, readers can refer to Chapter 1.4.

Aside from the intrinsic difficulty of isoform identification due to the short read lengths and the huge number of candidate isoforms, the excess biases in RNA-seq experiments further complicate the isoform discovery problem. Ideally, RNA-seq reads are expected to be uniformly distributed within each isoform. However, the observed distribution of RNA-seq reads significantly violates the uniformity assumption due to multiple sources of biases. The most commonly acknowledged bias source is the different levels of GC content in different regions of an isoform. The GC content bias was first investigated by Dohm et al. [114], and a significantly positive correlation was observed between read coverage and GC contents. Another work later showed that the effect of GC content tends to be sample-specific [115]. Another major bias source is the positional bias, which causes the uneven read coverage at different relative positions within an isoform. As a result of the positional bias, reads are more likely to be generated from certain regions of an isoform, depending on experimental protocols, e.g., whether cDNA fragmentation or RNA fragmentation is used [116,117]. Failing to correct these biases will likely lead to high false discovery rates in isoform discovery and unreliable statistical results in downstream analyses.

Current computational methods account for the non-uniformity of reads using three main approaches: adjusting read counts summarized in defined genomic regions to offset the non-uniformity biases [116,118], assigning a weight to each single read to adjust for biases [119], and incorporating the biases as model parameters in likelihood-based methods [58,59,61,120] for isoform discovery or abundance estimation.

Despite continuous efforts on developing effective computational methods to identify full-length isoforms from the next-generation RNA-seq data, the existing methods still suffer from low accuracy for genes with complex splicing structures [72, 93]. A comprehensive assessment has shown that methods achieving good accuracy in identifying isoforms for *D. melanogaster* (34,776 annotated isoforms) and *C. elegans* (61,109 annotated isoforms) fail to maintain good performance for *H. sapiens* (200,310 annotated isoforms) [72,121]. Although it is generally believed that deeper sequencing will lead to better isoform discovery results, the improvement is not significant in *H. sapiens*, compared with *D. melanogaster* and *C. elegans*, due to the complex splicing structures of human genes [18, 72]. Moreover, despite increasing accuracy of identified isoforms evaluated at the nucleotide level and the exon level, it remains challenging to improve the isoform-level performance. In other words, even when all sub-isoform elements (i.e., short components of transcribed regions such as exons) are correctly identified, accurate assembly of these elements into full length isoforms remains a big challenge.

Motivated by the observed low accuracy in identifying full-length isoforms solely from RNA-seq data, researchers have considered leveraging information from reference annotations (e.g., Ensembl [121], GENCODE [75], and UCSC Genome Browser [102]) to aid isoform discovery. Existing efforts include two approaches. In the first approach, methods extract the coordinates of gene and exon boundaries, i.e., known splicing sites, from annotations and then assemble novel isoforms based on the exons or subexons of every gene [47, 62, 67]. In the second approach, methods directly incorporate all the isoforms in annotations by simulating faux-reads from the annotated isoforms [122]. However, the above two approaches have strong limitations in their use of annotations. The first approach does not fully use annotation information because it neglects the splicing patterns of annotated isoforms, and these patterns could assist learning the relationship between short reads and full-length isoforms. The second approach is unable to filter out non-expressed annotated isoforms because researchers lack prior knowledge on which annotated isoforms are expressed in an RNA-seq sample; hence, its addition of unnecessary faux-reads will bias the isoform discovery results and lose control of the false discovery rate.

In this chapter, we propose a more statistically principled approach, AIDE (**A**nnotation-assisted **I**soform **D**iscovery and abundance **E**stimation), to leverage annotation information in a more advanced manner to increase the precision and robustness of isoform discovery. Our approach is rooted in statistical model selection, which takes a conservative perspective to search for the smallest model that fits the data well, after adjusting for model complexity. In our context, a model corresponds to a set of candidate isoforms, and a more complex model contains more isoforms. Our rationale is that a robust and conservative computational method should only consider novel isoforms as credible if adding them would significantly better explain the observed RNA-seq reads than using only the annotated isoforms. AIDE differs from many existing approaches in that it does not aim to find all seemingly novel isoforms. It enables controlling false discoveries in isoform identification by employing a statistical testing procedure, which ensures that the discovered isoforms make statistically significant contributions to explaining the observed RNA-seq reads. Specifically, AIDE learns gene and exon boundaries from annotations and also selectively borrows information from the annotated isoform structures using a stepwise likelihood-based selection approach. Instead of fully relying on the annotation, AIDE identifies non-expressed annotated isoforms and remove them from the identified isoform set. Moreover, AIDE simultaneously estimates the abundance of the identified isoforms in the process of isoform reconstruction.

3.2 Methods

3.2.1 Isoform discovery and abundance estimation using AIDE

The AIDE method is designed to independently identify and quantify the RNA isoforms of each gene. Suppose that a gene has m non-overlapping exons and J candidate isoforms. If no filtering steps based on prior knowledge is applied to reduce the set of candidate isoforms, J equals $2^m - 1$, the number of all possible combinations of exons into isoforms. The observed data are the n RNA-seq reads mapped to the gene: $\boldsymbol{R} = \{r_1, \ldots, r_n\}$. The parameters we would like to estimate are the isoform proportions $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$, where α_j is the proportion of isoform j among all the isoforms (i.e., the probability that a random read is from isoform j) and $\sum_{j=1}^{J} \alpha_j = 1$. We also introduce hidden variables $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ to denote the isoform origins of the n reads, with $Z_i = j$ indicating that read r_i is from isoform j, and $\mathbb{P}(Z_i = j) = \alpha_j$, for $i = 1, \ldots, n$.

The joint probability of read r_i and its isoform origin Z_i can be written as:

$$\begin{aligned}
\mathbb{P}\left(r_i, Z_i | \boldsymbol{\alpha}\right) &= \prod_{j=1}^{J} \mathbb{P}\left(r_i, Z_i = j | \boldsymbol{\alpha}\right)^{\mathbb{I}\{Z_i = j\}} \\
&= \prod_{j=1}^{J} \left[\mathbb{P}\left(r_i | Z_i = j\right) \alpha_j\right]^{\mathbb{I}\{Z_i = j\}} \\
&\triangleq \prod_{j=1}^{J} \left(h_{ij} \alpha_j\right)^{I_{ij}} ,
\end{aligned} \tag{3.1}$$

where $I_{ij} \triangleq \mathbb{I}\{Z_i = j\}$ indicates whether read r_i is from isoform j, and $h_{ij} \triangleq \mathbb{P}\left(r_i | Z_i = j\right)$ is the generating probability of read r_i given isoform j, calculated based on the read generating mechanism.

3.2.1.1 Read generating mechanism

We have defined h_{ij} as the generating probability of read r_i given isoform j. Specifically, if read r_i is not compatible with isoform j (read r_i contains regions not overlapping with isoform j, or vice versa), then $h_{ij} = 0$; otherwise,

$$\begin{aligned}
h_{ij} &= \mathbb{P}(\text{starting position of } r_i \mid \text{isoform } j)\mathbb{P}(\text{fragment length of } r_i \mid \text{isoform } j) \\
&\triangleq P_{ij}^{s} P_{ij}^{f} .
\end{aligned} \tag{3.2}$$

In the literature, different models have been used to calculate the starting position distribution P^s and the fragment length distribution P^f. Most of these models are built upon a basic model:

$$\begin{aligned}
P_{ij}^{s} &= 1/L_j , \\
P_{ij}^{f} &= \frac{1}{\sqrt{2\pi}\sigma_f} \exp\left\{-\frac{(l_{ij} - \mu_f)^2}{2\sigma_f^2}\right\} ,
\end{aligned} \tag{3.3}$$

where L_j is the effective length of isoform j (the isoform length minus the read length), and l_{ij} is the length of fragment i given that read r_i comes from isoform j. However, this basic model does not account for factors like the GC content bias or the positional bias. Research has shown that these biases affect read coverages differently, depending on the specific experimental protocols. For example, reverse-transcription with poly-dT oligomers results in an over-representation of reads in the 3' ends of isoforms, while reverse-transcription with random hexamers results in an under-representation of reads in the 3' ends of isoforms [123]. Similarly, different fragmentation protocols have varying effects on the distribution of reads within an isoform [117].

Given these facts, we decide to use a non-parametric method to estimate the distribution P^s of read starting positions, because non-parametric estimation is capable of accounting for the differences in the distribution due to different protocols. We use a multivariate kernel regression to infer P^s from the reads mapped to the annotated single-isoform genes. Suppose there are a total of c_s exons in the single-isoform genes. For $k' = 1, \ldots, c_s$, we use $q_{k'}$ to denote the proportion of reads whose starting positions are in exon k', among all the reads mapped to the gene containing exon k'. Given any gene with J isoforms, suppose there are c_j exons in its isoform j, $j = 1, \ldots, J$. We estimate the conditional probability that a read r_i starts from exon k ($k = 1, 2, \ldots, c_j$), given that the read is generated from isoform j, as:

$$P^s_{ij}(b_i = k) \propto \frac{\sum_{k'=1}^{c_s} \prod_{d=1}^{3} \frac{1}{h_d} K\left(\frac{x_{kd} - x_{k'd}}{h_d}\right) q_{k'}}{\sum_{k'=1}^{c_s} \prod_{d=1}^{3} \frac{1}{h_d} K\left(\frac{x_{kd} - x_{k'd}}{h_d}\right)}\,, \text{ such that } \sum_{k=1}^{c_j} P^s_{ij}(b_i = k) = 1\,, \qquad (3.4)$$

where b_i denotes the (random) index of the exon containing the starting position of the read r_i. When $b_i = k$, we use x_{k1}, x_{k2}, and x_{k3} to denote the GC content, the relative position, and the length of exon k, respectively. The GC content of exon k, x_{k1}, is defined as the proportion of nucleotides G and C in the sequence of exon k. The relative position of exon k, x_{k2}, is calculated by first linearly mapping the genomic positions of isoform j to $[0, 1]$ (i.e., the start and end positions are respectively mapped to 0 and 1) and then locating the mapped position of the center position of the exon. For example, if isoform j spans from position 100 to position 1100 in a chromosome, and the center position of an exon is 200 in the same chromosome, then the relative position of this exon is 0.1. The meaning of $x_{k'd}$'s ($k' = 1, \ldots, c_s$; $d = 1, \ldots, 3$) are be defined in the same way. The kernel function $K(\cdot)$ is set as the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$. h_d denotes the bandwidth of dimension d and is selected by cross validation. The estimation procedure of P_{ij}^s is implemented through the R package `np` [124].

As for the fragment length (P^f), we assume that it follows a truncated log normal distribution. This is because RNA fragments that are too long or too short are filtered out in the library preparation step before the sequencing step. In addition, the empirical fragment length distribution is usually skewed to right instead of being symmetric (Figure B.2). Therefore, a truncated log normal distribution generally fits well the empirical distribution of fragment lengths:

$$P_{ij}^f = \begin{cases} \dfrac{\sqrt{2}\exp\left(-\left[\log\left(\frac{l_{ij}}{m_f}\right)\right]^2 \Big/ 2\sigma_f^2\right)}{\sqrt{\pi}\sigma_f\left[\operatorname{erf}\left(\frac{1}{\sqrt{2}\sigma_f}\log\left(\frac{t_u^f}{m_f}\right)\right)-\operatorname{erf}\left(\frac{1}{\sqrt{2}\sigma_f}\log\left(\frac{t_l^f}{m_f}\right)\right)\right]l_{ij}}, & \text{if } l_{ij} \in [t_l^f, t_u^f] \\ 0, & \text{otherwise} \end{cases}, \tag{3.5}$$

where m_f and σ_f are the median and the shape parameters of the distribution, respectively; t_l^f is the lower truncation threshold and t_u^f is the upper truncation threshold. The function $\operatorname{erf}(\cdot)$, the "error function" encountered in integrating the normal distribution, is defined as $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt$.

3.2.1.2 The probabilistic model and parameter estimation in AIDE

Given the aforementioned settings, the joint probability of all the observed and hidden data is

$$\mathbb{P}\left(\boldsymbol{R}, \boldsymbol{Z} | \boldsymbol{\alpha}\right) = \prod_{i=1}^{n} \mathbb{P}\left(r_i, Z_i | \boldsymbol{\alpha}\right) = \prod_{i=1}^{n}\prod_{j=1}^{J}\left(h_{ij}\alpha_j\right)^{I_{ij}} = \prod_{i=1}^{n}\prod_{j=1}^{J}\left(P_{ij}^s P_{ij}^f \alpha_j\right)^{I_{ij}}, \tag{3.6}$$

where P_{ij}^s and P_{ij}^f are defined in equations (3.4) and (3.5). The complete log-likelihood is

$$\ell\left(\boldsymbol{\alpha} | \boldsymbol{R}, \boldsymbol{Z}\right) = \sum_{i=1}^{n}\sum_{j=1}^{J} I_{ij}\log\left(P_{ij}^s P_{ij}^f \alpha_j\right). \tag{3.7}$$

However, as $\boldsymbol{Z}$ and the resulting I_{ij}'s are unobservable, the problem of isoform discovery becomes to estimate $\boldsymbol{\alpha}$ via maximizing the log-likelihood based on the observed data:

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} \ell\left(\boldsymbol{\alpha}|\boldsymbol{R}\right) \\ &= \arg\max_{\boldsymbol{\alpha}} \log \left[\prod_{i=1}^{n} \left(\sum_{j=1}^{J} P^s_{ij} P^f_{ij} \alpha_j\right)\right] \\ &= \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{J} P^s_{ij} P^f_{ij} \alpha_j\right) , \end{aligned} \tag{3.8}$$

subject to $\alpha_j \geq 0$ and $\sum_{j=1}^{J} \alpha_j = 1$. To directly solve problem (3.8) is not easy, so we use on the EM algorithm along with the complete log-likelihood (3.7), and it follows that we can iteratively update the estimated isoform proportions as $\alpha_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} {P^s_{ij} P^f_{ij} \alpha_j^{(t)}} / {\sum_{j'=1}^{J} P^s_{ij'} P^f_{ij'} \alpha_{j'}^{(t)}}$. As the algorithm converges, we obtain the estimated isoform proportion $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_J)'$.

3.2.2 Stepwise selection in AIDE

If we directly consider all the $J = 2^m - 1$ candidate isoforms in problem (3.8) and calculate $\hat{\boldsymbol{\alpha}}$, the problem is unidentifiable when $J > n$. Even when $J \leq n$, this may lead to many falsely discovered isoforms, because the most complex model with all the possible candidate isoforms would best explain the observed reads. Therefore, instead of directly using the EM algorithm to maximize the log-likelihood with all the possible candidate isoforms, we perform a stepwise selection of isoforms based on the likelihood ratio test (LRT). This approach has two advantages. On the one hand, we can start from a set of candidate isoforms with high confidence based on prior knowledge, and then we can sequentially add new isoforms to account for reads that cannot be fully explained by existing candidate isoforms. For example, a common case is to start with annotated isoforms. On the other hand, the stepwise selection by LRT intrinsically introduces sparsity into the isoform discovery process. Even though the candidate isoform pool can be huge when a gene has a large number of exons, the set of expressed isoforms is usually much smaller in a specific biological sample. LRT can assist us in deciding a termination point where adding more isoforms does not further improve the likelihood.

The stepwise selection consists of steps with two opposite directions: the forward step and the backward step (Figure 3.1). The forward step aims at finding a new isoform to best explain the RNA-seq reads and significantly improve the likelihood given the already selected isoforms. The backward step aims at rectifying the isoform set by removing the isoform with the most trivial contribution among the selected isoforms. Since stepwise selection is in a greedy-search manner, some forward steps, especially those taken in the early iterations, may not be the globally optimal options. Therefore, backward steps are necessary to correct the search process for a better solution path.

We separate the search process into two stages. We use stepwise selection to update the identified isoforms at both stages, but the initial isoform sets and the candidate sets are different in the two stages. Stage 1 starts with a single annotated isoforms that explains the most number of reads, and it only considers the annotated isoforms as candidate isoforms. Stage 1 stops when the the forward step can no longer finds an isoform to add to the identified isoform set, i.e., the LRT does not reject the null hypothesis given the p-value threshold. Stage 2 starts with the isoforms identified in stage 1, and considers all the possible isoforms, including the annotated isoforms not chosen in stage 1, as the candidate isoforms (Figure 3.1). The initial isoform set is denoted as $S_1^{(0)}$ (stage 1) or $S_2^{(0)}$ (stage 2), the candidate isoform set is denoted as C_1 (stage 1) or C_2 (stage 2), and the annotation set is denoted as A. At stage 1, the candidate set and initial set are respectively defined as $C_1 = A$ and $S_1^{(0)} = \left\{\arg\max_{j \in A}(\text{number of reads compatible with isoform } j)\right\}$. Suppose the stepwise selection completes after t_1 steps at stage 1, and the estimated isoform proportions after step t $(t = 1, 2, \ldots, t_1)$ are denoted as $\hat{\boldsymbol{\alpha}}^{(t)} = \left(\hat{\alpha}_1^{(t)}, \ldots, \hat{\alpha}_J^{(t)}\right)'$. Note that $\forall j \notin C_1$, $\hat{\alpha}_j^{(t)} \equiv 0$ at stage 1. At stage 2, the initial set and the candidate set are respectively defined as $S_2^{(0)} = \left\{j : \hat{\alpha}_j^{(t_1)} > 0\right\}$ and $C_2 = \{1, 2, \ldots, J = 2^m - 1\}$.

We introduce how to perform forward and backward selection based on a initial isoform set $S^{(0)}$ and a candidate set C. We omit the stage number subscripts for notation simplicity. At both stages, we first use the EM algorithm to estimate the expression levels of the initial isoform set $S^{(0)}$: $\hat{\boldsymbol{\alpha}}^{(0)} = \arg\max_{\boldsymbol{\alpha}} \ell\left(\boldsymbol{\alpha}|\boldsymbol{R}\right) = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log\left(\sum_{j \in S^{(0)}} P_{ij}^{s} P_{ij}^{f} \alpha_j\right)$, subject to $\alpha_j \geq 0$ if $j \in S^{(0)}$, $\alpha_j = 0$ if $j \notin S^{(0)}$, and $\sum_{j \in S^{(0)}} \alpha_j = 1$.

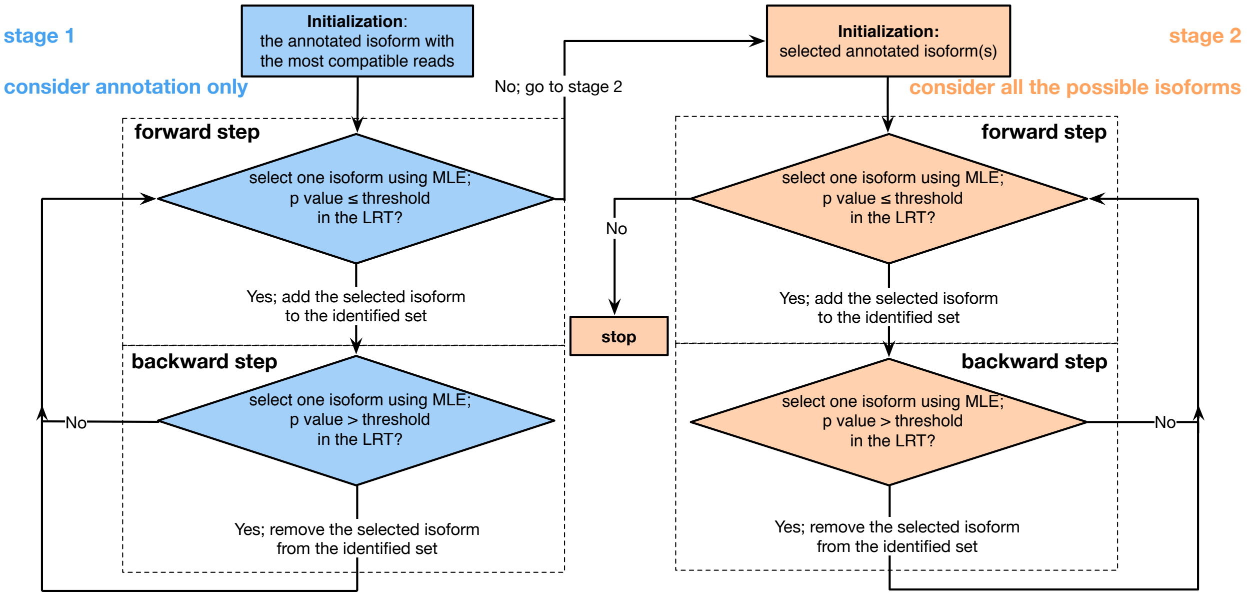

stage 1
consider annotation only
Initialization: the annotated isoform with the most compatible reads
forward step
select one isoform using MLE; p value ≤ threshold in the LRT?
Yes; add the selected isoform to the identified set
backward step
select one isoform using MLE; p value > threshold in the LRT?
No
Yes; remove the selected isoform from the identified set
No; go to stage 2
No
stop
stage 2
consider all the possible isoforms
Initialization: selected annotated isoform(s)
forward step
select one isoform using MLE; p value ≤ threshold in the LRT?
Yes; add the selected isoform to the identified set
backward step
select one isoform using MLE; p value > threshold in the LRT?
No
Yes; remove the selected isoform from the identified set

Figure 3.1: Workflow of the stepwise selection procedure of the AIDE method. Stage 1 starts with a single annotated isoform compatible with the most reads, and stage 2 starts with the annotated isoforms selected at stage 1. In the forward step at both stages, AIDE identifies the isoform that mostly increases the likelihood, and it uses the LRT to decide whether this increase is statistically significant. If significant, AIDE adds this isoform to its identified isoform set; otherwise, AIDE terminates the current stage. In the backward step at both stages, AIDE finds the isoform in its identified set such that the removal of this isoform decreases the likelihood the least, and it uses the LRT to decide whether this decrease is statistically significant. If insignificant, AIDE removes this isoform from its identified set; otherwise, AIDE keeps the identified set. AIDE stops when the forward step at stage 2 no longer adds any candidate isoform to the identified set.

3.2.2.1 Forward step

The identified isoform set at step t is denoted as $S^{(t)} = \{j : \hat{\alpha}_j^{(t)} > 0\}$. The log-likelihood at step t is $\ell^{(t)} = \ell\left(\hat{\boldsymbol{\alpha}}^{(t)} | \boldsymbol{R}\right) = \sum_{i=1}^{n} \log\left(\sum_{j \in S^{(t)}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t)}\right)$. At step $(t+1)$, we consider adding one isoform $k \in C \backslash S^{(t)}$ into $S^{(t)}$ as a forward step. Given $S^{(t)}$ and k, we estimate the corresponding isoform proportions as

$$\hat{\boldsymbol{\alpha}}^{(t,k)} = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)} \cup \{k\}} P_{ij}^{s} P_{ij}^{f} \alpha_j \right) , \tag{3.9}$$

subject to $\alpha_j \geq 0$ if $j \in S^{(t)} \cup \{k\}$, and $\alpha_j = 0$ otherwise. Then we choose the isoform

$$k^* = \arg\max_{k \in C \setminus S^{(t)}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)} \cup \{k\}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t,k)} \right) , \tag{3.10}$$

which maximizes the likelihood among all the candidate isoforms. Then the log-likelihood with the addition of this isoform k^* becomes $\ell^* = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)} \cup \{k^*\}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t,k^*)} \right)$. To decide whether to follow the forward step and add isoform k^* to the identified isoform set, we use the LRT to test the null hypothesis ($H_0 : S^{(t)}$ is the true isoform set from which the RNA-seq reads were generated) against the alternative hypothesis ($H_a : S^{(t)} \cup \{k^*\}$ is the true isoform set). Under H_0 we asymptotically have $-2(\ell^{(t)} - \ell^*) \sim \chi^2(1)$. If the null hypothesis is rejected at a pre-specified significance level, then $S^{(t+1)} = S^{(t)} \cup \{k^*\}$, $\hat{\boldsymbol{\alpha}}^{(t+1)} = \hat{\boldsymbol{\alpha}}^{(t,k^*)}$, and the log-likelihood is updated as $\ell^{(t+1)} = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t+1)} \right)$. Otherwise, $S^{(t+1)} = S^{(t)}$, $\hat{\boldsymbol{\alpha}}^{(t+1)} = \hat{\boldsymbol{\alpha}}^{(t)}$, and $\ell^{(t+1)} = \ell^{(t)}$.

3.2.2.2 Backward step

Suppose we add an isoform to the identified isoform set at step t, the updated isoform set is then denoted as $S^{(t+1)}$. We subsequently consider removing one isoform $k \in S^{(t+1)}$ at a backward step. Given $S^{(t+1)}$ and k, we estimate the corresponding isoform proportions as

$$\hat{\boldsymbol{\alpha}}^{(t+1,-k)} = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)} \setminus \{k\}} P_{ij}^{s} P_{ij}^{f} \alpha_j \right) ,$$

subject to $\alpha_j \geq 0$ if $j \in S^{(t+1)} \setminus \{k\}$, and $\alpha_j = 0$ otherwise. Then we choose the isoform

$$k^- = \arg\max_{k \in S^{(t+1)}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)} \setminus \{k\}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t+1,-k)} \right) ,$$

which maximizes the likelihood among all the isoforms in $S^{(t+1)}$. Then the log-likelihood with the removal of this isoform k^- becomes $\ell^- = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)} \setminus \{k^-\}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t+1,-k^-)} \right)$.

To decide whether to follow the backward step and remove isoform k^- from the identified isoform set, we use the LRT to test the null hypothesis (H_0 : $S^{(t+1)} \backslash \{k^-\}$ is the true isoform set from which the RNA-seq reads were generated) against the alternative hypothesis (H_a : $S^{(t+1)}$ is the true isoform set). Under H_0 we asymptotically have $-2(l^- - l^{(t+1)}) \sim \chi^2(1)$. If the null hypothesis is not rejected at a pre-specified significance level, then $S^{(t+2)} = S^{(t+1)} \backslash \{k^-\}$, $\hat{\boldsymbol{\alpha}}^{(t+2)} = \hat{\boldsymbol{\alpha}}^{(t+1,-k^-)}$, and the log-likelihood is updated as $\ell^{(t+2)} = \sum_{i=1}^{n} \log\left(\sum_{j \in S^{(t+2)}} P_{ij}^{s} P_{ij}^{f} \hat{\alpha}_j^{(t+2)}\right)$. Otherwise, $S^{(t+2)} = S^{(t+1)}$, $\hat{\boldsymbol{\alpha}}^{(t+2)} = \hat{\boldsymbol{\alpha}}^{(t+1)}$, and $\ell^{(t+2)} = \ell^{(t+1)}$.

At both stages 1 and 2, we iteratively consider the forward and backward steps and stop the algorithm at the first time when a forward step no longer adds an isoform to the identified set (Figure 3.1). To determine whether to reject a null hypothesis in a LRT, we set a threshold on the p-value. The default threshold is 0.01 divided by the number of genes considered for isoform discovery. Unlike the thresholds set on the FPKM values or isoform proportions in other methods, this threshold on p-values allows users to tune the AIDE method based on their desired level of statistical significance. A larger threshold generally leads to more discovered isoforms and a better recall rate, while a smaller threshold leads to fewer discovered isoforms that are more precise.

3.3 Results

The AIDE method utilizes the likelihood ratio test to identify isoforms via a stepwise selection procedure, which gives priority to the annotated isoforms and selectively borrows information from their structures. AIDE achieves simultaenous isoform discovery and abundance estimation based on a carefully constructed probabilistic model of RNA-seq read generation, and the stepwise selection process consists of model parameter estimation and likelihood ratio tests (Figure 3.1). We first conducted a proof-of-concept simulation study to verify the efficiency and accuracy of the AIDE method (Appendix A.8). This study demonstrates the effectiveness of AIDE in removing non-expressed annotated isoforms and identifying novel isoforms with higher accuracy. Second, we used a transcriptome-wide study to evaluate the performance of AIDE and three other widely used methods (Cufflinks, SLIDE, and StringTie) provided with various read coverages and annotations of different quality. Third, we further assessed the accuracy of these four methods on real human and mouse RNA-seq data. We also experimentally validated the discovery of AIDE and showed that it can identify isoforms with biological relevance. Fourth, we compared the performance of these methods with the results from the long-read sequencing technologies. Finally, we evaluated the isoform abundance estimation results of these four methods using a benchmark Nanostring dataset. In all the above studies, AIDE demonstrated its advantages in achieving the highest precision in isoform discovery and the best accuracy in isoform quantification among the four methods.

3.3.1 AIDE outperforms state-of-the-art methods in simulations

We compared AIDE with three other state-of-the-art isoform discovery methods, Cufflinks [49], StringTie [67], and SLIDE [62], in a simulation setting that well mimicked real RNA-seq data analysis. The four methods tackle the isoform discovery task from different perspectives. Cufflinks assembles isoforms by constructing an overlap graph and searching for isoforms as sparse paths in the graph. SLIDE utilizes a regularized linear model, and demonstrated precise results in large-scale comparisons [72]. StringTie uses a network-based algorithm, and achieves the best computational efficiency and memory usage among the existing methods [67]. Unlike all these three methods, our method AIDE, built upon a likelihood model and stepwise selection, converts the annotation-assisted isoform discovery problem into a statistical variable selection problem.

To conduct a fair assessment of the four methods, we simulated RNA-seq datasets using the R package `polyester` [117], which uses both built-in models and real RNA-seq datasets to generate synthetic RNA-seq data that exhibit similar properties to those of real data. We simulated eight human RNA-seq datasets with eight read coverages (10x, 20x, . . . , 80x) and pre-determined isoform fractions (Appendix A.4). An "nx" coverage means that an exonic genomic locus is covered by n reads on average. We compared the accuracy of the four methods supplied with annotations of different quality. In real scenarios, annotations contain both expressed (true) and non-expressed (false) isoforms in a specific RNA-seq sample. Annotations might miss some expressed isoforms in that RNA-seq sample because alternative splicing is known to be condition-specific and widely diverse across different conditions. Therefore, it is critical to evaluate the extent to which different methods rely on the accuracy of annotations, specifically the annotation purity and completeness, which we define as the proportion of expressed isoforms among the annotated ones and the proportion of annotated isoforms among the expressed ones, respectively. We constructed nine sets of synthetic annotations with varying purity and completeness (Table 3.1).

Table 3.1: Synthetic annotations. Purity and completeness of the nine sets of synthetic annotations were calculated based on the truly expressed isoforms in the simulated data.

synthetic annotation set	1	2	3	4	5	6	7	8	9
purity	40%	40%	40%	60%	60%	60%	80%	80%	80%
completeness	40%	60%	80%	40%	60%	80%	40%	60%	80%

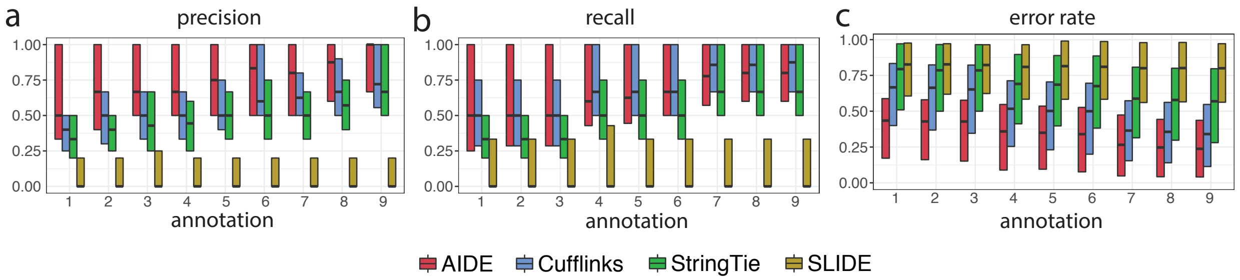


Figure 3.2: Isoform discovery and abundance estimation results of AIDE, Cufflinks, StringTie, and SLIDE on simulated RNA-seq data with 10x coverage. Each boxplot gives the 1st quantile, median, and 3rd quantile of the per-gene results given each set of synthetic annotation. **a:** precision of isoform discovery; **b:** recall of isoform discovery; **c:** error rates of abundance estimation.

We first compared AIDE and the other three methods Cufflinks, SLIDE, and StringTie in terms of their isoform discovery accuracy at the individual gene level. Regarding both precision (i.e., the proportions of expressed isoforms in the discovered isoforms) and recall (i.e., the proportions of discovered isoforms in the expressed isoforms), AIDE outperforms the other three methods with all the nine sets of synthetic annotations (Figures 3.2, B.3, and B.4). Especially with the less accurate synthetic annotation sets 1-4, AIDE demonstrates its clear advantages in precision and recall thanks to its stepwise selection strategy, which prevents AIDE from being misled by the wrongly annotated isoforms. When the read coverage is 10x and the annotations have 40% purity (sets 1-3), the median precision of AIDE is as high as the 3rd-quantile precision of Cufflinks. In addition, AIDE achieves high precision ($> 75\%$) much more frequently than the other three methods (Figure 3.2a). In terms of the recall, AIDE and Cufflinks exhibit better capability in correctly identifying truly expressed isoforms than StringTie and SLIDE. AIDE achieves high recall ($> 75\%$) in more genes when the annotation purity is 40%, and its recall rates are close to those of Cufflinks when the annotation purity is increased to 60% and 80% (Figure 3.2b). We also found that the annotation purity is more important than the annotation completeness for isoform discovery (Figure 3.2). This observation suggests that if practitioners have to choose between two annotation sets, one with high purity but low completeness and the other with low purity but high completeness, they should use the former annotation set as input into AIDE.

As a concrete example, for the human gene *DPM1*, annotation set 1 has a 67% purity and a 67% completeness, and annotation set 9 has a 60% purity and a 100% completeness. Thanks to its capacity to selectively incorporate information from the annotated isoforms, AIDE successfully identified the shortest truly expressed isoform, which is missing in the annotation set 1 (Figure B.5). With annotation set 9 that contains two non-expressed isoforms, AIDE also correctly identified the three truly expressed isoforms (Figure B.6). In contrast, Cufflinks, StringTie, and SLIDE missed some of the truly expressed isoforms with both annotation sets. Specifically, they missed the shortest expressed isoform not in annotation set 1, and they identified too many non-expressed isoforms with the less pure annotation set 9.

We also summarized the genome-wide average precision, recall, and F scores of AIDE and the other three methods at three different levels: base, exon, and isoform levels, with each of

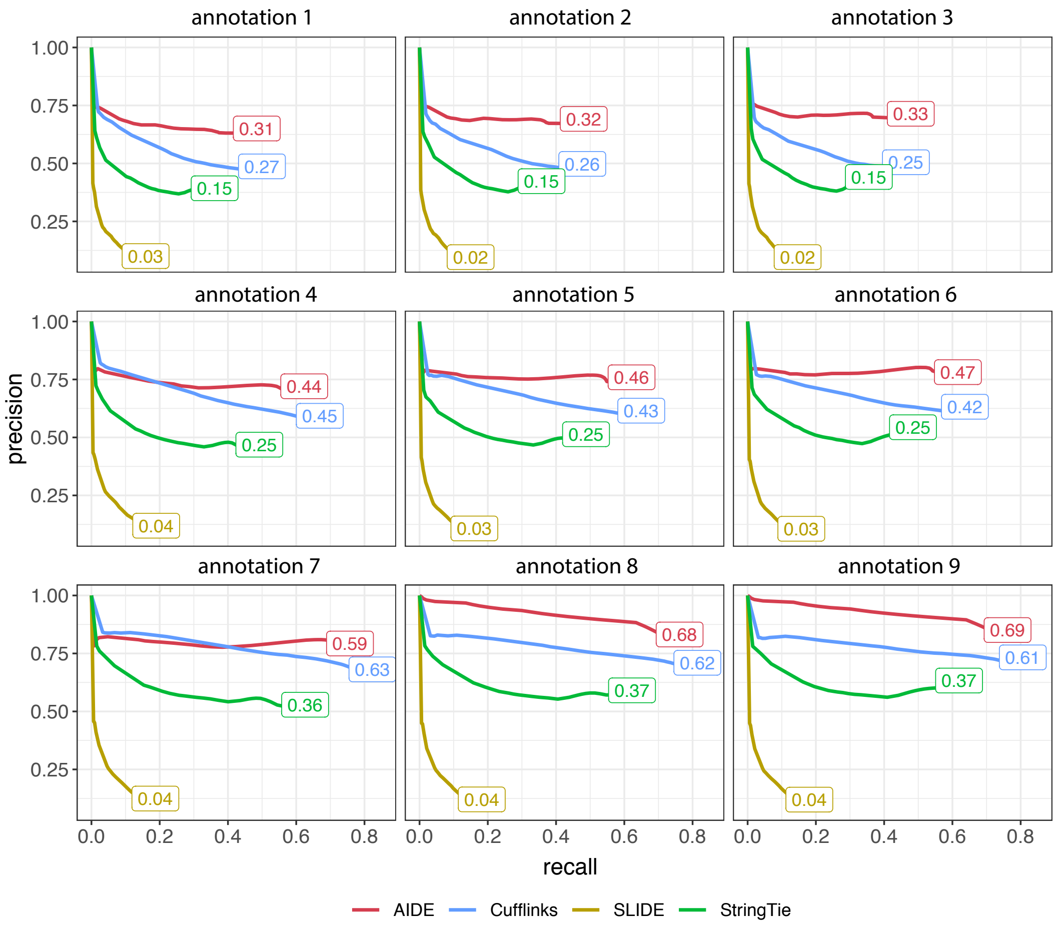


Figure 3.3: Precision-recall curves of AIDE and the other three isoform discovery methods in simulation. Given each synthetic annotation set, we applied AIDE, Cufflinks, StringTie, and SLIDE for isoform discovery, and summarized the expression levels of the predicted isoforms using the FPKM values. The precision-recall curves were obtained by thresholding the FPKM values of the predicted isoforms. The AUC of each method is also marked in the plot. The shown results are based on RNA-seq data with a 10x coverage.

the nine synthetic annotation sets (Figure B.7, Appendix A.5). All the four methods have high accuracy at the base and exon levels regardless of the annotation accuracy. However, even when exons are correctly identified, it remains challenging to accurately assemble exons into full-length isoforms. At the isoform level, AIDE achieves the best precision and F scores, as well as recall rates comparable to Cufflinks. In addition to the reconstruction accuracy based on the initial output of each method, we also compared the precision-recall curves of different methods by applying varying thresholds on the estimated isoform expression levels

(Figure 3.3). Regardless of the annotation quality, AIDE achieves the highest precision when all the methods lead to the same recall. These results demonstrate the advantage of AIDE in achieving high precision and low false discovery rates in isoform discovery.

As the true isoform proportions were specified in this simulation study, we also compared AIDE with the other three methods in terms of their accuracy in isoform abundance estimation. We use $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$ to denote the proportions of J possible isoforms enumerated from a given gene's known exons, and we use $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_J)'$ to denote the estimated proportions by a method. We define the estimation error rate as: $e(\hat{\boldsymbol{\alpha}}) = \frac{1}{2}\sum_{j=1}^{J} |\alpha_j - \hat{\alpha}_j|$. The error rate is a real value in $[0, 1]$, with a value 0 representing a 100% accuracy. With all the nine synthetic annotation sets, AIDE achieves the overall smallest error rates (Figure 3.2c).

3.3.2 AIDE improves isoform discovery in real data studies

We performed another transcriptome-wide comparison of AIDE and the other three methods to further validate the robustness and reproducibility of AIDE based on real data. Since transcriptome-wide benchmark data are unavailable for real RNA-seq experiments, we used the isoforms in the GENCODE annotation as a surrogate basis for evaluation [75]. For every gene, we randomly selected half of the annotated isoforms and input them as partial annotations into every isoform discovery method. We collected from public data repository three human embryonic stem cell (ESC) datasets and three mouse bone marrow-derived macrophage (BMDM) datasets (Table C.2). For each gene, we applied AIDE, Cufflinks, StringTie, and SLIDE to these six datasets for isoform discovery with partial annotations, and we evaluated the accuracy of each method by comparing their identified isoforms with the complete set of annotated isoforms. Although the annotated isoforms are not equivalent to the truly expressed isoforms in the six samples from which RNA-seq data were generated, the identified isoforms, if accurate, are supposed to largely overlap with the annotated isoforms given the quality of human and mouse annotations. Especially if we assume the human and mouse annotations are unions of known isoforms expressed under various well-studied biological conditions including human ESCs and mouse BMDMs, it is reasonable to use those annotations to estimate the precision of the discovered isoforms, i.e., what proportions of the discovered isoforms are expressed.

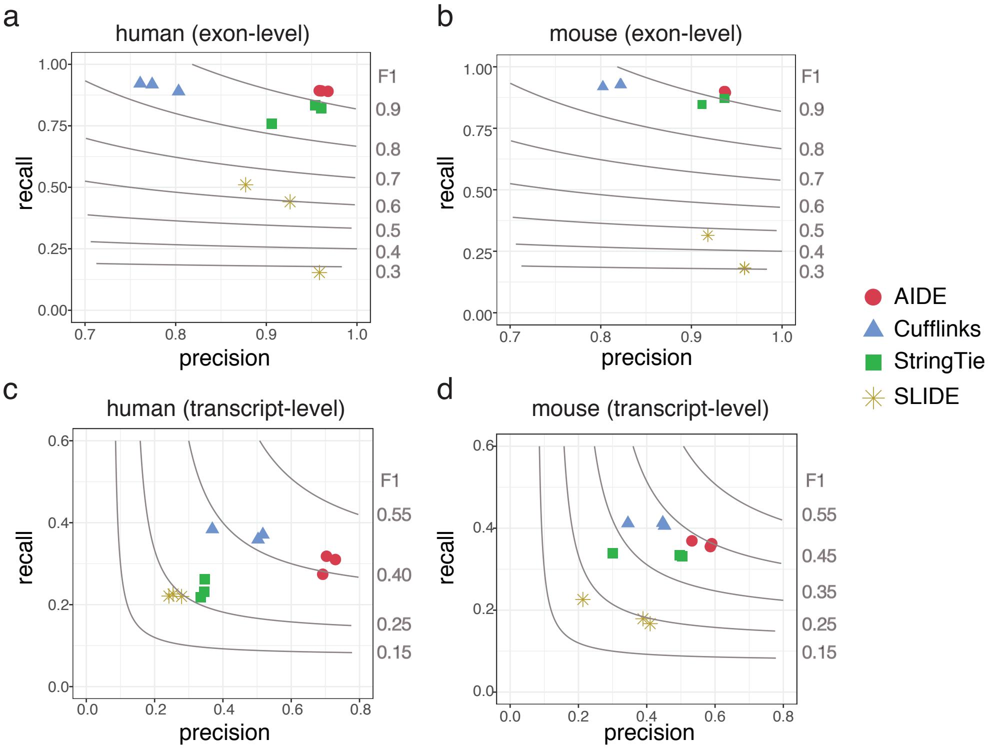


Figure 3.4: Comparison of AIDE and the other three methods in real data studies. **a**: exon-level accuracy on the human ESC samples; **b**: exon-level accuracy on the mouse BMDM samples; **c**: transcript-level accuracy on the human ESC samples; **d**: transcript-level accuracy on the mouse BMDM samples. The gray contours denote the F scores, as marked on the right of each panel.

We summarized the isoform discovery accuracy of the four methods on each dataset at both the exon level and the transcript level (Figure 3.4). At the exon level, AIDE has the highest precision and recall on all the six datasets, achieving F scores greater than 90% (Figure 3.4a-b). The second best method at the exon level is StringTie. Since connecting exons into the full-length isoforms is much more challenging than simply identifying the individual exons, all the methods have lower accuracy at the isoform level than at the exon level. While having slightly lower recall rates than Cufflinks, AIDE achieves the highest precision ($\sim$ 70% on human data and $\sim$ 60% on mouse data) at the isoform level (Figure

3.4c-d). Moreover, when all the methods achieve the same recall after thresholding the estimated isoform expression levels, AIDE has the largest precision on all the six samples (Figure B.8). In the three mouse BMDM samples, AIDE identified novel isoforms for the genes *MAPKAPK2*, *CXCL16*, and *HIVEP1*, which are known to play important roles in macrophage activation [125–127]. Since AIDE has higher precision rates than the other three methods, these novel isoforms found by AIDE are worth investigating in macrophage studies.

AIDE is able to achieve more precise isoform discovery than existing methods because it utilizes a statistical model selection principle to determine whether to identify a candidate isoform as expressed. We used two example genes *ZBTB11* and *TOR1A*, to illustrate the superiority of AIDE over the other three methods (Figure 3.5). The genome browser plots clearly show that AIDE identifies the annotated isoforms with the best precision, while the other methods either miss some annotated isoforms or lead to too many false discoveries. We also used these two genes to show that AIDE is robust to the choice of the p-value threshold used in the LRTs. The default choice of the threshold is $\frac{0.01}{\text{total number of genes}}$, which is 4.93×10^{-7} for human samples and 4.54×10^{-7} for mouse samples. We used AIDE to identify isoforms with different thresholds and tracked how the results change while the threshold decreases from 10^{-2} to 10^{-10} (Figure B.9). As expected, AIDE tends to discover slightly more isoforms with a larger threshold, and it becomes more conservative with a smaller threshold. However, the default threshold leads to accurate results for these four genes, and the discovered isoform set remains stable around the default threshold.

3.3.3 PCR-Sanger sequencing validates the effectiveness of AIDE

Since AIDE and Cufflinks have demonstrated higher accuracy than other methods in the assessment of genome-wide isoform discovery, we further evaluated the performance of these two methods on a small cohort of RNA-seq datasets using polymerase chain reaction (PCR) followed by Sanger sequencing. We applied both AIDE and Cufflinks to five breast cancer RNA-seq datasets for isoform identification with GENCODE annotation version 24. After

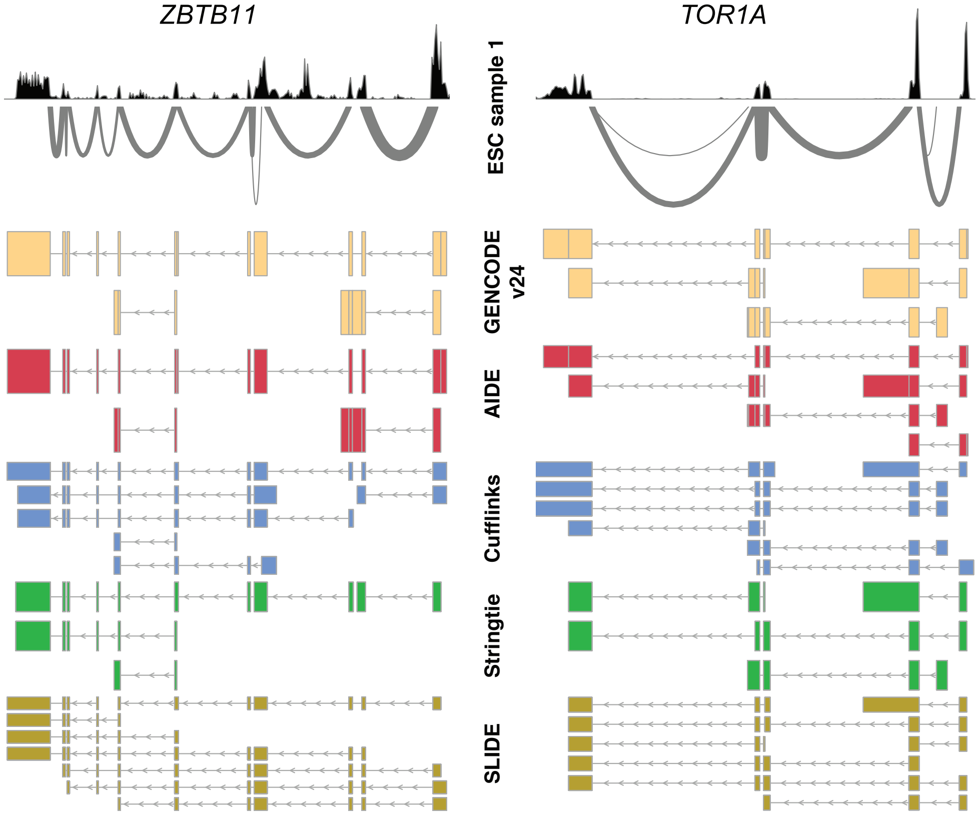


Figure 3.5: Isoform discovery for the human genes *ZBTB11* and *TOR1A* based on real data. The histogram and the sashimi plot denote the RNA-seq reads mapped to the two genes in the human ESC sample 1 (Table C.2). Isoform discovery was based on the GENCODE human annotation version 24. AIDE achieves the best accuracy among the four methods.

comparing the genome-wide isoform discovery results, we randomly selected ten genes that have annotated transcripts uniquely predicted by only AIDE or Cufflinks with FPKM > 2 for experimental validation (Appendix A.6). We summarized the genes into two categories: six genes with annotated isoforms identified only by Cufflinks but not by AIDE (category 1), and four genes with annotated isoforms identified only by AIDE but not by Cufflinks (category 2). For four genes in category 1, *MTHFD2*, *NPC2*, *RBM7*, and *CD164*, the experimental validation found that the isoforms uniquely predicted by Cufflinks were false positives (Figure 3.6a-d). Specifically, both AIDE and Cufflinks correctly identified the full-length isoforms *MTHFD2-201*, *NPC2-207*, *RBM7-203*, *CD164-003* for the four genes, respectively. However,

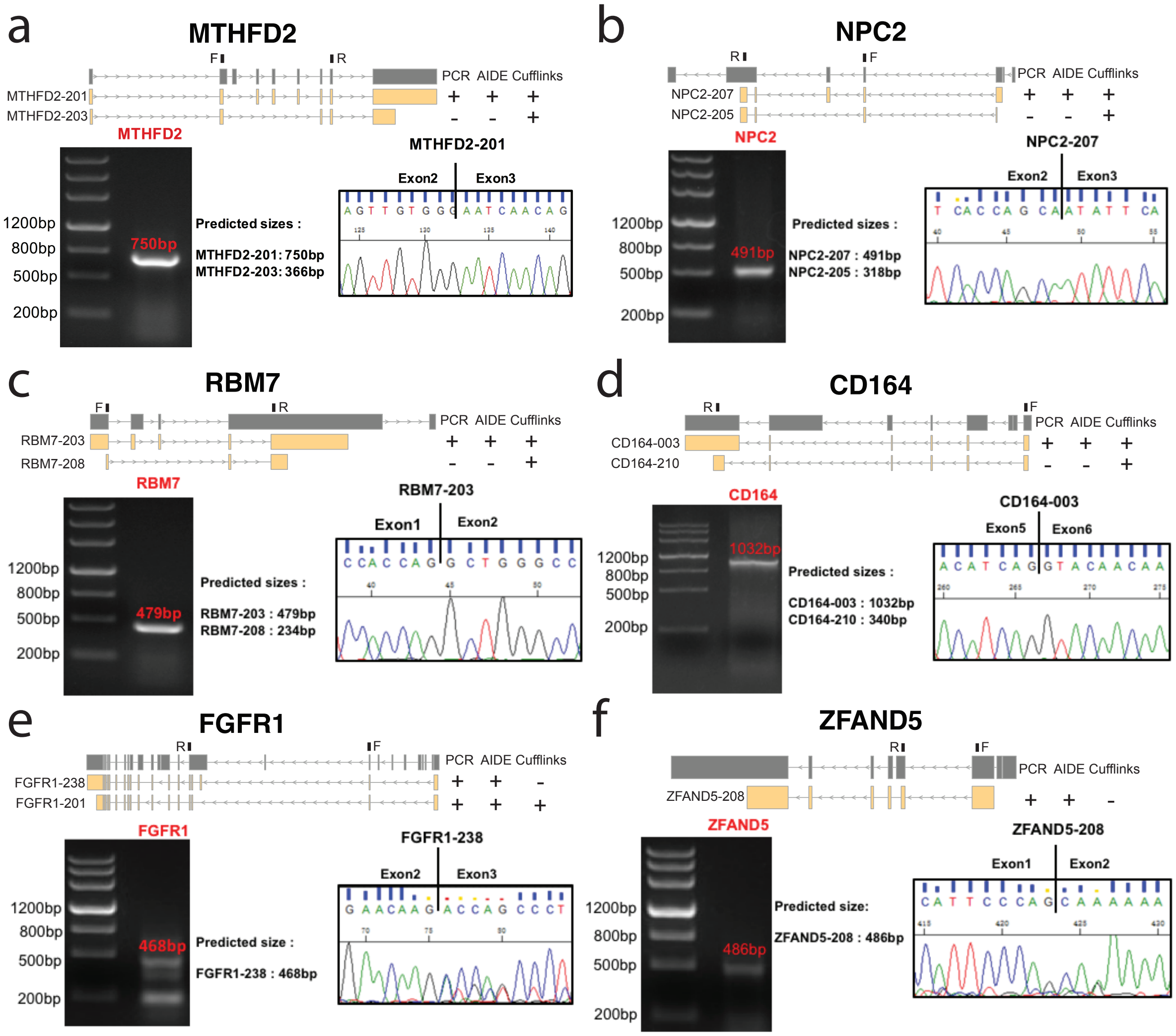


Figure 3.6: Experimental validation of isoforms predicted by AIDE and Cufflinks. Isoforms of genes *MTHFD2* (**a**), *NPC2* (**b**), *RBM7* (**c**), *CD164* (**d**), *FGFR1* (**e**), and *ZFAND5* (**f**) were validated by PCR and Sanger sequencing. The isoforms to validate (yellow) are listed under each gene (dark gray), with + / − indicating whether an isoform was / was not identified by PCR or a computational method. The forward (F) and reward (R) primers are marked on top of each gene. For each gene, the agarose gel electrophoresis result demonstrates the molecular lengths of PCR products, and the chromatography of Sanger sequencing confirms that the sequence around the exon-exon junction is unique to the PCR-validated isoform.

the isoforms *MTHFD2-203*, *NPC2-205*, *RBM7-208*, *CD164-210* predicted only by Cufflinks were all false discoveries. The validation results of category 1 indicate the potential of AIDE to effectively reduce false positive prediction of full-length isoforms compared with Cufflinks. For two genes in category 2, we validated the isoforms uniquely predicted by AIDE as true positives (Figure 3.6e-f). AIDE correctly identified isoforms *FGFR1-238* and *FGFR1-201* for the *FGFR1* gene, as well as isoform *ZFAND5-208* for the *ZFAND5* gene. On the other hand, Cufflinks only identified *FGFR1-201* and missed the other two isoforms. The validation results of category 2 suggest that AIDE also has good recall performance in this case study.

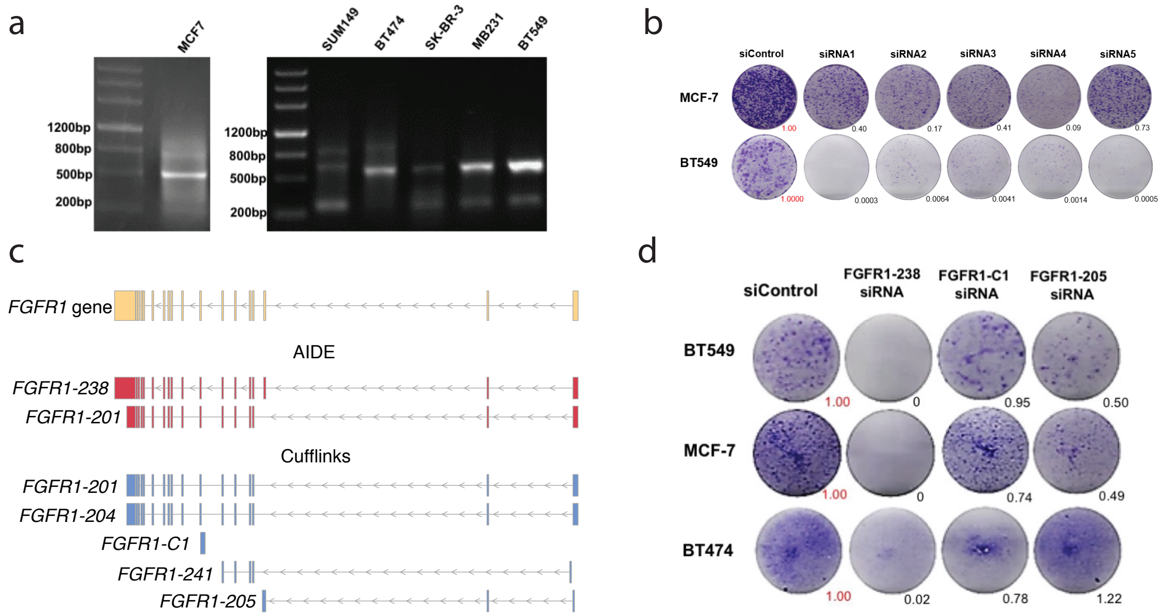


Figure 3.7: Biological functions of the isoform *FGFR1-238*. **a**: PCR experiments validated the expression of *FGFR1-238* in breast cancer cell lines MCF7, SUM149, BT474, SK-BR-3, MB231, and BT549. **b**: Long-term colonogenic assay with lipo3000 controls ("siControl") and *FGFR1-238* knockdowns. Tumor growths relative to the siControl were quantified by the ImageJ software [128]. **c**: *FGFR1* isoforms identified by AIDE and Cufflinks. **d**: Long-term colonogenic assay with siControl (negative control), si-*FGFR1-238* (positive control), si-*FGFR1-205*, and si-*FGFR1-C1*. Tumor growths relative to the siControl were quantified by the ImageJ software.

3.3.4 AIDE identifies isoforms with biological relevance

We investigated the biological function of *FGFR1-238*, an isoform predicted by AIDE but not by Cufflinks. Since *FGFR1-238* was identified in the breast cancer RNA-seq samples, we evaluated its function in breast cancer development by a loss-of-function assay. In detail, we validated the expression of *FGFR1-238* in breast cancer cell lines MCF7, BT549, SUM149, MB231, BT474, and SK-BR-3 using PCR (Figure 3.7a), and we designed primers to uniquely amplify a sequence of 533 bp in its 18-th exon. Results show that high levels of *FGFR1-238* were detected in cell lines MCF7, BT549, MB231, and BT474 (Figure 3.7a). Next, we designed five small interfering RNAs (siRNAs) that specifically target the unique coding sequence of *FGFR1-238* (Appendix A.7). Then we studied the dependence of tumor cell growth on the expression of *FGFR1-238* by conducting a long-term (10 days) cell proliferation assay in the presence or absence (control) of siRNA knockdown. Our experimental results clearly show that the knockdown of the *FGFR1-238* isoform inhibits the survival of MCF7 and BT549 cells (Figure 3.7b).

To further validate the specific biological function of isoform *FGFR1-238*, we also designed two siRNAs targeting two isoforms *FGFR1-205* and *FGFR1-C1* (novel) which were predicted by Cufflinks (Figure 3.7c). The expressions of *FGFR1-205* and *FGFR1-C1* were validated in three breast cancer cell lines, BT549, MCF7, and BT474, by PCR experiments. The three isoforms were knocked down in the host mammalian cells BT549, MCF7, and BT474 by RNA interference, respectively. The colonogenic assay shows that only the deletion of *FGFR1-238* but not *FGFR1-205* or *FGFR1-C1* obviously impacted long term cell survival (Figure 3.7d). Therefore, *FGFR1* and especially its isoform *FGFR1-238* could be promising targets for breast cancer therapy, implying the ability of AIDE in identifying full-length isoforms with biological functions in pathological conditions.

To further compare AIDE and other reconstruction methods in identifying isoforms with biological functions, we also applied AIDE, Cufflinks, and StringTie to the RNA-seq data of three melanoma cell lines. As one of the most predominant driver oncogenes, the tumorigenic function of the *NRAS* gene was well documented [130]. Recently, some novel isoforms of

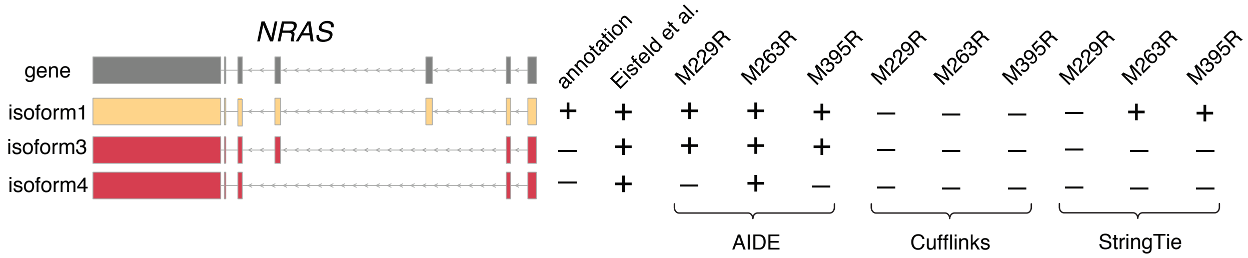


Figure 3.8: *NRAS* isoforms predicted by AIDE, Cufflinks, and StringTie. *NRAS* isoforms in the GENCODE annotation, reported by Eisfeld et al. [129], and discovered by AIDE, Cufflinks, or StringTie in three melanoma BRAF inhibitor resistant cell lines: M229R, M263R, and M395R.

the *NRAS* gene were experimentally identified using quantitative PCR and shown to have potential roles in cell proliferation and malignancy transformation [129]. Except for the annotated isoform 1, the other two isoforms identified by Eisfeld et al. have not been included in the GENCODE [75] or Ensembl [121] annotation (Figure 3.8). We applied both AIDE and the other two reconstruction methods to the RNA-seq data of melanoma cell lines [131, 132]. Our results showed that (1) the two novel *NRAS* isoforms, in addition to the annotated isoform 1, were identified by AIDE; (2) only isoform 1 was identified in two out of the three cell lines by StringTie; 3) none of them was identified by Cufflinks (Figure 3.8). These results again demonstrate the potential of AIDE as a powerful bioinformatics tool for isoform discovery from short-read sequencing data.

3.3.5 AIDE achieves the best consistency with long-read sequencing

We conducted another transcriptome-wide study to evaluate the isoform discovery methods by comparing their reconstructed isoforms (from the next-generation short RNA-seq reads) to those identified by the third-generation long-read sequencing technologies, including Pacific Biosciences (PacBio) [76] and Oxford Nanopore Technologies (ONT) [133]. Even though PacBio and ONT platforms have higher sequencing error rates and lower throughputs compared to the NGS technologies, they are able to generate much longer reads (1-100 kbp) to capture multiple splicing junctions [134]. Here we used the full-length transcripts identified from the PacBio or ONT sequencing data as a surrogate gold standard to evaluate the isoform discovery methods.

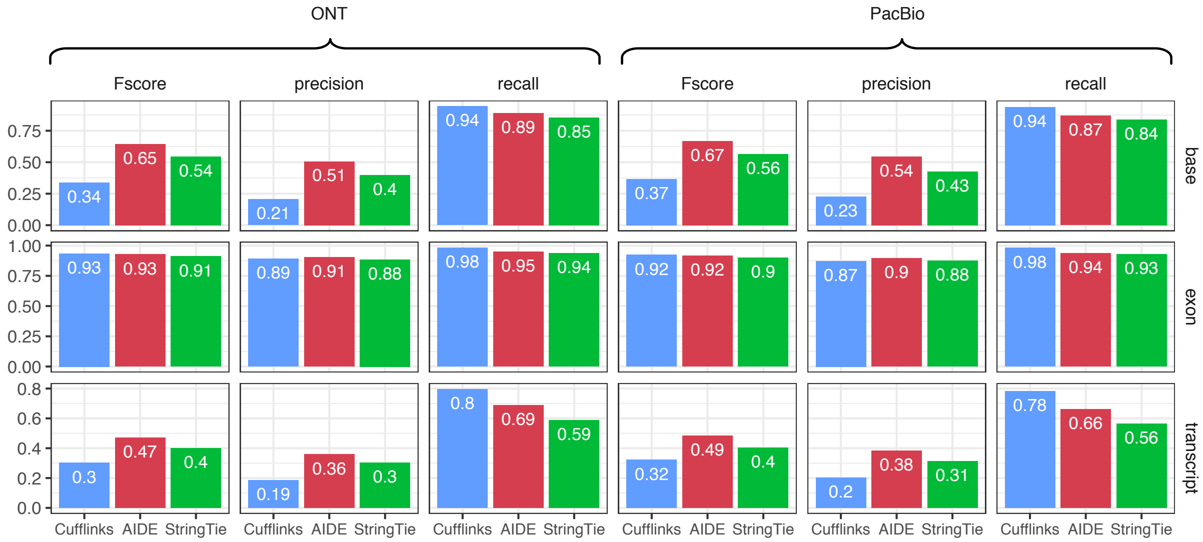


Figure 3.9: Evaluation of isoform discovery methods based on long-read technologies. **a**: The F score, precision, and recall of Cufflinks, AIDE, and StringTie calculated based on isoforms identified by ONT. **b**: The F score, precision, and recall of Cufflinks, AIDE, and StringTie calculated based on isoforms identified by PacBio.

We applied AIDE, Cufflinks, and StringTie to a next-generation RNA-seq sample of human ESCs, and compared their identified isoforms with those discovered from PacBio or ONT data generated from the same human ESC sample [134]. SLIDE requires its input RNA-seq reads to have the same mapped read length and is thus not applicable to this dataset. The comparison results based on ONT and PacBio are highly consistent: AIDE achieves the best precision and overall accuracy (F score) at both the base level and the transcript level (Figure 3.9). The fact that all the three methods have high accuracy at the exon level but much lower accuracy at the transcript level again indicate the difficulty of assembling exons into full-length isoforms based on short RNA-seq reads. Among all the three methods, AIDE is advantageous in achieving the best precision in full-length isoform discovery.

3.3.6 AIDE improves isoform abundance estimation

Since the structures and abundance of expressed isoforms are unobservable in real RNA-seq data, we also seek to evaluate the performance of different methods by comparing their estimated isoform expression levels in the FPKM unit with the NanoString counts, which could serve as benchmark data for isoform abundance when PCR validation is not available [56,72,104,135]. We expect an accurate isoform discovery method to discover a set of isoforms close to the expressed isoforms in an RNA-seq sample. If the identified isoforms are accurate, the subsequently estimated isoform abundance is more likely to be accurate and agree better with the NanoString counts.

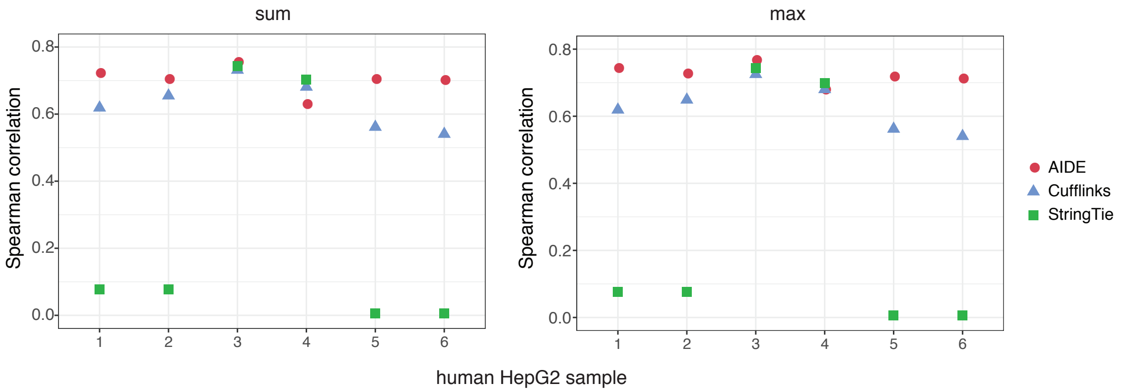


Figure 3.10: Spearman correlation coefficients between the estimated isoform expression and the benchmark NanoString counts. **a**: For every probe, the sum of the expression levels of its corresponding isoforms was used in the calculation. **b**: For every probe, the maximum of the expression levels of its corresponding isoforms was used in the calculation.

We therefore applied AIDE, Cufflinks, and StringTie to six samples of the human HepG2 immortalized cell line with both RNA-seq and NanoString data [72] (Table C.2). We supplied all the three methods with the GENCODE annotation (version 24) [75] for isoform discovery. SLIDE requires its input RNA-seq reads to have a unique read length after mapping and is thus not applicable to the six mapped HepG2 RNA-seq datasets. Since the NanoString nCounter technology is not designed for genome-wide quantification of RNA molecules, the

HepG2 NanoString datasets have measurements of 140 probes corresponding to 470 isoforms of 107 genes. We first found the isoforms compatible with every probe, and we then compared the sum or the maximum of the estimated abundance of these isoforms with the count of that probe. For each HepG2 sample, we calculated the Spearman correlation coefficient between the estimated isoform abundance ("sum" or "max") and the NanoString probe counts to evaluate the accuracy of each method (Figure 3.10). AIDE has the highest correlations in five out of the six samples, suggesting that AIDE achieves more accurate isoform discovery as well as better abundance estimation in this application. It is also worth noting that all three methods have achieved high correlation with the Nanostring counts for samples 3 and 4, since these two samples have the longest reads (100 bp) among all the samples.

3.4 Discussion

We propose a new method AIDE to improve the precision of isoform discovery and the accuracy of isoform quantification from the next-generation RNA-seq data, by selectively borrowing alternative splicing information from annotations. AIDE identifies isoforms in a stepwise manner while placing priority on the annotated isoforms, and it performs statistical testing to automatically determine what isoforms to retain. We demonstrate the efficiency and superiority of AIDE compared to three state-of-the-art methods, Cufflinks, SLIDE, and StringTie, on multiple synthetic and real RNA-seq datasets followed by an experimental validation through PCR-Sanger sequencing, and the results suggest that AIDE leads to more precise discovery of full-length RNA isoforms and more accurate isoform abundance estimation. In an evaluation based on the third-generation long-read RNA-seq data, AIDE also leads to the most consistent isoform discovery results than the other methods do.

In addition to reducing false discoveries, AIDE is also demonstrated to identify full-length RNA isoforms with biological functions in disease conditions. First, we assessed the biological relevance of the isoform *FGFR1-238*, which was only identified by AIDE, using a loss-of-function assay. We selected two breast cancer cell lines that had this isoform expressed and experimentally proved that cell proliferation was inhibited with this isoform being knocked down. Second, we applied both AIDE and Cufflinks to the RNA-seq data of melanoma cell lines. Only AIDE was able to detect two novel isoforms of the *NRAS* gene, which were reported to play a role in the drug resistance mechanism of *BRAF*-targeted therapy.

Even though long reads generated by PacBio and ONT have advantages over next-generation short RNA-seq reads for assembling full-length RNA isoforms, it remains necessary to improve computational methods for short-read-based isoform discovery. First, wide application of the long-read sequencing technologies is still hindered by their lower throughput, higher error rate, and higher cost per base [76]. Meanwhile, the short-read sequencing technology is still the mainstream assay for transcriptome profiling. Second, a huge number of next-generation RNA-seq datasets have been accumulated over the past decade. Considering that many biological or clinical samples used to generate those datasets are precious and no longer available for long-read sequencing, the existing short-read data constitute an invaluable resource for studying RNA mechanisms in these samples. Therefore, an accurate isoform discovery method will be indispensable for studying full-length isoforms from these data. Meanwhile, we also expect that with increased availability of long read data, we will be better equipped to compare and evaluate the reconstruction methods for short-read data.

To the best of our knowledge, AIDE is the first isoform discovery method that identifies isoforms by selectively leveraging information from annotations based on a statistically principled model selection approach. The stepwise likelihood ratio testing procedure in AIDE has multiple advantages. First, AIDE only selects isoforms that significantly contribute to the explanation of the observed reads, leading to more precise results and reduced false discoveries than those of existing methods. Second, the forward steps allow AIDE to start from and naturally give priority to the annotated isoforms, which have higher chances to be expressed. Meanwhile, the backward steps allow AIDE to adjust its previously selected isoforms given the newly added isoforms so that all the selected isoforms together better explain the observed reads. Third, the testing procedure in AIDE allows the users to adjust the conservatism and precision of the discovered isoforms according to their desired level of statistical significance. Because of these advantages, AIDE identifies fewer novel isoforms at a higher precision level than previous methods do, making it easier for biologists to experimentally validate the novel isoforms. In applications where the recall of isoform discovery is of great importance, users can increase the p-value threshold of AIDE.

Through the application of AIDE to multiple RNA-seq datasets, we demonstrate that selectively incorporating annotated splicing patterns, in addition to simply obtaining gene and exon boundaries from annotations, greatly helps isoform discovery. The stepwise selection procedure of AIDE also differentiates it from the methods that directly assume the existence of all the annotated isoforms. The application of AIDE has lead us to interesting observations that could benefit both method developers and data users. First, we find that a good annotation can help reduce the need for deep sequencing depths. AIDE has been shown to achieve good accuracy on datasets with low sequencing depths when supplied with accurately annotated isoforms, and its accuracy is comparable to that based on deeply sequenced datasets. Second, we find it more important for an annotation to have high purity than to have high completeness, in order to improve the isoform discovery accuracy of AIDE and the other methods we compared with in our study. Ideally, instead of using all the annotated isoforms, a better choice is to use a filtered set of annotated isoforms with high confidence. This requires annotated isoforms to have confidence scores, which unfortunately are unavailable in most annotations. Therefore, how to add confidence scores to annotated isoforms becomes an important future research question, and answering this question will help the computational prediction of novel isoforms.

When identifying differential splicing patterns between RNA-seq samples from different biological conditions, a well-established practice is to first estimate the isoform abundance in each sample, and then perform statistical testing to discover differentially expressed isoforms [136,137]. However, as we have demonstrated in both synthetic and real data studies, existing methods suffer from high risks of predicting false positive isoforms, i.e., estimating non-zero expression levels for unexpressed isoforms in a sample. Such false positive isoforms will severely reduce the accuracy of differential splicing analysis, leading to inaccurate comparison results between samples under two conditions, e.g., healthy and pathological samples. In contrast, AIDE's conservative manner in leveraging the existing annotations allows it to identify truly expressed isoforms at a greater precision and subsequently estimate isoform abundance with a higher accuracy. We expect that the application of AIDE will increase the accuracy of differential splicing analysis, lower the experimental validation costs, and lead to new biological discoveries at a higher confidence level.

The probabilistic model of AIDE is very flexible and can incorporate reads of varying lengths and generated by different platforms. The non-parametric approach to learning the read generating mechanism makes AIDE a data-driven method, and it does not depend on specific assumptions of the RNA-seq experiment protocols. Therefore, a natural extension of AIDE is to combine the short but more accurate reads from the next-generation technologies with the longer but more error-prone reads generated by new sequencing technologies such as PacBio [76] and Nanopore [78]. Joint modeling of the two types of reads has the potential to greatly improve the overall accuracy of isoform detection [138], since AIDE is shown to have better precision than existing methods, and longer RNA-seq reads capture more splicing junctions and can further improve the recall rate of AIDE. Aside from the stepwise selection procedure used by AIDE, another possible way to incorporate priority on the annotated isoforms in the probabilitic model is to add regularization terms only on the unannotated isoforms. However, this approach is less interpretable compared with AIDE, since the regularization terms lack direct statistical intepretations as the p-value threshold does. Moreover, this approach may lose control of the FDR when the annotation has a low purity. Another future extension of AIDE is to jointly consider multiple RNA-seq samples for more robust and accurate transcript reconstruction. It has been shown that it is often possible to improve the accuracy of isoform quantification by integrating the information in multiple RNA-seq samples [56, 69, 70]. Therefore, by extending AIDE to combine the consistent information from multiple technical or biological samples, it is likely to achieve better reconstruction accuracy, and enable researchers to integrate publicly available and new RNA-seq samples for transcriptome studies.

3.5 Acknowledgments

This chapter is based on my on-going work with Dr. Jessica Li, Dr. Hubing Shi, Ms. Shan Li, Dr. Xin Tong, and Dr. Ling Deng. I would like to thank Dr. Li and Dr. Shi for their supervision of this research. I also thank Ms. Shan Li and Dr. Shi for performing the validation experiments and for the valuable discussion. We especially thank Dr. Kin Fai Au and Dr. Yunhao Wang for providing their next- and third-generation human ESC RNA-seq data and processed RNA isoform information from the third-generation data. We would also like to thank Dr. Alexander Hoffmann for his insightful feedbacks on our studies.

CHAPTER 4

Introduction to Single-cell RNA Sequencing Analysis

4.1 Background

Bulk RNA-seq technologies have been widely used for transcriptome profiling to study transcriptional structures, splicing patterns, and gene and transcript expression levels [1]. However, it is important to account for cell-specific transcriptome landscapes in order to address biological questions such as cell heterogeneity and gene expression stochasticity [139]. Despite its popularity, bulk RNA-seq does not allow people to study cell-to-cell variation in terms of transcriptomic dynamics. In bulk RNA-seq, cellular heterogeneity cannot be addressed since signals of variably expressed genes would be averaged across cells. Fortunately, single-cell RNA sequencing (scRNA-seq) technologies are now emerging as a powerful tool to capture transcriptome-wide cell-to-cell variability (Figure 4.1) [88, 140, 141]. ScRNA-seq enables the quantification of intra-population heterogeneity at a much higher resolution, potentially revealing dynamics in heterogeneous cell populations and complex tissues [142].

The rapid development of scRNA-seq technologies offer unprecedented opportunities for investigating transcriptional mechanisms underlying biological and medical phenomena at the individual-cell resolution [143–145]. The scRNA-seq technologies have enabled researchers to investigate fundamental biomedical questions such as cellular composition of various tissues and cell types [146, 147], cell differentiation trajectories [148, 149], and spatial and temporal dynamics of single cells [150, 151]. Important discoveries have been made from scRNA-seq data and advanced our understanding of diseases such as neurological disorders [152, 153] and tumorigenesis [154, 155].

4.2 Currently available scRNA-seq technologies

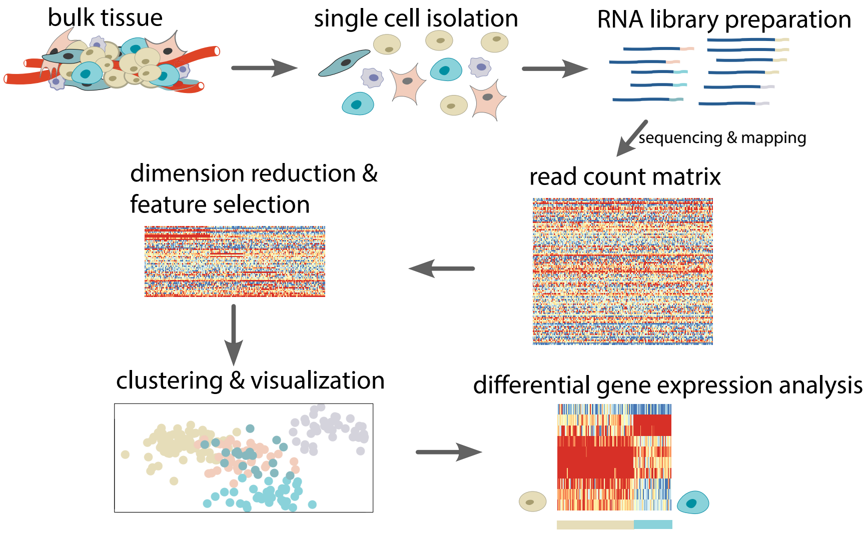


Figure 4.1: Workflow of a scRNA-seq analysis. A typical pipeline of scRNA-seq experiments and analysis involve the following major steps: isolation of single cells and RNA molecules from bulk tissue, library preparation and sequencing, read mapping to obtain read count matrix (rows for genes and columns for cells), dimensionality reduction and feature selection, clustering analysis and visualization, downstream investigation such as DE analysis.

Since the first scRNA-seq study was published in 2009 [156], more than twenty scRNA-seq experimental protocols have been developed [157–162]. These scRNA-seq technologies differ in one or more aspects including cell isolation, cell lysis, reverse transcription, transcript coverage, and the availability of unique molecular identifiers (UMIs) [163]. Among these different aspects, one important factor that needs to be considered for downstream analysis is transcript coverage. Tag-based protocols, such as Drop-seq [159] and Seq-Well [161], allow the integration of UMIs [164] to detect and quantify unique transcripts, but they can only capture and sequence the 3'-end or 5'-end of the RNA transcripts. In contrast, full-length protocols, such as Smart-seq2 [157] and Fluidigm C1, are able to capture full-length transcripts and allow the addition of the External RNA Control Consortium (ERCC) spike-ins [162, 165].

Compared to the tag-based protocols, full-length protocols are advantageous for transcript-

and exon- level studies in order to understand single-cell alternative splicing [166] and isoform usage [167]. For gene-level analyses, the tag-based protocols are usually designed to obtain a broad but shallow view of gene expression across many cells, while the full-length protocols provide a deeper and more accurate account of gene expression in fewer cells. In gene-level analyses, the first step after read mapping is often to summarize a gene expression matrix where rows represent genes and columns represent cells. Each element in the matrix summarizes the number of RNA-seq reads (for full-length protocols) or the number of unique UMIs (for tag-based protocols) mapped to a gene in an individual cell. The remaining of this chapter focuses on gene-level analyses of scRNA-seq data.

One important characteristic of scRNA-seq gene expression data is the *dropout* phenomenon where a gene is expressed in one cell but its RNA transcripts are not detected by the sequencing experiments [168]. Usually these events occur due to the low amounts of mRNA in individual cells, and thus a truly expressed transcript may not be detected during sequencing in some cells. If one gene's transcripts in a cell are not detected in a scRNA-seq experiment, its expression will be represented by a false zero count in the resulting gene expression matrix. This characteristic of scRNA-seq is shown to be protocol-dependent. For instance, the number of cells that can be analyzed with one chip is usually no more than a few hundreds on the Fluidigm C1 platform, with around 1-2 million reads per cell. On the other hand, protocols based on droplet microfluidics can parallelly profile more than $10,000$ cells, but with only 100-200k reads per cell [169]. Hence, there is usually a much higher dropout rate in scRNA-seq data generated by the droplet microfluidics than the Fluidigm C1 platform. New droplet-based protocols, such as inDrop [158] or 10x Genomics [160], have improved molecular dectection rates but still have relatively low sensitivity compared with the microfluidics technologies, without accounting for sequencing depths [170]. Statistical methods for scRNA-seq gene expression data need to take the dropout issue into consideration, and otherwise they may present varying efficacy when applied to data from different protocols.

4.3 Dimensionality reduction and feature selection

Since scRNA-seq experiments may involve thousands of genes and hundreds to millions of cells, scRNA-seq gene expression matrices are with a high dimensionality. Therefore, dimensionality reduction or feature selection approaches are usually applied to scRNA-seq data before downstream analysis to reduce the noises, speed up calculations, and assist visualization (Figure 4.1) [163, 171].

Dimensionality reduction approaches are widely used in single-cell studies to provide visualization in exploratory analysis and to help identify important patterns in single-cell transcriptomes. Their main goal is to project individual cells from the high-dimensional gene expression measurement space to a low-dimensional latent space. Principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (t-SNE) [14] are the two most frequently used methods used on single-cell gene expression data. PCA is a linear projection method, and it is often applied to visualize major cell clusters [172] or development trajectories [173]. Pseudotime inference methods such as TSCAN [174] and Waterfall [175] also rely on PCA to achieve dimensionality reduction before inferring the underlying temporal order of the individual cells. On the other hand, t-SNE is a non-linear approach that tends to well preserve local clustering structures, and it has been widely applied to help identify patterns in complex cell populations [159, 160].

More recent dimensionality reduction methods have been specifically designed for scRNA-seq data. For example, ZIFA, representing zero-inflated factor analysis, is a dimensionality reduction method that explicitly accounts for the presence of dropouts [176]. ZIFA uses a latent variable model and extends the factor analysis framework with an additional zero-inflation modulation layer. In the ZIFA model, the dropout probability (i.e., the probability that a gene's transcripts not being detected) is assumed to be a function of a gene's latent expression level, $p_0 = \exp(-\lambda x^2)$, where λ is the exponential decay parameter and x denotes the gene expression level. This assumption is based on the empirical observation that the dropout rate of a gene depends on the expected expression level of that gene in the cell population [176]. Another matrix factorization method, ZINB-WaVE, uses a zero-inflated negative binomial (ZINB) model to extract low-dimensional signals from scRNA-seq data [177]. ZINB-WaVE has the flexibility to include both gene- and cell-level covariates to adjust for batch effects or gene sequence effects on read counts. It assumes that the observed read count is a random variable following a ZINB distribution, while both the mean parameter and the zero proportion depend on the gene- and cell- level covariates through regression models.

In addition to dimensionality reduction, feature selection is another way to circumvent very high-dimensional problems. Researchers use this approach to identify a set of genes that are most relevant to the underlying structures of single cells, e.g., genes that are highly variable across cell subpopulations or genes that specifically expressed in one or a few cell subtypes. People have used gene expression variance, coefficient of variance, or Gini index to rank genes for this purpose [178–180]. It was also shown that the Gini index is a more appropriate criterion than the coefficient of variance for selecting rare cell-type-specific genes [180]. The dimensionality reduction and feature selection approaches are usually combined in practice before downstream analyses (e.g., clustering) are performed [181]. However, as there are no unified guidelines for choosing parameters, such as the number of gene features or the number of components to retain for the projected space, inconsistency may arise from different analysis pipelines even if they use the same statistical models.

4.4 Identification of cell subpopulations

A key goal of many scRNA-seq studies is to identify cellular subpopulations (e.g., cell subtypes and cell states) that cannot be defined by bulk data [143, 163, 171]. Especially, single cell analysis provides an opportunity to systematically define cell subpopulations as opposed to using criteria including marker protein expression or morphology [143]. Single cell studies that incorporate prior biological knowledge often use the expression levels of known marker genes to characterize major cell types. For example, the *GAD1* and *GAD2* genes are used as markers of amacrine cells [159], while the *CD3D* and *NKG7* genes are respectively used as markers of T cells and natural killer cells [160]. However, many cell types or subtypes have few well-established marker genes, and the limited number of marker genes may not fully capture the diverse facets of cell identities. Therefore, most studies rely on unsupervised clustering methods to identify cell subpopulations and putative cell types (Figure 4.1). This task needs to be carried out after careful normalization to remove technical variation such as batch effects in scRNA-seq data [182].

Clustering methods developed for scRNA-seq data are usually categorized by the statistical models on which they are based. The first type of scRNA-seq clustering method is based on the k-means algorithm. For example, the method SC3 aims to improve the accuracy and robustness of cell clustering through a consensus approach [183]. SC3 applies the k-means algorithm to the first d eigenvectors of the cell-cell distance matrices, and it obtains multiple clustering solutions by varying the value of d. It finally combines all clustering solutions into a consensus matrix, summarizing how often each pair of cells is assigned to the same cluster. Another method RaceID2 applies the k-medoids algorithm to a correlation-based distance matrix, and it infers the cluster number based on the saturation of the average within-cluster variation [184].

The second type of scRNA-seq clustering method is based on hierarchical clustering. CIDR is the first clustering method that incorporates imputation of dropout values, but the imputed expression value of a particular gene in a cell changes each time when the cell is paired up with a different cell [185]. The hierarchical clustering algorithm is then applied on the cell-cell distances after dimensionality reduction. In a study of mouse cortical cells, Tasic et al. first selected genes that were differentially expressed between preliminary clusters, and then applied the hierarchical clustering algorithm using only the DE genes [186].

The third type of scRNA-seq clustering method is based on graphs. The SNN-Cliq method [187] uses the ranking of cells' common k-nearest-neighbors (KNNs) to construct a graph with cells as nodes. It then identifies cell clusters in the graph by finding the maximal cliques, which are fully connected subgraphs not contained in a larger clique. Another two methods, PhenoGraph [188] and Seurat [181], also construct a KNN graph before cell clustering. They both use the Louvain method [189] to uncover cell clusters from the KNN graphs.

The current practice of single cell clustering analysis is usually interactive, as there lacks unified criteria to identify cell outliers, determine cluster numbers, and annotate cell clusters with biological information. Researchers often rely on visualization of cell cluster similarities and marker gene expression levels to evaluate the results based on existing knowledge. The clustering analysis may be repeated with different parameters until a reasonable result is obtained. However, with on-going efforts including the Human Cell Atlas [190], databases of known cell types and their gene expression characteristics may become available in the near future.

4.5 Differential gene expression analysis

After identifying cell subpopulations using statistical and computational methods, it is necessary to analyze the gene expression characteristics and differences among these subpopulations, so as to define cell identities. Similar as for bulk RNA-seq data, a main approach to comparing two cell subpopulations is to find DE genes through principled statistical tests. Since scRNA-seq data are highly heterogeneous and sparse due to the dropout events, commonly used Poisson or Negative Binomial models for bulk RNA-seq data cannot be directly applied on scRNA-seq data. Therefore, new strategies and methods have been developed for performing differential gene expression analysis on single-cell data [191].

We first introduce the basic framework for testing the mean difference of gene expression in different cell subpopulations (Figure 4.1). We consider two cell subpopulations $k = 1, 2$, each with J_k cells. $Y_{k,ij}$ denotes the expression level of gene i in the jth cell of subpopulation k. Depending on the specific methods, the expression levels could be measured as read counts (e.g., in SCDE [168]) or log-transformed TPM values (e.g., in MAST [192]). A common assumption is that

$$Y_{k,ij} \sim \lambda_{ki} f_0(\boldsymbol{\phi}_{0,ki}) + (1 - \lambda_{ki}) f_1(\theta_{ki}, \boldsymbol{\phi}_{1,ki}) \tag{4.1}$$

where λ_{ki} is the probability of gene i being a dropout in subpopulation k. If gene i is a dropout, we assume that its expression level follows distribution f_0 with one or more parameters $\boldsymbol{\phi}_{0,ki}$; if gene i is not a dropout, we assume that its expression level follows distribution f_1 with true expression level θ_{ki} and other necessary parameter(s) $\boldsymbol{\phi}_{1,ki}$. This assumption uses a dropout component (f_0) to account for the zero or very low read counts resulting from the dropout events. For example, SCDE specifies f_0 as a Poisson distribution and f_1 as a Negative Binomial distribution. Another scRNA-seq DE method MAST assumes f_0 to be δ_0, a point mass at 0, and f_1 to be a Normal distribution. The DGE analysis is then carried out by testing

$$H_0 : \theta_{1i} = \theta_{2i} \text{ vs. } H_1 : \theta_{1i} \neq \theta_{2i} \tag{4.2}$$

for each gene i. As in bulk DE analysis (Chapter 1.3.1), shrinkage estimation of gene expression variance can be incorporated into the framework to borrow information across genes and increase the robustness of estimation [192]. To apply the previously developed DE methods on scRNA-seq data, Van den Berge et al. [193] proposed an approach to estimate observation weights from a zero inflated NB model before the statistical testing step. The approach can be combined with DESeq2 [30] and edgeR [28] to detect DE genes in single cell data.

To better account for the cellular heterogeneity in terms of gene expression, new DE methods have been proposed to test for differences beyond traditional differential mean expression. For example, scDD [194] aims to identify differential distributions in terms of differential mean expression, differential proportions of cells within each distribution component, and differential modality. DEsingle uses a zero-inflated NB distribution to fit scRNA-seq data in each cell subpopulation and tests if there is difference in the zero component or the NB component of the distribution, or in both [195].

4.6 Discussion

Aside from differential gene expression analysis, analyses of pseudotime order and branching, gene co-regulation, and spatial inference are also common goals involved in single-cell genomics. The statistical models for scRNA-seq data will evolve as new technologies are developed. Currently, single-cell genomics still face great statistical and computational challenges given the extensive technical noises and the high dimensionality of scRNA-seq data. In Chapters 5 and 6, we will discuss two statistical challenges in detail to improve the analysis of scRNA-seq data. In Chapter 5, we discuss the imputation of dropout values to assist biological discovery based on single-cell gene expression. In Chapter 6, we discuss the experimental design of scRNA-seq experiments to achieve a balanced trade-off between exploring the depth or breadth of transcriptome information in single cells.

CHAPTER 5

Accurate Imputation for Single-cell RNA Sequencing Data

5.1 Introduction

The emerging scRNA-seq technologies enable the investigation of transcriptomic landscapes at single-cell resolution. However, ScRNA-seq data analysis is complicated by excess zero counts, the so-called dropouts due to low amounts of mRNA sequenced within individual cells [168]. This characteristic of scRNA-seq is shown to be protocol-dependent. For example, the number of cells that can be analyzed with one chip is usually no more than a few hundreds on the Fluidigm C1 platform, with around 1-2 million reads per cell. On the other hand, protocols based on droplet microfluidics can parallelly profile more than $10,000$ cells, but with only 100-200k reads per cell [169]. Hence, there is usually a much higher dropout rate in scRNA-seq data generated by the droplet microfluidics than the Fluidigm C1 platform. Statistical methods developed for scRNA-seq data need to take the dropout issue into consideration, in order to have sufficient efficacy when applied to data from different protocols.

Methods for analyzing scRNA-seq data have been developed from different perspectives, such as clustering, cell type identification, and dimensionality reduction. Some of these methods address the dropout events by implicit imputation while others do not. CIDR is the first clustering method that incorporates imputation of dropout values, but the imputed expression value of a particular gene in a cell changes each time when the cell is paired with a different cell [185]. The pairwise distances between every two cells are later used for clustering. Seurat is a computational strategy for spatial reconstruction of cells from single-cell gene expression data [181]. It infers the spatial origins of individual cells from the cell expression profiles and a spatial reference map of landmark genes. It also includes an imputation step to infer the expression of landmark genes based on highly variable genes. ZIFA is a dimensionality reduction model specifically designed for zero-inflated single-cell gene expression analysis [176]. The model is built upon an empirical observation: dropout rate of a gene depends on its mean expression level in the population and, and ZIFA accounts for dropout events using factor analysis.

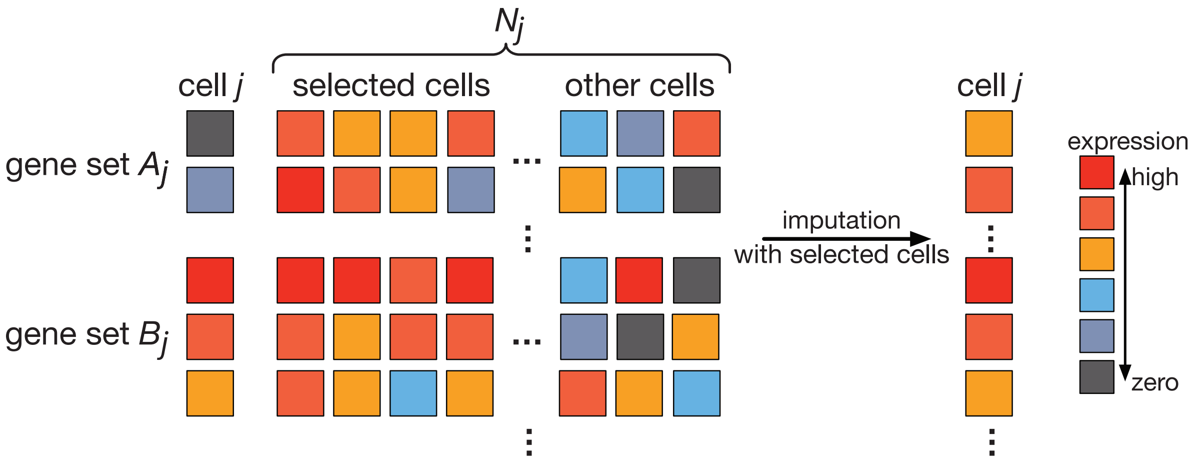


Figure 5.1: A toy example illustrating the workflow in the imputation step of scImpute, described in Equations (5.5)-(5.6). The scImpute method first learns each gene's dropout probability in each cell by fitting a mixture model. Next, scImpute imputes the highly probable dropout values in cell j (gene set A_j) by borrowing information of the same gene in other similar cells, which are selected based on gene set B_j (not severely affected by dropout events).

Since most downstream analyses on scRNA-seq, such as differential gene expression analysis, identification of cell-type-specific genes, and reconstruction of differentiation trajectory, rely on the accuracy of gene expression measurements, it is important to correct the false zero expression due to dropout events in scRNA-seq data by model-based imputation methods. To our knowledge, MAGIC is the first available method for explicit and genome-wide imputation of single-cell gene expression profiles [196]. MAGIC imputes missing expression values by sharing information across similar cells, based on the idea of heat diffusion. A key step in this method is to create a Markov transition matrix, constructed by normalizing the similarity matrix of single cells. In the imputation of a single cell, the weights of the other cells are determined through the transition matrix. Another imputation method, SAVER [197], borrows information across genes using a Bayesian approach to estimate (unobserved) true expression levels of genes. Both MAGIC and SAVER would alter all gene expression levels including those not affected by dropouts, and this would potentially introduce new biases into the data and possibly eliminate biologically meaningful variation. It is also inappropriate to treat all zero counts as missing values, since some of them may reflect true biological non-expression. Therefore, we propose a new imputation method for scRNA-seq data, scImpute, to simultaneously determine which values are affected by dropout events in data and perform imputation only on dropout entries. To achieve this goal, scImpute first learns each gene's dropout probability in each cell based on a mixture model. Next, scImpute imputes the highly probable dropout values in a cell by borrowing information of the same gene in other similar cells, which are selected based on the genes unlikely affected by dropout events (Figure 5.1).

5.2 Methods

5.2.1 Data processing and normalization

The input of our method is a count matrix $\boldsymbol{X}^C$ with rows representing genes and columns representing cells, and our eventual goal is to construct an imputed count matrix with the same dimensions. We start by normalizing the count matrix by the library size of each sample (cell) so that all samples have one million reads. Denote the normalized matrix by $\boldsymbol{X}^N$, we then make a matrix $\boldsymbol{X}$ by taking $\log 10$ transformation with a pseudo count 1.01:

$$X_{ij} = \log_{10}(X_{ij}^N + 1.01); \quad i = 1, 2, ..., I,\ j = 1, 2, ..., J, \tag{5.1}$$

where I is the number of genes and J is the number of cells. The pseudo count is added to avoid infinite values in parameter estimation. The advantage of the logrithmic transformation is to prevent a few large observations from being extremely influential. In addition, the transformed values allow for greater flexibility of the modeling.

5.2.2 Dectection of cell subpopulations and outliers

Since scImpute borrows information of the same gene from similar cells to impute the dropout values, a critical step is to first determine which cells are from the same subpopulation. Due to excess zero counts in scRNA-seq data, it is difficult to accurately cluster cells into true cell types using an unsupervised approach. Hence, the goal of this step is to find a candidate pool of "neighbors" for each cell. In a subsequent imputation step, scImpute will select similar cells from the candidate neighbors. Suppose that scImpute clusters the cells in a dataset into K subpopulations in this step. For each cell, its candidate neighbors are the other cells in the same cluster.

(1) PCA is performed on the matrix $\boldsymbol{X}$ for dimensionality reduction and the resulting matrix is denoted as $\boldsymbol{Z}$, where columns representing cells and rows representing PCs. The purpose of dimensionality reduction is to reduce the impact of large portions of dropout values. The PCs are selected such that at least 60% of the variance in data could be explained.

(2) Based on the PCA-transformed data $\boldsymbol{Z}$, the distance matrix $\boldsymbol{D}_{J\times J}$ between the cells could be calculated. For each cell j, we denote its distance to the nearest neighbor as l_j. For the set $\boldsymbol{L} = \{l_1, ..., l_J\}$, we denote its first quartile as Q_1, and third quartile as Q_3. The outlier cells are those cells which do not have any close neighbors:

$$O = \{j : l_j > Q_3 + 1.5(Q_3 - Q_1)\}. \tag{5.2}$$

For each outlier cell, we set its candidate neighbor set $N_j = \emptyset$. The outlier cells could be a result of experimental/technical errors, but they may also represent real biological variation such as rare cell types. We would not impute the gene expression values in the outlier cells, nor use them to impute gene expression values in other cells.

(3) The non-outlier cells $\{1, ..., J\}\backslash O$ are clustered into K groups by spectral clustering [198]. We denote $g_j = k$ if cell j is assigned to cluster k ($k = 1, ..., K$). Hence, cell j has the

candidate neighbor set $N_j = \{j' : g_{j'} = g_j, j' \neq j\}$.

5.2.3 Identification of dropout values

Once we obtain the transformed gene expression matrix $\boldsymbol{X}$ and the candidate neighbors of each cell j (N_j), the next step is to infer which genes are affected by the dropout events in which cells. We construct a statistical model to systematically determine whether an expression value comes from a dropout event or not. With the existence of dropout events, most genes have a bimodal expression pattern across similar cells, and that pattern can be described by a mixture model of two components (Figure 5.2). The first component is a Gamma distribution used to account for the dropouts, while the second component is a Normal distribution to represent the actual gene expression levels. For each gene, the proportions and parameters of the two components could be different in various cell types, so we construct a separate mixture model for each cell subpopulation.

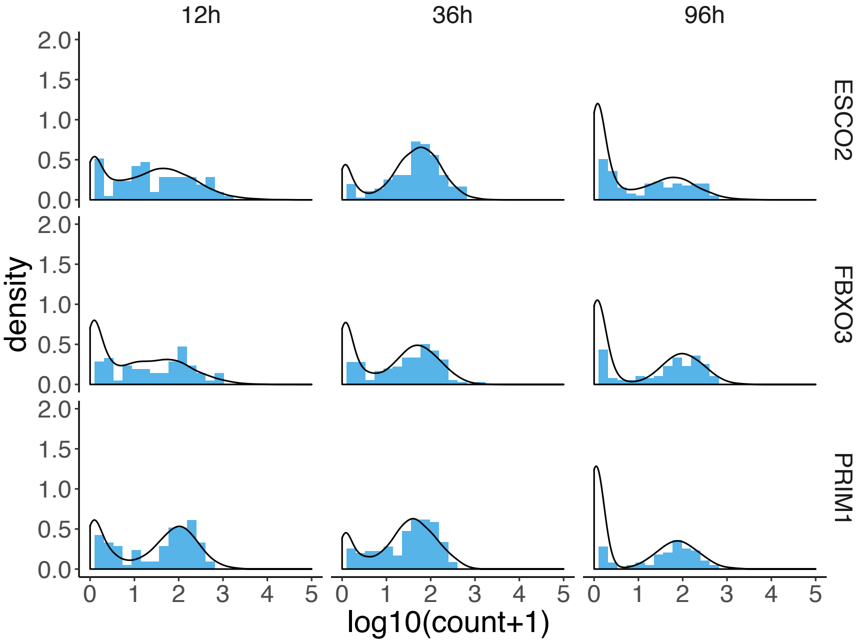


Figure 5.2: Three example genes (*ESCO2*, *FBXO3*, and *PRIM1*) for comparison of observed and fitted expression distribution in three different cell types. The results are based on the human ESC scRNA-seq data [173]. Blue histograms represent observed distributions, and black lines represent fitted distributions by scImpute.

For each gene i, its expression in cell subpopulation k is modeled as a random variable $X_i^{(k)}$ with density function

$$f_{X_i^{(k)}}(x) = \lambda_i^{(k)} \mathrm{Gamma}\left(x; \alpha_i^{(k)}, \beta_i^{(k)}\right) + \left(1 - \lambda_i^{(k)}\right) \mathrm{Normal}\left(x; \mu_i^{(k)}, \sigma_i^{(k)}\right) , \tag{5.3}$$

where $\lambda_i^{(k)}$ is gene i's *dropout rate* in cell subpopulation k, $\alpha_i^{(k)}, \beta_i^{(k)}$ are the shape and rate parameters of Gamma distribution, and $\mu_i^{(k)}, \sigma_i^{(k)}$ are the mean and standard deviation of Normal distribution. The intuition behind this mixture model is that if a gene has high expression and low variation in most cells within a cell subpopulation, a zero count is more likely to be a dropout value; on the other hand, if a gene has constantly low or medium expression with high variation, then a zero count may reflect real biological variability. An advantage of this model is that it does not assume an empirical relationship between dropout rates and mean gene expression levels, as Kharchenko et al. [168] did, allowing more flexibility in model estimation. The parameters in the mixture model can be estimated by the EM algorithm and we denote their estimates as $\hat{\lambda}_i^{(k)}$, $\hat{\alpha}_i^{(k)}$, $\hat{\beta}_i^{(k)}$, $\hat{\mu}_i^{(k)}$, and $\hat{\sigma}_i^{(k)}$. It follows that the *dropout probability* of gene i in cell j, which belongs to subpopulation k, can be estimated as

$$d_{ij} = \frac{\hat{\lambda}_i^{(k)} \mathrm{Gamma}\left(X_{ij}; \hat{\alpha}_i^{(k)}, \hat{\beta}_i^{(k)}\right)}{\hat{\lambda}_i^{(k)} \mathrm{Gamma}\left(X_{ij}; \hat{\alpha}_i^{(k)}, \hat{\beta}_i^{(k)}\right) + \left(1 - \hat{\lambda}_i^{(k)}\right) \mathrm{Normal}\left(X_{ij}; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{(k)}\right)} . \tag{5.4}$$

In summary, each gene i has an overall *dropout rate* $\hat{\lambda}_i^{(k)}$ in cell subpopulation k, which does not depend on individual cells within the subpopulation. Gene i also has *dropout probabilities* d_{ij}'s, which may vary among different cells. Ghazanfar et al. [199] also used a Gamma-Normal mixture model to analyze scRNA-seq data but only applied it to categorize non-zero expression into low expression or high expression values.

5.2.4 Imputation of dropout values

We impute the gene expression levels cell by cell. For each cell j, we select a gene set A_j in need of imputation based on the genes' dropout probabilities in cell j: $A_j = \{i : d_{ij} \geq t\}$, where t is a threshold on dropout probabilities. We also have a gene set $B_j = \{i : d_{ij} < t\}$ that have accurate gene expression with high confidence and do not need imputation. We first learn cells' similarities through gene set B_j. Then we impute the expression of genes in set A_j by borrowing information from the same gene's expression in other similar cells learned from B_j. Figure B.10c give some real data examples of dropout probabilities in cells, showing that it is reasonable to divide genes into two sets. To learn the cells similar to cell j based on gene set B_j, we use the non-negative least squares (NNLS) regression:

$$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}^{(j)}}{\arg\min} \, ||\mathbf{X}_{B_j,j} - \mathbf{X}_{B_j,N_j}\boldsymbol{\beta}^{(j)}||_2^2, \text{ subject to } \boldsymbol{\beta}^{(j)} \geq \mathbf{0}\,, \tag{5.5}$$

where N_j represents the indices of cells that are candidate neighbors of cell j. The response $\boldsymbol{X}_{B_j,j}$ is a vector representing the B_j rows in the j-th column of $\boldsymbol{X}$, the design matrix $\boldsymbol{X}_{B_j,N_j}$ is a sub-matrix of $\boldsymbol{X}$ with dimensions $|B_j| \times |N_j|$, and the coefficients $\boldsymbol{\beta}^{(j)}$ is a vector of length $|N_j|$. It is worth noting that NNLS itself has the property of leading to a sparse estimate $\hat{\boldsymbol{\beta}}^{(j)}$, whose components may have exact zeros [200], so NNLS can be used to select similar cells of cell j from its neighbors N_j. Finally, the estimated coefficients $\hat{\boldsymbol{\beta}}^{(j)}$ from the set B_j are used to impute the expression of gene set A_j in cell j:

$$\hat{X}_{ij} = \begin{cases} X_{ij}, & i \in B_j, \\ X_{i,N_j}\hat{\boldsymbol{\beta}}^{(j)}, & i \in A_j. \end{cases} \tag{5.6}$$

We construct a separate regression model for each cell to impute the expression of genes with high dropout probabilities. This method simultaneously determines the values that need imputation, and would not introduce bias to the high expressions of accurately measured genes. Since the identified dropouts are corrected by scImpute, the proportion of zero expression is reduced in the imputed data. However, scImpute does not inflate all the zero expressions, and some genes remain to have bimodal distributions after the imputation. Therefore, scImpute takes a relatively conservative approach to impute dropouts, attempting to avoid introducing biases and retain the stochasticity of gene expression.

The application of scImpute involves two parameters. The first parameter is K, which determines the number of initial clusters to help identify candidate neighbors of each cell. The imputation results do not heavily rely on the choice of K since scImpute uses a model-based method to select similar cells in a later stage. However, setting K to a value close to the true number of cell subpopulations can assist the selection of similar cells. The second parameter is a threshold t, and the imputation is only applied to the genes with dropout probabilities larger than t in a cell to avoid over-imputation. The sensitivity analysis based on the mouse embryo data [201] suggests that scImpute is robust to varying parameter values (Figure B.10a-b). Especially, the choice of parameter t only affects a minute fraction of genes (Figure B.10c).

5.2.5 Generation of simulated scRNA-seq data

We suppose there are three cell types c_1, c_2, and c_3, each with 50 cells, and there are $20,000$ genes in total. In the gene population, only 810 genes are truly differentially expressed, with one third having higher expression in each cell type respectively. We directly generate genes' $\log 10$-transformed read counts as expression values. First, mean expression levels of the $20,000$ genes are randomly drawn from a Normal distribution with mean 1.8 and standard deviation 0.5. Similarly, standard deviations of gene expression are randomly drawn from a Normal distribution with mean 0.6 and standard deviation 0.1. These parameters are estimated from the real scRNA-seq data of mouse embryo cells [201]. Second, we randomly draw 270 genes and shift their mean expression in cell type c_1 by multiplying it with an integer randomly sampled from $\{2, 3, \ldots, 10\}$; we also create 270 highly expression genes for each of cell types c_2 and c_3 in the same way. Next, the expression values of each gene in the 150 cells are simulated from Normal distributions defined by the mean and standard deviation parameters obtained in the first two steps. We refer to the resulting gene expression data as the *complete data*. Finally, we suppose the dropout rate of each gene follows a double exponential function $\exp(-0.1 \times \text{mean expression}^2)$, as assumed in [176]. Zero values are then introduced into the simulated data for each gene based on a Bernoulli distribution defined by the dropout rate of the gene, resulting in a gene expression matrix with excess zeros and in need of imputation. We refer to the gene expression data after introducing zero values as the *raw data*. This generation process of gene expression values does not directly follow the mixture model used in scImpute, so that we use this simulation to investigate the efficacy and robustness of scImpute in a fair way.

5.3 Results

5.3.1 scImpute recovers gene expression affected by dropouts

A key reason for performing imputation on scRNA-seq data is to recover biologically meaningful transcriptome dynamics in single cells so that we can more accurately determine cell identity and identify DE genes among different cell types. We first use three studies to illustrate scImpute's efficacy in imputing gene expressions. All the imputation results are obtained without using true cell type information unless otherwise noted.

First, we show that scImpute recovers the true expression of the ERCC spike-in transcripts [202], especially low abundance transcripts that are impacted by dropout events. The ERCC spike-ins are synthesized RNA molecules with known concentrations, which serve as gold standards of true expression levels, so the read counts can be compared with the true expression for accuracy evaluation. The dataset we used contains 3005 cells from the mouse somatosensory cortex region [203]. After imputation, the median correlation among the 57 transcripts' read counts and their true concentration increases from 0.92 to 0.95, and the minimum correlation increases from 0.81 to 0.89. The read counts and true concentrations also present a stronger linear relationship in each single cell (Figure 5.3).

Second, we show that scImpute correctly imputes the dropout values of 892 annotated cell-cycle genes in 182 ESCs that have been staged for cell-cycle phases (G1, S, and G2M) [149]. These genes are known to modulate the cell cycle and are expected to have non-zero expression during different stages of the cell cycle. Before imputation, 22.5% raw counts of the cell-cycle genes are zeros, which are highly likely due to dropouts. After imputation, most of the dropout values are corrected, and true dynamics of these genes in the cell cycle are revealed (Figure B.11). The imputed counts also better represents the true biological variation in these cell-cycle genes (Figure 5.4).

Third, we use a simulation study to illustrate the efficacy of scImpute in enhancing the identification of cell types. We simulated expression data of three cell types c_1, c_2, and c_3, each with 50 cells, and 810 among $20,000$ genes are truly differentially expressed (Chapter

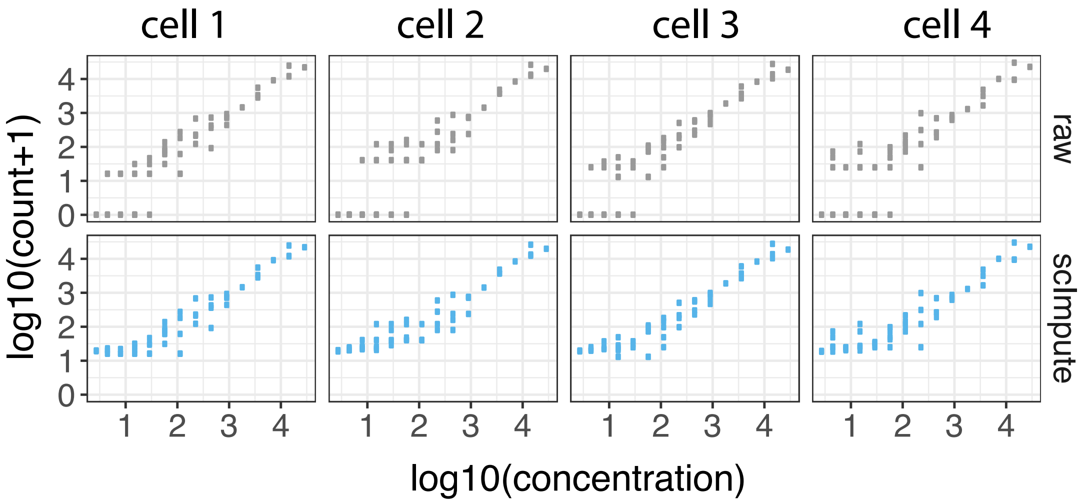


Figure 5.3: The ERCC spike-ins' log 10(count+1) and log 10(concentration) in four randomly selected mouse cortex cells. Imputed data presents a better linear relationship between the true concentration and the observed counts.

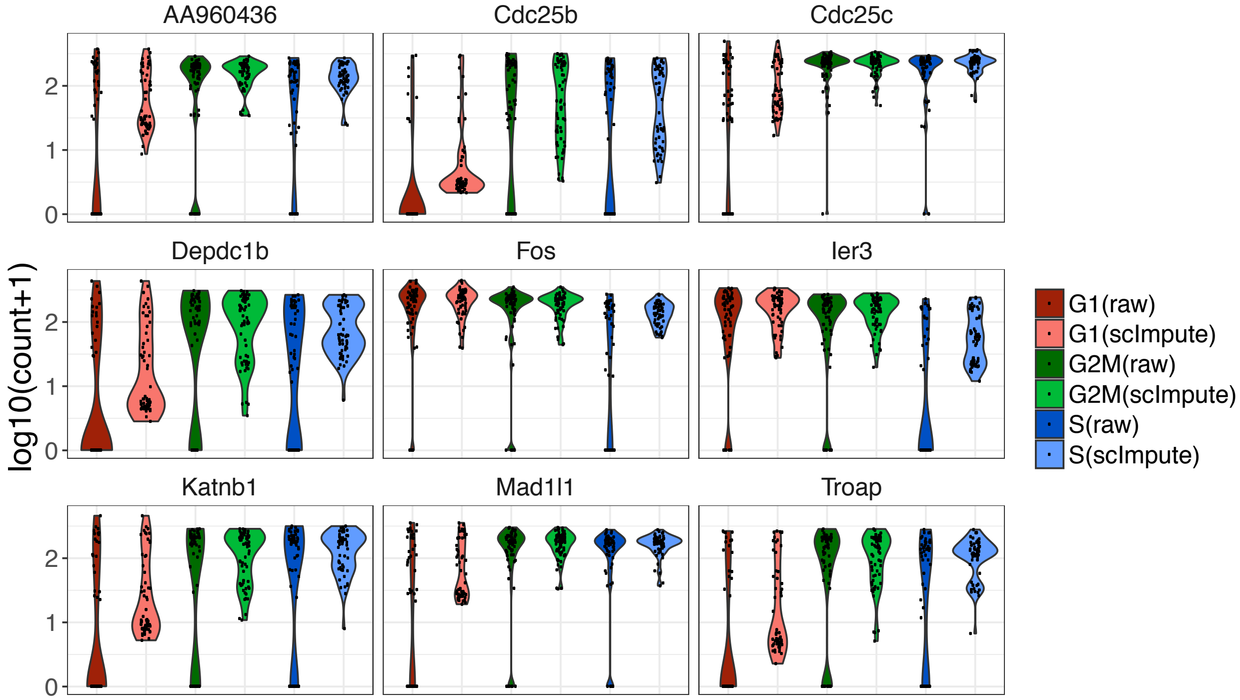


Figure 5.4: Violin plots showing the log 10(count+1) of nine cell cycle genes in the three phases (G1, S, and G2M). The scImpute method has corrected the dropout values of cell cycle genes.

5.2.5). Even though the three cell types are clearly distinguishable after we applied PCA to the complete data, they become less separated in the raw data with dropout events. The within-cluster sum-of-squares calculated based on the first two PCs increases from 94 in complete data to 2646 in raw data. However, the relationships among the 150 cells are clarified after we applied scImpute. The other two methods MAGIC and SAVER are also able to distinguish the three cell types, but MAGIC introduces artificial signals that largely alter the data and thus the PCA result, while SAVER only slightly improves the clustering result over that of the raw data (Figure 5.5). In addition, the dropout events obscure the differential pattern and thus increase the difficulty of detecting DE genes. The imputed data by scImpute leads to a clearer comparison between the up-regulated genes in different cell types, while the imputed data by MAGIC and SAVER fail to recover this pattern (Figure 5.5). We also assessed how the prevalence of dropout values influences the performance of scImpute. As expected, the DE analysis based on the imputed data has increased accuracy as the dropout proportion decreases. Yet scImpute still achieves $>$ 80.0% area under the curve even when the proportion of zero count in raw data is 75.0% (Figure B.12).

5.3.2 scImpute improves the identification of cell subpopulations

To illustrate scImpute's capacity in aiding the identification of cell subtypes or cell subpopulations, we applied our method to two real scRNA-seq datasets. The first one is a smaller dataset of mouse preimplantation embryos [201]. It contains RNA-seq profiles of 268 single cells from 10 developmental stages. Partly due to dropout events, 70.0% of read counts in the raw count matrix are zeros. To illustrate the dropout phenomenon, we plot the $\log 10$-transformed read counts of two 16-cell stage cells as an example in Figure 5.6. Even though the two cells come from the same stage, many expressed genes have zero counts in only one cell. This problem is alleviated in the imputed data by scImpute, and the Pearson correlation between the two cells increases from 0.72 to 0.82 (Figure 5.6), especially due to the decreased number of genes only expressed in one cell. MAGIC achieves an even higher correlation (0.95) but also introduces very large counts that do not exist in the raw data. Biological variation between the two cells is likely lost in the imputation process of MAGIC.

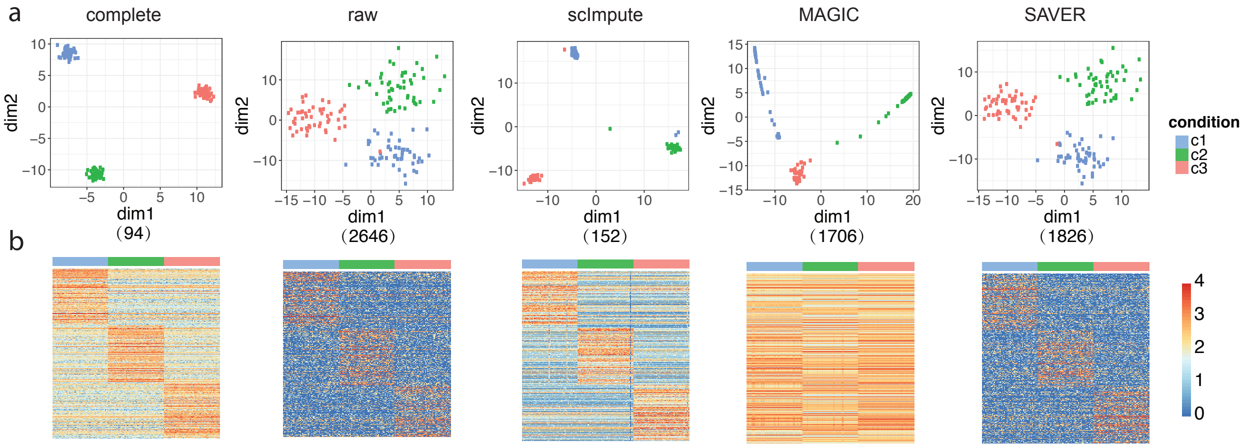


Figure 5.5: The scImpute method corrects dropout values and helps define cellular identity in the simulated data. **a**: The first two PCs calculated from the complete data, raw data, and imputed data by scImpute, MAGIC, and SAVER. Numbers in the parantheses are the within-cluster sum of squares calculated based on the first two PCs. The within-cluster sum of squares is defined as $\sum_{k=1}^{3}\sum_{j=1}^{50}||\boldsymbol{y}_{kj}-\bar{\boldsymbol{y}}_{k\cdot}||^2$, where $\bar{\boldsymbol{y}}_{k\cdot}=\frac{\sum_{j=1}^{50}\boldsymbol{y}_{kj}}{50}$ and $\boldsymbol{y}_{kj}$ is a vector of length 2, denoting the first two PCs of cell j in cell type c_k. **b**: The expression profile of the 810 true DE genes in the complete, raw, and imputed data.

On the other hand, SAVER's imputation does not have a clear impact on the data.

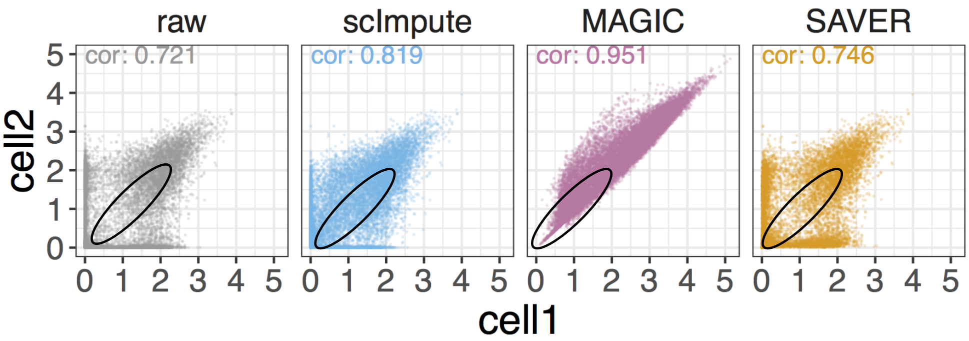


Figure 5.6: Raw and imputed gene expression levels of two mouse embryonic cells from the 16-cell stage. Correlation between the two cells is marked on the top-left of each scatter plot.

We compared the imputation results by investigating the clustering accuracy in the first two PCs. Although it is possible to differentiate the major developmental stages from the raw data, the imputed data by scImpute output more compact clusters (Figure 5.7). MAGIC gives a clean pattern of developmental stages, but it has a high risk of removing biologically meaningful variation, given that many cells of the same stage have almost identical scores in the first two PCs. We then compared the clustering results of the spectral clustering algorithm [198] on the first two PCs. Since the true cluster labels include several sub-stages in embryonic development, we used different numbers of clusters, $k = 6, 8, 10, 12$ and 14. The results were evaluated by four different measures: adjusted rand index [204], Jaccard index [205], normalized mutual information (nmi) [206], and purity. The four measures are all between 0 and 1, with 1 indicating perfect match between the clustering result and the true stages. All the four measures indicate that scImpute leads to the best clustering result as compared with no imputation and the imputation by MAGIC or SAVER (Figure B.13). This result suggests that scImpute improves the clustering of cell subpopulations by imputing dropout values in scRNA-seq data.

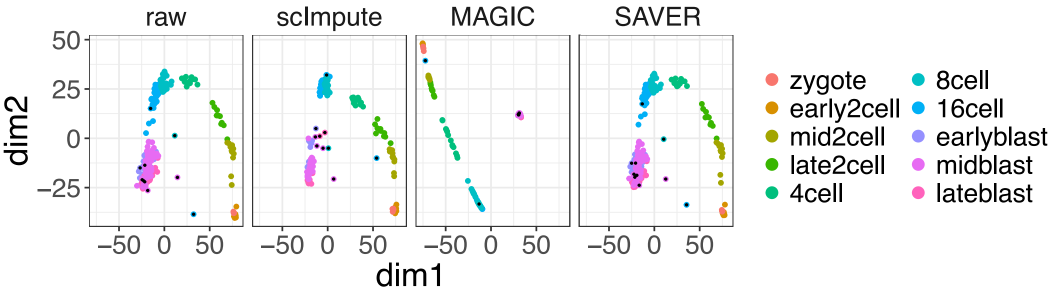


Figure 5.7: The scImpute method improves cell subpopulation clustering in the mouse embryonic cells. The scatter plots show the first two PCs obtained from the raw and imputed data of mouse embryonic cells. The black dots mark the outlier cells detected by scImpute.

We also applied scImpute to a large dataset generated by a high-throughput droplet-based system [160]. The dataset contains 4500 peripheral blood mononuclear cells (PBMCs) of nine immune cell types, with 500 cells of each type. In the raw data, 92.6% read counts are exactly zeros. Given dimensionality reduction by t-SNE [14], the cytotoxic and naïve cytotoxic T cells are clustered together, and the other four types of T cells are not separated. After scImpute's imputation, the cytotoxic (label 11) and naïve cytotoxic T (label 8) cells are separated into two groups, and the naïve T cells (label 5) and memory T cells (label 3) are now distinguishable from the remaining T cells (Figure 5.8). This evidence shows the strong ability of scImpute to identify cell subpopulations despite missing cell type information. On the other hand, MAGIC does not improve the clustering of cells in the same type (Figure B.14), and we could not obtain SAVER's results after running the program overnight. After the imputation by scImpute, the monocyte cells are grouped into one large and two small clusters, and we found that the three clusters reveal dynamics of two signature genes, *FCER1A*, which accumulates during the dendritic cell differentiation from monocytes [207], and *S100A8*, whose expression differs between subsets of human monocytes [208]. The large cluster (label 10) is characterized by high expression of *S100A8* and moderate expression of *FCER1A*; one of the small clusters (label 1) presents high expression of both *S100A8* and *FCER1A*, while in the other small cluster (label 2) *FCER1A* is largely non-expressed.

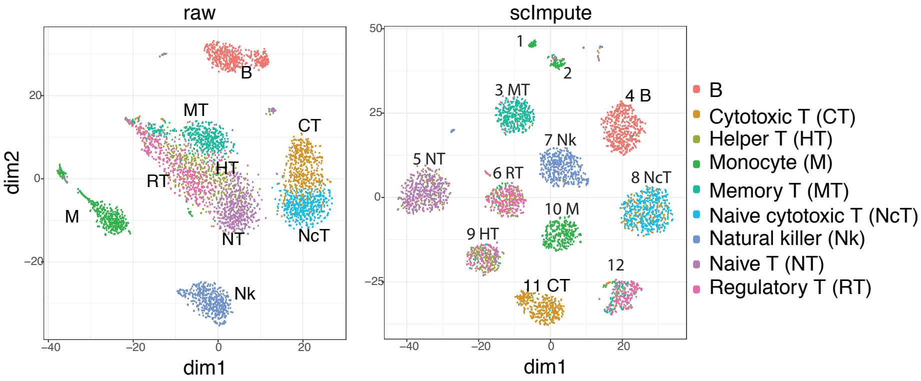


Figure 5.8: The scImpute method helps identify cell subpopulations in the PBMC dataset. The first two dimensions of the t-SNE results were calculated from raw and imputed PBMC dataset. Numbers marked on the imputed data are cluster labels. Cell type information is displayed for major clusters.

5.3.3 scImpute assists differential gene expression analysis

ScRNA-seq data provide insights into the stochastic nature of gene expression in single cells, but suffer from a relatively low signal-to-noise ratio compared with bulk RNA-seq data. Thus an effective imputation method should lead to a better agreement between scRNA-seq and bulk RNA-seq data of the same biological condition on genes known to have little cell-to-cell heterogeneity. To evaluate whether the DE genes identified from single-cell data are more accurate after imputation, we utilized a real dataset with both bulk and single-cell RNA-seq experiments on human ESCs and definitive endorderm cells (DECs) [173]. This dataset includes 6 samples of bulk RNA-seq (4 of H1 ESC and 2 of DEC) and 350 samples of scRNA-seq (212 of H1 ESC and 138 of DEC). The percentages of zero expression are 14.8% in bulk data and 49.1% in single-cell data.

We applied scImpute, MAGIC, and SAVER to impute the gene expression for each cell type respectively, and then performed DE analysis on the raw data and the imputed data. We used the R package DESeq2 [30] to identify DE genes from the bulk data, and the R packages DESeq2 and MAST [192] to identify DE genes from the scRNA-seq data. Inspecting the top 200 DE genes from the bulk data, we found that their expression profiles in the scRNA-seq data have stronger concordance with those in the bulk data after imputation by scImpute. We applied different thresholds to the FDRs of genes in the bulk data to obtain a DE gene list for every threshold. The same thresholds were applied to the FDRs of genes calculated from the raw and imputed scRNA-seq data to obtain DE gene lists respectively. Then we compared the DE gene lists obtained from the scRNA-seq data with those from the bulk data (i.e., the standard) to calculate precision, recall, and F scores (Figure B.15). The scImpute method leads to more similar DE genes to those from the bulk data, and achieves around 10% higher F scores compared with results on the raw data. We found that scImpute makes a good balance between the precision and recall rate, while MAGIC has low precision, and SAVER has low recall and is barely distinguishable from no imputation. We conclude that scImpute is preferred when users have a priority on the overall accuracy of the DE genes.

A comparison between the expression profiles of DEC and ESC marker genes [173, 209, 210] shows that the imputed data by scImpute best reflect the gene expression signatures by removing undesirable technical variation resulted from dropouts (Figure 5.9a). To determine if the DE genes identified in scRNA-seq data are biologically meaningful, we performed gene ontology (GO) enrichment analysis [211]. In the $\sim$ 300 DEC up-regulated genes that are only detected in the imputed data by scImpute but not in the raw data, enriched GO terms are highly relevant to the functions of DECs (Figure 5.9b). However, in the $\sim$ 300 DEC up-regulated genes that are only detected in the raw data, enriched GO terms are general

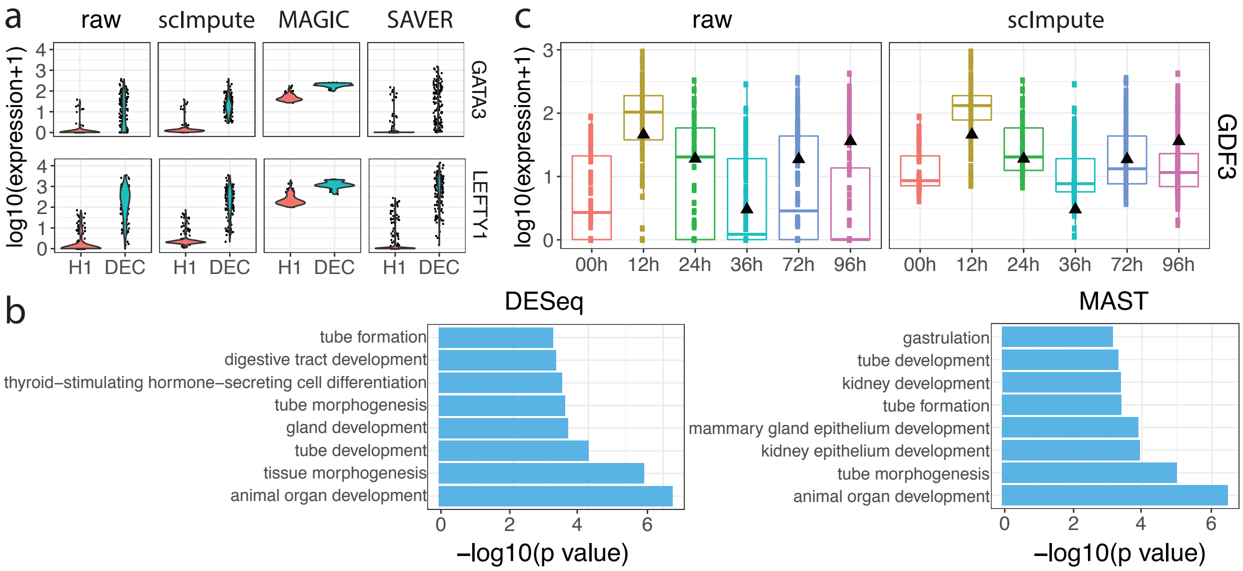


Figure 5.9: The scImpute method improves differential gene expression analysis and reveals expression dynamics in time-course experimments. **a**: Raw and imputed expression levels of two marker genes of DEC. **b**: Selected GO terms enriched in the DEC up-regulated genes that can be only detected (by DESeq2 or MAST) in the imputed data by scImpute, but not in the raw data. **c**: Time-course expression patterns of the gene *GDF3*, which is annotated with the GO term "endoderm development". Black triangles mark the gene's expression in bulk data.

and not characteristic of DECs. These results also demonstrate that scImpute can facilitate the usage of DE methods that were not designed for single-cell data.

5.3.4 scImpute recovers gene expression temporal dynamics

Aside from the data we used in differential expression analysis, Chu et al. [173] also generated bulk and single-cell time-course RNA-seq data profiled at 0, 12, 24, 36, 72, and 96 h of differentiation during DEC emergence. We utilize this dataset to show that scImpute can help recover the DE signals that are difficult to identify in the raw time-course data, and reduce false discoveries resulted from dropouts. We first applied scImpute to the raw scRNA-seq data with true cell type labels, and then studied how the time-course expression patterns

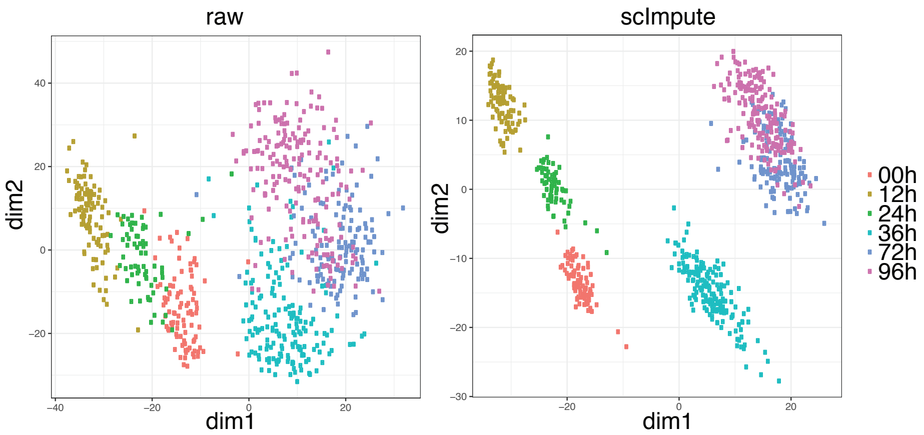


Figure 5.10: The first two PCs calculated from raw and imputed time-course ESC data.

change in imputed data. The imputed data better distinguishes cells of different time points (Figure 5.10), suggesting that imputed read counts reflect more accurate transcriptome dynamics along the time course. Even though the scRNA-seq data present more biological variation than the bulk data, it is reasonable to expect that the average gene expression signal across cells in scRNA-seq should correlate with the signal in bulk RNA-seq. For a genome wide comparison, the imputed data have significantly higher Pearson correlations with the bulk data. We studied 70 genes associated with the GO term "endoderm development" [212] and found that a subset of these genes that are likely affected by dropout events show higher expression and better consistency with the bulk data after the imputation by scImpute (Figures 5.9c). Similarly, we also studied the marker genes of DECs [173, 209, 210] and these genes' expression levels at the time point 96h are recovered by scImpute even though they have a median read count of zero in the raw data (Figure B.16).

5.4 Discussion

In this chapter, we propose a statistical method scImpute to address the dropout events prevalent in scRNA-seq data. scImpute focuses on imputing the missing expression values of dropout genes, while retaining the expression levels of genes that are largely unaffected by dropout events. Hence, scImpute can reduce technical variation resulted from scRNA-seq and better represent cell-to-cell biological variation, while it also avoids introducing excess bias during its imputation process. To achieve the above goal, scImpute first learns each gene's dropout probability in each cell by fitting a mixture model for each cell type. Next, scImpute imputes the highly probable dropout values of genes in a cell by borrowing information of the same gene in other similar cells, which are selected based on the genes not severely affected by dropout events. Comprehensive studies on both simulated and real data suggest that compared with the raw data, the imputed data by scImpute better presents cell type identity and lead to more accurate DE analysis results.

An attractive advantage of scImpute is that it can be incorporated into most existing pipelines or downstream analyses of scRNA-seq data, such as normalization [141, 213], differential expression analysis [168, 192], as well as clustering and classification [185, 187]. The scImpute method takes the raw read count matrix as input and outputs an imputed count matrix of the same dimensions, so it can be seamlessly combined with other computational tools without data reformatting or transformation. However, new analyzing tools specifically designed for the imputed data by scImpute may have improved performance over existing methods developed for raw scRNA-seq data, by incorporating features such as smaller proportion of zero expression, dropout rates, and dropout probabilities estimated by the mixture models. Another important feature of scImpute is that it only involves two parameters that can be easily understood and selected. The first parameter K denotes the potential number of cell populations. It can be selected based on clustering of the raw data and the resolution level desired by the users. If users are only interested in the differences among the major clusters, they could use a relatively small K, and scImpute can borrow more information among individual cells; otherwise, users can select a relatively large K, and scImpute would be more conservative in the imputation process. The second parameter is a threshold t on dropout probabilities. We show in a sensitivity analysis that scImpute is robust to different parameters (Figure B.10), and a default threshold value of 0.5 is sufficient for most scRNA-seq datasets. Moreover, cell type information is not necessary for the scImpute method. When cell type information is available, separate imputation on each cell type is expected to produce more accurate results. However, as illustrated by the simulation and real data studies, scImpute is able to infer cell-type-specific expression even when the true labels are not supplied.

The scImpute method scales up well when the number of cells increases, and the computation efficiency can be largely improved if a filtering step on cells can be performed based on biological knowledge. Aside from computational complexity, another future direction is to further improve imputation efficiency when dropout rates in raw data are severely high, as with the droplet-based technologies. Imputation task becomes more difficult when proportion of missing values increases. More complicated models that account for gene similarities may yield more accurate imputation results, but the prevalence of dropout events may require additional prior knowledge on similar genes to assist modeling. Despite the availability of computational methods that directly model zero-inflation in data [168, 192], scImpute takes the imputation perspective to improve the data quality, and its applicability is not restricted to a specific task. Hence, scImpute is an useful tool that benefits all types of scRNA-seq downstream analyses.

5.5 Acknowledgments

This chapter is based on my joint work with Dr. Jessica Li [214]. I would like to thank Dr. Li for her advice and support. We are grateful to Douglas Arneson, Feiyang Ma, Dr. Robert Modlin, Dr. Matteo Pellegrini, and Dr. Xia Yang at the University of California, Los Angeles, for providing insightful discussions. We thank Dr. Mark Biggin at Lawrence Berkeley National Laboratory for his suggestions on our work. We also thank Dr. Daria Merkurjev for assisting with the data collection.

CHAPTER 6

Experimental Design for Single-cell RNA Sequencing

6.1 Introduction

Since the first scRNA-seq study was published in 2009 [156], more than twenty scRNA-seq experimental protocols have been developed. An effective experimental design requires careful consideration of the target research question as well as the experimental budget, and a typical design in practice consists of two steps. First, researchers need to select a proper protocol among the available ones, and the primary consideration is the choice between a tag-based protocol that allows the integration of UMIs [164] and a full-length protocol that captures full-length transcripts and allows the addition of the ERCC spike-ins [165]. The tag-based protocols (e.g., Drop-seq [159]) are usually used to obtain a broad but shallow view of the transcriptomes across many cells, while the full-length protocols (e.g., Smart-seq2 [157]) provide a deeper account of the gene expression in fewer cells. For example, a study about gene expression dynamics during stem cell differentiation requires accurate gene expression measurements, so it should opt for a full-length protocol. In contrast, in a study aiming to identify a previously unknown cell phase during the differentiation, it is necessary to sequence a large number of cells using a tag-based protocol to capture the transient phases. In the second step, to optimize an experiment with a selected protocol and a fixed budget, researchers need to choose between exploring the depth or breadth of transcriptome information, which sums up to determining the appropriate number of cells to sequence.

However, in contrast to the classical experimental design [215] guided by certain theoretical optimality (e.g., the maximum power of a statistical test), the scRNA-seq experimental design is impeded by various sources of data noises, making a reasonable theoretical analysis tremendously difficult [88, 176]. Especially, scRNA-seq data are characterized by excess zeros resulted from the dropout events, in which a gene is expressed in a cell but its mRNA transcripts are undetected. As a result, many commonly used statistical assumptions are not directly applicable to modeling scRNA-seq data. For example, Baran-Gale et al. proposed using a Negative Binomial (NB) model to estimate the number of cells to sequence, so that the resulting experiment is expected to capture at least a specified number of cells from the rarest cell type [216]. However, the estimation accuracy depends on the idealized NB model assumption, which real scRNA-seq data usually do not closely follow (Figure 6.1). There is also a theoretical investigation of the cell-depth trade-off based on the Poisson assumption of gene read counts and a specific list of genes of interests [217]. In contrast to model-based design approaches [218], multiple scRNA-seq studies used descriptive statistics to provide qualitative guidance instead of well-defined optimization criteria for experimental design [162, 219]. However, since the descriptive statistics were proposed from different perspectives, their resulting experimental designs are difficult to unify to guide practices. For example, one study reported that the sensitivity of most protocols saturates at approximately one million reads per cell [220], while another study found that the saturation occurs at around 4.5 million reads per cell [170]. The reason for this discrepancy is that the two studies defined the sensitivity in different ways: the first study used the gene detection rate while the second study used the minimum number of input RNA molecules required for confidently detecting a spike-in control [202].

In this chapter, we propose a statistical simulator scDesign for optimizing scRNA-seq experimental design from the perspective of detecting DE genes between two biological conditions (determined before an experiment) or two cell states (inferred after an experiment), a major scRNA-seq data analysis task. Given a pre-defined significance level (e.g., an FDR), the power of a scRNA-seq experiment for detecting DE genes is jointly determined by the sensitivity of detecting gene expression, the accuracy of measuring gene expression, and the number of cells sequenced for each cell state. For each protocol and a specified total sequencing depth (i.e., the total number of reads in a scRNA-seq experiment), the cell-wise

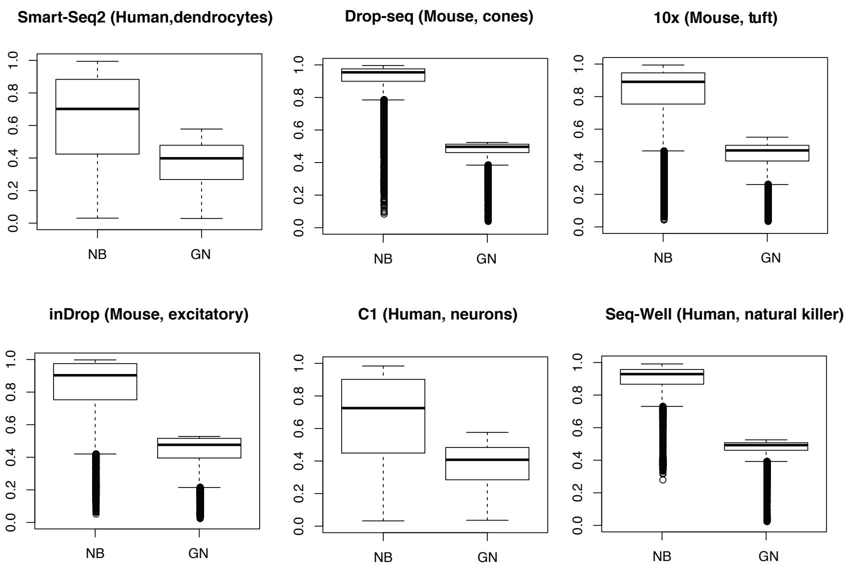


Figure 6.1: Comparison of the Negative Binomial (NB) model and the Gamma-Normal (GN) model used in scImpute. Both models were used to fit six scRNA-seq datasets from different protocols, as listed in Table C.3. For each gene, the Kolmogorov-Smirnov (KS) distance between the empirical and fitted gene expression distributions was calculated and summarized in the boxplots. In all the scenarios, the GN model leads to smaller KS distances.

sequencing depth (i.e., the expected number of reads per cell) decreases as the cell number increases [144]. However, existing power analysis methods for scRNA-seq experiments unrealistically assume a fixed cell-wise sequencing depth, which does not change as the cell number varies [220, 221]. Therefore, the practical scRNA-seq experimental design calls for a new approach that accounts for various characteristics and constraints of a real scRNA-seq experiment.

Our scDesign method provides a simulation-based experimental design framework that has multiple advantages in real practice. First, scDesign is protocol- and data-adaptive. It learns scRNA-seq data characteristics from rapidly accumulating public scRNA-seq data generated under diverse settings. For example, 1976 series of scRNA-seq datasets are currently available in the Gene Expression Omnibus (GEO) database [222]. There are also newly developed scRNA-seq databases such as SCPortalen (70 studies with 67, 146 cells) [223], scRNASeqDB (36 studies with 8910 cells) [224], and the Single Cell Portal (43 studies with

$496,366$ cells). Second, scDesign generates synthetic data that well mimic real scRNA-seq data under the same experimental settings, providing a basis for using its synthetic data to guide practical scRNA-seq experimental design. Third, scDesign is flexible in accommodating user-specific analysis needs. Users can apply scDesign to evaluate the performance of downstream analyses, such as gene differential expression and cell clustering, under various experimental settings at no experimental cost. Assisted by the evaluation results, users will be able to design a scRNA-seq experiment according to their specified criteria.

6.2 Methods

6.2.1 The statistical framework of scDesign

We develop scDesign based on a statistical generative framework that utilizes both existing real scRNA-seq data and reasonable assumptions mimicking various experimental processes. In contrast to the existing simulation methods for scRNA-seq data, scDesign constructs a Gamma-Normal mixture model to account for dropout events. This is motivated by the successful applications of our imputation method, scImpute, for recovering dropout gene expression values in scRNA-seq data (as discussed in Chapter 5). This mixture model allows scDesign to overcome the dropout hurdle in learning the key gene expression characteristics from real scRNA-seq data (Figure 6.1), so that scDesign generates synthetic data highly similar to real data in multiple aspects.

Depending on whether the task is to design a scRNA-seq experiment to sequence one or two batches of cells, scDesign has the corresponding one-state mode (Figure 6.2a) or the two-state mode (Figure 6.2b). In the one-state mode, scDesign leverages the information in a real scRNA-seq dataset from one biological condition (e.g., treatment or control) or one cell state (e.g., T cells) to generate a single scRNA-seq dataset given an experimental setting, i.e., a pre-specified total sequencing depth and a cell number. From the real scRNA-seq dataset, scDesign first estimates two cell-wise and three gene-wise parameters, which jointly define the key characteristics of scRNA-seq data. Second, scDesign simulates ideal

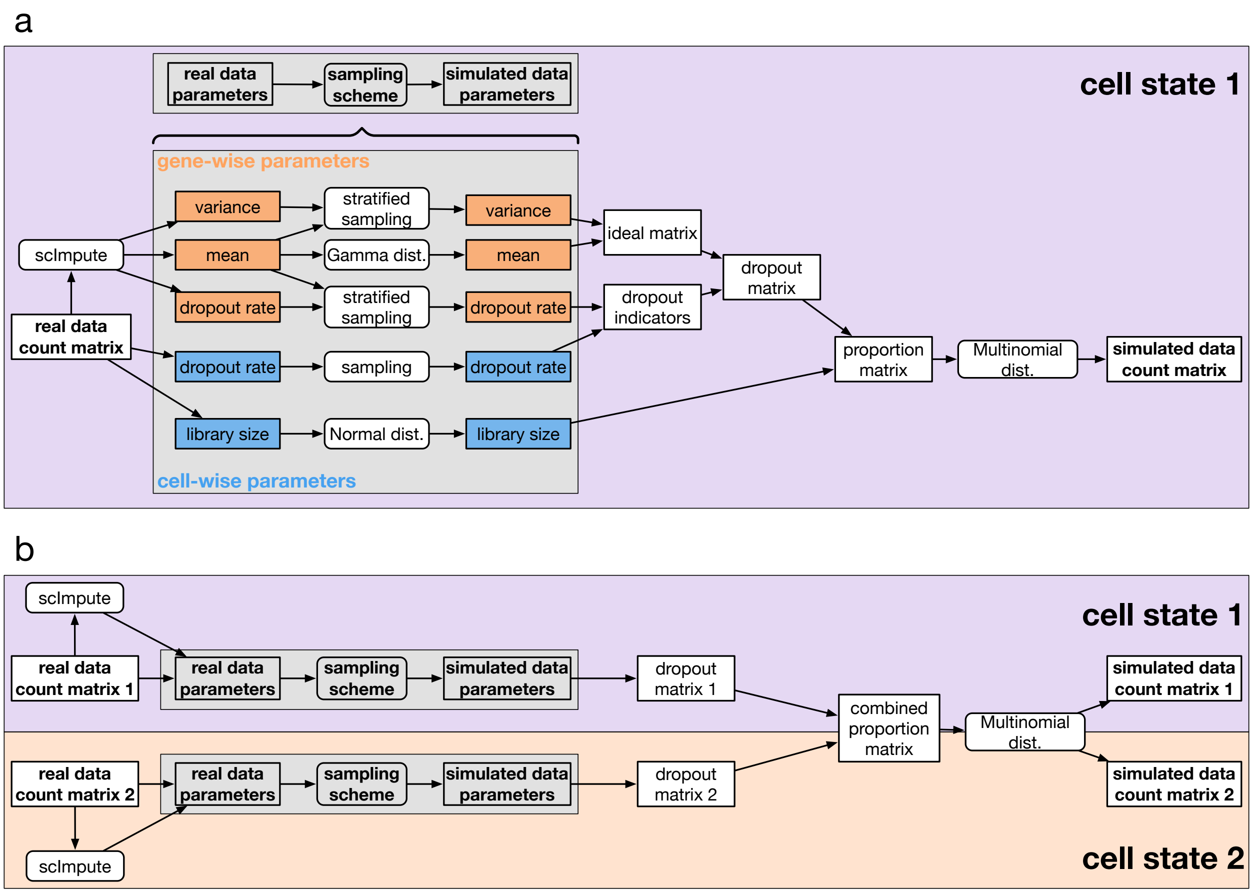


Figure 6.2: The statistical framework of scDesign. **a**: The simulation process of one count matrix from a single cell state (one-state mode). **b**: The joint simulation process of two count matrices from two different cell states (two-state mode).

gene expression levels for new cells of the same biological condition or cell state based on the estimated gene expression parameters. Third, scDesign introduces missing values to mimic the actual dropout events in a scRNA-seq experiment. Fourth, scDesign outputs a synthetic gene expression matrix with entries as read counts. In the two-state mode, scDesign leverages the information in two real scRNA-seq datasets from different biological conditions or cell states to generate two datasets given an experimental setting. In this two-state mode, the simulation by scDesign mimics an experiment where two groups of cells from two biological conditions or cell states are sequenced together (Figure 6.2b). Similar to the one-state mode, scDesign independently simulates ideal gene expression levels for new cells of the two cell states, introduces dropout values based on the estimated dropout parameters of each state, and generates observed read counts by accounting for the fact that RNA molecules from the two batches of cells compete to be sequenced. Finally, scDesign outputs two gene expression count matrices, one for each cell condition or state. It is worth noting that the scDesign framework is directly generalizable to more than two biological conditions or cell states.

6.2.2 scDesign for scRNA-seq data simulation

We first describe how scDesign generates simulated RNA-seq data given existing real scRNA-seq data from a certain cell state. These simulated count matrices capture the characteristics of real count matrices, so they can be used to assist the development of computational methods and evaluate the performance of those methods under user-specified settings. We introduce how to simulate a single count matrix below, and introduce how to simulate multiple count matrices following a differentiation path in Appendix A.9.

Given a real single-cell count matrix with I genes and J_0 cells, our goal is to generate a new count matrix with I genes and J cells, under the constraint that the new matrix has a total of S reads (Figure 6.2a). Both J and S are user-specified parameters. This resembles the real scenario where both the cell number and the total read number need to be specified before a scRNA-seq experiment.

(1). Estimate parameters from real scRNA-seq data.

We denote the real single-cell count matrix by $\boldsymbol{X}^{\text{real}}$, whose I rows and J_0 columns represent the genes and cells, respectively. About the two cell-wise parameters, for each cell j we estimate its library size as

$$\hat{s}_{0j} = \sum_{i=1}^{I} X_{ij}^{\text{real}},\ j = 1, \ldots, J_0\,, \tag{6.1}$$

and its cell-wise dropout rate as

$$\hat{q}_{0j} = \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}\{X_{ij}^{\text{real}} = 0\},\ j = 1, \ldots, J_0\,. \tag{6.2}$$

Then we fit the cell library sizes $\hat{s}_{01}, \ldots, \hat{s}_{0J_0}$ using a truncated Normal distribution, and the estimated mean and standard deviation (SD) are denoted as $\hat{\mu}_s$ and $\hat{\sigma}_s$, respectively.

To estimate the three gene-wise parameters, we first normalize the read counts given their corresponding library sizes and then perform a logarithmic transformation on the normalized values. The transformed matrix is denoted as $\boldsymbol{X}^{\log}$, where

$$X_{ij}^{\log} = \log_{10}\left(\frac{\text{median}\{\hat{s}_{01}, \ldots, \hat{s}_{0J}\}}{\hat{s}_{0j}} X_{ij}^{\text{real}} + 1.01\right) . \tag{6.3}$$

Using the Gamma-Normal mixture model described in the scImpute method, we estimate the gene-wise dropout rate as well as mean and SD of gene expression. The mixture model considers the expression levels of gene i as independently and identically distributed random variables, $X_{i1}^{\log}, \ldots, X_{iJ_0}^{\log}$, following the density function

$$f_i(x) = \lambda_{0i}\,\text{Gamma}\left(x; \alpha_{0i}, \beta_{0i}\right) + \left(1 - \lambda_{0i}\right)\text{Normal}\left(x; \mu_{0i}, \sigma_{0i}^2\right) , \tag{6.4}$$

where λ_{0i} is gene i's dropout rate, α_{0i} and β_{0i} are the shape and rate parameters of the Gamma distribution, and μ_{0i} and σ_{0i} are the mean and SD of the Normal distribution. The Gamma component describes the gene expression distribution when the dropout events occur, while the Normal component represents the distribution of actual gene expression levels. We used multiple real scRNA-seq datasets to demonstrate that this mixture model outperforms the widely used NB model in terms of goodness-of-fit to real data (Figure 6.1). The parameters in this model can be estimated by the EM algorithm and the resulting dropout rate, mean, and SD estimates are denoted as $\hat{\lambda}_{0i}$, $\hat{\mu}_{0i}$, and $\hat{\sigma}_{0i}$, respectively. We then use a Gamma distribution to fit the estimated mean expression levels $\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}$ and denote the estimated shape and scale parameters as $\hat{k}_0$ and $\hat{\theta}_0$.

To summarize, we estimate two cell-wise parameters including the cell library size $\hat{s}_{0j}$ and the cell-wise dropout rate $\hat{q}_{0j}$ ($j = 1, \ldots, J_0$), and estimate three gene-wise parameters including the mean expression $\hat{\mu}_{0i}$, the SD $\hat{\sigma}_{0i}$, and the gene-wise dropout rate $\hat{\lambda}_{0i}$ ($i = 1, \ldots, I$).

(2). Simulate ideal gene expression values.

In this step, we simulate the ideal expression values independently for each gene without considering varying cell library sizes and the dropout issue. For each gene i, we first simulate its mean expression $\mu_i \sim \text{Gamma}(\hat{k}_0, \hat{\theta}_0)$, then we simulate its SD by stratified sampling from the binned observations, which we could process from the real count matrix. Specifically, we divide the estimated gene mean expression values $\{\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}\}$ into B intervals, and we use $\hat{\mu}_{0(k)}$ to denote the k-th order statistic of $\{\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}\}$. Then, the first interval is $\left(-\infty, \hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)}-\hat{\mu}_{0(1)}}{B}\right]$, the B-th interval is $\left(\hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)}-\hat{\mu}_{0(1)}}{B}(B-1), +\infty\right)$, and the b-th interval is $\left(\hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)}-\hat{\mu}_{0(1)}}{B}(b-1), \hat{\mu}_{0(1)} + \frac{\hat{\mu}_{0(I)}-\hat{\mu}_{0(1)}}{B}b\right]$ $(1 < b < B)$. We define $\hat{z}_{0i} = b$ if $\hat{\mu}_{0i}$ belongs to the b-th bin, and similarly we define $z_i = b$ if μ_i belongs to the b-th bin. We simulate the SD σ_i of gene i by sampling from the stratified gene SDs estimated from the real data: $\sigma_i \sim \text{Uniform}(\{\hat{\sigma}_{0i'} : \hat{z}_{0i'} = z_i, i' = 1, \ldots, I\})$. Finally, we generate the ideal expression matrix $\boldsymbol{X}^{\text{ideal}}$, where $X_{ij}^{\text{ideal}} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_i, \sigma_i^2)$, $j = 1, \ldots, J$.

(3). Introduce dropout events.

In this step, we introduce dropout values into the synthetic count matrix, while accounting for the variability of both gene-wise and cell-wise dropout rates. The cell-wise dropout rate in a synthetic cell j is simulated as $q_j \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\{\hat{q}_{01}, \ldots, \hat{q}_{0J_0}\})$, $j = 1, \ldots, J$. For each gene i $(i = 1, \ldots, I)$, we simulate its gene-wise dropout rate λ_i by sampling one value from the stratified dropout rates estimated from the real data: $\lambda_i \sim \text{Uniform}(\{\hat{\lambda}_{0i'} : \hat{z}_{0i'} = z_i, i' = 1, \ldots, I\})$. Then, we simulate the number of dropout events for gene i: $n_i \sim \text{Binomial}(J, \lambda_i)$. In other words, gene i is affected by the dropout events in n_i cells. These n_i cells are sampled without replacement from the cell population $\{1, 2, \ldots, J\}$, with cell j being selected with probability $\frac{q_j}{\sum_{j=1}^{J} q_j}$. We denote the sampling results by I_{ij}, with $I_{ij} = 1$ indicating that gene i is a dropout in cell j and $I_{ij} = 0$ indicating that gene i is successfully amplified in cell j. Then we obtain the synthetic count matrix with dropout events $\boldsymbol{X}^{\text{drop}}$, where $X_{ij}^{\text{drop}} = \left[10^{X_{ij}^{\text{ideal}} \mathbb{I}\{I_{ij}=0\}} - 1.01\right]$, and $[x]$ is the nearest integer to x.

(4). Simulate the final count matrix.

We first simulate the library size of each synthetic cell j: $s_j \overset{\text{i.i.d.}}{\sim} \text{Normal}(\hat{\mu}_s, \hat{\sigma}_s^2)$, $j = 1, \ldots, J$, and then we calculate the expected proportion of each entry in the count matrix $P_{ij} = {}^{s_j X_{ij}^{\text{drop}}}/_{\sum_{i=1}^{I} \sum_{j=1}^{J} s_j X_{ij}^{\text{drop}}}$. Finally, we obtain the final synthetic count matrix $\boldsymbol{X}^{\text{syn}}$, which is constrained by the sequencing depth S, by simulating its counts from the multinomial distribution:

$$(X_{11}^{\text{syn}}, ..., X_{1J}^{\text{syn}}, ..., X_{I1}^{\text{syn}}, ..., X_{IJ}^{\text{syn}}) \sim \text{Multinomial}\left(S, (P_{11}, ..., P_{1J}, ..., P_{I1}, ..., P_{IJ})\right) . \quad (6.5)$$

6.2.3 scDesign for scRNA-seq experimental design

The scDesign method aims to determine the best number of cells to sequence given a fixed sequencing depth, such that the resulting RNA-seq data are optimized for differential gene expression analysis. We denote the two real count matrices as $\boldsymbol{X}^{\text{real1}}$, with I rows (genes) and J_{01} columns (cells), and $\boldsymbol{X}^{\text{real2}}$, with I rows (genes) and J_{02} columns (cells). Without loss of generality, we assume that the two matrices, which represent two cell states, have the same genes listed in the same order. We introduce how to simulate a synthetic count matrix for each state with scDesign in two scenarios, and the procedure is then repeated with varying cell numbers to obtain synthetic data for power analysis (Appendix A.10).

Scenario A. In scenario A, we assume that cells from the two cell states are prepared as separate libraries and sequenced independently. Given $\boldsymbol{X}^{\text{real1}}$ and $\boldsymbol{X}^{\text{real2}}$, the goal of scDesign is to generate a synthetic count matrix with I genes and J_1 cells for state 1, and a synthetic count matrix with I genes and J_2 cells for state 2. Cell states 1 and 2 have sequencing depths of S_1 and S_2, respectively. For each state g ($g = 1, 2$), we follow Chapter 6.2.2 to simulate a count matrix $\boldsymbol{X}_{I \times J_g}^{\text{syn},g}$. The only difference is in step (2), where we directly set $\mu_i^g = \hat{\mu}_{0i}^g$ and $\sigma_i^g = \hat{\sigma}_{0i}^g$, $i = 1, ..., I$, instead of simulating new parameters. This is to ensure that the rows in the two simulated matrices still represent the same set of real genes, and the power analysis based on the simulated data is biologically meaningful.

Scenario B. Now we consider the case where the two cell states are jointly sequenced. Suppose that the two cell states are mixed in one biological sample, and the experimental setting is that J cells are to be sequenced to generate S RNA-seq reads in total. We assume that the two cell states present in fractions of p_1 and p_2 in the sample, respectively. That is, $0 < p_1 < 1$, $0 < p_2 < 1$, and $p_1 + p_2 \leq 1$. When $p_1 + p_2 < 1$, there are more than two cell states present in the same sample. The goal of scDesign in scenario B is to simulate count matrices for the two selected cell states, based on a real count matrix of each state.

(1). Determine cell numbers.

We denote the numbers of cells from state 1, state 2, and the remaining states as J_1, J_2, and J_r, respectively. These numbers are sampled from a Multinomial distribution: $(J_1, J_2, J_r) \sim \text{Multinomial}\left(J, (p_1, p_2, 1 - p_1 - p_2)\right)$.

(2). Simulate count matrices with dropout events.

Following steps (1)-(3) in Chapter 6.2.2, we simulate two count matrices $\boldsymbol{X}^{\text{drop1}}_{I\times J_1}$ and $\boldsymbol{X}^{\text{drop2}}_{I\times J_2}$ for cell states 1 and 2, respectively. The only difference is in step (2), where we directly set $\mu_i^g = \hat{\mu}_{0i}^g$ and $\sigma_i^g = \hat{\sigma}_{0i}^g$, $i = 1, ..., I$, to ensure that the rows in the synthetic count matrices represented the same set of real genes.

(3). Simulate the final count matrices.

We first simulate the library sizes of the cells in the two states: $s_j^1 \sim \text{Normal}\left(\hat{\mu}_s^1, (\hat{\sigma}_s^1)^2\right)$, $j = 1, ..., J_1$; $s_j^2 \sim \text{Normal}\left(\hat{\mu}_s^2, (\hat{\sigma}_s^2)^2\right)$, $j = 1, ..., J_2$, where $\hat{\mu}_s^1$ and $\hat{\sigma}_s^1$ are estimated from $\boldsymbol{X}^{\text{real1}}$, and $\hat{\mu}_s^2$ and $\hat{\sigma}_s^2$ are estimated from $\boldsymbol{X}^{\text{real2}}$. Then we combine the two count matrices to obtain the expected proportion matrix $\boldsymbol{P}_{I\times(J_1+J_2)}$:

$$P_{ij} = \frac{Z_{ij}}{\sum_{i=1}^{I}\sum_{j'=1}^{J_1} s_{j'}^1 X_{ij'}^{\text{drop1}} + \sum_{i=1}^{I}\sum_{j''=1}^{J_2} s_{j''}^2 X_{ij''}^{\text{drop2}}}\,, \tag{6.6}$$

where $Z_{ij} = s_j^1 X_{ij}^{\text{drop1}}$ if $1 \leq j \leq J_1$, and $Z_{ij} = s_{j-J_1}^2 X_{i(j-J_1)}^{\text{drop2}}$ if $J_1 < j \leq J_1 + J_2$. The first J_1 columns and the last J_2 columns in $\boldsymbol{P}$ give the expected proportions of genes in cell states 1 and 2, respectively. We further assume that the total number of reads from the two states together is $S_0 = [S(J_1 + J_2)/J]$, where $[x]$ denotes the nearest integer to x. Then we simulate the final count matrix $\boldsymbol{X}^{\text{syn}}_{I\times(J_1+J_2)}$ constrained by the sequencing depth from a Multinomial distribution: $\left(\boldsymbol{X}^{\text{syn}}_{11}, ..., \boldsymbol{X}^{\text{syn}}_{1(J_1+J_2)}, ..., \boldsymbol{X}^{\text{syn}}_{I1}, ..., \boldsymbol{X}^{\text{syn}}_{I(J_1+J_2)}\right) \sim \text{Multinomial}\left(S_0, (P_{11}, ..., P_{1(J_1+J_2)}, ..., P_{I1}, ..., P_{I(J_1+J_2)})\right)$. The final count matrices of cell state 1 and state 2 are $\boldsymbol{X}^{\text{syn},1}_{I\times J_1}$ ($X^{\text{syn},1}_{ij} \triangleq X^{\text{syn}}_{ij}$) and $\boldsymbol{X}^{\text{syn},2}_{I\times J_2}$ ($X^{\text{syn},2}_{ij} \triangleq X^{\text{syn}}_{i(j+J_1)}$), respectively.

6.3 Results

6.3.1 scDesign captures key characteristics of scRNA-seq data

We first demonstrate that scDesign accurately captures six key characteristics of real scRNA-seq data, so it serves as a reliable data simulator to assist scRNA-seq experimental design and to compare relevant computational methods. To assess the simulation performance of scDesign as compared with four other simulation methods, splat, powsimR, Lun, and scDD, we compared the simulated data generated by each method with the real data from various protocols. Both splat and powsimR are software packages for simulating scRNA-seq data [221, 225]; Lun denotes the simulation design introduced by Lun et al. [226]; scDD denotes the simulation method designed to evaluate the DE method scDD [194]. We considered six experimental protocols, Smart-seq2 [157], Drop-seq [159], 10x Genomics [160], Fluidigm C1 (SMARTer) [227], inDrop [158], and Seq-Well [161], and we collected three real scRNA-seq gene expression datasets of distinct cell types from each protocol (Table C.3). In summary, we used 18 real count matrices of 17 cell types from human and mouse to evaluate the five simulation methods.

For each real count matrix, we randomly split the cells into two subsets of equal sizes, one used to estimate gene expression parameters and simulate a new count matrix with the same dimensions, and the other used to evaluate the simulation results. We compared each pair of real and simulated count matrices in terms of six summary statistics, including four gene-wise statistics (the count mean, the count variance, the count coefficient of variation (cv), and the gene-wise zero proportion) and two cell-wise statistics (the library size and the cell-wise zero proportion). Our results show that scDesign well mimics real scRNA-seq experiments based on all six experimental protocols, even though those protocols generate data with distinct properties. For example, data from Smart-seq2 and Fluidigm C1 have relatively larger library sizes and smaller count cvs (Figure 6.3a), while data from the other four protocols have smaller library sizes and larger zero proportions (Figure B.17). We measured the similarity between each summary statistics' empirical distributions in real and the

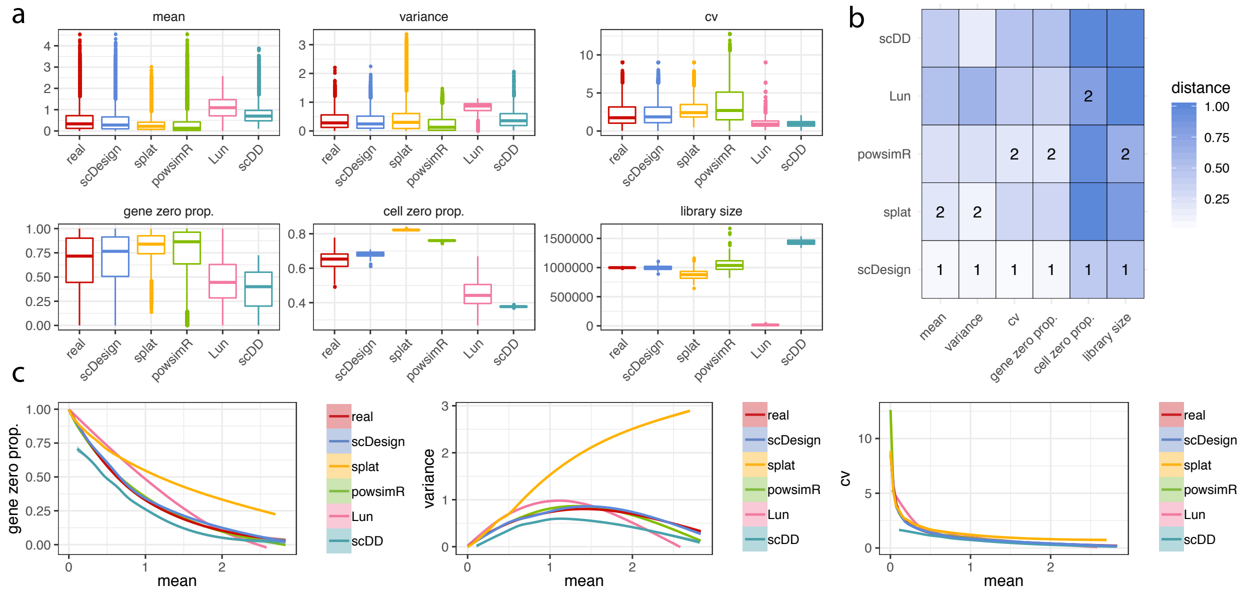


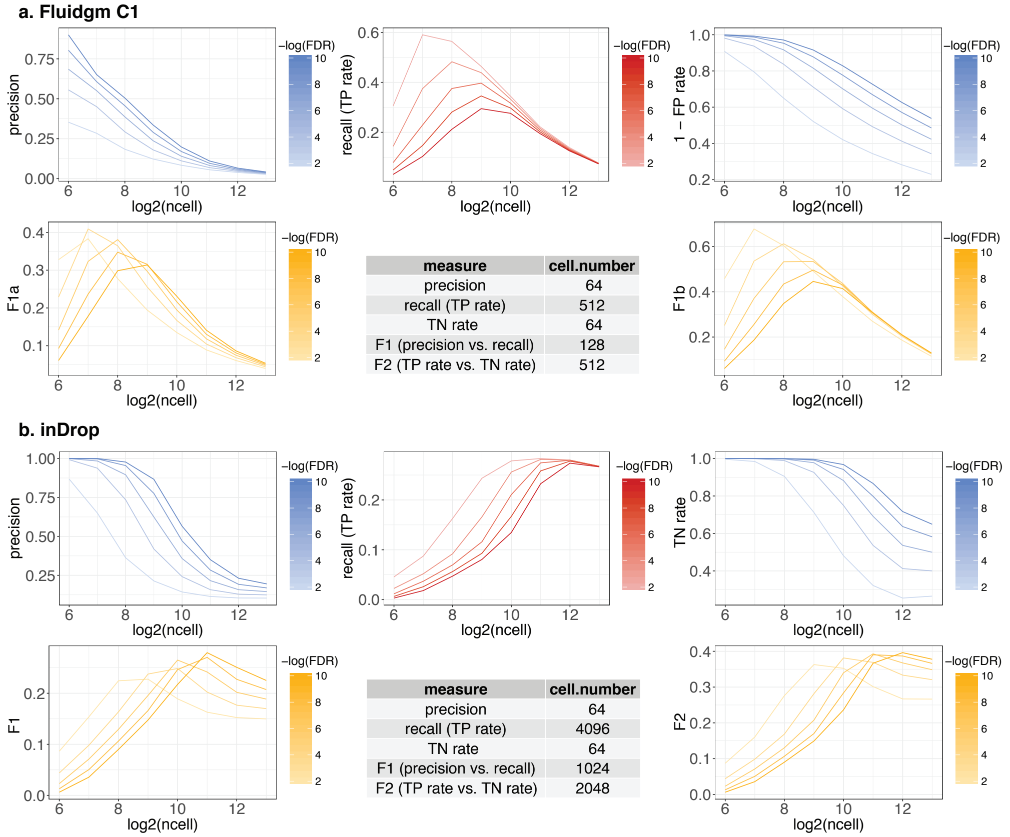
Figure 6.3: Comparison of scRNA-seq simulation methods based on the Smart-seq2 protocol. **a:** The gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in real and simulated monocyte datasets. **b:** The KS distances between the six statistics in the real and simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmap. **c:** The empirical relationships between the key statistics in the real and simulated data.

corresponding simulated data, using the Kolmogorov-Smirnov (KS) distance, whose value is between 0 and 1 and a smaller value indicates greater similarity. Comparing the KS distances of the five methods, we found that scDesign performs the best for five protocols: Smart-seq2, Fluidigm C1, Seq-Well, Drop-seq, and inDrop (Figure 6.3b), while scDesign and powsimR perform comparably for 10x Genomics (Figure B.17). In summary, scDesign is ranked the best in 84 comparisons and the second best in 20 comparisons, among all the 108 comparisons (six statistics for each of the 18 datasets). In addition, our results also show that scDesign is able to preserve the relationships between genes' expression mean and expression variance, expression cv, and zero proportion (Figure 6.3c). The demonstrated advantage of scDesign is rooted in its ability to incorporate both parametric and non-parametric methods to simulate scRNA-seq data. By constructing a mixture model to account for the dropout events, scDesign explicitly models the gene-wise parameters from the real data. When generating cell-wise parameters for the simulated cells, scDesign uses different sampling techniques for each parameter to capture its distribution characteristic. In terms of the method stability, scDesign, Lun, and splat successfully generated simulated data for all the 18 datasets, while scDD encountered errors with five datasets, and powsimR had errors with four datasets.

6.3.2 scDesign guides rational scRNA-seq experimental design

Given a fixed sequencing depth in designing a scRNA-seq experiment, scDesign assists users to predict the optimal number of cells for sequencing. In the context of gene differential expression analysis of two cell states, the cell number is optimal if its resulting scRNA-seq data lead to the most accurate detection of DE genes, where the accuracy depends on a user-specified criterion, e.g., a statistical test's power given a significance level. We consider two scenarios: (A) cells from the two cell states are prepared as two separate libraries and sequenced independently; (B) cells from the two cell states are prepared in the same library and sequenced together. Scenario A includes many studies that investigated cells collected at two differentiating time points, cells of the same tissue type from patients and healthy subjects, or cells of the same type but exposed to different experimental treatments [228,229]. The experimental design under scenario A aims to select the optimal cell numbers simultaneously for two libraries, so that the subsequent DE analysis becomes the most accurate given a user-specified criterion. Scenario B includes many scRNA-seq studies that sequenced an *in vivo* tissue sample, e.g., the peripheral blood mononuclear cell sample [160], which is composed of a mixture of cell subtypes. In scenario B, DE analysis is performed on a pair of known or putative cell subtypes within the sequenced sample.

In scenario A, the constraints are the total sequencing depths of the two cell states, and scDesign aims to determine the optimal cell number of each cell state, among a set of candidate cell numbers. The scDesign method simulates a new count matrix of each state based on a real count matrix of the same state, for each pre-specified sequencing depth and cell number. Once obtaining the simulated count matrices corresponding to various candidate cell numbers, scDesign assesses the accuracy of DE gene identification using five metrics: precision, recall, true negative rate, F1 score (the harmonic mean of precision and recall), and F2 score (the harmonic mean of true negative rate and recall). We applied scDesign to optimize the designs of 14 example experiments (Table C.4). In every experiment, we set the sequencing depth to 100 million reads, and considereded eight candidate cell numbers per cell state: 64, 128, 256, 512, 1024, 2048, 4096, and 8192. The DE genes between two cell states were identified using a two-sample t test.



measure	cell.number
precision	64
recall (TP rate)	512
TN rate	64
F1 (precision vs. recall)	128
F2 (TP rate vs. TN rate)	512

measure	cell.number
precision	64
recall (TP rate)	4096
TN rate	64
F1 (precision vs. recall)	1024
F2 (TP rate vs. TN rate)	2048

Figure 6.4: Power analysis for DE studies comparing astrocytes and oligodendrocytes (scenario A). The thresholds on the FDRs to identify DE genes are denoted in the color legends. The table summarizes the optimal cell number according to each metric. **a:** the Fluidigm C1 protocol; **b:** the inDrop protocol.

Our results suggest that given a selected criterion in the DE analysis, the optimal cell number is jointly determined by multiple technical factors, including the experimental protocol and the variation introduced by sequencing, as well as biological factors, such as the intra- and inter- state cellular heterogeneity. Two factors are notable. First, when cells of the same two states are sequenced, the optimal cell number varies with protocols. For example, between two types of glial cells: astrocytes and oligodendrocytes, 512 cells per state is the optimal cell number that maximizes the recall in DE analysis when Fluidigm C1 is used, but the number becomes 4096 per state when inDrop is used (Figure 6.4). If users choose F1 score as the criterion, the optimal cell number per state is 128 and 1024 for Fluidigm C1 and inDrop, respectively. Therefore, Fluidigm C1 and inDrop require vastly different cell numbers to reach the same level of accuracy in DE analysis, and inDrop generally needs more cells than Fluidigm C1. This result is reasonable, since inDrop is a tag-based protocol that is advantageous in capturing more cells but disadvantageous in measuring each cell accurately, compared with the full-length protocol Fluidigm C1. Second, under the same protocol, the optimal cell number depends on the transcriptome similarity of the two cell states. For instance, with Smart-seq2, 512 cells need to be sequenced per state to maximize the recall in identifying DE genes between two dendrocyte subtypes, but only 256 cells per state are needed when dendrocytes are compared with monocytes (Table C.4). If the goal is to maximize the F2 score, the optimal cell number for comparing the two dendrocyte subtypes remains 512 per state, but the number reduces to 128 for comparing dendrocytes with monocytes. It is worth noting that the optimal cell number for both comparisons becomes 64, the smallest candidate cell number, when the criterion is the precision or the true negative rate (Table C.4). The reason is that only the genes with strong DE signals are detectable with a small sample size (cell number) in any statistical testing. Hence, with a reasonable lower bound on the cell number, the DE genes detected at a smaller cell number have a higher precision. Unlike the precision, the largest recall in DE analysis is mostly achieved at a medium to large cell number. In all the experimental designs we evaluated, the recall rate of DE genes first increases with the cell number and then decreases after reaching a peak. These results demonstrate the trade-off between the cell number and the cell-wise library size in scRNA-seq experiments. A combination of a small cell number and a large cell-wise library size ensures the identification of the DE genes with strong DE signals (i.e., achieving a high precision rate), but the small cell number may prohibit the detection of the DE genes with small to medium DE signals (i.e., sacrificing the recall rate). On the other hand, a combination of a reasonably large cell number and a small cell-wise library size increases the recall rate in detecting DE genes but compromises the precision rate due to high dropout rates. We also performed the DE analysis by replacing the two-sample t test with a scRNA-seq DE method MAST [192]. The optimal cell number remains 64 per state when the criterion is the precision. The optimal cell numbers defined by the recall have small differences from the t test results, but the scale and trend remain largely consistent.

In scenario B, the constraint is the total sequencing depth of one experiment with at least two cell states, and the goal is to determine the optimal total cell number for that experiment given a criterion in DE analysis. We apply scDesign to simulate a new count matrix of each cell state based on a real count matrix from the same state, with pre-specified total sequencing depth, total cell number, and cell proportions of the two cell states of interest. We applied scDesign to evaluate the designs of 12 example experiments (Table C.5). In every experiment, we set the sequencing depth to 100 million reads and considered six cell numbers: 512, 1024, 2048, 4096, 8192, and $16,384$. We estimated the cell proportions of the two cell states from the corresponding real data. In practical applications of scDesign, the cell state proportions can be inferred from pilot studies or public data [159, 161].

In contrast to scenario A, the optimal total cell number in scenario B depends on an additional factor: the cell state proportions, aside from the technical and biological factors we have discussed. The two cell states of interest may be present in various proportions depending on biological conditions and experimental protocols, and larger cell state proportions in general reduce the demand of a larger total cell number. For example, the estimated cell state proportions of astrocytes and oligodendrocytes in a human brain sample are 19.2% and 14.9%, respectively [230], and 1024 cells are needed to maximize the recall with Fluidigm C1 (Table C.5). In a mouse visual cortex sample, however, the estimated proportions of the same two cell types are 8.8% and 13.1%, respectively, and $16,384$ cells are required to achieve the highest recall with inDrop (Table C.5). Given an experimental protocol, the optimal total cell number depends on both the two cell state proportions and the magnitude of gene expression differences between the two cell states. For example, the proportions

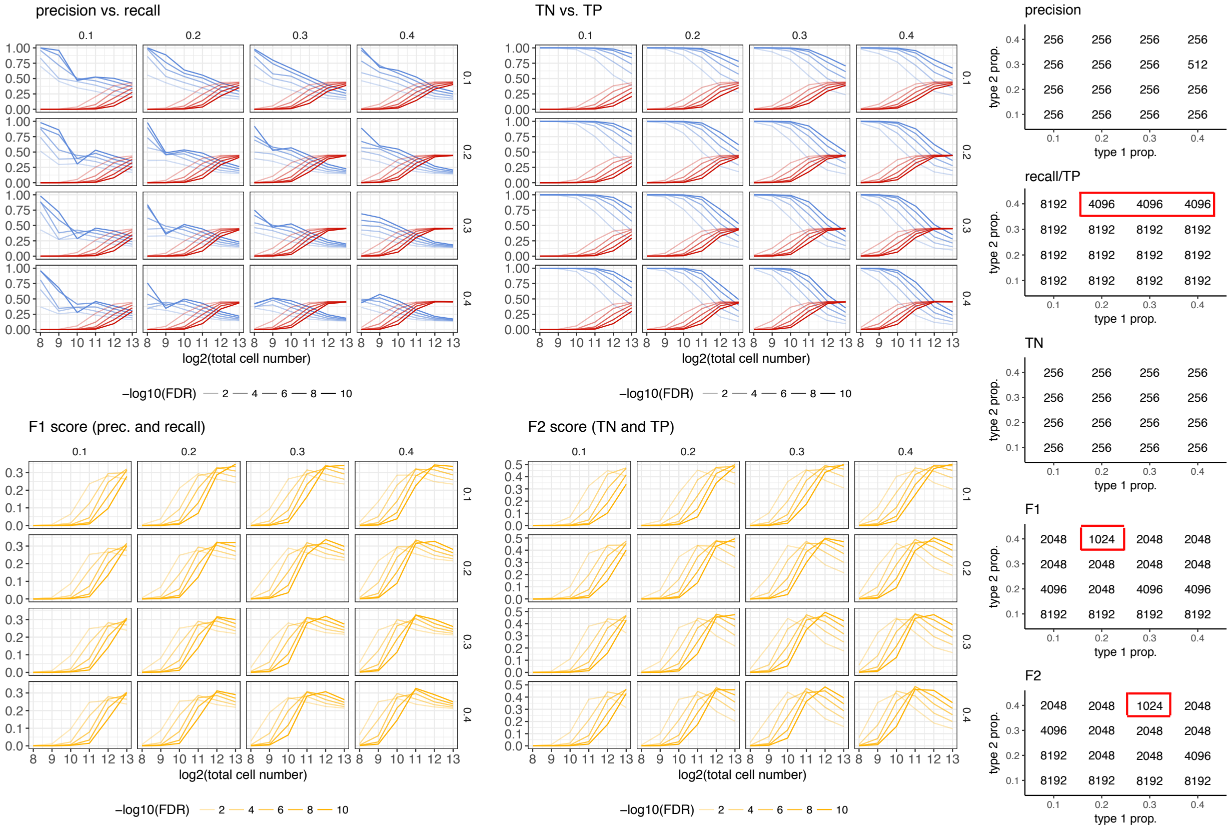


Figure 6.5: Power analysis for DE studies comparing CD4 and B cells with the Seq-Well protocol (scenario B). The thresholds on the FDRs to identify DE genes are denoted in the color legends. Different proportions (0.1, 0.2, 0.3, and 0.4) of CD4 and B cells were considered in the experimental design. For the three metrics of recall, F1, and F2, the smallest cell numbers leading to the best DE accuracy are marked in red boxes.

of CD4 cells, CD8 cells, and B cells in a human peripheral blood mononuclear sample are 17.2%, 10.2%, and 7.3%, respectively [161]. Two important facts about this experiment are: first, the proportion of CD8 cells is higher than the proportion of B cells; second, the magnitude of gene expression differences is larger between CD4 and B cells than between CD4 and CD8 cells. With the Seq-Well protocol, the DE analysis of CD4 vs. B cells only needs 4096 cells to achieve the highest F1 score. On the other hand, the DE analysis of CD4 vs. CD8 requires 16,384 cells to maximize the F1 score (Table C.5). To further assess the effect of cell state proportions on DE analysis, we synthesized CD4 and B cells with multiple hypothetical cell proportions: $10\%, 20\%, 30\%, \text{and } 40\%$ (Figure 6.5), among which the mixture of 40% B cells and 20-30% CD4 cells led to the minimum cell number required to maximize the recall and precision. Determining the optimal cell state proportions given a total cell number is especially useful when the cell states of interest can be enriched by fluorescence-activated cell sorting [228] or flow cytometry [231] before the sequencing step.

6.3.3 scDesign demonstrates reproducibility across studies

In addition to evaluating the results of scDesign across different cell types and scRNA-seq protocols, we also analyzed the experimental designs of the same cell types and protocols but different datasets, in attempt to assess the reproducibility of scDesign.

First, we applied scDesign to optimize the pairwise DE analysis between the oligodendrocyte precursor cells (OPCs) and three other brain cell types: differentiation-committed oligodendrocyte precursors (COPs), myelin-forming oligodendrocytes (MFOs), and newly formed oligodendrocytes (NFOs). Two real datasets were collected for each cell type, and the two datasets were generated using the Fluidigm C1 protocol but from different brain regions: dorsal horn and hypothalamus [232]. We applied scDesign in scenario A, assuming a total sequencing depth of 50 million reads for each cell type. In each experiment, we assumed that the libraries of the two cell types have the same number of cells, and we considered five candidate cell numbers per cell type: 64, 128, 256, 512, and 1024. The experimental designs based on the two brain regions lead to highly consistent results. Both designs show that in a DE analysis between OPCs and COPs, the optimal number of each cell type is 64 if selected by precision or true negative rate, 512 by F2 score, and 1024 by recall or F1 score (Figure 6.6); to better compare OPCs and MFOs or NFOs, the optimal number of each cell type is 64 if selected by precision, recall, or true negative rate, 128 by F2 score, and 64 by F1 score (Figure B.19). In fact, not only do the two designs identify the same optimal cell number in each case, but they also reveal highly consistent trends regarding how DE accuracy changes as the number of sequenced cells increases (Figures 6.6 and B.19). This example demonstrates the reproducibility of scDesign when taking input data from biological replicates.

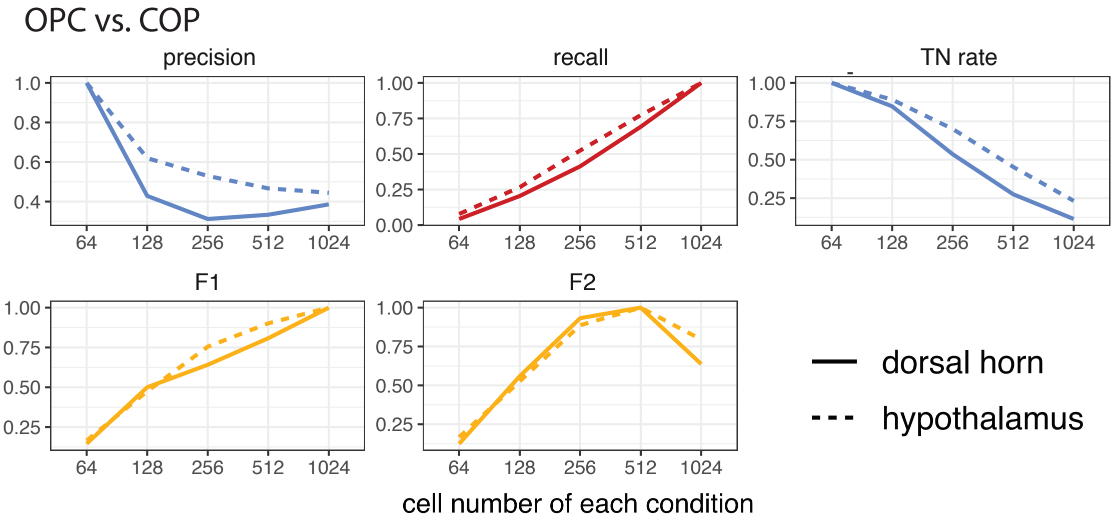


Figure 6.6: Reproducibility of scDesign based on data from different brain regions. The DE studies compared OPCs and COPs based on scRNA-seq data from two brain regions: dorsal horn and hypothalamus. When identifying the DE genes, the threshold set on the FDR rate was 10^{-10}. The y-axis of each line are divided by the maximum value of that line for normalization.

Second, we applied scDesign to optimize the pairwise DE analysis between three retina cell types: muller glia, amacrine, and rods. Two real datasets were collected from each cell type, and the two datasets were generated using the Drop-seq protocol in two independent studies [159, 229]. We applied scDesign in scenario A, assuming a total sequencing depth of 100 million reads for each cell type. In each experiment, we assumed that the libraries of the two cell types have the same number of cells and considered six candidate cell numbers per cell type: 64, 128, 256, 512, 1024, and 2048. The experimental designs based on the two single-cell studies lead to highly similar results (Figure B.20). As in the first example, both designs identified the same optimal cell number regardless of the DE criterion used. The only exception was in the comparison between the muller glia and rods: using the Macosko et al. data, the best cell number is 512 by F1 score and 1024 by F2 score, while using the Shekhar et al. data, the best cell number is 1024 by F1 score and 512 by F2 score. Such discrepancy is not suprising since we only evaluated a few candidate cell numbers, and the two input datasets inevitably differ in qualities as they came from two studies. Overall, this example demonstrates the reproducibility of scDesign when taking input data from independent studies to design scRNA-seq experiments.

6.3.4 scDesign assists scRNA-seq method development

In addition to assisting single-cell experimental design, scDesign can also simulate scRNA-seq data to benchmark various computational methods for differential gene expression analysis, single cell clustering analysis, dimensionality reduction of gene expression data, etc. Due to excess zeros resulting from dropout events and the fact that each gene's expression level in each cell is only measured once, the ground truth of individual genes' expression levels in single cells cannot be accurately estimated from scRNA-seq data. Also, cellular identities of individual cells are difficult to pre-determine in most experiments. Lacking the aforementioned ground truth encumbers the development of computational methods to decipher information from scRNA-seq data. Direct evaluation of computational methods relies on experimental validation, which is often unavailable for computationalists, and indirect interpretation from downstream analysis is used instead as a not-so-ideal substitute. Empowered by its ability to generate synthetic scRNA-seq data that well mimic real scRNA-seq data and have ground truth information, scDesign provides a flexible framework to benchmark computational methods for various scRNA-seq data analysis tasks.

We first demonstrate the application of scDesign to evaluating DE methods. We considered a baseline DE method, i.e., the two-sample t test, and four DE methods (MAST [192], SCDE [168], DESeq2 [30], and edgeR [28]) specifically designed for scRNA-seq data. Here both DESeq2 and edgeR denote their single-cell-adapted versions, where gene expression values are weighted by the weights estimated from a ZINB model before the statistical testing step [233]. We evaluated scDesign using real scRNA-seq data of six cell types: dendrocytes (Smart-seq2), oligodendrocytes (Fluidigm C1), interneurons (inDrop), retinal ganglions (Drop-seq), enterocytes (10x Genomics), and natural killer cells (Seq-Well). Based on the real data of each cell type, we simulated a pair of count matrices, with one matrix representing the given cell type and the other including up-regulated and down-regulated genes (Appendix A.9). In the first setting, we set the percentage of DE genes to 5% and sampled the fold changes of those DE genes' expression values uniformly from the interval $[2, 5]$. Then we evaluated the performance of the five DE methods by comparing the areas under their precision-recall curves (Figure B.21). With Smart-seq2 and Fluidigm C1, MAST and SCDE were the only two methods that achieved better accuracy than the two-sample t test, but overall the three methods had comparable precision and recall. With inDrop and 10x Genomics, edgeR became the best DE method, followed by MAST and SCDE. With Drop-seq and Seq-Well, the most accurate method was SCDE, and the baseline two-sample t test had poor performance. These simulation results suggest that scRNA-seq data from the 10x, inDrop, Drop-seq, and Seq-Well protocols need more specialized statistical modeling in the DE analysis, compared with Smart-seq2 and Fluidigm C1. In the second setting, we set the percentage of up-regulated and down-regulated genes in each comparison to 10% and sampled the fold changes of these DE genes uniformly from the interval $[4, 5]$. Since the magnitude of fold changes increased, the DE methods overall demonstrated improved accuracy, but the relative accuracy of the five DE methods was consistent with that under the first setting.

We next demonstrated the application of scDesign to comparing dimensionality reduction methods. We considered four methods: PCA, tSNE [14], independent component analysis (ICA) [234], and ZINB-WaVE [177]. We evaluated scDesign based on the same real scRNA-seq data used in the comparison of DE methods. Using the real data of each cell type, we simulated a set of synthetic count matrices, representing multiple cell states following a differentiation path (Appendix A.9). For the Smart-seq2 and Fluidigm C1 protocols, we simulated four cell states with two states each having 80 cells and the other two each having 50 cells. For the other four protocols, we simulated five cell states with two states each having 300 cells and the other three each having 100 cells. In each case, we first simulated the cell state at the starting point of differentiation based on the real data, and then we simulated each of the three subsequent cell states with 1% of up-regulated and down-regulated genes from its previous state. In addition, we sampled the fold changes of those DE genes' expression values uniformly from $[2, 5]$. After we applied the dimensionality reduction methods on each dataset, the hierarchical clustering method was applied on the first two dimensions, and the Jaccard index between the computed cell classes and the true cell conditions was calculated. Among the four dimensionality reduction methods, ZINB-WaVE had the best performance in grouping cells into biologically meaningful clusters based on the C1 data, followed by the tSNE method. Based on the Smart-Seq2 data, PCA had the best performance in the 2D space, followed by ZINB-WaVE. However, the comparison results were different for droplet-based protocols. The tSNE method led to the most accurate cell clusters for the Drop-seq, inDrop, 10x, and Seq-Well protocols, followed by ICA and PCA. In spite of the clustering performance, another factor worth noting is that PCA, ICA, and ZINB-WaVE generate comparable cell-cell distances after dimensionality reduction, but tSNE does not. The above results demonstrate the capacity of scDesign in helping developers evaluate competing computational methods for the same purpose (e.g., DE analysis or dimensionality reduction), and in selecting the appropriate method for analyzing scRNA-seq data from a specific protocol.

6.4 Discussion

The scRNA-seq technologies have become an essential tool for studying various biological and biomedical problems, but one unresolved challenge is how to balance the trade-off between exploring the depth or breadth of transcriptome information in experimental design. We introduce scDesign, the first statistical and computational simulator that enables rational and practical scRNA-seq experimental design. By integrating statistical assumptions and real scRNA-seq datasets from public repositories into its generative framework, scDesign is able to mimic the real experimental processes and simulate synthetic scRNA-seq datasets that well capture gene expression characteristics in real data. In addition, scDesign is a flexible and reproducible simulator that is capable of modeling protocol-specific scRNA-seq data generated under multiple biological and experimental conditions. We conducted a comprehensive comparison of scDesign and four other scRNA-seq simulation methods based on datasets from 17 different cell types and six experimental protocols. The comparison suggests that scDesign generates synthetic data with the largest resemblance to real scRNA-seq data regardless of cell types and protocols.

Using its simulated data, scDesign performs power analysis on DE analysis to provide a quantitative and objective standard for designing future experiments. In the context of differential gene expression analysis between two cell states, scDesign suggests an optimal cell number given a fixed sequencing depth, in the trade-off between a deeper sequencing of a smaller number of cells or a shallower sequencing of a larger number of cells. Specifically, we demonstrated the application of scDesign in two scenarios, where cells from the two states are sequenced as two separate libraries or as one pooled library. We evaluated the experimental designs for 14 and 12 scRNA-seq studies under the two scenarios, respectively. Our results for the first time demonstrate how the optimal experimental design for DE analysis depends on the scRNA-seq protocol and the intra and inter cell state transcriptome heterogeneity. In addition, our results revealed a general phenomenon that a deeper sequencing of a smaller number of cells leads to a higher precision in DE analysis. In contrast to the precision, maximizing the recall of DE analysis requires finding a balance between the cell-wise sequencing depth and the cell number, because our results show that the recall first increases and then decreases as we increase the cell number with the total sequencing depth fixed. The scDesign method enables researchers to design effective scRNA-seq experiments without pre-experimental costs in an objective manner, for example, guided by the expected power in downstream DE analysis. In addition, we demonstrate that scDesign leads to reproducible designs for target cell states given data generated in different studies.

Aside from enhancing future experimental designs, another contribution of scDesign is to assist computational method development for scRNA-seq data. Since large-scale benchmark data are not yet available, computationalists typically rely on scRNA-seq data from public repositories to evaluate new methods. However, quality control and normalization of real data are themselves ongoing research questions, making the results in many method papers not comparable nor reproducible [220,235]. To tackle this challenge, scDesign allows users to generate synthetic scRNA-seq data with user-specified experimental protocols, sequencing depths, cell states, cell numbers, as well as pre-specified differentially expressed genes. Given that scDesign generates synthetic data with known truth and well mimicking real data, users can leverage its synthetic data to comprehensively evaluate computational and statistical methods in a flexible, reproducible, and comparable way. For example, we compared five DE methods (the two-sample t test, MAST, SCDE, DESeq2, and edgeR) and four dimensionality reduction methods (PCA, tSNE, ICA, and ZINB-WaVE) using synthetic data generated by scDesign. Those comparison results provide useful guidance for researchers to select the most appropriate computational method to analyze real data.

We expect scDesign to assist scRNA-seq experimental design for a vast array of available experimental protocols. It incorporates real scRNA-seq data into its statistical framework to make flexible decisions based on the protocol and cell states used in the target study. If the real data of the two cell states are not generated from the same experiment, it is recommended to correct the batch effect before applying scDesign [236, 237]. To extend scDesign's ability to evaluate experimental designs for cell states whose scRNA-seq data are not yet publicly available, a future direction is to incorporate bulk RNA-seq data of the same type as a surrogate to estimate the gene expression parameters. Otherwise, pilot experiments need to be conducted to collect data for experimental design, which is also a widely adopted practice [238]. Another future extension of scDesign is to find the optimal design in the context of other types of downstream analyses besides the differential gene expression analysis, such as the detection of novel cell sub-types or the recovery of temporal transcriptome trajectories [218]. For instance, we may jointly learn the proportions and the gene expression profiles of multiple cell states from real scRNA-seq data and use them as input into our simulation framework to evaluate how the power of detecting rare cell types changes with experimental parameters. Given time-series scRNA-seq data, the scDesign framework can be modified to conduct ANOVA or more advanced statistical analysis to objectively select cell numbers for multiple time points. It is also possible to generalize the simulation framework of scDesign to account for more complex trajectories in the cell differentiation process. We expect scDesign to be an effective bioinformatic tool that assists rational scRNA-seq experiment design based on specific research goals and benchmarks competing scRNA-seq computational methods.

CHAPTER 7

Conclusions

Over the past decade, RNA sequencing technologies have revolutionized transcriptome analysis for interpreting genome function and investigating molecular bases for various disease phenomena. The bulk RNA-seq technologies enable the analysis of pooled populations of cells to understand the molecular constituents of cells and tissues, and have become an essential tool for understanding development and disease [1]. The recent few years have witnessed powerful advances in single cell sequencing technologies, allowing transcriptomic analysis to be carried out at a single-cell resolution. The emerging scRNA-seq technologies enable the quantification of transcriptomic dynamics in heterogeneous cell populations and complex tissues, offering unprecedented opportunities for investigating transcriptional mechanisms underlying organ development and disease [145]. This disseration aims to discuss the statistical challenges and potential solutions in the analysis of RNA-seq data, for better understanding both bulk and single-cell transcriptomics.

The first part of this dissertation focuses on discussing and developing statistical methods for bulk RNA-seq data. Chapter 1 provides an overview of the modeling and analysis of next-generation RNA-seq data from a statistical perspective. We summarize state-of-the-art computational methods for bulk RNA-seq data analysis at four different levels: sample, gene, transcript, and exon levels, and we discuss the key statistical considerations involved in these methods. We expect the discussion to be useful to both statisticians focusing on methodology development and applied bioinformaticians interested in understanding the commonly used statistical models.

Chapters 2 and 3 are devoted to addressing the statistical challenges involved in transcript-level analysis of RNA-seq data for more accurate isoform identification and quantification.

In Chapter 2, we propose a statistical model, MSIQ, to more accurately estimate isoform expression levels using multiple RNA-seq datasets. It tackles the challenge of accurate isoform quantification by aggregating information from multiple RNA-seq samples derived from the same biological condition. The unique advantage of MSIQ lies in its ability to identify and give higher weights to multiple RNA-seq samples that share similar transcriptome information, which we define as the consistent group. The MSIQ method not only provides a tool to measure isoform expression levels under different biological conditions, but also supports better reuse of public RNA-seq data by evaluating consistency of multiple samples from the same biological condition. In Chapter 3, we propose a statistical method, AIDE, to improve the precision of isoform discovery by selectively borrowing alternative splicing information from existing annotations. Based on a carefully constructed likelihood model, AIDE iteratively identifies isoforms in a stepwise manner while placing priority on the annotated isoforms, and it uses statistically principled tests to automatically filter the identified isoforms. We have demonstrated the superior performance of MSIQ and AIDE using both simulated and real RNA-seq datasets, compared to the state-of-the-art methods. In addition, the high precision of AIDE was confirmed based on a comparison to the long-read sequencing results and multiple experimental validations using PCR. We have also shown the ability of AIDE in identifying full-length isoforms with biological functions in pathological conditions, such as melanoma and breast cancer.

The MSIQ and AIDE models can be considered as umbrella frameworks that can be easily extended to incorporate more detailed modeling procedures to describe read generating processes using likelihood functions. The extension would allow flexible assumptions given additional experimental details or data features. As high-quality scRNA-seq data from full-length protocols become available, MSIQ and AIDE might be modified to answer transcript-level questions in single cell studies. Another future extension is to integrate MSIQ and AIDE to simultaneously identify and quantify mRNA isoforms from multiple RNA-seq samples. By aggregating information from multiple samples at the stage of isoform discovery, the joint model has the potential to further improve the accuracy and robustness of isoform expression analysis. We expect that the application of MSIQ and AIDE will help us better understand the isoform-level dynamics of RNA contents in different cells, tissues, and developmental stages. The greater accuracy of the identified isoforms and their estimated expression levels will shed light on transcriptional regulatory mechanisms of genetic diseases, thus assisting biomedical researchers in designing targeted therapies.

The second part of this dissertation focuses on discussing and developing statistical methods for single-cell RNA-seq data. Chapter 4 provides an overview of the currently available scRNA-seq technologies and summarizes common statistical analyses for scRNA-seq data, including dimensionality reduction and feature selection, clustering analysis to identify cell subpopulations, and differential gene expression analysis. The dicussion reveals a need for new statistical tools to assist rational experimental design and to help denoise scRNA-seq data, in order to make reliable and reproducible scientific discoveries.

Since computational analyses are complicated by the dropout events prevalent in scRNA-seq data, we propose the scImpute method in Chapter 5 to impute the missing gene expression levels resulting from technical limitations. A key feature of scImpute is that it focuses on imputing the missing expression values of dropout genes, while retaining the expression levels of genes that are largely unaffected by dropout events. To achieve this goal, scImpute first learns each gene's dropout probability in each cell by fitting a mixture model for each cell subpopulation. Next, scImpute imputes the highly probable dropout values of genes in a cell by borrowing information of the same genes' expression in other similar cells. We have demonstrated using both simulated and real scRNA-seq data that scImpute is able to recover gene expression levels affected by dropout events, improve the identification of cell subpopulations, and assist differential gene expression analysis. Following the development of scImpute, new computational methods have been proposed for imputation of scRNA-seq data. Given that large-scale, error-free scRNA-seq data are not yet available for benchmarking, it remains critical to assess the performance of computational methods from perspectives that have biologically meaningful interpretations. As scRNA-seq data of improved quality become available, we will be better equipped to perform comprehensive and fair comparisons of scRNA-seq computational methods.

In Chapter 6, we introduce scDesign, the first statistical simulator that enables rational scRNA-seq experimental design, to address the unresolved challenge about how to balance the trade-off between exploring the depth or breadth of transcriptome information in scRNA-seq experiments. The scDesign method provides a flexible and reproducible simulator that is capable of modeling protocol-specific scRNA-seq data generated under multiple biological conditions, and we have shown that it is able to mimic the real experimental processes and simulate synthetic scRNA-seq data that well capture gene expression characteristics in real data. In addition, scDesign is able to suggest an optimal cell number given a fixed sequencing depth, in the trade-off between a deeper sequencing of a smaller number of cells or a shallower sequencing of a larger number of cells, in the context of differential gene expression analysis between two cell states. Future extension of scDesign is to find the optimal design for other biological questions besides the identification of DE genes, such as the detection of novel cell subtypes or the recovery of temporal transcriptome trajectories.

In conclusion, we have discussed four statistical challenges and our proposed solutions for the analysis of bulk or single-cell RNA-seq data. We believe that these problems demonstrate the essential role of statistics in the interpretation of large-scale genomic data. Multiple statistical and computational methods are often available for the same analysis purpose of RNA-seq data, but there is hardly a method that is superior in every application scenario. Therefore, we would like to emphasize the necessity for method developers to demonstrate the reproducibility and interpretability of new methods. This will help the users apply these methods in proper scenarios and ensure that the key statistical assumptions are not violated, leading to biologically meaningful discoveries. As sequencing technologies develop, the collection of further data will enable advances in statistical methods to help answer emerging biological and biomedical questions.

APPENDIX A

Supplementary text

A.1 Approaches to RNA-seq data summary

The RNA-seq reads are mapped to the reference genome, represented by genomic positions covered by each read. That is, if a read has a total length $2c$ (the left and right end each have length c), it is represented by a set of genomic positions $\{y_1, \ldots, y_{2c}\}$. However, efficient data summary is needed to preserve most relevant information for isoform quantification while controlling the computational complexity at manageable level [100]. To our knowledge, there are three existing approaches designed for data summary:

- Approach 1 [53]: represent the read by $(y_1, y_{2c})'$;
- Approach 2 [62]: represent the read by $(y_1, y_c, y_{c+1}, y_{2c})'$;
- Approach 3 [100]: represent the read by $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, the indices of the exons that completely or partially overlap with each end of the read:

$$\begin{aligned} \boldsymbol{s}_1 &= \{i : \boldsymbol{G}_i \cap \{y_1, \ldots, y_c\} \neq \emptyset, i = 1, \ldots, N\}, \\ \boldsymbol{s}_2 &= \{i : \boldsymbol{G}_i \cap \{y_{c+1}, \ldots, y_{2c}\} \neq \emptyset, i = 1, \ldots, N\}, \end{aligned} \tag{A.1}$$

  where $\boldsymbol{G}_i$ denotes the set of genomic positions in exon i .

We compare these three approaches on the hypothetical gene in Figure 2.2 and summarize their results in Table A.1. It is obvious that Approach 1 loses more information than Approach 2. In this specific case, Approach 3 gives clearer indication of the possible isoform origins of the three reads, because it captures the mapping information inside the left and right ends, which are missed by Approaches 1 and 2. However, Approach 3 removes the actual genomic positions of reads, which are necessary for estimating the fragment length corresponding to each paired-end read. Recognizing the comparative advantages of Approaches 2 and 3, we use a combination of Approaches 2 and 3 as our data summary method.

Table A.1: Comparison of the three RNA-seq data summary approaches. Examples of three summarized reads and their encoded isoform origin information are listed.

Read	Approach 1		Approach 2		Approach 3	
	data	origin	data	origin	data	origin
1	(231,559)	1,2	(231,280,510,559)	1,2	{1}{4}	1,2
2	(100,578)	1,2	(100,199,460,578)	2	{1}{3,4}	2
3	(50,537)	1,2	(50,149,370,537)	1,2	{1}{2,3,4}	2
⋮	⋮		⋮		⋮	

A.2 Proof of Lemma 2.1

We first introduce some results from [239] to assist the proof of Lemma 2.1.

Theorem A.1. Suppose $(\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(t)}, \dots)$, $\boldsymbol{X}^{(t)} \in O \subseteq \mathbb{R}^n$ be a Markov chain with transition kernel K w.r.t a σ-finite measure ν, and π is an invariant distribution of this Markov chain. If the transition kernel K is π-irreducible and aperiodic, then for all $\boldsymbol{x} \in Q = \{\boldsymbol{x} \in O : \pi(\boldsymbol{x}) > 0\}$, as $t \to \infty$,

(i) $|K^{(t)}(\boldsymbol{x}, \cdot) - \pi| \to 0$, where $K^{(t)}(\boldsymbol{x}, \cdot)$ is the density of $\boldsymbol{x}^{(t)}$ given $\boldsymbol{X}^{(0)} = \boldsymbol{x}$;

(ii) for real-valued, π-integrable function f,

$$\frac{1}{t} \sum_{i=1}^{t} f\left(\boldsymbol{x}^{(i)}\right) \to \int_O f(\boldsymbol{x}) \pi(\boldsymbol{x}) d\nu(\boldsymbol{x}) \text{ a.s..}$$

Lemma A.1. If ν is discrete, then K is well-defined and π-irreducibility of K is a sufficient condition for the results of Theorem A.1.

Lemma A.2. If ν is n-dimentional Lebesgue measure, and π is lower semi-continuous at 0, then K is well-defined.

With the results introduced above, we can prove Lemma 2.1.

Proof. For simplicity, let $\boldsymbol{X} = (\boldsymbol{E}, \boldsymbol{Z}, \gamma)'$, a vector of $n = D + \sum_{d=1}^{D} n_d + 1$ dimensions, where $\boldsymbol{E} = (E_1, \ldots, E_D)'$ and $\boldsymbol{Z} = (Z_1^{(1)}, \ldots, Z_{n_1}^{(1)}, \ldots, Z_1^{(D)}, \ldots, Z_{n_D}^{(D)})'$. We denote the joint density of $\boldsymbol{X}$ as $\pi(\boldsymbol{x}|\boldsymbol{R}, \boldsymbol{\lambda}, a, b)$. In MSIQ, the transition kernel of the Markov chain is formed by the Gibbs sampler: $K(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)}) = \prod_{l=1}^{n} \pi \left(x_l^{(t+1)} | \{x_j^{(t)}\}_{j>l}, \{x_k^{(t+1)}\}_{k<l} \right)$.

We know that $\pi(\boldsymbol{x}|\boldsymbol{R}, \boldsymbol{\lambda}, a, b)$ is discrete w.r.t. X_i ($i = 1, 2, \ldots, n-1$) and is continuous w.r.t. X_n. According to Lemmas A.1 and A.2 [239], $K(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})$ is a well-defined kernel and π is an invariant distribution of the Markov chain applied by K. We also know that the conditional probabilities of $X_d = E_d$ ($d = 1, \ldots, D$) or $X_n = \gamma$ on everything else are always positive. For each $X_{D+i+\sum_{k=1}^{d-1} n_k} = Z_i^{(d)}$, if we define its conditional distribution (2.15) on the domain $\{j : q_{ij}^{(d)} \neq 0\}$, its conditional probability is also always positive. The positive properties of these conditional distributions naturally lead to the positivity of the transition kernel $K(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})$, and thus the π-irreducibility of K.

Now since $K(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})$ is π-irreducible, by Theorem A.1 for all $\boldsymbol{x} \in Q = \{\boldsymbol{x} : \pi(\boldsymbol{x}|\boldsymbol{R}, \boldsymbol{\lambda}, a, b) > 0\}$, $K^{(t)}(\boldsymbol{x}, \cdot)$ converges to $\pi(\boldsymbol{x})$: $|K^{(t)}(\boldsymbol{x}, \cdot) - \pi(\boldsymbol{x}|\boldsymbol{R}, \boldsymbol{\lambda}, a, b)| \to 0$ as $t \to \infty$, where $K^{(t)}(\boldsymbol{x}, \cdot)$ is the density of $\boldsymbol{x}^{(t)}$ given $\boldsymbol{X}^{(0)} = \boldsymbol{x}$. Therefore, $\hat{\theta}_d^{\mathrm{MSIQ}}$ converges to $\mathbb{E}(E_d|\boldsymbol{R}, \boldsymbol{\lambda}, a, b)$:

$$\lim_{T\to\infty} \hat{\theta}_d^{\mathrm{MSIQ}} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E_d^{(t)} = \mathbb{E}(E_d|\boldsymbol{R}, \boldsymbol{\lambda}, a, b). \tag{A.2}$$

Now we can prove the convergece of $\hat{\boldsymbol{\alpha}}^{\mathrm{MSIQ}}$. From model (2.9) it is easy to see that $\boldsymbol{\alpha}|\boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{R}, \lambda, a, b \sim \text{Dirichlet}(\lambda_1', \ldots, \lambda_J')$, where $\lambda_j' = \lambda_j + \sum_{d=1}^{D} \left(E_d \sum_{i=1}^{n_d} I_{i,j}^{(d)} \right)$. Thus the conditional posterior mean of α_j:

$$\mathbb{E}(\alpha_j|\boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{R}, \lambda, a, b) = \mathbb{E}(\alpha_j|\boldsymbol{X}, \boldsymbol{R}, \lambda, a, b) = \frac{\lambda_j + \sum_{d=1}^{D} \left(E_d \sum_{i=1}^{n_d} I_{i,j}^{(d)} \right)}{\sum_{j=1}^{J} \lambda_j + \sum_{d=1}^{D} E_d n_d} \triangleq g_j(\boldsymbol{X}). \tag{A.3}$$

Then the isoform proportion $\alpha_j^{(t)}$ defined in equation (2.17) can be written as: $\alpha_j^{(t)} = g_j(\boldsymbol{X}^{(t)})$. As a result,

$$\lim_{T\to\infty} \hat{\alpha}_j^{\mathrm{MSIQ}} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \alpha_j^{(t)} = \mathbb{E}_\pi[g_j(\boldsymbol{X})] = \mathbb{E}(\alpha_j|\boldsymbol{R}, \lambda, a, b), \tag{A.4}$$

and $\lim_{T\to\infty} \hat{\boldsymbol{\alpha}}^{\mathrm{MSIQ}} = \mathbb{E}(\boldsymbol{\alpha}|\boldsymbol{R}, \lambda, a, b)$. □

A.3 Sensitivity analysis of the MSIQ model

When calculating the MSIQ estimator in Chapter 2.3.1, we set the hyper-parameters as $a = 7$ and $b = 2$. These two parameters will influence the probability of assigning an RNA-seq sample to the consistent group. In order to illustrate that MSIQ is robust to these parameters, we present a sensitivity analysis here.

We randomly selected 500 genes out of the 3421 fly genes and carried out simulation based on these fly genes. For each gene, we considered five scenarios and independently simulated ten RNA-seq samples each with 500 paired-end reads by following the procedure described in Chapter 2.3.1. We estimated the isoforms proportions of these genes based on different values of a and b, and then compare the REE in different settings. In the calculation, we either fixed $a = 7$ and vary b among $\{2, 3, 4, 5, 6\}$, or d $b = 2$ and vary a among $\{4, 5, 6, 7, 8, 9, 10\}$. Meanwhile, the fragment length was set as 150 bp and the read length was set as 50 bp. As shown in Figure A.1, the REE rates in all five scenarios are largely invariant when parameters a and b take different values. This result justifies the robustness of MSIQ to the prior parameters.

A.4 Simulation for comparing isoform reconstruction methods

We considered 18,960 protein-coding genes from the human GENCODE annotation (version 24) [75]. For each gene, we set the proportions of isoforms not in the GENCODE database to 0. As for the annotated isoforms in GENCODE, their isoform proportions were simulated from a symmetric Dirichlet distribution with parameters $(1/\lceil \frac{J}{2} \rceil, \ldots, 1/\lceil \frac{J}{2} \rceil)'$, where J denotes the number of annotated isoforms for a given gene. When simulating the RNA-seq reads, we treated these simulated proportions as the pre-determined ground truth. Next, for each target read coverage among the eight choices (10x, 20x, ... , 80x), we used the R package polyester [117] to simulate one RNA-seq sample given the pre-determined isoform proportions. All the simulated RNA-seq samples contained paired-end reads with 100 bp length.

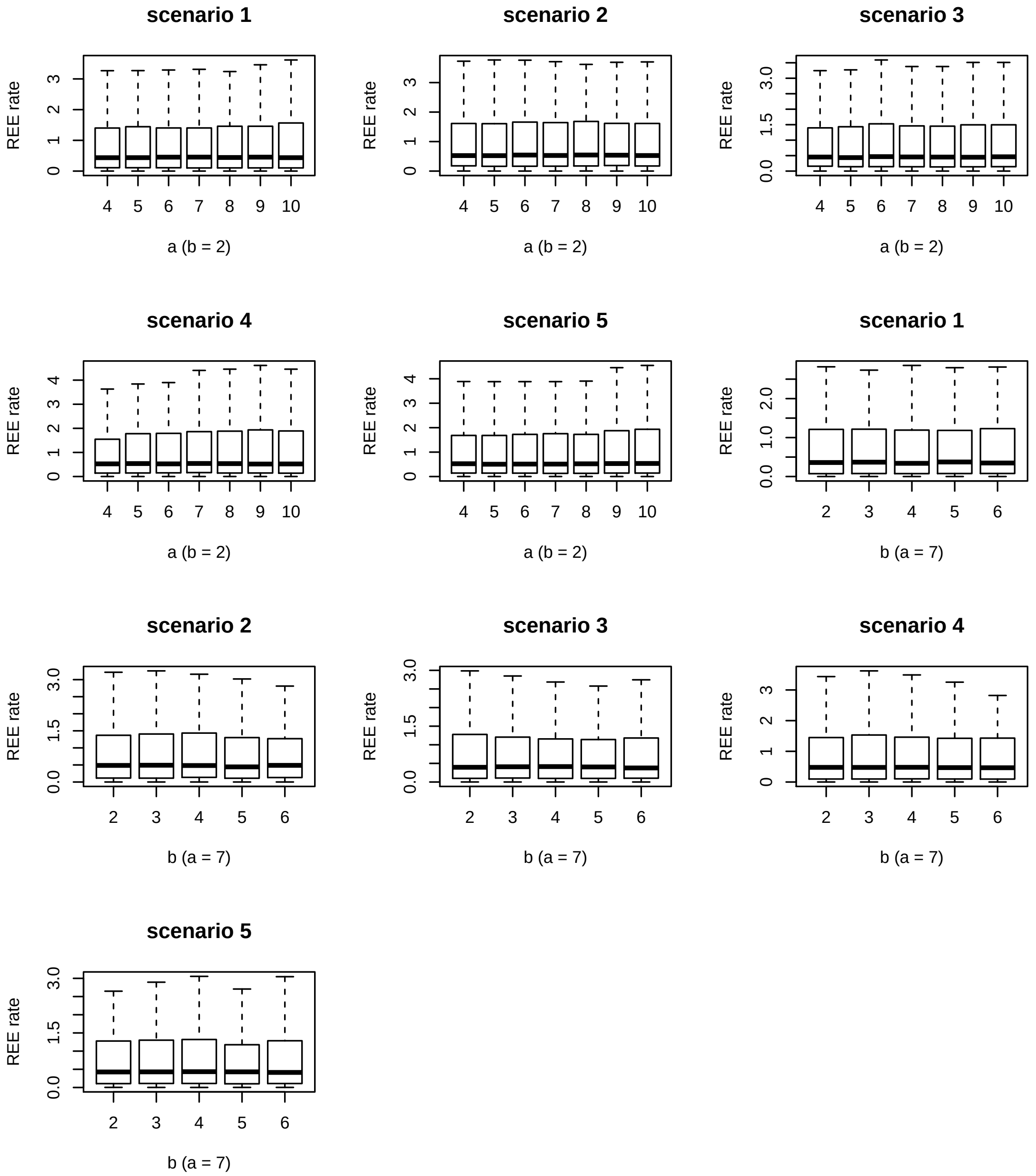


Figure A.1: REE of MSIQ on 500 fly genes in scenarios 1-5. The first five boxplots corresponde to the case when $b = 2$ and a varies between $\{4, 5, 6, 7, 8, 9, 10\}$; the last five boxplots correspond to the case when $a = 7$ and b varies between $\{2, 3, 4, 5, 6\}$.

A.5 Calculation of precision and recall in isoform reconstruction

We denote the true isoform set of an RNA-seq sample as S and the discovered isoform set as D. We use $B(s)$ and $E(s)$ to denote the number of bases and the number of exons in isoform s, respectively. We calculated the isoform-level precision and recall of isoform discovery by comparing the two sets S and D. When comparing the isoforms s_1 in S and s_2 in D, we allowed a small difference by requiring the number of mismatch bases to be smaller than $0.99B(s_1)$. The isoform-level precision and recall were calculated as

$$\text{precision}^{\text{isoform}} = \frac{|D \cap S|}{|D|}, \quad \text{recall}^{\text{isoform}} = \frac{|D \cap S|}{|S|}. \tag{A.5}$$

The exon-level precision and recall were calculated as

$$\text{precision}^{\text{exon}} = \frac{\sum_{s \in D \cap S} E(s)}{\sum_{s \in D} E(s)}, \quad \text{recall}^{\text{exon}} = \frac{\sum_{s \in D \cap S} E(s)}{\sum_{s \in S} E(s)}. \tag{A.6}$$

The base-level precision and recall rates were calculated as

$$\text{precision}^{\text{base}} = \frac{\sum_{s \in D \cap S} B(s)}{\sum_{s \in D} B(s)}, \quad \text{recall}^{\text{base}} = \frac{\sum_{s \in D \cap S} B(s)}{\sum_{s \in S} B(s)}. \tag{A.7}$$

A.6 Validation of isoforms by PCR and Sanger sequencing

Biopsy was collected freshly and the total RNAs were extracted with the RiboPure Kit (Ambion). The reverse transcription reactions were performed using the RevertAid First-Strand cDNA Synthesis System kit (ThermoFisher) . With the cDNAs as templates and primers from TSINGKE, the PCR procedure (95°C 5 min, 95°C 30 s, 55°C 30 s, 72°C 1 min to 2 min, 40 to 50 cycles, 72°C 5 min for extension) was conducted using a ThermoFisher PCR system. PCR products were purified using Gel Extraction Kit (OMEGA) followed by Sanger sequencing with their special forward primers. Finally, the sequencing results were analyzed using the Sequence Scanner software.

A.7 Validation of isoform functions by colongenic assay

Breast cancer cell lines BT549, MB231, SUM149, BT474, SK-BR-3, and MCF-7 were cultured in the State Key Laboratory of Biotherapy, West China Hospital. The cells' total RNAs were extracted with the RiboPure Kit (Ambion), and the reverse transcription reactions were performed using the RevertAid First-Strand cDNA Synthesis System kit (ThermoFisher). With the cDNAs as templates and primers from TSINGKE, the PCR procedure (95°C 5 min, 95°C 30 s, 55°C 30 s, 72°C 1 min, 40 cycles, 72°C 5 min for extension) was conducted using a ThermoFisher PCR system. PCR products were purified using the Gel Extraction Kit (OMEGA) followed by Sanger sequencing with special forward primers. The five siRNAs specifically targeting *FGFR1-238* were synthesized at Shanghai GenePharma. The cells' colonegenic assays lasted for 10-12 days.

A.8 AIDE improves isoform discovery via stepwise selection

We conducted a proof-of-concept simulation study to verify the efficiency and accuracy of our proposed AIDE method. We used this study to show why simply performing forward selection is insufficient and how stepwise selection leads to more precise and robust isoform discovery results. Here we considered $2,262$ protein-coding genes from the human GENCODE annotation (version 24) [75]. We treated the annotated isoforms as the true isoforms and simulated paired-end RNA-seq reads from those isoforms with pre-determined abundance levels. For every gene, we applied AIDE, which uses stepwise selection, and its counterpart AIDEf, which only uses forward selection, to discover isoforms from the simulated reads.

To evaluate the robustness of AIDE to the accuracy of annotation, we considered three types of annotation sets: (1) "N" (no) annotations: no annotated isoforms were used; (2) "I" (inaccurate) annotations: the annotated isoforms consisted of half of the randomly selected true isoforms and the same number of false isoforms; (3) "A" (accurate) annotations: the annotated isoforms consisted of half of the randomly selected true isoforms.

The simulation results averaged from the $2,262$ genes show that AIDE and AIDEf perform the best when the “A” annotations are supplied, and they have the worst results with the “N” annotations, as expected (Figure A.2). Given the “A” annotations, AIDE and AIDEf have similarly good performance. However, when supplied with the “I” or the “N” annotations, AIDE has much better performance than AIDEf. Notably, the performance of AIDE with the “I” annotations is close to that with the “A” annotations, demonstrating the robustness of AIDE to inaccurate annotations. On the other hand, AIDEf has decreased precision when the “I” annotations are supplied because forward selection is incapable of removing non-expressed annotated isoforms from its identified isoform set. Given the “N” annotations, AIDE also has better performance than AIDEf. These results suggest that choosing stepwise selection over forward selection is reasonable because perfectly accurate annotations are usually not available in real scenarios.

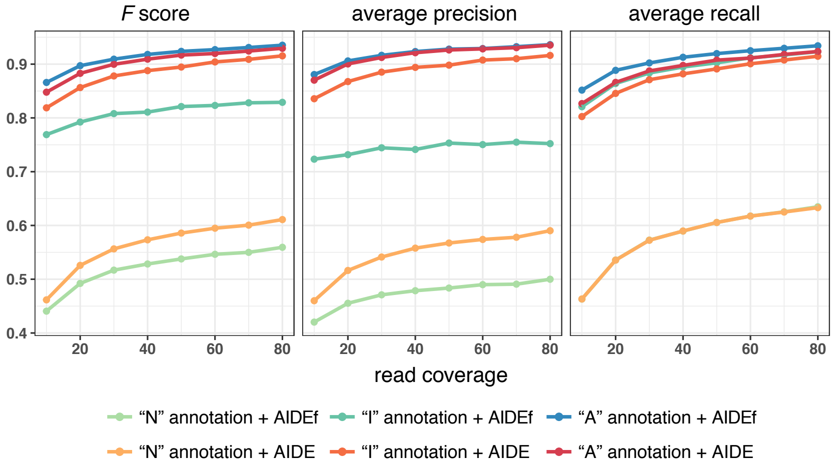


Figure A.2: Comparison of AIDE (stepwise selection) and AIDEf (forward selection only) across the $2,262$ genes in simulation. For AIDE and AIDEf, each measure is calculated based on three types of annotations and RNA-seq samples with varying read coverages. The horizontal axis denotes the average per-base coverage of RNA-seq reads.

Figure A.2 also suggests that both approaches exhibit improved performance with all the three types of annotations as the read coverages increase. Higher read coverages help the most when the “N” annotations are supplied, but its beneficial effects become more negligible with the “A” annotations. When the coverages increase from 10x to 80x, the F scores of AIDE increase by 32.6%, 12.3%, and 9.5% with the “N”, “I”, and “A” annotations, respectively. Moreover, we observe that the F scores of AIDE with the "N" annotations and the 80x coverage are approximately 30% lower than the F scores with the "A" annotations and the 30x coverage. This suggests that accurate annotations can assist isoform discovery and reduce the costs for deep sequencing depths to a large extent.

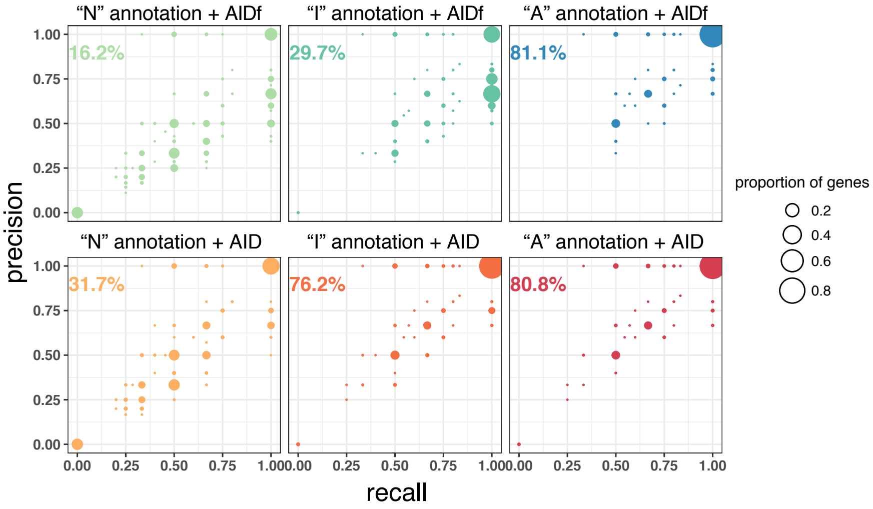


Figure A.3: Comparison of AIDE (stepwise selection) and AIDEf (forward selection only) in terms of the per-gene precision and recall in simulation, with 80x read coverage and three types of annotations. The circle sizes are proportional to the fractions of genes with the corresponding precision and recall.

We also summarized the precision and recall rates of AIDE at the individual gene level (Figure A.3). When the "A" annotations are supplied, both AIDE and AIDEf achieve 100% precision and recall for over 80% of the genes. When the "N" or "I" annotations are supplied, we observe a 2.0- or 2.6- fold increase in the number of genes with 100% precision and recall from AIDEf to AIDE. These results again demonstrate the effectiveness of AIDE in removing non-expressed annotated isoforms and identifying novel isoforms with higher accuracy due to its use of statistical model selection principles.

A.9 Simulating count matrices following a differentiation path

Given a real dataset with I genes and J_0 cells, the goal of this section is to generate G ($G \geq 2$) new count matrices, each of which has I genes, J synthetic cells, and a total of S reads. The synthetic data should represent G cell states following a specified differentiation path with known DE genes, such that these data serve as a good basis for benchmarking single-cell data analysis and method development. When generating the G synthetic count matrices, we assume that the G cell states follow a differentiation path, with a p_{up} proportion of up-regulated genes and a p_{down} proportion of down-regulated genes from state g to state $g+1$ ($g = 1, \ldots, G-1$).

(1). Estimate parameters from real scRNA-seq data.

As described in Chapter 6.2.2, from the real count matrix $\boldsymbol{X}^{\text{real}}_{I\times J_0}$, we obtain the following parameter estimates: (1) the mean $\hat{\mu}_s$ and the standard deviation $\hat{\sigma}_s$ of the Normal distribution used to model the cell library sizes; (2) the cell-wise dropout rates $\hat{q}_{01}, \ldots, \hat{q}_{0J_0}$; (3) the gene-wise dropout rate $\hat{\lambda}_{0i}$, mean $\hat{\mu}_{0i}$, and standard deviation $\hat{\sigma}_{0i}$ of gene i, $i = 1, \ldots, I$. A Gamma distribution is used to fit the estimated gene mean expression $\hat{\mu}_{01}, \ldots, \hat{\mu}_{0I}$, and the estimated shape and scale parameters are denoted as $\hat{k}_0$ and $\hat{\theta}_0$, respectively. The above parameter estimates are then used to simulate the expression parameters of state 1, while the parameters of state $g+1$ depended on the parameters of its previous state g.

(2). Simulate gene mean expression values of the G states.

In this step, we simulate the log-scale mean gene expression values under each cell state, without considering dropout events. We assume that from state g to state $g+1$, the proportions of up-regulated and down-regulated genes are p_{up} and p_{down}, respectively. The fold changes of gene mean expression levels are independently and uniformly distributed within $[f_l, f_u]$.

We used μ_i^g to denote the mean expression of gene i in cell state g. For cell state 1, we simulate μ_i^1 from the Gamma distribution: $\mu_i^1 \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\hat{k}_0, \hat{\theta}_0)$, $i = 1, \ldots, I$. Then given $\mu_1^g, \ldots, \mu_I^g$, we simulate $\mu_1^{g+1}, \ldots, \mu_I^{g+1}$ ($g = 1, \ldots, G-1$) as follows. We first simulate the number of up-regulated genes n^g_{up}, and the number of down-regulated genes n^g_{down} from a Multinomial distribution: $\left(n^g_{\text{up}}, n^g_{\text{down}}, I - n^g_{\text{up}} - n^g_{\text{down}}\right) \sim \mathrm{M}\left(I, (p_{\text{up}}, p_{\text{down}}, 1 - p_{\text{up}} - p_{\text{down}})\right)$. We randomly draw the $n^g_{\text{up}} + n^g_{\text{down}}$ DE genes from the gene population $\{1, \ldots, I\}$ without replacement and denote

$$d^g_i = \begin{cases} 1, & \text{if gene } i \text{ is up-regulated} \\ -1, & \text{if gene } i \text{ is down-regulated} \cdot \\ 0, & \text{otherwise} \end{cases} \tag{A.8}$$

Then, we simulate μ^{g+1}_I, the mean expression of gene i in state $g + 1$:

$$\mu^{g+1}_i = \begin{cases} \mu^g_i + \log_{10} f^g_i & \text{if } d^g_i = 1 \\ \mu^g_i - \log_{10} f^g_i & \text{if } d^g_i = -1 \;, \\ \mu^g_i, & \text{otherwise} \end{cases} \tag{A.9}$$

where $f^g_i \overset{\text{i.i.d}}{\sim} \text{Uniform}[f_l, f_u])$.

(3). Simulate the count matrices.

With the mean gene expression $\mu^g_1, \ldots, \mu^g_I$, we simulate the count matrix $\boldsymbol{X}^{\text{syn},g}$ under each state g independently following steps (2)-(4) in Chapter 6.2.2.

A.10 Power analysis of DE detection with scDesign

We have introduced two scenarios in experimental designs. If the two cell states are sequenced separately, the design needs specification of the sequencing depth S and the cell numbers J_1 in state 1 and J_2 in state 2. If the two cell states are sequenced together, the design needs specification of the sequencing depth S and the total cell number J. The goal of power analysis is to determine the best choice of cell number(s) to optimize the downstream DE analysis between two cell states, given a fixed S.

Given $\boldsymbol{X}^{\text{real1}}_{I \times J_{01}}$ and $\boldsymbol{X}^{\text{real2}}_{I \times J_{02}}$ from two different cell states, for each gene i we estimated its mean expression values in the mixture model as $\hat{\mu}^g_{0i}$ and $\hat{\sigma}^g_{0i}$ for state g ($g = 1, 2$) (see Chapter

6.2.2). Then we calculated an effect score of gene i to denote its differential expression strength: $h_i = {}^{|\hat{\mu}^1_{0i} - \hat{\mu}^2_{0i}|}/_{(\hat{\sigma}^1_{0i} + \hat{\sigma}^2_{0i})}$. The top N genes with the largest h_i's are used as the true DE genes to be compared with the detected DE genes from the simulated data, and this gene set is denoted as A^0.

Given an experimental design, we simulated B count matrices $\{\boldsymbol{X}^{\text{syn},11}, \ldots, \boldsymbol{X}^{\text{syn},1B}\}$ for cell state 1, and B count matrices $\{\boldsymbol{X}^{\text{syn},21}, \ldots, \boldsymbol{X}^{\text{syn},2B}\}$ for cell state 2. By performing DE analysis on $\boldsymbol{X}^{\text{syn},1b}$ and $\boldsymbol{X}^{\text{syn},2b}$, we identified a DE gene set A^b. Denoting the gene population set as Ω, we calculated five accuracy metrics: precision (a^b_1), recall (a^b_2), true negative rate (a^b_3), F1 (a^b_4), and F2 (a^b_5):

$$a^b_1 = \frac{|A^0 \cap A^b|}{|A^b|},\ a^b_2 = \frac{|A^0 \cap A^b|}{|A^0|},\ a^b_3 = 1 - \frac{|A^b \setminus A^0|}{|\Omega \setminus A^0|},\ a^b_4 = 2\frac{a^b_1 \times a^b_2}{a^b_1 + a^b_2},\ a^b_5 = 2\frac{a^b_2 \times a^b_3}{a^b_2 + a^b_3}\,. \tag{A.10}$$

Then we averaged each of the five metrics calculated over the B sets of data as $a_i = \frac{1}{B}\sum_{b=1}^{B} a^b_i$, $i = 1, \ldots, 5$. Finally, we repeated the above steps for each candidate cell number and selected the cell number that maximizes the user-specified metric among the five metrics.

For analyses presented in Chapter 6.3, we set $N = 1000$ and $B = 100$. The DE method used in the simulation were the two sample t test, which was applied to the non-zero gene expression values, and MAST [192]. In real data applications, users are suggested to use the DE method of their choice for the experimental design.

APPENDIX B

Supplementary figures

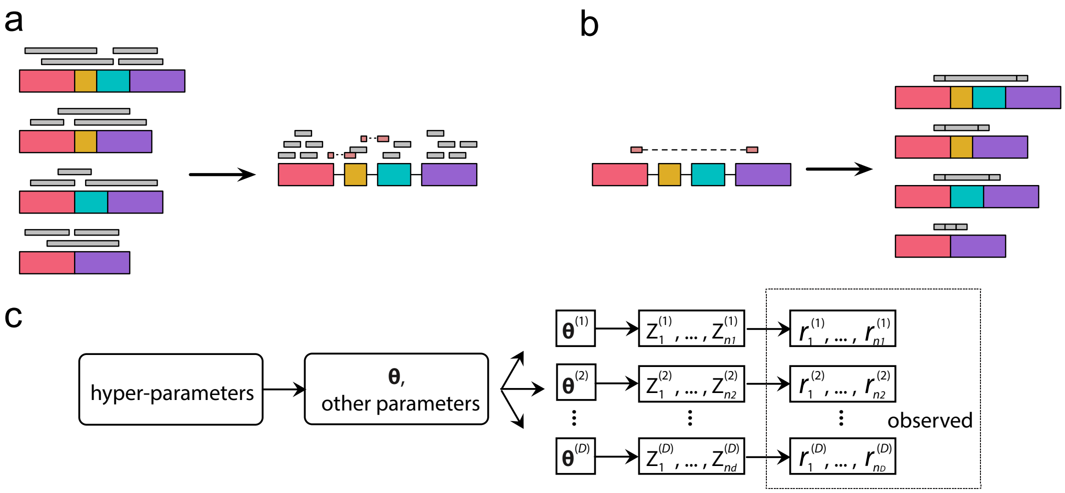

a
b
c
hyper-parameters
$\boldsymbol{\theta}$, other parameters
$\boldsymbol{\theta}^{(1)}$
$Z_1^{(1)}, \dots, Z_{n1}^{(1)}$
$r_1^{(1)}, \dots, r_{n1}^{(1)}$
$\boldsymbol{\theta}^{(2)}$
$Z_1^{(2)}, \dots, Z_{n2}^{(2)}$
$r_1^{(2)}, \dots, r_{n2}^{(2)}$
observed
$\boldsymbol{\theta}^{(D)}$
$Z_1^{(D)}, \dots, Z_{nd}^{(D)}$
$r_1^{(D)}, \dots, r_{nD}^{(D)}$

Figure B.1: Illustration of isoform reconstruction and quantification challenges. **a**: Taken this 4-exon gene as an example, the observed RNA-seq reads were sequenced from fragments of the true but unobservable isoforms. The read length is fixed in each experiment, but the fragment lengths can vary. Since only the two ends of each fragment are sequenced as paired-end reads, this leads to information loss in RNA-seq experiments. **b**: Given the paired-end reads mapped to the 4-exon gene (one end mapped to the first exon and the other end mapped to the fourth exon), the inferred fragment length could be different when assuming different isoform origin of the read. **c**: An example Bayesian framework to estimate the population isoform proportions $\boldsymbol{\theta}$ of a gene given D samples. $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(D)}$ are considered as the realization of $\boldsymbol{\theta}$ in D samples. $Z_i^{(d)}$ denotes the isoform origin of read $r_i^{(d)}$ in sample d. Only the reads $r_i^{(d)}$'s are observed information. Other random variables are hidden, and parameters need estimation.

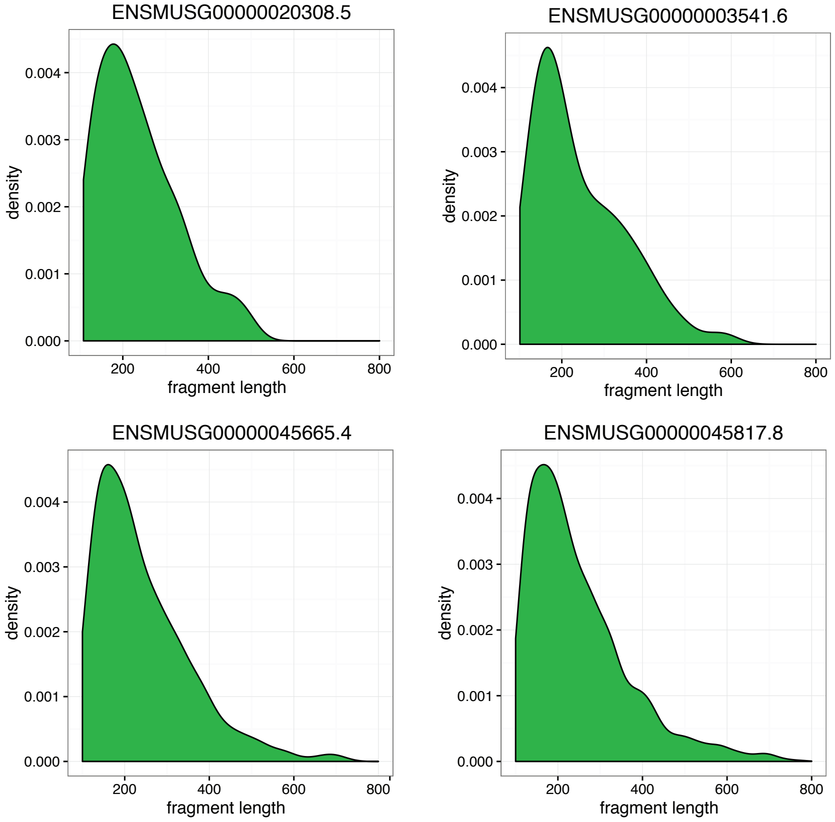


Figure B.2: The distribution of fragment length in real RNA-seq data. The empirical fragment length distribution of four example mouse genes in the mouse bone marrow-derived macrophage dataset (Table C.2).

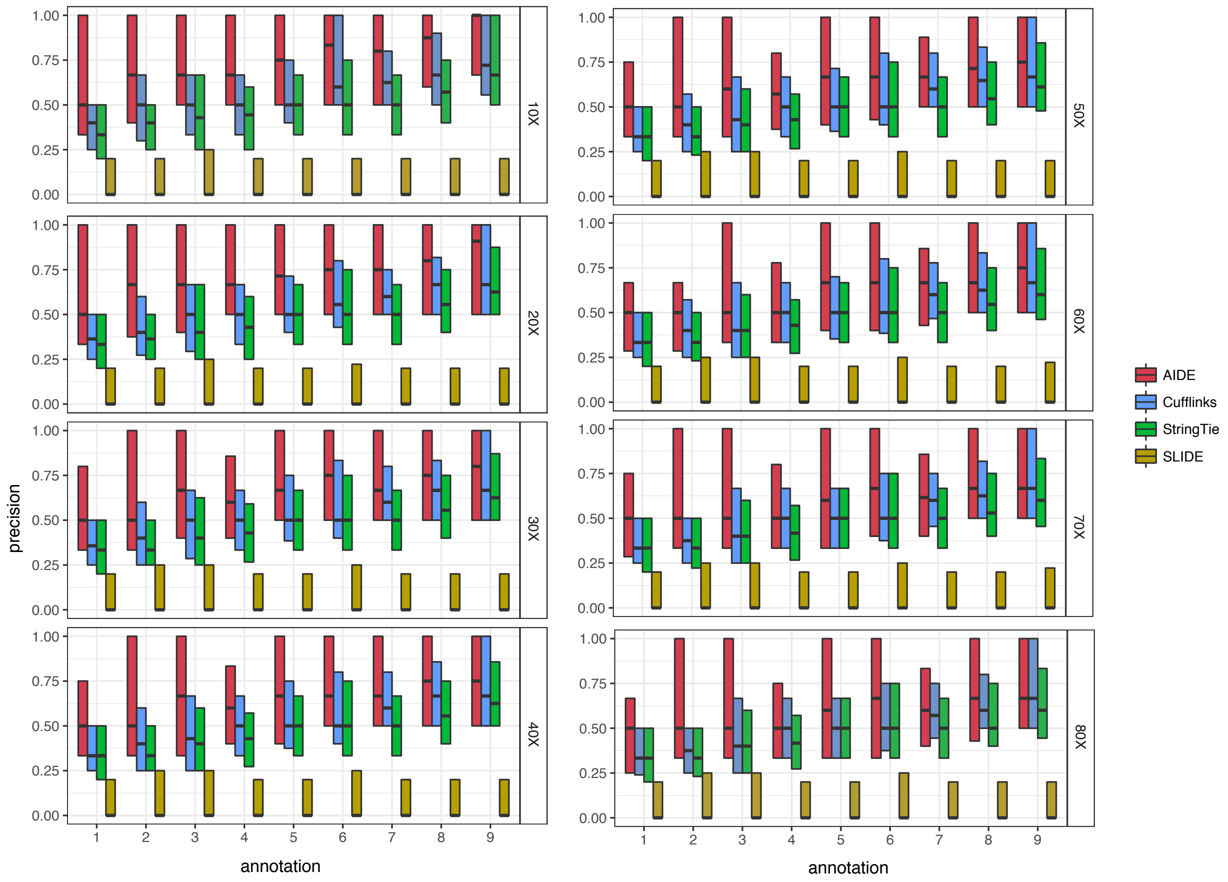


Figure B.3: Gene-level precision of AIDE and the other three isoform discovery methods in simulation. Each boxplot gives the 1st quantile, median, and 3rd quantile of the gene-level precision given the corresponding synthetic annotation set and read coverage.

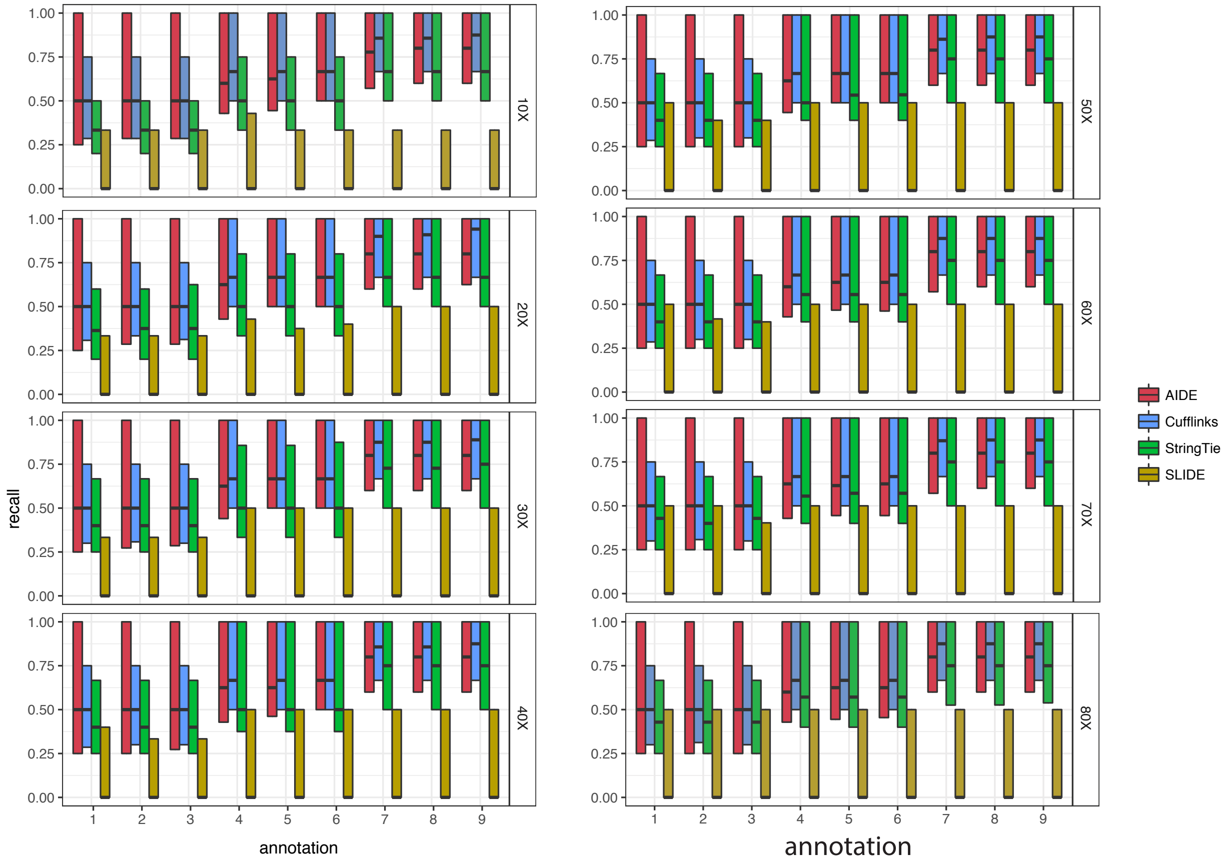


Figure B.4: Gene-level recall of AIDE and the other three isoform discovery methods in simulation. Each boxplot gives the 1st quantile, median, and 3rd quantile of the gene-level recall given the corresponding synthetic annotation set and read coverage.

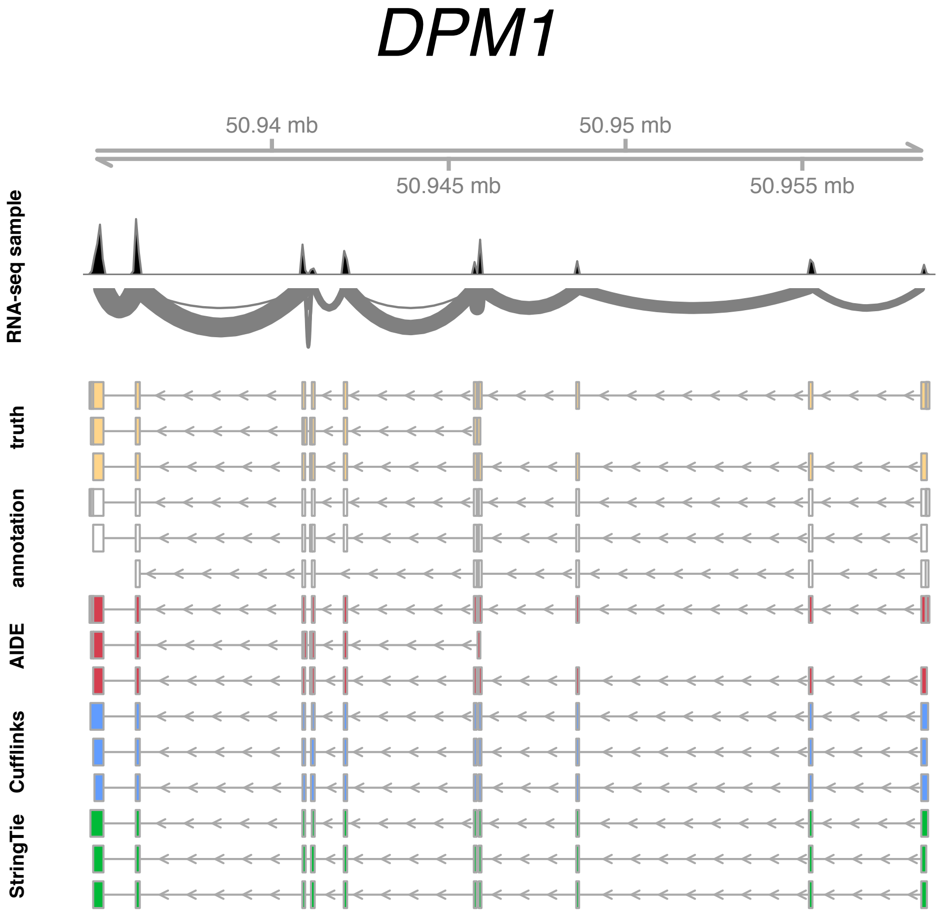


Figure B.5: Isoform discovery for the human gene *DPM1* with synthetic annotation set 1. The histogram and the sashimi plot denote the RNA-seq reads mapped to the *DPM1* gene. The annotation (white) for this gene has a 67% purity and a 67% completeness, compared with the truly expressed isoforms (yellow). AIDE, Cufflinks, and StringTie each discovered three isoforms, but only AIDE was able to identify the shortest isoform missing in the annotation. SLIDE reported 17 isoforms, which are not displayed in the plot.

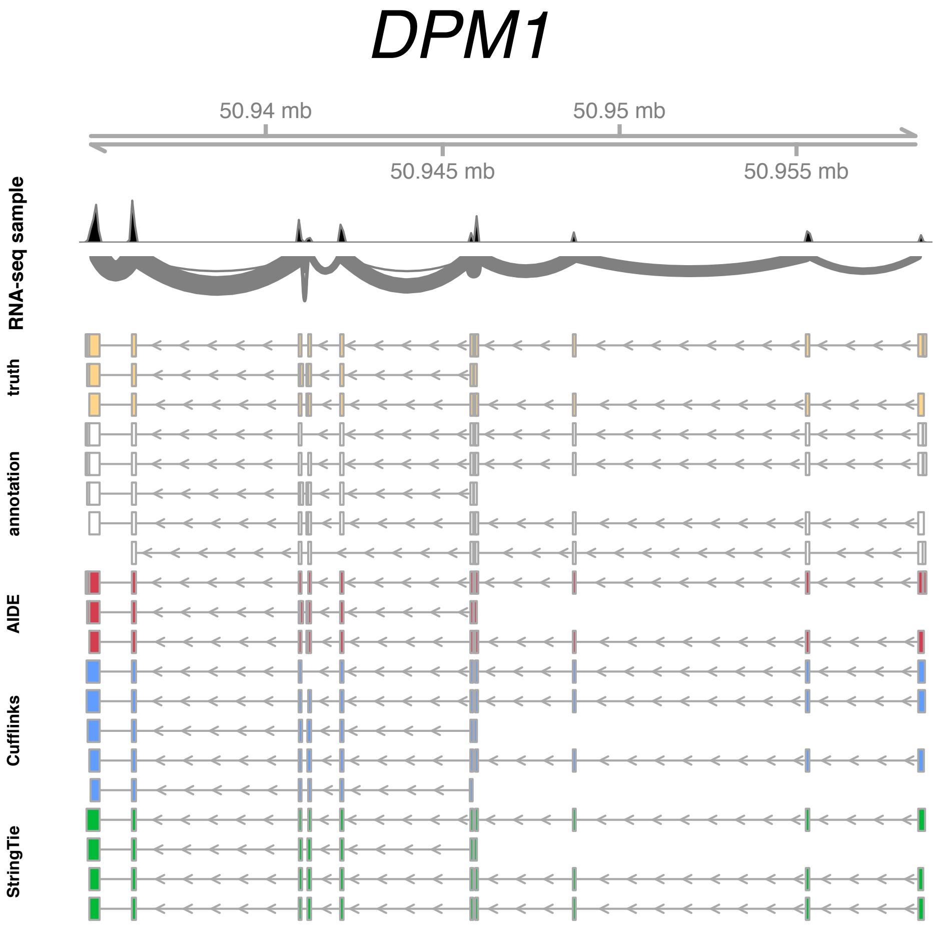


Figure B.6: Isoform discovery for the human gene *DPM1* with synthetic annotation set 9. The histogram and the sashimi plot denote the RNA-seq reads mapped to the *DPM1* gene. The annotation (white) for this gene has a 60% purity and a 100% completeness, compared with the truly expressed isoforms (yellow). AIDE, Cufflinks, and StringTie respectively discovered three, five, and four isoforms, and only AIDE was able to identify the three true isoforms with 100% accuracy. SLIDE reported 20 isoforms, which are not displayed in the plot.

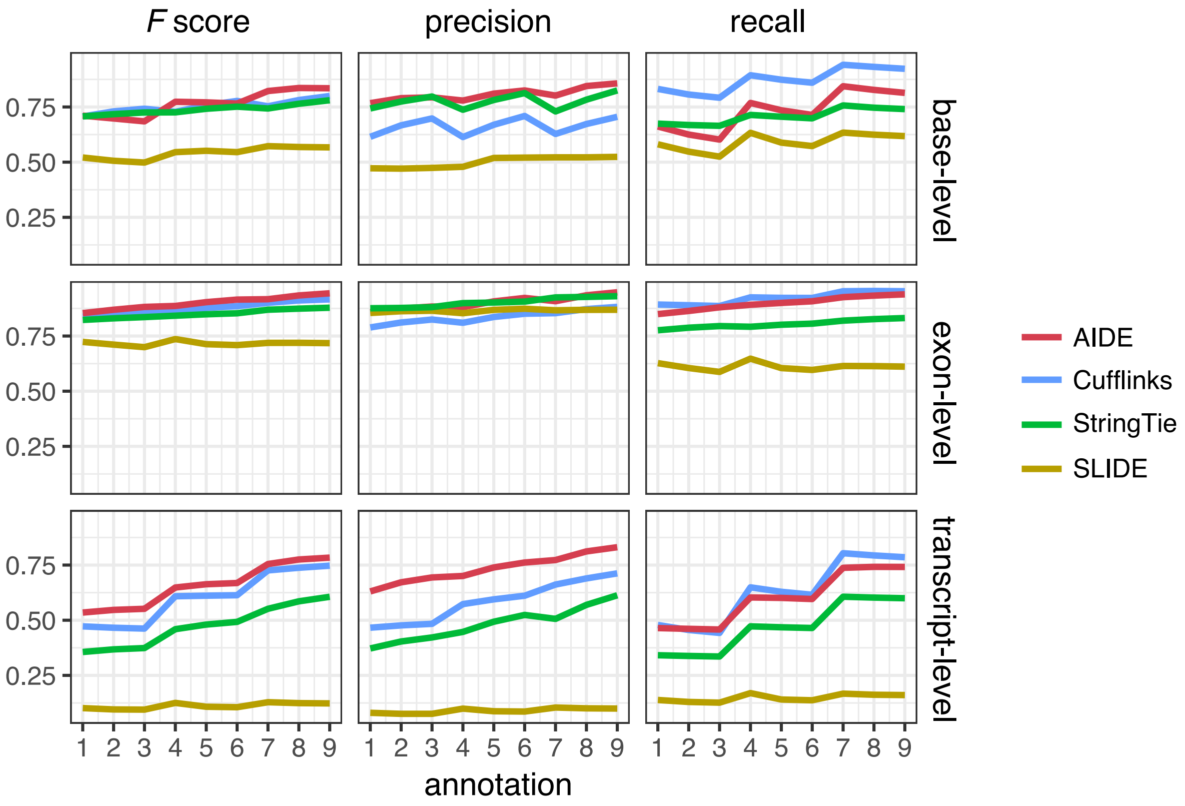


Figure B.7: Comparison between AIDE and the other three isoform discovery methods in simulation. The genome-wide average performance of AIDE, Cufflinks, StringTie, and SLIDE given each of the nine synthetic annotation sets was summarized. The base-level, exon-level, and transcript-level precision, recall, and F scores averaged across the human genes were calculated based on the RNA-seq data with a 10x coverage.

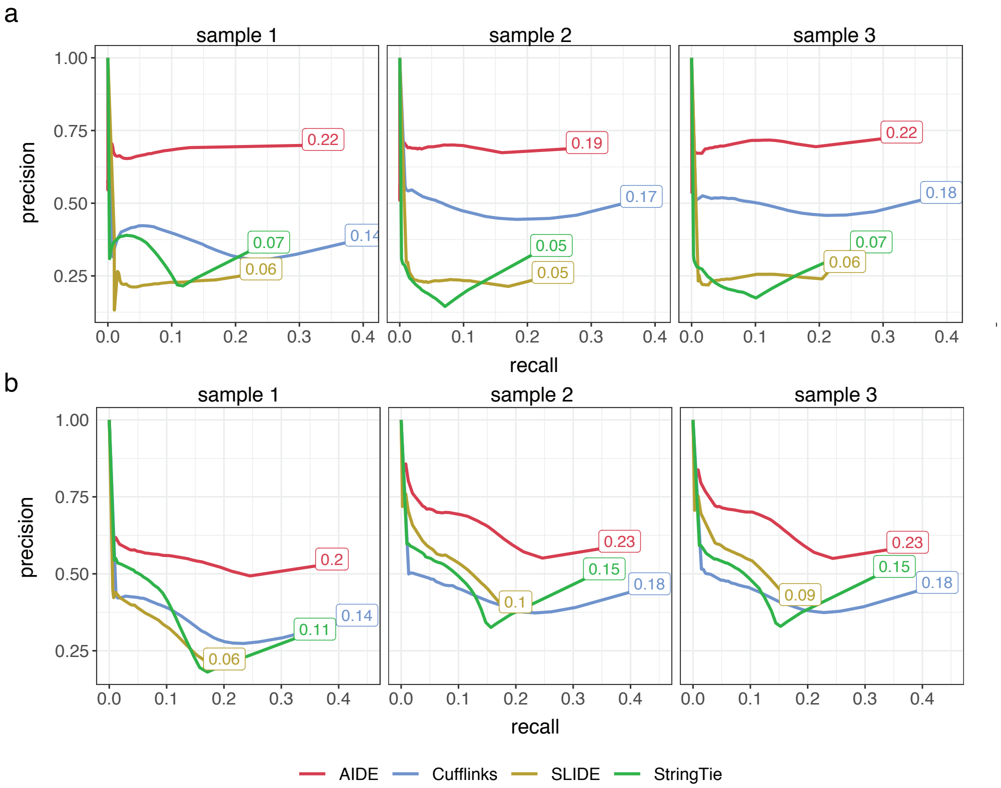


Figure B.8: Precision-recall curves of AIDE and the other three isoform discovery methods in real data studies. We applied AIDE, Cufflinks, StringTie, and SLIDE on three human ESC samples (**a**) and three mouse BMDM samples (**b**). The estimated expression levels of the predicted isoforms were summarized using the FPKM values. The precision-recall curves (at isoform-level) were obtained by thresholding the FPKM values of the predicted isoforms. The corresponding AUC of each method is also marked in the plot.

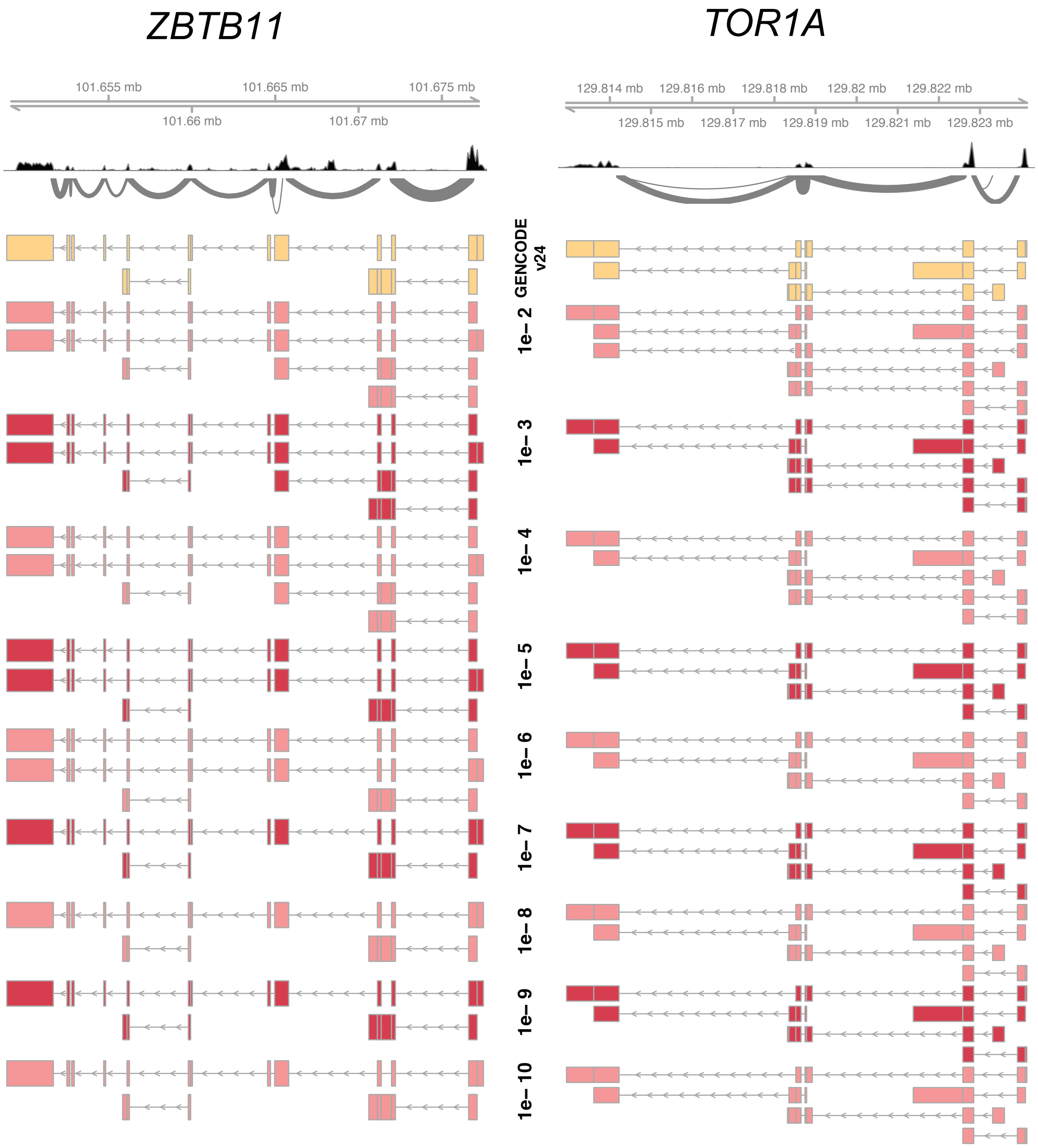


Figure B.9: Isoform discovery for the human genes *ZBTB11* and *TOR1A* given different p-value thresholds. The histogram and the sashimi plot denote the RNA-seq reads mapped to the two genes in the human ESC sample 1 (Table C.2). Isoform discovery was based on the GENCODE human annotation version 24. The threshold on the p-values resulted from the likelihood ratio tests decreased from 10^{-2} to 10^{-10}.

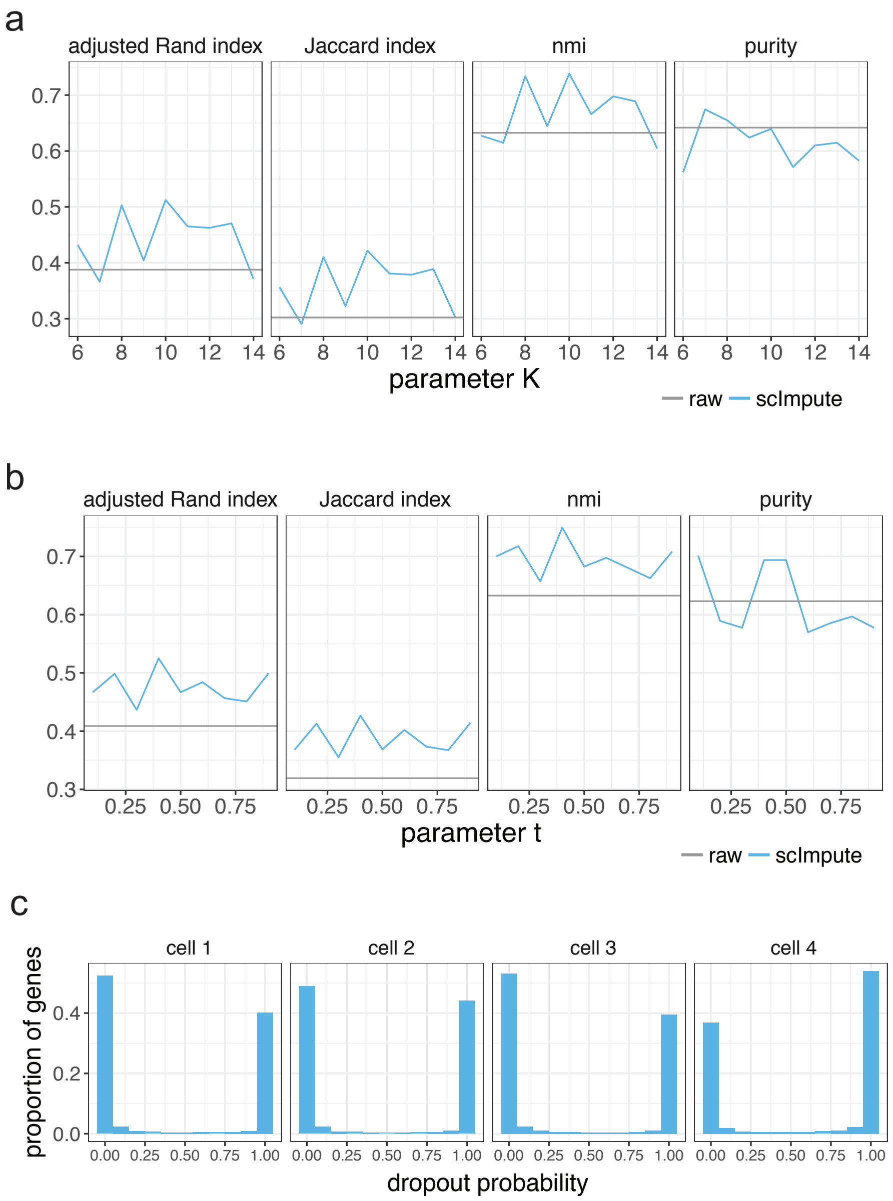


Figure B.10: Sensitivity analysis based on the mouse embryo data [201]. **a**: Clustering results of imputed data when different values of parameter K were used in scImpute. **b**: Clustering results of imputed data when different values of parameter t were used in scImpute. **c**: The distributions of dropout probabilities in four randomly selected cells from the mouse embryo data. Most genes had dropout probabilities very close to either 0 or 1.

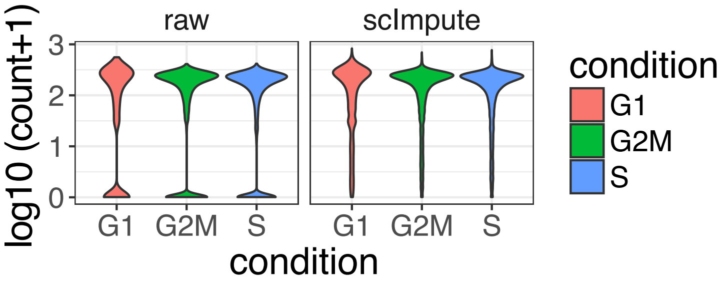


Figure B.11: Expression levels of cell cycle genes before and after imputation. Violin plots show the log 10(count+1) of the 892 cell cycle genes in the three phases (G1, G2M, and S). This comparison result shows that scImpute has successfully imputed the dropout expression values of cell cycle genes.

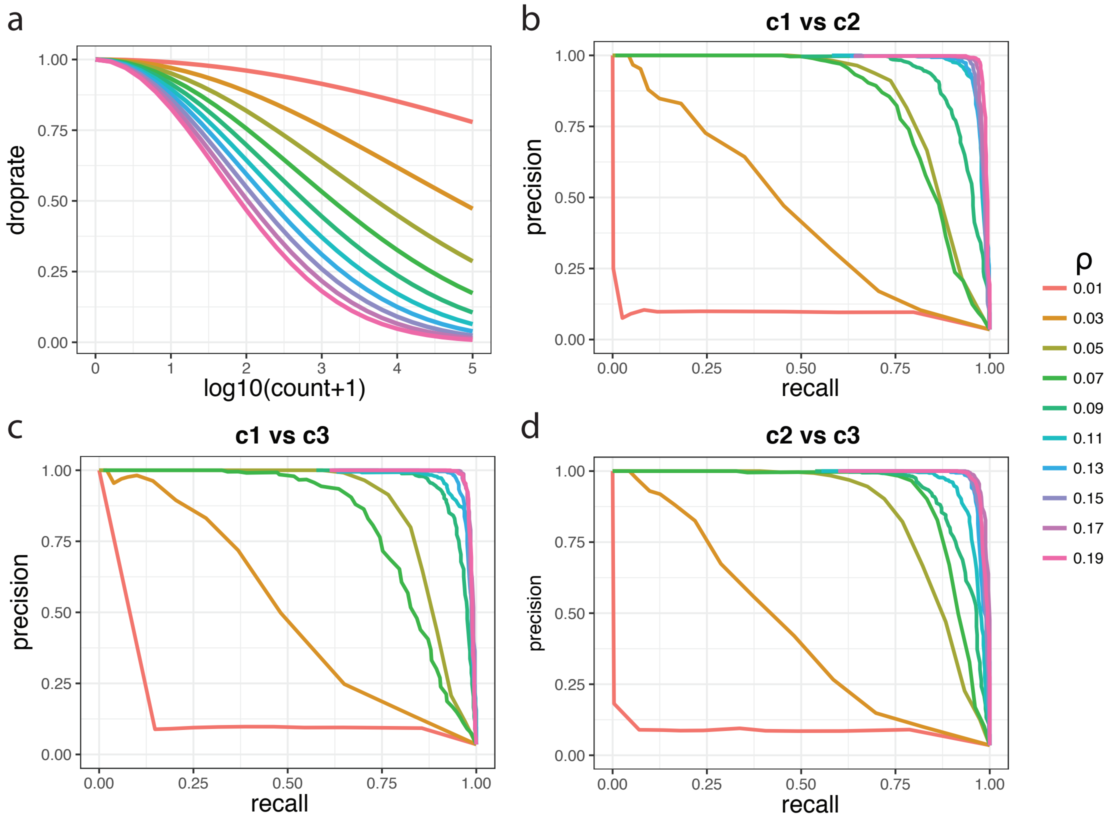


Figure B.12: Performance of scImpute given different dropout rates in simulated raw data. **a**: The theoretical dropout rates determined by the double exponential function $\exp(\rho \log 10(\text{count}+1)^2)$, with ρ varying from 0.01 to 0.19 by a step size of 0.02. **b-d**: The precision-recall curves for the identification of DE genes from the imputed data.

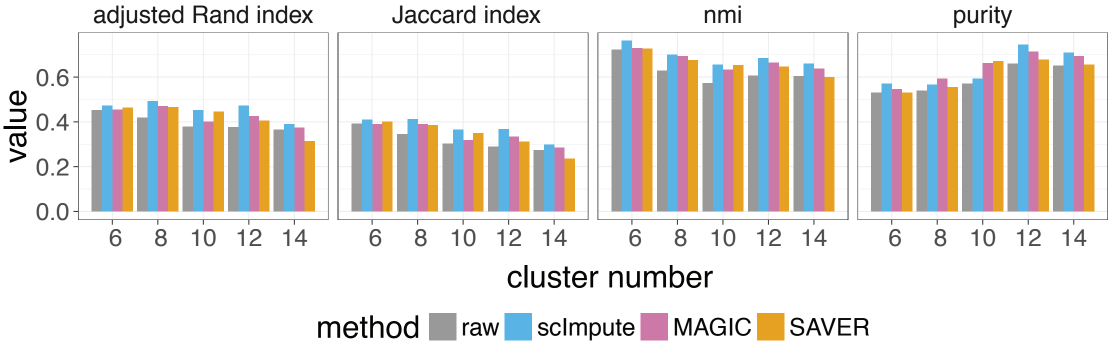


Figure B.13: The adjusted Rand index, Jaccard index, nmi, and purity scores of clustering results based on raw and imputed data. Clustering was performed by the spectral clustering algorithm on the single cells' scores in the first two PCs.

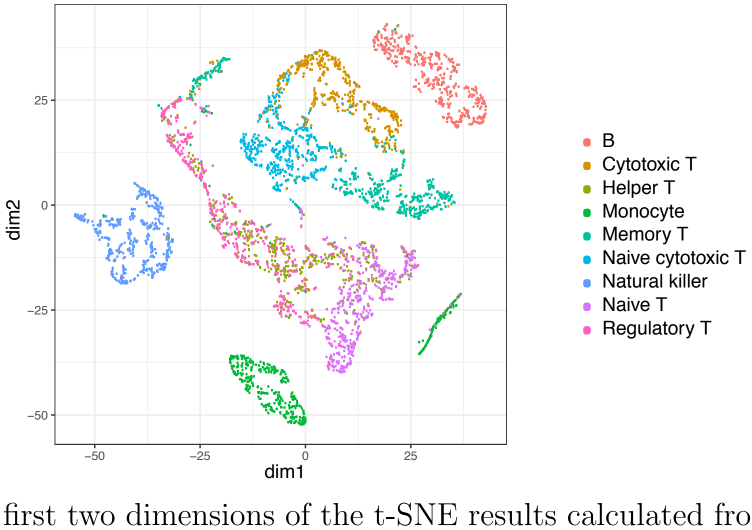


Figure B.14: The first two dimensions of the t-SNE results calculated from imputed PBMC data by MAGIC.

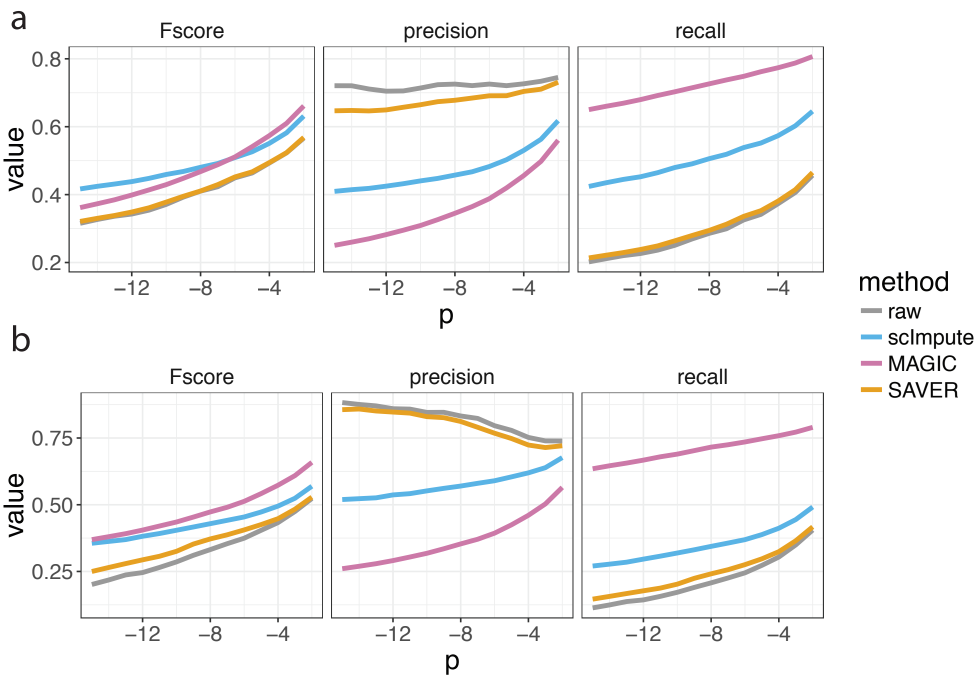


Figure B.15: Comparison of DE analysis between bulk and single-cell data. **a**: p-values on both bulk and single-cell data were calculated using DESeq2 [30]. **b**: p-values on bulk data were calculated using DESeq2 and p-values for single-cell data were calculated using MAST [192].

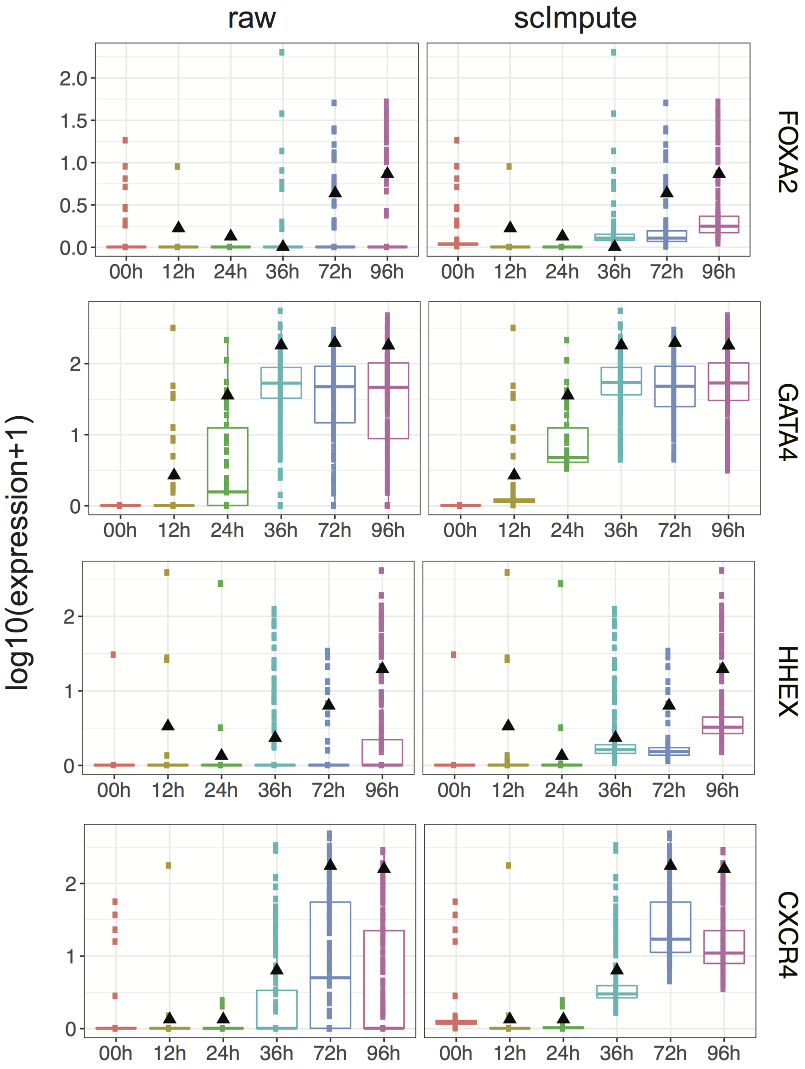


Figure B.16: Time-course expression patterns of four marker genes of DECs. The genes' $\log 10(\text{count} + 1)$ in both raw and imputed data are summarized in the boxplots. Black triangles mark the genes expression in bulk data.

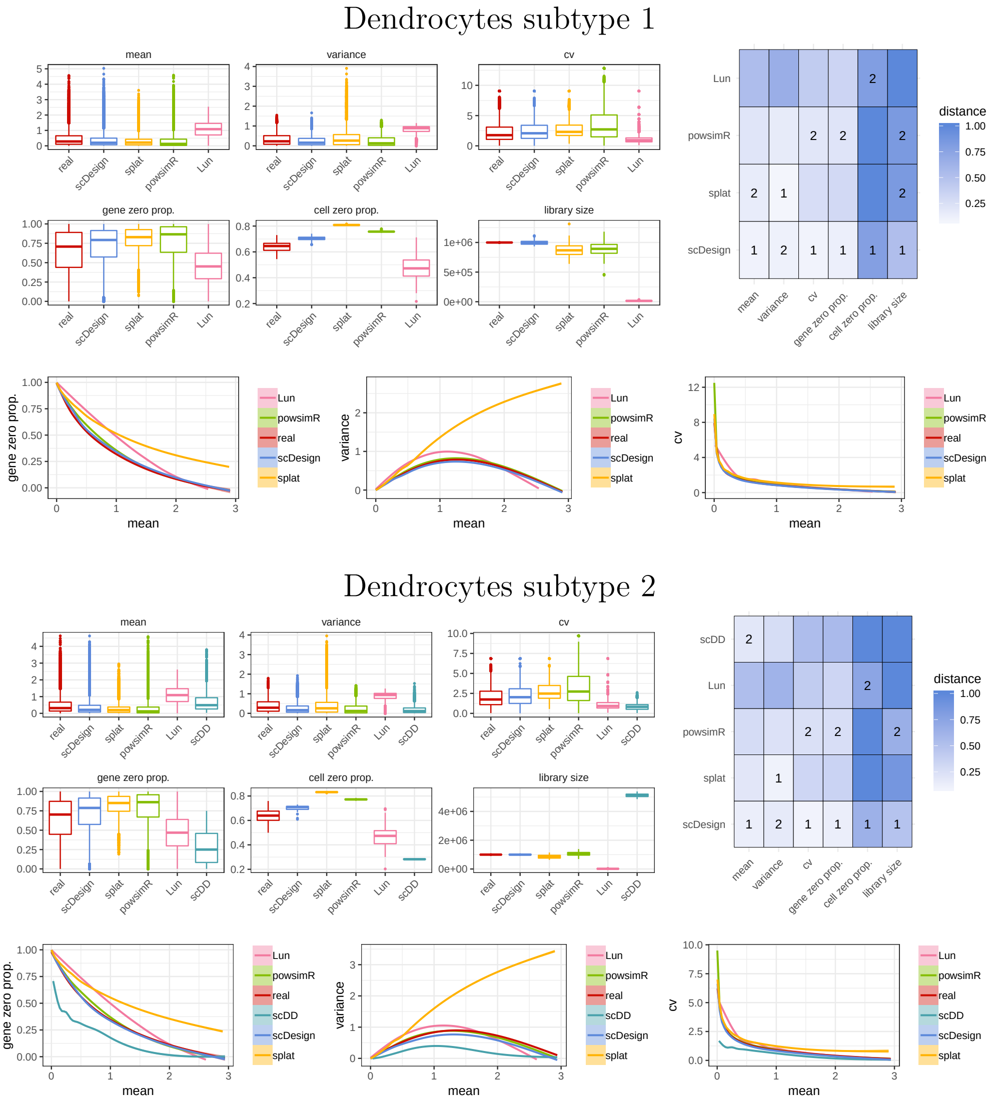


Figure B.17: Comparison of scDesign and the other four simulation methods based on the Smart-seq2 protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. The line plots demonstrate the empirical relationships between the key statistics in real and simulated data. Note that scDD failed to simulate data for the dendrocytes subtype1 dataset.

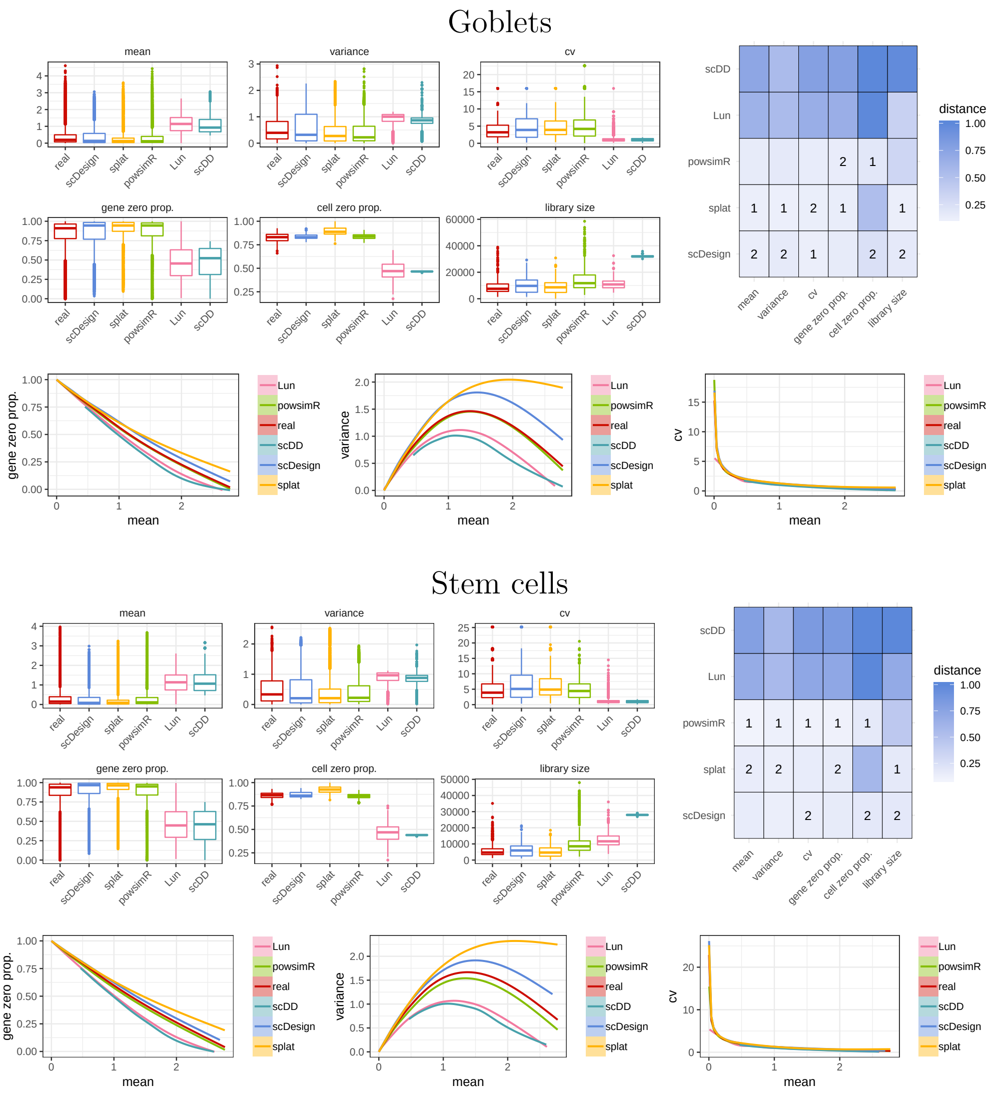


Figure B.18: Comparison of scDesign and the other four simulation methods based on the 10x Genomics protocol. The boxplots display the gene-wise expression mean, expression variance, expression coefficient of variation, zero proportion, and the cell-wise zero proportion and library size in both real and simulated datasets. The heatmaps display the KS distances between the six statistics in the real data and in the simulated data. The best and second best simulation methods with respect to each statistic are respectively marked with 1 and 2 in the heatmaps. The line plots demonstrate the empirical relationships between the key statistics in real and simulated data.

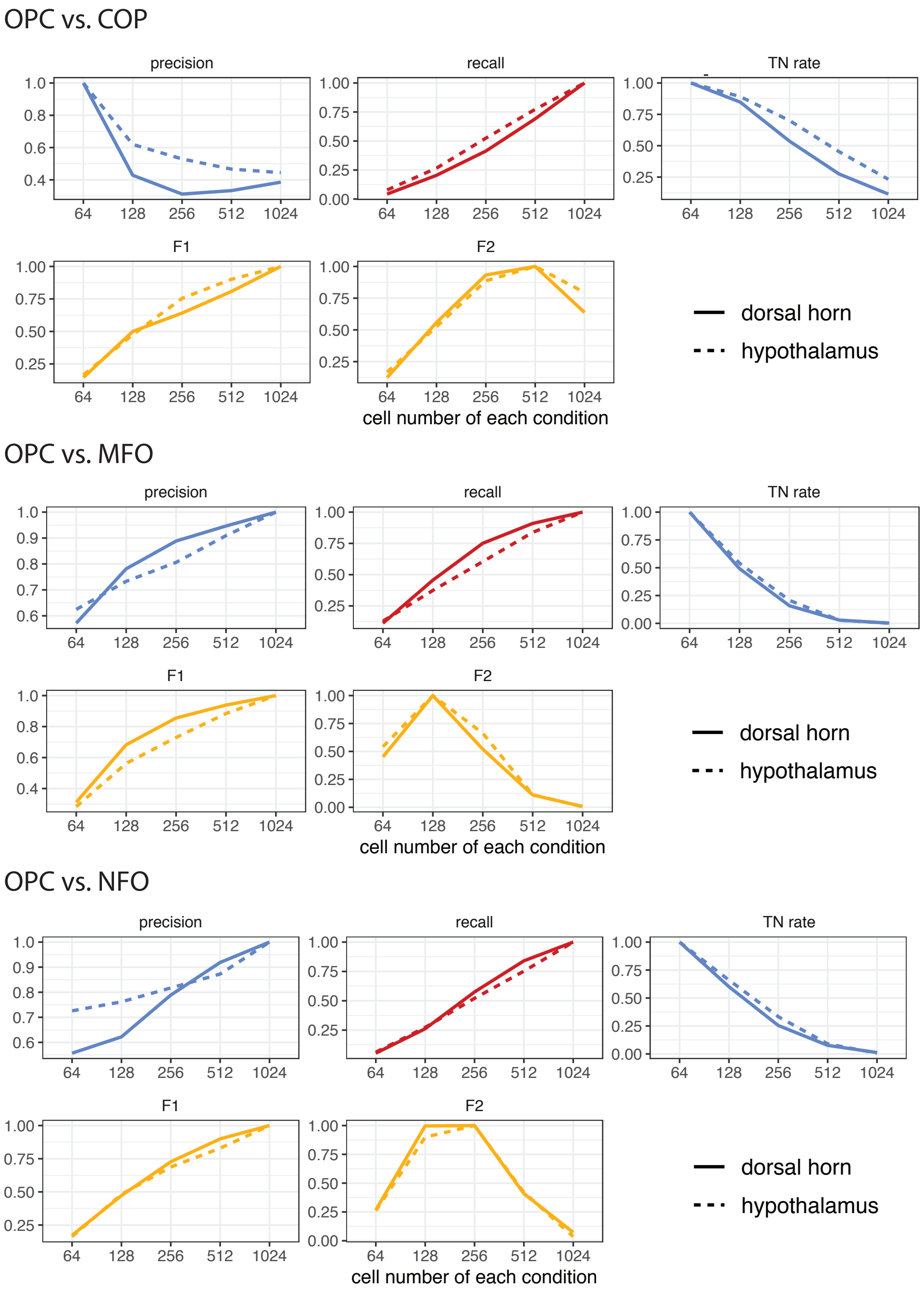


Figure B.19: Reproducibility of scDesign based on data from different brain regions. The DE studies compared OPC and three other cell types based on scRNA-seq data from two brain regions: dorsal horn and hypothalamus. When identifying the DE genes, the threshold set on the FDR rate was 10^{-10}. The y-axis of each line are divided by the maximum value of that line for normalization.

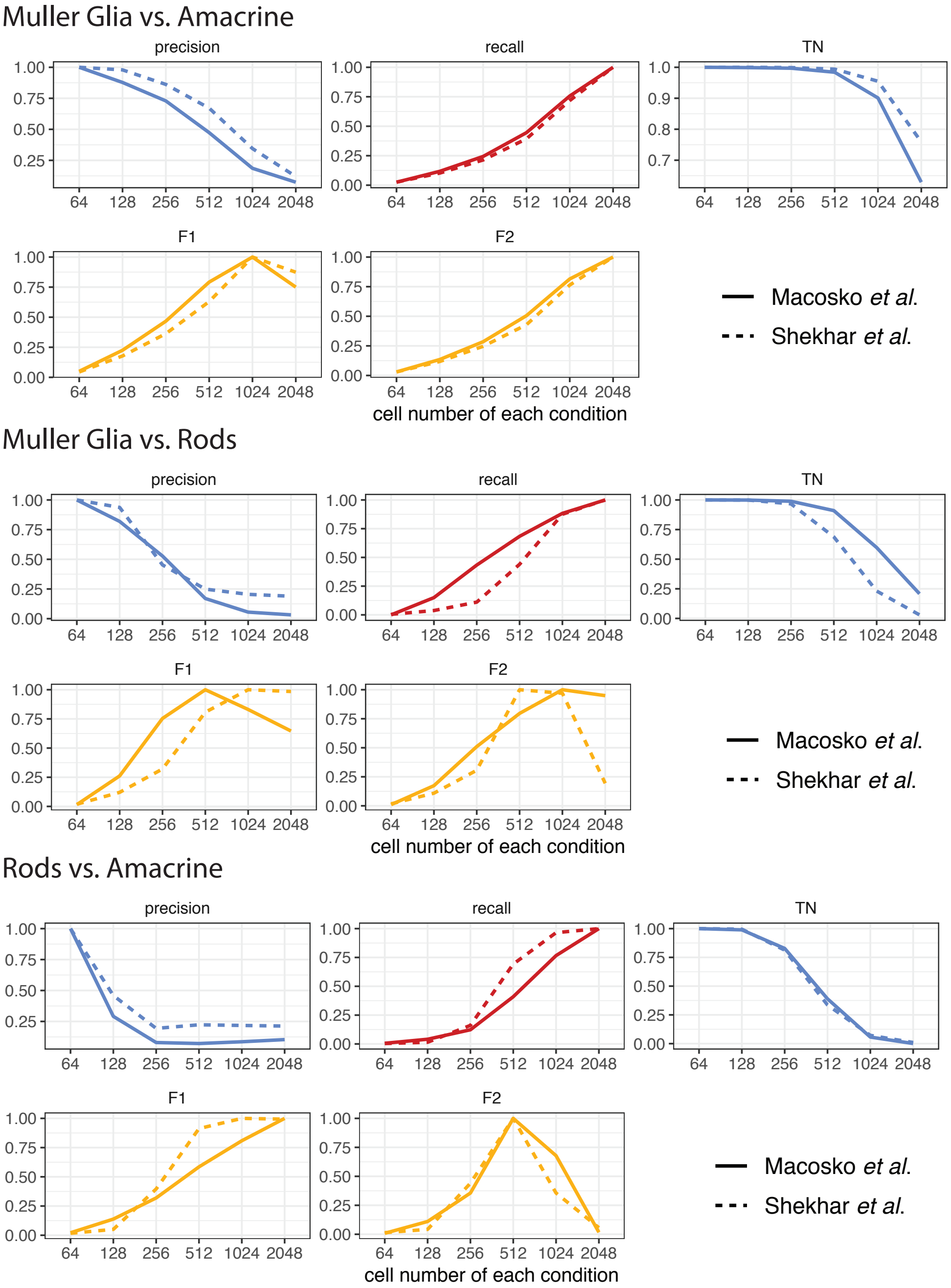


Figure B.20: Reproducibility of scDesign based on data from different studies. The DE studies compared three types of retina cells based on scRNA-seq data from two studies: Macosko *et al.* and Shekhar *et al.* When identifying the DE genes, the threshold set on the FDR rate was 10^{-10}. The y-axis of each line are divided by the maximum value of that line for normalization.

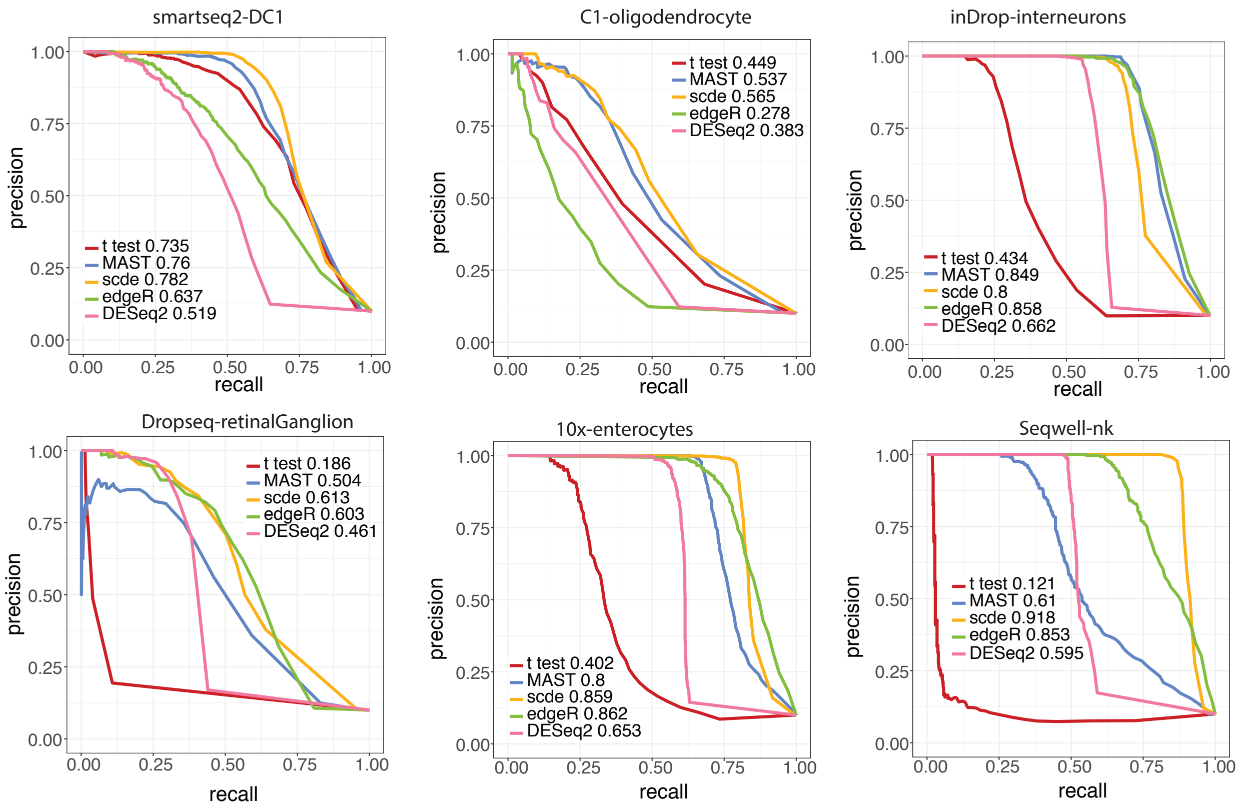


Figure B.21: Comparison of scRNA-seq DE methods. The precision-recall curves of the five DE methods were summarized for the six scRNA-seq protocols, respectively. Corresponding AUCs are shown in the plots.

APPENDIX C

Supplementary Tables

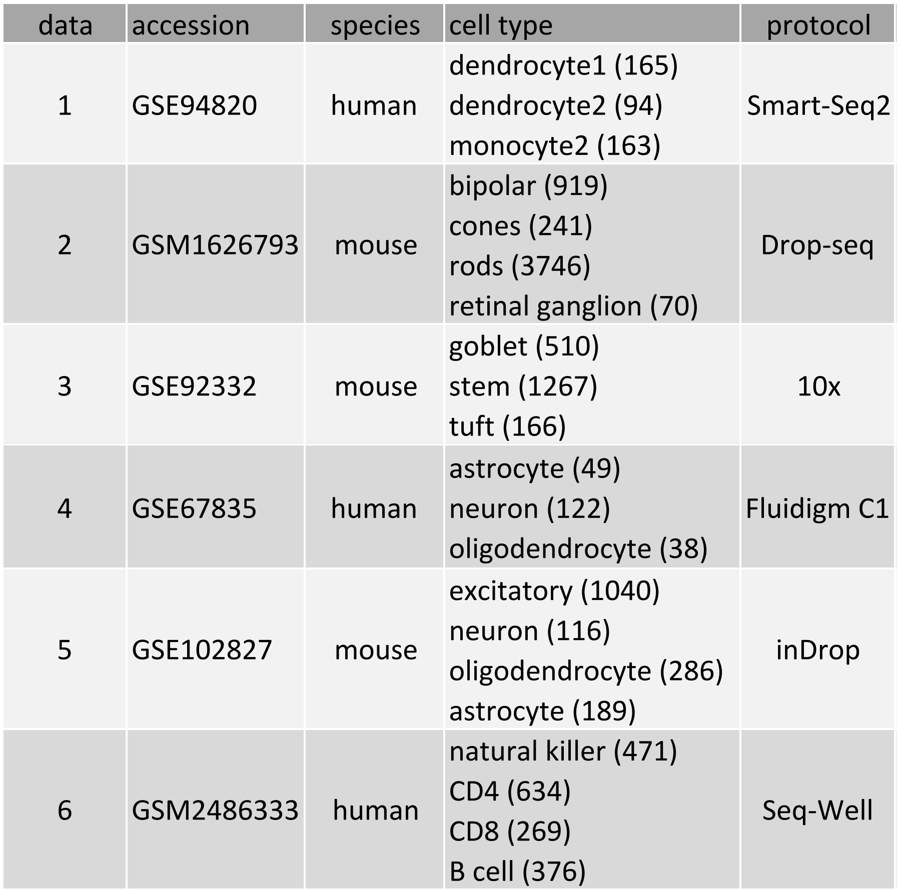
Table C.1: Description of the RNA-seq datasets used in Chapter 2.

replicate ID	cell/tissue	data type	read length	accession number
1	hESC	real data	76×2	GSM758566
2	hESC	real data	35×2 & 36×2	GSM517435
3	hESC	real data	35×2 & 36×2	GSM517435
4	hESC	real data	76×2	GSM958733
5	hESC	real data	101×2	GSM958743
6	hESC	real data	101×2	GSM1153528
7	Brain	real data	50×2	GSE19166
8	Brain	real data	50×2	GSE19166
9	Brain	real data	50×2	GSE19166
10-14	/	simulated data	36×2	/
15-19	/	simulated data	76×2	/
20-24	/	simulated data	101×2	/
1	HepG2	real data	50×2	ENCFF084JYA
2	HepG2	real data	50×2	ENCFF790CFB
3	HepG2	real data	38×2	ENCFF916YZY
4	HepG2	real data	38×2	ENCFF179TFY
5	HepG2	real data	50×2	ENCFF168NGI
6	HepG2	real data	50×2	ENCFF711DJN

Table C.2: Description of real RNA-seq datasets used in Chapter 3.

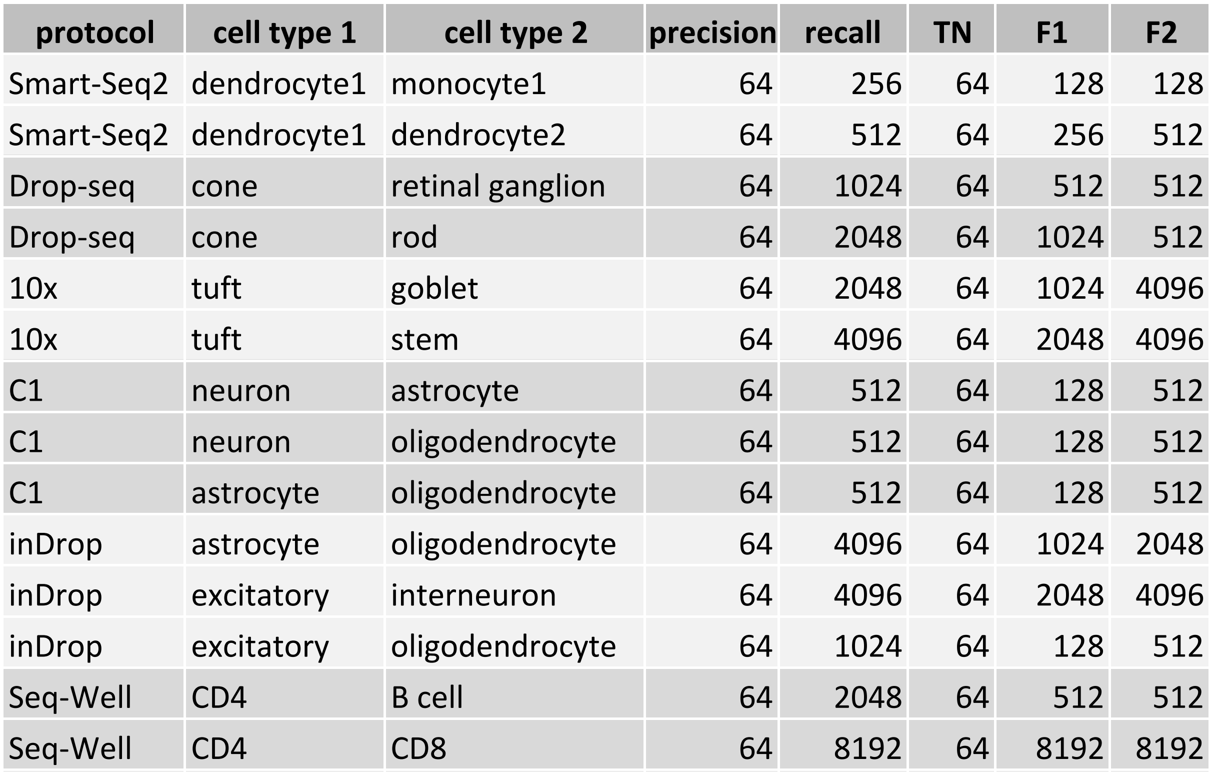
sample	cell type	read length	accession number
1	HepG2	50×2	ENCFF084JYA
2	HepG2	50×2	ENCFF790CFB
3	HepG2	100×2	ENCFF916YZY, ENCFF800YJR
4	HepG2	100×2	ENCFF179TFY, ENCFF782TAX
5	HepG2	76×2	ENCFF168NGI
6	HepG2	76×2	ENCFF711DJN
sample	cell type	read length	accession number
1	human ESC	76×2	GSE90225
2	human ESC	76×2	GSE33480
3	human ESC	101×2	GSE47626
sample	cell type	read length	accession number
1	mouse BMDM	100×2	ENCSR614DLJ
2	mouse BMDM	100×2	ENCSR822FMG
3	mouse BMDM	100×2	ENCSR614KOV

Table C.3: Description of real scRNA-seq datasets used in Chapter 6.



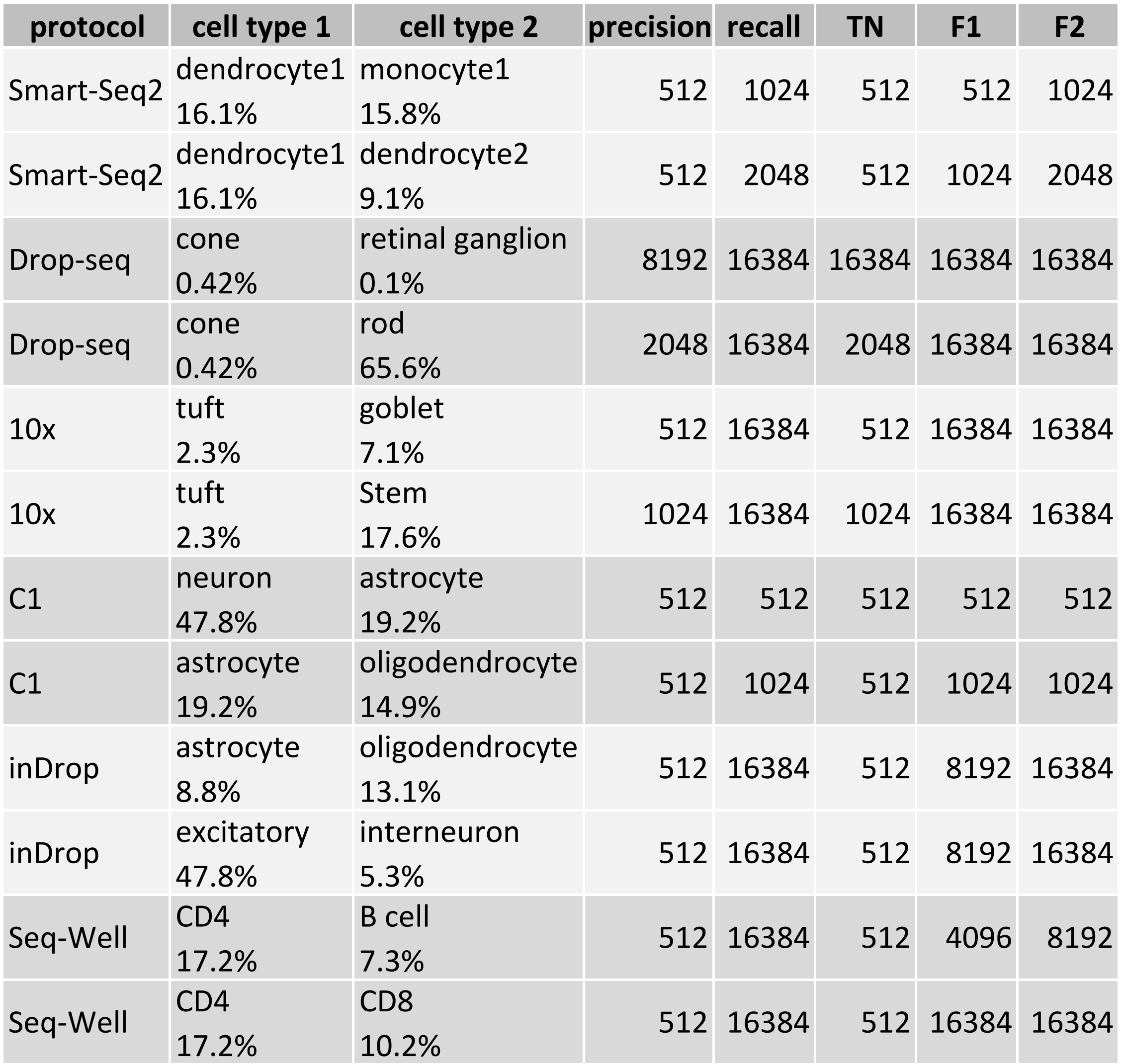
data	accession	species	cell type	protocol
1	GSE94820	human	dendrocyte1 (165) dendrocyte2 (94) monocyte2 (163)	Smart-Seq2
2	GSM1626793	mouse	bipolar (919) cones (241) rods (3746) retinal ganglion (70)	Drop-seq
3	GSE92332	mouse	goblet (510) stem (1267) tuft (166)	10x
4	GSE67835	human	astrocyte (49) neuron (122) oligodendrocyte (38)	Fluidigm C1
5	GSE102827	mouse	excitatory (1040) neuron (116) oligodendrocyte (286) astrocyte (189)	inDrop
6	GSM2486333	human	natural killer (471) CD4 (634) CD8 (269) B cell (376)	Seq-Well

Table C.4: The optimal cell numbers in experimental design (scenario A) using t tests. Five DE accuracy metrics were calculated in every comparison: precision, recall, TN (true negative rate), F1 score, and F2 score. For each metric, the smallest cell number that leads to the highest accuracy was recorded as the optimal number.



protocol	cell type 1	cell type 2	precision	recall	TN	F1	F2
Smart-Seq2	dendrocyte1	monocyte1	64	256	64	128	128
Smart-Seq2	dendrocyte1	dendrocyte2	64	512	64	256	512
Drop-seq	cone	retinal ganglion	64	1024	64	512	512
Drop-seq	cone	rod	64	2048	64	1024	512
10x	tuft	goblet	64	2048	64	1024	4096
10x	tuft	stem	64	4096	64	2048	4096
C1	neuron	astrocyte	64	512	64	128	512
C1	neuron	oligodendrocyte	64	512	64	128	512
C1	astrocyte	oligodendrocyte	64	512	64	128	512
inDrop	astrocyte	oligodendrocyte	64	4096	64	1024	2048
inDrop	excitatory	interneuron	64	4096	64	2048	4096
inDrop	excitatory	oligodendrocyte	64	1024	64	128	512
Seq-Well	CD4	B cell	64	2048	64	512	512
Seq-Well	CD4	CD8	64	8192	64	8192	8192

Table C.5: The optimal cell numbers in experimental design (scenario B) using t tests. Five DE accuracy metrics were calculated in every comparison: precision, recall, TN (true negative rate), F1 score, and F2 score. For each measure, the smallest cell number that leads to the highest accuracy was recorded as the optimal number. The cell type proportions are listed under the cell labels.



protocol	cell type 1	cell type 2	precision	recall	TN	F1	F2
Smart-Seq2	dendrocyte1 16.1%	monocyte1 15.8%	512	1024	512	512	1024
Smart-Seq2	dendrocyte1 16.1%	dendrocyte2 9.1%	512	2048	512	1024	2048
Drop-seq	cone 0.42%	retinal ganglion 0.1%	8192	16384	16384	16384	16384
Drop-seq	cone 0.42%	rod 65.6%	2048	16384	2048	16384	16384
10x	tuft 2.3%	goblet 7.1%	512	16384	512	16384	16384
10x	tuft 2.3%	Stem 17.6%	1024	16384	1024	16384	16384
C1	neuron 47.8%	astrocyte 19.2%	512	512	512	512	512
C1	astrocyte 19.2%	oligodendrocyte 14.9%	512	1024	512	1024	1024
inDrop	astrocyte 8.8%	oligodendrocyte 13.1%	512	16384	512	8192	16384
inDrop	excitatory 47.8%	interneuron 5.3%	512	16384	512	8192	16384
Seq-Well	CD4 17.2%	B cell 7.3%	512	16384	512	4096	8192
Seq-Well	CD4 17.2%	CD8 10.2%	512	16384	512	16384	16384

REFERENCES

[1] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[2] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, 9(1):e78644, 2014.

[3] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Tyler Alioto, Jonas Behr, Paul Bertone, Regina Bohnert, Davide Campagna, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12):1185, 2013.

[4] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, 2013.

[5] Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R Gruber, Georges Martin, and Mihaela Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, 16(1):1–26, 2015.

[6] Nicolas J Tourasse, Jonathan RM Millet, and Denis Dupuy. Quantitative RNA-seq meta-analysis of alternative exon usage in C. elegans. *Genome Research*, 27(12):2120–2128, 2017.

[7] Ruiqi Gao and Jingyi Jessica Li. Correspondence of D. melanogaster and C. elegans developmental stages revealed by alternative splicing characteristics of conserved exons. *BMC Genomics*, 18(1):234, 2017.

[8] Michelle N Arbeitman, Eileen EM Furlong, Farhad Imam, Eric Johnson, Brian H Null, Bruce S Baker, Mark A Krasnow, Matthew P Scott, Ronald W Davis, and Kevin P White. Gene expression during the life cycle of Drosophila melanogaster. *Science*, 297(5590):2270–2275, 2002.

[9] Anamaria Necsulea, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C Baker, Frank Grützner, and Henrik Kaessmann. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–640, 2014.

[10] Wei Vivian Li, Yiling Chen, and Jingyi Jessica Li. TROM: A testing-based method for finding transcriptomic similarity of biological samples. *Statistics in Biosciences*, 9(1):105–136, 2017.

[11] Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[12] Aaron D Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1):51–59, 1978.

[13] Juan Zhao, Yiwei Zhou, Xiujun Zhang, and Luonan Chen. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences*, 113(18):5130–5135, 2016.

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[15] Trevor F Cox and Michael AA Cox. *Multidimensional scaling.* Chapman and Hall/CRC, 2000.

[16] Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792, 2017.

[17] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

[18] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.

[19] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[20] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[21] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[22] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[23] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[24] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[25] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

[26] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):3158, 2013.

[27] Joshua S Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy A Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10(1):221, 2009.

[28] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[29] Thomas J Hardcastle and Krystyna A Kelly. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, 2010.

[30] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[31] Danni Yu, Wolfgang Huber, and Olga Vitek. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics*, 29(10):1275–1282, 2013.

[32] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

[33] Mark A Van De Wiel, Gwenaël GR Leday, Luba Pardo, Håvard Rue, Aad W Van Der Vaart, and Wessel N Van Wieringen. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128, 2013.

[34] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014.

[35] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[36] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687, 2017.

[37] Nicholas J Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, Tom Owen-Hughes, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, 2016.

[38] Jerzy Neyman and Egon S Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20(1/2):175–240, 1928.

[39] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[40] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B*, 57(1):289–300, 1995.

[41] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[42] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[43] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[44] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[45] Martin Oti, Jeroen van Reeuwijk, Martijn A Huynen, and Han G Brunner. Conserved co-expression for candidate disease gene prioritization. *BMC Bioinformatics*, 9(1):208, 2008.

[46] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.

[47] Stefan Canzar, Sandro Andreotti, David Weese, Knut Reinert, and Gunnar W Klau. CIDANE: Comprehensive isoform discovery and abundance estimation. *Genome Biology*, 17(1):1, 2016.

[48] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.

[49] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[50] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, 2013.

[51] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525, 2016.

[52] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22, 1977.

[53] Jing Zhang, C-C Jay Kuo, and Liang Chen. WEMIQ: an accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics*, page btu757, 2014.

[54] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 2017.

[55] Aziz M Mezlini, Eric JM Smith, Marc Fiume, Orion Buske, Gleb L Savich, Sohrab Shah, Sam Aparicio, Derek Y Chiang, Anna Goldenberg, and Michael Brudno. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research*, 23(3):519–529, 2013.

[56] Wei Vivian Li, Anqi Zhao, Shihua Zhang, Jingyi Jessica Li, et al. MSIQ: Joint modeling of multiple RNA-seq samples for accurate isoform quantification. *The Annals of Applied Statistics*, 12(1):510–539, 2018.

[57] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.

[58] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, 34(12):1287–1291, 2016.

[59] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[60] Zheng Xia, Jianguo Wen, Chung-Che Chang, and Xiaobo Zhou. NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics*, 12(1):162, 2011.

[61] Regina Bohnert and Gunnar Rätsch. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Research*, 38(suppl_2):W348–W351, 2010.

[62] Jingyi Jessica Li, Ci-Ren Jiang, James B Brown, Haiyan Huang, and Peter J Bickel. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*, 108(50):19867–19872, 2011.

[63] Wei Li, Jianxing Feng, and Tao Jiang. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *Journal of Computational Biology*, 18(11):1693–1707, 2011.

[64] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010.

[65] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[66] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, 2010.

[67] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015.

[68] Xi Wang, Zhengpeng Wu, and Xuegong Zhang. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *Journal of Bioinformatics and Computational Biology*, 8(supp01):177–192, 2010.

[69] Yen-Yi Lin, Phuong Dao, Faraz Hach, Marzieh Bakhshi, Fan Mo, Anna Lapuk, Colin Collins, and S Cenk Sahinalp. CLIIQ: Accurate comparative detection and quantification of expressed isoforms in a population. In *Algorithms in Bioinformatics*, pages 178–189. Springer, 2012.

[70] Jonas Behr, André Kahles, Yi Zhong, Vipin T Sreedharan, Philipp Drewe, and Gunnar Rätsch. MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, 29(20):2529–2538, 2013.

[71] Elsa Bernard, Laurent Jacob, Julien Mairal, and Jean-Philippe Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, page btu317, 2014.

[72] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, Paul Bertone, RGASP Consortium, et al. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12):1177–1184, 2013.

[73] Jie Wu, Martin Akerman, Shuying Sun, W Richard McCombie, Adrian R Krainer, and Michael Q Zhang. SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, 27(21):3010–3016, 2011.

[74] Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601, 2014.

[75] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome research*, 22(9):1760–1774, 2012.

[76] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.

[77] Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, et al. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153, 2008.

[78] Ashley Byrne, Anna E Beaudin, Hugh E Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M DuBois, E Camilla Forsberg, Mark Akeson, and Christopher Vollmers. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, 8:16027, 2017.

[79] Kin Fai Au, Vittorio Sebastiano, Pegah Tootoonchi Afshar, Jens Durruthy Durruthy, Lawrence Lee, Brian A Williams, Harm van Bakel, Eric E Schadt, Renee A Reijo-Pera, Jason G Underwood, et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences*, 110(50):E4821–E4830, 2013.

[80] Christoph Bleidorn. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, 14(1):1–8, 2016.

[81] Gokul Ramaswami, Wei Lin, Robert Piskol, Meng How Tan, Carrie Davis, and Jin Billy Li. Accurate identification of human Alu and non-Alu RNA editing sites. *Nature Methods*, 9(6):579–581, 2012.

[82] Jae Hoon Bahn, Jae-Hyung Lee, Gang Li, Christopher Greer, Guangdun Peng, and Xinshu Xiao. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Research*, 22(1):142–150, 2012.

[83] Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, John R Prensner, Joseph R Evans, Shuang Zhao, et al. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 47(3):199–208, 2015.

[84] Hadas Hezroni, David Koppstein, Matthew G Schwartz, Alexandra Avrutin, David P Bartel, and Igor Ulitsky. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Reports*, 11(7):1110–1122, 2015.

[85] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010.

[86] Daniel E Zak, Adam Penn-Nicholson, Thomas J Scriba, Ethan Thompson, Sara Suliman, Lynn M Amon, Hassan Mahomed, Mzwandile Erasmus, Wendy Whatney, Gregory D Hussey, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet*, 387(10035):2312–2322, 2016.

[87] R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476, 2010.

[88] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610–620, 2015.

[89] Wei Vivian Li and Jingyi Jessica Li. Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology*, 6(3):195–209, 2018.

[90] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[91] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, 42(D1):D756–D763, 2014.

[92] Kasper D Hansen, Zhijin Wu, Rafael A Irizarry, and Jeffrey T Leek. Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, 29(7):572–573, 2011.

[93] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):1, 2016.

[94] Yuting Ye and Jingyi Jessica Li. NMFP: a non-negative matrix factorization based preselection method to increase accuracy of identifying mRNA isoforms from RNA-seq data. *BMC Genomics*, 17(1):127, 2016.

[95] Meena K Sakharkar, Vincent TK Chow, and Pandjassarame Kangueane. Distributions of exons and introns in the human genome. *In Silico Biology*, 4(4):387–393, 2004.

[96] Lior Pachter. Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889*, 2011.

[97] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, 2014.

[98] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*, 35(4):319–321, 2017.

[99] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[100] David Rossell, Camille Stephan-Otto Attolini, Manuel Kroiss, and Almond Stöcker. Quantifying alternative splicing from paired-end RNA-sequencing data. *The Annals of Applied Statistics*, 8(1):309, 2014.

[101] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083, 2012.

[102] Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, 2015.

[103] Meghana M Kulkarni. Digital multiplexed gene expression analysis using the NanoString nCounter system. *Current Protocols in Molecular Biology*, 94(1):25B–10, 2011.

[104] Pierre-Luc Germain, Alessandro Vitriolo, Antonio Adamo, Pasquale Laise, Vivek Das, and Giuseppe Testa. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Research*, 44(11):5054–5067, 2016.

[105] Mateusz G Adamski, Patryk Gumann, and Alison E Baird. A method for quantitative analysis of standard and high-throughput qPCR expression data based on input sample quantity. *PloS One*, 9(8):e103917, 2014.

[106] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, 11(1):41–46, 2014.

[107] Iain C Macaulay and Thierry Voet. Single cell genomics: advances and future perspectives. *PLoS Genet*, 10(1):e1004126, 2014.

[108] Claudia Ghigna, Cristina Valacca, and Giuseppe Biamonti. Alternative splicing and tumor progression. *Current Genomics*, 9(8):556–570, 2008.

[109] Joan E Hooper. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Human genomics*, 8(1):3, 2014.

[110] Guey-Shin Wang and Thomas A Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, 2007.

[111] Daniel Mordes, Xiaoyan Luo, Amar Kar, David Kuo, Lili Xu, Kazuo Fushimi, Guowu Yu, Paul Sternberg Jr, and Jane Y Wu. Pre-mRNA splicing and retinitis pigmentosa. *Molecular Vision*, 12:1259, 2006.

[112] Natalia N Singh and Ravindra N Singh. Alternative splicing in spinal muscular atrophy underscores the role of an intron definition model. *RNA Biology*, 8(4):600–606, 2011.

[113] Sagar Chhangawala, Gabe Rudy, Christopher E Mason, and Jeffrey A Rosenfeld. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology*, 16(1):131, 2015.

[114] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105–e105, 2008.

[115] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, 2011.

[116] Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):1, 2010.

[117] Alyssa C Frazee, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, 2015.

[118] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, 12(1):1, 2011.

[119] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131–e131, 2010.

[120] Hui Jiang and Julia Salzman. A penalized likelihood approach for robust estimation of isoform expression. *Statistics and Its Interface*, 8(4):437, 2015.

[121] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2017.

[122] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[123] Francesca Finotello, Enrico Lavezzo, Luca Bianco, Luisa Barzon, Paolo Mazzon, Paolo Fontana, Stefano Toppo, and Barbara Di Camillo. Reducing bias in RNA sequencing data: a novel approach to compute counts. *BMC Bioinformatics*, 15(1):S7, 2014.

[124] Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

[125] Anne Limbourg, Johann von Felden, Kumaravelu Jagavelu, Kashyap Krishnasamy, L Christian Napp, Piyushkumar R Kapopara, Matthias Gaestel, Bernhard Schieffer, Johann Bauersachs, Florian P Limbourg, et al. Map-kinase activated protein kinase 2 links endothelial activation and monocyte/macrophage recruitment in arteriogenesis. *PloS One*, 10(10):e0138542, 2015.

[126] Liping Zhang, Limei Ran, Gabriela E Garcia, Xiaonan H Wang, Shuhua Han, Jie Du, and William E Mitch. Chemokine CXCL16 regulates neutrophil and macrophage infiltration into injured muscle, promoting muscle regeneration. *The American Journal of Pathology*, 175(6):2518–2527, 2009.

[127] Joachim L Schultze and Susanne V Schmidt. Molecular features of macrophage activation. In *Seminars in immunology*, volume 27, pages 416–423. Elsevier, 2015.

[128] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. NIH image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671, 2012.

[129] Ann-Kathrin Eisfeld, Sebastian Schwind, Kevin W Hoag, Christopher J Walker, Sandya Liyanarachchi, Ravi Patel, Xiaomeng Huang, Joseph Markowitz, Wenrui Duan, Gregory A Otterson, et al. NRAS isoforms differentially affect downstream pathways, cell growth, and cell transformation. *Proceedings of the National Academy of Sciences*, 111(11):4179–4184, 2014.

[130] Vikas K Goel, Alexander JF Lazar, Carla L Warneke, Mark S Redston, and Frank G Haluska. Examination of mutations in BRAF, NRAS, and PTEN in primary cutaneous melanoma. *Journal of Investigative Dermatology*, 126(1):154–160, 2006.

[131] Gatien Moriceau, Willy Hugo, Aayoung Hong, Hubing Shi, Xiangju Kong, C Yu Clarissa, Richard C Koya, Ahmed A Samatar, Negar Khanlou, Jonathan Braun, et al. Tunable-combinatorial mechanisms of acquired resistance limit the efficacy of BRAF/MEK cotargeting but result in melanoma drug addiction. *Cancer Cell*, 27(2):240–256, 2015.

[132] Chunying Song, Marco Piva, Lu Sun, Aayoung Hong, Gatien Moriceau, Xiangju Kong, Hong Zhang, Shirley Lomeli, Jin Qian, C Yu Clarissa, et al. Recurrent tumor cell-intrinsic and-extrinsic alterations during MAPKi-induced melanoma regression and early adaptation. *Cancer Discovery*, 7(11):1248–1265, 2017.

[133] Alexander S Mikheyev and Mandy MY Tin. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, 2014.

[134] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6, 2017.

[135] Gary K Geiss, Roger E Bumgarner, Brian Birditt, Timothy Dahl, Naeem Dowidar, Dwayne L Dunaway, H Perry Fell, Sean Ferree, Renee D George, Tammy Grogan, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology*, 26(3):317, 2008.

[136] Fatemeh Seyednasrollah, Asta Laiho, and Laura L Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16(1):59–70, 2013.

[137] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562, 2012.

[138] Shuhua Fu, Yingke Ma, Hui Yao, Zhichao Xu, Shilin Chen, Jingyuan Song, Kin Fai Au, and Bonnie Berger. IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*, 1:9, 2018.

[139] Andrew McDavid, Greg Finak, Pratip K Chattopadyay, Maria Dominguez, Laurie Lamoreaux, Steven S Ma, Mario Roederer, and Raphael Gottardo. Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, 29(4):461–467, 2012.

[140] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 42(14):8845–8860, 2014.

[141] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6):e1004333, 2015.

[142] Serena Liu and Cole Trapnell. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research*, 5, 2016.

[143] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145, 2016.

[144] Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75, 2017.

[145] S Steven Potter. Single-cell RNA sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology*, 14(8):479, 2018.

[146] Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4):593–607, 2016.

[147] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

[148] Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E Sun, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593, 2013.

[149] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[150] Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C Marioni. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, 33(5):503, 2015.

[151] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236, 2013.

[152] Paul W Hook, Sarah A McClymont, Gabrielle H Cannon, William D Law, A Jennifer Morton, Loyal A Goff, and Andrew S McCallion. Single-cell RNA-Seq of mouse dopaminergic neurons informs candidate gene selection for sporadic parkinson disease. *The American Journal of Human Genetics*, 102(3):427–446, 2018.

[153] Nathan G Skene, Julien Bryois, Trygve E Bakken, Gerome Breen, James J Crowley, Héléna A Gaspar, Paola Giusti-Rodriguez, Rebecca D Hodge, Jeremy A Miller, Ana B Muñoz-Manchado, et al. Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics*, page 1, 2018.

[154] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[155] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016.

[156] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377, 2009.

[157] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096, 2013.

[158] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[159] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[160] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8, 2017.

[161] Todd M Gierahn, Marc H Wadsworth II, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14(4):395, 2017.

[162] Dominic Grün and Alexander van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810, 2015.

[163] Geng Chen and Tieliu Shi. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10:317, 2019.

[164] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72, 2012.

[165] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1):63, 2016.

[166] Yan Song, Olga B Botvinnik, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L Xu, and Gene W Yeo. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular Cell*, 67(1):148–161, 2017.

[167] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mRNA quantification and differential analysis with census. *Nature Methods*, 14(3):309, 2017.

[168] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

[169] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1):44–73, 2017.

[170] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, 14(4):381, 2017.

[171] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, page 1, 2019.

[172] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8:15081, 2017.

[173] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17(1):173, 2016.

[174] Zhicheng Ji and Hongkai Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic Acids Research*, 44(13):e117–e117, 2016.

[175] Jaehoon Shin, Daniel A Berg, Yunhua Zhu, Joseph Y Shin, Juan Song, Michael A Bonaguidi, Grigori Enikolopov, David W Nauen, Kimberly M Christian, Guo-li Ming, et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*, 17(3):360–372, 2015.

[176] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241, 2015.

[177] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284, 2018.

[178] Michael I Love, Simon Anders, Vladislav Kim, and Wolfgang Huber. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*, 4, 2015.

[179] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637, 2014.

[180] Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biology*, 17(1):144, 2016.

[181] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

[182] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565, 2017.

[183] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. SC3: consensus clustering of single-cell rna-seq data. *Nature Methods*, 14(5):483, 2017.

[184] Dominic Grün, Mauro J Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik Jansen, Hans Clevers, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19(2):266–277, 2016.

[185] Peijie Lin, Michael Troup, and Joshua WK Ho. CIDR: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biology*, 18(1):59, 2017.

[186] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335, 2016.

[187] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015.

[188] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[189] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[190] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.

[191] Tianyu Wang, Boyang Li, Craig E Nelson, and Sheida Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, 20(1):40, 2019.

[192] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic,

et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, 2015.

[193] Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I Love, Davide Risso, Jean-Philippe Vert, Mark D Robinson, Sandrine Dudoit, and Lieven Clement. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biology*, 19(1):24, 2018.

[194] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, 2016.

[195] Zhun Miao, Ke Deng, Xiaowo Wang, and Xuegong Zhang. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34(18):3223–3224, 2018.

[196] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

[197] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. Gene expression recovery for single cell rna sequencing. *bioRxiv*, 2017.

[198] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[199] Shila Ghazanfar, Adam J Bisogni, John T Ormerod, David M Lin, and Jean YH Yang. Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. *BMC Systems Biology*, 10(5):11, 2016.

[200] Martin Slawski, Matthias Hein, et al. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.

[201] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[202] Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, 2011.

[203] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

[204] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[205] Glenn W Milligan and Martha C Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458, 1986.

[206] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[207] Natalija Novak, Carmen Tepel, Susanne Koch, Klaudia Brix, Thomas Bieber, and Stefan Kraft. Evidence for a differential expression of the FcεRIγ chain in dendritic cells of atopic and nonatopic donors. *Journal of Clinical Investigation*, 111(7):1047, 2003.

[208] Alexandru Schiopu and Ovidiu S Cotoi. S100A8 and S100A9: DAMPs at the crossroads between innate immunity, traditional risk factors, and cardiovascular disease. *Mediators of Inflammation*, 2013, 2013.

[209] Pei Wang, Ryan T Rodriguez, Jing Wang, Amar Ghodasara, and Seung K Kim. Targeting SOX17 in human embryonic stem cells creates unique strategies for isolating and analyzing developing endoderm. *Cell Stem Cell*, 8(3):335–346, 2011.

[210] Pei Wang, Kristen D McKnight, David J Wong, Ryan T Rodriguez, Takuya Sugiyama, Xueying Gu, Amar Ghodasara, Kun Qu, Howard Y Chang, and Seung K Kim. A molecular signature for purified definitive endoderm guides differentiation and isolation of endoderm from mouse and human embryonic stem cells. *Stem Cells and Development*, 21(12):2273–2287, 2012.

[211] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, 2009.

[212] Judith A Blake, Janan T Eppig, James A Kadin, Joel E Richardson, Cynthia L Smith, and Carol J Bult. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, 45(D1):D723–D729, 2017.

[213] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, 2017.

[214] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.

[215] Gerry P Quinn and Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002.

[216] Jeanette Baran-Gale, Tamir Chandra, and Kristina Kirschner. Experimental design for single-cell RNA sequencing. *Briefings in Functional Genomics*, 17(4):233–239, 2017.

[217] Martin J Zhang, Vasilis Ntranos, and David Tse. One read per cell per gene is optimal for single-cell RNA-Seq. *bioRxiv*, 2018.

[218] Bianca Dumitrascu, Karen Feng, and Barbara E Engelhardt. GT-TS: Experimental design for maximizing cell type discovery in single-cell data. *bioRxiv*, 2018.

[219] Simone Rizzetto, Auda A Eltahla, Peijie Lin, Rowena Bull, Andrew R Lloyd, Joshua WK Ho, Vanessa Venturi, and Fabio Luciani. Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Scientific Reports*, 7(1):12781, 2017.

[220] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, 65(4):631–643, 2017.

[221] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

[222] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[223] Imad Abugessaisa, Shuhei Noguchi, Michael Böttcher, Akira Hasegawa, Tsukasa Kouno, Sachi Kato, Yuhki Tada, Hiroki Ura, Kuniya Abe, Jay W Shin, et al. SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Research*, 46(D1):D781–D787, 2017.

[224] Yuan Cao, Junjie Zhu, Peilin Jia, and Zhongming Zhao. scRNASeqDB: A database for RNA-Seq based gene expression profiles in human single cells. *Genes*, 8(12):368, 2017.

[225] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, 2017.

[226] Aaron TL Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, 2016.

[227] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10):1053, 2014.

[228] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos

Tanay, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[229] Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, 2016.

[230] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

[231] A Yu Yen-Rei, Emily G OKoren, Danielle F Hotten, Matthew J Kan, David Kopin, Erik R Nelson, Loretta Que, and Michael D Gunn. A protocol for the comprehensive flow cytometric analysis of immune cells in normal and inflamed murine non-lymphoid tissues. *PloS One*, 11(3):e0150606, 2016.

[232] Sueli Marques, Amit Zeisel, Simone Codeluppi, David van Bruggen, Ana Mendanha Falcão, Lin Xiao, Huiliang Li, Martin Häring, Hannah Hochgerner, Roman A Romanov, et al. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, 352(6291):1326–1329, 2016.

[233] Koen Van den Berge, Charlotte Soneson, Michael I Love, Mark D Robinson, and Lieven Clement. zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications. *bioRxiv*, 2017.

[234] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[235] Davis J McCarthy, Kieran R Campbell, Aaron TL Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017.

[236] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421, 2018.

[237] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411, 2018.

[238] Aniruddha Chatterjee, Antonio Ahn, Euan J Rodger, Peter A Stockwell, and Michael R Eccles. A guide for designing and analyzing RNA-Seq data. In *Gene Expression Analysis*, pages 35–80. Springer, 2018.

[239] Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications*, 49(2):207–216, 1994.