Title
Statistical methods for analyzing mRNA isoform variation in large-scale RNA-seq data

Author
Demirdjian, Levon

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Statistical methods for analyzing mRNA isoform variation in large-scale RNA-seq data

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

by

Levon Demirdjian

2018

ABSTRACT OF THE DISSERTATION

Statistical methods for analyzing mRNA isoform
variation in large-scale RNA-seq data

by

Levon Demirdjian
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2018
Professor Ying Nian Wu, Chair

Alternative splicing (AS) is a major source of cellular and functional complexity in the eukaryotic transcriptome and plays a critical role in many developmental processes and diseases. Variations in AS are an important factor in disease-causing mutations, and it is hypothesized that over half of all known disease-causing mutations affect splicing patterns. Next-generation RNA sequencing (RNA-seq) technology has enabled the accumulation of large-scale sequencing data from diverse human tissues and populations and has provided an important resource for discovering variations in AS, yet the size and complexity of large-scale RNA-seq datasets continue to pose significant data analysis challenges to researchers. In this work, we propose new statistical methodologies that more effectively leverage complex RNA-seq data structures for studying AS.

In the first part of this work, we propose a sensitive and robust methodology called PAIRADISE for detecting genetic and allelic variation of alternative splicing in population-scale transcriptome datasets. PAIRADISE uses a novel statistical framework to detect allele-specific alternative splicing (ASAS) from population-scale RNA-seq data. A key feature of PAIRADISE is a statistical model that aggregates ASAS signals across multiple replicates of a given individual or multiple individuals in a population. PAIRADISE

consistently outperforms alternative statistical models in simulation studies, and boosts the power of ASAS detection when applied to replicate or population-scale RNA-seq data.

Next, we introduce the rMATS-Iso statistical framework for quantifying AS in modules with complex patterns of AS using replicate RNA-seq data. Importantly, rMATS-Iso leverages an EM algorithm to disambiguate short RNA-seq reads which may be consistent with multiple mRNA isoforms. As a result, rMATS-Iso can accommodate complex patterns of AS within a splicing module where transcripts can be defined by any combination of exons, splice site choices, etc. In addition, rMATS-Iso uses a likelihood ratio test to detect differential splicing between sample groups, and quantifies the extent to which each individual isoform contributes to the overall difference.

In conjunction with the continued development of next-generation sequencing methods, we anticipate that both PAIRADISE and rMATS-Iso will have broad utilities in elucidating the landscape of alternative splicing variation as well as other forms of mRNA isoform variation in human populations.

The dissertation of Levon Demirdjian is approved.

Yi Xing

Jingyi Li

Ariana Anderson

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2018

The deepest solace lies in understanding,

This ancient unseen stream,

A shudder before the beautiful

-Nightwish

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

If imitation is the highest form of flattery, then I can only hope to one day emulate the generosity and support that my advisors Ying Nian Wu and Yi Xing have given me over these past five years. I owe a special debt of gratitude to Shihao Shen who took me under his wing during my early days in the Xing lab and showed me all the ropes. Shihao's guidance has been remarkable and has undoubtedly helped me develop into a better statistician. I also want to thank Jessica Li and Ariana Anderson for their helpful and insightful suggestions that greatly improved the quality of this dissertation.

I have worked with a number of incredible collaborators over the last few years including Keith Heinzerling, Ray-Bing Chen, George J. Huffman, Majid Mojirsheibani, Hrayr Karagueuzian, and everybody in the Xing lab. I am particularly grateful to Yaping Zhou, my mentor at NASA Goddard Space Flight Center, for her encouragement and support both during and after my internship.

My friends in the Statistics department have added much needed levity into my daily routine. Thank you Guani, Min, Medha, and Hao for all of the stimulating (and not-so-stimulating) discussions and for introducing me to all of the hipster coffee joints around Westwood. My newfound caffeine addiction is on you guys.

I want to thank my parents for their superhuman patience and encouragement, and for giving me the freedom to carve out my own path in life. It takes an incredible degree of strength and trust to guide your children towards success without projecting your own personal aspirations onto them, and I will always appreciate the time and energy that my parents put towards achieving this goal. Finally, to my brother and sister Vahan and Meghrie: never stop pushing yourselves toward higher and greater limits.

VITA

2010 B.A in Economics, *summa cum laude*, California State University Northridge.

2011–2013 Teaching Assistant, Mathematics Department, California State University Northridge.

2013 M.S. in Mathematics, California State University Northridge.

2016 Intern, Climate and Radiation Lab, NASA Goddard Space Flight Center.

2014–2018 Teaching Assistant, Statistics Department, UCLA.

PUBLICATIONS

L. Demirdjian, M. Mojirsheibani (2018). *Classification on convex sets in the presence of missing covariates.* arXiv: 1805.00450v1 [math.ST]

L. Demirdjian, Y. Zhou, G.J. Huffman (2018). *Statistical modeling of extreme precipitation with TRMM data.* J. Appl. Meteor. Climatol. 57:1, DOI: 10.1175/JAMC-D-17-0023.1

L. Demirdjian (2017). *The Promise: When Truth Overshadows Power.* Significance Magazine, John Wiley & Sons Ltd., www.significancemagazine.com/culture/575-when-truth-overshadows-power.

E. Park, J. Guo, S. Shen, **L. Demirdjian**, Y.N. Wu, L. Lin, Y. Xing (2017). *Population and allelic variation of A-to-I RNA editing in human transcriptomes.* Genome Biology 18:143, DOI: 10.1186/s13059-017-1270-7

L. Demirdjian, M. Mojirsheibani (2017). *Kernel classification with missing data and the choice of smoothing parameters.* Statist. Papers. DOI: 10.1007/s00362-017-0883-y

K. Heinzerling, **L. Demirdjian**, Y.N. Wu, S. Shoptaw (2016). *Single nucleotide polymorphism near CREB1, rs7591784, is associated with pretreatment methamphetamine use frequency and outcome of outpatient treatment for methamphetamine use disorder.* J. Psychiatr. Res. 74:22-9, DOI: 10.1016/j.jpsychires.2015.12 .008

CHAPTER 1

Introduction

Nearly seventy years ago and soon after the discovery of DNA, it was a commonly held belief that each gene in the eukaryotic genome coded for a specific protein. According to this view, the information contained within genes was transferred to RNA and eventually into proteins, with a one-to-one correspondence between input (gene) and output (protein). A consequence of this assumption was the necessity of a vast number of genes, nearly 100,000, to explain the extraordinary degree of phenotypic complexity in mammals, yet all signs pointed to the fact that the actual number of genes was much smaller (Nilsen and Graveley, 2010; Pertea and Salzberg, 2010). To make matters even more complicated, additional observational evidence raised two other puzzles that this theory could not explain. First, there was simply too much DNA in the eukaryotic genome to correspond to the expected number of genes. Second, it was observed that RNA located within the nucleus of vertebrate cells was much longer than the messenger RNA (mRNA) in the cytoplasm, yet the tail ends of both the nuclear RNA and cytoplasmic mRNA had the same structure (they both had the same 5' cap structure and poly(A) tail at the RNA's 3' end) (Berk, 2016; Nilsen and Graveley, 2010; Sharp, 2005). Clearly then, this theory of information transfer from DNA to protein needed to be refined to comport with contemporary empirical findings.

About three decades after these questions had been raised, scientists discovered that a single gene coded for two distinct forms of immunoglobulin in antibodies (Alt et al., 1980; Early et al., 1980; Nilsen and Graveley, 2010). Though the mechanisms driving

this phenomenon were not understood at the time (the phenomenon itself was considered to be unusual), an increasing body of evidence corroborated the finding that one gene could code for a multitude of structurally and functionally distinct proteins. Over the next few decades, experiment after experiment would confirm that this seemingly unusual mechanism, alternative splicing, was not unusual at all; in fact, recent studies have shown that alternative splicing is in fact a very common phenomenon, with nearly all multi-exon human genes undergoing some form of alternative splicing (Pan et al., 2008; Wang et al., 2008). In conjunction with processes such as the use of alternative transcription start sites, RNA editing and post-translational modification, alternative splicing enables an incredible degree of phenotypic diversity that is a hallmark of all complex organisms (Nilsen and Graveley, 2010). Notably, the discovery of alternative splicing explained all of the puzzles that had been plaguing scientists nearly thirty years earlier.

1.1 Transcriptomics and Alternative Splicing

The transcriptome is the set of all RNA molecules within a biological sample. There can be significant variation in the transcriptomes of different tissues and cell types and these variations eventually manifest themselves as changes in the proteins that the RNA molecules code for. Before RNA is coded into protein, however, it is first converted into mRNA, and through a process known as alternative splicing, a single gene can result in multiple distinct mRNA transcripts, each of which is referred to as an mRNA isoform. The field of transcriptomics aims to study these variations and their broader biological consequences, e.g. their implications in disease phenomena.

Formally, pre-mRNA alternative splicing (AS) is a complex biological process involving the differential use of exons (protein-coding regions of a gene) and splice sites, enabling a number of distinct mRNA isoforms to be produced from a single gene (Fig 1.1A) (Nilsen and Graveley, 2010; Sharp, 1994). Note that AS is different than constitutive splicing,

where splicing proceeds the same way in every pre-mRNA of a given gene. The basic patterns of AS include exon skipping, alternative 5' and 3' splice site choice, mutually exclusive exons, intron retention, and alternative splicing coupled with alternative first or last exons (Fig 1.1B). Exon skipping is the most common AS event in humans and is characterized by the inclusion or removal of an exon (often referred to as a cassette exon) from a pre-mRNA molecule, resulting in two distinct mRNA isoforms of which only one contains the cassette exon. In addition to the basic patterns of AS, there exist more complex patterns that further increase the number of possible mRNA isoforms generated by a single gene. In extreme cases, such complex AS behavior can generate thousands or even tens of thousands of distinct mRNA isoforms. A notable, albeit extreme example is illustrated by the Drosophila melanogaster gene DSCAM (Down syndrome cell adhesion molecule). The DSCAM gene contains four clusters of mutually exclusive exons (i.e. exactly one exon from each cluster is transcribed), where the clusters are composed of 12, 48, 33, and 2 alternative exons respectively. As a result, this single gene can code for up to 38,016 different mRNA isoforms, more than twice the number of genes in the species (Venables et al., 2012; Schmucker et al., 2000).

Alternative splicing is regulated by extensive RNA-protein interactions involving cis elements (elements within the pre-mRNA) as well as trans-acting factors - elements outside of the pre-mRNA like RNA binding proteins which interact with these cis elements (Wang and Burge, 2008; Fu and Ares, 2014). Cis elements involved in splicing can be divided into two categories: splicing signals, and splicing regulatory elements. Splicing signals include the 5' and 3' splice sites defining the boundaries between an intron and its upstream and downstream exon, respectively, and the branch site and polypyrimidine tract. These splicing signals are recognized by the spliceosome, the core splicing machinery, and play a critical role in exon and intron definition (Wang and Burge, 2008). In contrast, splicing regulatory elements include exonic splicing enhancers (ESEs), intronic splicing enhancers

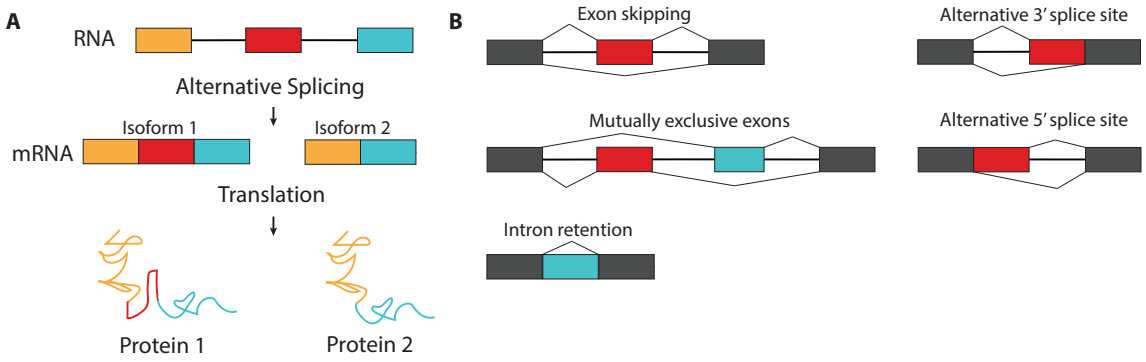

Figure 1.1: (A) Alternative splicing permits the synthesis of multiple distinct mRNA isoforms and proteins from a single gene. The orange and teal exons represent constitutive exons, while the red exon is an alternatively spliced exon. (B) The basic patterns of alternative splicing include exon skipping, alternative $3'/5'$ splice site selection, mutually exclusive exons, and intron retention. Black exons represent constitutive exons, while red and teal exons represent alternatively spliced exons. Horizontal black lines represent introns, and the line segments above and below the exons represent different patterns of splicing.

(ISEs), exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs), all of which are regulatory motifs that interact with RNA-binding proteins to regulate splicing (Wang and Burge, 2008; Wang and Cooper, 2007). The most thoroughly studied RNA-binding proteins involved in splicing regulation include SR proteins (named for their repeating regions of serine (S) and arginine (R) amino acids), which bind to both ESEs and ISEs to promote exon splicing, and heterogeneous nuclear ribo-nucleoproteins (hnRNPs), which bind to both ESSs and ISSs to suppress exon splicing (Nilsen and Graveley, 2010; Fu and Ares, 2014; Lin and Fu, 2007; Martinez-Contreras et al., 2007).

1.2 Alternative Splicing and Disease

Variations in alternative splicing are an important factor in disease-causing mutations, and it is hypothesized that over half of all known disease-causing mutations affect splicing patterns (Wang and Cooper, 2007; López-Bigas et al., 2005). Cancer and various neurodegenerative diseases have been shown to be associated with departures from normal splicing patterns. Notably, several cis-acting mutations that modify the splicing of cancer-relevant genes have been linked with cancer initiation and progression. For example, exon 7 skipping in the CDH17 gene (which codes for liver intestine cadherin, a cell-cell adhesion protein) has been associated with poor outcomes and high incidence of tumor recurrence in patients with hepatocellular carcinomas and gastric and pancreatic cancers (Srebrow and Kornblihtt, 2006). Importantly, alternative splicing occurs in every category of cancer hallmarks (Liu and Cheng, 2013), with differential splicing between cancer and healthy cells. Leveraging these differences in splicing patterns, recent advances in statistical modeling have demonstrated the clinical utility of using mRNA isoform ratios to predict cancer patient survival times (Shen et al., 2016). Proper identification of such changes in alternative splicing can be a key step in disease classification, and can ultimately enable the development of therapeutic approaches aimed at altering the splicing patterns of target genes. A promising example of this approach concerns spinal muscular atrophy, where artificial enhancement of exon 7 inclusion in the SMN2 gene has been associated with restoration of SMN protein back to normal levels in cultured human cells (Lorson et al., 2010; Hua et al., 2011; Kornblihtt et al., 2013).

1.3 Tissue Specific Alternative Splicing

Variations in alternative splicing play a vital role in tissue-specific functions, and tissue specific AS tends to be associated with functional changes in the underlying tissues (for

example, AS events in brain tissue are associated with neural-specific functions) (Kornblihtt et al., 2013). An interesting case is that of infrared detection in vampire bats. Alternative splicing of the TRPV1 (transient receptor potential cation channel V1) gene lowers thermal activation of the TRPV1 channel to boost vampire bats' infrared radiation detection capabilities. This AS event is both species specific and tissue specific, as this AS event occurs in the trigeminal ganglia of the bats but not in the dorsal root ganglia (Gracheva et al., 2011). Recognizing the importance of tissue-specific and species-specific splicing, it was long speculated that AS is an evolutionary advantage distinguishing humans from other primates and primates from other vertebrates. Empirical findings have corroborated this viewpoint by showing increased AS frequency in primate organ tissue compared to other vertebrate species. Moreover, of all species and tissue types studied, the human cerebellum displays the most pronounced degree of AS (Barbosa-Morais, Nuno L et al., 2012).

1.4 High-Throughput Analysis of Alternative Splicing

Before the advent of RNA-sequencing in the late 2000's, alternative splicing was quantified primarily using three technologies: reverse transcription polymerase chain reaction (RT-PCR), expressed sequencing tags (ESTs), and microarrays. In RT-PCR, reverse transcriptase enzymes are used to convert RNA molecules into their complementary DNA sequences, which are then amplified using PCR (Percifield et al., 2014; Farrell, 2010). Sequencing of ESTs, which are short segments of full-length mRNA sequences, provided an unprecedented view into the diversity of mRNA and the landscape of AS in humans (Lee et al., 2004; Xing et al., 2006). In contrast to low throughput techniques like RT-PCR and ESTs, splicing microarrays provide a high-throughput solution for studying the effects of splicing across biological conditions, though like RT-PCR, splicing microarrays are limited to known splicing events (Lee et al., 2004).

The recent development of next-generation RNA-sequencing (RNA-seq) technology has greatly facilitated the measurement and analysis of eukaryotic transcriptomes in a high-throughput manner, and has led to the discovery of novel splicing isoforms of known genes (Wang et al., 2009; Griffith et al., 2010). Briefly, an RNA-seq experiment entails converting a population of RNA to a library of complementary DNA (cDNA) fragments and attaching sequencing adaptors to each cDNA fragment. The molecules are then sequenced either from one end or both ends to obtain short sequences, which are subsequently aligned to a reference genome or transcriptome (Wang et al., 2009). Aligning the sequences to a reference genome or transcriptome is an important step that allows researchers to known how many sequences are generated from each exon in the region being studied.

RNA-seq can detect novel genes and mRNA isoforms, and the massively parallel nature of RNA-seq allows billions of short sequences, known as *reads*, to be generated in a single sequencing run (Wang et al., 2009); consequently, RNA-seq facilitates the quantitation of alternative splicing events. Due to the popularity of RNA-seq and continued reduction in the cost of sequencing, many large-scale RNA-seq datasets have been made publically available, including data from the Genotype-Tissue Expression (GTEx) study (Ardlie and Coauthors, 2015), the Geuvadis RNA-seq data of 445 B-lymphocyte cell lines from five populations (Lappalainen and Sammeth, 2013), as well data from The Cancer Genome Atlas (TCGA) study, which includes samples from over 10,000 individuals with 33 cancer types. Comparison of the splicing regulatory machinery between different tissues using RNA-seq has revealed substantial variation in splicing behavior between tissues (Wang and Cooper, 2007; Wang et al., 2009), underscoring the benefits of utilizing RNA-seq to compare AS between distinct biological conditions.

1.5 Quantifying Alternative Splicing Using RNA-seq

A commonly used metric for the quantification of alternative splicing is the PSI value ψ (Percent Spliced In), which is defined to be the percentage of a gene's mRNA transcripts which contain a given exon or splice site. For simple exon-skipping events, ψ values are estimated using RNA-seq reads supporting the inclusion isoform (i.e. reads mapping to the alternative exon body or to splice junctions of the alternative exon) or skipping isoform (i.e. reads joining the two constitutive exons) (Katz et al., 2010). For example, one of the simplest estimators of ψ is given by

$$\hat{\psi} = \frac{\ell_S I}{\ell_S I + \ell_I S}, \tag{1.1}$$

where I is the number of reads supporting the inclusion isoform, S is the number of reads supporting the skipping isoform, and where ℓ_I and ℓ_S are constants that account for the different lengths of both isoforms (discussed in more detail in the next chapter). The estimator $\hat{\psi}$ in (1.1) is often referred to as the *naive estimate* of ψ, since it ignores an important source of experimental variability: the total number of RNA-seq reads clearly affects the confidence in any estimate of PSI, with higher read coverage translating into more confident estimates of PSI, and thus any robust statistical framework modeling PSI values must consider this source of experimental variability. In fact, capturing and modeling this source of uncertainty has been shown to boost the results of subsequent statistical analyses (Shen et al., 2016; Katz et al., 2010; Shen et al., 2014).

A number of computational tools that utilize short-read RNA-seq data have recently been developed for quantifying and analyzing mRNA isoform expression and AS variation. These tools can be grouped into two categories: transcript-based tools, and event-based tools (see Park et al. (2018) for a comprehensive discussion comparing both approaches). Briefly, transcript-based tools aim to estimate the abundance and proportions of full-length mRNA isoforms. Methods falling into this category typically assign reads to their

corresponding full-length mRNA isoforms by using a generative statistical model in conjunction with an expectation-maximization (EM) type algorithm (Xing et al., 2006; Dempster et al., 1977) after aligning short RNA-seq reads to a reference transcriptome (though recent innovations in pseudoalignment (Bray et al., 2016) and lightweight mapping (Patro et al., 2017) algorithms have precluded the need for traditional read alignment). Event-based tools, on the other hand, seek to directly quantify AS and detect differential AS events using statistical methods. Some of the most popular event-based tools include MISO (Katz et al., 2010), rMATS (Shen et al., 2014), MAJIQ (Vaquero-Garcia et al., 2016), and SpliceTrap (Wu et al., 2011). Though these models differ in their underlying statistical assumptions, use of replicates, definition of splicing events etc., their estimates of psi values are largely in sync with one another.

1.6 Goal of this Dissertation

Though the development of RNA-seq technology has revolutionized transcriptome analysis, including the analysis of mRNA isoform variation, the size and complexity of large-scale RNA-seq datasets continue to pose significant data analysis challenges to researchers. To address this difficulty, we developed new statistical methodologies for the analysis of alternative splicing that leverage such complex data structures in more nuanced ways.

This dissertation primarily focuses on two complex data structures, the first of which is paired RNA-seq data, where the data are in the form of pairs of measurements taken on matched samples across groups. With the right analytical tools, paired RNA-seq data can help identify splicing differences among groups of matched samples, e.g. paired tumor vs. normal adjacent samples or heterozygous alleles in an allele-specific alternative splicing analysis. Importantly, we develop a sensitive and robust statistical methodology for paired RNA-seq data that addresses a gap in the existing literature on allele-specific alternative splicing, where the existing methods tend to be ad-hoc and fail to pool splicing

signals from multiple samples. Another challenge owing to the combinatorial nature of alternative splicing is the reconstruction of full-length isoforms from RNA-seq sequence fragments. RNA-seq reads do not capture the full-length mRNA sequence, but rather shorter fragments of the target sequence (Wang et al., 2009). As a result, it can be difficult to infer the underlying isoform structure that generated the RNA-seq read in the case where more than one isoform is compatible with the generated sequence. Furthermore, it is difficult to assess the functional impact of a splice variant without knowing its corresponding full-length transcript (Boue et al., 2002). To address this difficulty, we develop a new statistical framework called rMATS-Iso which is a generalization of the existing rMATS statistical model for differential alternative splicing analysis. By redefining the observed data to be isoform consistency counts instead of counts unique to each isoform, rMATS-Iso circumvents the central problem posed by ambiguous RNA-seq reads.

The methods we develop address both of these problems and provide a fundamental set of tools for elucidating the transcriptome. We are confident that these tools will help shed light on important biological processes and will ultimately enhance our understanding of human health.

CHAPTER 2

Detecting Allele-Specific Alternative Splicing from Population-Scale RNA-seq Data

The analysis of alternative splicing often involves a study design in which two samples are matched to one another; this type of design is natural, for example, when comparing splicing levels in paired tumor-normal samples or in pre-post therapy samples. In general, a paired design can increase statistical power by reducing individual level variation. In this chapter, we introduce a statistical and computational framework called PAIRADISE developed specifically for paired RNA-seq data, where the data are in the form of pairs of measurements taken on matched samples across sample groups. In particular, PAIRADISE fills an important methodological gap for studying the role of genetic variation in alternative splicing, namely, the analysis of allele-specific alternative splicing (ASAS). By treating the two alleles of an individual as paired, and multiple individuals sharing a heterozygous SNP as replicates, PAIRADISE frames ASAS detection as a statistical problem for identifying differential alternative splicing from RNA-seq data with paired replicates.

We begin this chapter with an introduction to genetic variation of alternative splicing, highlighting the differences between the sQTL and ASAS paradigms. Next, we introduce the PAIRADISE statistical framework, comparing and contrasting it to its precursor, the rMATS paired statistical model. We demonstrate the performance of PAIRADISE using a simulation study, where PAIRADISE outperforms every alternative statistical

model in every simulation setting. Finally, we apply PAIRADISE to replicate RNA-seq data including a population-scale RNA-seq dataset, and demonstrate the ability of PAIRADISE ASAS analysis to detect the effects of rare variants on alternative splicing.

2.1 Introduction

As discussed in Chapter 1, the relationship between aberrant alternative splicing (AS) and disease has been thoroughly studied over the last several decades and is relatively well understood. In contrast, the relation between naturally occurring variability in AS and phenotypic diversity and disease susceptibility in humans has only recently received commensurate attention, thanks in large part to the advent of RNA-seq and the accumulation of population-scale RNA-seq data for diverse human tissues and cell types (Park et al., 2018).

Genetic variation of alternative splicing, such as cis-acting sequence polymorphisms, can modulate complex traits and diseases in human individuals (Lu et al., 2012; Manning and Cooper, 2017). Splicing quantitative trait loci (sQTL) analysis is a widely used approach to uncover genetic variation of alternative splicing. In an sQTL analysis, the splicing level of a given exon or splice site is treated as a quantitative trait and tested for association with genotype across a population. A variety of computational tools have been developed for identifying sQTLs (Zhao et al., 2013; Ongen and Dermitzakis, 2015; Monlong et al., 2014; Jia et al., 2015; Yang et al., 2017), including GLiMMPS (Zhao et al., 2013), a generalized linear mixed model for sQTL association testing that accounts for the measurement uncertainty of mRNA isoform ratios in RNA-seq data. Analyses of population-scale RNA-seq and genotype data have revealed thousands of sQTLs in human genes, including numerous sQTLs associated with genome-wide association study (GWAS) signals of human traits or diseases (Park et al., 2018).

An alternative strategy for uncovering associations between sequence polymorphisms and alternative splicing is allele-specific alternative splicing (ASAS) analysis. ASAS analysis identifies differential splicing events between mRNA transcripts originating from two different haplotypes within an individual. Specifically, heterozygous SNPs present in mRNAs are used to assign RNA-seq reads to two alleles, and differential splicing between the two alleles is tested using RNA-seq read counts (Lappalainen and Sammeth, 2013; Li et al., 2012; Tilgner et al., 2014). A feature of the ASAS approach, compared with the sQTL approach, is that the two alleles of a single individual should share an identical cellular environment, thus splicing differences between the two alleles should arise from genetic effects. However, while a number of statistical models and computational tools have been developed for sQTL analysis (Zhao et al., 2013; Ongen and Dermitzakis, 2015; Monlong et al., 2014; Jia et al., 2015; Yang et al., 2017), rigorous methods for ASAS analysis are lacking. The approaches used in previous work were ad hoc and had important methodological limitations. ASAS was often discovered as allele-specific expression of individual exons, but such exon expression is itself confounded by allele-specific gene expression (Tilgner et al., 2014; Skelly et al., 2011; Van De Geijn et al., 2015). Moreover, ASAS events were detected in one cell line or individual at a time, by comparing isoform-specific read counts (e.g. Fisher exact test of exon inclusion vs skipping counts) between the two alleles (Lappalainen and Sammeth, 2013; Li et al., 2012; Tilgner et al., 2014). However, by performing ASAS analysis for each individual separately, signals from multiple individuals were not combined, likely reducing the statistical power.

2.2 The rMATS Paired Statistical Model

As mentioned above, there are a lack of robust statistical methods for the analysis of ASAS. A clever way to address this shortcoming is to frame the problem of ASAS detection as a specialized case of differential AS analysis with paired replicates; in this scenario, two

alleles within each individual are paired while replicate samples, or multiple individuals in a population, represent replicates. Importantly, setting the problem up in this manner allows one to leverage existing methods for differential AS analysis with paired replicates. One such popular tool is the rMATS paired statistical model (Shen et al., 2014), which utilizes a hierarchiecal design to simultaneously capture the estimation uncertainty in exon inclusion levels within individual samples/replicates, as well as the uncertainty across samples/replicates.

For the exon skipping type of AS event, the rMATS paired model utilizes RNA-seq reads specific to the exon inclusion and exon skipping isoforms in order to estimate exon inclusion levels in two sample groups, then uses a likelihood ratio test to detect differential splicing (see Figure 2.1 for a schematic diagram of the rMATS framework). For exon i, sample group $j = 1, 2$, and replicate $k = 1, \ldots, M$, let I_{ijk} denote the count of RNA-seq reads unique to the exon inclusion isoform and let S_{ijk} denote the count of reads unique to the exon skipping isoform. In addition, let ψ_{ijk} denote the corresponding exon inclusion level; rMATS considers ψ_{i1k} and ψ_{i2k} to be latent, unobserved variables specified by the following bivariate normal distribution:

$$\begin{bmatrix} \text{logit}(\psi_{i1k}) \\ \text{logit}(\psi_{i2k}) \end{bmatrix} \overset{iid}{\sim} \text{Normal}\left(\mu_i = \begin{bmatrix} \text{logit}(\psi_{i1}) \\ \text{logit}(\psi_{i2}) \end{bmatrix}, \Sigma_i = \begin{bmatrix} \sigma_{i1}^2 & \rho_i \sigma_{i1} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & \sigma_{i2}^2 \end{bmatrix} \right) \tag{2.1}$$

for $k = 1, \ldots, K$. The distribution in (2.1) models the variability in exon inclusion levels between replicates; notably, (2.1) utilizes a correlation parameter ρ_i to model the joint variability of $\text{logit}(\psi_{i1k})$ and $\text{logit}(\psi_{i2k})$, though the underlying biological interpretation of ρ_i is unclear. $\text{logit}(\psi_{i1}), \text{logit}(\psi_{i2}), \sigma_{i1}^2$, and σ_{i2}^2 represent group specific mean and variance parameters, respectively.

In addition to modeling the group specific exon inclusion levels, the rMATS paired model captures the variability in RNA-seq read counts for each sample group using the

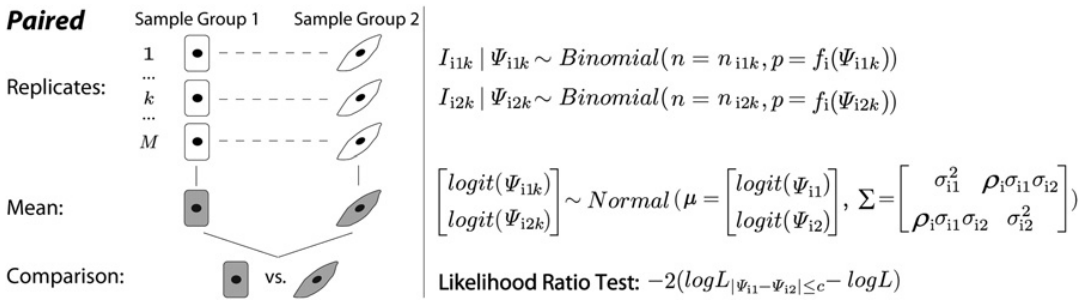

Figure 2.1: rMATS paired statistical model for differential alternative splicing analysis (figure reproduced from (Shen et al., 2014)). rMATS uses a hierarchical design to model the uncertainty resulting from the RNA-seq read coverage, as well as the variability in splicing levels across individuals/replicates from a population.

following binomial distributions:

$$I_{i1k}|\psi_{i1k} \sim \text{Binomial}\left(n_{i1k} = I_{i1k} + S_{i1k}, p_{i1k}(\psi_{i1k}) = \frac{\ell_{iI}\psi_{i1k}}{\ell_{iI}\psi_{i1k} + \ell_{iS}(1-\psi_{i1k})}\right),$$

$$I_{i2k}|\psi_{i2k} \sim \text{Binomial}\left(n_{i2k} = I_{i2k} + S_{i2k}, p_{i2k}(\psi_{i2k}) = \frac{\ell_{iI}\psi_{i2k}}{\ell_{iI}\psi_{i2k} + \ell_{iS}(1-\psi_{i2k})}\right), \qquad (2.2)$$

where ℓ_{iI} and ℓ_{iS} are the effective read lengths (i.e. number of unique isoform specific read positions) of the exon inclusion and exon skipping isoforms, respectively. p_{i1k} and p_{i2k} are deterministic length normalization functions which normalize the exon inclusion levels by the effective lengths of the isoforms:

$$p_{ijk}(\psi_{ijk}) = \frac{\ell_{iI}\psi_{ijk}}{\ell_{iI}\psi_{ijk} + \ell_{iS}(1-\psi_{ijk})}, \quad j = 1, 2.$$

The probability of observing an RNA-seq read corresponding to a specific isoform depends not only on the probability ψ, but also on the length (number of base pairs) of the isoform. For example, when $\psi = 0.5$ and in the absence in any bias in RNA-seq read coverage, we can expect more RNA-seq reads to correspond to the inclusion isoform simply by virtue of the fact that there are more read positions corresponding to the inclusion isoform than for

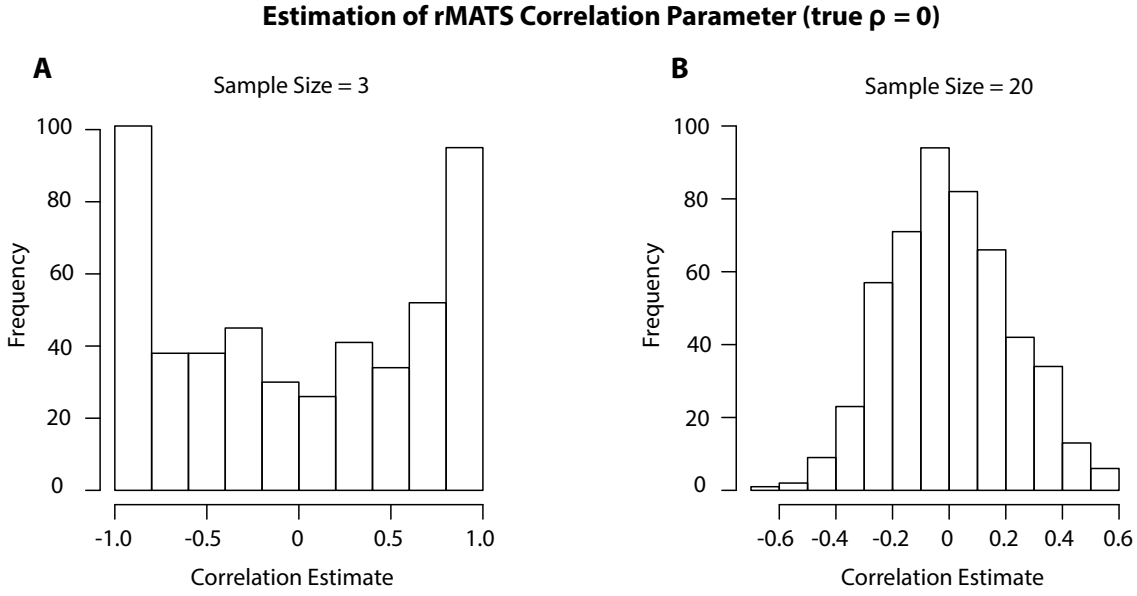

Figure 2.2: (A) When the number of samples/replicates is small, estimation of the rMATS paired model's correlation parameter is inaccurate and skewed towards 1 and -1. (B) This issue resolves with larger sample sizes. In both plots, the true correlation parameter ρ is equal to 0.

the skipping isoform. Thus, the function p_{ijk} is necessary to control for the confounding effects of the the different lengths of each isoform.

Though the rMATS paired model accounts for the distinct structure of paired RNA-seq data, it has several limitations. First, the model relies on the estimation of the correlation parameter ρ_i between the two paired sample groups, the interpretation of which is not straightforward. Moreover, when the number of replicates in a sample group is small, the estimation of the correlation parameter can be inaccurate (Figure 2.2). Finally, rMATS is limited to the RNA-seq analysis of only 5 types of pre-defined alternative splicing events.

2.3 The PAIRADISE Statistical Model

To address the limitations of the rMATS paired test, we developed PAIRADISE (Paired Analysis of Allelic Differential Splicing Events), a more general framework for testing count based ratio differences between sample groups. In addition to accounting for the uncertainty due to RNA-seq read coverage, PAIRADISE directly models the average difference in logit exon inclusion levels between two groups and circumvents the problem of estimating a correlation parameter. Moreover, PAIRADISE can be applied to several types of mRNA isoform variation, including alternative splicing, alternative polyadenylation, and RNA editing. Here we use alternative splicing, specifically exon skipping events, to illustrate the model and computational procedure.

PAIRADISE utilizes a hierarchical framework to detect ASAS by modeling the paired differences between the two alleles across a population. The PAIRADISE model simultaneously accounts for both the estimation uncertainty of alternative splicing levels in each allele within each individual (or replicate), and the variability in alternative splicing levels between alleles and across individuals (or replicates). The defining characteristic of the PAIRADISE statistical model is a simple and intuitive additive structure defining the variability in exon inclusion levels. More precisely, PAIRADISE assumes the logit transformed exon inclusion levels are given by

$$\begin{aligned} \text{logit}(\psi_{i1k}) &= \alpha_{ik} + \epsilon_{i1k}, \\ \text{logit}(\psi_{i2k}) &= \alpha_{ik} + \delta_i + \epsilon_{i2k}, \end{aligned} \tag{2.3}$$

where the subject effect for exon i, α_{ik}, is assumed to follow a normal distribution

$$\alpha_{ik} \overset{iid}{\sim} N(\mu_i, \sigma_i^2), \quad k = 1, \ldots, M; \tag{2.4}$$

in other words, $\alpha_{i1}, \ldots, \alpha_{iM}$ all follow the same normal distribution with mean μ_i and variance σ_i^2. In expression (2.3), we are assuming that

$$\epsilon_{i1k} \overset{iid}{\sim} N(0, \sigma_{i1}^2),$$
$$\epsilon_{i2k} \overset{iid}{\sim} N(0, \sigma_{i2}^2), \quad k = 1, \ldots, M, \tag{2.5}$$

and that ϵ_{i1k} and ϵ_{i2k} are independent of each other. The variable δ_i in (2.3) measures the expected difference between $\text{logit}(\psi_{i2k})$ and $\text{logit}(\psi_{i1k})$ conditional on α_{ik}, i.e.

$$\delta_i = E\left[\text{logit}(\psi_{i2k}) - \text{logit}(\psi_{i1k}) | \alpha_{ik}\right]. \tag{2.6}$$

Expression (2.3) illustrates how PAIRADISE decomposes the variability in exon inclusion levels into two sources: the variability that is *common* to both alleles/groups (given by the subject effect α_{ik}), and the variability that is *unique* to each allele/group (given by ϵ_{i1k} and ϵ_{i2k}). Equations (2.3), (2.4), and (2.5) imply the following joint distribution of $\text{logit}(\psi_{i1k})$ and $\text{logit}(\psi_{i2k})$:

$$\left(\begin{bmatrix} \text{logit}(\psi_{i1k}) \\ \text{logit}(\psi_{i2k}) \end{bmatrix} \middle| \alpha_{ik}, \sigma_{i1}, \sigma_{i2}, \delta_i \right) \sim N\left(\begin{pmatrix} \alpha_{ik} \\ \alpha_{ik} + \delta_i \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2 \end{pmatrix} \right). \tag{2.7}$$

The distribution in (2.7) illustrates a key advantage of PAIRADISE over the rMATS paired model: conditional on the subject effect α_{ik}, $\text{logit}(\psi_{i1k})$ and $\text{logit}(\psi_{i2k})$ are independent of eachother. Therefore, there is no need to estimate an additional correlation parameter as in the rMATS paired framework. Note that the likelihood function used in the optimization of the PAIRADISE model is based on the conditional distribution in (2.7), and not the marginal distributions given in 2.3 (more details about the optimization procedure are given in the appendix, section 2.11).

The second layer of the PAIRADISE statistical model, i.e. the distributions of observed RNA-seq read counts I_{i1k} and I_{i2k}, is the same as the corresponding layer in rMATS given in (2.2). A schematic diagram illustrating the PAIRADISE statistical model is given in Figure 2.3.

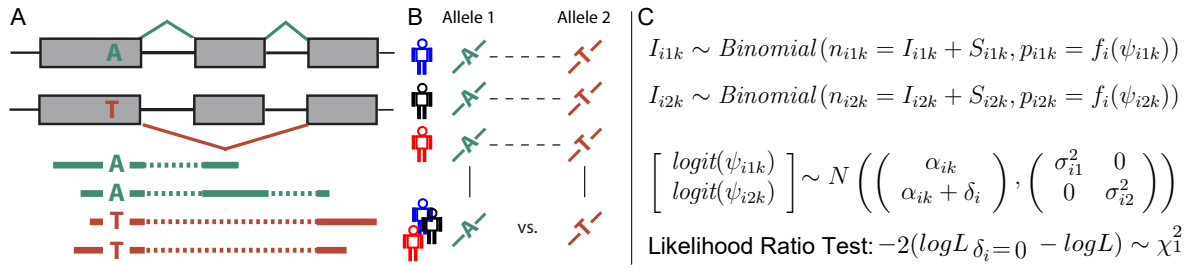

Figure 2.3: The PAIRADISE statistical framework for identifying allele-specific alternative splicing (ASAS). (A) ASAS analysis aims to identify differential alternative splicing between two alleles within an individual. Heterozygous SNPs are used to assign RNA-seq reads to specific alleles. (B) PAIRADISE aggregates ASAS signals across multiple replicates of a given individual or multiple individuals in a population. (C) Summary of the PAIRADISE statistical model.

In order to detect differential splicing between sample groups, PAIRADISE uses a likelihood ratio test to test the null hypothesis $\delta_i = 0$ against the alternative hypothesis $\delta_i \neq 0$ (a one sided alternative is also straightforward to implement). Since the variables $\text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k})$ and α_{ik} are regarded as latent (unobserved) variables, we utilize an optimization procedure that first calculates the maximum likelihood estimates (MLEs) of the observed data likelihood based on the current estimates of the latent variables, then using the current MLEs, updates the latent variables by maximizing the complete data likelihood. This procedure is performed under both the null and alternative hypotheses and iterated until the model parameters converge. The test statistics of the likelihood ratio test are then compared to a χ^2 distribution with one degree of freedom to derive the p-value. We utilize the Benjamini-Hochberg method to calculate the false discovery rates (FDRs) from p-values. Derivation of the likelihood functions and optimization algorithm are given in the appendix, Section 2.11.

2.4 Evaluating PAIRADISE Using a Simulation Study

To evaluate the performance of PAIRADISE, we compared it to four alternative statistical models using a simulation study. The four alternative models were the rMATS paired test (Shen et al., 2014), paired t-test, paired Wilcoxon signed-rank test, and Fisher's method. The paired t-test and paired Wilcoxon signed-rank test were conducted on point estimates of ψ values (which we denote by ψ_{naive}) derived from RNA-seq read counts, i.e.

$$\psi_{naive} = \frac{\ell_S I}{\ell_S I + \ell_I S}, \tag{2.8}$$

while ignoring the estimation uncertainty of ψ as influenced by sequencing coverage. Fisher's method uses Fisher's exact test on allele-specific read counts to obtain a p-value of ASAS for each individual separately, then uses the Fisher's combined probability test to aggregate p-values across all individuals. We designed a set of simulation studies with varying sample size (number of replicates) and variability among replicates, and measured the performance of each method by analyzing its receiver operating characteristic (ROC) curve for the task of classifying a simulated event as being differentially spliced versus non-differentially spliced.

More precisely, each simulation was performed by generating $5,000$ exon skipping events and varying the number of replicates ($M = 3, 5, 10, 20, 50$) as well as the variability among replicates, i.e. σ_{i1}, σ_{i2}, and σ_i. These standard deviations were chosen from the 1st, 2nd, and 3rd quartiles (corresponding to low, medium, and high variability) of their corresponding estimated distributions obtained from applying PAIRADISE to the Geuvadis CEU dataset (this dataset is described in more detail in section 2.6). Since the true values of the parameter δ_i were not known, to generate null ($\delta_i = 0$) and alternative ($\delta_i \neq 0$) cases, we set the middle 50% of the empirical estimates of δ_i to 0, then randomly sampled one value per event; as a result, roughly 50% of the events were generated from the null hypothesis of no splicing difference between groups. The remaining simu-

lation parameters, i.e. the total read counts n_{i1k} and n_{i2k}, effective lengths ℓ_{iI} and ℓ_{iS}, and mean logit inclusion level μ_i, were similarly obtained empirically from the Geuvadis dataset. The subject effects α_{ik} were sampled from the normal distribution in (2.4) using the empirically sampled parameter values. The logit exon inclusion values $\text{logit}(\psi_{i1k})$ and $\text{logit}(\psi_{i2k})$ were then sampled from the normal distributions given in (2.7) using the simulated value of α_{ik} and the empirically sampled parameter values. The read counts of the exon inclusion isoforms were then sampled from the binomial distributions given by (2.2) using the generated values for the exon inclusion levels as well as the sampled values for the total read counts and effective lengths. PAIRADISE and the other paired tests were applied to the simulated data to compute the p-value and FDR of differential splicing for each simulated event.

PAIRADISE outperformed all other statistical models in every simulation setting, based on the area under curve (AUC) of the ROC curve as well as the true positive rate (TPR) at 5% false positive rate (FPR; see Figure 2.4). The increased performance of PAIRADISE over other models was even more pronounced when the sample size was small. For example, in the simulations with 3 replicates and low variance, the AUC for PAIRADISE, rMATS paired test, paired t-test, paired Wilcoxon test, and Fisher's method were 81%, 74%, 74%, 71%, and 73%, respectively. PAIRADISE continued to outperform other methods in simulations with medium or high variance. We observed the same trend for the TPR at 5% FPR. Additionally, we note that in the low or medium variance setting, other models required roughly 2-3 times larger sample size to achieve the same AUC and TPR values as a sample size of 3 replicates for PAIRADISE. We also note that while almost all methods had better performance with increased sample size, Fisher's method had worse performance with large sample size in the medium and high variance settings. This is not surprising, as Fisher's method is particularly sensitive to outliers in large datasets (Loughin, 2004; Whitlock, 2005). Taken together, these

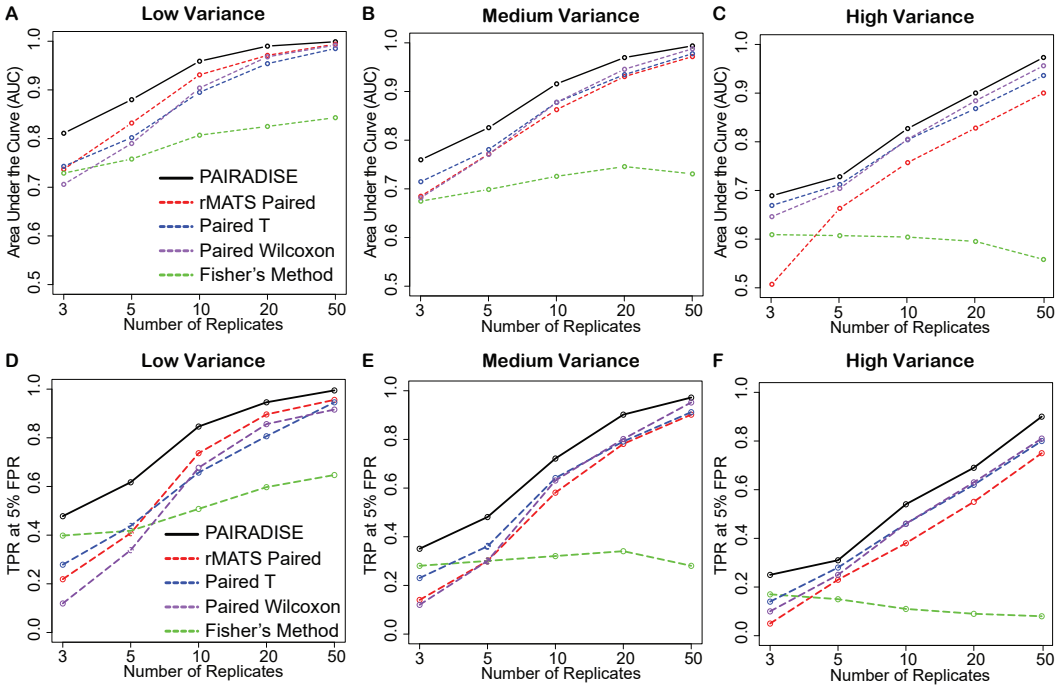

Figure 2.4: Simulation studies to compare the performance of PAIRADISE, rMATS paired model, paired t-test, paired Wilcoxon signed-rank test, and Fisher's method. (A-C) The area under curve (AUC) of all methods in simulation settings with the number of replicates equal to 3, 5, 10, 20, and 50, and three settings of variability (low, medium, high) sampled from the 1st, 2nd, and 3rd quartile of the empirical variance estimated from the Geuvadis dataset. (D-F) The true positive rate (TPR) at 5% false positive rate (FPR) of all methods in various simulation settings

simulation studies indicate that PAIRADISE outperforms other statistical models and requires fewer replicates to achieve the same level of performance.

2.5 Analysis of ASAS in the GM12878 Cell Line

To illustrate how PAIRADISE could be utilized to discover ASAS events in replicate RNA-seq data, we applied PAIRADISE to six RNA-seq biological replicates of the hu-

Gene	ASAS exon (hg19)	ASAS SNP	GWAS trait (SNP)
BAZ2A	-chr12:57024559-57024649	rs2255074	Mean platelet volume (rs2950390)
HLA-DOA	-chr6:32974856-32974992	rs365066	Platelet counts (rs399604)
HLA-DPA1	-chr6:33036795-33036984	rs1126543	Hepatitis B (rs3077)
ORC4	-chr2:148733470-148733544	rs897172	Urate levels (rs2307394)
PRR4	-chr12:11273608-11273779	rs2597984	Bitter taste perception (rs2708377)
SP110	-chr2:231037559-231037733	rs13018234	Obesity-related traits (rs13010639)
TLR1	-chr4:38806374-38806408	rs5743565	Alcohol consumption (rs4543123), Allergic sensitization (rs17616434), Asthma and hay fever (rs4833095), Helicobacter pylori serologic status (rs10004195)
TOMM7	-chr7:22857618-22857667	rs1054471	Fibrinogen (rs2286503)
TRDMT1	-chr10:17210839-17210916	rs2273734	Telomere length (rs10904887)
VAMP8	+chr2:85806134-85806290	rs1009	Prostate cancer (rs10187424)

Table 2.1: By combining the ASAS signals from all six GM12878 RNA-seq replicates, PAIRADISE identified 13 ASAS events linked ($r^2 > 0.8$) to GWAS signals.

man GM12878 B-lymphocyte cell line from a European female (more information about the data is provided in subsection 2.11.5). This setup allowed us to gauge the ability of PAIRADISE to aggregate ASAS signals over multiple RNA-seq replicates of a given individual. Using the SNP and haplotype information of GM12878, PAIRADISE identified 121 significant ASAS events at FDR $\leq 10\%$, of which 13 were in high ($r^2 > 0.8$) linkage disequilibrium (LD) with GWAS trait/disease-associated SNPs in the NHGRI GWAS catalog (a list of these SNPs can be found in Table 2.1).

By combining the ASAS signals from all six GM12878 RNA-seq replicates, PAIRADISE substantially increased the power of ASAS detection. Analyzing each individual replicate in isolation resulted in an average of 51 (between 30 to 61) significant ASAS events in each replicate; in contrast, applying PAIRADISE to all six replicates jointly resulted in 121 significant events. To validate the differential splicing events identified by PAIRADISE, we also performed an sQTL analysis using the GLiMMPS software (Zhao et al., 2013). More specifically, we applied GLiMMPS to the 89 CEU (Utah Residents with European

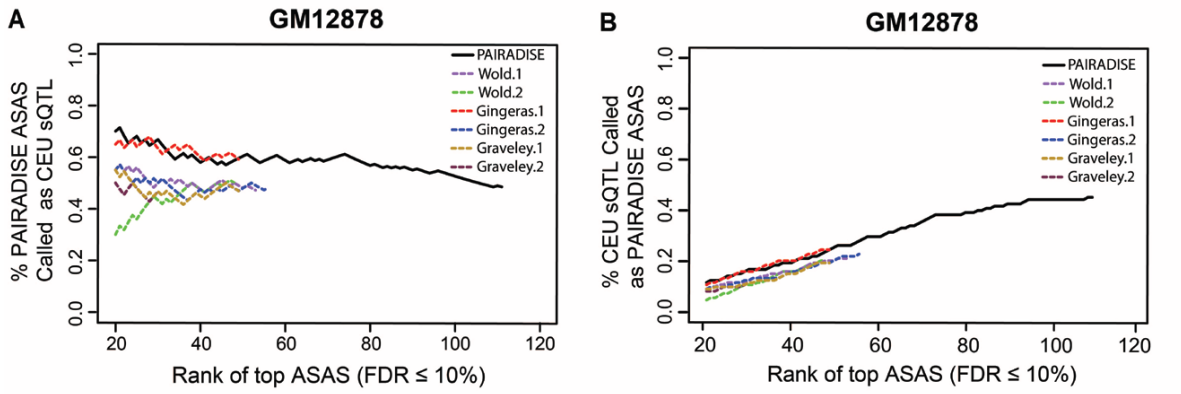

Figure 2.5: By aggregating signals from all six GM12878 RNA-seq replicates, PAIRADISE substantially boosts the power of ASAS detection. (A) The percentage of significant ASAS events that are also significant sQTL events against the ranking of ASAS events by PAIRADISE or in individual replicates. (B) The percentage of significant sQTL events that are also significant ASAS events against the ranking of ASAS events by PAIRADISE or in individual replicates. Both plots correspond to events analyzed by both PAIRADISE and GLiMMPS.

Ancestry) B-lymphocyte cell lines, whose RNA-seq and genotype data were available from the Geuvadis project.

Since PAIRADISE and GLiMMPS each use a different set of criteria for filtering out alternative spicing events before analysis, we restricted the present comparison to only focus on alternative exons surviving both sets of filters (see subsections 2.11.7 and 2.11.8 for more about the ASAS and sQTL specific filters we implemented). At FDR $\leq 10\%$ (permutation FDR based on 10 permutations), GLiMMPS identified 163 significant sQTLs in the CEU population, including 117 that were also analyzed by PAIRADISE for ASAS signals. To measure the concordance of the ASAS results with the sQTL results, we plotted the percentage of significant ASAS events that were also significant sQTL events (Figure 2.5A), as well as the percentage of significant sQTL events that were also signifi-

cant ASAS events (Figure 2.5B), against the ranking of ASAS events by PAIRADISE or in individual replicates. By aggregating all six biological replicates together, PAIRADISE identified two to four times as many significant ASAS events compared to the analysis of each individual replicate in isolation; moreover, using all six replicates together resulted in comparable and often higher concordance with the sQTL results. For example, 46% to 59% of ASAS events identified in individual replicates were also significant sQTLs. By contrast, 62% of the top 50 and 49% of all PAIRADISE ASAS events were also significant sQTLs.

2.6 Population-Scale Analysis of ASAS

We tested PAIRADISE on a population-scale RNA-seq dataset containing data from multiple populations and multiple individuals within those populations. Specifically, we applied PAIRADISE to the Geuvadis RNA-seq data of 445 B-lymphocyte cell lines taken from 89 CEU (Utah Residents with European Ancestry), 92 FIN (Finnish in Finland), 86 GBR (British in England and Scotland), 91 TSI (Toscani in Italia), and 87 YRI (Yoruba in Ibadan, Nigeria) individuals with both RNA-seq and genotype data. We tested PAIRADISE on each population separately, and at FDR $\leq$ 10%, PAIRADISE identified 111 ASAS events in CEU, 144 in FIN, 130 in GBR, 151 in TSI, and 180 in YRI respectively (Figure 2.6A). In addition to identifying 197 population-specific ASAS events, PAIRADISE identified 17 events that were significant in all five populations. There was a higher degree of overlap in ASAS events between the four European populations, while the YRI African population had the largest number, 84, of population-specific ASAS events.

As a basis for comparison, we also conducted an sQTL analysis on the Geuvadis dataset using GLiMMPS. At FDR $\leq$ 10%, GLiMMPS identified 163 significant sQTL events in CEU, 138 in FIN, 148 in GBR, 136 in TSI, and 212 in YRI (Figure 2.6B). GLiMMPS identified 301 population specific sQTL events, and 13 events that were significant in all

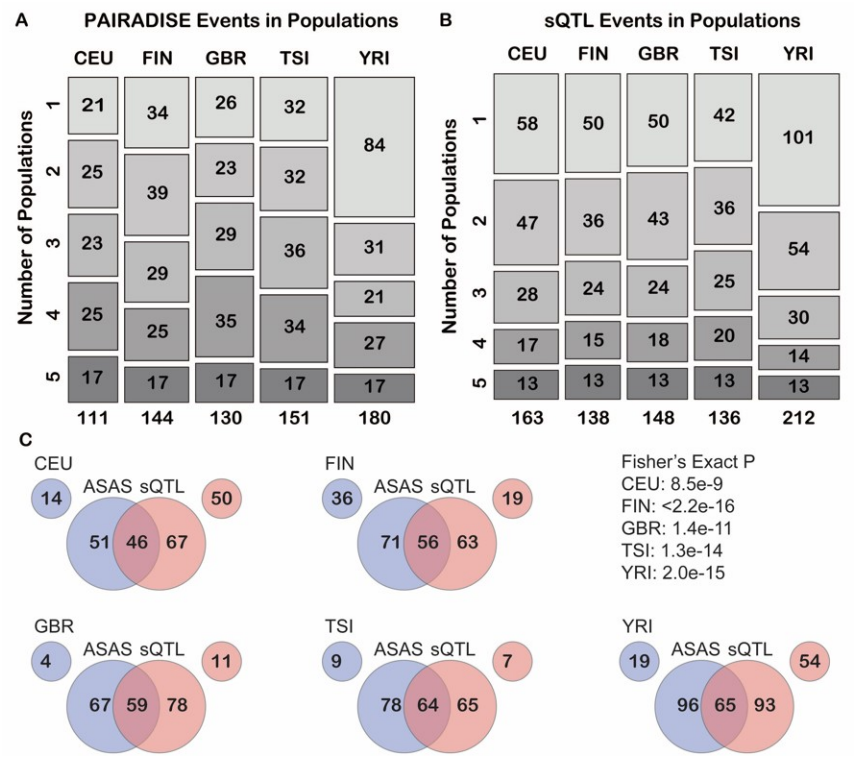

Figure 2.6: PAIRADISE analysis of ASAS in five Geuvadis populations. (A) Mosaic plot showing the number of significant ASAS events shared between five populations. Values in the top rectangles represent population specific ASAS events and values in the bottom rectangles represent ASAS events shared by all five populations. (B) Mosaic plot showing the number of significant sQTL events shared between five populations. Values in the top rectangles represent population specific sQTL events and values in the bottom rectangles represent sQTL events shared by all five populations. (C) Venn diagrams of ASAS events identified by PAIRADISE and sQTL events identified by GLiMMPS. The two outlying circles in each Venn diagram represent events only considered by one method due to method-specific filters and limitations. The p-values of overlap between ASAS events and sQTL events over random expectation (computed using Fisher's exact test) are provided for each population.

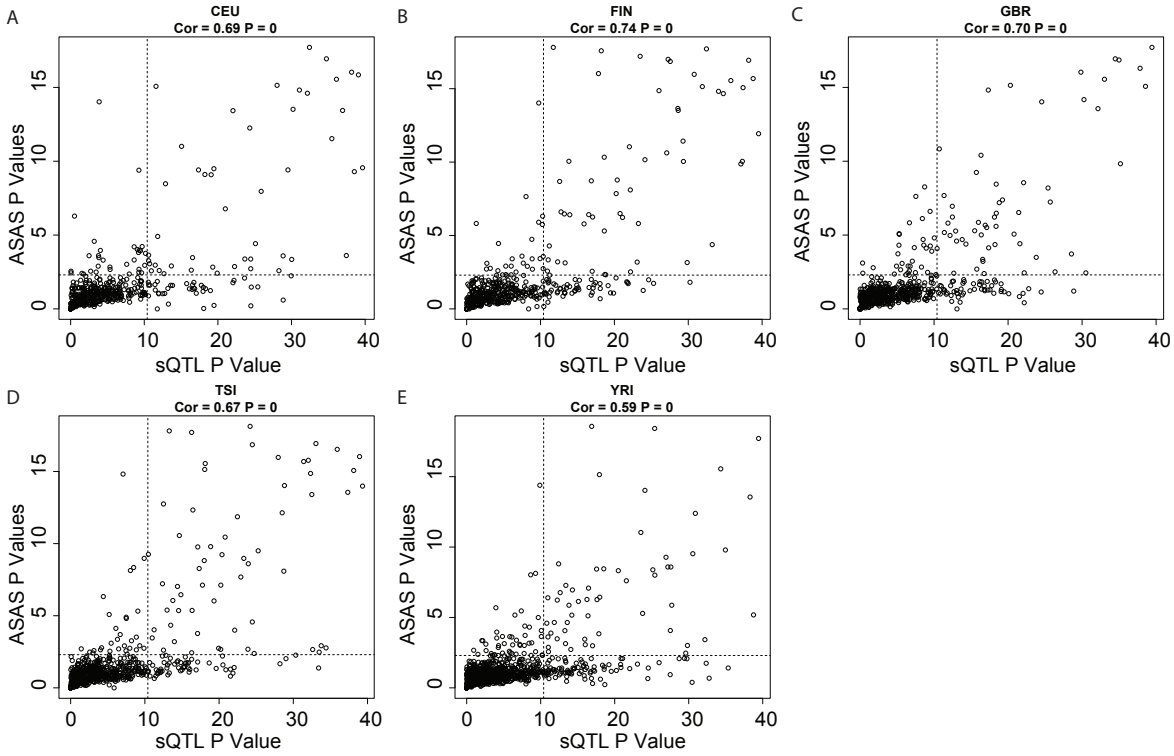

Figure 2.7: Concordance between sQTL p-values computed using GLiMMPS, and ASAS p-values computed using PAIRADISE on the five populations from the Guevadis dataset. The correlations ranged from 0.59 to 0.74 and were significant in all five populations.

five populations. As was the case for ASAS events, the YRI African population had the largest number, 101, of population-specific sQTLs.

Next, we compared the results of the ASAS analysis using PAIRADISE with the results of the sQTL analysis performed using GLiMMPs. There was a strong concordance between the ASAS signals detected by PAIRADISE and sQTL signals detected by GLiMMPS. Due to method-specific filters and limitations (discussed in subsections 2.11.7 and 2.11.8), certain splicing events in each population were only analyzed by one approach and not the other. For events analyzed by both approaches, the $-\log_{10}$ of ASAS p-values and sQTL p-values were correlated in all five populations (Pearson correlation $p \leq 2.2e^{-16}$, Figure 2.7). In addition, there was a significant overlap between significant

ASAS events and significant sQTL events in each population (Fig. 2.6C).

2.7 Functional Splicing Variation Identified by PAIRADISE

The ASAS events identified by PAIRADISE in Geuvadis often had important biological implications. For example, exon 4 of HLA-DQB1 (major histocompatibility complex, class II, DQ beta 1) had significant ASAS signals in both the CEU (Figure 2.8A) and YRI (Figure 2.8B) populations; the major G allele of SNP rs1049107 had significantly higher exon inclusion levels than the minor A allele across almost every individual heterozygous for this SNP in both populations. An sQTL analysis revealed a similar trend; individuals with the GG genotype had higher exon inclusion levels than heterozygous individuals, which in turn had higher exon inclusion levels than individuals with the AA genotype (see Figures 2.8C-D; also see Figure 2.8E for the sashimi plot of each genotype in the CEU population). The gene HLA-DQB1 encodes a cell surface receptor which plays an essential role in the proper functioning of the immune system (Shiina et al., 2009). The exon 4 skipping isoform of HLA-DQB1 lacks the sequence which encodes the transmembrane domain of DQβ, leading to the production of a soluble protein isoform that can modulate immune response and induce peripheral tolerance (Královičová et al., 2004).

A number of significant ASAS events were associated with SNPs identified in GWAS analyses. For example, PAIRADISE identified differential splicing in the A/G alleles of SNP rs1009 in exon 2 of VAMP8 (vesicle associated membrane protein 8) among CEU individuals (Figure 2.9A, PAIRADISE ASAS p-value = $3.7e^{-11}$). SNP rs1009 was also significantly associated with exon 2 splicing in an sQTL analysis (Figures 2.9B-C, GLiMMPS sQTL p-value = $6.8e^{-19}$). SNP rs1009 is in high LD ($r^2 = 0.97$) with GWAS SNP rs10187424; a GWAS analysis of $51,311$ individuals conducted by the international PRACTICAL consortium (Kote-Jarai and Coauthors, 2011) previously identified rs10187424 as being associated with prostate cancer susceptibility (Figure 2.9D).

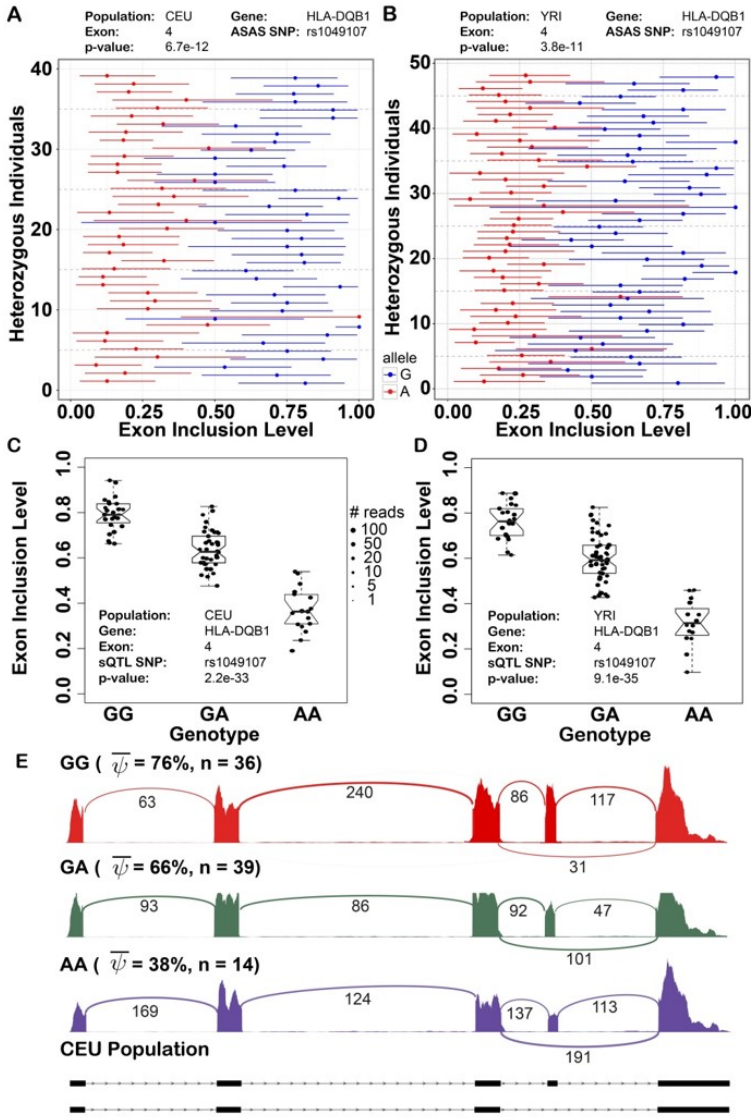

Figure 2.8: (A, B) A significant ASAS event involving SNP rs1049107 in the HLA-DQB1 gene identified by PAIRADISE in the CEU and YRI populations. The error bars around the exon inclusion levels represent 95% confidence intervals. (C, D) rs1049107 was also identified as a significant sQTL event by GLiMMPS; each dot represents the exon inclusion level of one individual, with dot sizes indicating the number of reads covering the splicing event for that individual. (E) Sashimi plots of the HLA-DQB1 gene with average exon read density and splice junction counts for the three genotypes in the CEU population.

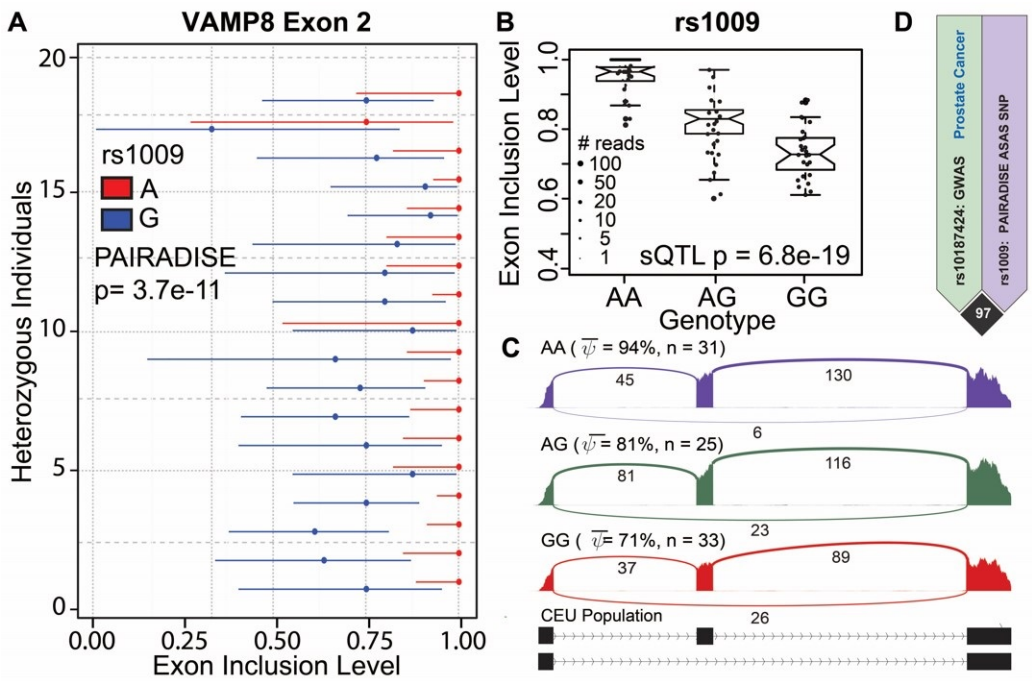

Figure 2.9: (A) A significant ASAS event identified by PAIRADISE corresponding to SNP rs1009 in exon 2 of the VAMP8 gene. The error bars around the exon inclusion levels represent 95% confidence intervals. (B) rs1009 was also identified as a significant sQTL by GLiMMPS. Each dot represents the exon inclusion level of one individual, with dot sizes indicating the number of reads covering the splicing event for that individual. (C) Sashimi plots of the VAMP8 gene with average exon read density and splice junction counts for the three genotypes in the CEU population. (D) rs1009 is in high LD ($r^2 = 0.97$) with GWAS SNP rs10187424.

In various genes, PAIRADISE identified ASAS events associated with multiple traits or diseases. One example is the exon 2 skipping event of TLR1 (toll like receptor 1), which shows consistent splicing differences between the A/G alleles of SNP rs5743565 in CEU individuals (Figure 2.10A, PAIRADISE ASAS $p = 5.2e^{-14}$). The same SNP was also significantly associated with exon 2 splicing in an sQTL analysis using GLiMMPS (Figures 2.10B-C, $p = 4.7e^{-31}$). Exon 2 of TLR1 corresponds to a 77-bp region in the

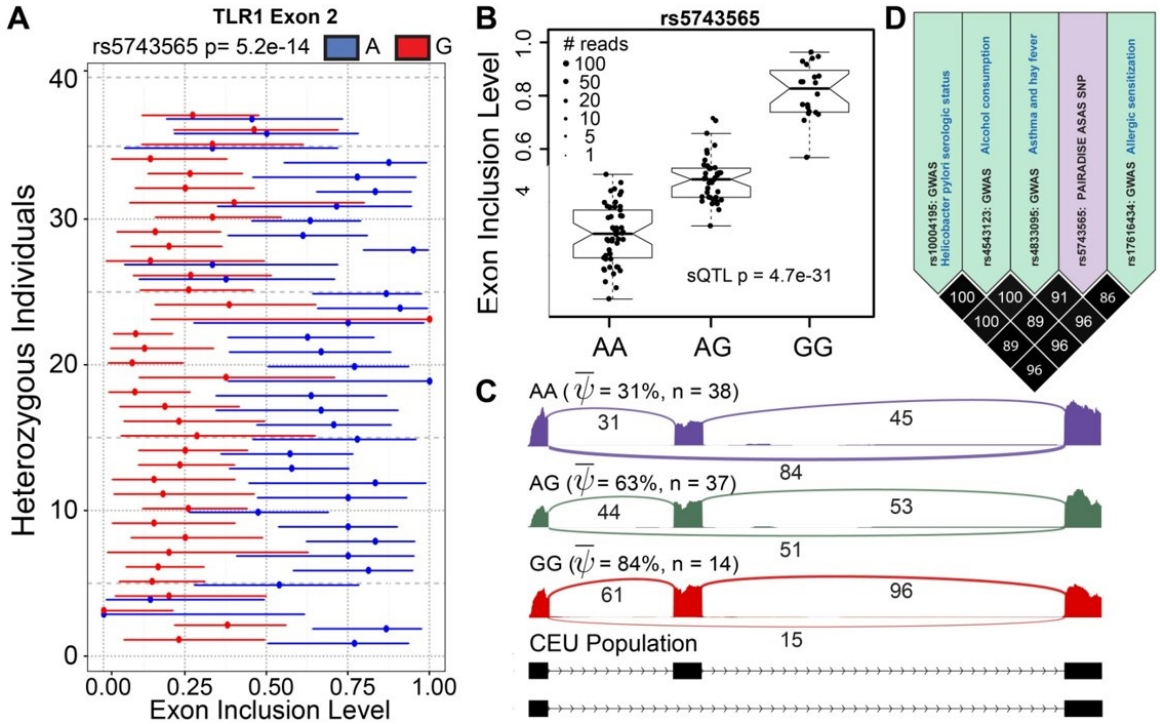

Figure 2.10: (A) A significant ASAS event identified by PAIRADISE corresponding to SNP rs5743565 in exon 2 of the TLR1 gene. The error bars around the exon inclusion levels represent 95% confidence intervals. (B) rs5743565 was also identified as a significant sQTL by GLiMMPS. Each dot represents the exon inclusion level of one individual, with dot sizes indicating the number of reads covering the splicing event for that individual. (C) Sashimi plots of the TLR1 gene with average exon read density and splice junction counts for the three genotypes in the CEU population. (D) rs5743565 is in high LD ($r^2 \geq 0.86$) with GWAS SNPs rs10004195, rs4543123, rs4833095, and rs17616434.

5'-untranslated region, and alternative splicing of this exon has been shown to alter TLR1 mRNA stability and steady-state level (Chang et al., 2006). SNP rs5743565 is in high LD with a number of GWAS SNPs: rs17616434 ($r^2 = 0.86$), associated with allergic sensitization (Bønnelykke and Coauthors, 2013); rs4833095 ($r^2 = 0.91$), associated with asthma (Daley et al., 2012); rs4543123 ($r^2 = 0.89$), associated with alcohol consumption (Kapoor et al., 2013); and rs10004195 ($r^2 = 0.89$), associated with helicobacter pylori

serologic status (Mayerle et al., 2013).

2.8 PAIRADISE Analysis of Rare Variants

One of the unique advantages of ASAS analysis relative to sQTL analysis is the ability to detect allelic differences in splicing levels in rare variants. sQTL analysis across individuals in a population is practically limited to focusing on common variants: for example, under Hardy-Weinberg equilibrium, only 0.04% of individuals can be expected to be homozygous for the minor allele for a variant with minor allele frequency (MAF) 2%, thus precluding the possibility of an sQTL analysis except for studies with especially large sample sizes. ASAS analysis does not face the same limitation since the percent of individuals expected to be heterozygous is 3.9%, roughly a 100-fold increase in expected frequency.

Among the significant ASAS events detected by PAIRADISE in GM12878 were 10 genetically regulated exon skipping events associated with rare variants (MAF $< 5\%$ in the CEU population). For example, an exon skipping event in IFI16 (interferon gamma inducible protein 16) was found to be significantly associated with SNP rs62621173. As remarked above, this SNP would conventionally be filtered out of an sQTL analysis due to its low MAF of 2% in CEU (Hernandez et al., 2017). In contrast, among the six RNA-seq replicates of GM12878, ASAS analysis using PAIRADISE revealed consistent splicing differences in exon inclusion levels between the two alleles, generating a significant ASAS signal (Figure 2.11A, PAIRADISE ASAS $p = 1.5e^{-5}$), with the minor T allele associated with higher exon inclusion levels. IFI16 acts as a sensor for intracellular DNA, thus playing a role in innate immunity (Unterholzner et al., 2010). The exon 9 skipping isoform of IFI16 contains one less copy of the 56-amino acids serine–threonine–proline (S/T/P)-rich spacer region within the protein product (Veeranki and Choubey, 2012). SNP rs62621173 has been linked with the age of onset of Alzheimer's disease (Vélez et al., 2016).

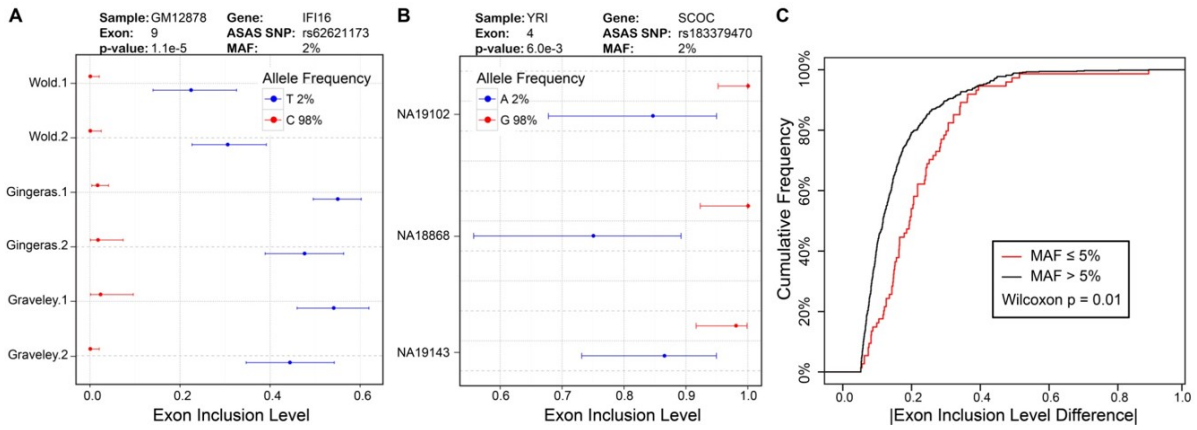

Figure 2.11: PAIRADISE identifies rare variants' effects on alternative splicing. (A) An ASAS event in the IFI16 gene with respect to SNP rs62621173 (CEU MAF: 2%; C: 98%, T: 2%) identified from the six RNA-seq replicates of GM12878. The error bars around the exon inclusion levels represent 95% confidence intervals. (B) An ASAS event in the SCOC gene with respect to SNP rs183379470 (CEU MAF: 2%; G: 98%; A: 2%) identified from three YRI individuals in Geuvadis. (C) The cumulative distribution function comparing the allelic difference of exon inclusion levels for ASAS events associated with rare variants (MAF $\leq$ 5%) or common variants (MAF $>$ 5%) in the five Geuvadis populations.

In addition to rare variants identified in GM12878, PAIRADISE also identified 63 significant ASAS events associated with rare variants in the five populations of the Geuvadis data. One example is a significant ASAS event in the SCOC (short coiled-coil protein) gene, which was associated with the rare variant rs183379470 (MAF = 2%), identified from three individuals in the YRI population (Figure 2.11B). The major G allele had an average exon inclusion level of 98% compared to 83% for the minor A allele.

Since fewer individuals will have rare variants compared to common variants, we can expect to see larger effect sizes for significant rare variant ASAS events in Geuvadis. Indeed the data support this hypothesis: the average allelic difference in exon inclusion levels was 19% for ASAS associated rare variants, as compared to 12% for common vari-

ants (Figure 2.11C; two-sided Wilcoxon $p = 0.01$). This phenomenon does not extend to GM12878, however, underscoring the fact that in general, larger effect sizes are not necessary for rare variant ASAS events to be significant. In GM12878, the average allelic difference in exon inclusion levels was 21% for ASAS associated rare variants, as compared to 19% for common variants (two-sided Wilcoxon $p = 0.88$).

2.9 Tumor-Specific Splicing Analysis

Since PAIRADISE is indeed a generic framework for testing differential splicing between matched sample groups (and not necessarily limited to ASAS), as a proof of concept we perform a tumor-specific splicing analysis using paired tumor/normal cell RNA-seq data from The Cancer Genome Atlas (TCGA). More specifically, we applied PAIRADISE to TCGA data for 12 cancer types, where the numbers of replicates in each dataset ranged from 11 (ESCA, Esophageal Carcinoma) to 107 (BRCA, Breast Invasive Carcinoma).

At FDR $\leq$ 10%, PAIRADISE identified $3,181$ differential splicing events across all tumor types (Table 2.2); moreover, over $1,300$ splicing events were significant in more than one tumor type, including 2 differential splicing events found to be significant for 10 different tumors (Table 2.3). These cases are highlighted in Figure 2.12, corresponding to two exon skipping events in the SMARCA4 gene and CALD1 gene respectively. In both events, the exon inclusion levels are consistently larger in tumor cells than in normal cells across nearly every tumor type. The SMARCA4 gene encodes a protein which is a member of the SWI/SNF family of proteins. Proteins from this group are believed to regulate transcription of certain genes through helicase and ATPase activities which modify the chromatin structure around those genes (Barutcu et al., 2016). The CALD1 gene encodes a calmodulin and actin binding protein and has been show to play an important role in smooth muscle and nonmuscle contraction. Variations in the CALD1 gene have been associated with diseases including Mixed Endometrial Stromal and Smooth

Tumor Type	**# Paired Samples**	**# Sig Exons**
BLCA	19	71
BRCA	107	1316
COAD	26	223
ESCA	11	10
HNSC	40	211
KICH	25	350
KIRC	65	1015
KIRP	32	202
LIHC	50	178
LUAD	53	553
LUSC	50	1091
THCA	59	520

Table 2.2: Number of differential alternative splicing events for each tumor type.

# Tumor Types	**# Sig Exons**
10	2
9	5
8	16
7	32
6	72
5	145
4	216
3	287
2	588
1	1818

Table 2.3: Number of common splicing events across multiple tumor types.

Muscle Tumor, Kidney Leiomyosarcoma, and glioma (Zheng et al., 2004). Though a proof of concept, these results further demonstrate how PAIRADISE can identify consistent splicing differences across samples and groups.

2.10 Discussion

To address the lack of robust statistical methods for ASAS analysis, we have introduced the PAIRADISE statistical framework for detecting ASAS from population-scale RNA-seq and genotype data. PAIRADISE provides a powerful tool for elucidating the genetic variation and phenotypic association of alternative splicing and frames the problem of ASAS detection as that of identifying differential alternative splicing from RNA-seq data with

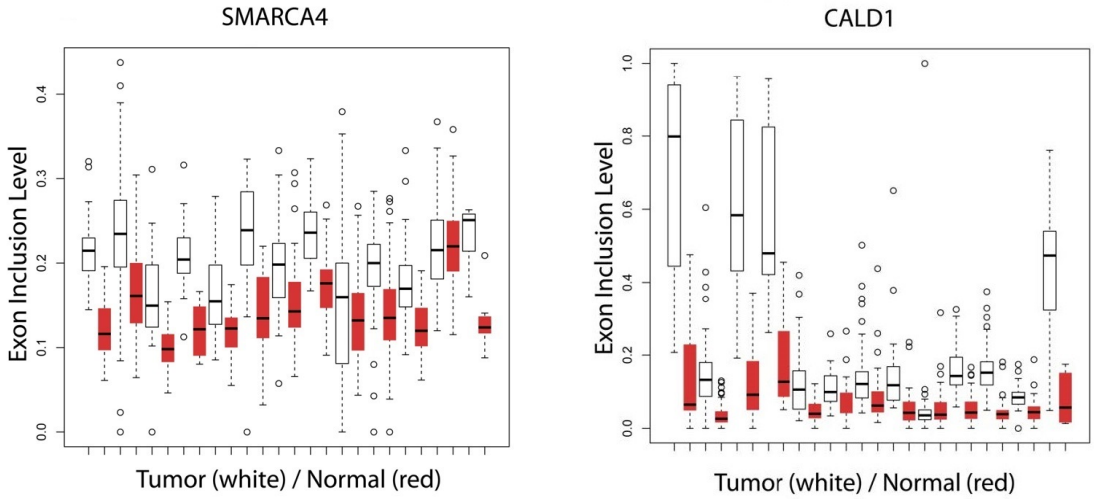

Figure 2.12: Two differential splicing events were significant for 10 different tumors. Left) Differential splicing event in the SMARCA4 gene. Exon inclusion levels are consistently higher in tumor cells relative to normal cells. Right) Differential splicing event in the CALD1 gene. Exon inclusion levels are consistently higher in tumor cells relative to normal cells. Each pair of boxplots corresponds to one tumor type.

paired replicates; leveraging the pairing structure of the two alleles allows PAIRADISE to identify consistent allelic differences in alternative splicing across multiple biological replicates or multiple individuals from a population. Our simulation studies demonstrate that PAIRADISE outperforms alternative statistical models for ASAS analysis; the gains in performance are especially large when the number of replicates is small, with other methods requiring up to 2-3 times as many samples to achieve the same performance as PAIRADISE (Figure 2.4A). We also demonstrate how PAIRADISE can increase the power of ASAS detection in a single individual by aggregating the ASAS signals from the six biological replicates of the GM12878 B-lymphocyte cell line. Compared to the analysis of each RNA-seq replicate in isolation, the PAIRADISE analysis of six replicates in combination generates two to four times as many significant ASAS events, at a comparable and often higher level of concordance with sQTL signals in the CEU population. We highlight

a particular advantage of PAIRADISE by demonstrating its ability to detect the effects of rare genetic variants on alternative splicing using both single-individual (GM12878) and population-scale (Geuvadis) RNA-seq datasets. Finally, since PAIRADISE is indeed a generic model for testing for differences in count-based ratios between matched pairs, we demonstrate the broad applicability of the PAIRADISE statistical model by using it to detect differential alternative splicing events on matched tumor-normal RNA-seq data from TCGA.

There are several limitations to the PAIRADISE statistical model. Alternative splicing levels may themselves depend on other factors such as other cis SNPs or the concentration or activity of trans-acting splicing regulators. This information could be integrated into PAIRADISE by adding an additional layer into the hierarchical framework that captures the relation between individual-specific allelic diffferences and certain covariates. Note that a larger population sample size would be needed to identify such covariates that affect the magnitude of ASAS signals across individuals. Another limitation of ASAS analysis with PAIRADISE is that it requires heterozygous SNPs to be outside of the alternative exon to enable allele-specific read assignment, but simultaneously requires SNPs to be close enough to the alternative splicing event for both to be detected on the same RNA-seq read. SNPs which are located within the alternative exon or too far from the alternative exon cannot be analyzed by PAIRADISE using short-read RNA-seq data. Finally, other than accounting for the total read coverage of different samples, PAIRADISE does not account for any differences in the quality of biological samples like those owing to the different protocols used by different labs. Li et al. (2018) demonstrated the importance of explicitly accounting for the heterogeneity in the quality of RNA-seq samples using a Bayesian hierarchical framework. We believe that this is a promising direction for future work.

2.11 Appendix

2.11.1 Derivation of the Likelihood Function

PAIRADISE aims to test whether there is a significant difference in the means of the distributions of logit(ψ_{i1k}) and logit(ψ_{i2k}). Adopting the notation of hypothesis testing, PAIRADISE performs the following test:

$$H_0 : \delta_i = 0$$

$$H_a : \delta_i \neq 0.$$

In the PAIRADISE framework, I_{i1k} and I_{i2k} are the observed data and ψ_{i1k}, ψ_{i2k}, and α_{ik} are all regarded as latent, unobserved variables. In order to make inference about the parameter δ_i, we must first derive an expression for the observed data likelihood. For notational simplicity, we first set $\theta_i = (\delta_i, \sigma_{i1}, \sigma_{i2}, \sigma_i, \mu_i)$. For a given exon i, the complete data likelihood function (the likelihood of the observed and latent variables) is given by

$$\prod_{k=1}^{M} f(I_{i1k}, I_{i2k}, \text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k}), \alpha_{ik} | \theta_i) =$$

$$\prod_{k=1}^{M} f(I_{i1k}|\psi_{i1k}) \cdot f(I_{i2k}|\psi_{i2k}) \cdot f(\text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k})|\alpha_{ik}, \delta_i, \sigma_{i1}, \sigma_{i2}) \cdot f(\alpha_{ik}|\mu_i, \sigma_i), \quad (2.9)$$

where

$$f(I_{i1k}|\psi_{i1k}) = c_1 \cdot \left[\frac{\ell_{iI}\psi_{i1k}}{\ell_{iI}\psi_{i1k} + \ell_{iS}(1-\psi_{i1k})}\right]^{I_{i1k}} \cdot \left[\frac{\ell_{iS}(1-\psi_{i1k})}{\ell_{iI}\psi_{i1k} + \ell_{iS}(1-\psi_{i1k})}\right]^{S_{i1k}},$$

$$f(I_{i2k}|\psi_{i2k}) = c_2 \cdot \left[\frac{\ell_{iI}\psi_{i2k}}{\ell_{iI}\psi_{i2k} + \ell_{iS}(1-\psi_{i2k})}\right]^{I_{i2k}} \cdot \left[\frac{\ell_{iS}(1-\psi_{i2k})}{\ell_{iI}\psi_{i2k} + \ell_{iS}(1-\psi_{i2k})}\right]^{S_{i2k}},$$

$$f(\text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k})|\alpha_{ik}, \delta_i, \sigma_{i1}, \sigma_{i2}) =$$

$$c_3 \cdot \frac{1}{\sigma_{i1}\sigma_{i2}} \exp\left(-\frac{(\text{logit}(\psi_{i1k}) - \alpha_{ik})^2}{2{\sigma_{i1}}^2}\right) \cdot \exp\left(-\frac{(\text{logit}(\psi_{i2k}) - \alpha_{ik} - \delta_i)^2}{2{\sigma_{i2}}^2}\right),$$

$$f(\alpha_{ik}|\mu_i, \sigma_i) = c_4 \cdot \frac{1}{\sigma_i} \exp\left(-\frac{1}{2{\sigma_i}^2}(\alpha_{ik} - \mu_i)^2\right),$$

and where the constants c_1, c_2, c_3 and c_4 do not depend on the model parameters or latent variables. Note that we are using the same function $f(\cdot)$ to represent pdfs/pmfs of different variables for notational clarity. Also note that the distributions of $\text{logit}(\psi_{i1k})$ and $\text{logit}(\psi_{i2k})$ are based on the conditional distribution given in (2.7) rather than the marginal distributions in (2.3). To derive an expression for the observed data likelihood, we can integrate the latent variables out of the complete data likelihood in expression (2.9) to obtain

$$\prod_{k=1}^{M} f(I_{i1k}, I_{i2k}|\theta_i)$$

$$= \prod_{k=1}^{M} \int f(I_{i1k}, I_{i2k}, \text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k}), \alpha_{ik}|\theta_i) d\,\text{logit}(\psi_{i1k}) \cdot d\,\text{logit}(\psi_{i2k}) \cdot d\alpha_{ik}. \quad (2.10)$$

Since there is no closed-form expression for the integral in (2.10), we proceed by using Laplace's method to obtain an approximation of this integral. Briefly, our application of Laplace's method uses a second-order Taylor expansion around the MLEs of $\alpha_{ik}, \text{logit}(\psi_{i1k})$, and $\text{logit}(\psi_{i2k})$ to approximate the observed data likelihood. Let $f_1 = \log(f)$, and let $\widehat{\alpha_{ik}}$, $\text{logit}(\widehat{\psi_{i1k}})$, and $\text{logit}(\widehat{\psi_{i2k}})$ be the MLEs of α_{ik}, $\text{logit}(\psi_{i1k})$, and $\text{logit}(\psi_{i2k})$. Then for $k = 1, \ldots, M$,

$$\int f(I_{i1k}, I_{i2k}, \text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k}), \alpha_{ik}|\theta_i) d\,\text{logit}(\psi_{i1k}) \cdot d\,\text{logit}(\psi_{i2k}) \cdot d\alpha_{ik}$$

$$= \int \exp(f_1(I_{i1k}, I_{i2k}, \text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k}), \alpha_{ik}|\theta_i)) d\,\text{logit}(\psi_{i1k}) \cdot d\,\text{logit}(\psi_{i2k}) \cdot d\alpha_{ik}$$

$$= \int \exp\{f_1(I_{i1k}, I_{i2k}, \text{logit}(\widehat{\psi_{i1k}}), \text{logit}(\widehat{\psi_{i2k}}), \widehat{\alpha_{ik}}|\theta_i)$$

$$+ \frac{1}{2} \begin{bmatrix} \text{logit}(\psi_{i1k}) - \text{logit}(\widehat{\psi_{i1k}}) \\ \text{logit}(\psi_{i2k}) - \text{logit}(\widehat{\psi_{i2k}}) \\ \alpha_{ik} - \widehat{\alpha_{ik}} \end{bmatrix}' \Sigma_{ik} \begin{bmatrix} \text{logit}(\psi_{i1k}) - \text{logit}(\widehat{\psi_{i1k}}) \\ \text{logit}(\psi_{i2k}) - \text{logit}(\widehat{\psi_{i2k}}) \\ \alpha_{ik} - \widehat{\alpha_{ik}} \end{bmatrix}$$

$$+ o((\text{logit}(\psi_{i1k}) - \text{logit}(\widehat{\psi_{i1k}}))^2) + o((\text{logit}(\psi_{i2k}) - \text{logit}(\widehat{\psi_{i2k}}))^2)$$

$$+ o((\alpha_{ik} - \widehat{\alpha_{ik}})^2)\} \quad d\,\text{logit}(\psi_{i1k}) \cdot d\,\text{logit}(\psi_{i2k}) \cdot d\alpha_{ik}$$

$$\approx (2\pi)^{3/2}(-|\Sigma_{ik}|)^{-1/2} \exp\{f_1(I_{i1k}, I_{i2k}, \text{logit}(\widehat{\psi_{i1k}}), \text{logit}(\widehat{\psi_{i2k}}), \widehat{\alpha_{ik}}|\theta_i)\}, \tag{2.11}$$

where we have used the fact that

$$\left[\left.\frac{\partial f_1(z_{ik})}{\partial\,\text{logit}(\psi_{i1k})}\right|_{\text{logit}(\widehat{\psi_{i1k}})}, \quad \left.\frac{\partial f_1(z_{ik})}{\partial\,\text{logit}(\psi_{i2k})}\right|_{\text{logit}(\widehat{\psi_{i2k}})}, \quad \left.\frac{\partial f_1(z_{ik})}{\partial \alpha_{ik}}\right|_{\widehat{\alpha_{ik}}} \right] = \mathbf{0}$$

for $z_{ik} := (\text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k}), \alpha_{ik})$. The Hessian matrix Σ_{ik} in (2.11) is given by

$$\Sigma_{ik} = \begin{bmatrix} \dfrac{\partial^2 f_1(z_{ik})}{\partial\,\text{logit}^2(\psi_{i1k})} & \dfrac{\partial^2 f_1(z_{ik})}{\partial\,\text{logit}(\psi_{i1k})\partial\,\text{logit}(\psi_{i2k})} & \dfrac{\partial^2 f_1(z_{ik})}{\partial\,\text{logit}(\psi_{i1k})\partial\alpha_{ik}} \\ \dfrac{\partial^2 f_1(z_{ik})}{\partial\,\text{logit}(\psi_{i1k})\partial\,\text{logit}(\psi_{i2k})} & \dfrac{\partial^2 f_1(z_{ik})}{\partial\,\text{logit}^2(\psi_{i2k})} & \dfrac{\partial^2 f_1(z_{ik})}{\partial\,\text{logit}(\psi_{i2k})\partial\alpha_{ik}} \\ \dfrac{\partial^2 f_1(z_{ik})}{\partial\alpha_{ik}\partial\,\text{logit}(\psi_{i1k})} & \dfrac{\partial^2 f_1(z_{ik})}{\partial\alpha_{ik}\partial\,\text{logit}(\psi_{i2k})} & \dfrac{\partial^2 f_1(z_{ik})}{\partial\alpha_{ik}^2} \end{bmatrix}, \tag{2.12}$$

where each of the second-order partial derivatives is evaluated at the MLEs $\widehat{\alpha_{ik}}$, $\text{logit}(\widehat{\psi_{i1k}})$, and $\text{logit}(\widehat{\psi_{i2k}})$. Note that the determinant of the above Hessian matrix is always negative (shown in subsection 2.11.4). Combining (2.10) and (2.11) yields an expression for the observed data likelihood:

$$\prod_{k=1}^{M} f(I_{i1k}, I_{i2k}|\theta_i) \approx c_5 \prod_{k=1}^{M} (-|\Sigma_{ik}|)^{-1/2} f(I_{i1k}, I_{i2k}, \text{logit}(\widehat{\psi_{i1k}}), \text{logit}(\widehat{\psi_{i2k}}), \widehat{\alpha_{ik}}|\theta_i)$$

or

$$\sum_{k=1}^{M} \log f(I_{i1k}, I_{i2k}|\theta_i) \approx$$

$$\sum_{k=1}^{M} \left\{ f_1(I_{i1k}, I_{i2k}, \text{logit}(\widehat{\psi_{i1k}}), \text{logit}(\widehat{\psi_{i2k}}), \widehat{\alpha_{ik}}|\theta_i) - \frac{1}{2}\log(-|\Sigma_{ik}|) \right\} + c_6 \tag{2.13}$$

for some constants c_5 and c_6.

2.11.2 Optimization

Next, we outline an iterative procedure that will produce estimates $(\hat{\delta}_i, \hat{\sigma_{i1}}, \hat{\sigma_{i2}}, \hat{\mu}_i, \hat{\sigma}_i)$ based on the observed data log likelihood in (2.13). For $k = 1, \ldots, M$, initialize $\text{logit}(\psi_{i1k})$ and $\text{logit}(\psi_{i2k})$ from the individual binomial distribution of each replicate:

$$\widehat{\text{logit}(\psi_{i1k}^{(0)})} = \text{logit}\left(\frac{I_{i1k}\ell_{iS}}{I_{i1k}\ell_{iS} + S_{i1k}\ell_{iI}}\right), \quad \widehat{\text{logit}(\psi_{i2k}^{(0)})} = \text{logit}\left(\frac{I_{i2k}\ell_{iS}}{I_{i2k}\ell_{iS} + S_{i2k}\ell_{iI}}\right).$$

Since $\text{logit}(\psi_{i1k}) = \alpha_{ik} + \epsilon_{i1k}$, one can set $\widehat{\alpha_{ik}}^{(0)} = \widehat{\text{logit}(\psi_{i1k}^{(0)})}$. Next, let $t \leftarrow 1$ and proceed through the following steps:

Step 1: Estimate the MLEs of the observed data likelihood based on the estimated values of $\widehat{\text{logit}(\psi_{i1k}^{(t-1)})}, \widehat{\text{logit}(\psi_{i2k}^{(t-1)})}$, and $\widehat{\alpha_{ik}}^{(t-1)}$. That is, maximize expression (2.13):

$$(\hat{\delta}_i^{(t)}, \hat{\sigma_{i1}}^{(t)}, \hat{\sigma_{i2}}^{(t)}, \hat{\mu}_i^{(t)}, \hat{\sigma}_i^{(t)}) =$$

$$\underset{\delta_i, \sigma_{i1}, \sigma_{i2}, \mu_i, \sigma_i}{\text{argmax}} \sum_{k=1}^{M} \left(f_1(I_{i1k}, I_{i2k}, \widehat{\text{logit}(\psi_{i1k}^{(t-1)})}, \widehat{\text{logit}(\psi_{i2k}^{(t-1)})}, \widehat{\alpha_{ik}}^{(t-1)} | \theta_i) - \frac{1}{2}\log(-|\Sigma_{ik}^{(t-1)}|) \right).$$

$\Sigma_{ik}^{(t-1)}$ is the Hessian matrix given in (2.12) where the partial derivatives are evaluated using the estimates $\widehat{\alpha_{ik}}^{(t-1)}, \widehat{\text{logit}(\psi_{i1k}^{(t-1)})}$, and $\widehat{\text{logit}(\psi_{i2k}^{(t-1)})}$ (a formal expression for $|\Sigma_{ik}^{(t)}|$ is given in the next subsection).

Step 2: For $k = 1, \ldots, M$, update the estimates $\widehat{\alpha_{ik}}^{(t)}, \widehat{\text{logit}(\psi_{i1k}^{(t)})}$, and $\widehat{\text{logit}(\psi_{i2k}^{(t)})}$ based on the complete data likelihood (2.9) and the latest MLEs of $\hat{\delta}_i^{(t)}, \hat{\sigma_{i1}}^{(t)}, \hat{\sigma_{i2}}^{(t)}, \hat{\mu_i}^{(t)}, \hat{\sigma_i}^{(t)}$:

$$(\widehat{\text{logit}(\psi_{i1k}^{(t)})}, \widehat{\text{logit}(\psi_{i2k}^{(t)})}, \widehat{\alpha_{ik}}^{(t)}) =$$

$$\underset{\text{logit}(\psi_{i1k}),\text{logit}(\psi_{i2k}),\alpha_{ik}}{\text{argmax}} \Bigg(A(\psi_{i1k}) + B(\psi_{i2k}) + C(\alpha_{ik}) + D(\psi_{i1k}, \psi_{i2k}, \alpha_{ik}) \Bigg),$$

where

$$A(\psi_{i1k}) = I_{i1k} \log \left(\frac{\ell_{iI}\psi_{i1k}}{\ell_{iI}\psi_{i1k} + \ell_{iS}(1-\psi_{i1k})} \right) + S_{i1k} \log \left(\frac{\ell_{iS}(1-\psi_{i1k})}{\ell_{iI}\psi_{i1k} + \ell_{iS}(1-\psi_{i1k})} \right),$$

$$B(\psi_{i2k}) = I_{i2k} \log \left(\frac{\ell_{iI}\psi_{i2k}}{\ell_{iI}\psi_{i2k} + \ell_{iS}(1-\psi_{i2k})} \right) + S_{i2k} \log \left(\frac{\ell_{iS}(1-\psi_{i2k})}{\ell_{iI}\psi_{i2k} + \ell_{iS}(1-\psi_{i2k})} \right),$$

$$C(\alpha_{ik}) = -\frac{1}{2\hat{\sigma_i}^{2(t)}}(\alpha_{ik} - \hat{\mu_i}^{(t)})^2,$$

$$D(\psi_{i1k}, \psi_{i2k}, \alpha_{ik}) = -\frac{(\text{logit}(\psi_{i1k}) - \alpha_{ik})^2}{2\hat{\sigma_{i1}}^{2(t)}} - \frac{(\text{logit}(\psi_{i2k}) - \alpha_{ik} - \hat{\delta_i}^{(t)})^2}{2\hat{\sigma_{i2}}^{2(t)}}.$$

Step 3: Let $t \leftarrow t+1$ and go to step 1. Iterate between steps 1 and 2 until the difference in log likelihoods between consecutive iterations is smaller than some threshold ϵ, say $\epsilon = 10^{-2}$. Use an optimization algorithm (e.g. L-BFGS-B or BOBYQA) to optimize the likelihood function with the parameters $\sigma_{i1}, \sigma_{i2}, \sigma_i$ constrained within $(0, \infty)$, and $\alpha_{ik}, \mu_i, \delta_i, \text{logit}(\psi_{i1k}), \text{logit}(\psi_{i2k})$ unconstrained.

The above optimization procedure is performed for two cases: the unconstrained model, and the model constrained under the null hypothesis (i.e. the model with $\delta_i = 0$). The likelihood-ratio test statistic then asymptotically (in M) follows a χ^2 distribution with 1 degree of freedom:

$$-2(\log L_{\delta_i=0} - \log L) \sim \chi^2_1,$$

where $L_{\delta_i=0}$ is the likelihood function under the null hypothesis and L is the likelihood function under the alternative hypothesis.

2.11.3 Computing the Hessian Σ_{ik}

The expressions for the partial derivatives in the Hessian matrix Σ_{ik} given in (2.12), evaluated at $\widehat{\alpha_{ik}}, \text{logit}(\widehat{\psi_{i1k}}), \text{logit}(\widehat{\psi_{i2k}})$, are given by

$$\frac{\partial^2 f_1(z_{ik})}{\partial \,\text{logit}^2\,(\psi_{i1k})} = \frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i1k}(\hat{\psi}_{i1k}-1)(I_{i1k}+S_{i1k})}{[\ell_{iI}\hat{\psi}_{i1k}+\ell_{iS}(1-\hat{\psi}_{i1k})]^2} - \frac{1}{\sigma_{i1}^2} \tag{2.14}$$

$$\frac{\partial^2 f_1(z_{ik})}{\partial \,\text{logit}^2\,(\psi_{i2k})} = \frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i2k}(\hat{\psi}_{i2k}-1)(I_{i2k}+S_{i2k})}{[\ell_{iI}\hat{\psi}_{i2k}+\ell_{iS}(1-\hat{\psi}_{i2k})]^2} - \frac{1}{\sigma_{i2}^2} \tag{2.15}$$

$$\frac{\partial^2 f_1(z_{ik})}{\partial \alpha_{ik}^2} = -\left[\frac{1}{\sigma_{i1}^2}+\frac{1}{\sigma_{i2}^2}+\frac{1}{\sigma_i^2}\right] \tag{2.16}$$

$$\frac{\partial^2 f_1(z_{ik})}{\partial \alpha_{ik}\partial \,\text{logit}(\psi_{i1k})} = \frac{\partial^2 f_1(z_{ik})}{\partial \,\text{logit}(\psi_{i1k})\partial \alpha_{ik}} = \frac{1}{\sigma_{i1}^2}$$

$$\frac{\partial^2 f_1(z_{ik})}{\partial \alpha_{ik}\partial \,\text{logit}(\psi_{i2k})} = \frac{\partial^2 f_1(z_{ik})}{\partial \,\text{logit}(\psi_{i2k})\partial \alpha_{ik}} = \frac{1}{\sigma_{i2}^2}$$

$$\frac{\partial^2 f_1(z_{ik})}{\partial \,\text{logit}\,(\psi_{i1k})\partial \,\text{logit}\,(\psi_{i2k})} = \frac{\partial^2 f_1(z_{ik})}{\partial \,\text{logit}\,(\psi_{i2k})\partial \,\text{logit}\,(\psi_{i1k})} = 0.$$

The determinant $|\Sigma_{ik}^{(t)}|$ is therefore given by the expression

$$\left|\Sigma_{ik}^{(t)}\right| = D(\sigma_{i1},\sigma_{i2}) + E(\sigma_{i1},\sigma_{i2})\cdot F(\sigma_{i1},\sigma_{i2},\sigma_i),$$

where

$$D(\sigma_{i1},\sigma_{i2}) = \frac{1}{(\sigma_{i1}^2)^2}\left(\frac{1}{\sigma_{i2}^2} - \frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i2k}^{(t)}(\hat{\psi}_{i2k}^{(t)}-1)(I_{i2k}+S_{i2k})}{[\ell_{iI}\hat{\psi}_{i2k}^{(t)}+\ell_{iS}(1-\hat{\psi}_{i2k}^{(t)})]^2}\right),$$

$$E(\sigma_{i1},\sigma_{i2}) = \left(\frac{1}{\sigma_{i1}^2} - \frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i1k}^{(t)}(\hat{\psi}_{i1k}^{(t)}-1)(I_{i1k}+S_{i1k})}{[\ell_{iI}\hat{\psi}_{i1k}^{(t)}+\ell_{iS}(1-\hat{\psi}_{i1k}^{(t)})]^2}\right),$$

$$F(\sigma_{i1}, \sigma_{i2}, \sigma_i) = \frac{1}{(\sigma_{i2}^2)^2} + \left(\frac{1}{\sigma_{i1}^2} + \frac{1}{\sigma_{i2}^2} + \frac{1}{\sigma_i^2}\right)\left(\frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i2k}^{(t)}(\hat{\psi}_{i2k}^{(t)} - 1)(I_{i2k} + S_{i2k})}{[\ell_{iI}\hat{\psi}_{i2k}^{(t)} + \ell_{iS}(1 - \hat{\psi}_{i2k}^{(t)})]^2} - \frac{1}{\sigma_{i2}^2}\right).$$

2.11.4 Proof that the Determinant of Σ_{ik} is Negative

To ease notation, rewrite the Hessian in (2.12) as

$$\Sigma_{ik} = \begin{bmatrix} x_1 & 0 & \frac{1}{\sigma_{i1}^2} \\ 0 & x_2 & \frac{1}{\sigma_{i2}^2} \\ \frac{1}{\sigma_{i1}^2} & \frac{1}{\sigma_{i2}^2} & x_3 \end{bmatrix}$$

where x_1, x_2 and x_3 are defined as in (2.14), (2.15) and (2.16) (we ignore the indices i and k for additional clarity). Next, let

$$a_1 = \frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i1k}(\hat{\psi}_{i1k} - 1)(I_{i1k} + S_{i1k})}{[\ell_{iI}\hat{\psi}_{i1k} + \ell_{iS}(1 - \hat{\psi}_{i1k})]^2}$$

and

$$a_2 = \frac{\ell_{iI}\ell_{iS}\hat{\psi}_{i2k}(\hat{\psi}_{i2k} - 1)(I_{i2k} + S_{i2k})}{[\ell_{iI}\hat{\psi}_{i2k} + \ell_{iS}(1 - \hat{\psi}_{i2k})]^2}$$

so that

$$x_1 = a_1 - \frac{1}{\sigma_{i1}^2}$$

and

$$x_2 = a_2 - \frac{1}{\sigma_{i2}^2}.$$

It follows that

$$\det(\Sigma_{ik}) = (x_3 - 1)x_1x_2 + \det \begin{bmatrix} x_1 - \frac{1}{\sigma_{i1}^4} & -\frac{1}{\sigma_{i1}^2\sigma_{i2}^2} \\ -\frac{1}{\sigma_{i1}^2\sigma_{i2}^2} & x_2 - \frac{1}{\sigma_{i2}^4} \end{bmatrix}$$

$$= x_1x_2x_3 - \left[\frac{x_1}{\sigma_{i2}^4} + \frac{x_2}{\sigma_{i1}^4}\right]$$

$$= -\left[\frac{1}{\sigma_i^2} + \frac{1}{\sigma_{i1}^2} + \frac{1}{\sigma_{i2}^2}\right]\left[(a_1 - \frac{1}{\sigma_{i1}^2})(a_2 - \frac{1}{\sigma_{i2}^2})\right] - \left[\frac{a_1}{\sigma_{i2}^4} - \frac{1}{\sigma_{i1}^2\sigma_{i2}^4} + \frac{a_2}{\sigma_{i1}^4} - \frac{1}{\sigma_{i2}^2\sigma_{i1}^4}\right]$$

$$= -\frac{a_1a_2}{\sigma_{i1}^2} + \frac{a_1}{\sigma_{i1}^2\sigma_{i2}^2} + \frac{a_2}{\sigma_{i1}^4} - \frac{1}{\sigma_{i1}^4\sigma_{i2}^2} - \frac{a_1a_2}{\sigma_{i2}^2} + \frac{a_1}{\sigma_{i2}^4} + \frac{a_2}{\sigma_{i1}^2\sigma_{i2}^2} - \frac{1}{\sigma_{i1}^2\sigma_{i2}^4}$$

$$- \frac{a_1a_2}{\sigma_i^2} + \frac{a_1}{\sigma_i^2\sigma_{i2}^2} + \frac{a_2}{\sigma_i^2\sigma_{i1}^2} - \frac{1}{\sigma_i^2\sigma_{i1}^2\sigma_{i2}^2} - \left[\frac{a_1}{\sigma_{i2}^4} + \frac{a_2}{\sigma_{i1}^4} - \frac{1}{\sigma_{i1}^2\sigma_{i2}^4} - \frac{1}{\sigma_{i2}^2\sigma_{i1}^4}\right]$$

$$= -a_1a_2\Big(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_{i1}^2} + \frac{1}{\sigma_{i2}^2}\Big) + a_1\Big(\frac{1}{\sigma_{i1}^2\sigma_{i2}^2} + \frac{1}{\sigma_i^2\sigma_{i2}^2}\Big)$$

$$+ a_2\Big(\frac{1}{\sigma_{i1}^2\sigma_{i2}^2} + \frac{1}{\sigma_i^2\sigma_{i1}^2}\Big) - \frac{1}{\sigma_i^2\sigma_{i1}^2\sigma_{i2}^2} < 0,$$

which follows since $a_1, a_2 < 0$.

2.11.5 Description of Data

GM12878

The RNA-seq data from six replicates of the GM12878 B-lymphocyte cell lines were generated by the following labs (sample IDs from the ENCODE project are given in parentheses): Brenton Graveley's lab at UConn (sample ENCSR000AEF, 2 replicates), Barbara Wold's lab at Caltech (sample ENCSR000AEG, 2 replicates), Thomas Gingeras' lab at CSHL (sample ENCSR000AED, 2 replicates).

Guevadis

We also used RNA-seq and genotype data from the Geuvadis dataset of B-lymphocyte cell lines of 445 individuals from five populations (Lappalainen and Sammeth, 2013). Genotype data for these individuals were from the Phase 3 of 1000 Genomes Project (release 05-02-2013) (Auton and Coauthors, 2015).

2.11.6 Allele-Specific Alignment of RNA-seq Data

The PAIRADISE computational pipeline takes two inputs: FASTQ files of RNA-seq data, and VCF files of phased genotype data. In addition, PAIRADISE also uses a human reference genome, a GTF file of gene/transcript annotations and a list of RNA editing sites that are masked for allele-specific read assignment. More details about the PAIRADISE running parameters, as well as download links and annotation files are provided at our lab's website (https://github.com/Xinglab/PAIRADISE). The PAIRADISE statistical model is available as a stand-alone R package and forms the final stage of our computational pipeline.

Our pipeline first performs allele-specific read mapping onto alternative splicing events using rPGA (version 2.0.0, https://github.com/Xinglab/rPGA). First, the reference genome is personalized based on the phased genotype data of each individual. For each individual, the reference genome is modified at each SNP position to carry the alleles of that particular individual. This process yields one personal reference genome for each haplotype. Second, RNA-seq reads are aligned to both personal genomes using STAR (Dobin et al., 2013) (version 2.5.3a, https://github.com/alexdobin/STAR/archive/2.5.3a.tar.gz), allowing 6 mismatches and restricting splice junctions to canonical splice sites only. The third step involves allele-specific read assignment: for each uniquely mapped read, we first identify all heterozygous SNPs covered by the read, and note whether the read carries the first or second haplotype allele at each base. Reads carrying haplotype 1 (or 2) alleles at the majority of the heterozygous SNP positions are assigned to haplotype 1 (or 2). Reads which fail to meet either of these requirements are removed.

2.11.7 Allele-Specific Read Assignment

To detect alternative splicing events, the allele-specific bam files mapped onto the two haplotypes were merged together and input to rMATS (version 3.2.5) (Shen et al., 2014).

To ensure a consistent set of alternative splicing events across all samples, the merged allele-specific bam files of all samples were used together in the rMATS analysis.

In an allele-specific alternative splicing analysis using RNA-seq, reads must be assigned to one of the two alleles of a heterozygous SNP. For a given exon skipping event and heterozygous SNP at one of the flanking constitutive exons, a read must cover both the SNP and a splice junction (either exon skipping or exon inclusion) to be assigned to one of the two isoforms (see Figure 2.3A). In the situation where a read covers multiple heterozygous SNPs at either of the flanking exons, we counted that read separately for each SNP. Note that we excluded any SNPs within the alternatively spliced exon from any further analysis since they can only be detected from the exon inclusion isoform. We also filtered out alternative splicing events for which the average ψ values across all individuals were less than 5% or greater than 95% for both alleles, as well as events with less than 10 total reads on average across all individuals for both alleles.

2.11.8 sQTL Analysis of the Five Geuvadis Populations

To perform an sQTL analysis on the five Geuvadis populations, we first processed the RNA-seq data from these populations using rMATS to generate a consistent set of alternative splicing events across all of the populations. Next, we applied the following filtering criteria across the alternative splicing events in each population separately:

1. the average ψ values across all individuals was between 5% and 95%
2. the average total read count of all individuals was greater than or equal to 10
3. the range of ψ values across all individuals was greater than 20%
4. more than 20% of the individuals had non-zero read counts.

The GLiMMPS statistical model (Zhao et al., 2013) was then used to discover sQTLs by

testing for association between genotype and exon inclusion levels with SNPs within 200kb upstream or downstream of alternative exons. As an additional filtering criterion, SNPs with minor allele frequency (MAF) less than 5% were removed from the analysis (MAFs were estimated for each population separately). Finally, the GLiMMPS sQTL p-values for each alternative splicing event were defined to be the p-values of the SNP with the most significant association within the 200kb window. The sQTL FDRs were estimated based on 10 permutations (Zhao et al., 2013) and linkage disequilibrium correlations between SNPs were calculated by the 1000 Genomes Project (HapMap release #27). GWAS traits and associated SNPs were collected from the NHGRI-EBI GWAS catalog (version 01-14-16) (MacArthur et al., 2017).

CHAPTER 3

Quantifying Alternative Splicing Variation in Multi-Isoform, Complex Splicing Modules

Thus far, we have exclusively been dealing with basic patterns of alternative splicing such as studying differential isoform expression from simple exon skipping events. As discussed in Chapter 1, patterns of alternative splicing are often more complex than those highlighted in Figure 1.1, and thus it is imperative to develop computational models that can accommodate these commonly encountered patterns of isoform variation. To make matters even more complicated, RNA-seq reads are often ambiguous in the sense that there is no aspect of the read which uniquely identifies it as having been generated from one specific isoform. For example, reads fully contained within a constitutive exon of a simple exon skipping event could have been generated either by the exon inclusion or exon skipping mRNA isoforms.

In this chapter, we propose rMATS-Iso, a generalization of the rMATS statistical framework, and the first event-based tool which can detect differential alternative splicing in splicing modules with complex splicing patterns using replicate RNA-seq data. The rMATS-Iso statistical model utilizes a hierarchical framework to account for both the estimation uncertainty in ψ values within individual replicates as due to RNA-seq read coverage, as well as the variation in ψ values among replicates. rMATS-Iso leverages an EM algorithm to disambiguate short RNA-seq reads which may be consistent with multiple mRNA isoforms. As a result, rMATS-Iso can accommodate complex patterns

of alternative splicing within a splicing module where transcripts can be defined by any combination of exons, splice site choices, etc. In addition to quantifying isoform composition within individual sample groups, rMATS-Iso utilizes a likelihood ratio test to identify differential splicing between two sample groups. Once differential splicing has been identified, rMATS-Iso further quantifies the extent to which each individual transcript contributes to the overall difference between groups.

We begin the chapter by reviewing the problem of transcript reassembly from partial sequencing observations, tracing the evolution of the problem over the last few decades. Next, we introduce the rMATS-Iso statistical model, and demonstrate its performance using two simulation studies. Finally, we apply rMATS-Iso to identify differential alternative splicing using RNA-seq data generated from the PC3E (epithelial) and GS689 (mesenchymal) cell lines. All technical derivations appear in the appendix at the end of the chapter.

3.1 Introduction

The reconstruction of full-length isoforms from sequence fragments poses a unique challenge owing to the combinatorial nature of alternative splicing. RNA-seq reads do not capture the full-length mRNA sequence, but rather shorter fragments of the target sequence (Wang et al., 2009). As a result, it can be difficult to infer the underlying isoform structure that generated the RNA-seq read in the case where more than one isoform is compatible with the generated sequence (Figure 3.1). Furthermore, it is difficult to assess the functional impact of a splice variant without knowing its corresponding full-length transcript (Boue et al., 2002).

Before the advent of RNA-seq, Xing et al. (2004) outlined several of the contemporary challenges like those mentioned above that made analyzing the transcriptome more

difficult. Though Xing et al. (2004) focused on ESTs in their analysis, the challenges they highlighted were by no means limited to ESTs; in fact, the problems they discussed would soon become very relevant for alternative sequencing technologies like RNA-seq. These challenges included: 1) The existence of many different biological processes generating multiple mRNA isoforms. These processes include alternative splicing, alternative polyadenylation, and RNA editing. 2) Genetic polymorphisms which themselves may interact with the processes mentioned above. 3) The fact that observed sequencing reads (in our case RNA-seq reads) are short snapshots of the true data-generating process (full-length transcript). 4) The existence/analysis of complex splicing modules. 5) Random experimental errors. 6) The question of how to best integrate information from multiple different sources (e.g. ESTs, full-length mRNAs, genomic sequences) to produce the most biological meaningful results.

Xing et al. (2004) referred to these problems collectively as the "multiassembly problem" and proposed a two-tiered methodolgy to address it. The first stage involved using short-length EST fragments to reconstruct the full-length isoform. To represent all the different ways which the exons in the target region could be spliced together, the authors employed the framework of splice graphs introduced by Heber et al. (2002). Briefly, splice graphs utilize a graph representation in order to represent a gene's structure and patterns of alternative splicing. More specifically, splice graphs use nodes to represent the exons within a gene or splicing module, and directed edges between nodes to represent different splicing events. To traverse the graph and infer the most likely set of isoforms which could have generated the observed data (i.e. sequencing reads), Xing et al. (2004) then employed a dynamic programming algorithm known as heaviest bundling developed in Lee (2003) and Lee et al. (2002). The second stage of the multiassembly problem involved implementing a heuristic approach to evaluate the transcript sequences for completeness and what the authors referred to as distinguishing "productive isoforms" from EST artifacts.

Despite the successes of splice-graph based approaches for transcript reassembly, a number of problems facing isoform reconstruction still remained. First of all, splice graph transversal algorithms like heaviest bundling were not necessarily well suited for dealing with complex patterns of alternative splicing. For example, heaviest bundling could not guarantee an optimal traversal in the presence of coupled edges in the splice graph (which could represent splicing events such as mutually exclusive exons). In addition, an important and unanswered question remained as to how to account for sequence fragments which were consistent with a number of distinct isoforms. Realizing the need for an explicitly probabilistic approach, Xing et al. (2006) developed a statistical framework and EM algorithm for estimating the probability of each transversal across the splice graph. More specifically, the authors developed a hierarchical statistical model assumed to be generating EST reads, thus explicitly capturing the sources of uncertainty when categorizing sequence fragments. This algorithm has served as the basis for isoform level estimation by popular RNA-seq tools such as Cufflinks (Trapnell et al., 2013) and RSEM (Li and Dewey, 2011). Note that though the model was developed with EST reads in mind, it can easily be extended to handle RNA-seq data.

The statistical model of Xing et al. (2006) consists of two layers: the observed data, and a layer of latent, unobserved variables assumed to be generating the observed data. Suppose for a given gene or splicing module that there are a total of n different isoforms, and let $\psi = (\psi_1, \ldots, \psi_n)$ denote the probabilities that the reads (e.g. RNA-seq reads) will be generated by each isoform; here $\sum_{f=1}^{n} \psi_f = 1$, and $0 \leq \psi_f \leq 1$ for $f = 1, \ldots, n$. In this framework, ψ is assumed to be a fixed parameter vector that is to be estimated from the data. The observed data consist of K sequence observations $O_1, \ldots, O_K$. The K sequence observations can be encoded into a $K \times n$ indicator matrix Z, where $Z_{if} = 1$ if the i^{th} sequence was generated from isoform f and $Z_{if} = 0$ otherwise. In other words, the i^{th} row of Z indicates which isoform generated O_i. Of course, Z is unobserved since

each sequence can be consistent with multiple isoforms, so another matrix Y is introduced into the model. Instead of denoting which isoform generated each sequence, the entries of Y denote which isoforms are *consistent* with each sequence (i.e. which isoforms *could have* generated each sequence). More formally, $Y_{if} = 1$ if O_i is consistent with isoform f and $Y_{if} = 0$ otherwise. The indicator matrix Y contains the observed data. Under this framework, the observed data likelihood function is given by

$$L(\psi|Y) = \sum_{i=1}^{K} \log \left(\sum_{f=1}^{n} y_{i,f} \psi_f \right) \tag{3.1}$$

and an EM algorithm is used to find the maximum likelihood estimates of ψ. The model also includes a likelihood ratio test designed to test for differences in ψ values between two sample groups (e.g. two distinct tissues).

Despite the successes of this method, it faces a number of limitations. First, the model is specifically designed to handle reads from one sample (e.g. biological replicate) at a time; therefore, there is no way to use EST/RNA-seq reads being generated from multiple samples. As a result, even though the model accounts for differences in ψ values between distinct tissues/biological conditions, it ignores the variability in ψ values between different samples from the same tissue/biological condition. Another problem is that the structure of the splicing module is not explicitly incorporated into the statistical model, nor is there any length-normalization of the probabilities ψ_f to adjust for the effective lengths of each of the different isoforms. Therefore, estimates of ψ values for larger isoforms are biased upwards relative to smaller isoforms. Finally, the statistical model implicitly assumes that knowing that a read is consistent with an isoform gives no additional information about whether or not the same read is consistent with another isoform. In other words, the patterns of isoform compatibility are assumed to be independent of one another. In reality, the relation between these "consistency patterns" is more complex; for example, in the case of a simple exon-skipping event, a read covering

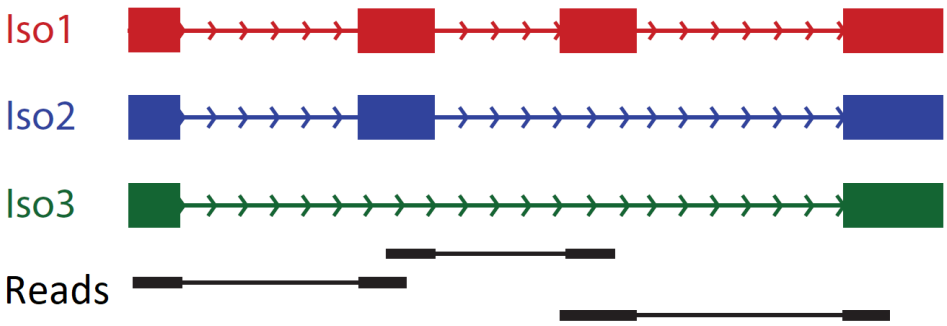

Figure 3.1: rMATS-Iso is a multi-isoform generalization of the rMATS statistical and computational framework. rMATS-Iso can accommodate alternative splicing modules with more than two isoforms, as well as RNA-seq reads which are consistent with more than one isoform. Here, reads 1 and 3 were generated from isoform 1, while read 2 could have been generated by isoforms 1 or 2.

part of the alternative exon is necessarily inconsistent with the exon skipping isoform. A more realistic statistical framework should capture such structure in the underlying splicing module / sequence fragment relation.

3.2 The rMATS-Iso Statistical Model

To address the shortcomings of the model in Xing et al. (2006), we developed a new statistical framework called rMATS-Iso. rMATS-Iso is a multi-isoform generalization of the rMATS statistical framework and provides a more realistic platform for estimating isoform abundance. rMATS-Iso quantifies the uncertainty in isoform inclusion levels due to RNA-seq read coverage as well as variability between biological samples. More specifically, rMATS-Iso utilizes a hierarchical design to model the variability in isoform inclusion levels among individual samples from the same biological condition. Moreover, rMATS-Iso models the variability in RNA-seq read counts within each individual while simultaneously accounting for the compatibility of RNA-seq reads with multiple isoforms. The principal

novelty of rMATS-Iso is that it defines the observed data to be RNA-seq read counts of observed *isoform consistency patterns* instead of read counts uniquely corresponding to each isoform. Figure 3.1 shows a fictional example of a splicing module with three distinct isoforms and how RNA-seq reads can be consistent with several different isoforms.

3.2.1 Modeling the Between-Sample Variability

To model the variability in isoform inclusion levels among samples, let ψ_{kf} denote the isoform inclusion level of the f^{th} transcript in sample k, and let $\psi_k = (\psi_{k1}, \ldots, \psi_{kn})$ denote the vector of isoform inclusion levels. Here, $0 \leq \psi_{kf} \leq 1$ for all $f = 1, \ldots, n$, and $\sum_{f=1}^{n} \psi_{kf} = 1$. rMATS-Iso assumes that the vectors for each sample are independent, identically distributed draws from a Dirichlet distribution:

$$\psi_k = (\psi_{k1}, \ldots, \psi_{kn}) \overset{iid}{\sim} \text{Dirichlet}(\alpha), \quad k = 1, \ldots, K, \tag{3.2}$$

where K is the total number of samples. Here, by samples we are referring to multiple different instances drawn from the same population (e.g. multiple biological replicates taken from the same tissue in a given individual, or multiple draws from the same tissue type taken across individuals within a population). The ψ_k values in equation 3.2 are regarded as latent, unobserved variables, and $\alpha = (\alpha_1, \ldots, \alpha_n)$ is an $n-$dimensional parameter vector of positive real numbers.

The Dirichlet parameter α in equation 3.2 fully specifies the joint probabilistic behavior of all isoforms' inclusion levels, e.g. their means, variances, and higher-order moments. Moreover, since the marginal distributions of a Dirichlet random vector follow a beta distribution, the parameter α fully specifies the marginal distributions of each individual ψ_f. Estimates of α will therefore capture both the joint variability in ψ, as well as the marginal variability in each ψ_f.

3.2.2 Modeling the Within-Sample Variability

If each RNA-seq read uniquely identified one transcript, then a straightforward way to model the variability in RNA-seq reads would be to place a multinomial distribution on the read counts corresponding to each transcript. Conjugacy of the Dirichlet and Multinomial distributions would imply a Dirichlet likelihood, and inference could easily be carried out by using the expectation-maximization (EM) algorithm (Dempster et al., 1977) to estimate the parameter in equation 3.2. In practice, however, the situation is more complicated since short RNA-seq reads may be consistent with multiple different transcripts (Figure 3.1). As mentioned in the previous section, Xing et al. (2006) addressed a similar problem in the analysis of EST reads by developing a probabilistic model of the isoform reconstruction problem based on the EM algorithm. The key idea there was to classify reads as being consistent with each isoform, and then to regard the resulting "consistency matrices" as the observed data.

Inspired by the work of Xing et al. (2006), we expand upon this reformulation of the observed data by modeling the reads that are consistent with each *combination* of transcripts rather than the reads that are consistent with each *individual* transcript. To formalize this idea, first note that in general, when there are n distinct transcripts, there can be a total of $M = 2^n - 1$ possible combinations of transcripts, since

$$\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1$$

by the binomial theorem. We refer to each of these M combinations as a "consistency pattern" (or simply, a pattern). Each of the M patterns corresponds to a set of isoforms with which an RNA-seq read may be consistent. These patterns can be encoded into a binary pattern matrix P with M rows and n columns, where $P_{mf} = 1$ if isoform f is included in the m^{th} consistency pattern, and $P_{mf} = 0$ otherwise for $m = 1, \ldots, M$, and $f = 1, \ldots, n$. For example, in the case where there are a total of $n = 3$ distinct transcripts,

the pattern matrix P is given by:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{3.3}$$

Here, the first consistency pattern (i.e. the first row of P) is $(1, 0, 0)$, indicating that only transcript 1 is included in the first pattern - reads corresponding to this pattern would be reads which are only consistent with the first transcript. The fifth consistency pattern is $(1, 0, 1)$, indicating that only transcripts 1 and 3 are included in the fifth pattern - reads corresponding to this pattern would be reads which are consistent with transcripts 1 and 3, but not transcript 2.

Instead of modeling the read counts corresponding to each individual transcript, rMATS-Iso models the read counts corresponding to each consistency pattern in P. Since each individual sample can have RNA-seq reads of differing lengths, we define $\mathcal{R}_k = \{r_1, \ldots, r_{L_k}\}$ to be the set of unique read lengths for the reads corresponding to subject k. Here, L_k is the total number of distinct RNA-seq read lengths for sample k.

To define the observed data, let $Y_{kr} = (Y_{kr1}, \ldots, Y_{krM}) \in \mathbb{N}_0^M$ be the vector of counts of reads with length r corresponding to each of the M patterns; here $\mathbb{N}_0$ denotes the set of non-negative natural numbers. Finally, define

$$Y_k = \begin{bmatrix} — & Y_{kr_1} & — \\ — & Y_{kr_2} & — \\ & \vdots & \\ — & Y_{kr_{L_k}} & — \end{bmatrix}. \tag{3.4}$$

The r^{th} row, m^{th} column of Y_k denotes the number of reads of length r corresponding to

pattern m in sample k. rMATS-Iso utilizes the following multinomial distribution for the read counts given in (3.4):

$$Y_k|\psi_k \sim \prod_{r\in\mathcal{R}_k} Multinomial(Y_{kr}; R_{kr}, p_r(\psi_k)), \quad k = 1, \ldots, K. \tag{3.5}$$

In expression (3.5), $R_{kr} = \sum_{m=1}^{M} Y_{krm}$ and $p_r(\psi_k) = (p_{r1}(\psi_k), \ldots, p_{rM}(\psi_k))$. Here, $p_{rm}(\psi_k)$ denotes the probability that a read with length r will be consistent with the m^{th} pattern and is equal to

$$p_{rm}(\psi_k) = \sum_{f=1}^{n} \theta_{mf}^{(r)} \cdot \tilde{\psi}_{kf}, \tag{3.6}$$

where $\tilde{\psi}_{kf}$ is the length-normalized version of ψ_{kf} (explained in more detail in the appendix, section 3.8.1). $\theta_{mf}^{(r)}$ denotes the probability that a read with length r will correspond to pattern m, given the read was actually generated by isoform f.

3.2.3 Transforming the Isoform Probabilities into Consistency Probabilities

The probabilities $\theta_{mf}^{(r)}$ given in expression (3.6) provide a very important insight into how rMATS-Iso converts information regarding the probabilistic behavior of the individual isoforms into information regarding the probabilistic behavior of the observed counts of consistency patterns. Each $\theta_{mf}^{(r)}$ is a deterministic quantity determined entirely by the structure of the alternative splicing module and RNA-seq read length, and can be computed using a straightforward algorithm (a formal derivation of $\theta_{mf}^{(r)}$ can be found in the appendix, section 3.8.4). If we define the column vector $\theta_f^{(r)} = (\theta_{1f}^{(r)}, \ldots, \theta_{Mf}^{(r)})^T$ and stack all of the $\theta_f^{(r)}$ vectors into a matrix $\Theta^{(r)} = [\theta_1^{(r)}, \ldots, \theta_n^{(r)}]$, then the probability vector $p_r(\psi_k)$ in equation (3.5) can be written more compactly as

$$p_r(\psi_k) = \Theta^{(r)} \cdot \tilde{\psi}_k.$$

Seen from this perspective, $\Theta^{(r)}$ defines a linear mapping T such that

$$T : D_n \to D_M,$$

where $D_\ell = \{(x_1, \ldots, x_\ell) : \sum_{i=1}^{\ell} x_i = 1, 0 \leq x_i \leq 1 \ \ \forall i = 1, \ldots, \ell\}$. More intuitively, the matrix $\Theta^{(r)}$ transforms the probabilities $\tilde{\psi}$ from n dimensional "isoform" space into M dimensional "pattern" space via the transformation

$$T(\tilde{\psi}) = \Theta^{(r)} \cdot \tilde{\psi}.$$

The defining feature of the rMATS-Iso framework is that the data are observed in M dimensional pattern space with probabilities defined over D_M, whereas inference occurs in the original n dimensional isoform space with probabilities defined over D_n. As an aside, note that Equations (3.2) and (3.5) will not yield a conjugate likelihood function since $M \neq n$. It is also worth noting that our derivation of $\theta_{mf}^{(r)}$ assumes uniform read coverage within each splicing module; in other words, we assume reads are equally likely to occur in any given location within the module (see section 3.8.4 for more details regarding the computation of $\Theta^{(r)}$).

3.3 An EM Algorithm for Estimating the Model Parameters

Since the isoform inclusion levels ψ_{kf} are not observable, we use the EM algorithm to estimate the Dirichlet parameter α in Equation (3.2). The E step of the EM algorithm involves computing the conditional expectation of the joint log-likelihood of the observed and unobserved variables. In the case of rMATS-Iso, this expectation is given by

$$\sum_{k=1}^{K} E_{\psi_k}\left[\log P(Y_k, \psi_k)|Y_k, \alpha^{(t)}\right] = \sum_{k=1}^{K}\sum_{f=1}^{n} \alpha_f E_{\psi_k}\left[\log \psi_{kf}|Y_k, \alpha^{(t)}\right] - K \log B(\alpha) + c, \quad (3.7)$$

where $E_{\psi_k}[\cdot]$ denotes expectation with respect to ψ_k, $B(\alpha)$ is the multivariate beta function, and c is a constant that does not depend on the parameter α. $\alpha^{(t)}$ denotes the parameter value during the t^{th} iteration of the EM algorithm (the starting value $\alpha^{(0)}$ can be randomly initialized). Unfortunately, there is no closed form expression for the distribution $P(\psi_k|Y_k, \alpha^{(t)})$ required to compute the expectation on the right side of Equation

(3.7); however, using importance sampling, Equation (3.7) can be approximated as

$$\sum_{k=1}^{K} E_{\psi_k}\left[\log P(Y_k, \psi_k)|Y_k, \alpha^{(t)}\right] \approx \sum_{k=1}^{K}\sum_{f=1}^{n} \alpha_f \frac{\sum_{s=1}^{S} \log \psi_f^{(s)} \omega_k(\psi^{(s)}, Y_k)}{\sum_{s=1}^{S} \omega_k(\psi^{(s)}, Y_k)} - K \log B(\alpha) + c, \tag{3.8}$$

where $\psi^{(1)}, \ldots, \psi^{(S)} \overset{iid}{\sim} \text{Dirichlet}(\alpha^{(t)})$, and where $\omega_k(\psi^{(s)}, Y_k)$ are the importance ratios (more details about the importance sampling procedure are provided in the appendix, section 3.8.2).

The full EM-algorithm for estimating α is provided in Algorithm 1. The algorithm runs until the difference in consecutive parameter estimates is sufficiently small:

$$\|\alpha^{(t)} - \alpha^{(t-1)}\|^2 \leq \epsilon.$$

We use $\epsilon = 0.01$, and $S = 500$ throughout the manuscript. The optimization in Algorithm 1 can be performed using a constrained optimization routine such as the L-BFGS-B optimization algorithm. A formal derivation for each component of Algorithm 1 can be found in various sections of the appendix.

3.4 Testing for Differential Splicing Between Two Groups

In addition to estimating the distributions of isoform inclusion levels, rMATS-Iso can also be used to detect differential splicing between two sample groups. More specifically, we assume that the isoform inclusion levels for each sample group are drawn from the following distributions:

$$\psi_{11}, \ldots, \psi_{1K_1} \overset{iid}{\sim} \text{Dirichlet}(\alpha_1)$$
$$\psi_{21}, \ldots, \psi_{2K_2} \overset{iid}{\sim} \text{Dirichlet}(\alpha_2)$$

Using a likelihood-ratio test, rMATS-Iso tests whether there is a difference between α_1 and α_2, the Dirichlet parameters corresponding to sample groups 1 and 2, against the null

Algorithm 1: EM algorithm for estimating Dirichlet parameter α

Input : $Y_1, \ldots, Y_K$, $\Theta = \{\Theta^{(r_1)}, \ldots, \Theta^{(r_{L_k})}\}$, $\ell = \{\ell_1, \ldots, \ell_n\}$, S, T, ϵ

/* $Y_1, \ldots, Y_K$ are the multinomial pattern count matrices */

/* Θ contains the isoform to pattern probabilities */

/* ℓ contains the effective lengths of each isoform */

/* S is the number of samples to draw */

/* T is the number of EM iterations */

/* ϵ threshold for terminating EM algorithm */

Output: $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)$

Randomly initialize $\alpha^{(1)}$;

for $t = 1, \ldots, T$, **do**

  Draw $\psi^{(1)}, \ldots, \psi^{(S)} \overset{iid}{\sim} \text{Dirichlet}(\alpha^{(t)})$;

  Set $\tilde{\psi}_f^{(s)} \leftarrow \dfrac{\ell_f \psi_f^{(s)}}{\sum_{j=1}^n \ell_j \psi_j^{(s)}}$ for $s = 1, \ldots, S$ and $f = 1, \ldots, n$; // Length normalization

  Set $\omega_k(\psi^{(s)}, Y_k) \leftarrow \prod_{r \in \mathcal{R}_k} \prod_{m=1}^{M} \left(\sum_{j=1}^{n} \tilde{\psi}_j^{(s)} \theta_{mj}^{(r)} \right)^{Y_{krm}}$; // Importance ratios

  Set $\alpha^{(t+1)} \leftarrow \underset{\alpha}{\text{argmax}} \left(\sum_{k=1}^{K} \sum_{f=1}^{n} \alpha_f \left(\frac{\sum_{s=1}^{S} \log \psi_f^{(s)} \omega_k(\psi^{(s)}, Y_k)}{\sum_{s=1}^{S} \omega_k(\psi^{(s)}, Y_k)} \right) - K \log B(\alpha) \right)$ st $\alpha_f > 0 \;\; \forall f = 1, \ldots, n$;

  if $\|\alpha^{(t+1)} - \alpha^{(t)}\|^2 < \epsilon$ **then**

    continue;

  end

end

return $\alpha^{(t+1)}$;

hypothesis that $\alpha_1 = \alpha_2$:

$$H_0 : \alpha_1 = \alpha_2$$

$$H_a : \alpha_1 \neq \alpha_2.$$

Moreover, if the null hypothesis of equality between α_1 and α_2 is rejected, a p-value is assigned to each of the f transcripts to assess how strongly differences in the corresponding α_{1f} and α_{2f} contribute to the overall difference in splicing between groups. More details regarding the likelihood ratio test and derivation of individual-isoform p-values can be found in the appendix, section 3.8.3.

3.5 rMATS-Iso Simulation Studies

To assess the performance of rMATS-Iso, we performed a series of simulation studies using data generated from the Flux-Simulator (v 1.2.1) (Griebel et al., 2012). The Flux-Simulator software aims to reproduce the sources of variability present in real RNA-seq experiments such as those introduced during reverse transcription, RNA fragmentation, library preparation, and high-throughput sequencing. Thus, simulating data from the Flux-Simulator is more realistic than directly simulating from the rMATS-Iso statistical model, where no sources of experimental error are assumed to exist. Six samples were simulated in a 3 vs. 3 comparison of the PC3E (epithelial) and GS689 (mesenchymal) prostate cancer cell lines. Each sample had 200 million 101-bp paired-end reads. The transcript structure was based on the Ensembl Human GRCh37.87 GTF annotation. For each one of the six samples, the simulated true value of transcript expression was based on the transcription expression of one of the 3 PC3E samples and 3 GS689 samples respectively. RNA-seq data were generated using transcripts from the PC3E and GS689 cell lines, where reads from splicing modules were simulated for 3 replicates in each sample group / cell line. The number of transcripts within each splicing module ranged from 2 to

5. The data were subsequently input into the rMATS-Iso software package with the goal of quantifying splicing as well as detecting differential splicing events between the two groups. Modules containing transcripts for which no isoform-specific reads were possible, as well as modules with fewer than 100 reads in both sample groups were removed prior to the analysis, resulting in a total of 3,823 AS events being analyzed.

Since the Dirichlet parameters of the two sample groups were not set in advance, we defined whether two groups were differentially spliced according to the following criterion: modules where the largest absolute difference in average psi values between groups (averaged across replicates) for at least one isoform was greater than t_{alt} for some threshold $0.05 \leq t_{alt} \leq 0.15$ were set equal to alternative cases (differential splicing, i.e. $\alpha_1 \neq \alpha_2$). Modules where the absolute difference in average psi values between groups for every isoform was less than t_{null} for some threshold $0.01 \leq t_{null} \leq 0.05$ were defined to be null cases (no differential splicing, i.e $\alpha_1 = \alpha_2$). A similar definition was used to define null and alternative cases in individual isoforms within modules generated from the alternative hypothesis (individual isoforms within splicing modules generated from the null hypothesis were defined to be null-cases themselves). Null/alternative cases for individual isoforms were only defined for modules containing more than 2 isoforms.

The simulation results reveal that rMATS-Iso was able to accurately estimate isoform inclusion levels across splicing modules as well as identify differential splicing events (Figure 3.2). There was strong and significant correlation between the true psi values and those predicted by rMATS-Iso (Pearson's correlation coefficient $r = 0.98$, p-value $< 2.2e^{-16}$; Figure 3.2A). Moreover, the strengths of the correlations were consistent across modules with different numbers of isoforms, with correlations of $0.99, 0.98, 0.94, 0.95$ (all p-values $< 2.2e^{-16}$) for modules with 2,3,4, and 5 transcripts, respectively (Figure 3.3). To quantify the performance of rMATS-Iso at identifying differential splicing events, we computed the area under the curve (AUC) of the receiver operating characteristic curves (ROC) using

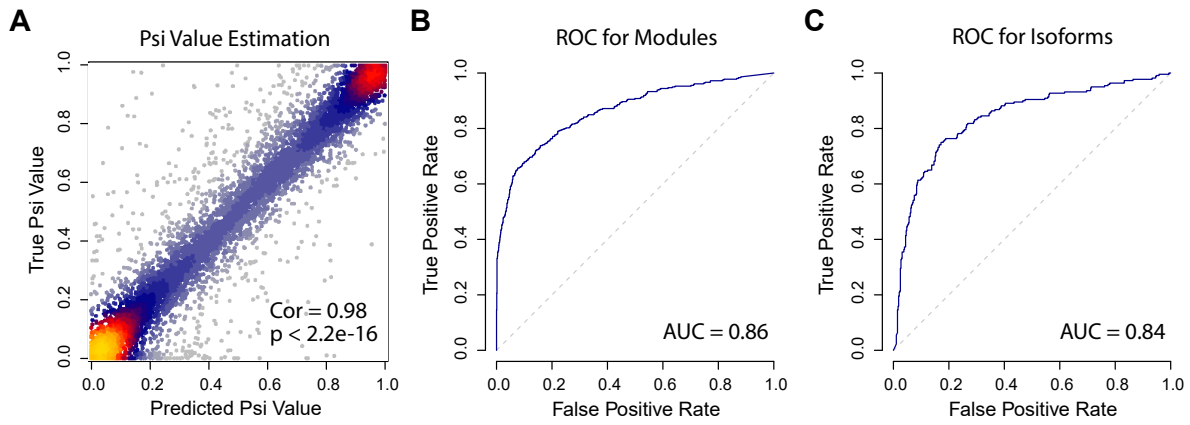

Figure 3.2: rMATS-Iso is able to accurately estimate isoform inclusion levels across splicing modules as well as identify differential splicing events using data simulated from the Flux-Simulator. (A) There is high and significant concordance between the true ψ values and those predicted by rMATS-Iso (Pearson's correlation coefficient $r = 0.98$, p-value $< 2.2e^{-16}$). (B) ROC curve for the task of identifying differential alternative splicing between splicing modules (AUC = 0.86). (C) ROC curve for the task of identifying pairwise differences in individual isoforms' inclusion levels (AUC = 0.84).

the p-values generated from rMATS-Iso. rMATS-Iso was effectively able to identify differential splicing events while maintaining robustness against non-significant differences in isoform inclusion levels between sample groups. For example, for the choice of thresholds $t_{null} = 0.01$ and $t_{alt} = 0.10$, rMATS-Iso achieves an AUC of 0.86 for identifying differential splicing events (Figure 3.2B), and an AUC of 0.84 for identifying differentially spliced isoforms within modules (Figure 3.2C). Table 3.1 shows the AUC results over a range of values for the thresholds t_{null} and t_{alt}.

Next, we performed another simulation where data were generated from the rMATS-Iso statistical model directly, using the same module structures as before. More specifically, for each splicing module from the Flux-Simulated dataset, we randomly decided (with 50% probability) whether or not that module was differentially spliced. If the

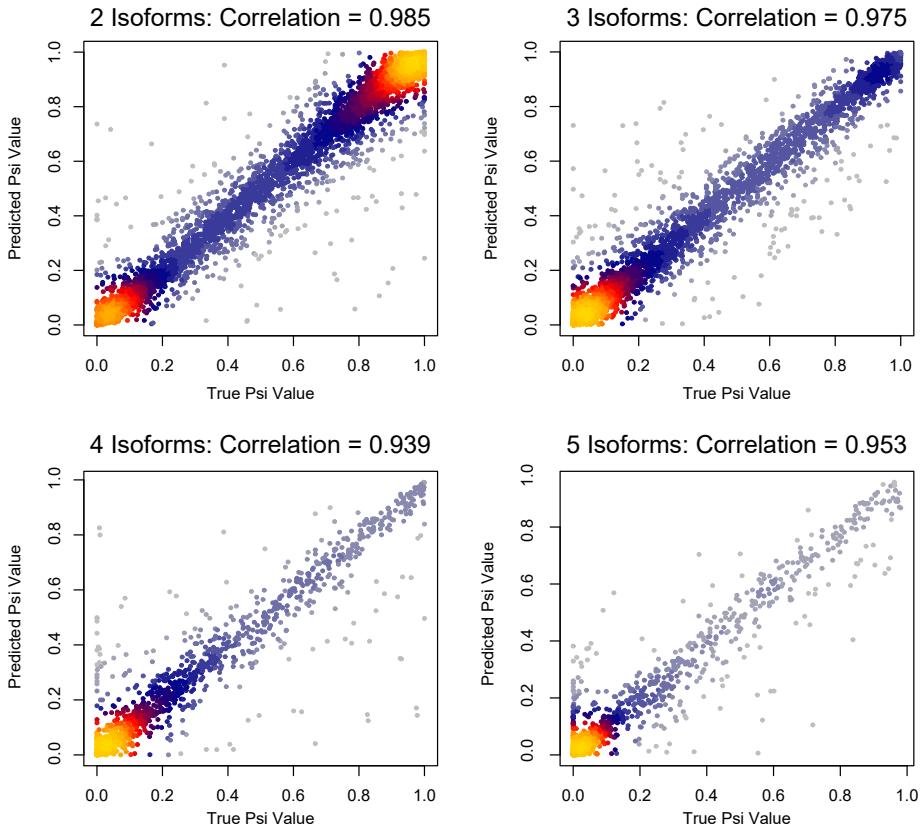

Figure 3.3: There is high and significant concordance between the true ψ values and those predicted by rMATS-Iso across modules with different numbers of isoforms (Pearson's correlation coefficient p-value $< 2.2e^{-16}$ for each plot).

module corresponded to differential splicing, we randomly sampled the parameter vectors α_1 and α_2 jointly from the empirical distributions of unconstrained parameter estimates obtained from the previous simulation, specifically from modules where the largest absolute difference in average psi values was greater than $t_{alt} = 0.10$. If the splicing module corresponded to the null hypothesis of no differential splicing, we randomly sampled $\alpha_1 = \alpha_2 = \alpha$ from the empirical distribution of constrained parameter estimates from modules where the absolute difference in average psi values between groups for every isoform was less than $t_{null} = 0.01$. For this simulation, we treated the total number of RNA-seq reads for each replicate, as well as the number of replicates in each

t_{alt} \ t_{null}	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**
0.05	0.818	0.795	0.780	0.769	0.760
0.10	0.863	0.845	0.832	0.824	0.816
0.15	0.896	0.882	0.873	0.866	0.861
0.20	0.904	0.894	0.887	0.883	0.879

Table 3.1: The area under curve (AUC) for the classification task in Figure 3.2B for different values of the thresholds t_{null} and t_{alt}.

sample group, K_1, K_2, as parameters and varied them from simulation to simulation ($R = 50, 100, 200, 400, 500, 1000$, and $K_1 = K_2 = 2, 3, 5, 10, 20, 50$).

The results are shown in Figure 3.4 and illustrate that increasing either the total number of reads or number of replicates increases the performance of rMATS-Iso in detecting differential splicing events. Figure 3.4 can also provide useful insight into performing differential splicing analysis with a fixed budget of replicates/reads. For example, if obtaining additional replicates is prohibitively costly for a particular study, significant gains in AUC can be obtained simply by increasing the sequencing depth, especially if the original sequencing depth is shallow.

3.6 Analysis of the PC3E and GS689 Cell Lines Using Long RNA-seq Reads

To illustrate the utility of rMATS-Iso, we analyzed an RNA-seq dataset generated from six independent samples of the PC3E (epithelial) and GS689 (mesenchymal) prostate cancer cell lines (three samples per cell line) discussed in the previous subsection. As part

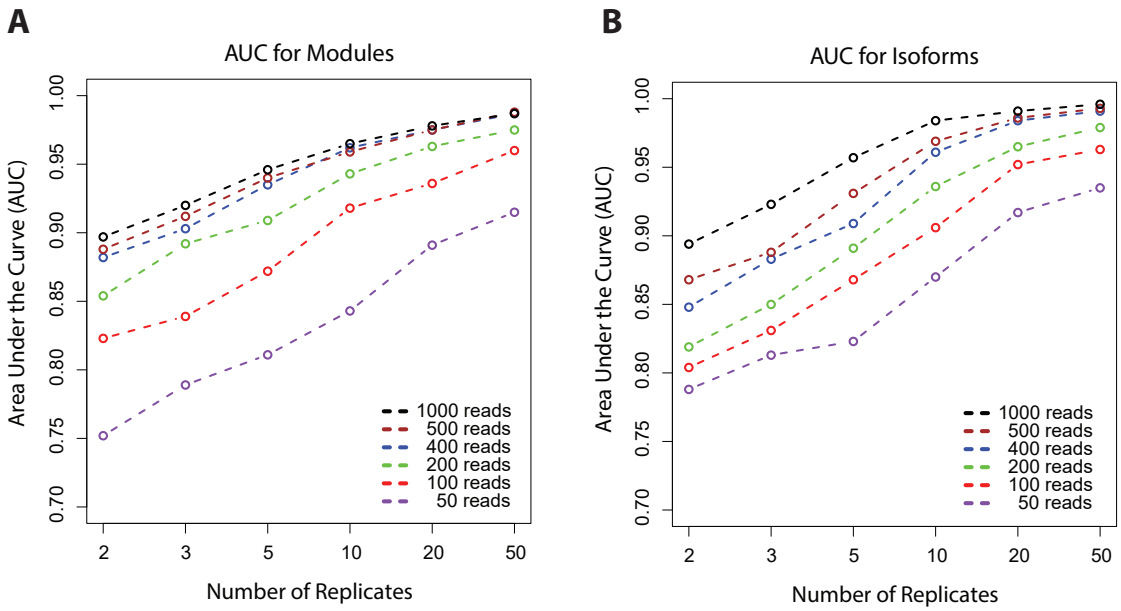

Figure 3.4: The performance of rMATS-Iso improves as more replicates and/or RNA-seq reads are added. (A) AUC for the task of identifying differential splicing events. (B) AUC for the task of identifying differences in individual isoforms' inclusion levels.

of the data processing pipeline, we first created a gene annotation file using both short RNA-seq reads as well as Pacific Biosciences long RNA-seq read data generated from all six samples of PC3E and GS689. Next, the resulting gene annotation file along with the alignment results were used to generate all of the necessary input for the rMATS-Iso statistical model, including a file containing the splicing module structures as well as files containing isoform consistency counts for each replicate in every splicing module. More details about the data processing pipeline are given in the appendix, section 3.8.5.

The most abundant splicing patterns, along with the corresponding numbers of significant differential splicing events, are shown in Figure 3.5. Among the 10 most abundant splicing patterns, rMATS-Iso identified 712 significant events from a total of 13, 211 events analyzed. These events include the basic patterns of alternative splicing such as exon skipping (456 significant events), alternative 3' (62 significant events) and 5' (15 significant

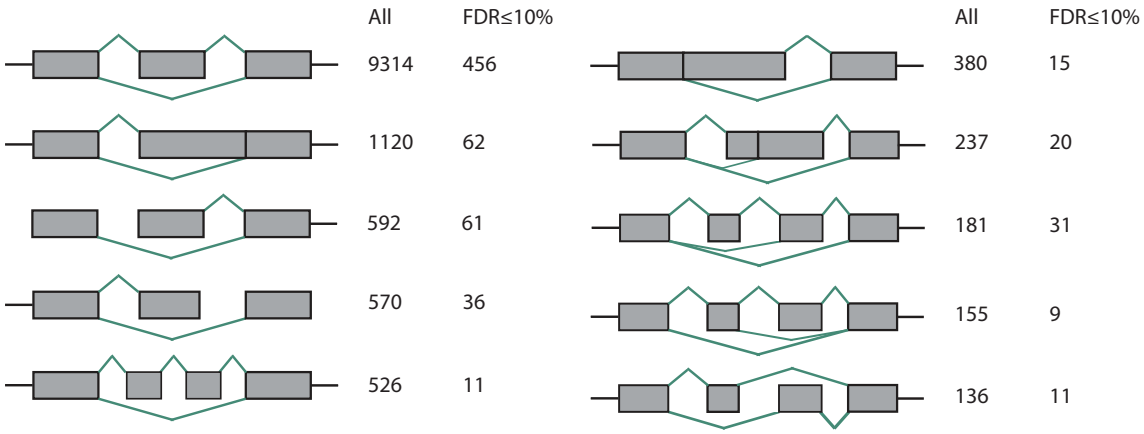

Figure 3.5: The most abundant splicing patterns in PC3E identified by rMATS-Iso along with the corresponding number of significant differential splicing events at FDR $\leq 10\%$. Among the most abundant splicing patterns, 712 significant events were identified from a total of $13,211$ events analyzed.

events) splice sites, alternative first exons (61 significant events) and alternative last exons (36 significant events), as well as more complex events such as exon skipping coupled with alternative 3' splice site (20 significant events), and mutually exclusive exons (11 significant events).

An example of a complex differential splicing event identified by rMATS-Iso is shown in Figure 3.6. This event corresponds to an exon skipping + alternative $5'$ splice site event in the FLNB gene (rMATS-Iso p-value $< 2.2e^{-16}$). There were significant differences in the inclusion levels of isoforms 2 and 3 between the two sample groups. The estimated mean inclusion level for isoform 2 is 76% in PC3E vs 24% in GS689 (p-value of isoform difference = 0.019), while the estimated mean inclusion level for isoform 3 is 1% in PC3E vs 73% in GS689 (p-value of isoform difference = 0.00015). The FLNB gene codes for the protein filamin B which forms the cytoskeleton and is important for normal cell growth; differential expression of the FLNB splice variants has previously been shown to be associated with cell survival and differentiation of giant cell bone tumors (Tsui et al.,

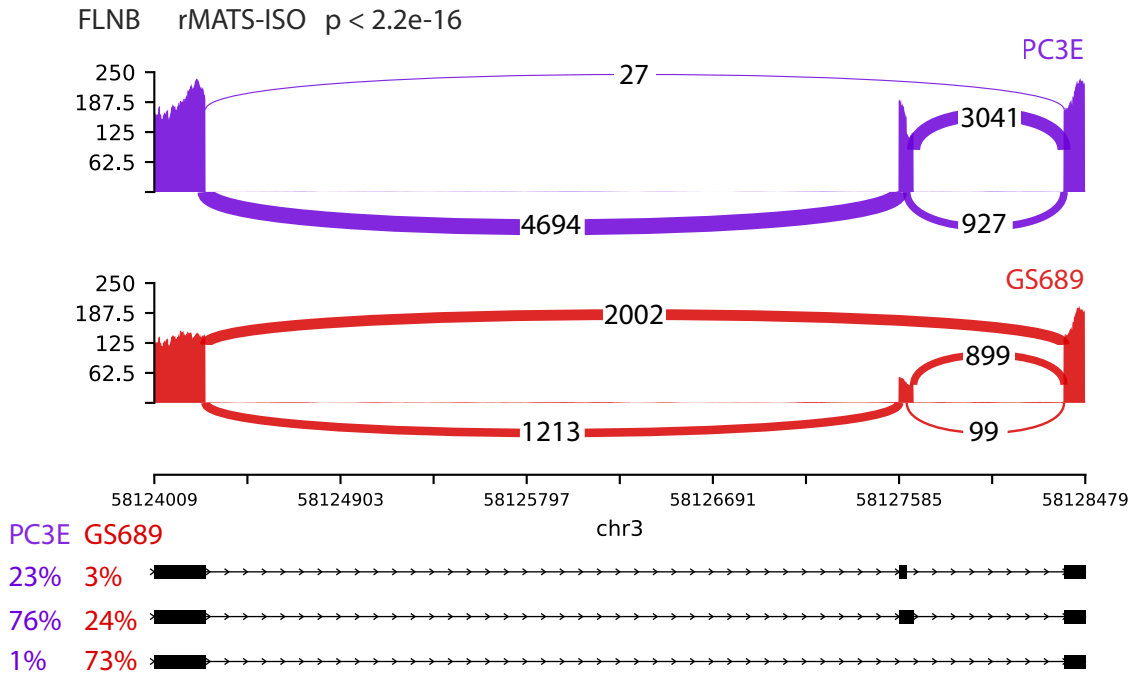

Figure 3.6: A significant differential splicing event identified by rMATS-Iso in the FLNB gene (rMATS-Iso p $< 2.2e^{-16}$). Top) Sashimi plots indicating the read counts corresponding to each exon junction in each group. Bottom) Mean isoform inclusion levels estimated using rMATS-Iso for each sample group.

2016).

Another differential alternative splicing event identified by rMATS-Iso is shown in Figure 3.7. This event corresponds to a mutually exclusive exon splicing event in the MYO1B gene (rMATS-Iso p-value $< 2.2e^{-16}$). The largest differences in isoform ratios between the PC3E and GS689 cell lines occur between isoforms 1 and 4 corresponding to inclusion of all 4 exons and skipping of both exons 2 and 3 respectively. The estimated mean inclusion level for isoform 1 is 6% in PC3E vs 56% in GS689 (p-value of isoform difference = 0.0002), while the estimated mean inclusion level for isoform 4 is 81% in PC3E vs 23% in GS689 (p-value of isoform difference = $7e^{-05}$). Aberrant expression of the MYO1B gene has been linked to cell migration and lymph node metastasis in patients

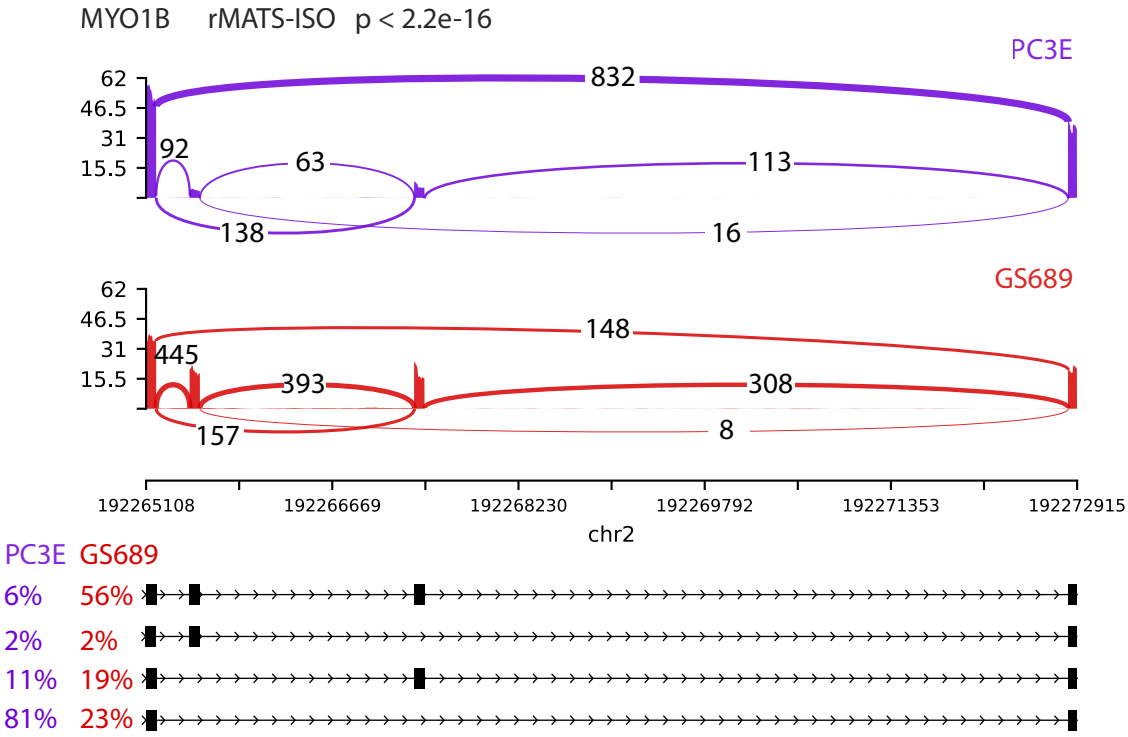

Figure 3.7: A significant differential splicing event identified by rMATS-Iso in the MYO1B gene (rMATS-Iso p $< 2.2e^{-16}$). Top) Sashimi plots indicating the read counts corresponding to each exon junction in each group. Bottom) Mean isoform inclusion levels estimated using rMATS-Iso for each sample group.

with certain classes of cancer (Ohmura et al., 2015).

3.7 Discussion

In this chapter, we have proposed a novel statistical framework, rMATS-Iso, to fill a methodological gap in the literature on quantifying isoform variation in complex splicing modules. rMATS-Iso accounts both for the estimation uncertainty in ψ values owing to RNA-seq read coverage, as well as for the variability in splicing levels between samples from the same biological population. Moreover, to address the possible ambiguity of RNA-

seq read counts, rMATS-Iso converts the underlying isoform probabilities from isoform space to pattern space where there is no problem or ambiguity defining the observed read counts. rMATS-Iso leverages a likelihood ratio test in order to detect differential splicing between sample groups, and utilizes a simulation-based approach to quantify individual-isoform differences between groups. Our simulation results reveal that rMATS-Iso accurately estimates the true psi values within the PC3E and GS689 cell lines, and that adding samples and/or increasing the RNA-seq read coverage increases the accuracy of detecting differential events.

There are several possible extensions to the rMATS-Iso model. First, our computation of the isoform-to-pattern probability matrix $\Theta^{(r)}$ assumes that RNA-seq reads are equally likely to be drawn across each feasible position in a given splicing module. In practice, the distribution of where RNA-seq reads occur across a module may be non-uniform, and a more realistic model to reflect this departure from uniformity may be more appropriate. In addition, our model assumes that alternative splicing levels are fully determined by factors within the splicing module. In reality, alternative splicing levels may depend on other factors and may even differ within sub-strata of the population of interest. Integrating this information as an additional layer in the hierarchical framework is a promising direction for future work.

3.8 Appendix

3.8.1 Normalizing Isoform Lengths

Since the effective lengths of each isoform (defined as the number of isoform-specific read positions) can differ between isoforms, the isoform inclusion levels ψ must be appropriately normalized to obtain the proportion of reads generated by each isoform. Let ℓ_f denote the effective length of isoform f (see (Shen et al., 2014) for more details on computing the

effective lengths for different types of alternative splicing events). Then for $f = 1, \ldots, n$, the proportion of reads generated from isoform f is given by

$$\tilde{\psi}_f = \frac{\ell_f \cdot \psi_f}{\sum_{j=1}^{n} \ell_j \cdot \psi_j}.$$

$\tilde{\psi}$ is used in equation 3.6 to adjust the multinomial read counts corresponding to each consistency pattern.

3.8.2 Approximating the Conditional Expectation of the Log-Likelihood

Let $\mathcal{R}_k = \{r_1, \ldots, r_{L_k}\}$ denote the set of unique read lengths for the reads corresponding to subject k. Combining Equations 3.2 and 3.5 yields the joint probability of Y_k and ψ_k, given by

$$P(Y_k, \psi_k; \alpha) \propto P(\psi_k; \alpha) P(Y_k|\psi_k) \propto \frac{1}{B(\alpha)} \prod_{f=1}^{n} \psi_{kf}^{\alpha_f - 1} \prod_{r \in \mathcal{R}_k} \prod_{m=1}^{M} p_{rm}(\psi_k)^{Y_{krm}},$$

where $B(\alpha)$ is the multivarite beta function

$$B(\alpha) = \frac{\prod_{f=1}^{n} \Gamma(\alpha_f)}{\Gamma\Big(\sum_{f=1}^{n} \alpha_f\Big)},$$

and where $p_{rm}(\psi_k)$ is defined as in 3.6. The log-likelihood function can therefore be written as

$$\sum_{k=1}^{K} \log P(Y_k, \psi_k; \alpha) = \sum_{k=1}^{K} \sum_{f=1}^{n} \alpha_f \log \psi_{kf} - K \log B(\alpha) + C_0, \tag{3.9}$$

where C_0 is a constant that does not depend on α. Finally, taking the conditional expectation of expression 3.9 yields

$$\sum_{k=1}^{K} E_{\psi_k}\left[\log P(Y_k, \psi_k; \alpha)|Y_k, \alpha^{(t)}\right] = \sum_{k=1}^{K} \sum_{f=1}^{n} \alpha_f E_{\psi_k}\left[\log \psi_{kf}|Y_k, \alpha^{(t)}\right] - K \log B(\alpha) + C_0. \tag{3.10}$$

Since there is no closed form expression for $P(\psi_k|Y_k, \alpha^{(t)})$, we can use importance sampling to approximate the expectation in the above expression. More specifically, let $Q(\psi|\alpha^{(t)})$ denote the density function of a Dirichlet($\alpha^{(t)}$) random vector, and let $\psi^{(1)}, \ldots, \psi^{(S)}$ denote an iid sample drawn from $Q(\psi|\alpha^{(t)})$. Also, let

$$P(\psi^{(s)}|Y_k, \alpha^{(t)}) := c_p \cdot P_0(\psi^{(s)}|Y_k, \alpha^{(t)})$$
$$Q(\psi|\alpha^{(t)}) := c_q \cdot Q_0(\psi|\alpha^{(t)}),$$

where c_p and c_q are the normalizing constants of their respective density functions. Note that

$$\begin{aligned}\frac{P_0(\psi^{(s)}|Y_k, \alpha^{(t)})}{Q_0(\psi^{(s)}|\alpha^{(t)})} &\propto \frac{P(Y_k|\psi^{(s)}, \alpha^{(t)})Q_0(\psi^{(s)}|\alpha^{(t)})}{Q_0(\psi^{(s)}|\alpha^{(t)})} = P(Y_k|\psi^{(s)}) \\ &\propto \prod_{r\in\mathcal{R}_k}\prod_{m=1}^{M}\Big(\sum_{j=1}^{n}\tilde{\psi}_j^{(s)}\theta_{mj}^{(r)}\Big)^{Y_{krm}} \\ &:= \omega_k(\psi^{(s)}, Y_k). \qquad (3.11)\end{aligned}$$

The conditional expectation in 3.10 can now be written as

$$\begin{aligned}E_{\psi_k}\left[\log\psi_{kf}|Y_k, \alpha^{(t)}\right] &= \int \log\psi_{kf}P(\psi_k|Y_k, \alpha^{(t)})d\psi_k \\ &= \int\left(\log\psi_{kf}\cdot\frac{P(\psi_k|Y_k, \alpha^{(t)})}{Q(\psi_k|\alpha^{(t)})}\right)Q(\psi_k|\alpha^{(t)})d\psi_k \\ &= \frac{\int\left(\log\psi_{kf}\cdot\frac{P(\psi_k|Y_k, \alpha^{(t)})}{Q(\psi_k|\alpha^{(t)})}\right)Q(\psi_k|\alpha^{(t)})d\psi_k}{\int\frac{P(\psi_k|Y_k, \alpha^{(t)})}{Q(\psi_k|\alpha^{(t)})}Q(\psi_k|\alpha^{(t)})d\psi_k} \\ &= \frac{\int\left(\log\psi_{kf}\cdot\frac{P_0(\psi_k|Y_k, \alpha^{(t)})}{Q_0(\psi_k|\alpha^{(t)})}\right)Q(\psi_k|\alpha^{(t)})d\psi_k}{\int\frac{P_0(\psi_k|Y_k, \alpha^{(t)})}{Q_0(\psi_k|\alpha^{(t)})}Q(\psi_k|\alpha^{(t)})d\psi_k} \\ &= E_{\psi_k}\left[\log\psi_{kf}\cdot\frac{P_0(\psi_k|Y_k, \alpha^{(t)})}{Q_0(\psi_k|\alpha^{(t)})}\middle|\alpha^{(t)}\right] \Big/ E_{\psi_k}\left[\frac{P_0(\psi_k|Y_k, \alpha^{(t)})}{Q_0(\psi_k|\alpha^{(t)})}\middle|\alpha^{(t)}\right]\end{aligned}$$

$$\approx \frac{\sum_{s=1}^{S} \log \psi_f^{(s)} \cdot \omega_k(\psi^{(s)}, Y_k)}{\sum_{s=1}^{S} \omega_k(\psi^{(s)}, Y_k)} \tag{3.12}$$

when the sample size S is large enough.

3.8.3 Likelihood Ratio Test and Derivation of Individual-Isoform p-values

When there are two sample groups, rMATS-Iso assumes that the isoform inclusion levels from each group are drawn from the following distributions:

$$\psi_{11}, \ldots, \psi_{1K_1} \overset{iid}{\sim} \text{Dirichlet}(\alpha_1)$$
$$\psi_{21}, \ldots, \psi_{2K_2} \overset{iid}{\sim} \text{Dirichlet}(\alpha_2).$$

A likelihood ratio test is then used to test the hypothesis $\alpha_1 \neq \alpha_2$ against the null hypothesis that $\alpha_1 = \alpha_2$:

$$H_0 : \alpha_1 = \alpha_2$$
$$H_a : \alpha_1 \neq \alpha_2.$$

The log likelihood function for the two sample situation can be written as

$$\begin{aligned} L(\alpha_1, \alpha_2 | Y_1, Y_2, \psi_1, \psi_2) &= \sum_{k=1}^{K_1} \log P(Y_{1k}, \psi_{1k}; \alpha_1) + \sum_{k=1}^{K_2} \log P(Y_{2k}, \psi_{2k}; \alpha_2) \\ &= \sum_{k=1}^{K_1} \sum_{f=1}^{n} \alpha_{1f} \log \psi_{1kf} + \sum_{k=1}^{K_2} \sum_{f=1}^{n} \alpha_{2f} \log \psi_{2kf} \\ &\quad - K_1 \log B(\alpha_1) - K_2 \log B(\alpha_2) + C_1, \end{aligned}$$

where the constant C_1 does not depend on either α_1 or α_2. The EM algorithm proceeds analogously as before, yielding two estimates $\hat{\alpha}_1$, $\hat{\alpha}_2$ under the alternative (unconstrained) hypothesis and one estimate $\hat{\alpha}$ under the null (constrained) hypothesis. Under this setup, the likelihood ratio test statistic asymptotically follows a χ^2 distribution with n degrees of freedom.

Algorithm 2: Algorithm for computing isoform-specific p-values

Input : $\hat{\alpha}, \hat{\alpha}_1, \hat{\alpha}_2, T$

/* $\hat{\alpha}$ is the estimate of α under the null hypothesis */

/* $\hat{\alpha}_1$ is the estimate of α_1 under the alternative hypothesis */

/* $\hat{\alpha}_2$ is the estimate of α_2 under the alternative hypothesis */

/* T is the number of rounds of simulation to perform */

Output: $(p_1, \ldots, p_n)$, where p_f is the p-value for isoform f.

Set $D \leftarrow \dfrac{\hat{\alpha}_1}{\sum_{i=1}^{n} \hat{\alpha}_{1i}} - \dfrac{\hat{\alpha}_2}{\sum_{i=1}^{n} \hat{\alpha}_{2i}}$;

for $t = 1, \ldots, T$, **do**

  Draw $\psi_{1k}^{(t)} \overset{iid}{\sim} \text{Dirichlet}(\hat{\alpha})$ for $k = 1, \ldots, K_1$;

  Draw $\psi_{2k}^{(t)} \overset{iid}{\sim} \text{Dirichlet}(\hat{\alpha})$ for $k = 1, \ldots, K_2$;

  Set $D^t \leftarrow \bar{\psi}_1^{(t)} - \bar{\psi}_2^{(t)} = \dfrac{1}{K_1} \sum_{k=1}^{K_1} \psi_{1k}^{(t)} - \dfrac{1}{K_2} \sum_{k=1}^{K_2} \psi_{2k}^{(t)}$;

end

for $f = 1, \ldots, n$, **do**

  Set $p_f \leftarrow \dfrac{1}{T} \sum_{t=1}^{T} I\{D_f^t < -|D_f|\} + \dfrac{1}{T} \sum_{t=1}^{T} I\{D_f^t > |D_f|\}$;

end

return $p = (p_1, \ldots, p_n)$;

If the null hypothesis $\alpha_1 = \alpha_2$ is rejected, we may be interested in quantifying how strongly each individual isoform contributes to the differences observed in the data. With this goal in mind, we use a simulation-based strategy that computes p-values for each isoform; a small p-value indicates that an isoform contributes significantly to the difference in overall isoform inclusion levels (of course, there is no need for such a procedure if there

are only 2 isoforms). First, note that under the null hypothesis $\alpha_1 = \alpha_2 = \alpha$,

$$E[\psi_{jf}] = \frac{\alpha_f}{\sum_{i=1}^{n} \alpha_i} \approx \frac{\hat{\alpha}_f}{\sum_{i=1}^{n} \hat{\alpha}_i} \approx \frac{1}{K_j} \sum_{k=1}^{K_j} \psi_{jkf} = \bar{\psi}_{jf}, \quad j = 1, 2.$$

Taking advantage of the above approximation, the p-values for each isoform can be approximated using Algorithm 2. We set the default number of simulations to $T = 100,000$.

3.8.4 Computing the Transcript to Pattern Probabilities

The probabilities $\theta_{mf}^{(r)}$ in Equation 3.6 represent the probability that a read will correspond to pattern m, given the read was generated from isoform f. Each $\theta_{mf}^{(r)}$ value is a fixed (non-random) quantity and can be determined by the structure of the alternative splicing module and RNA-seq read length. These probabilities can be calculated using the following three steps:

Step 1: Find the effective length of each isoform.

Step 2: Find the effective length of each pattern.

Step 3: Normalize the effective length of each pattern by the effective length of each isoform.

The following scheme can be used to compute $\Theta^{(r)}$:

Step 1: *Find all of the exon bodies and junctions in which a read will be consistent with a given isoform.*

Let $I = \{1, \ldots, n\}$ and for isoform $f = 1, \ldots, n$, do:

(A) Let $I_f = \{i \in I;$ exon i is included in isoform $f\}$, and let $n_f = |I_f|$.

(B) Let $E_f^r = \{i \in I_f; e_i \geq r\}$, where e_i is the length of exon i. E_f^r tracks all exons which can fully contain a read of length r.

(C) For every $k \in \{2, \ldots, n_f\}$, defined $S_{fk} = \{(i_1, \ldots, i_k); i_j \in I_f \text{ for } j = 1, \ldots, k\}$, i.e. S_{fk} contains all k-tuples of indices from I_f. Let $S_f = \cup_{k=2}^{n_f} S_{fk}$. We would like to

define a subset of S_f that tracks all feasible exon junctions within isoform f. With this goal in mind, define J_f^r to be the subset of S_f such that for all $z \in J_f^r$,

(a) $z_j < z_{j+1}$ for all $j = 1, \ldots, \ell_z - 1$, where ℓ_z is the length of z. Furthermore, the exons indexed by z_j and z_{j+1} are adjacent in isoform f.

(b) $r \leq \sum_{j=1}^{\ell_z} e_{z_j} < e_{z_1} + r - 1$; this condition states that the total length of the exons indexed by z is greater than or equal to the read length r, and that at least one read can cover both the first and last exons indexed by z.

Each ordered pair (i, j) in J_f^r corresponds to the junction between exons i and j. k-tuples with $k \geq 3$ indicate a "multi-exon junction", i.e. a configuration for which a read crosses exactly $k - 1$ junctions.

Step 2: *Identify the exon bodies and junctions in which a read will uniquely correspond to each isoform consistency pattern. Then find the effective length, i.e. the total number of unique read positions, of each consistency pattern.*

Enumerate all $M = 2^n - 1$ consistency patterns into an $M \times n$ matrix P as defined in section 3.2.

(A) For $m = 1, \ldots, M$, let

$$H_m = \{f \in I; P_{mf} = 1\}.$$

For all $f \in H_m$, define

$$E_{fm}^r := \bigcap_{j \in H_m^c} (E_f^r - E_j^r),$$

$$J_{fm}^r := \bigcap_{j \in H_m^c} (J_f^r - J_j^r).$$

Finally, let

$$\tilde{E}_{fm}^r := \bigcap_{f \in H_m} E_{fm}^r,$$

$$\tilde{J}^r_{fm} := \bigcap_{f \in H_m} J^r_{fm}.$$

Intuitively, $\tilde{E}^R_{fm}$ corresponds to exon bodies in which a read will uniquely correspond to pattern m. Similarly, $\tilde{J}^r_{fm}$ corresponds to junctions for which a read will uniquely correspond to pattern m.

(B) For $m = 1, \ldots, M$, let

$$B^r_m = \sum_{i \in \tilde{E}^r_m} e_i - |\tilde{E}^r_m| \cdot (r-1),$$

where e_i is the length of exon i. B^r_m is the number of exon body read positions unique to pattern m.

(C) For each $m = 1, \ldots, M$ and every tuple $z_j = (z_{j1}, \ldots, z_{jk_j})$ in $\tilde{J}^r_m$, where $j \in \{1, \ldots, |\tilde{J}^r_m|\}$ and where k_j can vary from tuple to tuple, let

$$e^{\text{middle}}_j = \begin{cases} \sum_{t=2}^{k_j - 1} e_{z_{jt}} & \text{if} \quad k_j \geq 3 \\ 0 & \text{otherwise,} \end{cases}$$

and set

$$U^r_m = \sum_{j=1}^{|\tilde{J}^r_m|} \min\left(e_{z_{j1}}, e_{z_{jk_j}}, r - e^{\text{middle}}_j - 1\right).$$

U^r_m is the total number of read positions corresponding to every tuple in $\tilde{J}^r_m$.

(D) For $m = 1, \ldots, M$, let $n^r_m = |B^r_m| + |U^r_m|$, and let $N^r = (n^r_1, \ldots, n^r_M)$.

Step 3: *Normalize the effective length of each pattern by the effective length of each isoform included in that pattern to obtain the transcript-to-pattern matrix* $\Theta^{(r)}$.

For $f = 1, \ldots, n$, let $P_f = (P_{f1}, \ldots, P_{fM})$ denote the f^{th} column of the pattern matrix P. Then the f^{th} column of $\Theta^{(r)}$ can be computed as

$$\Theta^{(r)}_f = \frac{P_f \cdot N^r}{\sum_{i \in I_f} e_i - r + 1}.$$

3.8.5 Generating Alternative Splicing Modules and Compatibility Matrices

rMATS-Iso takes a gene annotation file in GTF format and alignment result file in sorted BAM format as input and generates two types of output files: ".IsoExon" and ".IsoMatrix". Only the first bam input file has a corresponding ".IsoExon" output file, which contains the splicing module information of all the input bam files. Each bam input file has a corresponding ".IsoMatrix" output file, which contains the isoform read compatibility information of each individual bam file.

rMATS-Iso builds a splicing graph within each gene. In the splicing graph, exons are represented as nodes and splice junctions between two exons are represented as directed edges between two nodes (directed from the 5' splice site to the 3' splice site). Redundant exons that have same 5' and 3' splice sites are merged into one unique exon. In addition, one virtual start node and one virtual end node are added to the splicing graph for the completeness of the splicing module's definition. We add one virtual edge between the virtual start node and each exon node that does not have any incoming edges. Similarly, virtual outgoing edges are added between terminal exon nodes and the virtual end node. For each junction edge in the splicing graph, rMATS-Iso calculates the corresponding supporting junction read count. By default, only uniquely mapped and properly paired reads are counted during this step. We then remove all junction edges that do not have enough supporting junction reads (the default value is set to 1 read).

The concept of "splicing modules" was previously defined in (Hu et al., 2013). rMATS-Iso implements a method which is similar to that of (Hu et al., 2013) to find valid splicing modules. All the splicing module information is stored in the ".IsoExon" file. For each valid splicing module, all possible paths between nodes are generated to represent all of the possible transcript isoforms. Next, the aligned reads are compared with all of the enumerated isoforms. If a read alignment is consistent with an isoform, meaning that the read may have been generated from that transcript, we say that the read is compatible

with the isoform. A single read could be compatible with multiple isoforms. All of the read compatibility information of each BAM file is stored in the corresponding ".IsoMatrix" file.

CHAPTER 4

Conclusion

We have developed two new statistical methodologies for quantifying and better understanding mRNA isoform variation. First, we introduced PAIRADISE, a method for detecting allele-specific alternative splicing (ASAS) from RNA-seq data. Unlike conventional approaches that detect ASAS events one sample at a time, PAIRADISE uses a statistical model that aggregates ASAS signals across multiple individuals in a population. By treating the two alleles of an individual as paired, and multiple individuals sharing a heterozygous SNP as replicates, PAIRADISE formulates ASAS detection as a statistical problem for identifying differential alternative splicing from RNA-seq data with paired replicates. PAIRADISE outperforms alternative statistical models in simulation studies, and boosts the power of ASAS detection in replicate or population-scale RNA-seq data. Additionally, PAIRADISE ASAS analysis detects the effects of rare variants on alternative splicing.

Next we introduced rMATS-Iso, a multi-isoform generalization of the rMATS statistical and computational framework, for quantifying alternative splicing variation from complex splicing events. rMATS-Iso addresses a commonly encountered difficulty associated with RNA-seq data, namely the difficulty of assigning short RNA-seq reads to one particular mRNA isoform. To address this ambiguity of RNA-seq reads, rMATS-Iso redefines the observed data to be counts of isoform *consistency*, then uses an EM algorithm to estimate isoform probabilities. Our simulation studies reveal that rMATS-Iso provides accurate estimates of the true isoform inclusion levels, and that the performance

of rMATS-Iso improves as more samples and/or RNA-seq reads are added.

Both PAIRADISE and rMATS-Iso are generalizable across diseases and biological systems and will provide a fundamental set of tools for elucidating the transcriptome. We are confident that these methodologies will play a crucial role in highlighting the role of mRNA isoform variation in complex disease processes, and will ultimately help contribute a better understanding of human health.

Bibliography

Alt, F. W., Bothwell, A. L., Knapp, M., Siden, E., Mather, E., Koshland, M., and Baltimore, D. (1980). Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell*, 20(2):293–301.

Ardlie, K. G. and Coauthors (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Auton, A. and Coauthors (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Barbosa-Morais, Nuno L, Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencoew, B. J. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, 338:1587–1593.

Barutcu, A. R., Lajoie, B. R., Fritz, A. J., Mccord, R. P., Nickerson, J. A., Wijnen, A. J. V., Lian, J. B., Stein, J. L., Dekker, J., Stein, G. S., and Imbalzano, A. N. (2016). SMARCA4 regulates gene expression and higher-order chromatin structure in proliferating mammary epithelial cells. *Genome Res.*, 26:1188–1201.

Berk, A. J. (2016). Discovery of RNA splicing and genes in pieces. *Proc. Natl. Acad. Sci.*, 113(4):801–805.

Bønnelykke, K. and Coauthors (2013). Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nat. Genet.*, 45(8):902–906.

Boue, S., Vingron, M., Kriventseva, E., and Koch, I. (2002). Theoretical analysis of alternative splice forms using computational methods. *Bioinformatics*, 18 Suppl 2:S65–S73.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527.

Chang, J.-S., Huggett, J. F., Dheda, K., Kim, L. U., Zumla, A., and Rook, G. a. W. (2006). Myobacterium tuberculosis induces selective up-regulation of TLRs in the mononuclear leukocytes of patients with active pulmonary tuberculosis. *J. Immunol.*, 176(5):3010–3018.

Daley, D., Park, J. E., He, J. Q., Yan, J., Akhabir, L., Stefanowicz, D., Becker, A. B., Chan-Yeung, M., Bossé, Y., Kozyrskyj, A. L., James, A. L., Musk, A. W., Laprise, C., Hegele, R. G., Paré, P. D., and Sandford, A. J. (2012). Associations and interactions of genetic polymorphisms in innate immunity genes with early viral infections and susceptibility to asthma and asthma-related phenotypes. *J. Allergy Clin. Immunol.*, 130(6):1284–1293.

Dempster, A., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.*, 39(1):1–38.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell*, 20(2):313–319.

Farrell, R. E. (2010). RT-PCR: A Science and an Art Form. In *RNA Methodol. A Lab. Guid. Isol. Charact.*, chapter 18, pages 385–448. San Diego.

Fu, X. D. and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.*, 15(10):689–701.

Gracheva, E. O., Cordero-Morales, J. F., González-Carcacía, J. A., Ingolia, N. T., Manno, C., Aranguren, C. I., Weissman, J. S., and Julius, D. (2011). Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. *Nature*, 476(7358):88–92.

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, 40(20):10073–10083.

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y. C., Pugh, T. J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J. K., Chan, S. Y., Li, H. I., McDonald, H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G. B., Jones, S. J., Tai, I. T., and Marra, M. A. (2010). Alternative expression analysis by RNA sequencing. *Nat. Methods*, 7(10):843–847.

Heber, S., Alekseyev, M., Sze, S.-H., Tang, H., and Pevzner, P. A. (2002). Splicing graphs and EST assembly problem. *Bioinformatics*, 18(Suppl 1):S181–S188.

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2017). Singleton Variants Dominate the Genetic Architecture of Human Gene Expression. *bioRxiv*.

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., Monroy, A., Kuan, P. F., Hammond, S. M., Makowski, L., Randell, S. H., Chiang, D. Y., Hayes, D. N., Jones, C., Liu, Y., Prins, J. F., and Liu, J. (2013). DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.*, 41(2).

Hua, Y., Sahashi, K., Rigo, F., Hung, G., Horev, G., Bennett, C. F., and Krainer, A. R. (2011). Peripheral SMN restoration is essential for long-term rescue of a severe spinal muscular atrophy mouse model. *Nature*, 478(7367):123–126.

Jia, C., Hu, Y., Liu, Y., and Li, M. (2015). Mapping splicing quantitative trait loci in RNA-seq. *Cancer Inform.*, 14:45–53.

Kapoor, M., Wang, J. C., Wetherill, L., Le, N., Bertelsen, S., Hinrichs, A. L., Budde, J., Agrawal, A., Bucholz, K., Dick, D., Harari, O., Hesselbrock, V., Kramer, J., Nurnberger, J. I., Rice, J., Saccone, N., Schuckit, M., Tischfield, J., Porjesz, B., Edenberg, H. J., Bierut, L., Foroud, T., and Goate, A. (2013). A meta-analysis of two genome-wide association studies to identify novel loci for maximum number of alcoholic drinks. *Hum. Genet.*, 132(10):1141–1151.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12):1009–1015.

Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: A pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.*, 14(3):153–165.

Kote-Jarai, Z. and Coauthors (2011). Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat. Genet.*, 43(8):785–791.

Královičová, J., Houngninou-Molango, S., Krämer, A., and Vořechovský, I. (2004). Branch site haplotypes that control alternative splicing. *Hum. Mol. Genet.*, 13(24):3189–3202.

Lappalainen, T. and Sammeth, M. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Lee, C. (2003). Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 19(8):999–1008.

Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464.

Lee, C., Roy, M., and Coauthors (2004). Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.*, 5(7).

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

Li, G., Bahn, J. H., Lee, J. H., Peng, G., Chen, Z., Nelson, S. F., and Xiao, X. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.*, 40(13).

Li, W. V., Zhao, A., Zhang, S., and Li, J. J. (2018). MSIQ: Joint modeling of multiple RNA-SEQ samples for accurate isoform quantification. *Ann. Appl. Stat.*, 12(1):510–539.

Lin, S. and Fu, X.-D. (2007). SR Proteins and Related Factors in Alternative Splicing. *Exp. Med. Biol.*, pages 107–122.

Liu, S. and Cheng, C. (2013). Alternative RNA splicing and cancer. *Wiley Interdiscip. Rev. RNA*, 4(5):547–566.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.*, 579(9):1900–1903.

Lorson, C. L., Rindt, H., and Shababi, M. (2010). Spinal muscular atrophy: Mechanisms and therapeutic strategies. *Hum. Mol. Genet.*, 19(R1):111–118.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.*, 47(3):467–485.

Lu, Z.-X., Jiang, P., and Xing, Y. (2012). Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdiscip. Rev. RNA*, 3(4):581–592.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, 45(D1):D896–D901.

Manning, K. S. and Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nat. Rev. Mol. Cell Biol.*, 18(2):102–114.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.-F., Revil, T., and Chabot, B. (2007). hnRNP Proteins and Splicing Control. In *Adv. Exp. Med. Biol.*, volume 623, pages 123–147.

Mayerle, J., den Hoed, C. M., Schurmann, C., Stolk, L., Homuth, G., Peters, M. J., Capelle, L. G., Zimmermann, K., Rivadeneira, F., Gruska, S., Völzke, H., de Vries, A. C., Völker, U., Teumer, A., van Meurs, J. B. J., Steinmetz, I., Nauck, M., Ernst, F., Weiss, F.-U., Hofman, A., Zenker, M., Kroemer, H. K., Prokisch, H., Uitterlinden, A. G., Lerch, M. M., and Kuipers, E. J. (2013). Identification of Genetic Loci Associated With Helicobacter pylori Serologic Status. *J. Am. Med. Assoc.*, 309(18):1912–1920.

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.*, 5(4698).

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463.

Ohmura, G., Tsujikawa, T., Yaguchi, T., Kawamura, N., Mikami, S., Sugiyama, J., Nakamura, K., Kobayashi, A., Iwata, T., Nakano, H., Shimada, T., Hisa, Y., and Kawakami, Y. (2015). Aberrant Myosin 1b Expression Promotes Cell Migration and Lymph Node Metastasis of HNSCC. *Mol. Cancer Res.*, 13(4):721–731.

Ongen, H. and Dermitzakis, E. T. (2015). Alternative Splicing QTLs in European and African Populations. *Am. J. Hum. Genet.*, 97(4):567–575.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.*, 102(1):11–26.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419.

Percifield, R., Murphy, D., and Stoilov, P. (2014). Medium Throughput Analysis of Alternative Splicing by Fluorescently Labeled RT-PCR. In *Spliceosomal Pre-mRNA Splicing. Methods Mol. Biol. (Methods Protoc.)*, pages 299–313. Humana Press, Totowa, NJ.

Pertea, M. and Salzberg, S. L. (2010). Between a chicken and a number of human genes. *Genome Biol*, 11:206.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., Zipursky, S. L., Hughes, H., South, C. E. Y., Angeles, L., and Arbor, A. (2000).

Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *Cell*, 101:671–684.

Sharp, P. (1994). Split genes and RNA splicing. *Cell*, 77:805–815.

Sharp, P. (2005). The discovery of split genes and RNA splicing. *Trends Biochem. Sci.*, 30(6):279–281.

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. USA*, 111(51):E5593–601.

Shen, S., Wang, Y., Wang, C., Wu, Y. N., and Xing, Y. (2016). SURVIV for survival analysis of mRNA isoform variation. *Nat. Commun.*, 7(11548).

Shiina, T., Hosomichi, K., Inoko, H., and Kulski, J. K. (2009). The HLA genomic loci map: Expression, interaction, diversity and disease. *J. Hum. Genet.*, 54(1):15–39.

Skelly, D. a., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.*, 21:1728–1737.

Srebrow, A. and Kornblihtt, A. (2006). The connection between splicing and cancer. *J. Cell Sci.*, 119(13):2635–2641.

Tilgner, H., Grubert, F., Sharon, D., and Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA*, 1640(27):10–12.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, 31(1):46–53.

Tsui, J. C.-C., Lau, C. P.-Y., Cheung, A. C., Wong, K.-C., Huang, L., Tsui, S. K.-W., and Kumta, S. M. (2016). Differential expression of filamin B splice variants in giant cell tumor cells. *Oncol. Rep.*, 36(6):3181–3187.

Unterholzner, L., Keating, S. E., Baran, M., Horan, K. A., Jensen, S. B., Sharma, S., Sirois, C. M., Jin, T., Latz, E., Xiao, T. S., Fitzgerald, K. A., Paludan, S. R., and Bowie, A. G. (2010). IFI16 is an innate immune sensor for intracellular DNA. *Nat. Immunol.*, 11(11):997–1004.

Van De Geijn, B., Mcvicker, G., Gilad, Y., and Pritchard, J. K. (2015). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11):1061–1063.

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*, 5(e11752):1–30.

Veeranki, S. and Choubey, D. (2012). Interferon-inducible p200-family protein IFI16, an innate immune sensor for cytosolic and nuclear double-stranded DNA: Regulation of subcellular localization. *Mol. Immunol.*, 49(4):567–571.

Vélez, J. I., Lopera, F., Sepulveda-Falla, D., Patel, H. R., Johar, A. S., Chuah, A., Tobón, C., Rivera, D., Villegas, A., Cai, Y., Peng, K., Arkell, R., Castellanos, F. X., Andrews, S. J., Silva Lara, M. F., Creagh, P. K., Easteal, S., De Leon, J., Wong, M. L., Licinio, J., Mastronardi, C. A., and Arcos-Burgos, M. (2016). APOE∗E2 allele delays age of onset in PSEN1 E280A Alzheimer's disease. *Mol. Psychiatry*, 21(7):916–924.

Venables, J. P., Tazi, J., and Juge, F. (2012). Regulated functional alternative splicing in Drosophila. *Nucleic Acids Res.*, 40(1):1–10.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6.

Wang, G. S. and Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, 8(10):749–761.

Wang, Z. and Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(617):802–813.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, 18(5):1368–1373.

Wu, J., Akerman, M., Sun, S., McCombie, W. R., Krainer, A. R., and Zhang, M. Q. (2011). Splice Trap: A method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, 27(21):3010–3016.

Xing, Y., Resch, A., and Lee, C. (2004). The multiassembly problems: Reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.*, 14(3):426–441.

Xing, Y., Yu, T., Wu, Y. N., Roy, M., Kim, J., and Lee, C. (2006). An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.*, 34(10):3150–3160.

Yang, Q., Hu, Y., Li, J., and Zhang, X. (2017). ulfasQTL: An ultra-fast method of composite splicing QTL analysis. *BMC Genomics*, 18(Suppl 1):23–26.

Zhao, K., Lu, Z.-x., Park, J., Zhou, Q., and Xing, Y. (2013). GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.*, 14(7):R74.

Zheng, P. P., Sieuwerts, A. M., Luider, T. M., Van Der Weiden, M., Sillevis-Smitt, P. A., and Kros, J. M. (2004). Differential expression of splicing variants of the human caldesmon gene (CALD1) in glioma neovascularization versus normal brain microvasculature. *Am. J. Pathol.*, 164(6):2217–2228.