2019

Methods For Robust Quantification Of Rna Alternative Splicing In Heterogeneous Rna-Seq Datasets

Scott Simon Norton
University of Pennsylvania, scott.s.norton@outlook.com

Methods For Robust Quantification Of Rna Alternative Splicing In Heterogeneous Rna-Seq Datasets

Abstract

RNA alternative splicing is primarily responsible for transcriptome diversity and is relevant to human development and disease. However, current approaches to splicing quantication make simplifying assumptions which are violated when RNA sequencing data are heterogeneous. Influences from genetic and environmental background contribute to variability within a group of samples purported to represent the same biological condition. This work describes three methods which account for data heterogeneity when detecting differential RNA splicing between sample groups. First, a robust model is implemented for outlier detection within a group of purported replicates. Next, large RNA-seq datasets with high within-group variability are addressed with a statistical approach which retains power to detect changing splice junctions without sacricing specicity. Finally, applying these tools to call sQTLs in GTEx tissues has identified splicing variations associated with risk loci for cardiovascular disease and anomalous skeletal development. Each of these methods correctly handles the properties of heterogeneous RNA-seq data to improve precision and reduce false discovery rate.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Yoseph Barash

Second Advisor

Hongzhe Lee

Keywords

heterogeneous, methods, outlier, qtl, rna, splicing

Subject Categories

Bioinformatics

METHODS FOR ROBUST QUANTIFICATION OF RNA ALTERNATIVE SPLICING IN HETEROGENEOUS RNA-SEQ DATASETS

Scott Norton

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Yoseph Barash, Associate Professor of Genetics

Graduate Group Chairperson

Benjamin Voight, Associate Professor of Genetics

Dissertation Committee

Hongzhe Li, Professor of Biostatistics (chair)

Junhyong Kim, Professor of Biology

Sarah Tishkoff, David and Lyn Silfen University Professor Departments of Genetics and Biology

Hagen Tilgner, Assistant Professor of Neuroscience (Weill-Cornell School of Medicine)

METHODS FOR ROBUST QUANTIFICATION OF RNA ALTERNATIVE SPLICING IN HETEROGENEOUS RNA-SEQ DATASETS

ABSTRACT

METHODS FOR ROBUST QUANTIFICATION OF RNA ALTERNATIVE SPLICING IN HETEROGENEOUS RNA-SEQ DATASETS

Scott Norton

Yoseph Barash

RNA alternative splicing is primarily responsible for transcriptome diversity and is relevant to human development and disease. However, current approaches to splicing quantification make simplifying assumptions which are violated when RNA sequencing data are heterogeneous. Influences from genetic and environmental background contribute to variability within a group of samples purported to represent the same biological condition. This work describes three methods which account for data heterogeneity when detecting differential RNA splicing between sample groups. First, a robust model is implemented for outlier detection within a group of purported replicates. Next, large RNA-seq datasets with high within-group variability are addressed with a statistical approach which retains power to detect changing splice junctions without sacrificing specificity. Finally, applying these tools to call sQTLs in GTEx tissues has identified splicing variations associated with risk loci for cardiovascular disease and anomalous skeletal development. Each of these methods correctly handles the properties of heterogeneous RNA-seq data to improve precision and reduce false discovery rate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

This section reviews the fundamentals of RNA splicing biology, the tools and technologies that exist to measure RNA splicing in tissues and changes between tissues, and the inherent challenges that are addressed by the work supporting the later chapters.

1.1. Splicing biology

1.1.1. Eukaryotic mRNA transcripts are processed by the spliceosome

Gene expression is the essential mechanism by which a cell synthesizes the proteins it needs to function. These proteins are encoded as genes in the cell's DNA, which are transcribed into messenger RNA (mRNA) by an RNA polymerase (RNA Polymerase II in mammals) and translated into proteins by ribosomes. Cells have additional layers of regulation on top of this, which allow it to control the abundance of each protein to adapt to its immediate needs. In complex organisms, the patterns of regulation vary between cell types and result in different genes being expressed at different levels, which results in differences in cellular function.

Eukaryotes in particular have a complex gene structure, with its sequence consisting of "exons" and "introns". In general, mature mRNAs consist of only exons, which contain the linear sequence that directly encodes for the protein product. As such, the introns must be removed from the nascent mRNA transcript (pre-mRNA), and the process by which this occurs is called "splicing". During processing of the pre-mRNA, which commonly occurs co-transcriptionally (Herzel et al., 2017), a complex of RNA and proteins called the "spliceosome" assembles on the nascent RNA transcript at specific points on the intron and flanking exons. The components of this machinery and the process it mediates are summarized in a review by Matlin et al. (2005). Briefly, the spliceosome contains five catalytic small nuclear ribonucleoproteins (snRNPs) termed U1, U2, U4, U5, and U6 in complex with large proteins. The assembly of the basal spliceosome is guided by consensus motifs at the 5' (recognized by U1) and 3' (recognized by U2 auxiliary factor (U2AF)) splice sites as well as the branch point adenine residue (located in the intron some 30 nt

upstream from the 3' splice site, bound by splicing factor 1 (SF1)) and an 18+-residue polypyrimidine tract at the 3' end of the intron (also bound by U2AF). Additional splicing regulatory proteins (SRps), such as those in the CELF and RBFOX families (Sun et al., 2012; Chen and Manley, 2009; Gazzara et al., 2017), can bind to the mRNA, either in the intron or in either flanking exon, to enhance or inhibit spliceosomal assembly and activity. Once formed, the basal spliceosome recruits additional RNAs and proteins, including U2 snRNP itself, to bring the 5' and 3' splice sites into proximity with one another. This proximity mediates the ATP-driven excision of the intron in two major catalytic steps: nucleophilic attack of the branch point adenine on the 5' splice site, cleaving the 5' end and creating the lariat intermediate; and nucleophilic attack of the 3' end of the upstream exon on the 3' splice site (which also has a conserved sequence motif), freeing the intron from the rest of the transcript and joining the exons together (Herzel et al., 2017). This process is repeated iteratively for each intron to be removed.

Splicing is one of three major forms of post-transcriptional or co-transcriptional processing of mRNAs in eukaryotes (Beyer and Osheim, 1988; Tilgner et al., 2012; Ameur et al., 2011). It is also essential for the 5' end of the transcript to be "capped" with a 5-methylguanine residue. This 5' cap is recognized by the eukaryotic translation initiation complex (specifically, by a protein called EIF4F in humans), and is therefore required for the protein to be translated. Additionally, the 3' end of the transcript is adorned with a polyadenosine (poly-A) tail. This tail is associated with transcript stability in both eukaryotes and prokaryotes. In eukaryotes, it serves as the binding substrate for poly-A binding proteins (PABP).

mRNAs are not the only class of transcripts subject to this processing. Recently, several long noncoding RNAs (lncRNAs) have been classified. These RNA Polymerase II products are capped, spliced, and polyadenylated in the same manner as mRNAs. However, as the name suggests, the mature transcripts do not code for any protein product, nor are they exported to the cytoplasm where the ribosomes reside. Instead, the mature lncRNA transcripts are localized to nuclear subcompartments where they perform their roles in

chromatin and transcriptional regulation (Cao, 2014).

1.1.2. Alternative splicing contributes to proteomic and functional diversity

Differential binding of the aforementioned SRps affect the degree to which a given splice site is used. The consequence of this is the differential inclusion and exclusion of exons and exonic segments in the mature transcript, a phenomenon termed "alternative splicing". Figure 1 depicts a toy example of a gene with two transcript isoforms, one in which the red exon (circled) is included, and one where it is skipped. The functional consequence is the presence or absence of the red domain on the protein product. If, for example, this red domain is a ligand binding site, the choice of whether to include or skip the coding exon will affect the response of the protein product to that ligand, which impacts how the cell as a whole responds to stimulus. In some cases, an alternative transcript may not yield a viable protein product at all. This usually happens because the alternative transcript has a premature termination codon, introduced either in the alternatively included exon or intron, or as a result of a frame shift. These transcripts are normally targeted for nonsense-mediated decay (NMD) (Lykke-Andersen and Jensen, 2015). Additionally, alternative splicing can result in differential inclusion of regulatory domains on the mRNA itself, such as RNA-binding protein (RBP) and micro RNA (miRNA) binding sites, whether due to sequence alone or a change in the mRNA's secondary structure that affects binding of the aforementioned. While this change would not affect the translated protein product, it does impact where the protein is expressed and at what level.

Alternative splicing is particularly abundant in higher mammals, with an estimated 90% of multi-exon genes in humans, both coding and non-coding (Deveson et al., 2017), showing evidence of multiple transcript isoforms (Wang and Cooper, 2007). Many of these alternative isoforms are cell-type specific, and the regulation of alternative isoform expression is tightly regulated during various stages of organism development. Moreover, dysregulation of alternative splicing has been observed in several diseases, including familial and sporadic forms of frontotemporal dementia (FTD) and Alzheimer's disease, and various cancers (Forman et al., 2006; Bai et al., 2013; Arnold, 2013; Colombo et al., 2014; Zhang et al., 2013;

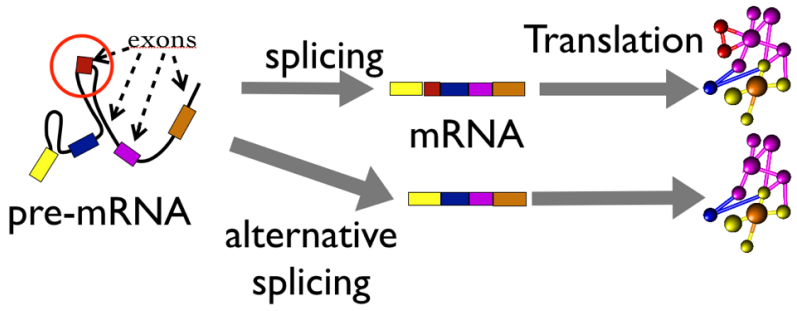


Figure 1: Cartoon illustrating mRNA alternative splicing. The red exon (circled) can either be spliced in to yield the top isoform or spliced out to yield the bottom isoform. Each isoform is translated into a different protein, one with the red domain, and one without.

Colak et al., 2013; Dvinge and Bradley, 2015). In order to better understand the precise mechanisms driving alternative splicing differentiation and disease, it is necessary to accurately quantify splicing in tissue samples and measure splicing changes between biological or experimental conditions.

1.2. RNA splicing quantification

RNA splicing quantification is a rapidly evolving field accelerated by the advent of RNA-seq in 2008 (Nagalakshmi et al., 2008; Mortazavi et al., 2008). This chapter summarizes some of the major technological developments in measuring RNA levels in tissue extracts, and the evolution of methods for quantifying alternative splicing.

1.2.1. Important technological developments for measurement of RNA abundance

Prior to the advent of RNA-seq, two biochemical techniques predominated for measuring RNA alternative splicing. The higher-throughput of these is the microarray, in which fluorescently-labeled DNA oligonucleotides (oligos) anneal to complementary RNAs. Alternative transcript abundance for a given gene can be measured by designing two or more probes with different fluorescent tags, each complementary to a sequence unique to one of the transcript isoforms. After washing away unbound probes, the abundance of each isoform can be interpreted by imaging from the relative intensity of each fluorescent frequency. This technique can be performed for multiple genes or multiple samples in parallel by loading each well on a specially-designed wellplate with either different input samples or different hybridization probes (Schena et al., 1996). The signal in each well can be made easier to

detect by reverse-transcribing (RT) the mRNA transcript to complementary DNA (cDNA) followed by a polymerase chain reaction (PCR) to amplify the cDNA product (RT-PCR).

The microarray technique has some technical limitations. First, the throughput of the technique is restricted to the number of wells on the wellplate used for the analysis. Additionally, because the hybridization probes must be designed before the microarray analysis is performed, this technique cannot be used for *de novo* discovery of splice isoforms. Lastly, the fluorescent intensity can be difficult to interpret in a quantitative fashion - the signal can either be too low for the sensor to distinguish from background, potentially resulting in false negative detection, or so high that it saturates the sensor, frustrating differential analysis.

Carefully-performed RT-PCR is a popular low-throughput technique for local splicing quantification and is considered to be the gold standard among biochemical assays. RT-PCR primers are designed to flank an alternative splicing event, which is defined in terms of the alternatively-included exon, exonic fragment, or retained intron. Classically, splicing events include cassette exons, alternative 5' donor or 3' acceptor sites, mutually exclusive exons, and intron retentions, as depicted in Figure 2. RT-PCR is assessed by measuring the relative intensity of two or more bands on a Northern blot, which correspond to the known fragment sizes for each splice form. In the case of a cassette event, amplified fragments including the alternative exon will be larger and migrate less distance on the gel than fragments skipping the exon. The relative intensities of the two bands are summarized as a single quantity PSI (Ψ), or "percent spliced in", for the alternative fragment. Splicing can be directly compared between two experimental conditions as the change in Ψ (Delta PSI, $\Delta\Psi$). This approach is advantageous in that it accurately captures local splicing decisions in a quantitative fashion. However, RT-PCR is labor-intensive and low-throughput, and shares microarrays' restriction to known splicing variations.

1.2.2. RNA-seq-based methods facilitate high-throughput quantification and de-novo discovery of splicing variations

The advent of high throughput sequencing (Next-Generation Sequencing, or NGS) has revolutionized quantitative genomic and transcriptomic analysis. The predominant variation of this is the sequencing-by-synthesis technique employed on the Illumina platform. This technology directly reads the nucleotide sequences of each DNA molecule in the input sample. These "reads" can be mapped back to a reference genome or transcriptome using tools such as BowTie (Langmead et al., 2009), STAR (Dobin et al., 2013), or HISAT2 (Kim et al., 2015).

The technique itself, which is described in detail by Solexa Ltd. (Bennett, 2004; Slatko et al., 2018), requires special preparation of a DNA or cDNA library to produce DNA molecules to average 300 nucleotides in size with specialized adapters annealed to the 3' end of each molecule. The library is loaded onto a flow cell, which contains billions of clusters of oligonucleotides complementary to these adapters. The DNA is amplified on these clusters by the sequencer using the oligos as PCR primers. The final round of amplifcation is modified to add one fluorescently-labeled nucleotide at a time, up to a predetermined, protocol-dependent length. When incorporated, the sequencer reads the fluorescent signal simultaneously for all clusters on the flow cell, and interprets the sequence of images as the original DNA sequence at each cluster.

One key advantage of this technology is that each molecule in the input sample can be mapped directly to at most one read in the output. When multiple molecules originate from the same genomic region, the relative representation of that region can be interpreted from the number of reads mapping back to it. This is a digital count which, unlike fluorescent signal detection, has no upper bound, thus enabling the accurate quantification of both lowly-expressed and highly-expressed loci. Additionally, splice-aware mappers such as STAR and HISAT2 can map RNA reads to the genome instead of a transcriptome, allowing for *de-novo* discovery of splice junctions, virtually removing the requirement for a prior

transcript model.

A second benefit of high-throughput sequencing is that it can be multiplexed, allowing for the simultaneous evaluation of multiple samples. Multiplexing is achieved by incorporating unique DNA "barcodes" into the adapters annealed to the amplified DNA in each sample. During the image processing step, the sequencer separates the reads using the barcode as a key for which sample the reads originally came from.

High-throughput sequencing, including RNA-seq, is not without its limitations. First, because library preparation is a multistep process with the potential for loss at each step, some molecules from the original extract may not appear in the sequencer output. This is evidenced by an apparent read, fragment, or transcript count of 0 for some genomic regions (zero-inflation) (Rashid et al., 2011). Since the same can arise due to the molecule being absent in the original tissue sample (i.e. the gene or transcript is not expressed), zero counts must be interpreted with caution. Each step of the process additionally incorporates its own systematic biases, including a positional bias favoring the 3' end of transcripts and a bias towards sequences with high GC content. Finally, NGS technologies are designed to produce only short sequencing reads, typically between 100-150 bp in length. While this is sufficient to estimate gene expression, it makes quantification of whole transcript isoforms challenging. Sequencing the cDNA from both ends (paired-end sequencing) improves mapping accuracy, but the gains in performance for transcriptome quantification and assembly are modest (Song and Florea, 2013). These technical shortcomings must be addressed when processing RNA-seq experiments for expression or splicing quantification.

1.2.3. Splicing quantification from RNA-seq relies on the underlying model

Isoform quantifiers

Several tools and models have been proposed to address the unique challenges posed by using RNA-seq for transcript and splicing quantification. Chief among these are methods such as SALMON (Patro et al., 2017) and RSEM (Li and Dewey, 2011) which attempt to estimate relative abundances of whole transcript isoforms. In brief, these methods assume

a transcript model and attempt to assign RNA-seq reads to transcripts. This is more easily done in simple cases where two alternative transcripts share all but a small number of exons or parts of exons. However, many genes have three or more transcripts, some of which mutually share exons and splice junctions. The *CYP11B1* locus, for example, has five annotated transcript isoforms with substantial local overlap in sequence. In particular, the three longest isoforms are nearly impossible to distinguish from the two shortest isoforms given only reads starting within exon 9 (see Figure 2). Therefore, these algorithms must determine what fraction of reads mapping to shared exons come from which transcript isoforms.

Both RSEM and SALMON construct a joint parametric model of transcript abundance and read assignments to transcripts, and use expectation maximization (EM) to optimize this model. In EM, observations (reads) are first assigned their expected class labels (transcripts) based on the current state of the model parameters. Next, the parameters are updated to values which maximize the total likelihood of the joint model given the current label assignments. In the simplest case, the total number of reads assigned to a transcript is scaled (normalized) by the number of mappable positions on that transcript. The maximum likelihood estimate is then the fraction of normalized reads for each transcript (transcript-level expression) relative to the sum of normalized reads for all transcripts at that locus (total gene expression). The two steps of label assignment (expectation) and parameter update (maximization) are repeated until the joint likelihood of the model converges. The final optimized parameters are then used to provide transcript-level expression estimates as transcripts per million RNA molecules (TPM).

In general, transcript levels are highly informative of biological activity in cells. However, the task of quantifying isoform expression transcriptome-wide is complicated by the problem of *de-novo* transcript assembly from short RNA-seq reads. The main barrier to accurate transcriptome assembly from NGS data is determining whether two distant exons originate from the same mRNA transcript. This problem can be addressed by long-read sequencing

technologies such as PacBio and Oxford Nanopore. However, the current cost of running these platforms is still considerably high for decent coverage. Moreover, these technologies are more useful for detecting isoforms in the task of transcriptome definition than for isoform quantification.

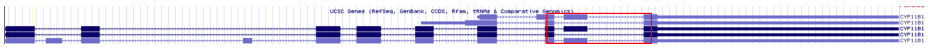


Figure 2: *CYP11B1* as an example of a complex gene locus. Image lifted from the UCSC hg19 Genome Browser, "UCSC Genes" track, zoomed to cover the *CYP11B1* locus on chromosome 8q24.3. The region in the red box highlights parts of exons 9-11. Reads originating from fragments mapping here can be used to distinguish between the three long and two short isoforms, but cannot be assigned to any one transcript. In particular, it is not possible to distinguish between isoforms 3 and 5 from these reads alone, as they do not capture the inclusion of exon 4.

Event quantifiers

Several methods for RNA quantification avoid the problem of transcript assembly altogether by framing RNA splicing in a more local context. The premise is that alternative splicing results in the alternative inclusion or exclusion of exons. In many cases, this affects the differential presence of functional or structural domains in the translated protein product or on the mRNA itself. The fraction of the RNA sample that includes the alternative exon, relative to the number of reads which exclude the exon, is often reported as the "percent spliced in" (Ψ) for that exon.

The alternative exclusion of whole exons is called "exon skipping" or a "cassette exon event", and is one of several classically-defined "splicing event" types in the literature (Wang et al., 2015). Other event types include alternative 5' splice sites, alternative 3' splice sites, mutually exclusive exons, and intron retention. These are depicted in Figure 4. These events are much easier to define from NGS data because they rely only on reads mapping uniquely to each variant of the event. Methods such as rMATS (Shen et al., 2014) and MISO (Katz et al., 2010) were developed to quantify Ψ for splicing events transcriptome-wide.

Some methods, such as SUPPA (Entizne et al., 2016; Trincado et al., 2018), take a hybrid approach wherein they use transcript-level quantifications to inform event-level quantifica-

tions. This approach takes normalized transcript counts (number of copies of each transcript per million of reads sequenced, or TPM) as estimated by SALMON or RSEM and, for each event, combines counts from all transcripts containing the same variation of the event to estimate Ψ. For example, in an exon skipping event, Ψ is estimated as the ratio between the sum of all TPM from transcripts containing the cassette exon and the sum of TPM from all transcripts annotated at that gene locus. While this resolves the ambiguity problem faced by SALMON and RSEM alone, it does not facilitate *de-novo* junction discovery.

While event quantifiers are fast and accurate, the definition of these events is limited in its ability to describe the full complexity of mammalian transcriptomes. Indeed, as much as 30% of observed transcriptome variations cannot be correctly quantified under this framework because they do not fit the underlying events model (Vaquero-Garcia et al., 2016). Notably, both rMATS and MISO rely on a user-provided transcriptome annotation from which splicing events are constructed. As a consequence, they do not detect novel splicing events (*de-novo* detection), which limits their ability to discover new splicing biology. The ability to detect *de-novo* splicing variations is important for studying low-abundance or developmentally-transient cell types as well as diseases affecting splicing regulation.

Cluster quantifiers

A third, more recent class of methods approaches the problem as one of junction or intron clustering. This approach groups alternative transcripts by the set of junctions or introns that share splice sites between them (see Figure 3). Methods following this approach include Whippet (Sterne-Weiler et al., 2018) and LeafCutter (Li et al., 2018). This represents a generalization on the classical event model in that splice junctions sharing at least one flanking constitutive exon between them can be explained as a single splicing event covering differential usage of subsets of junctions. This model reflects the observation that RNA splicing occurs via the "stepwise removal of introns from nascent pre-mRNA". Because this model examines splicing locally and is not constrained by classical event definitions, it can accurately quanitfy the remaining 30% of transcriptome complexity that is omitted by rMATS and MISO. Both Whippet and LeafCutter additionally detect and quantify

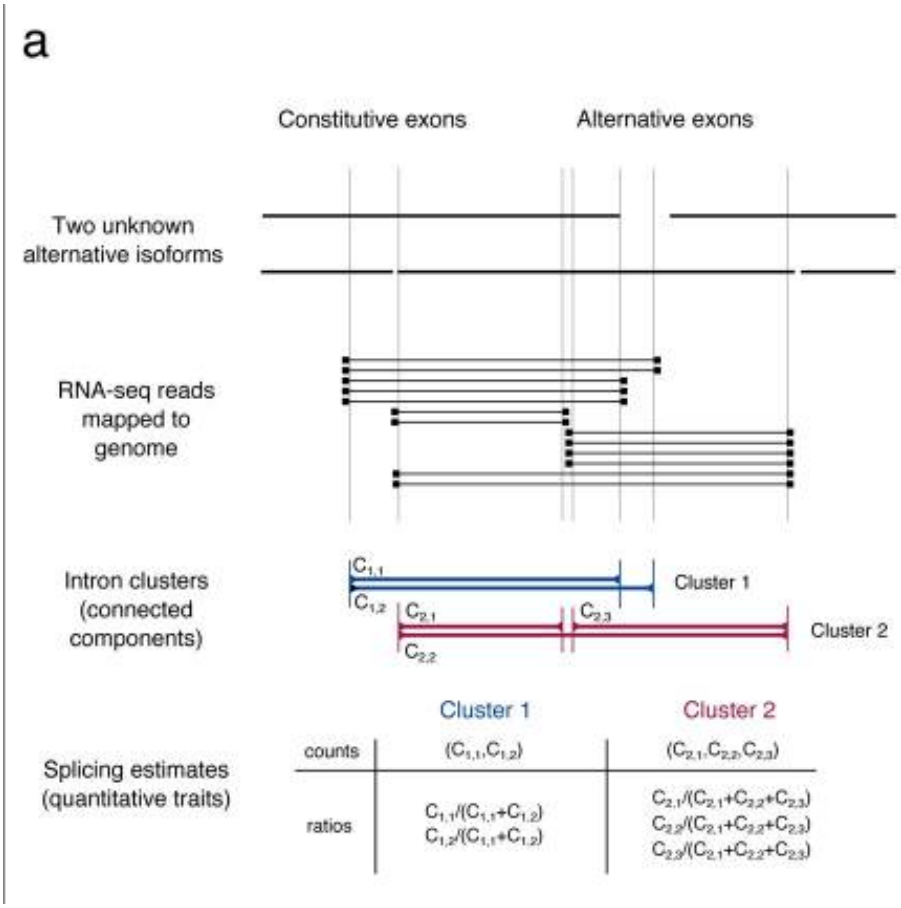


Figure 3: Figure 1a from Li et al. (2018). Differentially-excised introns are identified from spliced reads as flagged by the RNA-seq mapper. A cluster is defined from all excised introns which share at least one splice site.

unannotated exons and splice junctions from the input reads. Importantly, LeafCutter was designed for the discovery of splicing quantitative trait loci (sQTLs), allelic variants which associate with changes in splicing. However, neither approach detects unannotated or *de novo* retained introns, the likes of which are present in both normal (Schmitz et al., 2017) and disease (Dvinge and Bradley, 2015) tissue contexts.

Local splicing variation quantifier

The fourth class of methods explains alternative splicing in terms of "LSVs", which are conceptually similar to the aforementioned cluster-based models. The LSV definition can be interpreted by visualizing a gene model as a directed (5' to 3') graph, where the edges are splice junctions and the vertices are contiguous exonic fragments between adjacent

splice junctions. An LSV is defined as a split in this graph representing an individual splicing decision at a single splice site (Figure 5A). This framework naturally captures all the variations explained by classical binary splicing events (Figure 5B) but have the flexibility to describe non-classical and complex (3 or more splice junctions, Figure 5C) events that are misclassified or excluded by classical models. Additionally, alternative junction inclusion levels can be quantified from RNA-seq in a manner similar to traditional Ψ quantification.

MAJIQ (Model of Alternative Junction Inclusion Quantification) was developed around the LSV framework (Vaquero-Garcia et al., 2016; Norton et al., 2018). MAJIQ works by first building a splice graph model for each gene in the supplied reference annotation. This model captures all exons and splice junctions in the transcriptome annotation, and identifies points where two transcripts converge or diverge as LSVs. Optionally, MAJIQ supplements this model with evidence from the supplied RNA-seq alignments, adding new splice junctions and LSVs where there are enough reads to support them. Spliced reads flagged by the aligner are assigned uniquely to junctions in the splice graph. This process implements quality control measures such as probabilistic stack removal to remove PCR duplicates and parametric bootstrapping to capture the per-experiment, per-junction variance in mapped read levels. Next, a Bayesian model is applied to compute a posterior distribution of Ψ for each LSV junction using the bootstrapped read counts on top of a Jeffrey's prior. Evidence from replicate experiments is accumulated in this step to provide a more confident estimate of Ψ. An additional Bayesian prior is applied when estimating differential splicing between conditions. Finally, MAJIQ comes packaged with a visualization suite called VOILA which generates publication-ready illustrations of the splice graph model, read count distributions, and Ψ and $\Delta\Psi$ quantifications at each LSV (Figure 5D). Additionally, VOILA generates a human- and machine-readadble TSV file summarizing splicing quantifications at all LSVs.

Of the aforementioned methods that quantify alternative splicing in groups of RNA-seq experiments, all of them assume that these groups are *bona fide* replicates of an underlying biological condition. However, this assumption is not guaranteed to hold. Indeed, sample

groups can be heterogeneous for a number of known and unknown reasons. For instance, an experiment involving inbred mice can be confounded by differential food consumption between individual mice, mislabeling of tissue samples, or a shift in environmental conditions between batches of samples. Human population studies pose additional challenges for splicing quantification, as it is unethical to control for genetic and environmental variation in the study group the way one would with mice. One can attempt to compensate for this by expanding the sample size, but the underlying heterogeneity must still be accounted for properly. Genetic and environmental variants can covary with splicing, making it difficult to conclude whether an observed splicing variation is the result of the study condition or if it is confounded with an underlying trait.

The major contributions of this body of work are as follows. First, I developed a method for detecting splicing outliers in a group of purported replicate experiments. I extended that method to correct for outlier replicates at the Ψ level. This functionality, described as MAJIQout, was integrated into MAJIQ and is available starting in version 1.1. As part of developing MAJIQout we also developed an extensive set of evaluation criteria to assess algorithms for differential splicing quantification from RNASeq, something the community lacked. We then applied these tests to state of the art algorithms to assess their performance. Next, I addressed the impact of data heterogeneity and dataset size on splicing observations, and developed a framework for detecting differences in alternative splicing between large heterogeneous sample groups using robust rank-based statistics. We termed this framework MAJIQ-HET, and it is included as part of MAJIQ 2.0. Finally, I designed a pipeline that uses MAJIQ Ψ quantifications to call splicing quantitative trait loci (sQTLs). I applied this pipeline in collaboration with researchers at the University of Pennsylvania, the Children's Hospital of Philadelphia, and Erasmus University Medical Center to discover and validate splicing variations that associate with disease risk loci.

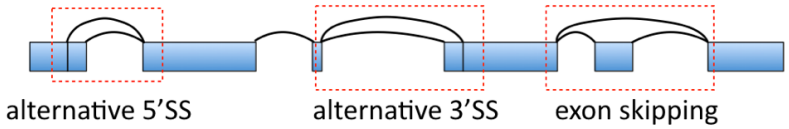


Figure 4: Schematic depicting some of the classical binary events that are quantified by event-based methods such as rMATS and MISO. From left to right: alternative 5' (donor) splice site, alternative 3' (acceptor) splice site, and exon skipping (cassette exon). Not depicted but still relevant are the cases of mutually exclusive exons where the decision is between two options for a cassette exon, and intron retention where the intervening intron is not spliced out.

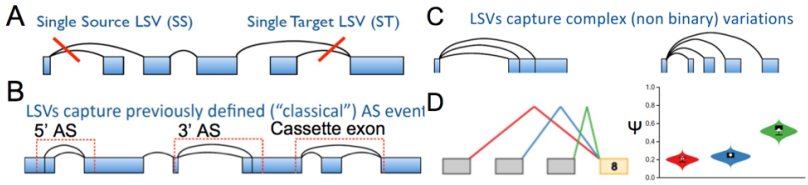


Figure 5: Overview of the LSV model. A: LSVs are visualized as splits in a splice graph. A single-source LSV shares the 5' node, and a single-target LSV shares the 3' node. B: Classically-defined binary splicing events are captured by two symmetric LSVs, one at the 5' node and one at the 3' node. C: The LSV model also explains more complex splicing variations, where three or more junctions share a common splice site. D: VOILA generates publication-ready visualizations of gene graphs, LSV structures, and $\Psi/\Delta\Psi$ quantifications.

CHAPTER 2 : Outlier detection and methods evaluations

This chapter details a new approach for outlier detection in RNA alternative splicing quantification, as well as best practices for evaluating and comparing method performance and accuracy. The original work is published in Norton et al. (2018).

2.1. Introduction

Data from present-day RNA-seq experiments face challenges relating to sequencing depth, transcriptome coverage, and biological and technical variability (Alamancos et al., 2014). A popular workaround for this is to sequence multiple replicates of the same experiment. Replicates can be either "biological" i.e. the data come from different organisms or tissue samples from the same condition group, or "technical" i.e. the same library is sequenced more than once. Generally, technical replicates grant consistency in observations and can improve power to detect true splicing changes. However, a single RNA-seq library can fail to capture low-abundance transcripts simply by chance. Biological replicate designs overcome this by constructing multiple libraryies, giving these transcripts a greater likelihood of representation in at least one experiment.

In each case, the information from replicate data is combined to detect biological signals. Many RNA splicing and transcript abundance quantifiers are designed to handle multiple replicates. rMATS (Shen et al., 2014), for example, implements a hierarchical logit-normal distribution to explain per-replicate Ψ in terms of a group mean Ψ and variance. MAJIQ (Vaquero-Garcia et al., 2016), meanwhile, models Ψ using a Bayesian framework where the read counts bootstrapped from each additional replicate updates a beta posterior model for the underlying group Ψ. Each of these approaches has its advantages and disadvantages for Ψ quantification, but both assume that the input samples are true replicates sharing an underlying distribution of event or junction Ψ. When the data violate that assumption, neither model is guaranteed to accurately represent the true underlying Ψ.

What does it mean for a group of experiments to violate this assumption of shared under-

ling Ψ? This dissertation discusses two such scenarios. The first, covered by this chapter, explores the case where one or more experiments is an outlier replicate. The second, explained in the next chapter, deals with larger sample groups that are true representatives of the underlying biological condition but exhibit a great deal of heterogeneity.

2.1.1. What is an outlier in the context of RNA alternative splicing?

Genomic datasets are grainy snapshots of biological samples, so some variance is to be expected. In addition to the irreducible portion of variance inherent in the technique, there are also genuine biological reasons why per-sample measurements differ. For bulk tissue RNA-seq, these can include differences in the cell-type composition of the tissue sample, fluctuations in gene expression within the sample, and specimen-specific environmental factors. In light of these known sources of variation, some disagreement between per-sample quantifications is tolerable. However, samples which deviate strongly from the group consensus - "outliers" - are not.

Statistical outliers with respect to an event are samples in an experiment which significantly skew the group estimate. Specifically, in the problem of splicing quantification from RNA-seq experiments, the distribution of mapped reads in an outlier deviates substantially from those of the remaining experiments in the group. This can manifest in the outlier reporting different splice junction inclusion levels than its fellow experiments, either relative to other junctions describing the same splicing event or in terms of total number of reads mapping to that event (depth of coverage). The impact of a read count outlier is of particular importance - if an experiment reports much fewer reads than the others in the group, it can reduce the apparent significance of changes in splicing ($\Delta\Psi$) between groups even if the outlier's point estimate of Ψ is in agreement with the group.

It is not atypical to observe some background divergence within a group of replicate experiments, even events at which one replicate disagrees quite strongly with the remainder. What makes a replicate a *bona fide* outlier is the prevalence of such disagreements across the event space. In order to determine this, one should test the suspected outlier across the en-

tire transcriptome, accumulating evidence of disagreement with the remaining experiments in the study group. Therefore, a robust outlier detection algorithm would define a metric of divergence between each replicate's quantification of an event and a group representative, and evaluate this metric on all events to determine whether one replicate is a serial offender in this regard.

2.1.2. How is outlier detection typically performed in the literature, and what are the possible shortcomings therein?

Statistical outliers are of significant concern when dealing with biological data, which is often noisy in and of itself. However, there is little to no discussion in the literature on how to handle it. Instead, researchers are left to use their own heuristics to determine whether a sample might be an outlier. Often, these heuristics carry hidden biases which impact anaylsis. A recent work by Conesa et al. (2016) suggests using PCA to query whether samples of the same condition cluster together, but admits that "no clear standard exists for biological replicates" as pertains to measuring within-group consistency. The work described in this chapter addresses this gap in the literature, and was originally published as Norton et al. (2018).

2.2. Algorithm

2.2.1. If we had some weights, how would we use them in MAJIQ?

Before we describe our approach to outlier detection, we first explain how such an approach should be applied in practice. Let us examine the original quantification model on a collection $\mathcal{T}$ of N experiments representing a biological condition. A splice graph is generated for $\mathcal{T}$ by the MAJIQ builder, and the junction-spanning RNA-seq reads from each experiment $t \in \mathcal{T}$ are assigned to LSV junctions. To control for sampling variance in read alignments across positions in the transcript, we bootstrap read counts for each junctions based on the number of reads starting at each nonzero position that is assigned to the LSV. Let us call these read counts $\{R_{i,j,t}\}$, where i is the index of the LSV, j is the junction index within LSV i, and t is the current experiment. Suppose that LSV i has J splice junctions. Prior to incorporating the $\{R_{i,j,t}\}$, the random variable $\Psi_{i,j}$ is assumed to have a Beta($\frac{1}{J}, \frac{J-1}{J}$)

distribution for each j. (The joint distribution of junction Ψ is a Dirichlet($\underbrace{\frac{1}{J}, \ldots, \frac{1}{J}}_{n \text{ times}}$). However, we consider the marginal distribution for each junction separately to make the model tractable.) Each experiment t contributes $R_{i,j,t}$ reads of evidence supporting junction j, and a combined $\sum_{j' \neq j} R_{i,j',t}$ reads supporting the other junctions. Thus the posterior distribution of $\Psi_{i,j} \mid t$ is

$$\Psi_{i,j} \mid t \sim \text{Beta}\left(R_{i,j,t} + \frac{1}{J}, \sum_{j' \neq j} R_{i,j',t} + \frac{J-1}{J}\right). \tag{2.1}$$

This posterior can be updated with the read counts from the remaining replicates in $\mathcal{T}$, so that the posterior distribution becomes

$$\Psi_{i,j} \mid \mathcal{T} \sim \text{Beta}\left(\sum_{t \in \mathcal{T}} R_{i,j,t} + \frac{1}{J}, \sum_{t \in \mathcal{T}} \sum_{j' \neq j} R_{i,j',t} + \frac{J-1}{J}\right). \tag{2.2}$$

Now, suppose we had used some heuristic to estimate for each $t \in \mathcal{T}$ the probability ρ_t that t is a *bona fide* member of $\mathcal{T}$. We can apply this knowledge by scaling the mapped read counts for each junction of each LSV for each t relative to its ρ_t as such:

$$\Psi_{i,j} \mid \mathcal{T}, \{\rho_t\}_{t \in \mathcal{T}} \sim \text{Beta}\left(\sum_{t \in \mathcal{T}} \rho_t R_{i,j,t} + \frac{1}{J}, \sum_{t \in \mathcal{T}} \sum_{j' \neq j} \rho_t R_{i,j',t} + \frac{J-1}{J}\right). \tag{2.3}$$

These ρ_t constitute the first computational objective of the outlier detection algorithm.

In practice, each of these three density functions accounts for one set of bootstrapped read count samples. To account for within-sample variance, this density is computed separately according to Equation 2.2 for each sample, and all these densities are averaged per LSV junction. The resulting density is encoded and processed downstream as a vector of binned probability masses.

2.2.2. L1 divergence between P(PSI) and group median

Here we define our metric for determining how well an experiment agrees with the rest of the group on junction Ψ. First, we need to determine a suitable representative distribution for the group consensus Ψ, henceforth labeled $(\Psi_{i,j} \mid \overline{\mathcal{T}})$. We choose to represent the group consensus using the group median, as this measure of center is known to be robust to outliers. To accomplish this, we define how to construct a median density from a set of random variables.

Definition 1 *Suppose a set of m random variables* $\{X_1, \ldots, X_m\}$ *have densities* $P(X_1 \leq x) = F_1(x)$ *and so forth. The median of these random variables, denoted* $\overline{X}$*, has density* $P(\overline{X} \leq x) = \overline{F}(x)$ *such that*

$$\overline{F}^{-1}(q) = \underset{1 \leq i \leq m}{\text{med}} F_i^{-1}(q)$$

for each $0 \leq q \leq 1$*.*

In the event that the random variables in Definition 1 are discrete rather than continuous, the inverse density functions F_i^{-1} can be interpolated i.e. linearly without egrigious loss of precision. The median density $P(\Psi_{i,j} \leq \psi \mid \overline{\mathcal{T}})$ can thus be computed from the per-replicate binned densities $P(\Psi_{i,j} \leq \psi \mid t)$.

Next, we define our measure of distance between probability densities. We acknowledge the existence and widespread use of Kullback-Leibler divergence (KL divergence) to quantify this. In short, the KL divergence between two discrete probability densities P and Q is defined as

$$D_{KL}(P||Q) = -\sum_i P(i) \log_2 \left(\frac{Q(i)}{P(i)} \right).$$

However, we choose not to use KL as our distance metric because it is not bounded above. Indeed, in a worst case scenario where the two distributions are spike-and-slab densities with non-overlapping spikes, the KL divergence tends toward positive infinity as the height of the slab and width of the spike both approach 0. Instead, we employ the L1-divergence metric, defined as such:

Definition 2 *If P and Q are two probabilitiy densities, then the L1 divergence between them is*

$$D_{L1}(P,Q) = \frac{1}{2}\sum_i |P(i) - Q(i)|.$$

The divergence measure constructed by this definition has two key properties. First, it is symmetric with respect to the random variables; that is, $D_{L1}(P,Q) = D_{L1}(Q,P)$, unlike KL divergence. Second, $D_{L1}(P,Q) \in [0,1]$ for any two probability densities P and Q. This property makes L1 divergence slightly easier to interpret.

Let $d_{i,j,t} = D_{L1}(\Psi_{i,j} \mid t, \Psi_{i,j} \mid \overline{\mathcal{T}})$ be the L1 divergence between the distribution of $\Psi_{i,j}$ informed by experiment t compared to the median density for group $\mathcal{T}$. Because roughly 70% of LSVs are binary in nature, we reduce this to a single value for each LSV by taking the junction with the maximum L1 divergence, that is, $d_{i,t} := \max_{1\le j\le J} d_{i,j,t}$.

2.2.3. Distribution of L1 divergences, and how it is used to construct global weights

Having computed the $d_{i,t}$ for each $t \in \mathcal{T}$ for each LSV, we can observe how they are distributed. Figure 6 depicts these distributions for two situations where a group of three mouse tissue experiments has a known outlier. When the group has no outlier, all three replicates have the same distribution as Replicate 1 and Replicate 2 (data not shown). While the distribution of L1 divergences for all three experiments has a spike at 0, only the outlier has an additional spike at 1, indicating an enrichment in highly-disagreeing LSVs.

We leverage this information in a robust manner by modeling the number of highly-disagreeing LSVs per LSV. Let K_t be the number of LSVs for which $d_{i,t} \ge \tau$ for some fixed τ, and $K_{\mathcal{T}}$ be the size of the union of such LSVs across all experiments in $\mathcal{T}$. In a study with no outliers, we expect these K_t to be around the same, that is, each replicate contributes equally to the total disagreement on LSV Ψ in the study. If all the replicates are "well-behaved", we could model this with a Binomial distribution, where

$$K_t \sim \text{Binomial}(K_{\mathcal{T}}, p),$$

where $p = \frac{1}{N}$ represents the equal proportion of highly-disagreeing LSVs counted for each replicate. In practice, we observe greater dispersion in K_t than can be explained by a binomial model. To capture this dispersion, we instead proposed a Beta Binomial model with $p \sim \text{Beta}(\frac{\alpha}{N}, \alpha(1 - \frac{1}{N}))$, where α is a fixed dispersion hyperparameter. Using this model, we can finally express the relative probability ρ_t that experiment t is a replicate of the condition underlying $\mathcal{T}$:

$$\rho_t = min\left(1, \frac{P_{BB}(X > K_t)}{P_{BB}(X > \frac{K_\mathcal{T}}{N})}\right),$$

where P_{BB} represents the probability of the expression under the aforementioned beta-binomial model.

2.2.4. Expected (replicates) distribution of L1 divergences, and how it is used to construct local weights

ρ_t generated as described can be applied according to Equation 2.3 to globally weight each experiment's contribution to junction Ψ proportional to how much we believe each is a true representative of $\mathcal{T}$. This approach is termed "MAJIQ-gw". However, Figure 6 indicates that even a strong outlier only disagrees on Ψ for a relatively small number of LSVs. While this global weighting scheme corrects those events, it comes at the cost of the contribution the outlier's reads make towards quantification of the events where it does agree with the group consensus. To counteract this, we devised a scheme for estimating local (per-LSV) weights. We start by summarizing the per-experiment distributions of $d_{i,t}$ into an empirical model that represents an average replicate of $\mathcal{T}$ using the ρ_t as weights:

$$P(d_{i,\mathcal{T}} = x) = \frac{\sum_t \rho_t P(d_{i,t} = x)}{\sum_t \rho_t}.$$

We then define a new variable $\nu_{i,t}$ to represent the "local" weight for LSV i as a likelihood ratio between the density of $d_{*,t}$ and $d_{*,\mathcal{T}}$ in a neighborhood of $d_{i,t}$. That is, if we let $\varepsilon > 0$, then

$$\nu_{i,t} = \min\left(1, \frac{P(|X - d_{i,\mathcal{T}}| < \varepsilon)}{P(|X - d_{i,t}| < \varepsilon)}\right).$$

These $\nu_{i,t}$ are used in the MAJIQ Ψ quantification model in the same way as the ρ_t:

$$\Psi_{i,j} \mid \mathcal{T}, \{\nu_{i,t}\}_{t\in\mathcal{T}} \sim \text{Beta}\left(\sum_{t\in\mathcal{T}} \nu_{i,t} R_{i,j,t} + \frac{1}{J}, \sum_{t\in\mathcal{T}} \sum_{j'\neq j} \nu_{i,t} R_{i,j',t} + \frac{J-1}{J}\right). \tag{2.4}$$

This approach of using local weights in MAJIQ is termed "MAJIQ-lw".

2.2.5. Synthetic introduction of an outlier into an otherwise clean dataset

To benchmark the performance of MAJIQ-gw and MAJIQ-lw for correcting outlier replicates, we devised two strategies for synthetically introducing an outlier into a group of biological replicate RNA-seq experiments. The first is a replicate swap, in which one experiment in the group is selected at random to be removed and replaced with an experiment representing a completely different condition or tissue in the same dataset. The second is a more complex procedure wherein a replicate is transformed to become an outlier. Briefly, one replicate is selected at random to receive a "synthetic perturbation". For each LSV with sufficient read support for quantification, Ψ is quantified for each junction. Next, a random subset of LSVs, representing a fraction θ of the set of quantifiable LSVs, is selected. For each LSV in this subset, the expected Ψ ($E[\Psi]$) for one junction is then shifted by adding or subtracting a fixed δ. Whether this δ is added or subtracted is determined by a Bern($E[\Psi]$) random variable. The remaining junctions $E[\Psi]$ are scaled linearly such that the $E[\Psi]$ for all junctions sum to 1. Finally, the original read counts for that LSV are shifted such that they now explain the new $E[\Psi]$. To simulate the effect of a read-depth outlier, all read counts for the perturbed experiment can be scaled by a factor of γ.

2.3. Evaluation metrics

When designing new methods for analyzing genomic and transcriptomic data, the developer should evaluate the performance of their method against others designed to accomplish the same task. The question of what constitutes a fair comparison, while crucial, is scarecely discussed in the literature. There are, of course, standard metrics for measuring method performance. One popular metric is the area under the receiver operating characteristic curve (AUROC), which summarizes the trend in true positive rate as the tolerance for false

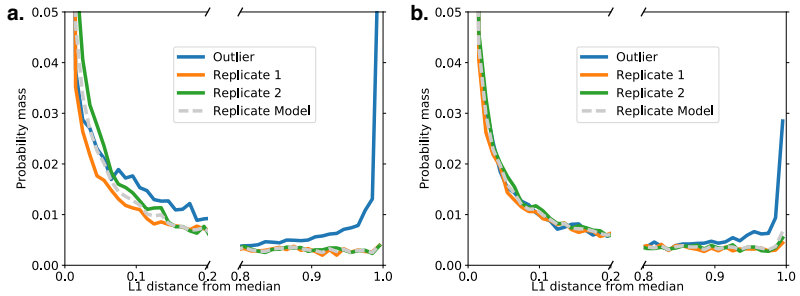


Figure 6: Distribution of L1 divergences over LSVs for a group of three mouse tissue replicates with an induced outlier. A: The outlier is introduced by replacing one replicate with an experiment from a different tissue. B: The outlier is introduced by shifting the expected Ψ of a subset of LSVs by up to 50% inclusion.

positives is relaxed. Normally, estimating false positive and true positive rates requires a ground truth of what is significant, which is generally not known *a priori* in real data. AUROC is therefore used more often when simulated data are available, as these can be custom generated with events preselected to be significant. For real data, we offer a metric called "intra-to-inter ratio" (IIR) as a proxy for false discovery rate, which is described in Subsection 2.3.2.

Irreproducible discovery rate (IDR) is a metric for evaluating methods on real data and is used by ENCODE for quality control in assessing ChIP-seq data and protocols. The idea behind IDR is that significant detections by a sound protocol should also be found significant if the same protocol is repeated on the same input. Hits that are not returned in the second replicate are deemed irreproducible. A complementary metric called "reproducibility ratio" (RR) was defined in Vaquero-Garcia et al. (2016), and a revised version of this is described in the next subsection.

2.3.1. Reproducibility ratio

In order to measure the internal consistency of differential splicing quantification tools, we define a metric called "reproducibility ratio" (RR). This metric is roughly complementary to the irreproducible discovery rate (IDR) often used to benchmark ChIP peak callers. In principle, a tool should report roughly the same events as significantly changing when presented with two different samples of the same comparison experiment. We formalize this

principle with the following procedure, which is applicable to any quantitative problem, not just splicing. For a given dataset with two or more replicates each of two distinct biological conditions, sample equal-sized, non-overlapping partitions from each condition. Run the tool on both partitions to measure differences, and rank the events in decreasing order of significance of those differences. Next, count the number of events in the ranking of one partition that pass your threshold of significance. This number, called N_A, is intrinsic to the algorithm and dataset in use. Finally, count the number of events in the top N_A of the first partition that also appear in the top N_A of the second partition. This is your reproduced count, R_A, and reproducibility ratio RR_A is this count as a fraction of the total number of events called significant in the first partition; that is,

$$RR_A = \frac{R_A}{N_A}.$$

For an unbiased algorithm, a higher RR_A is indicative of high confidence.

We also define $RR_A(n)$ for $n \leq N_A$ as a means of evaluating the reproducibility of the highest-confidence detections. In Vaquero-Garcia et al. (2016), this was formulated as the fraction of the top N_A ranked events in the first partition that are reproduced in the top n ranked events in the second partition. This meant that $RR_A(n)$ was bounded above by $\frac{n}{N_A}$, and methods were compared against each other by mapping $RR_A(n)$ against $\frac{n}{N_A}$. This definition was revised in Norton et al. (2018) to be the fraction of the top n ranked events in the first partition that are reproduced in the top n ranked events in the second partition. This new formulation made it easier to compare reproducibility ratio between methods for fixed n, and better emphasized the reproducibility of the very topmost events in the ranking.

2.3.2. Intra-to-Inter Ratio

Reproducibility ratio alone is not enough to indicate a method's performance on a dataset: a highly biased method can score high in reproducibility. To demonstrate that an algorithm is unbiased on real data, we present another metric called "intra-to-inter ratio" (IIR). This

metric rests on the principle that detection of significantly-changing events between two partitions of the same condition should be much less than detection between two partitions of different conditions. Moreover, any event called significantly different in a within-group comparison is likely a false positive; we therefore label such events as "putative false positives" (PFP). The procedure for estimating IIR is similar to the RR procedure described above. Briefly, the dataset is partitioned as before into two equal-sized subsets from each comparison group. The algorithm is used to count the number of significantly-different events between the two partitions of one group i.e. the number of PFP or N_{PFP}, and the number of significantly-changing events between one partition from each group (N_A). The IIR for this method is the ratio of the PFP count to the number of events that are changing between groups; that is,

$$IIR_A = \frac{N_{PFP}}{N_A}.$$

2.3.3. Real data

Model performance was benchmarked on two publicly-available mouse RNA-seq datasets. The first, published in Keane et al. (2011), covers six different body sites (hippocampus, lung, liver, spleen, heart, and kidney) with six replicates each. We had previously determined that for some tissues, one or two replicates did not have sufficient read coverage for splicing quantifications, so these were excluded from the analyses presented here. The second was provided by Zhang et al. (2014) and covers twelve different body sites (brown and white fat, cerebellum, heart, liver, lung, adrenal, brainstem, skeletal muscle, kidney, aorta, and hypothalamus). The mice in this study were trained to a 12-hour light, 12-hour dark cycle for a week and then held in 24-hour darkness. Sample collection for RNA-seq was performed every six hours starting 22 hours into the dark-only period. Each RNA-seq experiment was performed in technical duplicate; for our purposes, we consider sample pairs spaced 24 hours apart to be biological duplicates.

2.3.4. Synthetic data

By its definition, IIR can be interpreted as a proxy for false discovery rate (FDR) when the truth about what is changing is unknown. Nevertheless, it is important to demonstrate

performance on a controlled, simulated dataset where the ground truth is predetermined. The dataset generated for these evaluations should be as close as possible to the biology the tool is designed to measure. This is a significant challenge for RNA-seq, as the sources of biological noise observed in real data are difficult to simulate. Additionally, most RNA-seq simulators attempt to generate transcript abundances rather than the reads themselves. A notable exception is BEERS (Baruzzo et al., 2016), a simulator designed specifically to benchmark RNA-seq aligners. BEERS is a modular simulator where each module applies a different source of technical noise to the simulation model.

We employed BEERS to generate simulated datasets from 11 mouse hippocampus and liver tissue replicates obtained from Keane et al. (2011). Briefly, input transcript levels were estimated for each gene in the mouse genome using evidence from the original RNA-seq experiments. Gene-level expression were estimated empirically from the raw RNA-seq reads for each of the 41,113 genes in the ENSEMBL v75 mm10 annotation, so as to not bias these estimates towards any one model of transcript quantification. A subset of 3,055 genes was selected at random from the annotation to represent true differential splicing between the two tissues; the rest were assigned the same distribution of per-gene relative isoform abundance with some added noise to simulate biological variance. For each gene, Ψ values were estimated for the most complex LSV detected by MAJIQ in the annotation; *de-novo* events were not considered, as not all methods detect unannotated splicing events. The generated FASTA files were fed to the respective pipelines for splicing quantification according to the authors' recommendations. For rMATS and MAJIQ, which both require aligned BAMs, we mapped the simulated reads to prebuilt mm10 indices using STAR-2.5.3a with the option `--alignSJoverhangMin 8`.

2.3.5. Comparison to biochemical assays (RT-PCR)

All the algorithms presented here are designed to quickly and efficiently estimate splicing levels transcriptome-wide from high throughput sequencing data. A principal objective, then, must be to reproduce the accuracy of biochemical assays at scale. Quantifications derived from carefully-performed reverse-transcription polymerase chain reaction (RT-PCR)

experiments are often hailed as the gold standard for biochemical splicing quantification, but the procedure is quite labor-intensive and not scalable for transcriptome-wide analysis. In Vaquero-Garcia et al. (2016), we selected fifty LSVs with strong read support from Zhang et al. (2014) for follow-up by high-fidelity RT-PCR performed in triplicate. The RT-PCR quantifications were shown to correlate strongly with MAJIQ Ψ values, demonstrating the software's accuracy for Ψ quantification. This same principle is applied to compare the accuracy of Ψ quantifications of other methods.

2.4. Results

2.4.1. The impact of an outlier on differential splicing predictions and reproducibility

To evaluate the performance of MAJIQ on induced outliers, we used cerebellum and liver RNA-seq experiments from Zhang et al. (2014), and repeatedly generated outliers by perturbing a random cerebellum experiment as described in Section 2.2.5, varying θ, δ, and γ each time. We measured the impact of each of these parameters on the ρ_t for each experiment in the resulting group (Figure 7a,d,f). We further evaluated the effects on detection power (Figure 7b,e,g) and reproducibility ratio (Figure 7c,f,h) in a $\Delta\Psi$ comparison with the liver group. In these latter tests, we compared the previous unweighted MAJIQ model (MAJIQ-nw) to MAJIQ-gw, MAJIQ-lw, and a scenario where the outlier had been detected by some heuristic and removed from the evaluation (MAJIQ-rm).

In the cases where θ (Figure 7a-c) and δ (Figure 7d-f) were varied (with the remaining parameters fixed), the ρ_t for the outlier tends to tend towards 0 with increasing θ or δ (negative log tends towards ∞), whereas the ρ_t for the remaining replicates remains close to 1, demonstrating the algorithm's sensitivity and specificity to the outlier replicate. In addition, the number of events detected as differentially spliced between hippocampus and liver increases dramatically in response to increases in θ and δ, however the reproducibility of those events drops significantly. If we assume that MAJIQ is unbiased on well-behaved sample groups (an assumption we verify in later sections), this implies that the additional detections made in the presence of an outlier are false positives, underscoring the need for proper correction. Comparing between the four different MAJIQ models, we observe that

all three outlier correction models control for these false detections. However, MAJIQ-gw and MAJIQ-rm suffer slightly in that some detection is lost even under conditions where no outlier is induced i.e. either θ or δ is 0. MAJIQ-lw does not suffer from this loss of power - it detects the same number of LSVs as significantly changing as MAJIQ-nw does under no-outlier conditions. This observation highlights the robustness of the local correction to naturally-occurring variation, which the global correction does not account for.

2.4.2. Comparison between methods on synthetic data

The dataset from Keane et al. (2011) was used as input for simulating RNA-seq experiments using BEERS. The resulting FASTA files were quantified for differential splicing between simulated hippocampus and liver using SUPPA, rMATS, and MAJIQ with and without local-weights outlier correction. Events were called as significantly-changing based on the recommendations of the tools' respective authors. These results are depicted in Figure 8. The definition of what constitutes a splicing event depends on the tool in use, and the total number of events quantified (the sum of the #CHG, #NO_CHG, and #GREY columns) reflects this. SUPPA, for instance, attempts to quantify Ψ for all annotated binary events. rMATS, meanwhile, only considers splicing events that fit a limited model of local splicing variations. MAJIQ, meanwhile, is the only tool out of the three that attempts to quantify more complex splicing variations, though its event count is slightly lower than that of SUPPA due to the stringent quantifiability filters MAJIQ imposes. In MAJIQ's case, when a splicing event is ambiguously described by two or more LSVs, the ambiguity is resolved by taking the LSV reporting the highest $P(|\Delta\Psi| \geq 20\%)$. Based on the observations from Section 2.4.1, we perform all further comparative analyses on only the MAJIQ-nw and MAJIQ-lw models.

For each algorithm, we counted the number of splicing events reported as not changing (NO_CHG, i.e. $|\Delta\Psi| \leq 0.05$) and changing (CHG, i.e. $|\Delta\Psi| \geq 0.20$). Events in the interceding grey area (GREY) were excluded from further analysis.

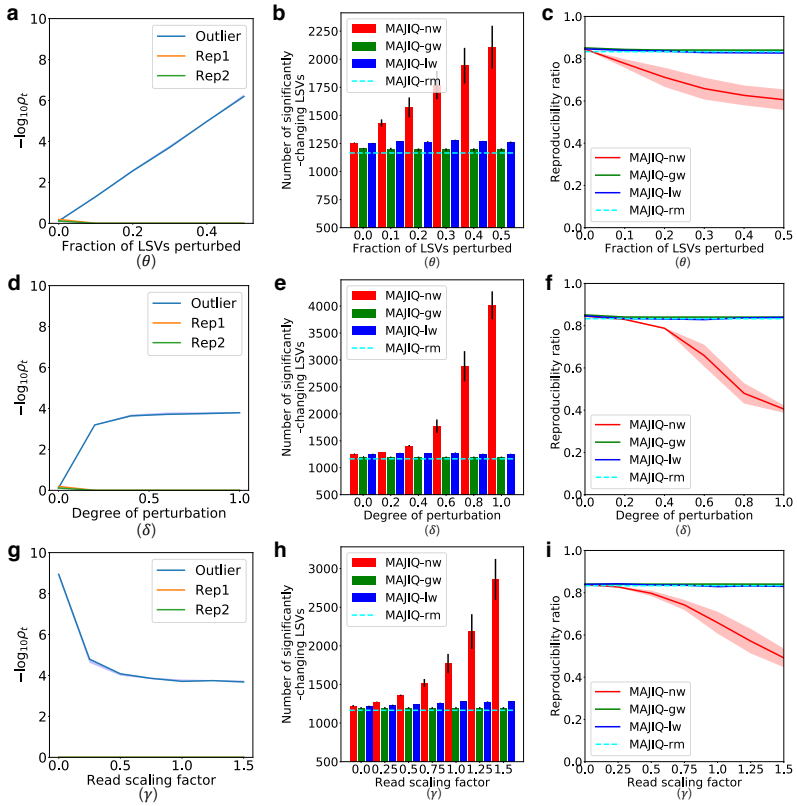


Figure 7: Synthetic perturbation of tissue replicates. In each test, three cerebella were compared against three livers from Zhang et al. (2014), but one of the cerebella was perturbed. MAJIQ-nw is the previous algorithm equivalent to fixed weights ($\rho_t = 1$). MAJIQ-rm is a control case where we assume some heuristic (e.g. PCA) was able to detect the outlier and remove it before executing the previous fixed-weights MAJIQ. **a,d,g.** Effect on ρ_t for the perturbed "outlier" (blue) and unperturbed replicates (Rep1,2 in green and orange). **b,e,h.** Effect on the number N_A of events detected to have $P(\Delta\Psi > 0.20) > 0.95$ between the cerebellum and liver samples. **c,f,i.** Effect on the reproducibility ratio $RR(N)$. **a-c.** θ, the fraction of LSVs perturbed, is varied between 0 and 0.5. At 0, the effect is the same as having no perturbation at all. **d-f.** δ, the maximal amount by which Ψ is perturbed, is varied between 0 and 1. $\delta = 0$ is equivalent to no perturbation. **g-i.** γ, the "read scaling factor", is varied between 0 and 1.5. When this factor is 0, it is functionally equivalent to a global weight of 0.

Naive

Tool	#CHG	#NO_CHG	#GREY	TP	TN	FP	FN	Sens	Spec	FDR	FNR
SUPPA	5611 ±50	16874 ±169	3661 ±77	1045 ±20	16394 ±170	480 ±25	4566 ±37	0.19 ±0.00	0.97 ±0.00	0.31 ±0.01	0.81 ±0.00
rMATS	512 ±16	4604 ±55	1350 ±30	254 ±2	4543 ±49	61 ±6	258 ±15	0.50 ±0.01	0.99 ±0.00	0.19 ±0.01	0.50 ±0.01
MAJIQ-nw	1133 ±29	15379 ±130	1509 ±4	979 ±20	15372 ±130	7 ±1	153 ±14	**0.86 ±0.01**	**1.00 ±0.00**	**0.01 ±0.00**	**0.14 ±0.01**
MAJIQ-lw	1133 ±29	15379 ±130	1509 ±4	977 ±21	15372 ±130	6 ±2	156 ±14	**0.86 ±0.01**	**1.00 ±0.00**	**0.01 ±0.00**	**0.14 ±0.01**

Realistic

Tool	#CHG	#NO_CHG	#GREY	TP	TN	FP	FN	Sens	Spec	FDR	FNR
SUPPA	5665 ±74	16955 ±173	3933 ±60	976 ±35	16390 ±163	564 ±10	4688 ±40	0.17 ±0.00	0.97 ±0.00	0.37 ±0.00	0.83 ±0.00
rMATS	495 ±8	4475 ±63	1320 ±18	239 ±8	4408 ±56	67 ±6	256 ±4	0.48 ±0.01	0.99 ±0.00	0.22 ±0.01	0.52 ±0.01
MAJIQ-nw	1097 ±25	13090 ±135	1396 ±4	913 ±19	13077 ±137	13 ±2	184 ±8	**0.83 ±0.00**	**1.00 ±0.00**	**0.01 ±0.00**	**0.17 ±0.00**
MAJIQ-lw	1097 ±25	13090 ±135	1396 ±4	912 ±21	13077 ±137	13 ±2	185 ±5	**0.83 ±0.00**	**1.00 ±0.00**	**0.01 ±0.00**	**0.17 ±0.00**

Realistic+swap

Tool	#CHG	#NO_CHG	#GREY	TP	TN	FP	FN	Sens	Spec	FDR	FNR
SUPPA	5696 ±27	16719 ±140	4008 ±63	705 ±50	16362 ±150	357 ±11	4990 ±33	0.12 ±0.01	0.98 ±0.00	0.34 ±0.02	0.88 ±0.01
rMATS	503 ±14	4376 ±60	1416 ±6	94 ±8	4349 ±55	26 ±4	408 ±6	0.19 ±0.01	0.99 ±0.00	0.22 ±0.02	0.81 ±0.01
MAJIQ-nw	1129 ±18	13233 ±190	1421 ±29	710 ±30	13199 ±203	34 ±13	419 ±44	0.63 ±0.03	**1.00 ±0.00**	0.05 ±0.02	0.37 ±0.03
MAJIQ-lw	1129 ±18	13233 ±190	1421 ±29	806 ±27	13202 ±204	30 ±14	323 ±40	**0.71 ±0.03**	**1.00 ±0.00**	**0.04 ±0.02**	**0.29 ±0.03**

Figure 8: Evaluation using "realistic" synthetic datasets. Each synthetic sample is created to match a real sample in terms of gene expression and a lower bound on transcriptome complexity. Three datasets were created: "Naive" with naive read simulation (uniform coverage, no errors, biases or indels); "Realistic" with more realistic read generation; "Realistic+swap" where one sample in an outlier. All datasets involve 3 biological replicates per group. Each method was evaluated using its own definition of AS events, which means the number of events is *not* comparable between methods. CHG are changing events ($|E[\Delta\Psi]| \geq 20\%$), NO_CHG are non changing events ($|E[\Delta\Psi]| \leq 5\%$), and GREY are events for which $20\% > |E[\Delta\Psi]| > 5\%$ and on which the algorithm is not evaluated. DEXSeq is not included here as it is not able to return dPSI values.

2.4.3. Comparison between methods on real data

To evaluate the various splicing methods on real data, we employed the cerebellum and liver tissue experiments from Zhang et al. (2014). We introduced an outlier by systematically replacing each cerebellum experiment with the time-matched experiment from skeletal muscle (swap). Figure 9 depicts the results from these comparisons. Figure 9a compares the reproducibility ratio (RR) achieved by MAJIQ, SUPPA, rMATS, and DEXSeq on these comparisons. For all methods, introducing the tissue swap reduced detection power, though the impact was least severe for MAJIQ. RR also dropped for all methods, however MAJIQ-lw retains an RR close to the no-outlier baseline. We note that DEXSeq reported an absurdly high number of significantly-changing events between conditions. DEXSeq does not report Ψ for splicing events; instead, it compares absolute expression levels at each exon in the transcriptome. Under this model, a change in gene-level expression between two conditions inevitably appears as a change in exon expression even if there is no variation in splicing.

Figure 9b,c compare the N NoSignal (i.e. #PFP) and IIR between methods as described when the groups are defined from biological replicates. Between the three methods that quantify Ψ, MAJIQ controls best for within-group variations, while SUPPA struggles the hardest. DEXSeq's #PFP and IIR are reported for completeness, however its counts suffer from the same caveat explained earlier. Nevertheless, we chose to represent this method here, as the events it detects are compatible with the definition of RR and IIR.

For Figure 9d, we matched the splicing events defined by SUPPA, rMATS, and MAJIQ with the 50 splicing events validated by RT-PCR in Vaquero-Garcia et al. (2016), and compared the $\Delta\Psi$ reported by each method in cerebellum vs. liver from Zhang et al. (2014). The same tissue swap procedure was employed to measure the impact of an outlier on the accuracy of splicing quantifications. On the control (no swap) comparison, rMATS and MAJIQ performed similarly in how well they correlate with RT-PCR $\Delta\Psi$, and both outperform SUPPA. As expected, this correlation drops significantly for all methods when an outlier is

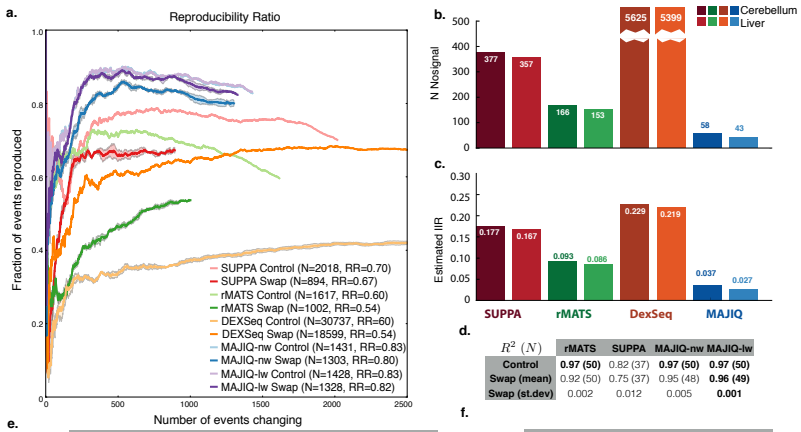
introduced, though the impact is least severe on MAJIQ-nw and MAJIQ-lw.

2.4.4. Evaluating method performance

A fair and unbiased suite of evaluation metrics is important when comparing one's method to others designed for the same task. To illustrate the principles of best practices in methods evaluation, we address some claims made by Li et al. (2018) in the course of evaluating their method LeafCutter against other splicing quantification software, including MAJIQ. These comments relate to Fig. 2 in the original paper (reproduced here as Figure 10) (Vaquero-Garcia et al. (2018), in review).

At issue are three claims made by Li et al. (2018) when comparing the performance of LeafCutter to that of MAJIQ, rMATS, and Cufflinks2. First, while LeafCutter and Cufflinks can quantify differential splicing between groups of 15 samples in under 10 hours, MAJIQ requires over 60 hours to quantify the same dataset, while rMATS runs out of memory. The authors cite this observation to claim that MAJIQ and rMATS do not scale well with increasing sample size. Second, the p-values returned by LeafCutter and rMATS are well-calibrated, however MAJIQ's are not. Finally, when measuring precision and recall on simulated data, LeafCutter performs strongly when calling both lowly- and highly-differentially spliced isoforms. In contrast, MAJIQ's true positive rate saturates under conditions of low simulated changes in splicing, rMATS struggles to exceed a TPR of 50%, and Cufflinks2's performance degrades slightly as the simulated change in alternative isoform abundance increases.

To address the first claim of non-scalability, we queried the software versions used in the original analyses. We determined that MAJIQ 0.9.2a (released in early 2016 alongside Vaquero-Garcia et al. (2016)) and rMATS 3.2.5 (released in August 2016) were employed. At the time these analyses were prepared for publication (mid-2017), two major upgrades were released for MAJIQ, with version 1.1 being current as of May 2017. We had shown that MAJIQ 1.1 was capable of handling datasets upwards of 700 samples within the span of a week. Around that time, Norton et al. (2018) was in preparation and includes comparative



RR_A (N_A)	DEXSeq	rMATS	SUPPA	MAJIQ-nw	MAJIQ-lw	MAJIQ-lw-relaxed
100% (80M)	0.60 (30737)	0.60 (1617)	0.70 (2018)	**0.83 (1431)**	**0.83 (1428)**	0.69 (2450)
50% (40M)	0.57 (22813)	0.56 (1296)	0.68 (1548)	**0.81 (801)**	**0.81 (794)**	0.64 (1829)
25% (20M)	0.54 (16104)	0.52 (931)	0.62 (1076)	**0.80 (360)**	0.79 (360)	0.60 (1246)

R^2 (N)	rMATS	SUPPA	MAJIQ-nw	MAJIQ-lw
Control	**0.97 (50)**	0.82 (37)	**0.97 (50)**	**0.97 (50)**
Swap (mean)	0.92 (50)	0.75 (37)	0.95 (48)	**0.96 (49)**
Swap (st.dev)	0.002	0.012	0.005	**0.001**

R^2 (N)	rMATS	SUPPA	MAJIQ-nw	MAJIQ-lw	MAJIQ-lw-relaxed
100% (80M)	**0.97 (50)**	0.84 (37)	**0.97 (50)**	**0.97 (50)**	**0.97 (50)**
50% (40M)	**0.97 (50)**	0.84 (37)	**0.97 (43)**	**0.97 (43)**	**0.97 (49)**
25% (20M)	0.95 (50)	0.85 (37)	**0.96 (20)**	**0.96 (20)**	**0.96 (47)**

Figure 9: Evaluation using real data. **(a)** Reproducibility ratio (RR) plots for detection of DS events between cerebellum and liver, with (Swap, dark line) or without (Control, faded line) a mislabeled muscle. The end of the line marks the point in the graph matching the number of events reported as significantly changing ($RR(N_A)$, see main text). Because DEXSeq report more than 3000 significantly-changing exonic segments, we present its extended RR curve in Supplementary Figures. **(b)** Number of events reported by each methods as significantly changing between two groups of biological replicates from the same condition (light color - liver, dark color - cerebellum). See Supplementary Material for equivalent plots when comparing groups of technical replicates. **(c)** The Inter to Intra Ratio (IIR), representing the ratio between the number of DS events reported when comparing biological or technical replicates (N_{PFP}) and the number of events reported when comparing similarly sized groups but from different conditions (N_A). See Supplementary Figures for equivalent plots when N_{PFP} is derived using technical replicates. **(d)** The same control and swap experimental setup as in (a) with accuracy assessed using 50 RT-PCR experiments from (Vaquero-Garcia et al., 2016). Values represent fraction of variations explained (R^2) and the number of events detected in parentheses. DEXSeq is not included here since it does not output Ψ values. Scatterplots are presented in Supplementary Figures. Mean and standard deviation derived by swapping out each of the cerebellum samples. **(e,f)** Repeating (a) and (d) but with subsets of the FASTQ files to test the effect of read coverage.

analyses with the newly-released rMATS 4.0 which implements parallel processing and a more efficient algorithm. We therefore postulated that the current versions of both software would scale much better on larger comparisons, refuting the claim in Li et al. (2018). We tested this hypothesis by running rMATS 4.0, three different releases of MAJIQ, and the LeafCutter release version to call differential splicing between the Genotype-Tissue Expression project (GTEx) (GTEx Consortium et al., 2017) v7 tissues using sample sets of increasing size, up to 15 cerebellum samples vs. 15 skeletal muscle samples (Figure 11a). As expected, rMATS 4.0 and MAJIQ 2.0 ran in a timeframe similar to that of LeafCutter.

The second claim of poor p-value calibration is addressed in part by Li et al. (2018) themselves (Supplementary Note 2.2). The authors elected to use $\hat{p} = 1 - P(|\Delta\Psi| \geq 0.2)$ as a proxy for p-value. However, these two quantities do not express the same concept. A p-value, by definition, is the likelihood of observations under a null model. $\hat{p}$, meanwhile, is a posterior probability for $\Delta\Psi$ given the observed reads distribution. It is therefore natural that the assumptions made of p-values do not fit $\hat{p}$ in the same way.

In evaluating the third claim, we noted that the ROC calculations for MAJIQ were the result of executing the quantifier incorrectly for these purposes. Proper ROC estimates require the investigator to catalogue both positive and negative classifications. By default, MAJIQ only reports LSVs with at least one significantly-changing junction. To also return the non-changing LSVs, the user should pass `--show-all`. This flag was omitted by Li et al. (2018), leading to the appearance of unimpressive performance by MAJIQ. The correct assessment of MAJIQ is overlayed atop the original in (Figure 11c).

A second issue regarding the third claim is the simulation structure employed by Li et al. (2018). Briefly, 160 protein-coding genes with multiple isoforms each were selected at random from the human genome annotation (hg38/GRCh38). For each gene, the expression of a randomly-selected transcript isoform was increased by a fixed ratio. The expression values of these genes were supplied to polyester (Frazee et al., 2015) to generate RNA-seq reads. This framework has two major shortcomings. First, the authors chose to simulate changes

in splicing by adjusting isoform expression. However, LeafCutter, rMATS, and MAJIQ all measure splicing as changes in relative junction usage or intron excision levels, which are quantified differently. As an example, consider an example splice junction which is unique to one transcript isoform. Empirically, the Ψ for that junction would be

$$\Psi = \frac{n_j}{n_J},$$

where n_j is the number of reads mapping across that junction and n_J is the number of reads mapping across all junctions that are part of the same splicing event, LSV, or cluster. A 10% increase in expression of that one isoform would result in Ψ shifting to

$$\Psi^* = \frac{1.1n_j}{n_J + 0.1n_j}.$$

$\Delta\Psi$, therefore, would be

$$\Psi^* - \Psi = 0.1\Psi\frac{n_J - n_j}{n_J + 0.1n_j}.$$

In the event that $\Psi = 0.5$, this estimate gives $\Psi^* = 0.52$ and $\Delta\Psi = 0.02$. This small change in splicing would not be detected by tools that define significant $\Delta\Psi$ as those surpassing a fixed threshold i.e. 10%; this is evident in Figure 10c, where no method performs better than randomly guessing. On the opposite extreme, a 3X change in expression would result in $\Delta\Psi = 0.25$, which is almost trivial to detect.

Another issue we observed in the Li et al. (2018) analysis pertains to the synthetic data employed by the authors. We found the data to be unrealistic, hampering the ability to conlcude about algorithms' relative performance. The authors attempt to represent human transcriptomic variation using only 160 genes, however the v94 ENSEMBL release estimates that there are over 20,000 protein coding genes (Zerbino et al., 2018).

Additionally, polyester makes several simplifying assumptions about RNA-seq data that limit its ability to capture real variance in human RNA-seq. To demonstrate this, we used a strategy similar to that employed for outlier detection, where instead of using L_1

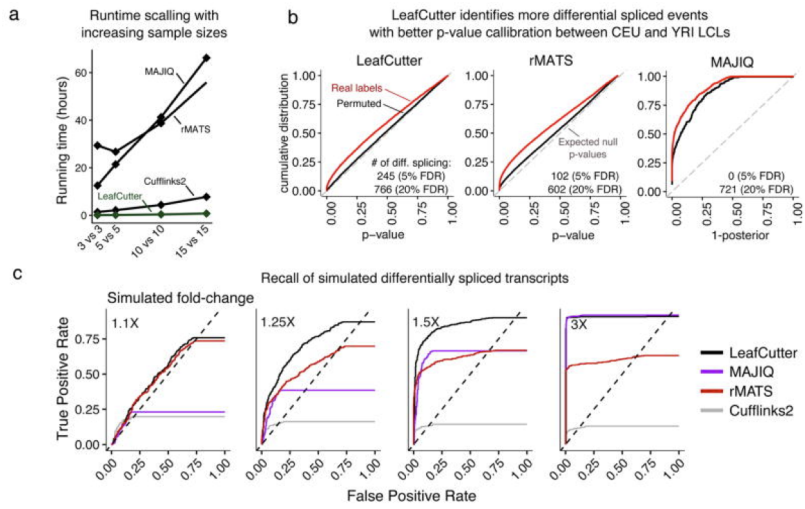


Figure 10: Reproduction of Figure 2 from Li et al. (2018).

divergence in Ψ, we measured the absolute deviation $\hat{\delta}$ in empirical junction ratios per LSV as reported by MAJIQ's builder. The empirical ratios were the same quantities used to seed the transcript ratios supplied to BEERS. We compared the distribution of $\hat{\delta}$ between mouse hippocampus replicates from Keane et al. (2011), simulated reads generated by BEERS, human cerebellum samples from GTEx v7, and the unspiked simulated samples used in Li et al. (2018) (Figure 12). An increase in within-group variance is evidenced by a right shift in the curve for that dataset. As expected, the GTEx samples exhibit significantly more within-group variance than either the mouse dataset or the simulations based thereon. Moreover, while the variance of the BEERS simulated replicates closely resembled their murine counterparts, the simulations used to benchmark LeafCutter do not approach a proper model of the variance observed in the real human samples they are meant to represent.

2.5. Discussion and conclusions

When measuring splicing in a set of technical or biological replicate samples, an outlier can skew quantifications and affect the quality of detected changes in alternative splicing. It is therefore important to detect outliers in an unbiased fashion. Since outlier behavior affects only a subset of splicing events, and because simply removing the outlier from analysis

wholesale weakens overall method performance, it is also important to account for the information encoded in the unaffected events while calling replicates which behave badly on an abundance of events. The algorithm presented in this chapter meets these specifications. Moreover, the strategy for correcting for these outliers was shown to be robust not only to perturbations of the underlying data, but to naturally-occuring variation as well.

We demonstrated the benefits of this outlier correction in evaluations on real data, and compared it to the behavior of other algorithms when an outlier is introduced. In general, MAJIQ is already fairly robust to outliers and experiences only a modest boost in performance from the correction model. The other tools, however, proved far more sensitive to the outlier when evaluated on reproducibility, detection power, putative false discovery rate, and correlation with RT-PCR quantifications.

The evaluation metrics defined in this chapter, particularly IIR, allow an unbiased comparison between splicing quantification tools, particularly when the events being described are similar in construction. That said, these metrics routinely disfavor DEXSeq, making it appear extremely sensitive but highly irreproducible. The proximal cause is the use of p-values for an absolute differential expression measure, which behaves quite differently from Ψ.

As a final point, we observed that within-group variance is an inherent property of the observed dataset, and takes characteristics from the original biological source. The work in this chapter primarily addresses the case of a misbehaving sample in a group of technical or biological replicates of inbred mouse tissues. Wild populations, free from these genetic and environmental controls, naturally exhibit more phenotypic variance. Human populations in particular display a great deal of heterogeneity resulting from generations of admixture between groups from around the globe, along with local adaptations from natural and artificial influencers in the environment. As such, the variance observed in RNA-seq from humans is greater than that in mouse littermates subject to uniform treatment, and we expect it to increase further in disease contexts. While the work presented in this chapter

deals with outliers in data of the latter sort, it leaves open the question of how to account for meaningful variance in heterogeneous human population and disease studies, which lately consist of hundreds if not thousands of independent RNA-seq experiments. We offer a modeling approach to the problem of handling large heterogeneous datasets for splicing quantification in the next chapter.

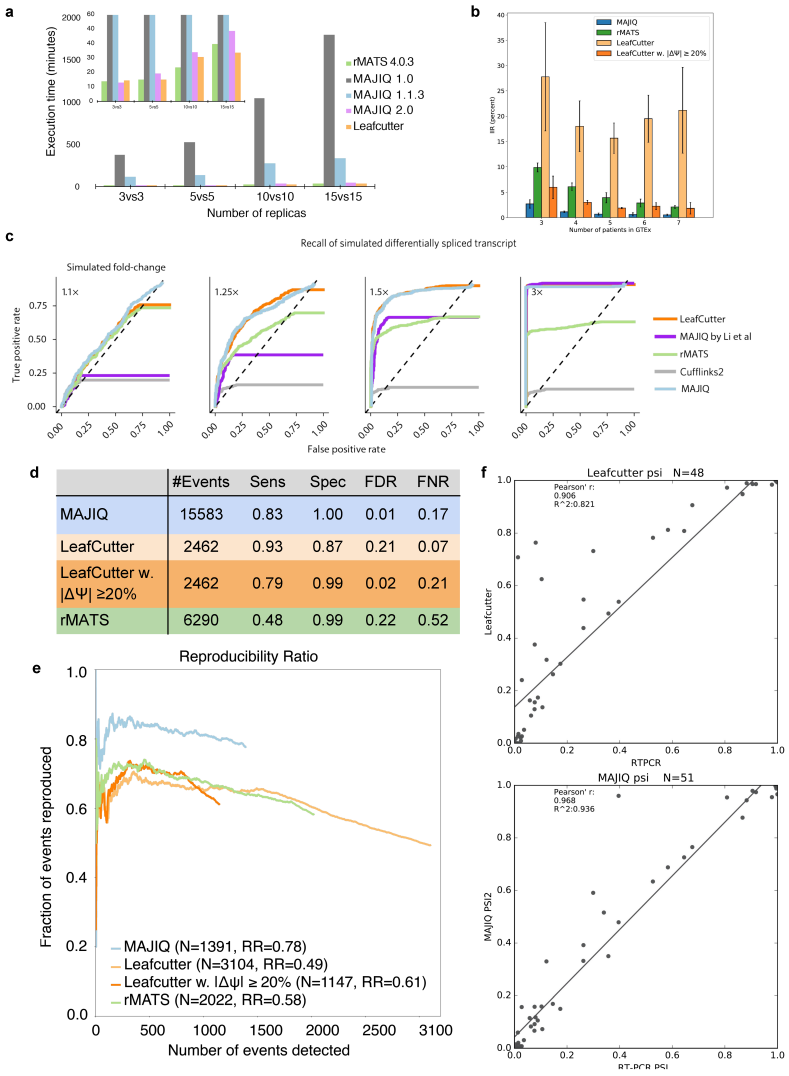


Figure 11: **Comparative evaluation**. Panels described on the following page.

Figure 11: **(a)** Running time for each algorithm, when comparing groups of different sizes. **(b)** The Intra to Inter Ratio (IIR) mean and standard deviation when using $3-7$ GTEx per tissue group (skeletal muscle).
The IIR, serving as a proxy for false discovery, represents the ratio between the number of differential events reported when comparing biological replicates of the same tissue (putative false positives), and the number of events reported when comparing similarly sized groups but from different conditions (here skeletal muscle and cerebellum, see main text and supplementary for details).
(c) The original ROC plots from Li *et al.* for evaluating each method's accuracy, with the correct execution of MAJIQ superimposed on them (blue line). The blue line was derived using scripts supplied by Li *et al.* for their data generation. The simulated data used in Li *et al.* (purple line) lacks high variability within a group compared to human tissue data from GTEx (red line) and does not match biological replicates from mouse tissues (blue line). Simulated data from Norton et al., made to mimic the blue line and used in (e) is shown in orange.
(d) Evaluation using "realistic" synthetic datasets: each synthetic sample is created to match a real sample in terms of gene expression and a lower bound on transcriptome complexity. This simulation does involve *de-novo* events which are not captured by rMATS or intron retention (not modeled by LeafCutter). All datasets involve 3 biological replicates per group. Each method was evaluated using its own definition of alternative splicing events, so events are not directly comparable between methods. Positive events were defined as those with ($|E[\Delta\Psi]| \geq 20\%$), and negative events were defined as those with a small difference between the groups of ($|E[\Delta\Psi]| \leq 5\%$).
(e) Reproducibility ratio (RR) plots for differentially spliced events between cerebellum and heart GTEx samples ($n = 5$ per group, as in Li *et al.*). The end of the line marks the point in the graph matching the number of events reported as significantly changing ($RR(N_A)$, see main text and supplementary). Events detected are not directly comparable as each algorithm uses a different definition for splicing events. The end of the line marks the point in the graph matching the number of events reported as significantly changing ($RR(N_A)$, see main text). Events detected are not directly comparable as each algorithm uses a different event definition.
(f) Evaluation of accuracy using RT-PCR experiments from Vaquero-Garcia et al. (2016). Both algorithms were used to quantify Ψ using RNA-seq from Zhang et al. (2014) and RNA from matching Liver tissue was used for validation.

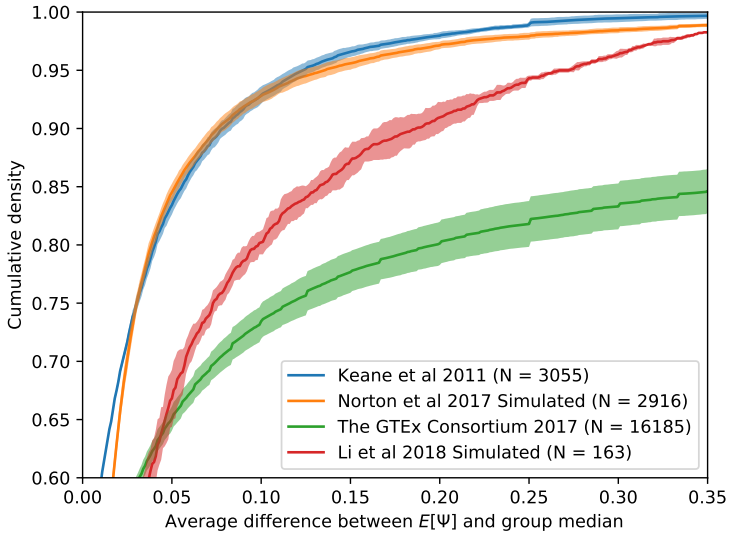


Figure 12: Distribution of absolute deviations in empirical Ψ for two real and two simulated datasets. N: Number of LSVs represented. Lines and ribbons represent mean and ± 1 standard deviation, respecitvely, of three random subsamples from the original dataset (5 samples each from Keane et al. (2011) and BEERS, 159 from GTEx, and 8 from Li et al. (2018)).

CHAPTER 3 : Large heterogeneous datasets

3.1. Introduction

The MAJIQ model was originally implemented to quantify differences in splicing between small groups of replicates, such as in a splice factor knockout or CRISPR mutation screen. The mouse tissue RNA-seq samples published in Keane et al. (2011) and Zhang et al. (2014) are examples of such clean datasets, and MAJIQ has been shown to perform well on these compared to competing methods. However, many datasets of interest for splicing quantification do not behave in this manner. The previous chapter discussed the situation where an outlier is present in one of the condition groups being compared. Larger datasets, in particular those from human population cohorts such as TCGA and GTEx, behave quite differently when handled in bulk. In addition to the surplus of data captured by these datasets, there also tends to be more variance between samples within a condition group because the samples are not biological replicates. This is an expected consequence of sampling from such a population exhibiting great genetic and environmental heterogeneity. Furthermore, splicing differences between conditions (i.e. case vs. control in a disease study) may be confounded with genomic variants and fixed known or unknown factors. While handling this variance and the potential contributors thereto is important, current methods for splicing quantification do not explicitly do so. Figure 12 represents the degree of within-group variability in a handful of datasets, including GTEx where this variability is known *a priori* to be heterogeneous relative to the other datasets depicted.

The assumptions made by existing methods ought to be reconsidered in the context of large heterogeneous datasets. As previously discussed, the kinds of datasets for which MAJIQ was originally designed are expected to have low variability between experiments from the same condition. From this expectation, one can reasonably assume that there exists a Ψ for each junction in each LSV which explains the distribution of reads in each sample within a group, and a $\Delta\Psi$ that explains the difference in read distributions at that LSV junction between two groups or between subsets of those two groups. The MAJIQ Ψ model

implements this intuition using a Bayesian framework, in which an noninformative prior density is assumed for Ψ of each LSV junction and updated using the junction-mapped read counts informed by each experiment in the condition group. Formally, for junction j in LSV i,

$$\Psi_{ij} \sim \text{Beta}\left(\alpha_{i0} + \sum_k R_{ijk}, \beta_{i0} + \sum_{\hat{j}} \sum_k R_{i\hat{j}k} - \sum_k R_{ijk}\right),$$

where R_{ijk} is the number of reads mapping across junction j of LSV i in replicate k. In the context of large heterogeneous datasets, this model has two major flaws. First, this model assumes that there is a single distribution of Ψ for each LSV which explains all samples in the group. While this is applicable to replicate samples, it does not hold for RNA-seq datasets sourced from individuals with distinct genetic and environmental backgrounds. A second concern with the MAJIQ model is its susceptibility to variation in local coverage. Disparities in local coverage can be the result of differences in total expression at that locus, or in sequencing depth between experiments. In general, samples with more reads mapping to an LSV tend to bias the group Ψ quantification more strongly towards the distribution inferred from considering that sample by itself. This effect is far more significant in situations where the assumption of shared underlying Ψ does not hold.

Many published splicing quantification methods, which were described in Chapter 2, are similarly tuned for smaller, non-heterogeneous datasets, and make assumptions specific to that setting. These methods have their own shortcomings when tested on large heterogeneous datasets. rMATS (Shen et al., 2014), for instance, implements a model of exon Ψ similar to a generalized linear mixed model (GLMM) with a logit link function. Under this framework, logit-transformed Ψ for a group of replicates follows a Gaussian distribution parameterized by a mean and variance. As demonstrated in Chapter 2, this makes rMATS more sensitive to outliers. Additionally, rMATS only quantifies splicing events which follow a classical binary structure; it does not quantify intron retention events, nor does it detect events with three or more splice junctions.

SUPPA (Trincado et al., 2018) quantifies RNA splicing using transcript expression estimates derived from the SALMON pseudoaligner. Differential splicing is then called using an empirical p-value for $\Delta\Psi$ between groups as estimated from transcript abundances. This approach relies heavily on a transcript annotation and accurate TPM estimates, which are hard to infer from short RNA-seq reads. Indeed, the only reads which can be unambiguously mapped to a single transcript are those that span junctions or exons unique to one transcript. Even then, the model does not account for unannotated transcripts, which again are difficult to detect accurately from current-generation sequencing data.

LeafCutter (Li et al., 2018) adopts a novel approach of clustering intron excision events together, then quantifying the distribution of reads along each path through these clusters. This concept is similar in principle to the LSV in that junctions - or introns, in the case of LeafCutter - are jointly considered in modeling and quantifying a given splicing event which may be complex. However, the intron cluster setup complicates the reporting of Ψ and $\Delta\Psi$. Indeed, LeafCutter does not estimate these. Another recent method, Whippet (Sterne-Weiler et al., 2018), implements a framework with elements from both MAJIQ's LSVs and LeafCutter's intron clusters. Briefly, contiguous splice graphs (CSGs) are inferred from the transcriptome and reads are mapped directly to the CSGs using STAR. The CSGs are broken down into alternative splicing (AS) graphs which represent all paths through the same event. Ψ for each exon in the AS graph is estimated from the total number of junction-mapped reads including and skipping the exon. In principle, this framework addresses the open question of how LSVs describing the same splicing event should be combined. However, the software estimates group Ψ by first collapsing all experiments in a group and estimates the ratio of total reads. By doing so, Whippet is unable to model within-sample or between-sample-with-group variance.

In order to account for the demands of large heterogeneous datasets for splicing quantification, a new method is necessary. Such a method should be able to efficiently handle the volume of samples in these datasets quickly and with a relatively small memory footprint.

Additionally, this method should give each sample an equal say in what the group Ψ quantification should be, so that the group result is not dominated by a handful of high-depth experiments. Moreover, it should take into account individual-level uncertainty in Ψ, perhaps by bootstrapping from a distribution estimated for each sample. Finally, the method should employ some robust statistic to represent the difference in Ψ between condition groups, to account for unknown confounding factors[1]. This chapter discusses an algorithm which has these qualities, and the implementation of this algorithm as "MAJIQ-HET".

3.2. Algorithm

The proposed algorithm builds on the existing Ψ quantification method in MAJIQ. Consider a dataset D with two condition groups T_1 and T_2. Let Ψ_{LSV} be the Ψ for each junction in the given LSV, and p_{LSV} be the p-value derived from a robust statistic *test* over the difference in Ψ_{LSV} between T_1 and T_2. After building the splicegraph S_D for D and associating all uniquely-mapped junction-spanning reads with LSVs in S, the following algorithm is run:

1: **for** LSV $\in S$ **do**
2: **for** $j \in range(1, 100)$ **do**
3: **for** experiment $t_i \in D$ **do**
4: Sample $\psi_i \sim P(\Psi_{\text{LSV}} \mid reads(t_i))$
5: **end for**
6: $p_j \leftarrow test(\{\psi_i\}_{i \in T_1}, \{\psi_i\}_{i \in T_2})$
7: **end for**
8: $p_{\text{LSV}} \leftarrow percentile(\{p_j\}, CL)$ [a]
9: **end for**

Algorithm 1: MAJIQ-HET algorithm

[a]The confidence level (CL) for the p-value distribution is a hyperparameter of the model. In the current implementation, this is fixed at 95%.

Algorithm 1 is implemented as MAJIQ-HET in the 2.0 release of MAJIQ, along with several robust statistical tests that satisfy line 6. The p-value estimation for these tests takes

[1]Heterogeneous datasets are very likely to have both known and unkown factors which are confounded with LSV Ψ. This chapter assumes that the input data have been cleaned i.e. to remove known confounding factors, so that at worst only latent factors remain, and a robust statistic is desired on top of that.

into account the size of each condition group. Note that the above algorithm satisfies the specifications laid out in Section 3.1. First, MAJIQ filters the total set of LSVs in the splicegraph to allow only those with, by default, at least 10 reads spanning 3 positions across each splice junction. Samples which do not meet this criteria for a given LSV are treated as missing values in the algorithm; they are removed from consideration in testing that LSV, and the effective size of each group in the statistical test is reduced to account for this. By design, the p-value distributions update to reflect the reduction in samples, so control of p-value confidence is built in to the algorithm. Furthermore, every quantifiable experiment counts equally in the HET evaluation. However, experiments with more reads mapping to the LSV will have greater confidence in its estimate of Ψ. The lower sampling variance from these experiments, compared to experiments with few reads mapping uniquely to the LSV, naturally reflects this.

The current implementation uses three rank-based test statistics: TNOM, InfoScore, and Wilcoxon Rank-Sums.[2] TNOM (Total Number Of Mistakes), which was previously implemented in ScoreGene (Kaminski and Friedman, 2002), tests the separability of two-class, one-dimensional data using a single threshold value to discriminate the two classes. For the current LSV, the set of labeled samples ψ_i is sorted by Ψ. For each possible threshold, every experiment with ψ_i to the left of that threshold is classified in one group, and everything to the right in the other. The number of classification errors is counted for both possible group assignments, and the lower of the two is the score for that threshold. The threshold with the lowest score is selected, and its score is the TNOM statistic for that set of samples. The p-value for that score follows an exact distribution whose parameters are the sizes of the true condition groups, and it can be computed efficitently using dynamic programming as described in Kaminski and Friedman (2002).

InfoScore is constructed similarly to TNOM. Again, labeled ψ_i samples are sorted by value, an optimal threshold is selected, and an exact p-value is computed using dynamic program-

[2]We also implemented the Student's t test, which is a classical parametric test intended as a benchmark. However, the public release of MAJIQ-HET does not include this test.

ming. This time, the score is the mutual information content between the threshold and the true labels, rather than the total classification error. This can lead to subtle differences in what the two tests consider to be a significant change. Finally, the Wilcoxon rank sums statistic is a standard rank-based statistic wherein each ψ_i sample is ranked by value, and the statistic is the sum of the ranks of the positively-labeled samples i.e. all the experiments from condition group 1.

3.2.1. Behavior on toy data

To illustrate how these tests might behave on real data, consider the toy dataset depicted in Figure 13a. Here each colored group has seven Ψ samples, with no clear overlap in their distributions. However, one red sample is an extreme outlier deep within the blue group. While there is a clear overall shift in Ψ between the two groups, a t-test for example would not consider this to be significant as it is sensitive to the outlier. TNOM and InfoScore, meanwhile, would choose an optimal split between the two groups, which intuitively is the most informative split. In this example, the best classification labels everything to the left of the split "blue", and everything to the right "red", mislabeling only one true red sample. The Wilcoxon rank-sums test would be slightly affected by the outlier as its rank would contribute significantly to pull the statistic towards the center of its distribution. However, note what happens in Figure 13b. Relative to a, the samples in b are shuffled slightly such that the rank sum is still the same, but the best TNOM is now much higher (3 instead of 1). In this scenario, we find that Wilcoxon rank-sums is robust to permutations of the data.

Figure 13: Toy example illustrating the rank-based statistics implemented in MAJIQ-HET.

3.3. Evaluations

3.3.1. Reproducibility in GTEx

To test the robustness of MAJIQ-HET in comparison to traiditonal $\Delta\Psi$, RNA-seq experiments were randomly selected from two tissues (Cerebellum and Skeletal Muscle) in GTEx v7, with up to 50 experiments from each tissue. MAJIQ $\Delta\Psi$ and MAJIQ-HET were then run on the comparison between the two groups. The results are depicted in Figure 14. In all tests, the number of events detected drops as sample size increases, but the reproducibility of those events goes up. Moreover, the HET tests show a significant improvement in reproducibility ratio as sample size increases, and consistently detect more events as changing compared to traditional $\Delta\Psi$. Unfortunately, the RR at 10-vs-10 is substantially lower for HET than for $\Delta\Psi$, highlighting the limitations of this approach on smaller comparisons.

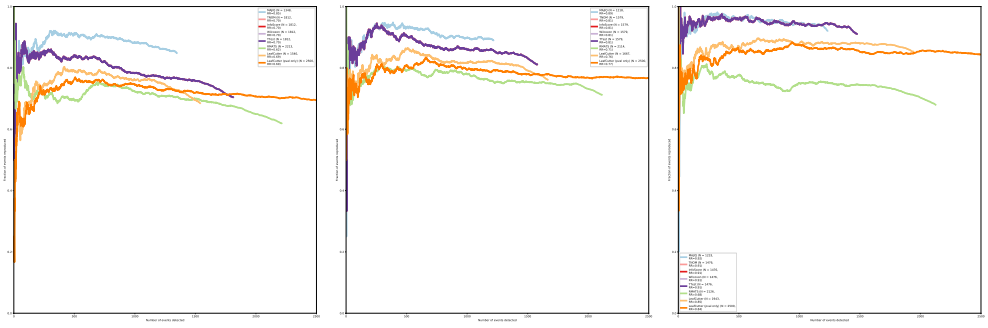


Figure 14: Reproducibility ratio of MAJIQ, MAJIQ-HET statistics, rMATS, and LeafCutter on comparisons of 10, 15, and 50 cerebellum samples with an equal number of muscle samples from GTEx. Events are ranked by expected $\Delta\Psi$ for each method.

3.3.2. IIR in GTEx

High RR can indicate strong consistency in a method as well as high bias. This bias would appear in a within-group comparison. To determine which is the more likely cause of the high RR observed above, IIR was computed for the same sample sets. Figure 18 shows a reduction in IIR as sample size increases, reaching 0 for all tests in a 50-vs-50 comparison. As before, the HET tests performed worse than $\Delta\Psi$ on the smaller comparisons.

We also compared the performance of MAJIQ $\Delta\Psi$ with that of MAJIQ-HET (Figure 16), rMATS and LeafCutter (Figure 17) on smaller subsets of GTEx comparisons, ranging from

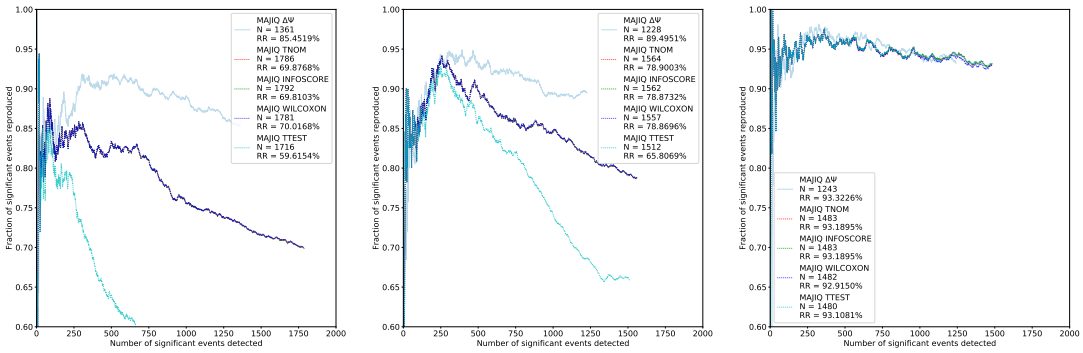


Figure 15: Reproducibility ratio of MAJIQ and MAJIQ-HET on comparisons of 10, 15, and 50 cerebellum samples with an equal number of muscle samples from GTEx. Here, MAJIQ-HET results are ranked by p-value, and MAJIQ results are ranked by $1 - P(|\Delta\Psi| \geq 20\%)$, rather than expected $\Delta\Psi$.

groups of 3 to 9. We note that the $N_{nosignal}$ for TNOM and InfoScore are both 0 for the 3-vs-3 comparison. This is caused not by the robustness of these statistics but the limitations of the p-value distribution for these tests: with 3 positive and 3 negative samples, the p-value for a perfect split (TNOM score of 0, InfoScore equal to the original data entropy), well above the significance cutoff of 0.05. Overall, the sensitivity of traditional MAJIQ on these small comparison groups was lower than that of rMATS and LeafCutter, in agreement with what was observed in Chapter 2. Interestingly, traditional MAJIQ is also less sensitive to within-group variations than MAJIQ-HET. While this may seem counterintuitive, we reiterate that HET is designed specifically for comparisons between large heterogeneous sample groups. This is highlighted in both the high RR and low IIR in the 50-vs-50 comparisons presented above.

We included the Student's T test in these comparisons as a benchmark for evaluating the rank-based tests. On these small comparisons, T-test gives the worst IIR, shortly behind Wilcoxon.

3.3.3. Overlaps between tests

To illustrate the intuition described in Subsection 3.2.1, upsets were computed between $\Delta\Psi$ and the HET tests for each comparison size. Figure 19 and Figure 20 show that in general, the majority of events called significant by $\Delta\Psi$ were also captured by all the HET tests.

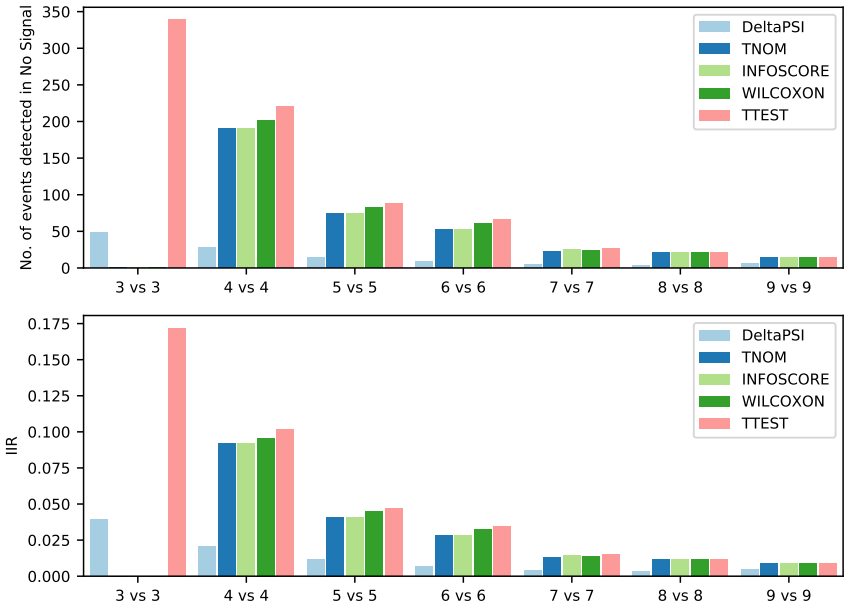


Figure 16: Robustness of MAJIQ Delta PSI to within-condition splicing changes in comparison with MAJIQ-HET on small subsets of GTEx. Top: Number of events detected significantly changing in a between-tissue comparison; Bottom: IIR.

The remainder is divided mainly between events called only by $\Delta\Psi$, and events called by all the HET tests but not by $\Delta\Psi$.

3.4. Simulated data

To determine whether MAJIQ-HET is able to recover true changes in splicing, we simulated RNA-seq experiments based on 120 donor samples from GTEx. Of these, 60 samples originated from skeletal muscle; the rest, from a collection of 12 brain subregions represented in GTEx. Transcript-level expression was estimated per gene by using the most complex LSV built from the annotation alone as a proxy for transcriptome complexity. These expression levels were input into BEERS (Baruzzo et al., 2016) to generate RNA-seq reads, which were mapped back to the hg19 reference genome using STAR-2.5.3a (Dobin et al., 2013). MAJIQ and MAJIQ-HET were then run to quantify splicing changes between varying sized groups of brain and muscle samples, and events called significant by these methods were used to compute statistics of sensitivity, specificity, false discovery rate, and false negative rate. We found that MAJIQ-HET had lower FDR and FNR on most comparisons, albeit

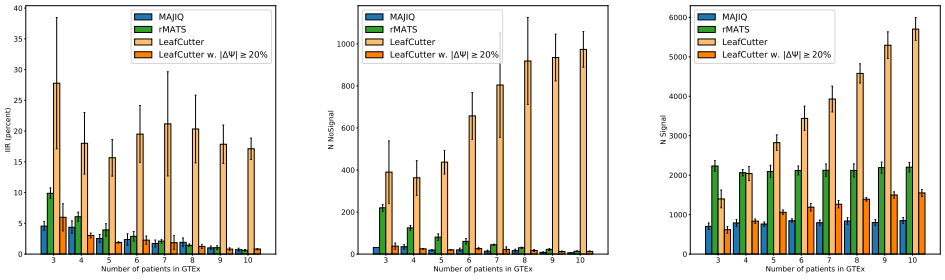


Figure 17: Performance of MAJIQ Delta PSI on small subsets of GTEx in comparison with rMATS and LeafCutter. Left: IIR; Middle: Number of events detected significantly changing between subsets of the same tissue; Right: Number of events detected significantly changing in a between-tissue comparison.

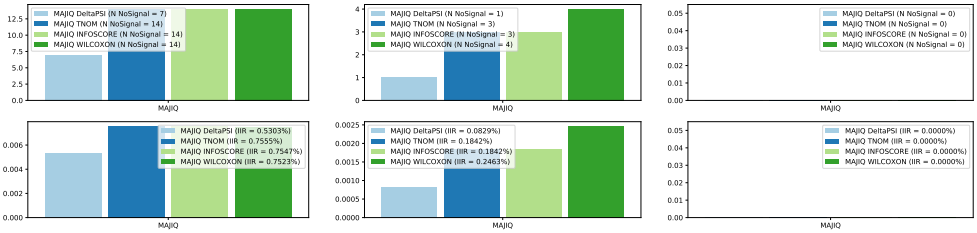


Figure 18: Intra-to-inter ratio of MAJIQ and MAJIQ-HET on comparisons of 10, 15, and 50 cerebellum samples with an equal number of muscle samples from GTEx. The no signal group is two disjoint sets of cerebellum samples.

with a slight drop in sensitivity (Table 1). Since earlier tests had shown improved performance when stipulating that all samples must be quantifiable for Ψ, we tested performance on simulated data under that constraint. We found that MAJIQ-HET retains its improved FDR and FNR while also showing a slight boost in sensitivity relative to traditional MAJIQ (Table 2).

3.5. Discussion and conclusions

MAJIQ-HET implements a suite of robust statistical tests for calling differences in splicing between large heterogeneous sample groups. By and large, the HET tests all agree on what events are changing significantly. This makes sense, as the non-parametric rank-based statistics are very similar in their construction and assumptions. The interesting cases are where they do not agree. In particular, the subset of events called significant by the HET tests but not by $\Delta\Psi$ could be explained as a difference in how high-depth experiments

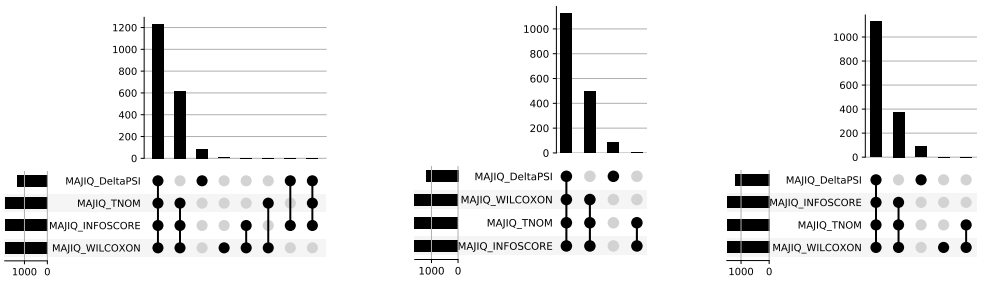


Figure 19: Upsets of MAJIQ and MAJIQ-HET on comparisons of 10, 15, and 50 cerebellum samples with an equal number of muscle samples from GTEx.

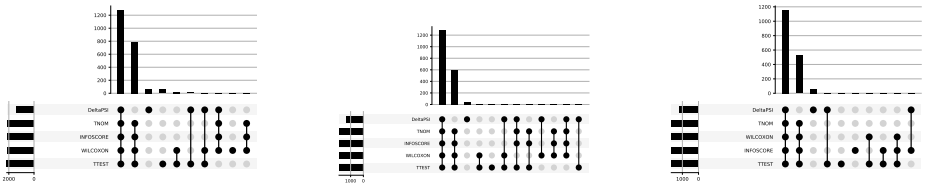


Figure 20: Upsets of MAJIQ and MAJIQ-HET on comparisons of 4, 6, and 9 cerebellum samples with an equal number of muscle samples from GTEx.

are handled. While an outlier with an overabundance of reads supporting the LSV would dominate the quantification of $\Delta\Psi$, the HET algorithm disallows this.

MAJIQ-HET is designed specifically to handle the challenges presented by large heterogeneous datasets. Its reproducibility and putative false positive rate are noticeably worse than that of traditional $\Delta\Psi$ quantification on small sample sets. This is not unexpected, as the nonparametric rank-based tests rely on a large sample size to reach significance even on large changes in Ψ. Traditional $\Delta\Psi$ is very well-equipped to handle small replicate comparisons, as evidenced in this and the previous chapter.

One of the underlying assumptions of the HET model was that the input data were adjusted for known and latent covariates prior to entering the pipeline. However, there are cases where the known confounding factors are study relevant. For example, chromatin features such as SNP genotype and epigenetic modifications may affect splice site selection at a nearby alternative event. We consider SNP genotype as an informative feature for splice

junction inclusion in Chapter 4.

CHAPTER 4 : Genotype-splicing associations

4.1. Introduction

4.1.1. Genome-wide association studies and QTL studies

Modern advances in genomics technologies facilitate discovery of genetic and epigenetic predictors of phenotype. This typically takes the form of a genome-wide association study (GWAS), in which a group of subjects with, for instance, a particular disease phenotype (cases) and a separate healthy cohort (controls) are genotyped at various loci throughout the genome, and allelic variants are tested for enrichment in the cases vs. the controls using a hypergeometric test. The first successful GWAS identified a single nucleotide polymorphism (SNP) in an intronic region where the G allele is enriched in myocardial infarction patients compared to the A allele (Ozaki et al., 2002).

As the cost of sequencing has gone down and the availability and quality of data has improved, it has become easier to measure loci that covary with quantitative traits (quantitative trait loci, or QTLs) such as BMI, blood pressure, and gene expression. The genotype tissue expression (GTEx) (GTEx Consortium et al., 2017) project is an ongoing effort to catalogue all human QTLs for tissue-specific gene expression (expression QTLs, or eQTLs). As of the v7 release in 2017, the project has published 11688 RNA-seq samples covering 53 different body sites from 635 independent donors. Of these, eQTLs have been called for 10294 RNA-seq samples in 48 body sites from 620 donors. However, the eQTL calls are based on gene-level expression quantifications, and do not reflect expression at the isoform level. As explained in the preceding chapters, splicing differences between conditions affect transcriptomic and proteomic structure, and have regulatory consequences besides. In addition, the regulation of alternative splicing is controlled by different processes and mechanisms as explained in Chapter 1. These differences motivate the testing of patient samples for splicing QTL (sQTL) in addition to the existing eQTL screens.

An sQTL is a genomic variant at which genotype correlates with some metric of alternative

splicing at a given gene locus. sQTLs are typically called by imputing, transforming, and normalizing splicing quantifications (either exon percent spliced in (PSI, Ψ) or relative isoform abundance) and calling QTLs using tools such as Matrix eQTL (Shabalin, 2012) or fastQTL (Ongen et al., 2016). LeafCutter in particular was designed to quantify RNA alternative splicing in a format compatible with both Matrix eQTL and fastQTL (Li et al., 2018). While fast and powerful, these methods do not quantify *de-novo* intron retention events or, in the case of rMATS (Shen et al., 2014), *de-novo* exon splicing events. Both of these features are implemented in MAJIQ, which has similar or better performance to rMATS and LeafCutter in evaluation testing. In addition, the commonly-used matrix-based methods for QTL calling handle missing values in Ψ by imputation. While this strategy makes sense for expression data, it may not be appropraite for splicing quantification. MAJIQ handles missing read data by refusing to quantify splicing events that fail to meet a minimum coverage threshold. A splicing event can be unquantifiable for a number of reasons: either the gene is not expressed in that sample, or the transcripts that are expressed skip the region containing the LSV in question. This behavior is propagated in the MAJIQ-HET workflow as described in Chapter 3 as dropout in the comparison groups.

With this in mind, we implemented an sQTL pipeline that uses MAJIQ Ψ quantifications as input. The sQTL pipeline is constructed as such:

1: **for** LSV $\in S$ **do**
2: Select junction Ψ from experiments with sufficient read support
3: Logit-transform Ψ
4: **for** SNP near LSV **do**
5: Impute genotype for quantifiable samples
6: Normalize Ψ and genotype to a Gaussian
7: Regress out known and inferred covariate vectors
8: Calculate test statistic for association (linear regression, F-test, generalized TNOM)
9: **end for**
10: **end for**
11: Correct p-values for FDR or FWER

Algorithm 2: MAJIQ-sQTL pipeline

Step 5 controls for missing values in Ψ in a manner similar to that implemented in MAJIQ-HET in the previous chapter. Briefly, if there are not enough reads to quantify Ψ in that sample, it is discarded from that analysis rather than imputed.

We exercised this pipeline in two studies investigating genetic effects on transcriptome phenotypes. The first study was a collaboration with the lab of Dr. Dan Rader[1], where we evaluated variants that were previously implicated by GWAS in cardiovascular disease to test for splicing regulation. Additionally, we performed a genome-wide assessment of genotype-splicing associations and measure the reproducibility of discovered sQTLs using an in-house patient cohort supplemented with related tissue samples from GTEx. The second study, spearheaded by the GenR project at Erasmus Medical Center (EMC) in Rotterdam in collaboration with Dr. Struan Grant[2], followed up on a GWAS metaanalysis for skeletal development timing in children. The sentinel SNP from the combined study, rs6410 ($p = 1.1 \times 10^{-11}$), is a synonymous variant positioned 15 base positions upstream of the 3' end of *CYP11B1* exon 1. *CYP11B1* is an adrenal-specific cytochrome p450 gene with

[1]Department of Medicine, Perelman School of Medicine
[2]Department of Pediatrics, Children's Hospital of Philadelphia

high conservation across mammals. Multiple alternative splice isoforms are annotated for *CYP11B1*, however they are not well-documented in the literature. One isoform alternatively includes cassette exons 2 and 4, which are in close physical proximity with rs6410. As exon 2 is an annotated cassette exon in *CYP11B1*, we suspected that rs6410 (or a variant in strong linkage disequilibrium with rs6410) could be acting as a splicing QTL. This suspicion was supported both by our sQTL analysis on GTEx and by our RT-PCR analysis on two separate donor cohorts (Figure 26, Figure 31).

The MAJIQ sQTL pipeline, all the analyses that were performed using that pipeline, and the *CYP11B1* RT-PCR analysis constitute my contributions to these projects.

4.2. Cardiovascular disease

Coronary artery disease (CAD) is a complex disease with an estimated 40-60% heritable component (McPherson Ruth and Tybjaerg-Hansen Anne, 2016). GWAS meta-analyses have reported over 100 genomic loci that are significantly-associated with CAD, with several also implicated in risk for other cardiovascular phenotypes (the CARDIoGRAMplusC4D Consortium et al., 2015; Nelson et al., 2017; Klarin et al., 2017; van der Harst Pim and Verweij Niek, 2018; Tada Hayato et al., 2014; Dichgans Martin et al., 2014; Pickrell et al., 2016; Chasman Daniel I. and Lawler Patrick R., 2017). These identified variants are non-coding SNPs whose gene regulatory targets are not yet known. While these GWAS significant SNPs have been tested for eQTL behavior in vascular cells, the tissues directly affected by CAD are heterogeneous, categorized by multiple cell types (Brænne Ingrid et al., 2015; Zhao Yuqi et al., 2016; Franzén et al., 2016). Gene expression and alternative splicing vary between cell types, including those that are closely related, highlighting the need for cell-type specific analyses. To that end, we profiled epithelial and smooth muscle coronary artery tissue samples from a small patient cohort for eQTLs and sQTLs. We then replicated identified eQTL and sQTL variants in GTEx using related tissue RNA-seq (Nürnberg et al., 2017, in review).

4.2.1. Methods

RNA was collected and sequenced from coronary artery epithelial (EC) and smooth muscle (SMC) cells obtained from 18 patients. Of these, thirteen patients had matching Illumina SNP genotype calls passing quality control. These data were processed with respect to the hg19/GRCh37 reference genome (Zerbino et al., 2018, Ensembl v75). Empirical covariates were inferred from genotype principal components only. Tissue-specific sQTLs were called according to the procedure described in Algorithm 2. A SNP was declared to be a significant sQTL for an LSV junction of the corrected p-value of association was less than 0.05. Additionally, genes reaching nominal significance for any LSV association ($p < 0.05$) were followed up in GTEx artery tissue types: aorta (N = 245), coronary artery (N = 140), and tibial artery (N = 353). As before, the RNA-seq and genotype calls from these GTEx samples were mapped to the human reference genome hg19/GRCh37. Both expression and genotype covariates were inferred. The same sQTL procedure was performed for each GTEx tissue on the subset of genes reaching nominal significance in the SMC/EC cohort, and p-values were corrected for multiple hypothesis testing as described. For both datasets, we additionally screened for colocalization between putative sQTLs and variants implicated by GWAS for seven different cardiovascular maladies, including migraine, coronary artery disease, and abdominal aortic aneurysm. The GWAS sentinel variants were considered along with every variant in strong linkage disequilibrium therewith ($\rho^2 \geq 0.80$ in GTEx).

4.2.2. Results

To examine the effect of GWAS loci for vascular disease on the relative abundance of RNA splice isoforms, we performed a genome-wide screen for sQTLs in HCAECs and HCASMCs. Splicing analysis was done using MAJIQ, which quantifies local splicing variations (LSVs) as percent spliced in (PSI) of alternatively-spliced mRNA segments. We identified 478 SNPs in 196 genes (SMC) and 1,028 SNPs in 359 genes (EC) which were nominally associated (p¡0.05) and passed 0.05 FDR correction at the gene level. Combined, these lists included 1399 unique SNPs in 512 genes. Next, we took 3,844 unique genes with at least one nominally significant sQTL in SMC or EC and tested those for sQTL using GTEx artery tissues.

Of those, 33057 SNPS in 3310 genes were validated in at least one GTEx tissue at FDR ¡ 0.05. We compared these lists with the set of SNPs and genes with putative sQTLs passing 0.05 FDR correction at the gene level in SMC and EC. 924 SNPs in 471 genes were reproduced in GTEx, whereas 475 SNPs in 41 genes were unique to the two cell types (Figure 21). Finally, we also performed genome wide sQTL analysis for the three artery GTEX tissues and identified 54298 unique SNPs in 7965 genes that passed 0.05 FDR correction at the gene level.

The vast majority of GWAS-associated loci for vascular phenotypes have not been functionally annotated. Colocalization analysis combines two different data sets to see if related phenotypes share genetic variants. If a SNP or signal colocalizes between two phenotypes (e.g. disease and changes of expression of nearby gene), then there is greater confidence that the variant may be relevant to disease. We therefore queried for overlap between published GWAS loci for vascular disease and sQTL loci identified above. First, for sQTL loci identified from HCASMCs and HCAECs with GWAS loci, we found one SNP at the TARS2 gene locus and 4 SNPs in 3 genes (YAP1, CFDP1, and STAT6) for HCASMCs and HCAECs, respectively, that passed 0.05 FDR correction at the gene level. All of these variants are in linkage disequilibrium (LD ≥ 0.8) with sentinel SNPs for genome-wide association with migraine ($p < 5 \times 10^{-8}$). Of these, rs167769 is both a sentinel variant for association with migraine and an sQTL associated with an alternative 5' splice site in the first exon of five of the six annotated transcripts of STAT6. This splicing variation affects the 5' UTR (sQTL FDR = 0.047) and accounts for approximately 7% increase in inclusion of the 18-nt extension. Second, from the GTEx genome wide sQTL analysis for artery tissues described above, we identified 38 SNPs in 12 genes, 17 SNPs in 7 genes, and 38 SNPs in 11 genes, in Aorta, Coronary, and Tibial artery respectively which overlapped GWAS SNPs or those in strong LD with those. Among these, 20 SNPs in 5 genes, 10 SNPs in 6 genes, and 29 SNPs in 8 genes, respectively, were the sentinel variants for their respective disease association studies. Of note, rs324011 is a significant sQTL for STAT6 in all three GTEx tissue types (sQTL FDR = 0.00351 in Aorta, 0.00349 in Coronary artery, and 0.00271 in Tibial artery).

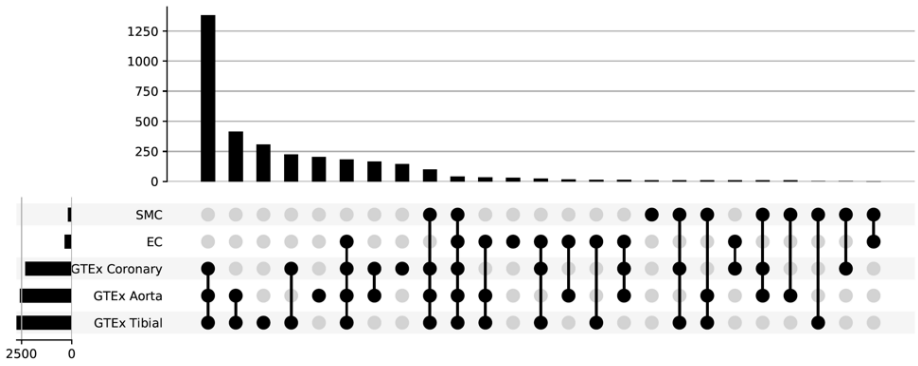


Figure 21: Upset of genes with sQTLs in the HCASMC (SMC), HCAEC (EC), and GTEx datasets (FDR ¡ 0.05, regression test). Vertical bars represent the count of unique genes per set. Below the bar graphs, each dot represents a dataset and intersecting sets are represented by lines connecting dots. Horizontal bars represent the total number of genes with putative sQTLs in each dataset.

This variant is in strong LD with the rs167769 discussed above (LD score = 0.943414), which associates with the same splicing variation in HCAEC but also in Aorta and Tibial artery. In addition, rs324011 was identified as a nominal eQTL for STAT6 in all three GTEx tissue types (eQTL p = 0.00139 in Aorta, 0.0453 in Coronary artery, and 4.17x10 -5 in Tibial artery). This overlap between sQTL and eQTL was observed for several of the aforementioned SNPs and may point to mechanistic connections such as splicing induced frameshifts that lead to nonsense mediated decay and result in lowered expression of genes.

For more details, see Table 3

4.3. Skeletal growth in children is GWAS-mapped to a locus physically located near an alternative splice site

Abnormal secretion of hormones from the adrenal gland can have adverse effects throughout the body. For example, Figure 25 depicts a pediatric case of adrenocortical carcinoma where the tumor secretes abnormal levels of sex hormones, triggering early onset of puberty. Among other phenotypes exhibited by this patient, one quantitative measure of this effect is the rate of skeletal development. This is readily measured by comparing an x-ray image of the child's hand to a set of age-based standards. In this case, the patient is 6 years of age, but the hand x-ray most closely matches the 10-year-old standard.

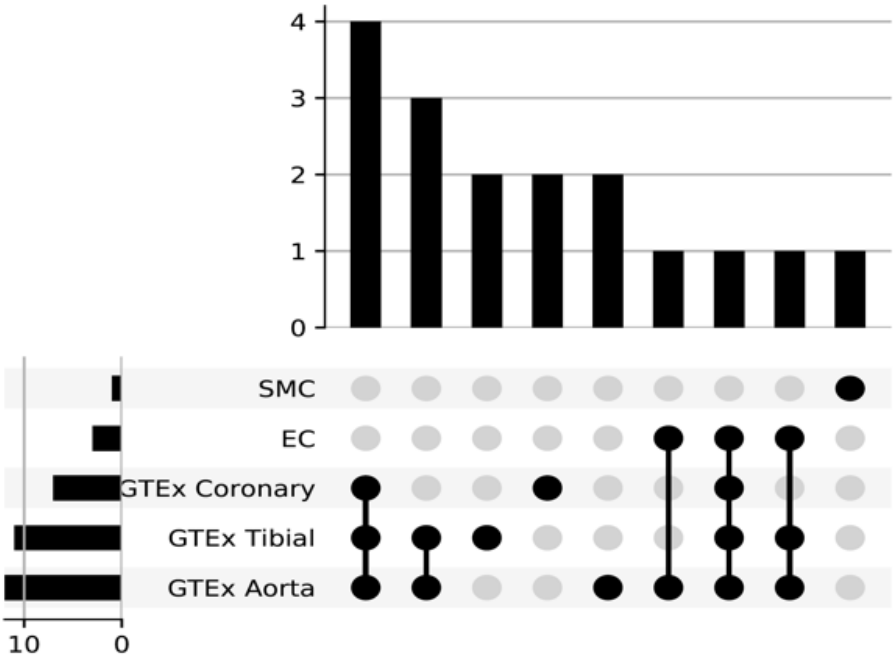


Figure 22: Upset of genes with sQTLs in the HCASMC/HCAEC and GTEx datasets that colocalize with any GWAS signal for association with cardiovascular disease.

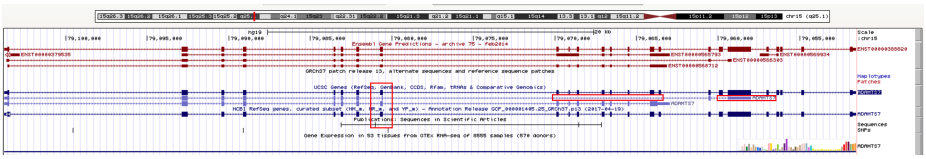


Figure 23: Diagram of the *ADAMTS7* locus in the UCSC genome browser, showing the exon structure. Outlined regions show the alternative 3' splice site (left), the tandem skipped exons (middle), and the 3' variation (right) on this isoform.

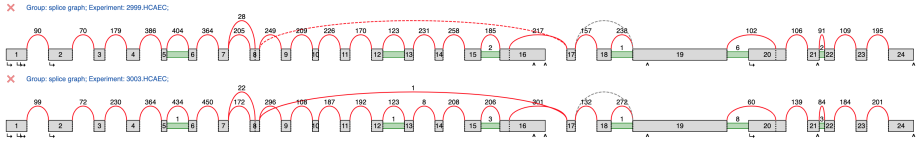


Figure 24: Splicegraph of *ADAMTS7* built from two representative samples in the in-house human coronary artery cohort. In all donors and both tissue types, few reads were found to map to the downstream 3' splice site in exon 8, and none were found to map to the junction skipping exons 9-16.

The trait of unusual skeletal growth rate is not limited to cancer patients, and may have a genetic underpinning. A genome-wide association study on 3510 children performed by the GenR project found a significantly associated variant in the coding region of *CYP11B1*, a cytochrome P450-like locus on chr8 almost exclusively expressed in the adrenal gland ($\beta = 0.15, p = 2.8 \times 10^{-10}$). This association was validated in a meta-analysis incorporating 1048 additional patients at the Children's Hospital of Philadelphia ($\beta = 0.14, p = 1.1 \times 10^{-11}$). The particular variant, rs6410, is a synonymous SNP located a few bases upstream of the 3' end of *CYP11B1* exon 1. This is particularly interesting because this locus has two annotated isoforms, with exons 2 and 4 being alternatively-included cassette exons. This suggests that rs6410, or a variant linked to it, may act as an sQTL. RNA-seq expression analysis published in GTEx v7 shows this locus to be expressed only in adrenal gland (median RPKM = 2262.1) and testis (GTEx Consortium et al., 2017, median RPKM = 2.8). Therefore, we sought to test whether LSV Ψ is associated with allelic variants in the vicinity of *CYP11B1*. We pursued two lines of evidence to this end. First, we employed our pipeline for sQTL discovery against all genomic variants at the *CYP11B1* locus as observed in GTEx adrenal gland samples. We then validated the findings from this pipeline using RT-PCR on two independent batches of patient adrenal samples.

4.3.1. Methods

sQTL screen

159 paired-end RNA-seq experiments from GTEx v6p adrenal donors were mapped to the hg19 genome annotation using STAR 2.5.2a (Dobin et al., 2013) with the option –alignSJoverhangMin 8. Sorted and indexed alignments were built into splicegraphs by MAJIQ 2.0(Vaquero-Garcia et al., 2016) build using only reads mapping entirely within the *CYP11B1* locus. LSV junctions were incorporated into the splicegraph if they were supported by at least 80 samples with a minimum of 3 reads across two start positions per sample. LSV junctions for each sample were then quantified using MAJIQ 2.0 psi, with a minimum of 10 reads across 3 start positions required for an LSV to be quantifiable in the sample.

We queried 1357 genomic variants within the *CYP11B1* locus as the sQTL search space. To limit the number of statistical tests performed, our pipeline imposes strict filters for quantifiability of LSV-variant associations. In brief, tested variants must have a minimum minor allele count of 5 and minimum minor allele frequency of 0.1 among the set of samples for which junction PSI was quantifiable. Consequently, a total 570 of LSV-SNP pairs were tested for significant association. For each LSV, missing genotypes were imputed from the set of PSI-quantifiable samples. PSI values were logit-transformed and quantile-normalized to a standard normal distribution. We used the subject's gender as a known potential confounding factor, and supplemented inferred latent covariates from genotype principal components and expression matrix factors generated by PEERS and published by GTEx. These confounding factors were regressed out from both the transformed PSI and the imputed genotypes. Finally, we computed the statistic of linear association between the PSI and genotype residuals against the null hypothesis of 0 slope. Association p-values were adjusted using a conservative Bonferroni correction with 392 significant associations involving two local splicing variations (see full list in Table S2).

RT-PCR validations

Total RNA was purified from each of 15 donor adrenal glands, nine supplied by the NIH and six provided by EMC. RNA transcripts were reverse transcribed using poly-dT oligos and random hexamer primers. Primers were designed within the bodies of exons 2 and 3 to amplify the alternative cassette exon 4 (see Table S1). PCR was run for 34 cycles, and amplified cDNA was size separated on a 10% acrylamide gel in TBE. Empirical PSI was estimated from the relative band intensity of the exon 4 inclusion product relative to the exon 4 skipping product, using ImageJ's gel analysis tool to quantify band intensities. The actual quantification was performed using a Python script to estimate the ratio between the background-adjusted height of the UV intensity peak corresponding to the inclusion band (the inclusion peak) and the sum of the inclusion peak and skipping peak heights:

$$\Psi_{empirical} = \frac{H_{265}}{H_{265} + H_{187}}.$$

In order to test for intron retention, additional primers were designed flanking the genomic boundary between exon 3 and the downstream intron. Amplified cDNA was size separated on a 10% agarose gel in TBE, and the observed bands were used to qualify presence or absence of the retained intron.

4.3.2. Results

We hypothesized rs6410 (chr8:143961005) is associated with changes of expression or isoform usage of CYP111B1 or nearby genes. GTEx v7p reported eQTLs for rs6410 included only the expression of *CYP11B1* in the adrenal gland (p = 0.00203484, beta = 0.133608 ± 0.0424836), an association that does not pass significance when multiple SNPs in the region are considered. In contrast we found several strong associations of rs6410 and other SNPs with splicing variations, i.e. sQTL.

In order to identify sQTLs at *CYP11B1*, we used 159 patient adrenal glands in GTEx v6p. RNA-seq experiments were mapped to the hg19 reference genome using STAR-2.5.2a, and inclusion levels of alternative splice junctions were quantified using MAJIQ 2.0 followed by an sQTL quantification pipeline (see Methods). This sQTL pipeline identified a total of 392 significant associations (see full list in Table S2) involving two local splicing variations (LSV). The first LSV involves a cassette exon downstream of the second exon of the canonical isoform (exon 4, chr8:143959172-143959250) (Figure 26). For this event, the sentinel GWAS hit of rs6410 achieves strong association ($p = 7.90 \times 10^{-16}$ with the skipping junction, $p = 1.30 \times 10^{-15}$ with the inclusion junction), though the strongest association is for rs10956995 ($p = 1.16 \times 10^{-23}$ with the inclusion junction). For rs6410 we find the common T allele is associated with skipping of the alternative 4 exon. Notably, the raw RNA-seq alignments clearly support this sQTL but also shows reads spanning the introns flanking exon 4 (Figure 27). We note that the measured changed in exon 4 inclusion levels is derived from junction-spanning reads mapped by STAR. The LSV in question is the 3' end of the cassette event, with splicing of exon 4 to exon 5 discernible by the change in read levels at the 3' end of exon 4. This indicates that the splicing events involved may be more complex than just cassette exons and can affect both function as well as expression levels of the gene

isoforms. Another possible explanation for these intronic reads are unannotated antisense transcripts, as the RNA-seq experiments were not strand-specific. These findings are also inline with additional bands observed in the RT-PCR validation experiments (see below).

The second LSV with strong sQTL associations involves an alternative 3' splice site in exon 9 of the canonical transcript isoform of *CYP11B1* (shorter 143956650-143956728 vs. longer 143956650-143956797 on chr 8, see Figure 29). Here the most significant association in GTEx is with rs4736311 (chr8:143952950, $p = 8.09 \times 10^{-26}$ with the shorter exon) while the sentinel GWAS hit rs6410, which is in LD with rs4736311, achieves $p = 5.01 \times 10^{-14}$. For rs6410, the common T allele is associated with the shorter exon 9. Notably, the annotated transcript containing the alternative junction also has its transcription start site at the 5' end of exon 6 (chr8:143957128-143957294). This splice isoform ablates both the mitochondrial localization signal and the substrate binding domain, the coding sequences for which are located upstream of exon 6.

To validate our findings we next performed RT-PCR experiments to quantify exon 4 skipping using RNA from six donors provided by Erasmus Medical Center (EMC). These six samples were distributed with one T/T, three T/C, and two C/C at rs6410. Inclusion levels of exon 4 correlated strongly with genotype at rs6410, and recapitulates the observation made in GTEx RNA-seq (Figure 26c). Evidence of intron retention was also observed via RT-PCR (Figure 30, Figure 28).

Finally, we also repeated the validation of the rs6410 sQTL using a second cohort of 9 donors from an NIH repository. The RT-PCR results from this cohort validates the general trend of inclusion levels of exon 4 (Figure 30) but we had difficulties achieving high enough amplification for those. This issue is likely the result of the donors' micronodular adrenal hyperplasia condition, which is associated with reduced expression of *CYP11B1* (Horvath and Stratakis, 2008).

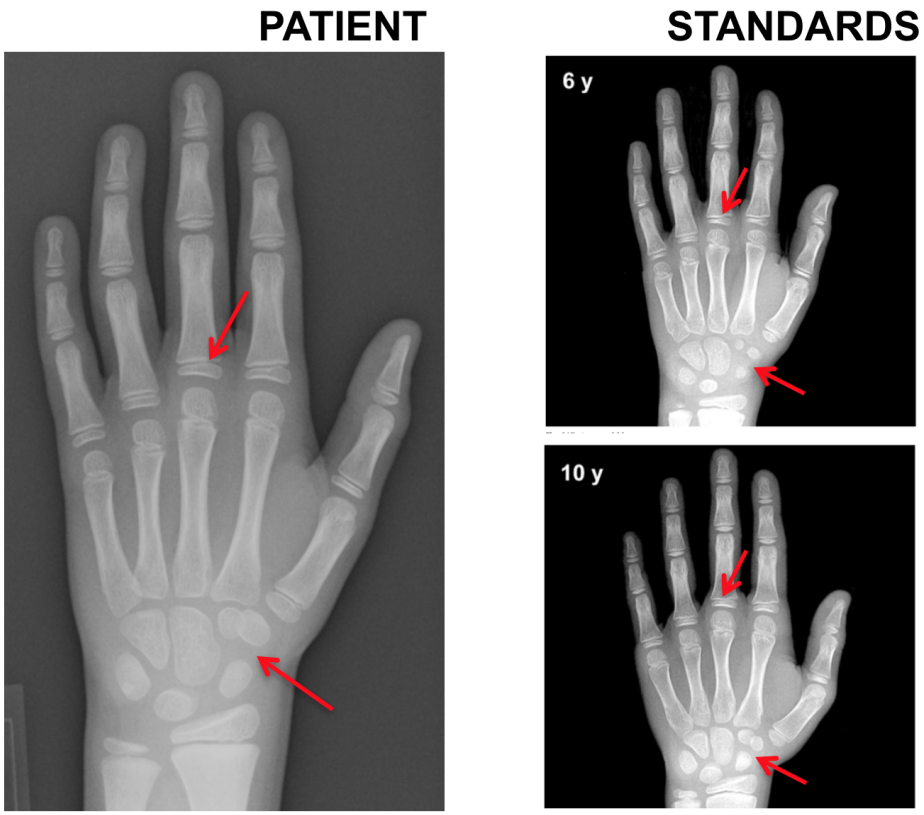


Figure 25: Example case of abnormal skeletal growth. The six-year-old patient presents with a sex-hormone-producing tumor and early onset of puberty. His hand x-ray (left) is morphologically more similar to the 10-year-old standard (bottom right) than it is to the six-year-old standard (top right) (Vicente Gilsanz and Osman Ratib, 2012, pp. 21-25).

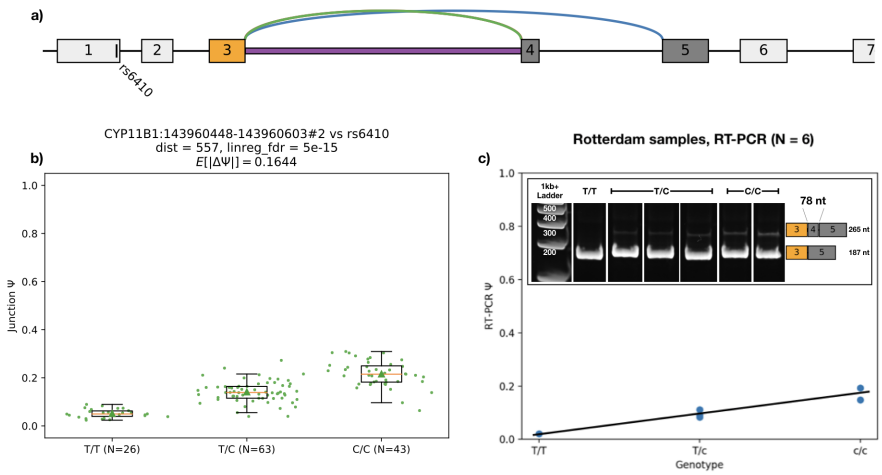


Figure 26: Association between rs6410 genotype and inclusion levels of cassette exon 4. a. Splice graph showing a local splicing variation (LSV) from exon 3 (orange). Blue junction represents skipping of exon 4; green junction represents inclusion of exon 4; and the purple rectangle represents retention of that intron. b. Scatterbox plot of MAJIQ PSI quantifications for the exon 4 inclusion junction (green), stratified by genotype in GTEx. The risk (C) allele is associated with increased levels of exon 4 inclusion. c. RT-PCR validation for the association between rs6410 genotype and exon 4 inclusion using six donor samples from EMC. Scatterplot shows the PSI quantifications for inclusion of exon 4 as computed from analysis of the gel electrophoresis image (inset). The trend matches that observed in the RNA-seq (b).

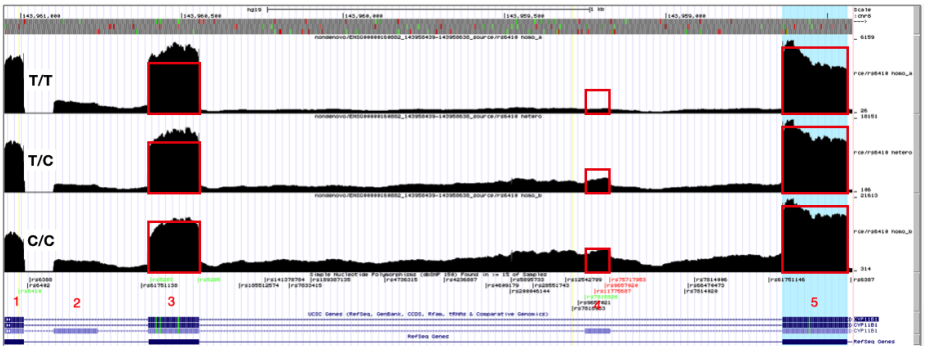


Figure 27: Pileup of aligned reads at the exon 4 skipping event, using a single representative of each genotype at rs6410. Exons 3, 4, and 5 are highlighted in red, with exon number labels printed below the pileups.

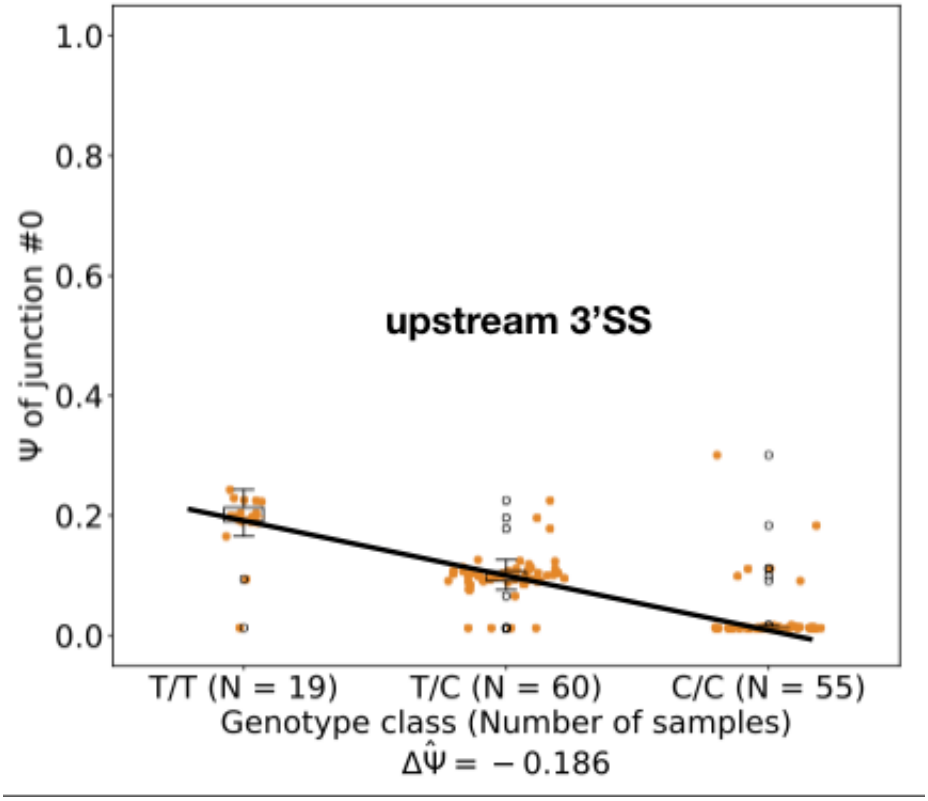


Figure 28: Scatterplot of Ψ for the alternative 3' splice site in *CYP11B1* exon 9, stratified by genotype at rs6392. Trend line is approximate. $p = 1.87 \times 10^{-27}$, F-test.

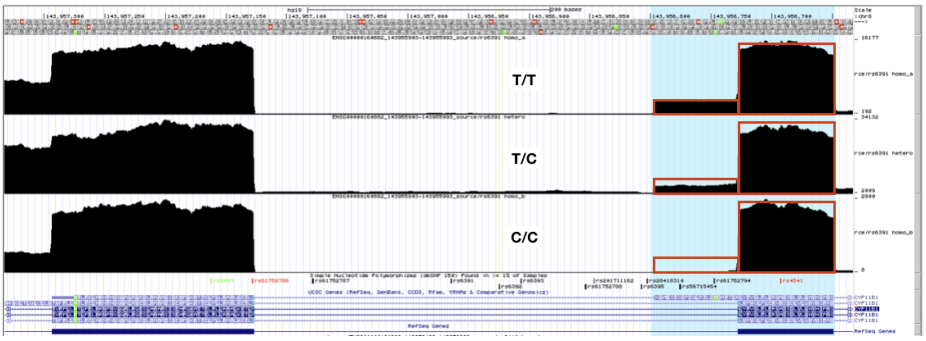


Figure 29: Pileup of aligned reads at the alternative 3' splice site in *CYP11B1* exon 9, using a single representative of each genotype at rs6392. Exons 8 and 9 are highlighted in red.

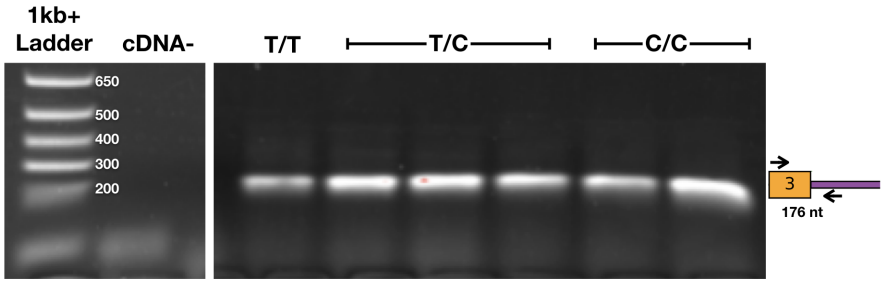


Figure 30: Amplification across the *CYP11B1* E3I3 boundary in the EMC donors, stratified by genotype at rs6410 as indicated across the top.

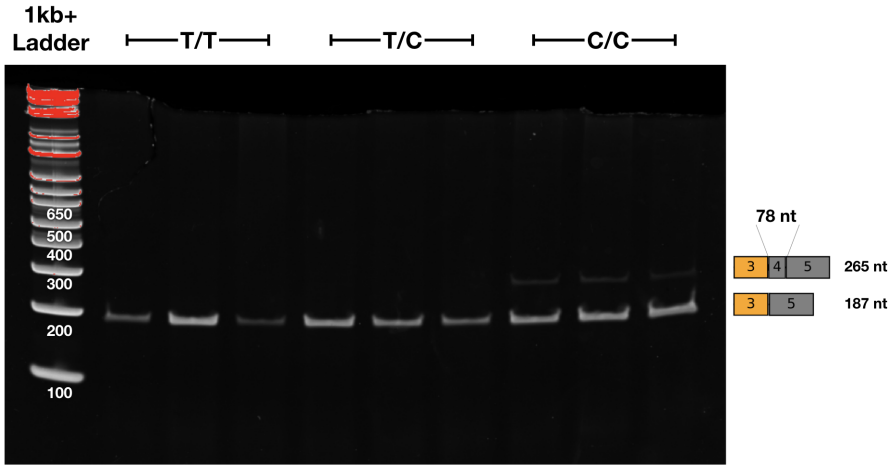


Figure 31: Amplification of the *CYP11B1* alternative exon 4 cassette event in the NIH donor cohort using the same primers as in Figure ??c. Amplification was inconsistent between samples and low overall, frustrating efforts to quantify Ψ.

4.4. Discussion and conclusions

Splicing QTL discovery is an important indicator towards the function of genomic variants. While existing methods are powerful to identify such variants in large cohorts, they often make assumptions and corrections which are more appropriate for expression TPM estimates than local splicing ratios or Ψ. We addressed this in our pipeline in several ways. The foremost contribution is the power and accuracy of MAJIQ in quantifying Ψ. Second, we handle missing values in Ψ by omitting them rather than imputing them. Lastly, we assume a statistical distribution designed to fit Ψ, rather than using a method that assumes a distribuion more suitable for TPMs or FPKMs.

We demonstrated the performance of this method by applying it to screen for splicing associations with variants implicated by GWAS meta-analyses in two disease studies. In the cardiovascular cohort, we uncovered *ADAMTS7* as a putative sQTL target with functional consequences on the transcriptome. We screened these cohorts for eQTLs and identified *ADAMTS7* as an eGene (the gene target of an eQTL) with allele-specific readout in cis. This splicing variation is particularly interesting in that the associated alternative 3' splice site results in a frame shift. The transcriptome annotation links the shorter exon with skipping of the next 9 exons, resolving the frame shift without introduction of a PTC. The alternative open reading frame deletes the majority of the functional motifs present in the canonical isoform. If translated, this isoform would manifest as a loss of function of *ADAMTS7*, offering a possible explanation for the observed association with CAD. However, RNA-seq evidence shows low abundance of reads mapping to this long skipping junction, suggesting 3'-to-5' degradation. Whether this is the result of nonsense-mediated decay or a cytoplasmic procedure is not clear. Proteomics analysis would be necessary to verify translation of this short isoform in risk or affected individuals.

In the bone growth association study, we identified two clusters of sQTLs near *CYP11B1* in strong linkage disequilibrium with each other. One affects inclusion of canonical exon 4, and the other impacts splice site selection in exon 9. While not explicitly considered in the

bioinformatics screen, we observed moderate coverage of the intronic regions flanking exon 4. However, the cDNA preparation for neither the RNA-seq nor RT-PCR experiments was strand-specific. It is possible that an unannotated (possibly non-coding) antisense transcript is responsible for the signal in these introns in both assays. An alternative explanation is that both screens are picking up partially-processed transcripts.

The annotated splicing variation introduces two in-frame cassette exons at the 5' end of the transcript. A robust structural analysis of the resulting protein isoform has not yet been published. However, the structure of the canonical *CYP11B1* dimer in complex with its substrate is available on the Protein Database (PDB) (Brixius-Anderko and Scott, 2018, PDB ID: 6M7X). In brief, the points where the cassette exons would be inserted are on the outer surface of the protein, distal from the steroid binding pocket. Moreover, the exon 2 cassette insertion is downstream from the mitochondrial localization signal present in the canonical isoform. The expected consequence of these cassette inclusions is a gain of function in *CYP11B1*, likely a novel binding motif on either the protein product or the mRNA itself. Such can be probed using CLIP-seq (for discovery of novel RNA-binding proteins) or co-immuniprecipitation followed by mass spectrometry.

Finally, we noted that the NIH cohort of nine donors had a diagnosed condition that likely affected expression of *CYP11B1*. This negatively impacted our ability to amplify *CYP11B1* for quantitative splicing analysis. Moreover, early attempts to quantify Ψ of *CYP11B1* exon 4 from these samples yielded irreproducible observations. The micronodular adrenal hyperplasia diagnosis did not come to light until late in the investigation, by which point we had successfully amplified this region in the EMC samples.

CHAPTER 5 : Conclusions

RNA alternative splicing is an important determinant of transcriptome complexity in humans, with an estimated 90% of multi-exon genes expressing alternative transcripts. Many of these alternative transcripts are cell-type specific, and aberrant expression of some has been associated with disease. The power of modern RNA sequencing technologies facilitates discovery of novel splice isoforms, as well as quantification of splice junction inclusion levels and changes between conditions. These techniques are not perfect, introducing biases and batch effects which complicate the accurate measurement of gene expression and splicing. Moreover, underlying sample variation can interfere with detection of splicing changes, leading to inflated false discovery rates. Despite being the state of the art, MAJIQ is not immune to these effects.

In this dissertation, I describe three tools which work with MAJIQ to account for data heterogeneity in various data contexts. Chapter 2 deals with the problem of outliers in small replicate studies. Briefly, splicing events in samples are scored based on how well their Ψ distributions agree with the group median for that event. The scores are used to scale down the sample reads at the LSV level, with harsher penalties for purported outliers. We demonstrated that this correction compensates for outliers in real data, stabilizing detection power and reproducibility relative to a clean comparison.

As part of designing and benchmarking the outlier detection and correction scheme, we simulated outliers by adjusting read counts relative to a fixed shift in Ψ for a subset of LSVs. This also served as an investigation into the effects of different kinds of outliers on Ψ quantification. We observed, for example, that increasing either the severity or abundance of misbehaving LSVs resulted in an inflated false positive count in uncorrected MAJIQ. We also observed that the effective total number of reads in the outlier sample has a profound effect on performance. When the effective depth was reduced, true positive detections were lost, whereas an increased effective depth caused false positives to accumulate. These

observations motivate more general questions about how data conditions affect model performance. Intuitively, a certain read count is required for adequate transcriptome coverage. In splicing analysis, the demand for junction-spanning reads, which represent a very small subset of the total fragment count, in turn demands deeper sequencing. The question, then, is how deep one would need to sequence before LSV detection begins to saturate. In addition, how many biological or technical replicates are needed to accurately quantify as much of the transcriptome as possible?

Chapter 3 lays the groundwork for addressing the latter of these questions. Here we address datasets with a large number of RNA-seq samples which are not necessarily replicates, for which MAJIQ's Bayesian assumption of shared underlying Ψ does not necessarily hold. We remedy this by implementing a framework of statistical tests over posterior bootstrap samples, with each input experiment serving as an independent reporter of Ψ per LSV. The purpose of this procedure is to capture the per-sample uncertainty in Ψ as a means of false discovery control, with some cost to speed. However, the procedure is currently completely naive to the reasons why one sample may be less certain about Ψ than another. There are known and latent confounding factors affecting gene expression and, potentially, alternative splicing, which MAJIQ-HET does not account for. These factors are accounted for on the Ψ level in the sQTL pipeline, and work is underway to implement confounding factors and bias correction in the MAJIQ pipeline.

A parallel problem in methods development is that of assessing methods performance. Our approach to this has been to design new metrics for assessing the performance of a method on a given dataset. The difficulty lies in the differences between methods, in this case how splicing events are defined. For example, for every splicing event that rMATS quantifies, MAJIQ identifies two congruent LSVs. The ratio is even greater when comparing between MAJIQ's LSVs and LeafCutter's intron clusters. In order to properly compare these methods, there needs to be a one-to-one mapping between events classified by all methods. Such a mapping would also improve communication of splicing events by making it easier to match

up splicing events described in the literature by two different software packages. This need has recently motivated two new projects in MAJIQ's development. First, a simplifier is in development to remove excess junctions that have enough read support to incorporate into the splice graph but have low inclusion levels across all samples. Second, an event classifier is planned to assign labels to clusters of LSVs that describe the same splicing event, similar to how intron clusters are defined by LeafCutter. These two additions to MAJIQ's pipeline will benefit end users as well, as the simplified, classified events will be easier to describe and prioritize for experimental follow-up.

Finally, Chapter 4 introduces a pipeline for associating genotype with MAJIQ splicing levels and applied it to two collaborative projects. I exercised this pipeline to call sQTLs relevant to two unrelated disease conditions, coronary artery disease and aberrant skeletal growth. In doing so, I identified a handful of disease-relevant loci which could potentially be explained by splicing variations and are therefore key targets for functional validation. While I was able to replicate one such variation by RT-PCR, namely the rs6410 association with inclusion of *CYP11B1* exon 4, the relevance of this splicing variation in the abnormal phenotype remains unclear. A likely hypothesis is that the alternatively included exons result in a gain of function in *CYP11B1*. Whether this new function is sequestration of factors which regulate the production of growth hormones, or an increase in cytochrome p450 efficiency, remains to be seen. Furthermore, the *ADAMTS7* splicing variation has yet to be validated experimentally. In addition to the alternative 3' splice site event captured by the sQTL screen, association between that and the downstream tandem skipping event and 3'-truncated transcript should be confirmed by RT-PCR. Additionally, translation of both the alternative isoforms of *CYP11B1* and *ADAMTS7* should be verified by targeted peptide analysis.

A natural application of this pipeline would be to perform genomewide, tissue-specific discovery of putative sQTLs in GTEx. However, there remain several inefficiencies that need to be addressed prior to large-scale application. The pipeline presented and exercised in

this work is a prototype which scales poorly to larger datasets. Ψ quantifications are read from the VOILA text output files, collated into Python objects, and associated with SNPs read at runtime from the variant call file. This results in a great deal of overhead both in disk operations, processor time, and memory demands. Much of this overhead can be resolved by integrating the sQTL pipeline into the MAJIQ workflow. In addition to removing multiple intermediate file I/O steps from the pipeline, this move would simultaneously necessitate the transpilation of this pipeline from Python to parallelized C++, which alone has improved the running time and memory efficiency of MAJIQ tenfold (see Figure 11a).

In summary, the methods I have developed and presented in this dissertation implement novel solutions for addressing data heterogeneity in RNA-seq splicing analysis. These tools can be applied in functional studies of splicing and transcriptome regulation, or to discover new disease mechanisms. MAJIQout and MAJIQ-HET are publicly available at `https://majiq.biociphers.org` as part of MAJIQ. The sQTL pipeline remains in development pending efficiency improvements to support genomewide screens on larger datasets.

APPENDIX

A.1. Attachments

Table 1: `validations_51pct.tsv`: Performance of MAJIQ and MAJIQ-HET on simulated GTEx samples, requiring that events only be quantifiable in a minimum of 51% of samples per group.

Table 2: `validations_100pct.tsv`: Performance of MAJIQ and MAJIQ-HET on simulated GTEx samples, requiring that an event be quantifiable in all input samples.

Table 3: `twist1-supp-tables.xlsx`: Summary of sQTLs in arterial tissues. See the README in the document itself.

BIBLIOGRAPHY

G. P. Alamancos, E. Agirre, and E. Eyras. Methods to Study Splicing from High-Throughput RNA Sequencing Data. In K. J. Hertel, editor, *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, pages 357–397. Humana Press, Totowa, NJ, 2014. ISBN 978-1-62703-980-2. doi: 10.1007/978-1-62703-980-2_26. URL `http://dx.doi.org/10.1007/978-1-62703-980-2_26`.

A. Ameur, A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, and L. Feuk. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural &Amp; Molecular Biology*, 18:1435, Nov. 2011. URL `https://doi.org/10.1038/nsmb.2143`.

Arnold. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *PNAS*, 110(8): E736–E745, Feb. 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1222809110. URL `http://www.pnas.org/content/110/8/E736`.

B. Bai, C. M. Hales, P.-C. Chen, Y. Gozal, E. B. Dammer, J. J. Fritz, X. Wang, Q. Xia, D. M. Duong, C. Street, G. Cantero, D. Cheng, D. R. Jones, Z. Wu, Y. Li, I. Diner, C. J. Heilman, H. D. Rees, H. Wu, L. Lin, K. E. Szulwach, M. Gearing, E. J. Mufson, D. A. Bennett, T. J. Montine, N. T. Seyfried, T. S. Wingo, Y. E. Sun, P. Jin, J. Hanfelt, D. M. Willcock, A. Levey, J. J. Lah, and J. Peng. U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 2013. doi: 10.1073/pnas.1310249110. URL `http://www.pnas.org/content/early/2013/09/09/1310249110.abstract`.

G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald, and G. R. Grant. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, 14: 135, Dec. 2016. URL `https://doi.org/10.1038/nmeth.4106`.

S. Bennett. Solexa ltd. *Pharmacogenomics*, 5(4):433–438, 2004. doi: 10.1517/14622416.5.4.433. URL `https://doi.org/10.1517/14622416.5.4.433`. PMID: 15165179.

A. L. Beyer and Y. N. Osheim. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes & Development*, 2(6):754–765, June 1988. doi: 10.1101/gad.2.6.754. URL `http://genesdev.cshlp.org/content/2/6/754.abstract`.

S. Brixius-Anderko and E. E. Scott. Structure of human cortisol-producing cytochrome P450 11b1 bound to the breast cancer drug fadrozole provides insights for drug design. *The Journal of biological chemistry*, Nov. 2018. ISSN 1083-351X 0021-9258. doi: 10.1074/jbc.RA118.006214.

Brænne Ingrid, Civelek Mete, Vilne Baiba, Di Narzo Antonio, Johnson Andrew D., Zhao Yuqi, Reiz Benedikt, Codoni Veronica, Webb Thomas R., Foroughi Asl Hassan, Hamby Stephen E., Zeng Lingyao, Trégouët David-Alexandre, Hao Ke, Topol Eric J., Schadt Eric E., Yang Xia, Samani Nilesh J., Björkegren Johan L.M., Erdmann Jeanette, Schunkert

Heribert, and Lusis Aldons J. Prediction of Causal Candidate Genes in Coronary Artery Disease Loci. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 35(10):2207–2217, Oct. 2015. doi: 10.1161/ATVBAHA.115.306108. URL https://doi.org/10.1161/ATVBAHA.115.306108.

J. Cao. The functional role of long non-coding RNAs and epigenetics. *Biological procedures online*, 16:11–11, Sept. 2014. ISSN 1480-9222. doi: 10.1186/1480-9222-16-11. URL https://www.ncbi.nlm.nih.gov/pubmed/25276098.

Chasman Daniel I. and Lawler Patrick R. Understanding AAA Pathobiology. *Circulation Research*, 120(2):259–261, Jan. 2017. doi: 10.1161/CIRCRESAHA.116.310395. URL https://doi.org/10.1161/CIRCRESAHA.116.310395.

M. Chen and J. L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, 10(11):741–754, Nov. 2009. ISSN 1471-0072. doi: 10.1038/nrm2777. URL http://dx.doi.org/10.1038/nrm2777.

R. Colak, T. Kim, M. Michaut, M. Sun, M. Irimia, J. Bellay, C. L. Myers, B. J. Blencowe, and P. M. Kim. Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. *PLoS Comput Biol*, 9(4):e1003030, Apr. 2013. doi: 10.1371/journal.pcbi.1003030. URL http://dx.doi.org/10.1371/journal.pcbi.1003030.

M. Colombo, M. J. Blok, P. Whiley, M. Santamariña, S. Gutiérrez-Enríquez, A. Romero, P. Garre, A. Becker, L. D. Smith, G. D. Vecchi, R. D. Brandão, D. Tserpelis, M. Brown, A. Blanco, S. Bonache, M. Menéndez, C. Houdayer, C. Foglia, J. D. Fackenthal, D. Baralle, B. Wappenschmidt, E. Díaz-Rubio, T. Caldés, L. Walker, O. Díez, A. Vega, A. B. Spurdle, P. Radice, and M. D. L. Hoya. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Human Molecular Genetics*, page ddu075, Feb. 2014. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddu075. URL http://hmg.oxfordjournals.org/content/early/2014/03/03/hmg.ddu075.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0881-8. URL http://dx.doi.org/10.1186/s13059-016-0881-8.

I. W. Deveson, S. A. Hardwick, T. R. Mercer, and J. S. Mattick. The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome. *Trends in Genetics*, 33 (7):464–478, July 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2017.04.004. URL http://www.sciencedirect.com/science/article/pii/S0168952517300690.

Dichgans Martin, Malik Rainer, König Inke R., Rosand Jonathan, Clarke Robert, Gretarsdottir Solveig, Thorleifsson Gudmar, Mitchell Braxton D., Assimes Themistocles L., Levi Christopher, O'Donnell Christopher J., Fornage Myriam, Thorsteinsdottir Unnur, Psaty Bruce M., Hengstenberg Christian, Seshadri Sudha, Erdmann Jeanette, Bis Joshua C., Peters Annette, Boncoraglio Giorgio B., März Winfried, Meschia James F.,

Kathiresan Sekar, Ikram M. Arfan, McPherson Ruth, Stefansson Kari, Sudlow Cathie, Reilly Muredach P., Thompson John R., Sharma Pankaj, Hopewell Jemma C., Chambers John C., Watkins Hugh, Rothwell Peter M., Roberts Robert, Markus Hugh S., Samani Nilesh J., Farrall Martin, Schunkert Heribert, null null, Gschwendtner Andreas, Bevan Steve, Chen Yu-Ching, DeStefano Anita L., Parati Eugenio A., Quertermous Tom, Ziegler Andreas, Boerwinkle Eric, Holm Hilma, Fischer Marcus, Kessler Thorsten, Willenborg Christina, Laaksonen Reijo, Voight Benjamin F., Stewart Alexandre F.R., Rader Daniel J., Hall Alistair S., and Kooner Jaspal S. Shared Genetic Susceptibility to Ischemic Stroke and Coronary Artery Disease. *Stroke*, 45(1):24–36, Jan. 2014. doi: 10.1161/STROKEAHA.113.002707. URL `https://doi.org/10.1161/STROKEAHA.113.002707`.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29 (1):15–21, 2013. doi: 10.1093/bioinformatics/bts635. URL `http://bioinformatics.oxfordjournals.org/content/29/1/15.abstract`.

H. Dvinge and R. K. Bradley. Widespread intron retention diversifies most cancer transcriptomes. *Genome medicine*, 7(1):45–45, May 2015. ISSN 1756-994X. doi: 10.1186/s13073-015-0168-9. URL `https://www.ncbi.nlm.nih.gov/pubmed/26113877`.

J. C. Entizne, J. L. Trincado, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras. Fast and accurate differential splicing analysis across multiple conditions with replicates. *bioRxiv*, 2016. doi: 10.1101/086876. URL `http://biorxiv.org/content/early/2016/11/10/086876`.

M. S. Forman, J. Farmer, J. K. Johnson, C. M. Clark, S. E. Arnold, H. B. Coslett, A. Chatterjee, H. I. Hurtig, J. H. Karlawish, H. J. Rosen, V. Van Deerlin, V. M.-Y. Lee, B. L. Miller, J. Q. Trojanowski, and M. Grossman. Frontotemporal dementia: Clinicopathological correlations. *Annals of Neurology*, 59(6):952–962, 2006. ISSN 1531-8249. doi: 10.1002/ana.20873. URL `http://onlinelibrary.wiley.com/doi/10.1002/ana.20873/abstract`.

O. Franzén, R. Ermel, A. Cohain, N. K. Akers, A. Di Narzo, H. A. Talukdar, H. Foroughi-Asl, C. Giambartolomei, J. F. Fullard, K. Sukhavasi, S. Köks, L.-M. Gan, C. Giannarelli, J. C. Kovacic, C. Betsholtz, B. Losic, T. Michoel, K. Hao, P. Roussos, J. Skogsberg, A. Ruusalepp, E. E. Schadt, and J. L. M. Björkegren. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science*, 353(6301): 827, Aug. 2016. doi: 10.1126/science.aad6970. URL `http://science.sciencemag.org/content/353/6301/827.abstract`.

A. C. Frazee, A. E. Jaffe, B. Langmead, and J. T. Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics (Oxford, England)*, 31(17): 2778–2784, Sept. 2015. ISSN 1367-4811 1367-4803. doi: 10.1093/bioinformatics/btv272.

M. Gazzara, M. Mallory, R. Roytenberg, J. Lindberg, A. Jha, K. Lynch, and Y. Barash. Ancient antagonism between CELF and RBFOX families tunes mRNA splicing outcomes. *Genome Research*, 27(8):1–11, May 2017. doi: 10.1101/gr.220517.117.

GTEx Consortium, F. Aguet, A. A. Brown, S. E. Castel, J. R. Davis, Y. He, B. Jo, P. Mohammadi, Y. Park, P. Parsana, A. V. Segrè, B. J. Strober, Z. Zappala, B. B. Cummings, E. T. Gelfand, K. Hadley, K. H. Huang, M. Lek, X. Li, J. L. Nedzel, D. Y. Nguyen, M. S. Noble, T. J. Sullivan, T. Tukiainen, D. G. MacArthur, G. Getz, A. Addington, P. Guan, S. Koester, A. R. Little, N. C. Lockhart, H. M. Moore, A. Rao, J. P. Struewing, S. Volpi, L. E. Brigham, R. Hasz, M. Hunter, C. Johns, M. Johnson, G. Kopen, W. F. Leinweber, J. T. Lonsdale, A. McDonald, B. Mestichelli, K. Myer, B. Roe, M. Salvatore, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, J. Bridge, B. A. Foster, B. M. Gillard, E. Karasik, R. Kumar, M. Miklos, M. T. Moser, S. D. Jewell, R. G. Montroy, D. C. Rohrer, D. Valley, D. C. Mash, D. A. Davis, L. Sobin, M. E. Barcus, P. A. Branton, N. S. Abell, B. Balliu, O. Delaneau, L. Frésard, E. R. Gamazon, D. Garrido-Martín, A. D. H. Gewirtz, G. Gliner, M. J. Gloudemans, B. Han, A. Z. He, F. Hormozdiari, X. Li, B. Liu, E. Y. Kang, I. C. McDowell, H. Ongen, J. J. Palowitch, C. B. Peterson, G. Quon, S. Ripke, A. Saha, A. A. Shabalin, T. C. Shimko, J. H. Sul, N. A. Teran, E. K. Tsang, H. Zhang, Y.-H. Zhou, C. D. Bustamante, N. J. Cox, R. Guigó, M. Kellis, M. I. McCarthy, D. F. Conrad, E. Eskin, G. Li, A. B. Nobel, C. Sabatti, B. E. Stranger, X. Wen, F. A. Wright, K. G. Ardlie, E. T. Dermitzakis, T. Lappalainen, F. Aguet, K. G. Ardlie, B. B. Cummings, E. T. Gelfand, G. Getz, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, K. J. Karczewski, M. Lek, X. Li, D. G. MacArthur, J. L. Nedzel, D. T. Nguyen, M. S. Noble, A. V. Segrè, C. A. Trowbridge, T. Tukiainen, N. S. Abell, B. Balliu, R. Barshir, O. Basha, A. Battle, G. K. Bogu, A. Brown, C. D. Brown, S. E. Castel, L. S. Chen, C. Chiang, D. F. Conrad, N. J. Cox, F. N. Damani, J. R. Davis, O. Delaneau, E. T. Dermitzakis, B. E. Engelhardt, E. Eskin, P. G. Ferreira, L. Frésard, E. R. Gamazon, D. Garrido-Martín, A. D. Gewirtz, G. Gliner, M. J. Gloudemans, R. Guigo, I. M. Hall, B. Han, Y. He, F. Hormozdiari, C. Howald, H. Kyung Im, B. Jo, E. Yong Kang, Y. Kim, S. Kim-Hellmuth, T. Lappalainen, G. Li, X. Li, B. Liu, S. Mangul, M. I. McCarthy, I. C. McDowell, P. Mohammadi, J. Monlong, S. B. Montgomery, M. Muñoz-Aguirre, A. W. Ndungu, D. L. Nicolae, A. B. Nobel, M. Oliva, H. Ongen, J. J. Palowitch, N. Panousis, P. Papasaikas, Y. Park, P. Parsana, A. J. Payne, C. B. Peterson, J. Quan, F. Reverter, C. Sabatti, A. Saha, M. Sammeth, A. J. Scott, A. A. Shabalin, R. Sodaei, M. Stephens, B. E. Stranger, B. J. Strober, J. H. Sul, E. K. Tsang, S. Urbut, M. van de Bunt, G. Wang, X. Wen, F. A. Wright, H. S. Xi, E. Yeger-Lotem, Z. Zappala, J. B. Zaugg, Y.-H. Zhou, J. M. Akey, D. Bates, J. Chan, L. S. Chen, M. Claussnitzer, K. Demanelis, M. Diegel, J. A. Doherty, A. P. Feinberg, M. S. Fernando, J. Halow, K. D. Hansen, E. Haugen, P. F. Hickey, L. Hou, F. Jasmine, R. Jian, L. Jiang, A. Johnson, R. Kaul, M. Kellis, M. G. Kibriya, K. Lee, J. Billy Li, Q. Li, X. Li, J. Lin, S. Lin, S. Linder, C. Linke, Y. Liu, M. T. Maurano, B. Molinie, S. B. Montgomery, J. Nelson, F. J. Neri, M. Oliva, Y. Park, B. L. Pierce, N. J. Rinaldi, L. F. Rizzardi, R. Sandstrom, A. Skol, K. S. Smith, M. P. Snyder, J. Stamatoyannopoulos, B. E. Stranger, H. Tang, E. K. Tsang, L. Wang, M. Wang, N. Van Wittenberghe, F. Wu, R. Zhang, C. R. Nierras, P. A. Branton, L. J. Carithers, P. Guan, H. M. Moore, A. Rao, J. B. Vaught, S. E. Gould, N. C. Lockart, C. Martin, J. P. Struewing, S. Volpi, A. M. Addington, S. E. Koester, A. R. Little, L. E. Brigham, R. Hasz, M. Hunter, C. Johns, M. Johnson, G. Kopen, W. F. Leinweber, J. T. Lonsdale, A. McDonald, B. Mestichelli, K. Myer, B. Roe, M. Salvatore, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, J. Bridge, B. A. Foster, B. M. Gillard,

E. Karasik, R. Kumar, M. Miklos, M. T. Moser, S. D. Jewell, R. G. Montroy, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, A. H. Undale, A. M. Smith, D. E. Tabor, N. V. Roche, J. A. McLean, N. Vatanian, K. L. Robinson, L. Sobin, M. E. Barcus, K. M. Valentino, L. Qi, S. Hunter, P. Hariharan, S. Singh, K. S. Um, T. Matose, M. M. Tomaszewski, L. K. Barker, M. Mosavel, L. A. Siminoff, H. M. Traino, P. Flicek, T. Juettemann, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, B. Craft, M. Goldman, M. Haeussler, W. J. Kent, C. M. Lee, B. Paten, K. R. Rosenbloom, J. Vivian, and J. Zhu. Genetic effects on gene expression across human tissues. *Nature*, 550: 204, Oct. 2017. URL `https://doi.org/10.1038/nature24277`.

L. Herzel, D. S. M. Ottoz, T. Alpert, and K. M. Neugebauer. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nature Reviews. Molecular Cell Biology*, Aug. 2017. ISSN 1471-0080. doi: 10.1038/nrm.2017.63.

A. Horvath and C. A. Stratakis. Unraveling the molecular basis of micronodular adrenal hyperplasia. *Current opinion in endocrinology, diabetes, and obesity*, 15(3):227–233, June 2008. ISSN 1752-2978. doi: 10.1097/MED.0b013e3282fe7416. URL `https://www.ncbi.nlm.nih.gov/pubmed/18438169`.

N. Kaminski and N. Friedman. Practical Approaches to Analyzing Results of Microarray Experiments. *American Journal of Respiratory Cell and Molecular Biology*, 27(2):125–132, Aug. 2002. ISSN 1044-1549. doi: 10.1165/ajrcmb.27.2.f247. URL `https://doi.org/10.1165/ajrcmb.27.2.f247`.

Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Meth*, 7(12):1009–1015, Dec. 2010. ISSN 1548-7091. doi: 10.1038/nmeth.1528. URL `http://dx.doi.org/10.1038/nmeth.1528`.

T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assuncao, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, Sept. 2011. ISSN 0028-0836. doi: 10.1038/nature10413. URL `http://dx.doi.org/10.1038/nature10413`.

D. Kim, B. Langmead, and S. L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12:357, Mar. 2015. URL `http://dx.doi.org/10.1038/nmeth.3317`.

D. Klarin, Q. M. Zhu, C. A. Emdin, M. Chaffin, S. Horner, B. J. McMillan, A. Leed, M. E. Weale, C. C. A. Spencer, F. Aguet, A. V. Segrè, K. G. Ardlie, A. V. Khera, V. K. Kaushik, P. Natarajan, CARDIoGRAMplusC4D Consortium, and S. Kathiresan. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration

pathways to coronary artery disease. *Nature Genetics*, 49:1392, July 2017. URL https://doi.org/10.1038/ng.3914.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, Mar. 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-3-r25. URL https://doi.org/10.1186/gb-2009-10-3-r25.

B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323. URL http://dx.doi.org/10.1186/1471-2105-12-323.

Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, and J. K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, Jan. 2018. ISSN 1546-1718. doi: 10.1038/s41588-017-0004-9. URL https://doi.org/10.1038/s41588-017-0004-9.

S. Lykke-Andersen and T. H. Jensen. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology*, 16:665, Sept. 2015. URL https://doi.org/10.1038/nrm4063.

A. J. Matlin, F. Clark, and C. W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, May 2005. ISSN 1471-0072, 1471-0080. doi: 10.1038/nrm1645. URL http://www.nature.com/doifinder/10.1038/nrm1645.

McPherson Ruth and Tybjaerg-Hansen Anne. Genetics of Coronary Artery Disease. *Circulation Research*, 118(4):564–578, Feb. 2016. doi: 10.1161/CIRCRESAHA.115.306566. URL https://doi.org/10.1161/CIRCRESAHA.115.306566.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5:621, May 2008. URL https://doi.org/10.1038/nmeth.1226.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881):1344, June 2008. doi: 10.1126/science.1158441. URL http://science.sciencemag.org/content/320/5881/1344.abstract.

C. P. Nelson, A. Goel, A. S. Butterworth, S. Kanoni, T. R. Webb, E. Marouli, L. Zeng, I. Ntalla, F. Y. Lai, J. C. Hopewell, O. Giannakopoulou, T. Jiang, S. E. Hamby, E. Di Angelantonio, T. L. Assimes, E. P. Bottinger, J. C. Chambers, R. Clarke, C. N. A. Palmer, R. M. Cubbon, P. Ellinor, R. Ermel, E. Evangelou, P. W. Franks, C. Grace, D. Gu, A. D. Hingorani, J. M. M. Howson, E. Ingelsson, A. Kastrati, T. Kessler, T. Kyriakou, T. Lehtimäki, X. Lu, Y. Lu, W. März, R. McPherson, A. Metspalu, M. Pujades-Rodriguez, A. Ruusalepp, E. E. Schadt, A. F. Schmidt, M. J. Sweeting, P. A. Zalloua,

K. AlGhalayini, B. D. Keavney, J. S. Kooner, R. J. F. Loos, R. S. Patel, M. K. Rutter, M. Tomaszewski, I. Tzoulaki, E. Zeggini, J. Erdmann, G. Dedoussis, J. L. M. Björkegren, EPIC-CVD Consortium, CARDIoGRAMplusC4D, The UK Biobank CardioMetabolic Consortium CHD working group, H. Schunkert, M. Farrall, J. Danesh, N. J. Samani, H. Watkins, and P. Deloukas. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics*, 49:1385, July 2017. URL `https://doi.org/10.1038/ng.3913`.

S. Norton, J. Vaquero-Garcia, and Y. Barash. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates.

S. S. Norton, J. Vaquero-Garcia, N. F. Lahens, G. R. Grant, and Y. Barash. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. *Bioinformatics*, 34(9):1488–1497, May 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx790. URL `http://dx.doi.org/10.1093/bioinformatics/btx790`.

S. Nürnberg, J. Vaquero-Garcia, S. Norton, M. Pjanic, S. Elwyn, J. Pluta, W. Zhao, S. Testa, Y. Barash, C. Brown, T. Quertermous, and D. Rader. Deep transcriptomic analysis of human vascular cells identifies risk genes for common vascular diseases, May 2017. URL `http://professional.heart.org/professional/EducationMeetings/MeetingsLiveCME/ATVBPVD/UCM_316902_ATVBPVD-Scientific-Sessions.jsp`.

H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics (Oxford, England)*, 32(10):1479–1485, May 2016. ISSN 1367-4811 1367-4803. doi: 10.1093/bioinformatics/btv722.

K. Ozaki, Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, Y. Nakamura, and T. Tanaka. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32:650, Nov. 2002. URL `https://doi.org/10.1038/ng1047`.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, Apr. 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4197. URL `http://www.nature.com/nmeth/journal/v14/n4/full/nmeth.4197.html`.

J. K. Pickrell, T. Berisa, J. Z. Liu, L. Ségurel, J. Y. Tung, and D. A. Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48:709, May 2016. URL `https://doi.org/10.1038/ng.3570`.

N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb. Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, 12(7):R67, Jul 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-7-r67. URL `https://doi.org/10.1186/gb-2011-12-7-r67`.

M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human

genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20):10614–10619, Oct. 1996. ISSN 0027-8424. doi: 10.1073/pnas.93.20.10614. URL https://www.ncbi.nlm.nih.gov/pubmed/8855227.

U. Schmitz, N. Pinello, F. Jia, S. Alasmari, W. Ritchie, M.-C. Keightley, S. Shini, G. J. Lieschke, J. J.-L. Wong, and J. E. J. Rasko. Intron retention enhances gene regulatory complexity in vertebrates. *Genome biology*, 18(1):216–216, Nov. 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1339-3. URL https://www.ncbi.nlm.nih.gov/pubmed/29141666.

A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, Apr. 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts163. URL https://doi.org/10.1093/bioinformatics/bts163.

S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601, 2014. doi: 10.1073/pnas.1419161111. URL http://www.pnas.org/content/111/51/E5593.abstract.

B. E. Slatko, A. F. Gardner, and F. M. Ausubel. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1):e59, Apr. 2018. ISSN 1934-3639. doi: 10.1002/cpmb.59. URL https://doi.org/10.1002/cpmb.59.

L. Song and L. Florea. CLASS: constrained transcript assembly of RNA-seq reads. *BMC bioinformatics*, 14 Suppl 5(Suppl 5):S14–S14, Apr. 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-S5-S14. URL https://www.ncbi.nlm.nih.gov/pubmed/23734605.

T. Sterne-Weiler, R. J. Weatheritt, A. J. Best, K. C. H. Ha, and B. J. Blencowe. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular cell*, 72(1):187–200.e6, Oct. 2018. ISSN 1097-4164 1097-2765. doi: 10.1016/j.molcel.2018.08.018.

S. Sun, Z. Zhang, O. Fregoso, and A. R. Krainer. Mechanisms of activation and repression by the alternative splicing factors RBFOX1/2. *RNA*, 18(2):274–283, 2012. URL http://rnajournal.cshlp.org/content/18/2/274.short.

Tada Hayato, Won Hong-Hee, Melander Olle, Yang Jian, Peloso Gina M., and Kathiresan Sekar. Multiple Associated Variants Increase the Heritability Explained for Plasma Lipids and Coronary Artery Disease. *Circulation: Cardiovascular Genetics*, 7(5):583–587, Oct. 2014. doi: 10.1161/CIRCGENETICS.113.000420. URL https://doi.org/10.1161/CIRCGENETICS.113.000420.

the CARDIoGRAMplusC4D Consortium, M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, T. R. Webb, L. Zeng, A. Dehghan, M. Alver, S. M. Armasu, K. Auro, A. Bjonnes, D. I. Chasman, S. Chen, I. Ford, N. Franceschini, C. Gieger, C. Grace, S. Gustafsson, J. Huang,

S.-J. Hwang, Y. K. Kim, M. E. Kleber, K. W. Lau, X. Lu, Y. Lu, L.-P. Lyytikäinen, E. Mihailov, A. C. Morrison, N. Pervjakova, L. Qu, L. M. Rose, E. Salfati, R. Saxena, M. Scholz, A. V. Smith, E. Tikkanen, A. Uitterlinden, X. Yang, W. Zhang, W. Zhao, M. de Andrade, P. S. de Vries, N. R. van Zuydam, S. S. Anand, L. Bertram, F. Beutner, G. Dedoussis, P. Frossard, D. Gauguier, A. H. Goodall, O. Gottesman, M. Haber, B.-G. Han, J. Huang, S. Jalilzadeh, T. Kessler, I. R. König, L. Lannfelt, W. Lieb, L. Lind, C. M. Lindgren, M.-L. Lokki, P. K. Magnusson, N. H. Mallick, N. Mehra, T. Meitinger, F.-u.-R. Memon, A. P. Morris, M. S. Nieminen, N. L. Pedersen, A. Peters, L. S. Rallidis, A. Rasheed, M. Samuel, S. H. Shah, J. Sinisalo, K. E. Stirrups, S. Trompet, L. Wang, K. S. Zaman, D. Ardissino, E. Boerwinkle, I. B. Borecki, E. P. Bottinger, J. E. Buring, J. C. Chambers, R. Collins, L. A. Cupples, J. Danesh, I. Demuth, R. Elosua, S. E. Epstein, T. Esko, M. F. Feitosa, O. H. Franco, M. G. Franzosi, C. B. Granger, D. Gu, V. Gudnason, A. S. Hall, A. Hamsten, T. B. Harris, S. L. Hazen, C. Hengstenberg, A. Hofman, E. Ingelsson, C. Iribarren, J. W. Jukema, P. J. Karhunen, B.-J. Kim, J. S. Kooner, I. J. Kullo, T. Lehtimäki, R. J. F. Loos, O. Melander, A. Metspalu, W. März, C. N. Palmer, M. Perola, T. Quertermous, D. J. Rader, P. M. Ridker, S. Ripatti, R. Roberts, V. Salomaa, D. K. Sanghera, S. M. Schwartz, U. Seedorf, A. F. Stewart, D. J. Stott, J. Thiery, P. A. Zalloua, C. J. O'Donnell, M. P. Reilly, T. L. Assimes, J. R. Thompson, J. Erdmann, R. Clarke, H. Watkins, S. Kathiresan, R. McPherson, P. Deloukas, H. Schunkert, N. J. Samani, and M. Farrall. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47:1121, Sept. 2015. URL `https://doi.org/10.1038/ng.3396`.

H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigó. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, Sept. 2012. doi: 10.1101/gr.134445.111. URL `http://genome.cshlp.org/content/22/9/1616.abstract`.

J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40, Mar. 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1417-1. URL `https://doi.org/10.1186/s13059-018-1417-1`.

van der Harst Pim and Verweij Niek. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research*, 122(3):433–443, Feb. 2018. doi: 10.1161/CIRCRESAHA.117.312086. URL `https://doi.org/10.1161/CIRCRESAHA.117.312086`.

J. Vaquero-Garcia, A. Barrera, M. R. Gazzara, J. González-Vallinas, N. F. Lahens, J. B. Hogenesch, K. W. Lynch, and Y. Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:e11752, Feb. 2016. ISSN 2050-084X. doi: 10.7554/eLife.11752. URL `http://elifesciences.org/content/5/e11752v2`.

J. Vaquero-Garcia, S. Norton, and Y. Barash. LeafCutter vs. MAJIQ and comparing soft-

ware in the fast-moving field of genomics. *bioRxiv*, page 463927, Jan. 2018. doi: 10.1101/463927. URL http://biorxiv.org/content/early/2018/11/08/463927.abstract.

Vicente Gilsanz and Osman Ratib. *Hand Bone Age: A Digital Atlas of Skeletal Maturity*. Springer, 2012. ISBN 978-3-642-23762-1.

E. T. Wang and A. T. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature*, 8:749–761, 2007. doi: 10.1038/nrg2164.

Y. Wang, J. Liu, B. O. Huang, Y.-M. Xu, J. Li, L.-F. Huang, J. Lin, J. Zhang, Q.-H. Min, W.-M. Yang, and X.-Z. Wang. Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158, Mar. 2015. ISSN 2049-9434. doi: 10.3892/br.2014.407. URL https://www.ncbi.nlm.nih.gov/pubmed/25798239.

D. R. Zerbino, P. Achuthan, W. Akanni, M. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, Jan. 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1098. URL http://dx.doi.org/10.1093/nar/gkx1098.

F. Zhang, M. Wang, T. Michael, and R. Drabier. Novel alternative splicing isoform biomarkers identification from high-throughput plasma proteomics profiling of breast cancer. *BMC Systems Biology*, 7(Suppl 5):S8, Dec. 2013. ISSN 1752-0509. doi: 10.1186/1752-0509-7-S5-S8. URL http://www.biomedcentral.com/1752-0509/7/S5/S8/abstract.

R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 2014. doi: 10.1073/pnas.1408886111. URL http://www.pnas.org/content/111/45/16219.abstract.

Zhao Yuqi, Chen Jing, Freudenberg Johannes M., Meng Qingying, null null, Rajpal Deepak K., and Yang Xia. Network-Based Identification and Prioritization of Key Regulators of Coronary Artery Disease Loci. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 36 (5):928–941, May 2016. doi: 10.1161/ATVBAHA.115.306725. URL https://doi.org/10.1161/ATVBAHA.115.306725.