2022

Improving Molecular Diagnosis Of Suspected Mendelian Disorders With Rna Splicing Analysis

Joseph Krittameth Aicher
University of Pennsylvania

Improving Molecular Diagnosis Of Suspected Mendelian Disorders With Rna Splicing Analysis

Abstract

Exome sequencing is the most advanced standard-of-care genetic test for people with suspected Mendelian disorders. Yet, the diagnostic rate of exome sequencing is only 31%. RNA sequencing (RNA-seq) is a promising molecular test for detecting potentially pathogenic changes in RNA splicing as part of obtaining a molecular diagnosis. In this dissertation, I develop new computational tools and perform analyses towards improving how we detect these potentially pathogenic changes in RNA splicing with the goal of improving the molecular diagnostic rate. First, in Chapter 1, I review background on how we diagnose patients and how RNA splicing and RNA-seq could be used to improve this process. Then, in Chapter 2, I describe my contributions to MAJIQ v2 as methodology to study RNA splicing from large and heterogeneous RNA-seq datasets. Afterwards, I use MAJIQ v2 in Chapter 3 to evaluate how tissue-specific expression and splicing affects what clinically-relevant splicing changes we can identify from clinically-accessible tissues. Then, in Chapter 4, I describe the limitations of MAJIQ v2 for our approach to detect splicing aberrations and the development and evaluation of MAJIQ v3 to address these challenges. With MAJIQ v3, I develop MAJIQ-CLIN in Chapter 5 to identify and prioritize splicing aberrations in patient RNA-seq data and compare our method to previous approaches. Finally, in Chapter 6, I discuss overall conclusions for the work and exciting areas for future work. Together, the work in this dissertation pushes forward how we can study and use RNA-seq to improve the diagnostic rate of patients with suspected Mendelian disorders.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Yoseph Barash

Second Advisor

Elizabeth J. Bhoj

Keywords

diagnostics, Mendelian disorders, molecular diagnosis, RNA-seq, splicing

Subject Categories

Bioinformatics | Biology | Genetics

IMPROVING MOLECULAR DIAGNOSIS OF SUSPECTED MENDELIAN DISORDERS WITH RNA SPLICING ANALYSIS

Joseph K. Aicher

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

2022

Supervisor of Dissertation	Co-Supervisor of Dissertation
Yoseph Barash	Elizabeth J. Bhoj
Associate Professor of Genetics	Assistant Professor of Pediatrics

Graduate Group Chairperson

Benjamin F. Voight, Associate Professor of Systems Pharmacology & Translational Therapeutics

Dissertation Committee

Maja Bucan, Professor of Genetics
Benjamin F. Voight, Associate Professor of Systems Pharmacology & Translational Therapeutics
Casey Greene, Professor of Biochemistry and Molecular Genetics, University of Colorado School of Medicine
Anne O'Donnell-Luria, Assistant Professor of Pediatrics, Harvard Medical School

IMPROVING MOLECULAR DIAGNOSIS OF SUSPECTED MENDELIAN DISORDERS WITH RNA SPLICING ANALYSIS

To Peter and Madeleine.

ACKNOWLEDGEMENT

This dissertation was made possible with the support of many people.

I am extremely grateful to my co-advisors Yoseph Barash and Elizabeth Bhoj. Their mentorship has been essential to my development as a scientist. To Yoseph and Elizabeth: thank you for your teaching, your support, and for keeping me on track.

I would like to express my deepest appreciation to my committee: Maja Bucan, Ben Voight, Casey Greene, and Anne O'Donnell-Luria. Their insight and suggestions have improved my work.

I also cannot overstate the support I have received from Penn GCB and the "best MSTP in the galaxy," in particular from Skip Brass, Ben Voight, Li-San Wang, Maureen Kirsch, and Maggie Krall for their guidance in navigating these programs. I am thankful to past and current members of the Barash and Bhoj labs. I have repeatedly received invaluable feedback and learned many new things from them during various meetings and informal interactions, especially Jordi Vaquero-Garcia, Anupama Jha, Caleb Radens, Divya Nair, and Laura Bryant.

I am also thankful for the sources of funding I have received for my physician-scientist training. Research reported in this dissertation was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number F30HD098803.

I would also like to extend my deepest gratitude to the numerous scientific mentors who cultivated my interest in mathematical and scientific research prior to joining the MD-PhD program, including: Manuel Lladser, Johan Van Hove, Elaine Spector-Christensen, Jerome Busenitz, William H Butler, Kabe Moen, Russell M Taylor II, Song Song, and Laura Reed.

My success would not have been possible without the unwavering love and support of my friends and family. I am thankful to my friends for their support along the way, including: Aileen Ren, Ary Swaminathan, David Kersen, Kevin Goff, Linda Zhou, Lindsey Fernandez, Rafi Fernandez,

Sai Phyo, and Tong Wang. I am extremely grateful to my parents, Thomas and Orapunn Aicher, for raising me and supporting me in all my endeavors. I am thankful to my older brother, Christopher Aicher, for being my role model, friend, and childhood mathematics instructor. I am also grateful to my in-laws for their support and advice: James and Patricia Bucher, Nicholas and Taylor Bucher-Dial, Brian Bucher, and Hoiyi Ng. Finally, I cannot begin to express my thanks and love to my wife, Bernadette Bucher, for her support and encouragement. Since our first encounter studying general topology, through our wedding at the start of my MD-PhD training and the births of our children, Peter and Madeleine, and through the present and future, our unwavering partnership in our family life and our detailed discussions of mathematics and computational science have been my foundation. The path of love is never smooth, but mine's continuous for you.

I am proud to have been supported by so many, and no words here would be enough to express my thanks.

ABSTRACT

IMPROVING MOLECULAR DIAGNOSIS OF SUSPECTED MENDELIAN DISORDERS WITH RNA SPLICING ANALYSIS

Joseph K. Aicher

Yoseph Barash

Elizabeth J. Bhoj

Exome sequencing is the most advanced standard-of-care genetic test for people with suspected Mendelian disorders. Yet, the diagnostic rate of exome sequencing is only 31%. RNA sequencing (RNA-seq) is a promising molecular test for detecting potentially pathogenic changes in RNA splicing as part of obtaining a molecular diagnosis. In this dissertation, I develop new computational tools and perform analyses towards improving how we detect these potentially pathogenic changes in RNA splicing with the goal of improving the molecular diagnostic rate. First, in Chapter 1, I review background on how we diagnose patients and how RNA splicing and RNA-seq could be used to improve this process. Then, in Chapter 2, I describe my contributions to MAJIQ v2 as methodology to study RNA splicing from large and heterogeneous RNA-seq datasets. Afterwards, I use MAJIQ v2 in Chapter 3 to evaluate how tissue-specific expression and splicing affects what clinically-relevant splicing changes we can identify from clinically-accessible tissues. Then, in Chapter 4, I describe the limitations of MAJIQ v2 for our approach to detect splicing aberrations and the development and evaluation of MAJIQ v3 to address these challenges. With MAJIQ v3, I develop MAJIQ-CLIN in Chapter 5 to identify and prioritize splicing aberrations in patient RNA-seq data and compare our method to previous approaches. Finally, in Chapter 6, I discuss overall conclusions for the work and exciting areas for future work. Together, the work in this dissertation pushes forward how we can study and use RNA-seq to improve the diagnostic rate of patients with suspected Mendelian disorders.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

1.1. Challenges in molecular diagnosis of Mendelian disorders

Exome sequencing is the most advanced standard-of-care genetic test for patients with suspected Mendelian disorders. The goal is to successfully determine a molecular diagnosis for each patient by identifying the pathogenic genetic variant(s) causing disease. Clinically, molecular diagnoses can lead to more accurate prognoses and improved clinical care for patients and their families. Scientifically, molecular diagnoses also inform fundamental biological research by suggesting novel genes and mechanisms in both normal development and disease pathogenesis. Unfortunately, a significant challenge in clinical and molecular genetics is that we are unable to provide a molecular diagnosis to most families of children tested with exome sequencing. Specifically, the molecular diagnostic rate of exome sequencing across various diagnostic laboratories averages only around 31% (Clark et al., 2018; Yang et al., 2014; Farwell et al., 2015; Retterer et al., 2016). Genome sequencing, where it has begun being implemented, has been reported to improve upon this diagnostic rate by around 10-15% (Clark et al., 2018; Alfares et al., 2018; Taylor et al., 2015). As a result, we are unable to provide a molecular diagnosis for the majority of patients tested with either exome or genome sequencing.

Exome sequencing measures an individual's personal genetic sequences at the exons and adjacent intronic regions for nearly all known genes, typically identifying tens of thousands of variants (Ross et al., 2020). In clinical practice, these variants must be prioritized by a computational pipeline to generate a short list of candidate variants that can be manually reviewed by genomic professionals. These pipelines rely on databases of known pathogenic variants (e.g., HGMD (Stenson et al., 2003), ClinVar (Landrum et al., 2014)) and predictive models of pathogenicity for new variants. To be useful for variant prioritization, predictive models must be highly sensitive and specific to ensure that the few causal pathogenic variants are scored highly versus the far more numerous benign variants. Clinical pipelines generally only

evaluate nonsynonymous and canonical splice-site variants, for which such predictive models exist. Although other intronic and synonymous exonic variants are known to cause disease, their pathogenicity is challenging to predict (Frésard and Montgomery, 2018; Gloss and Dinger, 2018). Consequently, existing clinical pipelines generally filter out such variants, resulting in missed diagnoses.

1.2. RNA splicing defects as a mechanism for Mendelian disease

One key mechanism by which intronic and exonic variants can cause disease is by altering RNA splicing (Scotti and Swanson, 2016; Wang and Cooper, 2007). RNA splicing is the process by which different segments of pre-mRNA are selectively included or excluded and removed as exons and introns to create a mature mRNA (from which proteins are translated) (Figure 1.1). This process is highly regulated across developmental stages and tissues, with over 90% of human genes undergoing alternative splicing, and is mediated by the spliceosome and numerous RNA-binding proteins (RBPs) that recognize different conserved sequence elements. Variants in these splice factors can thus alter splicing in trans, and intronic and exonic variants can alter splicing in *cis* by changing the strength of existing sequence elements or introducing cryptic ones. These different changes in splicing can alter protein function by inserting or removing part of the mRNA transcript (Figure 1.1). Furthermore, they can sometimes cause a frameshift and/or insertion of a premature termination codon, leading to loss of function. Such splicing-altering variants are known to cause Mendelian disorders (e.g., familial dysautonomia, Crouzon syndrome, etc.) and are associated with complex diseases such as Alzheimer's disease and cancer (Scotti and Swanson, 2016; Fenwick et al., 2014; Raj et al., 2018). It is estimated that 15-50% of human pathogenic variants alter splicing, with the lower number attributed to splice-sites alone and the higher number more challenging to pin down due to lack of appropriate data or tools and inherent biases in disease-variant databases for splice-site mutations (see above). A recent study suggests that these biases could cause existing clinical pipelines to miss the majority of splice-disrupting variants; Cheung et al. (2019) performed a massively parallel splicing reporter assay of 27,733 exonic and intronic variants in 2,198 human exons from ExAC and found that variants in core splice-sites only accounted for 17% of splice-disrupting variants, with the majority

of splice-disrupting variants occurring outside of the extended splice region. Recent years have seen a surge in splicing and isoform-specific focused therapeutics for rare genetic diseases such as spinal muscular atrophy and familial dysautonomia, as well as cancer (Singh and Cooper, 2012; Sinha et al., 2018; Seiler et al., 2018; Sotillo et al., 2015).

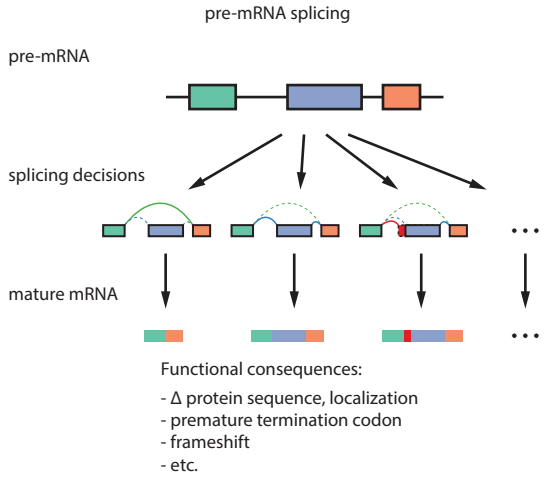


Figure 1.1: Different precursor mRNA (pre-mRNA) splicing decisions can have significant, potentially pathogenic, functional consequences.

Broadly, there are two approaches by which we can improve identification of variants causing potentially disease-causing alterations to RNA splicing: (1, DNA-first) better algorithms for predicting defects in RNA splicing from genetic sequence and (2, RNA-first) orthogonal approaches for directly measuring potentially disease-causing alterations to RNA splicing from patient RNA and matching back to genetic variants.

1.3. Algorithms to predict deleterious splicing changes

Given the importance of splicing and identifying pathogenic variants in general, many algorithms have been developed in recent years for variant annotation and prediction (see Table 1.1). Many algorithms (e.g., GWAVA, M-CAP, etc.) produce a score that can be used as a pathogenicity score for general variants. This score is sometimes trained directly on known pathogenic variants or indirectly using information about conservation/other annotations. While effective, they are not tailored specifically for splicing defects. S-CAP similarly produces a pathogenicity score but specifically focuses on predicting pathogenicity for variants that do not

Table 1.1: Algorithms for DNA-first splicing variant prioritization.

Model	Citation	Type of algorithm
GWAVA	Ritchie et al. (2014)	Pathogenicity scoring
CADD	Kircher et al. (2014)	Pathogenicity scoring
M-CAP	Jagadeesh et al. (2016)	Pathogenicity scoring
EIGEN	Ionita-Laza et al. (2016)	Pathogenicity scoring
LINSIGHT	Huang et al. (2017)	Pathogenicity scoring
S-CAP	Jagadeesh et al. (2018)	Pathogenicity scoring
HAL	Rosenberg et al. (2015)	Synthetic minigene
SPIDEX	Xiong et al. (2015)	Splicing code
	Jha et al. (2017)	Splicing code
MMSplice	Cheng et al. (2019)	Splicing code
SpliceAI	Jaganathan et al. (2019)	Splicing code
MTSplice	Cheng et al. (2021)	Splicing code

directly affect protein-coding sequence using splicing-related features. Other algorithms (e.g., HAL) use synthetic libraries but lack appropriate tissue context, and most lack mechanistic interpretation. Splicing code models offer great potential for variant prioritization as these can predict condition-specific splicing changes and offer mechanistic interpretation to disease-associated variants; however, these models have not yet been trained on genetic variants. Benchmarks for these different algorithms often apply different metrics or datasets, hindering the identification of the best algorithms for incorporation into clinical diagnosis pipelines. Completed and ongoing works I contributed to during my PhD are aimed to improve such methods (Jha et al., 2020).

1.4. Detection and quantification of RNA splicing using RNA-seq

However, the focus of this dissertation is directly measuring potentially disease-causing alterations to RNA splicing from patient RNA and matching back to genetic variants (the RNA-first approach). Clinical RNA-seq is one approach by which laboratories can identify splicing aberrations among other transcriptomic variations such as gene-expression outliers, allele-specific expression, and gene fusions. Indeed, previous work in several labs have demonstrated that RNA-seq can enable genetic diagnosis in patients previously unsolved by exome or genome sequencing (Cummings et al., 2017; Kremer et al., 2017; Gonorazky et al., 2019; Frésard et al.,

2019; Mertes et al., 2021; Jenkinson et al., 2020; Murdock et al., 2021).

RNA-seq describes a variety of methods which measure sequence from a given sample of RNA. The resulting sequences provide information about the presence and relative abundance of specific genomic features (e.g., genes, isoforms, splice sites, etc.) corresponding to observed sequences. Most commonly, input RNA is first enriched for mRNA, followed by reverse transcription to cDNA, and then sequencing with "short-reads" with Illumina technology (Stark et al., 2019). This approach is relatively inexpensive and yields short ($\sim$200bp) cDNA fragments which are sequenced with high accuracy. In contrast, alternative technology (e.g., PacBio, Oxford Nanopore) has more recently enabled sequencing of much longer cDNA (or direct sequencing of RNA) at lengths over 1-10kb, but currently with lower (but increasing) accuracy and throughput (Amarasinghe et al., 2020; Stark et al., 2019).

The resulting sequences can be used to quantify changes in gene isoform usage. Methods to quantify these changes can be divided broadly between methods that aim to quantify whole isoforms and those that quantify localized alternative splicing (AS) "events" within a gene. While quantifying all gene isoforms accurately across diverse conditions can be regarded as the grand challenge of transcriptomics, achieving this goal remains open due to several limiting factors. In the case of long reads technology, these factors include high error rate and high costs which do not allow researchers to capture enough reads from all isoforms. In the case of the more commonly used short reads technology, these limiting factors include the sparsity of reads, their positional bias, and the fact that reads usually cannot be assigned to a unique isoform. In addition, the composition of isoforms in a sample is typically unknown, requiring further inference of the existing isoforms or making simplifying assumptions such as a known transcriptome. These issues have led many researchers to focus on local AS "events" which can be more easily and accurately quantified from RNA-seq. AS events are quantified in terms of percent spliced in (PSI, denoted by Ψ), which is the relative ratio of isoforms including a specific splicing junction or retained intron. Traditionally, AS events have been studied only for a restricted set of the most common "types" (e.g., cassette exons). In a previous study,

Vaquero-Garcia et al. (2016) extended this set of AS event types using the formulation of local splicing variations (LSVs) and introduced MAJIQ as a software package for studying such LSVs. LSVs, which can be defined as splits in a gene splicegraph coming into or from a reference exon, allow researchers to capture not only previously defined AS types but also much more complex variations involving more than two alternative junctions. Furthermore, the LSV formulation, and similar definitions of local AS events suggested in subsequent works, also help incorporate and quantify unannotated (novel) splice junctions. Previous work comparing splicing across mouse tissues has shown that accounting for complex and novel variations results in an over 30% increase of detected differentially spliced events while maintaining the same level of reproducibility and experimental validation rates (Vaquero-Garcia et al., 2016). Importantly, capturing such unannotated splice variations is of particular importance for the study of disease such as cancer and neurodegeneration which often involve aberrant splicing.

Recent works have applied RNA-seq to identify splicing and other transcriptomic aberrations in RNA-seq data from patients with suspected Mendelian disorders (Cummings et al., 2017; Kremer et al., 2017; Gonorazky et al., 2019; Frésard et al., 2019; Mertes et al., 2021; Jenkinson et al., 2020; Murdock et al., 2021). In this setting, the goal is to identify changes that are unusual (i.e., outliers in the patient vs other samples) and subsequently correlate them with the patient's clinical phenotype and genetic variants. Focusing on splicing aberrations, some of these methods look at evidence of a novel spliced junction regardless of relative inclusion (Cummings et al., 2017; Gonorazky et al., 2019), while others quantify relative inclusion (e.g., PSI, etc.) (Kremer et al., 2017; Frésard et al., 2019; Mertes et al., 2021; Jenkinson et al., 2020; Murdock et al., 2021). They compare this evidence to a large control set of either external RNA-seq experiments (e.g., GTEx (GTEx Consortium et al., 2017)) or the other patient samples (leave-one-out), assuming that they have different splicing aberrations as the cause of their diseases. Some methods further model and correct for unobserved causes of sample covariation that could cause spurious outliers (Frésard et al., 2019; Mertes et al., 2021).

1.5. Gaps in knowledge

The focus of my dissertation research is to improve the molecular diagnostic rate for patients with suspected Mendelian disorders by identifying splicing aberrations from RNA-seq. Identifying these splicing aberrations from RNA-seq first requires an accurate and efficient approach for identifying and quantifying splicing changes in large (and often heterogeneous) RNA-seq datasets. Furthermore, tissue-specific expression and splicing mean that splicing aberrations in the clinically-relevant tissues (e.g., brain) will not always be measurable in the tissues which are clinically-available (e.g., fibroblasts). Finally, efficiently identifying splicing aberrations requires new methodology for identifying relevant novel changes in splicing events and quantifications and correcting for sample covariation due to known as well as unknown confounders.

In Chapter 2, I describe the new algorithms and performance comparisons I contributed to MAJIQ v2 which address challenges posed by large and heterogeneous RNA-seq datasets. In Chapter 3, I use MAJIQ v2 to understand the limitations of clinically-accessible tissues (CATs) for representing changes in non-CATs, informing which potentially pathogenic splicing changes we can hope to measure and which tissues we should prefer to sample in the clinical setting. In Chapter 4, I describe additional limitations we identified in MAJIQ v2 we needed to address to identify splicing aberrations in suspected Mendelian disorders, leading to MAJIQ v3. In Chapter 5, I describe the approach we developed for identifying splicing aberrations and comparisons to previous methodologies on patient data.

CHAPTER 2

RNA splicing analysis using heterogeneous and large RNA-seq datasets

2.1. Introduction

Despite previous demonstrations of MAJIQ's utility for analyzing AS (Vaquero-Garcia et al., 2016; Norton et al., 2018), we found MAJIQ along with many commonly used methods for AS events quantification to be ill-suited for handling heterogeneous and large RNA-seq datasets, such as GTEx and ENCODE. Datasets may involve anywhere from just a few to many thousands of samples each, and are typically heterogeneous as they often do not represent biological or technical replicates. The consequent increased splicing variability, illustrated in Figure 2.1a-b, can be the result of a multitude of factors, both experimental/technical (e.g., difference in sequencing machine), and biological (e.g., gender, age). While some confounding factors may be corrected with appropriate methods (Slaff et al., 2021), fully removing the observed variability in such data is unlikely and may also over-constrain the data, thus leading to a loss of true biological signal. These datasets pose several algorithmic, computational, and visualization challenges. First, the assumption of a shared PSI per LSV junction in a group, used by methods such as MAJIQ and LeafCutter, is violated in such data even when handling only a small dataset with few samples, leading to a potential increase in false positives and loss of power. Second, algorithms need to not only scale to thousands of samples efficiently but also to allow incrementally adding new samples as more data is acquired, and to support multiple group comparisons (e.g. multiple tissue comparisons across GTEx). Third, the increased complexity of the data requires efficient representation. Such efficient representation would allow users to capture the many unannotated splicing variations in the data, while at the same time simplifying its representation and quantification. Such simplification will allow filtering of lowly used splice junctions while also detecting possibly new sub-types of significant variations. Finally, efficient and user-friendly visualization is required to probe multiple sample groups as well as individual samples.

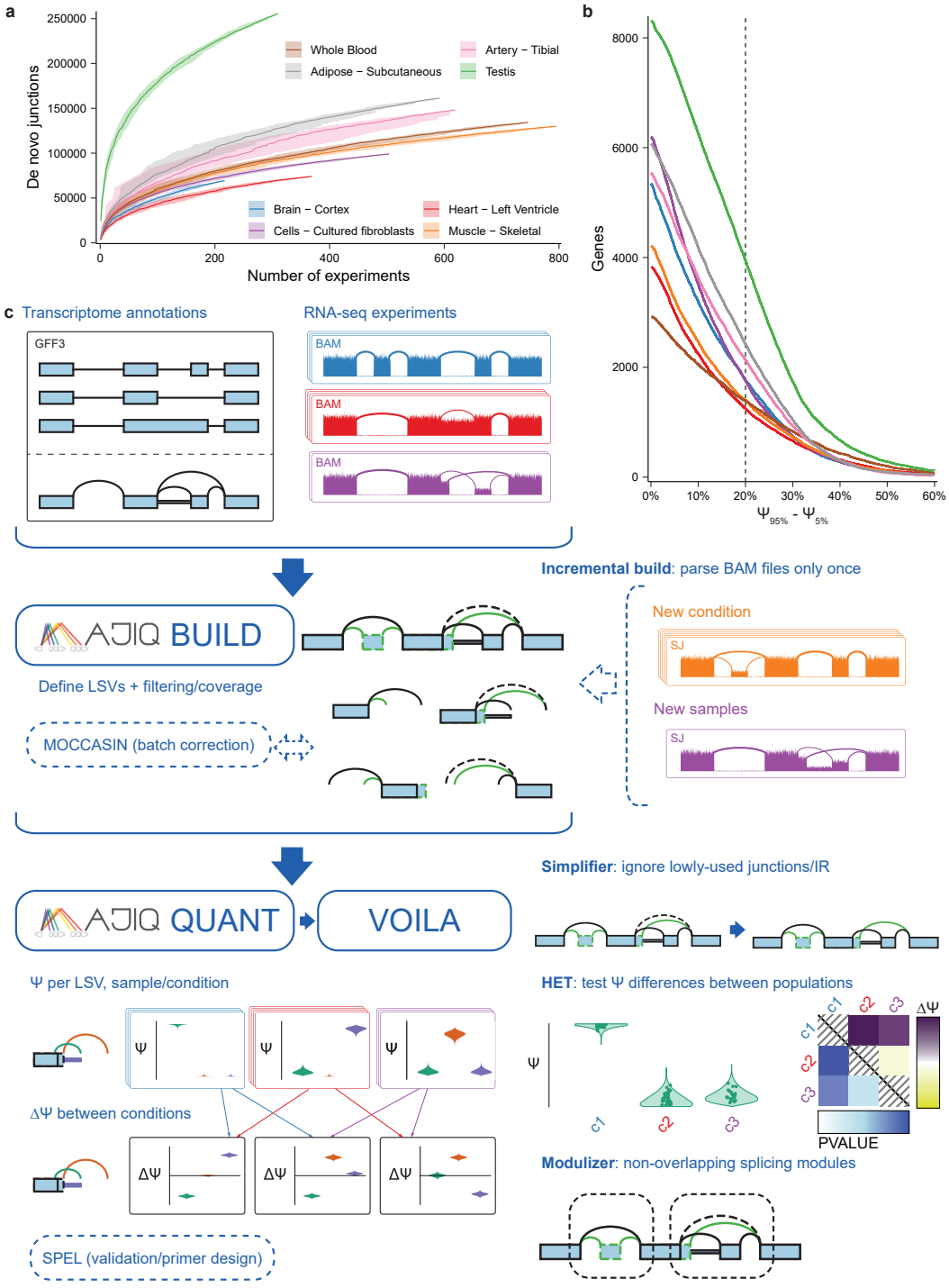


Figure 2.1: MAJIQ efficiently and accurately models, quantifies, and visualizes RNA splicing from large and complex RNA-seq datasets. (cont.)

Figure 2.1: (cont.) (a) The number of identified distinct unannotated *de novo* junctions increases with larger subsets of different tissues from GTEx. Lines show the median over 30 randomly selected permutations over experiments in each subset, confidence bands show the 5th to 95th percentiles over permutations of samples per tissue. (b) The number of genes with at least one junction where the difference between the 95th percentile and 5th percentile of PSI exceeds a given value for different tissues from GTEx (same tissues/colors as in Fig. 1a). Dashed vertical line indicates how many genes have a difference in PSI exceeding 20%. (c) MAJIQ combines annotated transcript databases and coverage from input RNA-seq experiments to build a model of each gene as a collection of exons connected by annotated and *de novo* junctions and retained introns (splicegraph). Junctions and retained introns sharing the same source or target exon form local splicing variations (LSVs). MAJIQ quantifies the relative inclusion of junctions and retained introns in each LSV in terms of percent spliced in (PSI, Ψ) and provides VOILA to make interactive visualizations of splicing quantifications with respect to each gene's splicegraph and LSV structures. MAJIQ v2 introduces an incremental build, which allows RNA-seq coverage to be read from BAM files only once to a coverage file (SJ), accelerating subsequent builds with different experiments. MAJIQ v2 introduces a simplifier, which can be used to reduce splicegraph/LSV complexity by ignoring lowly used junctions and retained introns. MAJIQ v2 introduces a new mode for quantification, HET, which compares PSI differences between populations of independent RNA-seq experiments and accounts for variable uncertainty per experiment. MAJIQ v2 introduces the modulizer, which allows performing analysis relative to non-overlapping splicing modules rather than LSVs.

To address these challenges, we developed an array of tools and algorithms included in the MAJIQ v2 package. In this chapter, I describe the MAJIQ v2 algorithms and my contributions to this work, including the nonparametric statistical tests for differential splicing (MAJIQ HET), the incremental splicegraph builder, and the MAJIQ v2 algorithm for quantifying intron retention.

2.2. Results

2.2.1. The MAJIQ v2 splicing analysis pipeline

To support RNA splicing analysis using large RNA-seq datasets, we implemented the set of tools and algorithms illustrated in Figure 2.1c. In the first step, the MAJIQ builder combines transcript annotations and coverage from aligned RNA-seq experiments in order to build an updated splicegraph for each gene which includes novel (unannotated) elements such as junctions, retained introns, and exons. Several user-defined filters can be applied at this stage to exclude junctions or retained introns which have low coverage or are not detected in enough samples in user-defined sample groups. Notably, per-experiment coverage is saved separately so that

it can be used in subsequent analyses without reprocessing aligned reads a second time (i.e. incremental build). This feature is highly relevant for large studies with incremental releases, such as ENCODE and GTEx, and also for individual lab projects where datasets or samples are added as the project evolves.

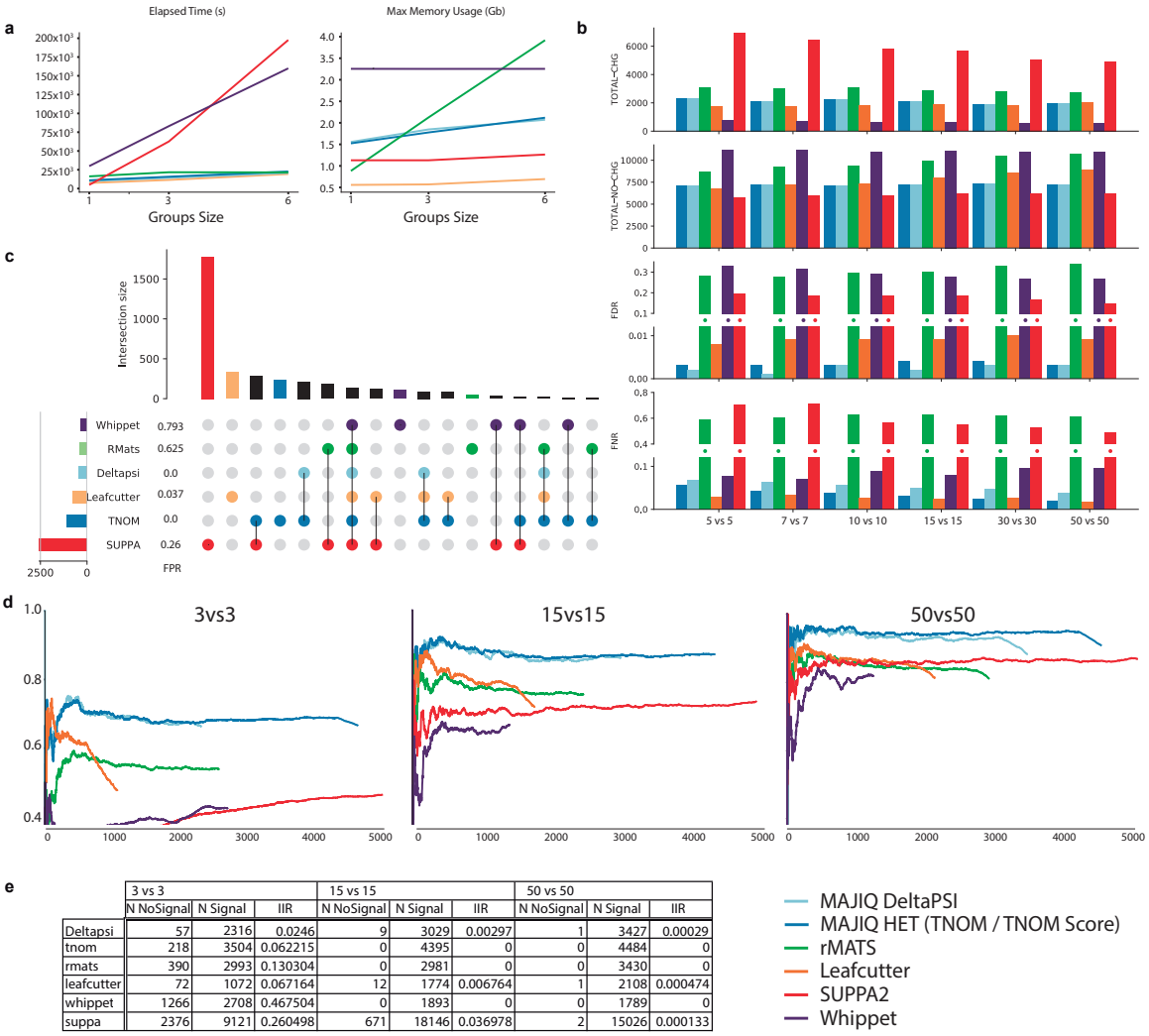
In the second step of the pipeline, the MAJIQ quantifier is executed. As in the original MAJIQ framework, splicing quantification is performed in units of LSVs. Briefly, an LSV corresponds to a split in gene splicegraphs coming into or out of a reference exon. Each LSV edge, corresponding to a splice junction or intron retention, is quantified in terms of its relative inclusion (PSI, $\Psi \in [0, 1]$) or changes in its relative inclusion between two conditions (dPSI, $\Delta\Psi \in [-1, 1]$). Given the junction spanning reads observed in each LSV, MAJIQ's Bayesian model results in a posterior distributions over the (unknown) inclusion level ($\mathbb{P}(\Psi)$), or the changes in inclusion levels between conditions ($\mathbb{P}(\Delta\Psi)$). This model accounts not only for the total number of reads but also for factors such as read distribution across genomic locations and read stacks. Given its Bayesian framework, the model can also output the confidence in inclusion change of at least C ($\mathbb{P}(|\Delta\Psi| > C)$), or the expectation over the computed posterior distributions ($\mathbb{E}[\Psi]$, $\mathbb{E}[\Delta\Psi]$). In this work, we introduce two new algorithms within the MAJIQ quantifier. The first involves how intron retention is quantified, allowing for much faster execution with higher accuracy (see Methods). The second addition is the implementation of additional test statistics, termed MAJIQ HET (heterogeneous). Conceptually, the original MAJIQ model assumes a shared (hidden) PSI value for a given group of samples and accumulates evidence (reads) across these samples to infer PSI. In contrast, MAJIQ HET quantifies PSI for each sample separately and then applies robust rank-based test statistics (TNOM, InfoScore, or Mann-Whitney U).

2.2.2. Performance evaluation

In order to assess MAJIQ HET, our new method for detecting differential splicing, we performed a comprehensive comparison to an array of commonly used algorithms using both synthetic and real data. We considered only algorithms capable of analyzing large datasets,

including the original MAJIQ algorithm (upgraded with the v2 code-base to enable efficient data processing), rMATS turbo (Shen et al., 2014), LeafCutter (Li et al., 2018), SUPPA2 (Trincado et al., 2018), and Whippet (Sterne-Weiler et al., 2018). Figure 2.2a shows processing time and memory when performing a multi-group, multi-sample comparison, typical for such datasets. In this case, we perform all pairwise comparisons between 10 tissue groups, and the number of samples in each group grows from 1 (10 total samples) to 6 (60 total samples). All algorithms are able to process such large datasets using only 0.5-4 GB of memory, an amount readily available on modern laptops. However, large differences exist in terms of running time, with SUPPA2 (55 hours) and Whippet (50 hours) taking substantially longer to analyze the larger dataset (6 samples per group, 60 total samples) compared to approximately 6 hours by rMATS, LeafCutter and MAJIQ v2.

Next, we assessed the accuracy of all algorithms using a large-scale synthetic dataset for comparing two tissue groups. This synthetic dataset, by far the largest of its kind to the best of our knowledge, was constructed to be "realistic" such that each synthetic sample was generated to mimic a real GTEx sample from either cerebellum or smooth muscle tissues (see Methods). All methods were required to report changing AS events which pass the method's statistical significance test and inferred to exhibit a substantial splicing change of at least 20% (see Methods). However, we note that since the various algorithms use significantly different definitions of AS events it is hard to compare those directly. For example, LeafCutter defines AS events as clusters of overlapping introns which may involve multiple 3'/5' alternative splice sites and skipped exons, while rMATS is limited to only classical AS events with two alternative junctions. Thus, to facilitate a comparative analysis, we resorted to comparing the various algorithms output at the gene rather than event level using the synthetic dataset shown in Figure 2.2b. First, we found SUPPA2 consistently reported over 6,000 differentially spliced genes, thousands more than any other method, while Whippet reported roughly 785 genes, significantly fewer than the other methods which reported over 2,000 changing genes (Fig. 2b top bar chart). Whippet, followed by rMATS, reported significantly more non-changing events. SUPPA2, rMATS, and Whippet all exhibited high FDR ranging around 15-30%, with the former two also



e

	3 vs 3			15 vs 15			50 vs 50		
	N NoSignal	N Signal	IIR	N NoSignal	N Signal	IIR	N NoSignal	N Signal	IIR
Deltapsi	57	2316	0.0246	9	3029	0.00297	1	3427	0.00029
tnom	218	3504	0.062215	0	4395	0	0	4484	0
rmats	390	2993	0.130304	0	2981	0	0	3430	0
leafcutter	72	1072	0.067164	12	1774	0.006764	1	2108	0.000474
whippet	1266	2708	0.467504	0	1893	0	0	1789	0
suppa	2376	9121	0.260498	671	18146	0.036978	2	15026	0.000133

Figure 2.2: MAJIQ v2 performance evaluation using synthetic and real data. (a) Time (left) and memory (right) consumption when analyzing multiple sample groups. Results shown are for running all pairwise differential splicing analysis between 10 tissue groups from GTEx v8 as the number of samples per group increases from 1 to 6 (x-axis). (b) Performance evaluation for differential splicing calls using simulated GTEx cerebellum and skeletal muscle samples and aggregated over genes (see main text and Methods). Metrics include the total number of genes reported as changing (TOTAL-CHG) or non changing (TOTAL-NO-CHG) by each method, and the associated FDR and FNR. Horizontal axis denotes the size of the groups. (c) Upset plot based on the 10vs10 analysis shown in (b). The bars on top represent the overlap between genes reported as differentially spliced by each method indicated below it. The bars and FPR values by each method name on the left refer to genes reported only by that method. (cont.)

Figure 2.2: (cont.) (d) Reproducibility ratio (RR) plots for real data, using GTEx cerebellum and liver samples. Analysis here is based on each method's reported list of splicing events (not genes) and unique scoring approach. X-axis is the ranked number of events reported by each method and Y-axis is the fraction of those events reproduced within the same number of top-ranking events when repeating the analysis using a different set of samples from the same tissue groups. The length of the line represents the total number of differentially splicing events reported by each method (see Methods for details). RR graphs are shown for comparing group sizes of 3 (left), 15 (middle), and 50 (right). (e) Intra-to-Inter Ratio (IIR) results for GTEx samples as in (d). IIR computes the ratio between the number of events reported as significantly changing when comparing two sample groups of the same type (N No Signal column) and the number of events reported as significantly changing when comparing groups of different types (here GTEx liver and cerebellum samples as in (d)).

exhibiting high FNR over 40%. Both MAJIQ and MAJIQ HET consistently maintained a lower false discovery rate compared to other algorithms (0.3%) and a low level of false negative rate which was similar to that of LeafCutter. On small sets, for example when using 5 samples per group, LeafCutter had a slightly lower FNR (2.5% vs 5.5% for HET), but MAJIQ exhibited lower FDR (0.03% vs 0.8%) while still reporting overall 34% more genes as changing (2,337 vs 1,739) and 6% more as non-changing (7,110 vs 6,713). It is also worth noting that the actual difference in the number of changing AS events reported by MAJIQ and LeafCutter is significantly higher, with 4,267 reported by MAJIQ vs. 2,169 by LeafCutter. This increased difference is mainly due to the increased resolution of event definition by MAJIQ. Specifically, MAJIQ uses the local splice variations formulation described above, while LeafCutter uses a definition of overlapping intronic regions which give rise to coarser event definition and can be sensitive to the coverage threshold used.

The significant differences between the methods described above raises the question how the reported sets of differentially spliced genes overlap. Figure 2.2c illustrates the result of such analysis when using 10 samples per group. Here, we looked at the intersection between different methods at the gene level and when a set was unique to a method (i.e. the underlying events are well defined) we also estimated the associated FPR. We found SUPPA2 reports a significantly higher number of unique genes (1,777) as differentially spliced but over a quarter of those are false positives. The next set sizes are those for LeafCutter (333), HET and SUPPA2 (288),

HET (230), and MAJIQ HET and PSI (214) with a FPR of 4% for the LeafCutter's unique set and close to 0 FPR for both MAJIQ's algorithms unique sets. rMATS and Whippet report significantly fewer unique genes with a high false positive rate of 62% and 79% respectively.

Next, we turned to assess performance on real GTEx data using several metrics. Here, unlike the synthetic data analysis which focused on comparative evaluation at the gene level, we focus on the actual AS events reported by each method. First, we used the reproducibility ratio (RR) statistic as shown in Figure 2d. The RR plots follow a similar procedure to that of irreproducible discovery rate (IDR) plots, used extensively to evaluate ChIP-seq peak callers (Li et al., 2011; Vaquero-Garcia et al., 2016). Briefly, RR plots answer the following simple question: given an algorithm A and a dataset D, if we rank all the events that algorithm A identifies as differentially spliced $(1, \dots, N_A)$, how many would be reproduced if you repeat this with dataset D', comprised of similar experiments using biological or technical replicates? The RR(n) plot, as shown in Figure 2.2d, is the fraction of those events that are reproduced (y-axis) as a function of $n \leq N_A$ (x-axis), with the overall reproducibility of differentially spliced events expressed as RR(N_A) (far right point of each curve in Figure 2.2d). In our RR analysis using groups of size 3 to 50 GTEx samples each, we found both MAJIQ and MAJIQ HET compared favorably to the other methods, but with the new HET algorithm exhibiting improved detection power resulting in a higher number of AS events at the same reproducibility level.

The second statistic we used for evaluating performance on real data is the intra-to-inter ratio (IIR) (Norton et al., 2018), which serves as a proxy for FDR on real data where the labels are unknown. Specifically, IIR computes the ratio between the number of differentially spliced events reported when comparing groups of the same condition (e.g. brain) and the number of events reported for similar group sizes of different conditions (e.g. brain vs liver). In our work, we found IIR to be a lower bound estimate of true FDR, though it lacks theoretical guarantees. In the analysis shown in Figure 2.2e, we found IIR to behave similarly to FDR on synthetic data with MAJIQ, MAJIQ HET, and LeafCutter exhibiting low IIR of 2%-6% even for small group sets of 5 samples, while rMATS, SUPPA2, and Whippet had an IIR of 13%, 26% and

46% respectively. However, unlike FDR on synthetic data, IIR dropped much more significantly, hitting practically zero for all methods for large sample groups. This result is to be expected since the IIR statistic compares sample groups of the same type, unlike the synthetic dataset described above where different tissues are compared.

The last component we included for assessing different methods' accuracy is a comparison to PSI quantifications using triplicates of RT-PCR assays, the gold standard in the RNA field. We previously produced over 100 such experiments from two different mouse tissues and showed MAJIQ compared favorably to SUPPA and rMATS (Vaquero-Garcia et al., 2016; Norton et al., 2018). Here, we extended this analysis to LeafCutter and found that MAJIQ's quantifications correlates significantly better with those of RT-PCR (see Fig. S2). We note that this analysis for LeafCutter was possible since all events we tested were simple cassette exon skipping, but it is not clear how to translate LeafCutter's output to actual PSI in the general case.

2.3. Discussion

MAJIQ v2 is the culmination of continuous development of MAJIQ since its original release in Vaquero-Garcia et al. (2016). The original MAJIQ, like many other algorithms, was designed for comparing relatively small groups of RNA-seq experiments from biological replicates. However, datasets nowadays can easily grow to hundreds and thousands of non-replicate samples. The sheer size and heterogeneous nature of such data poses challenges that go beyond just algorithmic efficiency. Additional challenges include the ability to capture but also simplify novel and complex splicing variations, the ability to define subtypes over such complex splicing events, and the ability to visualize and process such events and subtypes for downstream analysis. MAJIQ v2 is the only algorithm, to the best of our knowledge, that supports such features through efficient implementation of several algorithmic innovations we introduced either here (the simplifier, incremental build) or the full paper (Vaquero-Garcia et al., 2021) (modulizer, VOILA v2 visualization package). In addition, we perform extensive comparison of MAJIQ v2 to other algorithms and create a resource for reproducible algorithm comparison in the form of both data and software package. In the full paper, my co-authors further demonstrate the utility

of the new splicing analysis features by performing a detailed analysis of differential splicing between more than 2,300 samples from GTEx v8 brain subregions (Vaquero-Garcia et al., 2021). This is also demonstrated in our analysis of clinically-accessible and inaccessible tissues (next chapter).

The algorithmic contributions in this work include a new method to quantify *de novo* intron retention, an incremental build, addition of the MAJIQ HET statistics which do not assume a shared PSI between samples in a group, and the modulizer in VOILA. The resulting new features enhance splicing analysis, especially on larger datasets. For example, MAJIQ's incremental build saves much of the processing needed when adding new samples to existing repositories. Labs or centers can thus process data such as GTEx once, then efficiently add more relevant samples as needed. Furthermore, as these datasets get larger, we also expect to see more *de novo* junctions. These junctions increase the complexity of the splicegraph and the size of splicing events considered. The MAJIQ simplifier enables users to more finely control how this complexity enters the analysis.

With respect to performance, we showed MAJIQ v2 compares favorably to available methods in terms of efficiency, accuracy on synthetic data, and reproducibility on real RNA-seq data. Specifically, on synthetic data we found MAJIQ and LeafCutter were the only two tools that simultaneously demonstrated both low FDR and FNR when identifying genes with differential splicing. We note though that our usage of LeafCutter in this comparison included additional filtering for $\Delta\Psi > 20\%$ beyond the default p-value based filtering. This additional filtering was added as we found that the default LeafCutter settings performed much worse (Vaquero-Garcia et al., 2018). On real RNA-seq data from GTEx, MAJIQ's reproducibility was consistently higher than all other tools, particularly when comparing a small number of samples. When comparing MAJIQ HET to MAJIQ dPSI (from Vaquero-Garcia et al. (2016)), we found both to have similar reproducibility, but HET offered a significant increase in detection power. Finally, in terms of efficiency, we found MAJIQ v2 performed similarly to the most efficient tools in both memory and time. This is a notable achievement given that MAJIQ is the only tool amongst those that

offers detection and quantification of *de novo* intron retention, a computationally expensive yet important analysis as we discuss below.

The extensive evaluations we performed here serve not just to assess the specific tools we included, but as a service for the community. First, we created the largest synthetic RNA-seq dataset to date, with over 300 samples. In contrast to many other works, the data generated here was based on real life GTEx samples. It also does not reflect MAJIQ's model and was based instead on transcript-based quantifications by other algorithms (RSEM). As such, we would expect it to benefit tools that are built around a similar model (e.g. SUPPA). A second contribution is the evaluation package we created, validations-tools. This package allows users to not only reproduce our results but also to easily add future tools and repeat the analysis for future developers or for anyone who wants to assess performance on their own unique dataset. We highly recommend researchers and cores to take advantage of this as it is possible that on a dataset with other characteristics the various algorithms would perform differently. Finally, we note that the efforts to create reproducible results in genomics and specifically for tool development are constantly ongoing. We previously documented in detail issues we identified with using outdated software, software misuse, and lack of reproducibility that severely affected MAJIQ and other software assessment (Vaquero-Garcia et al., 2018).

Finally, our improved pipelines allowed us to map splicing variations across heterogeneous datasets of hundreds or thousands of experiments, as my co-authors demonstrate in the full paper on GTEx subregions (Vaquero-Garcia et al., 2021) and in the analysis of adult and fetal tissues described in the next chapter.

2.4. Methods

2.4.1. MAJIQ builder

In this subsection, we review how the MAJIQ builder prepares the structure and observations per experiment that are used for downstream splicing quantification as part of a scalable and principled approach to splicing analysis of large numbers of experiments. We describe the MAJIQ builder's new approach for estimating intron read rates, which allows junction and

intron coverage to be calculated once and reused efficiently for multiple analyses, unlike other methods that quantify intron retention. We also describe the MAJIQ simplifier, which reduces the complexity of the structural models of splicing used in quantification that especially arises from the analysis of large and heterogeneous datasets.

MAJIQ encodes the set of all possible splicing changes for a gene in terms of a splicegraph. A splicegraph is a graph-theoretic representation of a gene's splicing decisions from one exon to another, with exons as vertices and junctions and retained introns as distinct edges connecting exons. The exons of each gene are non-overlapping genomic intervals. Each junction has a source and target exon with a position within each exon, indicating the positions that are spliced together when the junction is used. Retained introns are between adjacent exons and indicate that intron retention between the exons is possible.

MAJIQ first constructs each gene's splicegraph by parsing transcript annotations from a GFF3 file. Exon boundaries and junctions from each transcript for a gene are combined in order to produce the minimal splicegraph that includes each transcript's annotated exons and junctions, splitting exons by retained introns to ensure that each junction starts and ends in different exons. MAJIQ then updates the splicegraph with novel junctions and introns found from processing input RNA-seq experiments' junction and intron coverage.

MAJIQ processes aligned input RNA-seq experiments to per-position junction and intron coverage in the following way. First, MAJIQ identifies reads with split alignments. The genomic coordinates of each split corresponds to a potential junction. Meanwhile, the coordinate of the split on the aligned read is the junction's "position" on the read. MAJIQ counts the number of reads for each junction from each possible position. Afterwards, MAJIQ identifies reads that contiguously intersect known or potential introns (i.e. reads that intersect the genomic coordinates between adjacent exons without splits within the intron boundaries). If the intron start is contained in the aligned read, the intron "position" is defined as for junctions (treating the exon/intron boundary as a junction with zero length). For aligned reads intersecting the intron but not the start, additional positions are defined by the genomic distances of the first

positions of the aligned reads to the intron start. These additional positions per intron increase the number of ways aligned reads can intersect introns in comparison to junctions. To adjust for this and model intron read coverage similarly to junction read counts, MAJIQ aggregates together adjacent intron positions to the equivalent number of possible positions per junction, taking the mean number of reads per reduced positions.

MAJIQ uses the obtained junction and intron coverage to update the splicegraph in the following way. Each potential junction is mapped to matching genes by prioritizing (1) genes that already contain the junction (i.e. annotated junctions) over (2) genes where both junction coordinates are within 400bp of an exon, which are prioritized over (3) genes where the junction is contained within the gene boundaries. The input experiments are divided into user-defined build groups. MAJIQ adds a novel junction to the splicegraph if there is sufficient evidence for its inclusion in one of the build groups. This happens when the total number of reads and total number of positions with at least one read exceeds the user-defined minimum number of reads and positions in at least a minimum number of experiments. MAJIQ adds novel exons or adjusts existing exon boundaries to accommodate the added novel junctions as previously described. Potential introns are added to the splicegraph under similar criteria, and their boundaries are adjusted or split to accommodate the adjusted or novel exon boundaries.

Since processed intron coverage is averaged over the entire original intronic region, we can carry over the same coverage as an estimate for all resulting splicegraph introns, which are contained in the original intron's boundaries. In contrast, MAJIQ's previous approach, which is also used by most other tools that quantify intron retention, quantified intron coverage using local counts of unsplit reads sharing the position of known junctions. These local counts must be calculated using information from all processed experiments (for all novel junctions), which requires samples to be reprocessed each time an analysis with different samples are performed. MAJIQ's new approach allows intron coverage to be processed once and used for multiple builds with potentially different intron boundaries. This enables MAJIQ's new incremental build feature, which saves intermediate files with junction and intron coverage that can be calculated once

and reused instead of BAM files for multiple builds. This reduces storage and time processing experiments that are part of multiple analyses.

While MAJIQ uses raw totals of read rates and number of nonzero positions for adding junctions and introns to the splicegraph, the MAJIQ builder performs additional modeling of per-position read rates for use in quantification. First, we mask positions with zero coverage and with outlier coverage. Outlier coverage is assessed under the observation that per-position read rates generally follow a Poisson distribution. For each junction/position, we use all other positions with nonzero coverage for that junction to estimate the Poisson rate parameter. Then, MAJIQ calls any position with an extreme right-tailed p-value (default 10^{-7}) under this model an outlier and ignores its contribution to coverage for quantification. Second, we perform bootstrap sampling of the total read rate over unmasked positions in order to model measurement error of true read rates. Under the assumption that each unmasked position is identically distributed, MAJIQ performs nonparametric sampling with replacement to draw from a distribution with identical mean and variance as the observed positions (see Section 2.4.7). Since we assume that our read rates are generally overdispersed relative to the Poisson distribution, MAJIQ replaces nonparametric sampling with Poisson sampling when the nonparametric estimate of variance is less than the mean (i.e. underdispersed).

MAJIQ performs quantification of splicing events modeled as LSVs, which are defined by a splicegraph. A source (target) LSV is defined for an exon as a choice over the incoming (outgoing) edges to (from) that exon from (to) a different exon. In general, only LSVs with at least two edges are considered. MAJIQ builder prepares output files with raw and bootstrapped coverage for each junction/intron in each LSV for quick use by downstream quantifiers.

We observed that builds from many build groups or with high coverage tend to have increasingly complex splicegraphs and LSVs with many junctions. Many of these junctions are often lowly used in all the samples but were included in the splicegraph because they had enough raw reads and positions (noisy novel junctions) or are part of an unused annotated transcript. This motivated the MAJIQ simplifier, which allows junctions and introns to be masked from the

final splicegraph used for quantification. After the splicegraph is constructed using all input build groups, MAJIQ calculates the ratio of the raw read rate for each junction/intron relative to the other junctions/introns in each LSV. If a junction has consistently low coverage in each of the build groups relative to the other choices in the two LSVs it can belong to, it is "simplified" and removed from the final splicegraph. This reduces the complexity of the final splicegraph and quantified LSVs, making output files smaller and downstream quantification more efficient.

In summary, the MAJIQ builder combines transcript annotations and input RNA-seq experiments in order to build a splicegraph encoding all possible splicing events consistent with both annotations and data and to prepare read coverage for quantification in terms of LSVs. The MAJIQ builder's new approach for estimating intron read rates allows junction and intron coverage to be calculated once and reused as part of an incremental build for multiple analyses, unlike other methods that quantify intron retention. The MAJIQ builder also introduces an approach for simplifying the complexity that arises in splicing events when processing large numbers of experiments. Overall, this allows the MAJIQ builder to produce structural models of possible splicing events and read coverage for downstream quantification that scale to the setting of large numbers of RNA-seq experiments.

2.4.2. MAJIQ quantifiers

MAJIQ provides three methods for quantifying RNA-seq experiments. MAJIQ PSI, MAJIQ dPSI, and MAJIQ HET, which we introduce in this paper. MAJIQ PSI and dPSI, which were previously described in Vaquero-Garcia et al. (2016), quantify groups of experiments that are assumed to be replicates with a shared true value of PSI per group. MAJIQ PSI estimates a posterior distribution of PSI (Ψ) for a single group, while MAJIQ dPSI compares these distributions for two groups in order to estimate a posterior distribution for dPSI ($\Delta\Psi$). MAJIQ HET compares two groups of samples but drops the replicate experiments assumption, enabling analysis of more heterogeneous samples. Instead, experiments are quantified individually and groups are compared under the assumption that the true values of PSI are identically distributed between the two groups.

All three pipelines share the same underlying machinery for inferring posterior distributions for Ψ. Formally, Ψ for a junction in an LSV is defined as the fraction of expressed isoforms using the junction out of all expressed isoforms containing the LSV. This fraction is not directly observable. Instead, we observe the number of reads aligned r_j to each junction j in the LSV. We model each r_j as a realization of a binomial distribution over the isoforms with probability Ψ_j:

$$r_j \sim \text{Binomial}\left(\sum_{j \in \text{LSV}} r_j, \Psi_j\right). \tag{2.1}$$

We take a Bayesian approach to integrate prior knowledge of Ψ, allowing for improved estimation when there is low read coverage. This requires a prior distribution on Ψ. We previously observed that most values of Ψ are nearly zero or one, which can be modeled using a generalization of the Jeffrey's prior for an LSV with J junctions:

$$\Psi_j \sim \text{Beta}\left(\frac{1}{J}, 1 - \frac{1}{J}\right). \tag{2.2}$$

This prior is conjugate to the binomial likelihood, allowing for efficient closed-form estimation of the posterior distribution of Ψ_j given the observed number of reads:

$$\Psi_j \Big| \left\{ r'_j : j' \in \text{LSV} \right\} \sim \text{Beta}\left(\frac{1}{J} + r_j, 1 - \frac{1}{J} + \sum_{j' \neq j} r'_j\right). \tag{2.3}$$

Since MAJIQ build obtains bootstrap replicates of observed read rates, we perform this posterior inference on each set of bootstrap replicate read rates to obtain an ensemble of posterior distributions.

For MAJIQ PSI, we obtain this ensemble of posteriors for replicate experiments by adding the observed read rates from the experiments that pass more stringent reads and position thresholds than the builder. MAJIQ PSI treats the average of the posterior distributions as a final distribution over Ψ. It reports point estimates of Ψ as the mean of this distribution ($\mathbb{E}[\Psi]$) and saves a discretized version of the distribution for visualization in VOILA.

MAJIQ dPSI takes this a step further by using the posterior distributions on Ψ_1, Ψ_2 for two groups in order to compute $\Delta\Psi = \Psi_2 - \Psi_1$ between the two groups. We start by computing the distribution of $\Delta\Psi$ under the assumption of independence of Ψ_1 and Ψ_2 by marginalizing the product of their distributions:

$$\mathbb{P}_{\text{ind}}(\Delta\Psi) = \sum_{\Psi_2 - \Psi_1 = \Delta\Psi} \mathbb{P}(\Psi)_1 \, \mathbb{P}(\Psi)_2 \,. \tag{2.4}$$

We know that Ψ_1 and Ψ_2 are not independent, so we integrate our knowledge that $\Delta\Psi$ is usually close to zero as a prior on $\Delta\Psi$. Following our previous work, we formulate our prior $\mathbb{P}_{\text{prior}}(\Delta\Psi)$ as a mixture of three components: (1) a spike around $\Delta\Psi = 0$, (2) a broader centered distribution around $\Delta\Psi = 0$, and (3) a uniform slab. We determine our final posterior distribution on $\Delta\Psi$ by adjusting $\mathbb{P}_{\text{ind}}(\Delta\Psi)$ by the prior and renormalizing:

$$\mathbb{P}(\Delta\Psi) \propto \mathbb{P}_{\text{ind}}(\Delta\Psi) \, \mathbb{P}_{\text{prior}}(\Delta\Psi) \,. \tag{2.5}$$

MAJIQ dPSI computes point estimates of $\Delta\Psi$ using the posterior mean of the distribution ($\mathbb{E}[\Delta\Psi]$) and identifies confidence of measured changes in inclusion as posterior probabilities $\mathbb{P}(|\Delta\Psi| > C)$.

MAJIQ HET takes a different approach for comparing inclusion between two groups of experiments. MAJIQ HET drops the assumption of replicate experiments to consider heterogeneity in Ψ between experiments within a group. Instead, MAJIQ HET assumes that the values of Ψ per experiment in each of the groups come from the same distribution. We evaluate this assumption using null hypothesis significance testing. Null hypothesis significance testing is performed using one (or more) of four tests: (1) Welch's two-sample t-test, (2) Mann-Whitney U test, (3) Total Number of Mistakes (TNOM) test, and (4) InfoScore test. Welch's two-sample t test and Mann-Whitney U test are well-documented elsewhere (Welch, 1947; Mann and Whitney, 1947). Our implementation of Mann-Whitney U test computes exact p-values when there are at most 64 experiments and computes asymptotic p-values using normal approximation with

tie and continuity correction for larger samples. Meanwhile, the InfoScore and TNOM tests are adapted from ScoreGenes (Barash et al., 2004). The TNOM test evaluates how well a single threshold on PSI can discriminate between the observed values in the two groups. The Total Number of Mistakes is the minimum number of misclassified observations under the best possible thresholds. The distribution on TNOM when the distributions are equal are calculated using the closed-form formula in Ben-Dor et al. (2002) to obtain p-values. Similarly, the InfoScore test evaluates how well a single threshold discriminates between groups, but, instead of measuring misclassifications directly, it identifies the threshold with the highest mutual information between the threshold and the true group labels. MAJIQ HET uses the dynamic programming algorithm in Ben-Dor et al. (2002) to evaluate the distribution of InfoScore under the null hypothesis in order to obtain p-values. All four tests require observed values of Ψ per experiment, which is not directly observed. MAJIQ HET accounts for variable uncertainty per experiment in our estimations of Ψ by repeated sampling of Ψ from the posterior distributions of quantified samples. MAJIQ HET computes the p-value for each repeated sample of Ψ over quantified experiments and reports the 95th-percentile over the resulting p-values. These p-value quantiles are not calibrated, so MAJIQ HET also computes p-values with the posterior means of Ψ. MAJIQ HET also reports the median of the observed posterior means of Ψ for each group. These p-values and the difference between the median observed posterior means are used together downstream in VOILA for the identification of high-confidence differentially spliced LSVs.

2.4.3. Sample selection from GTEx

We selected from GTEx in the following way. We required all samples to have a RIN score of greater than 6. For performance evaluation we chose to evaluate a comparison between cerebellum and skeletal muscle. We randomly selected 150 samples from both tissues, excluding the same donor from being selected in both tissues. Samples were downloaded as FASTQ or as BAM and converted to FASTQ depending on when they were released. Samples that were part of v7 are available on SRA, so they were downloaded using SRA Tools (v2.9.6) as FASTQ files. New samples from the v8 release were only available as BAMs on the cloud, so they were downloaded and converted to FASTQ using samtools (v1.9).

2.4.4. Simulated RNA-seq as ground truth

We used the expression quantification data from the GTEx v8 release as the basis for our simulations. Briefly, we downloaded publicly available gene- and transcript-level quantification tables for GTEx v8 from the GTEx portal (https://www.gtexportal.org/home/datasets). To match how the GTEx consortium performed these analyses, we downloaded the GRCh38 build of the reference genome sequence and gene models from v26 of the GENCODE annotation.

We selected 300 samples from GTEx to serve as the basis for 300 simulated samples, each real sample providing the expression distribution underlying one simulated sample (Table S3). To run BEERS, we first need to prepare four configuration files that are customized for the desired dataset: geneinfo, geneseq, intronseq, and feature quants. The geneinfo, genenseq, and intronseq files define the structure and sequence information for each simulated transcript. As a result, these three files are determined solely by the choice of reference genome build and annotation. The feature quant files are specific to each individual sample and define a distribution of transcript-level expression. First, we used the genome sequence and gene models to create the geneinfo, geneseq, and intronseq files. Since the genome is fixed across all simulated samples. We used the same set of these files to simulate all GTEx-derived samples. Next, we extracted TPM values for each sample from the GTEx transcript quantification table and used these distributions of TPM values to generate separate BEERS feature quant config files for each simulated sample. Lastly, to determine the total number of reads to simulate for each sample, we used the gene-level quantification file to count the total number of gene-mapping reads in each GTEx sample.

To simulated strand-specific reads with uniform coverage across no errors, substitutions, or intron retention events, we ran the BEERS simulator using the following command-line options: `-strandspecific -outputfq -error 0 -subfreq 0 -indelfreq 0 -intronfreq 0 -palt 0 -fraglength 100,250,500`.

We transformed ground-truth transcript abundances into ground-truth splicing quantifi-

cations for each splicing quantification tool, taking into account the tools' differing definitions of splicing events. First, we defined ground-truth abundances for each exon or junction by adding the abundances of all transcripts including the exon or junction. Then, for each tool, we adopted their splicing event definitions, mapping the exon/junction abundances to compute their splicing quantifications.

MAJIQ

MAJIQ reports splicing quantifications with respect to LSVs. Therefore, ground-truth values for PSI were calculated by dividing the ground-truth abundance of each junction by the sum of the ground-truth abundances for all junctions in each LSV.

rMATS

rMATS reports a different format file per event type. But since all of them are classical binary event types, all can be reduced to two paths events, inclusion and exclusion. Each file contains the exon that defines each of the ways, so we calculate the Ψ_{gt} as inclusion/(inclusion + exclusion) using the exon transcript combination to get the exons ground-truth abundances for all junctions in each LSV.

LeafCutter

LeafCutter reports splicing quantifications with respect to intron clusters composed of several junctions. Ground-truth values for LeafCutter's splicing ratios were calculated using ground-truth junction abundances, similar to MAJIQ.

SUPPA2

SUPPA2 reports classical events similarly to rMATS. So the approach we use here is similar to that tool. The main difference is that SUPPA2 reports the junctions coordinate in each one of the paths, so we use those junctions ground truth quantification to obtain the Ψ_{gt} as inclusion / (inclusion + exclusion).

Whippet

Whippet outputs a psi.gz that contains the psi quantification of an event. That PSI is their formulation of the quantification from inclusion and exclusion paths. Differently to SUPPA2 or rMATS, Whippet combines a set of junctions to define a path, emulating in that way a transcript (or a portion of it). So, in order to find Ψ_{gt} of those paths, we look for those transcripts that include all the junctions (and virtual junctions). We combine the expression of those transcripts to find the Ψ_{gt} of each path.

2.4.5. RNA-seq sample preprocessing before splicing analysis

We aligned RNA-seq reads from real and simulated GTEx samples to the human genome for splicing analysis with MAJIQ and other tools using the following procedure. Simulated GTEx samples were generated as pairs of FASTQ files. We performed quality and adapter trimming on each sample using TrimGalore (v0.4.5). Some tools require reads aligned to the genome. For these tools, we used STAR (v2.5.3a) to perform a two-step gapped alignment of the trimmed reads to the GRCh38 primary assembly with annotations from Ensembl release 94. Other tools required transcript quantifications relative to annotated transcripts. For these tools, we used Salmon (v0.14.0) using the trimmed samples to estimate transcript abundances.

2.4.6. Performance evaluations

We wrote a package of evaluation scripts, called validations-tools, in order to compare MAJIQ in terms of speed, memory footprint, accuracy, and reproducibility for each one of the following tools: rMATS, LeafCutter, SUPPA2, and Whippet. This package was written to allow future users to not only reproduce our results but to easily add future tools and repeat these kinds of analyses with different datasets.

We adjusted the tools parameters following recommendations by each tool's authors. Specific parameters are listed in Table S4. For these comparisons, we evaluated the methods' computational efficiency and ability to identify splicing differences.

First, we evaluated computational efficiency of the different methods. We evaluated

computational efficiency in terms of runtime and peak memory usage. Not all tools provide an extensive log of their execution, so, in order to measure wall time and memory usage, we used the output of `/usr/bin/time -v`. We ran each method for all pairs comparisons between 10 groups with increasing sample sizes on an Ubuntu Linux environment with 32 cores (Intel Xeon 2.7GHz and 64GB RAM).

Second, we evaluated the different methods' performance in quantifying splicing differences on simulated and real datasets. On the simulated datasets, where we know ground-truth differences in splicing between transcripts, we calculated true and false positive rates for the identification of splicing differences by each method. However, on real datasets, where no ground-truth is available, it is not possible to calculate true or false positive rates. Instead, we evaluated two metrics, reproducibility ratio (RR) and intra-to-inter ratio (IIR), on real (and simulated for comparison) data. The first metric, RR, measures the internal consistency of differential splicing tools. This internal consistency is reflected in the assumption that each tool should identify roughly the same events when repeating a comparison between two groups using different samples. We quantify this by performing two such comparisons and computing the fraction of the top n differentially-spliced events in the first comparison that are also in the top n events of the second comparison. This produces a "reproducibility-ratio" curve, RR(n) for the method as a function of the number of top events. If the first comparison yields N "significant" events, RR(N) is called the reproducibility ratio. For the specific case of MAJIQ, we note that in order to comparisons of LSV-type events more comparable to classic AS events such as used by rMATS, we filtered out overlapping LSVs (i.e. those that share junctions) in order to avoid double-counting classic AS events. For example, a classic exon-skipping event would have matching source and target LSVs that overlap. However, we note that this filtering only reduces N_A but does not affect the reproducibility curves (apart from extending to a different value of N_A) (Fig. S7). Although reproducibility of a method on real data is a scientifically important goal, it is not a sufficient goal because highly biased methods can be highly reproducible. To address this limitation, the second metric, IIR, is based on the principle that comparisons between (inter-) two groups should have many more significant events than comparisons within (intra-) a group.

Furthermore, significant events within the group are likely false positives. This is quantified by computing the ratio of the number of significant events from an intra-group comparison to the number of significant events from an inter-group comparison. We evaluated these metrics for each tool with varying sample sizes to identify which methods outperformed each other in different settings.

Event-level evaluations

In these evaluations we check reproducibility and accuracy of reported differentially spliced events by the various tools shown in Figure 2. As we describe in the main text, each tool defines alternative splicing events differently so that direct comparison of the events or their number between tools is not possible. Thus, when using real data each method was assessed by its own set of reported events to compute reproducibility ratios (RR) and intra-to-inter ratio (IIR) as in Figure 2d,e.

In contrast, when using GTEx based simulated data we do have the "ground truth" (denoted "gt" below) for the abundance of each transcript. We thus use these values to summarize Ψ and $\Delta\Psi$ observed in each method reported AS events and assess accuracy using the following definitions:

- True Positive: $\max \Delta\Psi_{\text{tool}} \geq 20\%$ and $\text{pvalue}_{\text{tool}} \leq 0.05$ and $\max \Delta\Psi_{\text{gt}} \geq 20\%$
- True Negative: $\max \Delta\Psi_{\text{tool}} < 5\%$ and $\text{pvalue}_{\text{tool}} > 0.05$ and $\max \Delta\Psi_{\text{gt}} < 5\%$
- False Positive: $\max \Delta\Psi_{\text{tool}} \geq 20\%$ and $\text{pvalue}_{\text{tool}} \leq 0.05$ and $\max \Delta\Psi_{\text{gt}} < 5\%$
- False Negative: $\max \Delta\Psi_{\text{tool}} < 5\%$ and $\text{pvalue}_{\text{tool}} > 0.05$ and $\max \Delta\Psi_{\text{gt}} \geq 20\%$
- Ambiguous: all other cases (when either $\Delta\Psi \in [5\%, 20\%)$ or when $\Delta\Psi$ and pvalue reported by the tool conflict),

where max is taken over all junctions/introns that belong to each AS event.

The above definitions were used to assess accuracy at the event level for each method,

as shown in Figure S1, and also served as the base for gene level evaluations described below.

Gene-level evaluations

To facilitate more direct comparison between the different methods shown in Figure 2 we aggregated each tool AS events and their respective annotation as TP, TN, FP, and FN as given above to assess gene level performance. Naturally, gene level labels of TP, TN, FP and FN are defined based on the events they contain. The gene level labels are easy to define as positive or negative when all AS events embedded in it are considered positive or negative respectively. The problem arises when a gene has some of its events as false positives and false negatives. In that case, we prioritize the labels according to the following order: FP, FN, TP, TN. This means for example that an occurrence of a false positive event in a gene (according to the method's specific event definition) would be counted as a false positive gene even if some other events were correctly labeled as true negative or even true positives. The rationale for this prioritization is that (a) positive events are expected to be rare and (b) we care the most about trying to validate or follow up on wrong hits (false positives) followed by missing true changes (false negatives).

2.4.7. Procedure for bootstrapped readrates from per-position reads

MAJIQ's bootstrapping procedure can be defined as follows. Without loss of generality, consider a single junction. For each RNA-seq read aligned with a split for this junction, we define the read's position relative to the junction (or vice-versa) with a position i and count the number of reads associated with each position, which we call S_i.

These raw readrates include biases that we would like to correct for; in particular, we define an explicit procedure for removing stacks by comparing the number of reads at each position against a Poisson model using the observed readrates at all other positions, which results in a set of stack-corrected nonzero readrates R_i for $i \in \{1, \dots, P\}$, where P is the number of nonzero positions after stack removal. These are the observed units for bootstrapping, so to emphasize:

$R_i \equiv$# of RNA-seq reads for i-th position, (observed readrates)

$i \in \{1, \ldots, P\}$. (nonzero positions after stack removal)

Other methods typically sum directly over positions R_i (really S_i since they generally also ignore read stacks) to produce a total junction readrate for use in quantification:

$$R \equiv \sum_{i=1}^{P} R_i. \qquad \text{(observed total junction readrate)}$$

Since we are unsatisfied with uncertainty/variance accounted for by directly using R, we generate samples from a bootstrap distribution over the P nonzero positions.

If we make the assumption that we are given the number of nonzero positions P and that the underlying readrate for each of these positions is independent and identically distributed with finite mean $\mathbb{E}[R_i] = \mu$ and variance $\mathbb{V}[R_i] = \sigma^2$, we can derive the mean and variance of our observed total readrate:

$$\begin{aligned}
\mathbb{E}[R] =& \mathbb{E}\left[\sum_{i=1}^{P} R_i\right] \\
=& \sum_{i=1}^{P} \mathbb{E}[R_i] \\
=& \mu P, && \text{(observed total readrate mean)} \\
\mathbb{V}[R] =& P\sigma^2. && \text{(observed total readrate variance)}
\end{aligned}$$

If we were able to take two samples for the observed total readrate (i.e. R and R'), their

difference has mean 0 and variance $2P\sigma^2$.

We define our bootstrapping procedure over observed nonzero reads $R_1, \ldots, R_P$ to generate bootstrapped total reads $\widehat{R}, \widehat{R}', \ldots$ such that the variance of the difference between bootstrap samples would be equivalent to that of the difference between two samples from the true distribution (i.e. $2P\sigma^2$). In order to do this, we take $P-1$ samples from $\{R_1, \ldots, R_P\}$ with replacement and scale their sum by $P/(P-1)$.

It is straightforward to see that the bootstrapped total readrate has the same mean as the observed total readrate. In order to prove that the variance of the difference between two sample matches, we note that the covariance $\text{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right)$ between any two draws from the observed per-position readrates with $Z_i \sim \text{Uniform}(P)$ is:

$$\begin{aligned}\text{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right) =& \mathbb{E}\left[\left(R_{Z_k} - \mu\right)\left(R_{Z_{k'}} - \mu\right)\right] \\ =& \mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] - \mu^2.\end{aligned}$$

. We note that $\mathbb{E}\left[R_i R_j\right] = \sigma^2\delta_{ij} + \mu^2$ (where δ_{ij} is the Kroencker delta). When $k = k'$, it follows that $\mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] = \sigma^2 + \mu^2$. Otherwise, the law of total expectation gives:

$$\begin{aligned}\mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] =& \mathbb{E}\left[\mathbb{E}\left[R_{Z_k} R_{Z_{k'}} \middle| Z_k, Z_{k'}\right]\right] && \text{(given } k \neq k'\text{)} \\ =& \frac{1}{P^2}\sum_{i=1}^{P}\sum_{j=1}^{P}\sigma^2\delta_{ij} + \mu^2 \\ =& \mu^2 + \frac{1}{P}\sigma^2.\end{aligned}$$

Combining the two cases, we have:

$$\begin{aligned}\mathbb{E}\left[R_{Z_k} R_{Z_{k'}}\right] =& \delta_{kk'}\left(\mu^2 + \sigma^2\right) + \left(1 - \delta_{kk'}\right)\left(\mu^2 + \frac{1}{P}\sigma^2\right) \\ =& \mu^2 + \frac{1}{P}\sigma^2 + \delta_{kk'}\frac{P-1}{P}\sigma^2. && \text{(second moment sampled readrate)}\end{aligned}$$

Therefore,

$$\text{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right) = \frac{1}{P}\sigma^2 + \delta_{kk'}\frac{P-1}{P}\sigma^2. \qquad \text{(covariance sampled readrate)}$$

Thus, the variance of the bootstrapped total readrate is

$$\begin{aligned}
\mathbb{V}\left[\widehat{R}\right] =& \frac{P^2}{(P-1)^2}\mathbb{V}\left[\sum_{k=1}^{P-1} R_{Z_k}\right] \\
=& \frac{P^2}{(P-1)^2}\sum_{k=1}^{P-1}\sum_{k'=1}^{P-1}\text{Cov}\left(R_{Z_k}, R_{Z_{k'}}\right) \\
=& \frac{P^2}{(P-1)^2}\sum_{k=1}^{P-1}\sum_{k'=1}^{P-1}\frac{1}{P}\sigma^2 + \delta_{kk'}\frac{P-1}{P}\sigma^2 \\
=& 2P\sigma^2. \qquad \text{(true bootstrap readrate variance)}
\end{aligned}$$

But we want the variance of the difference between two samples from the bootstrap procedure. So, we calculate the covariance between two distinct samples:

$$\begin{aligned}
\text{Cov}\left(\widehat{R}, \widehat{R}'\right) =& \frac{P^2}{(P-1)^2}\sum_{k=1}^{P-1}\sum_{k'=1}^{P-1}\frac{1}{P}\sigma^2 \\
=& P\sigma^2. \qquad \text{(covariance between samples of } \widehat{\mathcal{P}})
\end{aligned}$$

Therefore, we find that:

$$\begin{aligned}
\mathbb{V}\left[\widehat{R} - \widehat{R}'\right] =& 2\mathbb{V}\left[\widehat{R}\right] - 2\,\text{Cov}\left(\widehat{R}, \widehat{R}'\right) \\
=& 4P\sigma^2 - 2P\sigma^2 \\
=& 2P\sigma^2. \qquad \text{(bootstrap total readrate variance as difference)}
\end{aligned}$$

In practice, the observed nonzero positions can lead to a bootstrap distribution with variance less than its mean (underdispersed). We generally expect readrates to follow a Poisson or negative binomial (overdispersed) distribution, so in these cases, we fall back to parametric bootstrapping with a Poisson distribution with mean μP. Otherwise, we use the nonparametric

bootstrap sampling procedure as described above.

CHAPTER 3

Mapping splicing variations in clinically accessible and nonaccessible tissues

3.1. Introduction

The previous chapter described how we built MAJIQ v2 to analyze large and heterogeneous RNA-seq datasets as part of our goal of improving the molecular diagnosis of patients with suspected Mendelian disorders. In this chapter, we use the MAJIQ v2 methodology to answer the question of how tissue-specificity affects what splicing variations we can expect to capture in the clinical setting.

As laboratories move to measuring the transcriptome directly with RNA-seq, one challenge they face is tissue specificity. Tissue-specific expression is the most discussed complicating factor of RNA-based analysis, as a gene must be expressed in the tissue to be studied (Cummings et al., 2017; Gonorazky et al., 2019). Alternative splicing between tissues is less often addressed, and further complicates analysis. If a tissue other than the tissue of clinical interest is tested, a gene that is expressed in both tissues can still be spliced differently. Thus, splicing defects affecting the tissue of clinical interest might not be realized in the tested tissue despite the gene being expressed in both. Therefore, one tissue can be an adequate proxy for a gene's splicing in a different tissue only if it is both expressed and spliced similarly (Figure 3.1b).

Clinicians and researchers can only perform RNA-seq on tissues they have access to. In the clinical setting, these tissues are typically limited to those from blood or skin biopsies: whole blood, Epstein–Barr virus (EBV)-transformed lymphocytes, and fibroblasts. We refer to these three as clinically accessible tissues (CATs). At the same time, laboratories are often interested in pathology occurring in inaccessible tissues (non-CATs, e.g., brain, heart, etc.).

Several recent studies consider limitations of using RNA-seq from CATs for clinical diagnosis. Frésard et al. (2019) demonstrate that RNA-seq in whole blood can make some diagnoses in patients from diverse disease categories. However, Cummings et al. (2017) studying

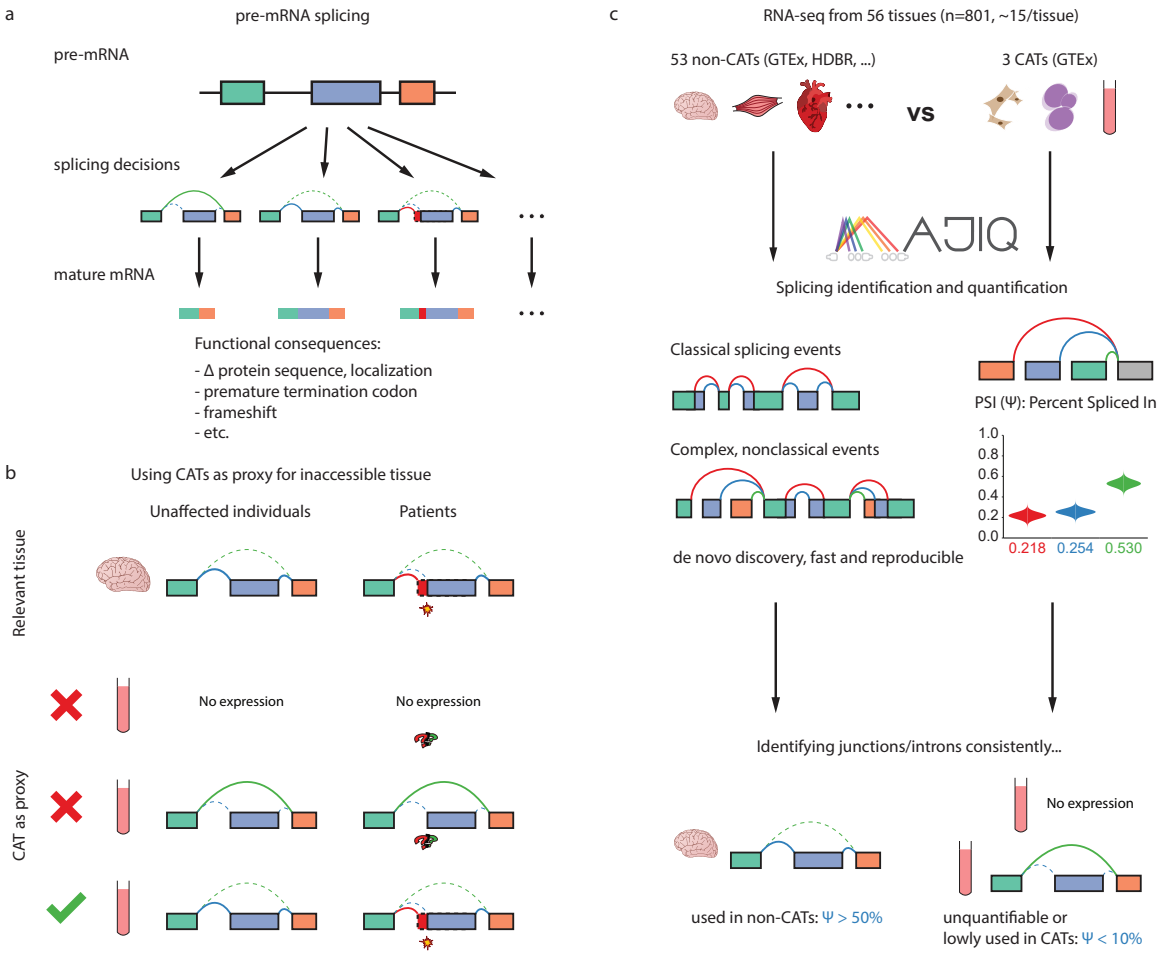


Figure 3.1: Identification of splicing events inadequately represented by clinically accessible tissues (CATs). (a) Different precursor mRNA (pre-mRNA) splicing decisions can have significant, potentially pathogenic, functional consequences. (b) Splicing events in inaccessible tissues (non-CATs) can only be adequately represented by accessible tissues as a proxy if the gene is both expressed and similarly spliced. (c) We used MAJIQ on RNA-seq samples from 56 different tissues to define and identify inadequately represented splicing events between inaccessible and accessible tissues.

a cohort of patients with neuromuscular disease, performed RNA-seq on skeletal muscle biopsies collected as part of the clinical diagnostic workup motivated by low gene expression of many known neuromuscular disease genes in whole blood and fibroblasts. Gonorazky et al. (2019) further show that they can identify aberrant splicing in muscle that they would not detect from fibroblasts from the same patients. While serving as an important proof of concept of the limitations of RNA-seq with CATs in neuromuscular disease genes, these studies raise the more general question: what are the limitations of RNA-seq in CATs across other non-CATs and genes in general, and how can one quantify them? An answer to this question could inform the selection of the best tissue to send for RNA-seq in clinical practice by evaluating the degree to which each CAT faithfully represents splicing in genes and tissues of interest for different patient phenotypes.

We address the question by considering splicing in nonaccessible versus accessible tissues in terms of splicing events. We model splicing events as local splicing choices either starting or ending at a single exon (the reference exon) in a given gene. These local choices involve splice junctions or retained introns that are included in the gene's transcripts. These splicing events include constitutive splice junctions and local splicing variations (LSVs), which are splicing events where the reference exons can be spliced to multiple RNA segments, thus allowing variation. We identify and quantify splicing events from RNA-seq data using the MAJIQ v2 toolkit for splicing detection, quantification, and visualization.

When using splicing in CATs as a proxy to splicing in some other tissue of interest, we consider three possible scenarios or splicing event categories (Figure 3.1b): (1) the event is unquantifiable in the CAT due to low gene expression and/or sequencing depth, (2) the event is quantifiable but spliced differently, and (3) the event is quantifiable and not spliced differently. We further focus on splicing events that are consistently included, meaning that they are similarly quantified in nearly all samples for a given tissue type. Naturally, categorizing events into these scenarios depends on the thresholds used to define them. Here, we define consistently spliced events in non-CATs to be events with a junction or retained intron with $\Psi > 50\%$ in more

than 85% of samples (Figure 3.1c). We emphasize finding the subset of these events that correspond with either of the first two scenarios where splicing measured in a CAT inadequately represents splicing in the non-CAT. We define these events as those that are unquantifiable or have $\Psi < 10\%$ in more than 85% of a CAT's samples.

In this work, we analyze 53 adult tissues in the Genotype-Tissue Expression Project (GTEx) (GTEx Consortium et al., 2017) and 3 fetal tissues from the Human Developmental Biology Resource (HDBR) (Lindsay et al., 2016) (cerebellum, cortex) and ArrayExpress accession E-MTAB-7031 (Pervolaraki et al., 2018) (heart) (Figure 3.1c). We map all transcriptome variations across these data sets, contrasting splicing between CATs and non-CATs. We make our analyses accessible as an online resource, which we call MAJIQ-CAT (https://tools.biociphers.org/majiq-cat). This online resource has been designed for clinicians and researchers interested in obtaining patient RNA-seq in the context of Mendelian disease. With MAJIQ-CAT, these users can explore how faithfully different CATs represent splicing in their specific genes and tissues of interest, informing their choice of patient tissue to collect. Finally, we discuss implications for RNA-seq in clinical practice and the need for alternative solutions for the genes and tissues that are inadequately represented by CATs.

3.2. Materials and Methods

3.2.1. Sample selection criteria

We used RNA-seq data for samples from 56 different tissue types: 53 adult tissues and 3 fetal tissues. We obtained samples for all 53 adult tissue types from GTEx (dbGaP accession phs000424). Meanwhile, we obtained samples for fetal cerebral cortex and cerebellum from HDBR (ArrayExpress accession E-MTAB-4840), and we obtained samples for fetal heart from ArrayExpress accession E-MTAB-7031.

All RNA-seq data have been previously described and were derived from tissues collected ethically. For GTEx, tissues from deceased donors are not legally classified as human subjects research under US Code of Federal Regulations Title 45, Part 46 (45 CFR 46) but were collected under written or recorded verbal authorization from next of kin, while tissues collected from

living donors were only included after full, written consent was obtained. HDBR is a tissue bank regulated by the UK Human Tissue Authority, and samples in HDBR were collected with appropriate maternal written consent and approval from a National Research Ethics Service (NRES) Committee. The fetal heart samples were acquired with informed written parental consent obtained from all subjects under approval of NHS Lothian, the University of Edinburgh Research Governance Hope, and the University of Leeds Ethical Committee.

We restricted sample selection from each of these data sets/tissue types using available metadata as follows. We restricted selection to unique donors per tissue type. This restriction was relevant to both GTEx and HDBR, which include donors contributing multiple sample per tissue type. Available HDBR metadata did not suggest criteria for preferring one sample over another, so we restricted selection to the first available sample per donor. However, GTEx metadata includes information on the number of megabases per sample, so we restricted selection to the sample with the largest size. GTEx metadata also included further information that we used; specifically, we further restricted selection to samples that (1) were hosted by National Center for Biotechnology Information (NCBI), (2) had matched genome sequencing data, (3) had an average spot length of 152bp, (4) were not flagged by GTEx as a sample to remove (SMTORMVE), and (5) had an RNA integrity number (RIN) score greater than 6.

Given these restrictions, we selected up to 15 samples per tissue type for further analysis. We chose 15 samples as the maximum number of samples per tissue group because preliminary analysis using MAJIQ indicated that reproducibility for tissue-specific differential splicing analysis saturates with around 15 samples in GTEx (data not shown). Consequently, when there were more than 15 samples meeting the above criteria for a particular tissue type, we randomly selected 15 samples among them. For the other tissue types, we kept all samples meeting criteria for further analysis.

3.2.2. Sample read alignment

We aligned RNA-seq reads from the selected samples to the human genome for splicing analysis with MAJIQ using the following procedure. We downloaded selected samples as FASTQ

files using SRA Tools (v2.9.6). We performed quality and adapter trimming on each sample using TrimGalore (v0.4.5). We used STAR (v2.5.3a) to perform a two-step gapped alignment of the trimmed reads to the GRCh38 primary assembly with annotations from Ensembl release 94.

3.2.3. Gene expression quantification

We quantified gene expression in each of the samples. We quantified transcript abundances in transcripts per million (TPM) using Salmon (v0.13.1) quasi-mapping on the trimmed reads for each sample with a transcriptome built from Ensembl release 94 annotations. We aggregated the quantifications to gene expression by taking the sum of abundances for the transcripts associated with each gene.

3.2.4. Splicing identification and quantification using MAJIQ

First, we used MAJIQ v2 with Ensembl release 94 annotations to identify/model the set of all possible annotated and *de novo* splicing events across our samples. We then quantified these splicing events for each sample, considering an event to be quantifiable for a given sample if it had at least one junction with at least ten supporting reads starting from at least three unique positions. We estimated the percent spliced in (PSI or Ψ) for the junctions and retained introns in each quantifiable splicing event. For the quantifiable LSVs, we used MAJIQ to estimate PSI. Meanwhile, we assigned $\Psi = 100\%$ to the quantifiable constitutive junctions, as they were the only choice for inclusion in their respective events.

Finally, we identified and filtered out ambiguous splicing events per sample. We defined ambiguous splicing events as events containing junctions or retained introns that were in quantified splicing events assigned to more than one gene. These ambiguous assignments occur because of the presence of overlapping genes, especially combined with the unstranded nature of the RNA-seq experiments we used. The resulting nonambiguous quantifications were used to identify relevant consistent and tissue-specific differences between CATs and non-CATs.

3.2.5. Identifying relevant splicing events

To determine the extent to which splicing in non-CATs is inadequately represented by splicing in CATs, we first defined which splicing events for each non-CAT we would consider changes in usage for. For each non-CAT, we consider the set of consistent splicing events, which are splicing events with a junction or retained intron that is highly included in nearly all samples for their tissue type. Specifically, we considered splicing events with a junction or retained intron quantified as $\Psi > 50\%$ in more than 85% of the samples for each non-CAT.

We then evaluated how well splicing quantified in CATs reflected splicing in these consistent splicing events. To do so, we identified the subset of these events for which usage in CATs was consistently low or unquantified. Specifically, we identified which events were either unquantifiable or had $\Psi < 10\%$ for the same junction or retained intron in more than 85% of the samples for each CAT. We call these splicing events inadequately represented in their respective CAT.

3.2.6. Analysis of genes with consistently used splicing events

We then aggregated information about these consistent and inadequately represented splicing events to their respective genes for each tissue. That is, we determined which genes had consistent splicing events in each non-CAT and the subset of these genes for which these events were inadequately represented for each CAT. We evaluated gene expression for the inadequately represented genes to assess how inadequately represented splicing related to low gene expression versus tissue-specific alternative splicing. We also evaluated which of the inadequately represented genes were annotated as disease-causing. We obtained our list of disease-causing genes by combining annotations from ClinVar (gene annotations from the table named "gene_condition_source:id") and HGMD 2018.3 (inferred from variants classified as DM (disease-causing mutation), the highest level of deleteriousness).

3.2.7. Data access and software

Sequencing data used for this analysis are available in dbGaP under accession phs000424 and ArrayExpress under accessions E-MTAB-4840 and E-MTAB-7031. Software versions, resources, and specific parameters used are listed in Table S1 of Aicher et al. (2020). The analysis was implemented for reproducible execution as a Snakemake pipeline. Links to source code for the analysis and online resource, MAJIQ-CAT, are listed in Table S2 of Aicher et al. (2020) and have been deposited to Zenodo.

3.3. Results

Our sample selection procedure yielded a data set with $n = 801$ RNA-seq samples for 53 non-CATs and 3 CATs. Seven hundred sixty-two samples came from GTEx, 30 fetal brain samples came from HDBR, and 9 fetal heart samples came from E-MTAB-7031. We selected and processed 15 samples for each tissue except for bladder, cervix (ectocervix and endocervix), fetal heart, and fallopian tube, where we selected all available samples that met our criteria.

Across all samples, we identified a total of 239,406 quantifiable splicing events (124,909 LSVs and 114,497 constitutive junctions) in 25,494 genes. Per sample, we quantified a median of 116,153 splicing events (65,481 LSVs and 50,719 constitutive junctions) in 12,872 genes. We then identified and removed ambiguous splicing events with junctions or retained introns associated with multiple genes, leaving a total of 223,590 splicing events (117,728 LSVs and 105,862 constitutive junctions) in 26,643 genes with a per-sample median of 107,174 splicing events (60,591 LSVs and 46,825 constitutive junctions) in 12,049 genes.

Among quantified LSVs and constitutive junctions, we identified in each non-CAT a median of 73,669 junctions or retained introns in 9,966 genes that were consistently used (Figure 3.2a). Looking at these same events in CATs, we found that 27.7% were inadequately represented in at least one CAT (3925 or 40.2% of genes); 4.4% were inadequately represented by all CATs (609 or 6.3% of genes).

We compared the quantities of inadequately represented splicing per CAT and non-

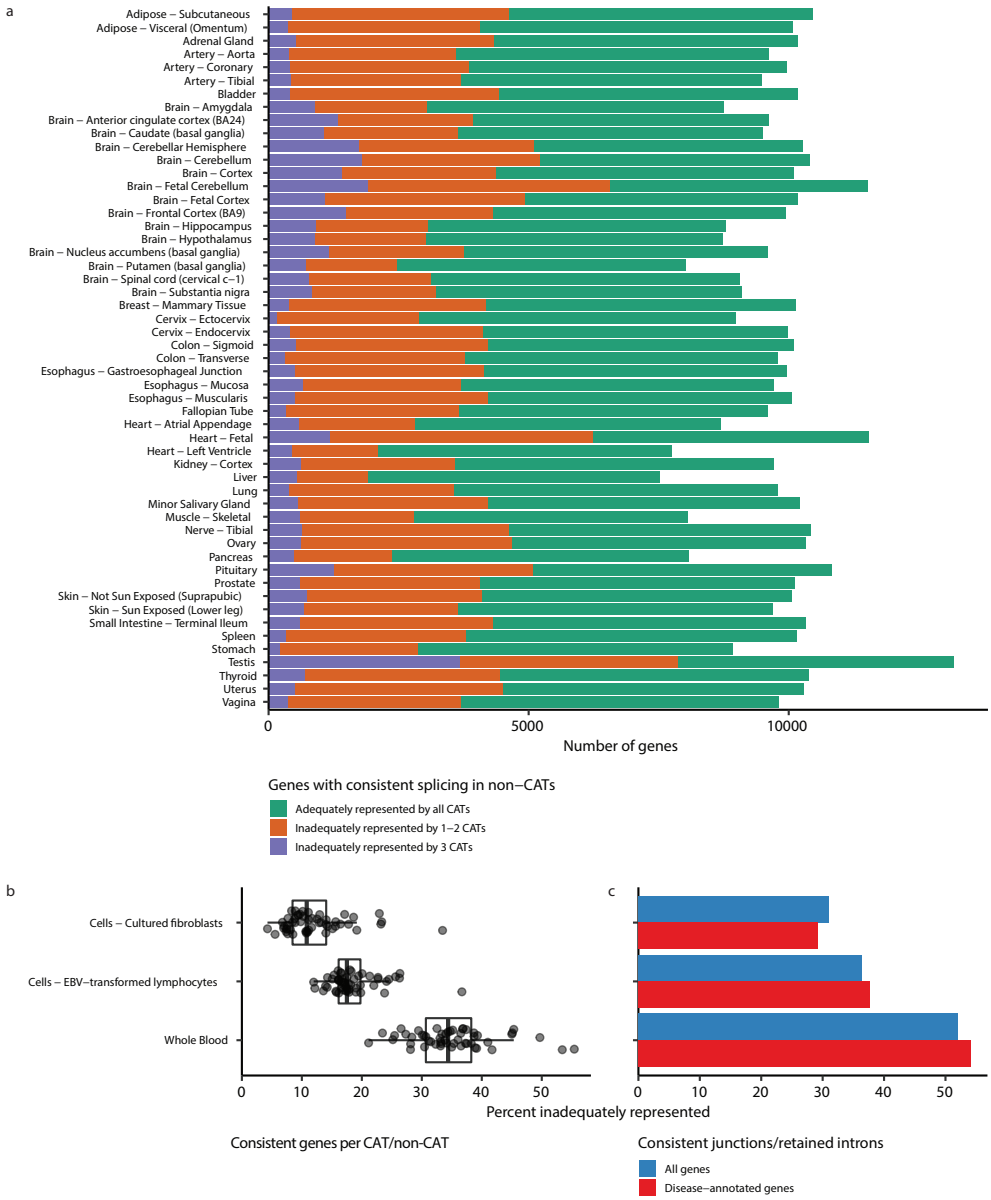


Figure 3.2: Mapping transcriptome variations identified in clinically accessible tissues (CATs) vs. non-CATs. (a) Of an average of 9966 genes with consistently spliced events per non-CAT, 3925 (40.2%) were inadequately represented in at least one CAT, with 609 (6.3%) being inadequately represented by all CATs. (b) The percentages of genes with consistently spliced events that were inadequately represented over the 53 non-CATs were lowest in fibroblasts and highest in whole blood. (c) The percentage of junctions/retained introns that were consistently used in at least one non-CAT that were inadequately represented by each CAT was lowest in fibroblasts and highest in whole blood.

CAT. The median percentage of genes with consistently spliced events that were inadequately represented across non-CATs was 10.8% for fibroblasts in comparison to 17.5% for EBV-transformed lymphocytes and 34.4% for whole blood (Figure 3.2b). The percentage was lowest in fibroblasts and highest in whole blood for each non-CAT except for spleen, for which the percentage was lowest in EBV-transformed lymphocytes. Considering all consistently spliced junctions/retained introns across non-CATs, we found that the percentage of inadequately represented splicing was also lowest in fibroblasts and highest in whole blood (Figure 3.2c).

We examined potential strategies for decreasing inadequately represented splicing compared with current practice. To evaluate the benefit of potentially acquiring and sequencing two CATs instead of one, we quantified for each CAT the percentages of inadequately represented events and genes that were not inadequately represented in the other two CATs (Figs. S3–S5 in Aicher et al. (2020)). We also considered how primary skin types, which are typically not accepted by clinical laboratories for nondermatologic conditions, would perform as alternative CATs by calculating their percentages of inadequately represented junctions and genes (Fig. S6 in Aicher et al. (2020)).

We further investigated the expression and pathogenicity of inadequately represented genes. The maximum median gene expression across inadequately representing CATs was less than 1 TPM in 52.1% of inadequately represented genes (Figure 3.3a). However, an average of 5.8% of inadequately represented genes per non-CAT (217 genes) are expressed with greater than 10 TPM. Meanwhile, a median of 29.2% of inadequately represented genes per non-CAT had variants annotated as disease-causing in either ClinVar or HGMD (Figure 3.3b).

To facilitate the interrogation of specific genes and splicing variations of interest by clinicians, we developed MAJIQ-CAT (https://tools.biociphers.org/majiq-cat). MAJIQ-CAT is an online resource that provides panels with which users can select genes and non-CATs to look at how well CATs represent splicing across their genes of interest, both globally (Figure 3.4a) and looking at individual splicing events in a specific gene (Figure 3.4b). Genes can be selected from predefined lists of genes (e.g., from ClinVar, ClinGen, etc.) or custom lists provided by the user

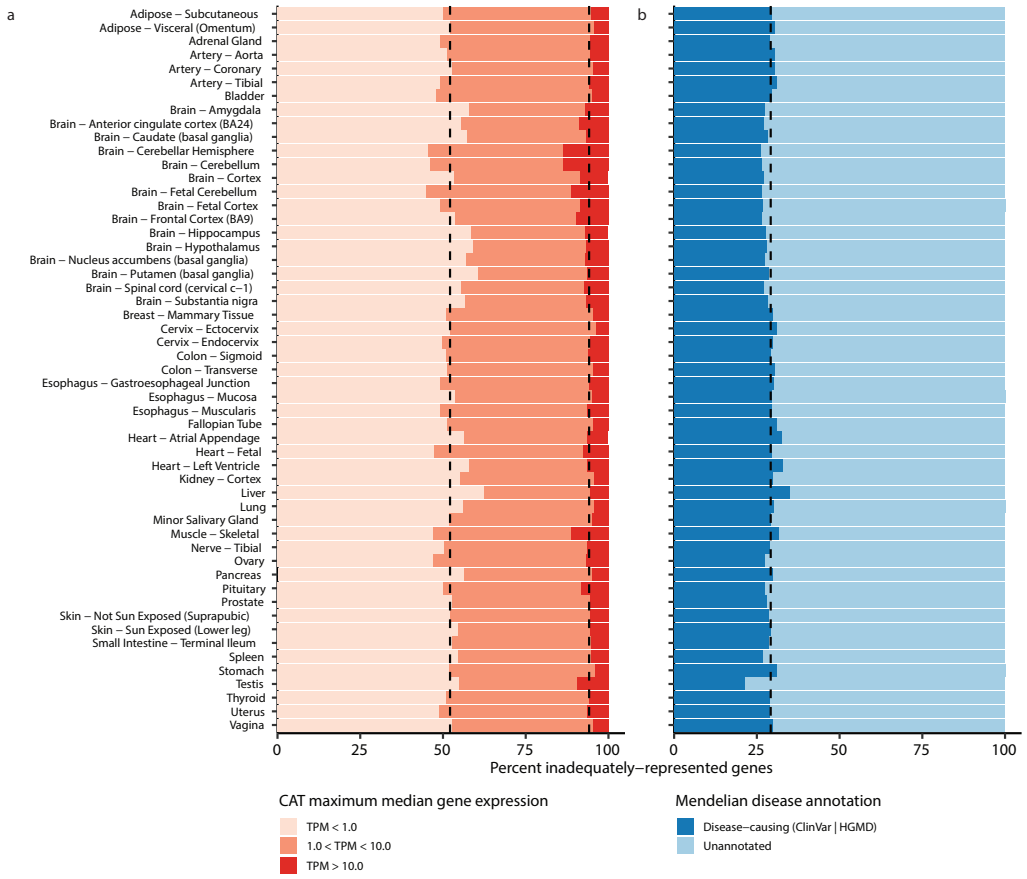


Figure 3.3: Expression and disease-gene relationship of inadequately represented genes. (a) The majority of inadequately represented genes are lowly expressed (TPM < 1) in clinically accessible tissues (CATs), but an average of 217 genes (5.8%) are well expressed (TPM > 10) in at least one inadequately representing CAT. (b) An average of 29.2% of inadequately represented genes are annotated as having disease-causing variants.

either interactively or by uploading a text file. Non-CATs can be selected similarly. Changes to these inputs automatically regenerate plots and tables describing the consistent and inadequately represented genes. Individual genes can further be explored by clicking their names to load an additional page that displays their tissue-specific gene expression and splicing events. For example, if a laboratory was interested in studying intellectual disability as a phenotype, they could focus on brain non-CATs and genes associated with the corresponding Human Phenotype Ontology (HPO) term (HP:0001249), finding 1232 genes with consistent splicing in at least one of the brain tissues (Figure 3.4a). They could further look for genes that are expressed but inadequately represented by filtering the table by expression; in this example, setting a minimum of TPM >10 yields a list of 139 genes. Clicking into one of the resulting genes (e.g., *MEF2C*) leads to another page with comparisons of splicing in the CATs and the brain tissues, demonstrating where the inadequately represented splicing events are and the distributions of PSI in each tissue for each event (Figure 3.4b). We developed additional, more detailed example scenarios to demonstrate how to use MAJIQ-CAT in the supplementary information of the published paper (Aicher et al., 2020).

3.4. Discussion

In this study, we present a comprehensive analysis of RNA splicing events that consistently occur in clinically inaccessible tissues, focusing on how corresponding events take place in clinically accessible tissues. While clinicians and scientists are often interested in what takes place in the inaccessible tissues as part of disease pathology, laboratories can only measure the accessible tissues as a proxy. Thus, these results inform clinicians and scientists as to where RNA-seq is limited, especially with respect to previously underappreciated tissue-specific splicing, and suggest when specific clinically accessible tissues should be preferred over others or when alternative approaches to clinical RNA-seq are needed. By making the results interactively accessible through MAJIQ-CAT, we enable clinicians and scientists to more directly explore how these limitations impact specific genes and tissues of interest to them.

Previous studies analyzing patient RNA-seq from CATs demonstrated that these data

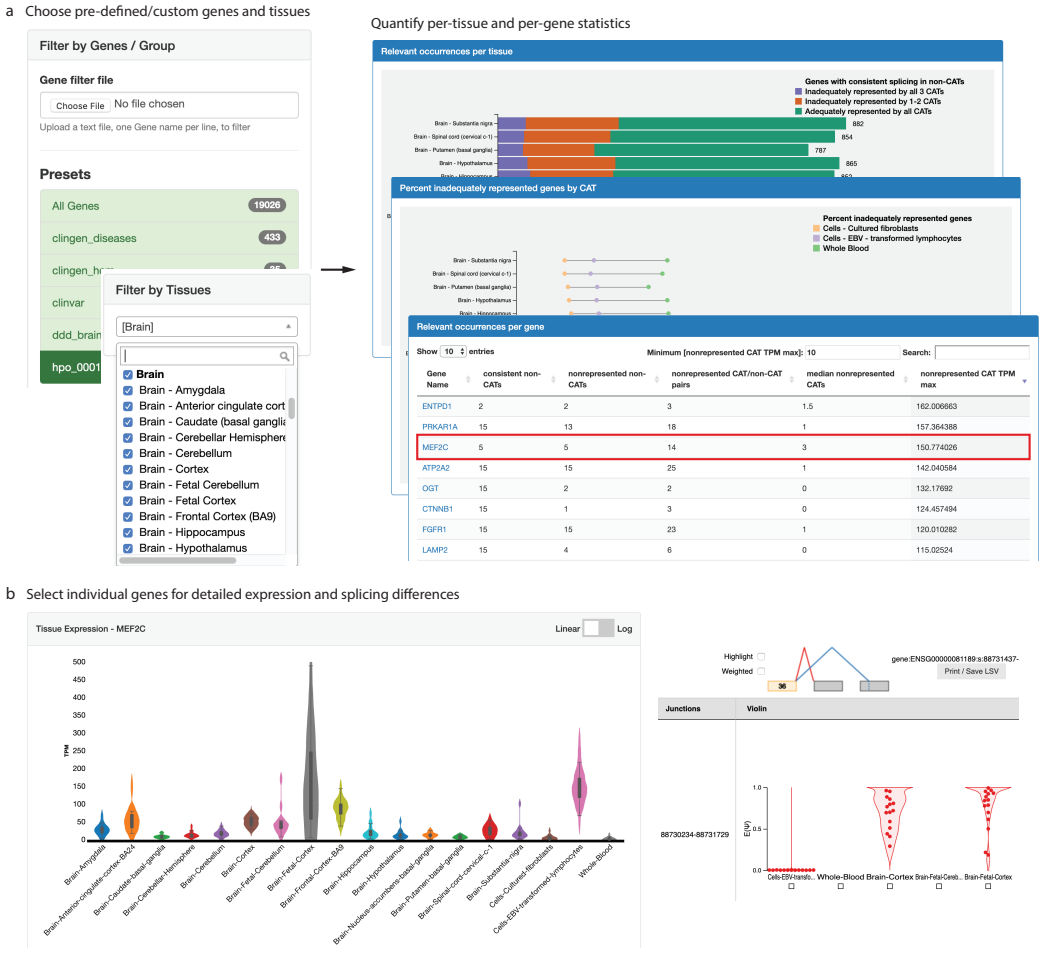


Figure 3.4: MAJIQ-CAT enables clinicians and scientists to explore inadequate representation of splicing by clinically accessible tissues (CATs) in specific genes and tissues of interest. (a) MAJIQ-CAT allows users to choose from predefined or custom gene sets and tissues (left) to quantify and understand the user-specific relevant limitations of RNA-seq in different accessible tissues (right). (b) Users can further explore individual genes for tissue-specific differences in gene expression and splicing. Shown here is a closer look at the gene MEF2C, with a violin plot of its expression in CATs and selected non-CATs (left) and violin plots of percent splicing inclusion (PSI) for one of its inadequately represented splicing events (right). See main text for more details.

could be used to identify rare disease genes and variants from a variety of disease categories (Kremer et al., 2017; Frésard et al., 2019). These studies demonstrate that although clinicians are typically limited to using CATs for patient RNA-seq, RNA-seq in those tissues can still improve the molecular diagnostic rate for suspected Mendelian disorders by identifying changes that were not identified using exome or genome sequencing alone. Our study provides an orthogonal, but related, result. Because we are typically limited to using CATs for patient RNA-seq, there are splicing events found in disease-relevant tissues and genes that will consistently be a blind spot in such studies.

Our study found that 40.2% of genes with consistent splicing events per non-CAT are inadequately represented by at least one CAT. This implies that clinicians and scientists interested in how one of the inadequately represented genes are spliced in the non-CAT in patients need to be careful about which clinically accessible tissues they measure as a proxy because at least one of the accessible tissues will not represent the splicing events well. We show that many of these genes are considered disease-causing (29.2% of the inadequately represented genes); thus, understanding these limitations is increasingly clinically relevant as RNA-seq enters clinical practice.

Considering these 40.2% of genes with inadequately represented splicing, the majority (52.1%) were associated with low gene expression (TPM < 1), as expected. However, we still find that 217 genes per non-CAT are highly expressed (TPM > 10) but spliced differently in CATs. The limitations of these genes for clinical RNA-seq would be missed by previous expression-first analyses, highlighting the novelty and impact of our splicing-first analysis.

For the other 59.8% of genes, we note that splicing in CATs may still not always adequately represent splicing in non-CATs. While they may not pass the stringent thresholds we set to define inadequately represented splicing present in most samples for a CAT ($\Psi < 10\%$ or unquantifiable in more than 85% of samples), splicing inclusion may take intermediate values or be highly variable between samples. Furthermore, even for splicing variations that are similar between tissues, they may still involve different tissue-specific regulation by different tissue-specific

factors. Thus, while we might expect variants in tissue-independent splicing sequence elements (e.g., canonical splice sites) to impact the different tissues similarly, variants in tissue-specific splicing enhancers or silencers could lead to tissue-specific defects that would not be represented by CATs.

It is also important to note that the results described here are dependent on the limitations and technical biases of current practices and technologies for poly-A selected RNA-seq. For example, sequencing with greater depths or read lengths than is typically done in common practice could potentially increase detection of lower-expressed genes/splicing events. Likewise, alternative and/or future approaches for mRNA isolation or sequencing will differentially impact detection of splicing across the genome. In particular, protocols including globin-depletion of whole blood would likely improve its performance as a CAT because globin genes account for the majority of expressed transcripts in GTEx whole blood. Since these data are not available in GTEx, we plan to evaluate globin-depleted whole blood as a CAT for a future update to MAJIQ-CAT. It will be important to re-evaluate differences in what we can detect between tissues as emerging technologies, such as long-read sequencing, enter common practice and replace current protocols for measuring clinical transcriptomes.

One important conclusion from the analysis performed here is that for the 3316 genes per non-CAT that are inadequately represented by one or two CATs, at least one CAT offers a better representation of the gene's splicing than the others. Thus, our study implies that researchers interested in one of these genes and tissues should have a preference for which clinically accessible tissue to collect. Summarizing across all genes with consistent splicing, we found that fibroblasts almost always had the lowest percentage of inadequately represented genes. Thus, our results suggest that researchers interested in all genes and tissues equally should prefer collecting patient fibroblasts if possible. However, clinicians and scientists are often interested in specific genes or tissues relevant to a specific biological process. Our online resource, MAJIQ-CAT, will enable clinicians, scientists, and laboratories to interactively explore which CATs are most relevant for representing the biology they care about and which genes and splicing events are most affected.

This work's predictions about fibroblasts being a more appropriate CAT for clinical analysis have since been corroborated by others in the literature. Murdock et al. (2021) analyzed patient RNA-seq from both whole blood and fibroblasts. In solved cases with both tissues, they were able to identify the putative causative defect in all the fibroblast samples but only half the whole blood samples.

Another important conclusion of this study is that there are 609 genes per non-CAT that are inadequately represented by all CATs. For these genes, using RNA-seq in any CAT as a proxy would have many limitations for studying splicing. In these cases, alternative approaches are likely necessary. One possible path forward is the use of in vitro differentiation or transdifferentiation of CRISPR-iPSCs or patient-derived cells toward tissue types of interest. Gonorazky et al. (2019) illustrated this possibility for transdifferentiated myotubes from patient fibroblasts as an alternative to skeletal muscle biopsy, although how these results would translate to other, more inaccessible, tissues remains to be explored. Another possible path forward is the use of *in silico* models of splicing (Barash et al., 2010b,a; Xiong et al., 2015; Jha et al., 2017; Zhang et al., 2019; Jaganathan et al., 2019; Cheng et al., 2019). Previous works in several labs have developed models of tissue-specific splicing but do not directly train models on genetic variants. Cheng et al. (2019) directly trains models to predict splicing changes using genetic variants but does not account for tissue specificity. More recent (Cheng et al., 2021) and future developments combining aspects of these models to produce predictions of tissue-specific splicing as a consequence of genetic variants could help us understand potential splicing defects in those genes where we do not have a good proxy. These alternative strategies could be combined with other orthogonal approaches, including predicted variant pathogenicity, to further advance detection of splicing variants in these inadequately represented genes.

In summary, in this study, we demonstrated and quantified the limitations of CATs to serve as a proxy for non-CATs for RNA splicing measured by RNA-seq. We highlighted how alternative splicing contributes to these limitations in addition to tissue-specific gene expression. In addition, we developed and have made available an online resource, MAJIQ-CAT, that will

allow clinicians and scientists to directly explore how these limitations affect specific genes and tissues of interest. MAJIQ-CAT will be of particular use for determining tissues to study for genes that are only inadequately represented in some but not all CATs. For the genes inadequately represented by all CATs, future work on alternative approaches to estimate splicing defects in patients will be necessary to improve clinical diagnoses.

CHAPTER 4

MAJIQ v3 addresses limitations of MAJIQ v2

4.1. Introduction

Chapter 3 used MAJIQ v2 developed in Chapter 2 to answer the question of how tissue-specificity affects what splicing variations we can expect to capture in the clinical setting. With actual analyses of RNA-seq data from patients with suspected Mendelian disorders (as discussed in Chapter 5), we found that MAJIQ v2 lacked several features that we needed to accurately and efficiently perform our analyses.

These missing features can be roughly divided into (1) better tools for comparing genomic features identified (e.g., splicegraphs) from different groups of input experiments, (2) valid intron retention coverage in rare cases of overlapping genes, and (3) additional measures for quantification of PSI. Addressing these limitations required the development of MAJIQ v3. In this chapter, I elaborate on these limitations, demonstrate how MAJIQ v3 addresses these limitations, and show how MAJIQ v3 provides additional performance improvements integrating features from MAJIQ v2 and MOCCASIN (Slaff et al., 2021).

4.2. Methods

4.2.1. MAJIQ incremental v3

In Chapter 2, we described incremental build (Figure 2.1c) as a new feature in MAJIQ v2. Incremental build parses aligned reads from BAM files to create "SJ files" with coverage over junctions and retained introns. These SJ files are used as input rather than BAM files to the MAJIQ build step. By doing so, the BAM files for experiments used in splicing analysis only need to be processed once, speeding up subsequent analyses. These SJ files are then organized into independent build groups which, along with GFF3 transcriptome annotations, are processed all at once in the MAJIQ v2 builder to simultaneously produce (1) a splicegraph, and (2) coverage over LSVs over this splicegraph suitable for quantification (Figure 4.1a). In MAJIQ v2, this is the only way of generating LSV coverage or splicegraphs.

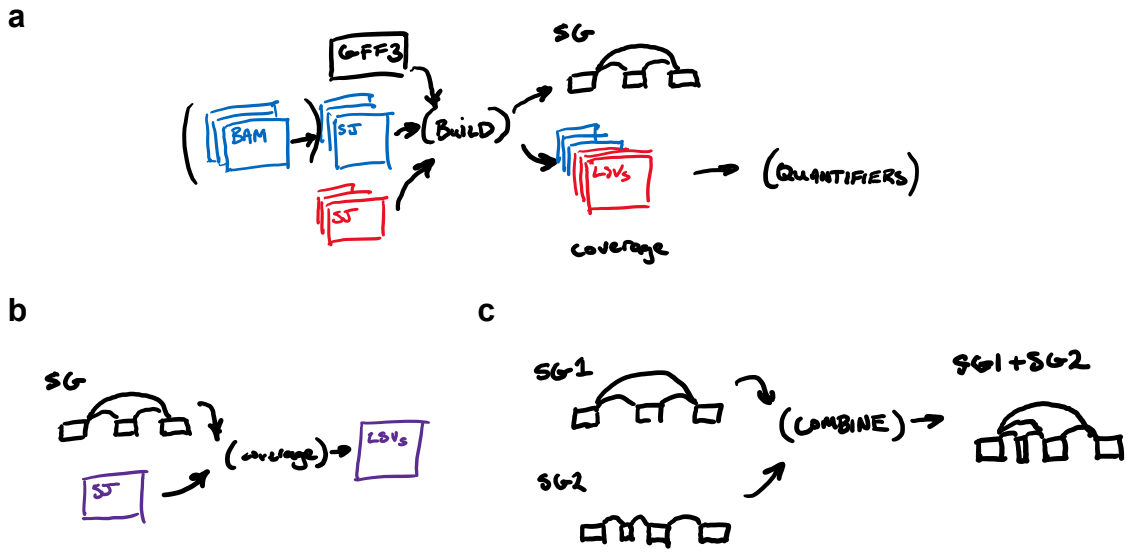


Figure 4.1: MAJIQ v3 extends MAJIQ v2 incremental build to splicegraphs and LSV coverage. (a) The MAJIQ builder takes annotated transcripts (GFF3) and SJ coverage from input experiments to produce (1) a splicegraph and (2) coverage over LSVs from that splicegraph. In MAJIQ v2, these steps are coupled together. (b) In MAJIQ v3, coverage over LSVs can be produced directly from an input splicegraph and SJ coverage. This improves parallelizability of splicing workflows and allows incremental quantification of new experiments for direct comparison with previous analyses. (c) In MAJIQ v3, new splicegraphs can be generated incrementally by combining 2 or more input splicegraphs. The combined splicegraph is structurally equivalent to a splicegraph built with the combined experiments used to create the input splicegraphs.

MAJIQ v3 extends incremental build to LSV coverage and splicegraphs. Conceptually, LSV coverage is a function of a splicegraph and coverage over junctions and retained introns (Figure 4.1b). In this sense, the MAJIQ builder first identifies the splicegraph and then uses it to define LSVs for coverage. MAJIQ v3 separates the MAJIQ builder into these two steps:

1. `majiq-build update`, which updates an input splicegraph (e.g., annotated splicegraph from GFF3) with SJ coverage from RNA-seq experiments,

2. `majiq-build psi-coverage`, which outputs coverage over LSVs defined by input splicegraphs from input SJ coverage.

By separating these steps, generating LSV coverage over many samples can be parallelized not only over multiple threads but over multiple machines on a cluster. Furthermore, LSV coverage can be generated incrementally: that is, for new RNA-seq experiments that were not used to define the input splicegraph. This incremental generation of LSV coverage is useful for the scenario where a researcher has thousands of samples which have previously been quantified with respect to a common set of LSVs. When the researcher obtains a few additional samples, MAJIQ v3 can directly generate coverage for (and then quantify) these new samples against the same set of LSVs. As a result, they can perform a quick comparison against the old dataset without reprocessing any of the preexistent samples. In contrast, for previous versions of MAJIQ, the approach would be to rebuild coverage over the combined set of samples and requantify all of the thousands of samples.

However, the incremental coverage and quantifications from the new approach would be over the original set of LSVs, rather than an updated set of LSVs that included evidence for novel junctions and retained introns from the new samples. The splicegraph combining the original and new input experiments can also be created incrementally without rerunning the MAJIQ builder on all input experiments. MAJIQ v3 introduces the `majiq-build combine` command, which takes two or more input splicegraphs to create a combined splicegraph (Figure 4.1). The combined splicegraph is defined by:

1. junctions: take the union of junctions found in each splicegraph, passing build/simplification filters if they were passed in any of the inputs,

2. exons: identify annotated exons, then use splice sites from updated junctions to identify novel exons and update existing exon boundaries,

3. retained introns: consider the intronic regions corresponding to annotated exon boundaries. Mark each region as passing build/simplification filters if any of the input splicegraphs has an intron overlapping the region that passed those filters. Then, define introns relative to the new updated exon boundaries, passing build/simplification filters if they were passed in the overlapping regions corresponding to annotated exon boundaries.

Combining input splicegraphs in this manner is equivalent to rerunning a build over the union of the build groups from the input splicegraphs. So, rather than running the builder again over all the input experiments, the researchers can instead build a splicegraph on just the new samples, followed by running `majiq-build combine` with the old and new splicegraphs.

After introducing operations to combine multiple splicegraphs, one could imagine other operations to manipulate and compare existing splicegraphs. For example, which introns and junctions in the combined splicegraph are found only in the new experiments? Similarly, we may expect that the combined splicegraph will have many of the same events as the old splicegraph. Since we already have quantifications over the old events, can we identify which events are shared vs not and perform quantifications only on the new/changed events? These questions motivated creating additional operations over splicegraphs akin to an algebra over splicegraphs, which we describe in Section 4.2.2.

4.2.2. Splicegraph algebra

MAJIQ v3 defines several operations on splicegraphs. One of these operations is combining splicegraphs as described in the previous subsection. Additional operations include identifying matched junctions, retained introns, and LSVs. These operations were motivated by two goals we had for the clinical pipeline. First, we wanted MAJIQ to identify novel junctions and

retained introns relative to control RNA-seq experiments (as well as annotated transcripts). In contrast, MAJIQ v2 always defines "*de novo*" status of introns and junctions relative to annotated transcripts. Second, when new samples (or cases) are added to an analysis (of controls), the existing samples have already been quantified with respect to the old set of LSVs. We also expect that most of these LSVs are unchanged in the new splicegraph. Rather than redundantly storing and requantifying these same LSVs, identifying matched LSVs between the new and old splicegraphs would permit producing coverage for (and subsequently quantifying) only the LSVs unique to the new splicegraph.

Identifying matched junctions is straightforward: does the other splicegraph have a junction for the same gene and coordinates. Identifying matched retained introns requires more care. Exonic, and thus intronic coordinates, are updated between splicegraphs to match novel junction splice sites. So, we cannot simply match intron coordinates. It is not enough to look for overlapping intron coordinates, either. Extended exon boundaries and novel exons can mask or split introns. So, we match introns from both splicegraphs to regions corresponding to boundaries of annotated exons (as done for combining splicegraphs). Two introns are matched if they share the same annotated intronic region. Finally, MAJIQ identifies matched events by searching for events belonging to the same gene which have junctions and introns with the same coordinates. In this case, we require intron coordinates to be identical because matched events are most useful for thinking about quantifications. Different intron coordinates could lead to different intron coverage and thus quantification.

Using these operations, MAJIQ v3 introduces the optional flag `--annotated` to its quantification commands to identify novel transcriptomic features relative to another splicegraph. to pass an additional splicegraph to its quantification commands. When used, MAJIQ accepts an additional splicegraph as input and uses the above operations to identify which junctions, retained introns, and events are novel relative to this splicegraph. Otherwise, MAJIQ v3 has the same behavior as MAJIQ v2 and marks junctions and retained introns with *de novo* status relative to the annotated transcripts. This could be used for the clinical setting by creating a

splicegraph including all RNA-seq samples except the patient of interest.

MAJIQ v3 also introduces the optional flag `--ignore-from` to its `psi-coverage` command to omit LSVs found in another splicegraph. When used, MAJIQ accepts an additional splicegraph as input and only outputs LSVs which are not found in this additional splicegraph. This can be used to quantify splicing with respect to a previous analysis in two passes (Figure 4.2).

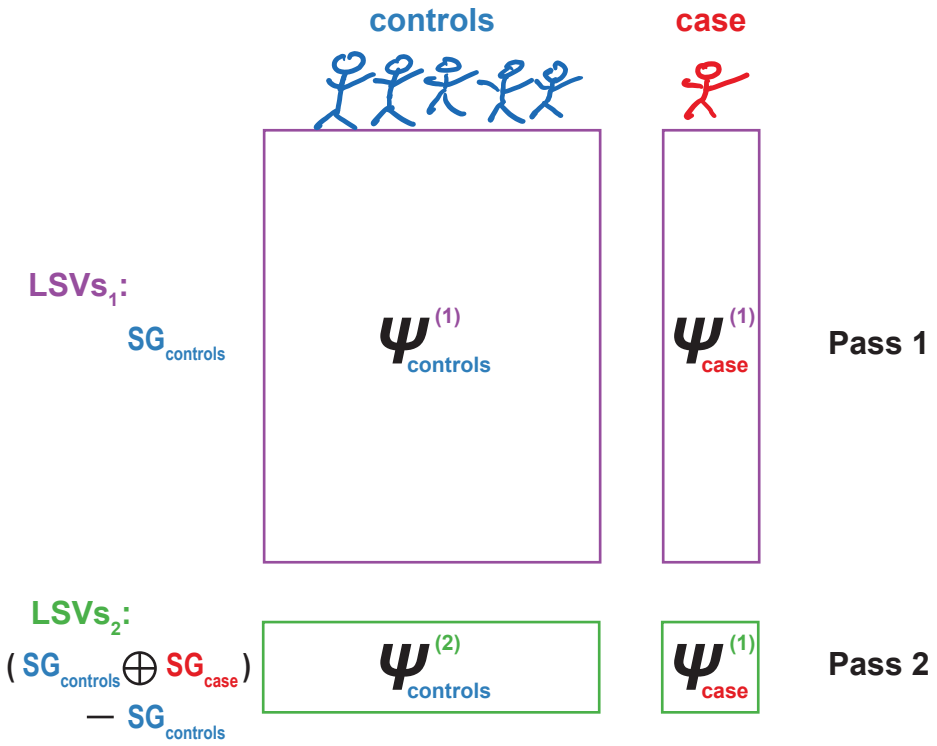


Figure 4.2: MAJIQ v3 splicegraph algebra enables analyses to be performed in two passes. Figure illustrates matrix of quantifications over experiments (columns) and LSVs (rows). Experiments are divided into controls (blue) and a case (red). LSVs are quantified from the combined splicegraph from the controls and case. That is, if $SG_{controls}$ is the splicegraph over controls and SG_{case} is the splicegraph over the case, the analysis is performed on their combination $SG_{controls} \bigoplus SG_{case}$. The LSVs for the combined splicegraph can be partitioned into LSVs that are the same in $SG_{controls}$ (purple) and those that are unique to their combination (green). Analysis on these LSVs can be described as a first and second pass. Analyses sharing the same build groups can be decomposed in this manner to compute quantifications on the first pass LSVs only once.

4.2.3. New model for retained intron coverage

MAJIQ v3 measures coverage over retained introns differently than MAJIQ v2. This difference addresses artifacts in intron coverage around overlapping genes found in MAJIQ v2. In MAJIQ v2, regions for intronic coverage are defined per gene. Each gene's regions are defined solely on the basis of the gene's exons. As a result, regions from different genes on the same contig can overlap, which leads to redundant time and space measuring intronic coverage for overlapping regions. More critically, these regions can overlap each others' exons. So, exonic coverage from one gene can be artifactually counted to another gene's introns. Figure 4.3a illustrates how this happens in MAJIQ v2.

In contrast, MAJIQ v3 measures coverage over non-overlapping regions (per strand, if the RNA-seq experiment is stranded) over each contig. The intronic regions from each gene are combined, and the exonic regions for each region are subtracted, as illustrated in Figure 4.3b. This excludes any exonic region from contributing to intronic coverage in a different gene. Additional care is taken to identify intronic regions that belong to "annotated introns". Annotated introns are part of some annotated transcript's exon that was split by junctions from other transcripts. We identify regions that correspond to annotated introns so that they are not counted to novel introns.

MAJIQ v3 measures coverage over these regions using the same procedure that MAJIQ v2 uses for gene introns (as described in Section 2.4.1). This procedure bins together positions for which aligned reads can overlap each region and counts overlapping unsplit reads per bin. Then, MAJIQ maps this coverage back to introns in the following way. For each intron, MAJIQ identifies the collection of these regions which overlap (in coordinates, strand, and annotation status). MAJIQ takes the weighted average of number of reads, number of read bins with at least one read, and bootstrapped coverage from each of the overlapping regions. The values from each region are weighted by the number of read positions corresponding to the region (i.e., region length plus the maximum read length of the RNA-seq experiment, adjusted for minimum overhang).

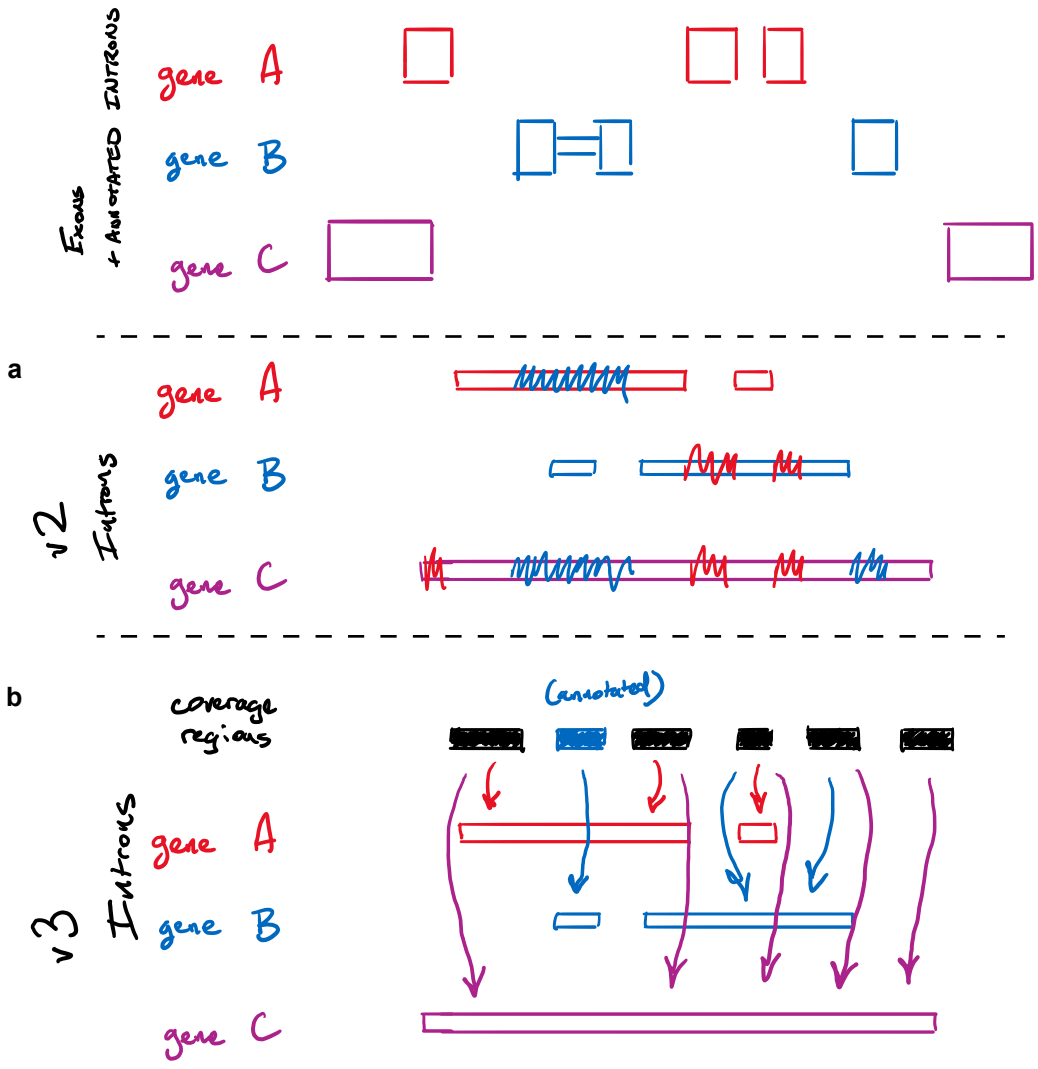


Figure 4.3: MAJIQ v3 measures intron coverage over non-overlapping regions which exclude all genes' exons. (a) Diagram illustraing how MAJIQ v2 measures coverage in overlapping regions which can overlap with each others' exons. Squiggly lines show where exonic coverage from other genes can lead to artifactual coverage. (b) MAJIQ v3 instead measures coverage over non-overlapping regions which exclude all exons. Coverage regions are defined by excluding all exonic regions and noting which regions originated from annotated introns. Coverage over introns is determined by averaging overlapping regions.

4.2.4. Corrected posterior standard deviation / variance

In order to identify single experiments as outliers for the clinical pipeline as described in Chapter 5, we want to directly use estimates of uncertainty of the measured posterior distributions. To enable this, MAJIQ v3 corrects how estimates of posterior distribution variance are calculated.

MAJIQ v2 attempts to calculate PSI posterior variance by the following approach. Recall that MAJIQ estimates M bootstrap replicates of beta posterior distributions for Ψ, parameterized by $\{(\alpha_m, \beta_m)\}_{m=1}^{M}$. MAJIQ v2 calculates variance by discretizing Ψ into 40 equally-sized bins over $[0, 1]$, and setting the probability of each bin to the average probability of the bin over bootstrap replicates. Then, MAJIQ v2 calculates variance with respect to the midpoints of the bins, weighted by these probabilities.

There are many challenges with this procedure for estimating the posterior variance, for which I will not elaborate further. However, if we imagine increasing the number of equally-sized bins, we see that this is an approximation of a uniform mixture of beta distributions. The variance of this distribution can be directly and exactly calculated much more efficiently.

This mixture distribution can be factored into its mixture components using the hidden random variable $Z \sim \text{Uniform}(\{1, \dots, M\})$, indicating which of the M bootstrap replicate posterior distributions Ψ is sampled from. Then, the conditional distribution of Ψ given Z is $\Psi|Z \sim \text{Beta}(\alpha_Z, \beta_Z)$, where (α_m, β_m) are the Beta distribution parameters for the m-th bootstrap replicate's posterior distribution. The law of total variance states that we can decompose the variance of Ψ into the variance of a conditional expectation and the expectation of a conditional variance. That is:

$$\text{Var}[\Psi] = \text{Var}[\mathbb{E}[\Psi|Z]] + \mathbb{E}[\text{Var}[\Psi|Z]].$$

These conditional expectations and variances of a Beta-distributed random variable have closed

form:

$$\mathbb{E}[\Psi|Z=m] = \frac{\alpha_m}{\alpha_m+\beta_m},$$
$$\text{Var}[\Psi|Z=m] = \frac{\alpha_m\beta_m}{(\alpha_m+\beta_m)^2(1+\alpha_m+\beta_m)}.$$

So, the variance of Ψ is:

$$\text{Var}[\Psi] = \text{Var}\left[\frac{\alpha_Z}{\alpha_Z+\beta_Z}\right] + \mathbb{E}\left[\frac{\alpha_Z\beta_Z}{(\alpha_Z+\beta_Z)^2(1+\alpha_Z+\beta_Z)}\right],$$

which can be computed by enumerating the means/variances of each component beta distribution and numerically taking their variance/mean. This is faster and correctly calculates the variance of the mixture of bootstrapped posterior distributions on PSI.

4.2.5. Approximate bootstrap coverage

MAJIQ v3 uses a smooth approximation to the uniform mixture of bootstrapped posteriors for posterior quantiles and samples. Modeling PSI directly as a finite mixture distribution over PSI causes challenges when total coverage at an LSV increases. As total coverage increases, the uncertainty of each mixture component becomes vanishingly small. Since we only sample a finite number of mixture components, the support of the actual mixture distribution becomes finite. This leads to negative consequences when sampling or taking quantiles from the distribution. Figure 4.4a,b shows mixture distribution with the same underlying mixture distribution means but corresponding to small and high number of reads, showing how jagged the distribution can become with increasing number of reads.

To resolve these difficulties, MAJIQ uses a smooth approximation of the mixture distribution when sampling (HET) and taking quantiles (CLIN). MAJIQ approximates the mixture distribution with a single beta distribution, with parameters set to match the mixture distribution's mean and variance. Figure 4.4c,d shows the smooth approximation's probability density in contrast to the original mixture.

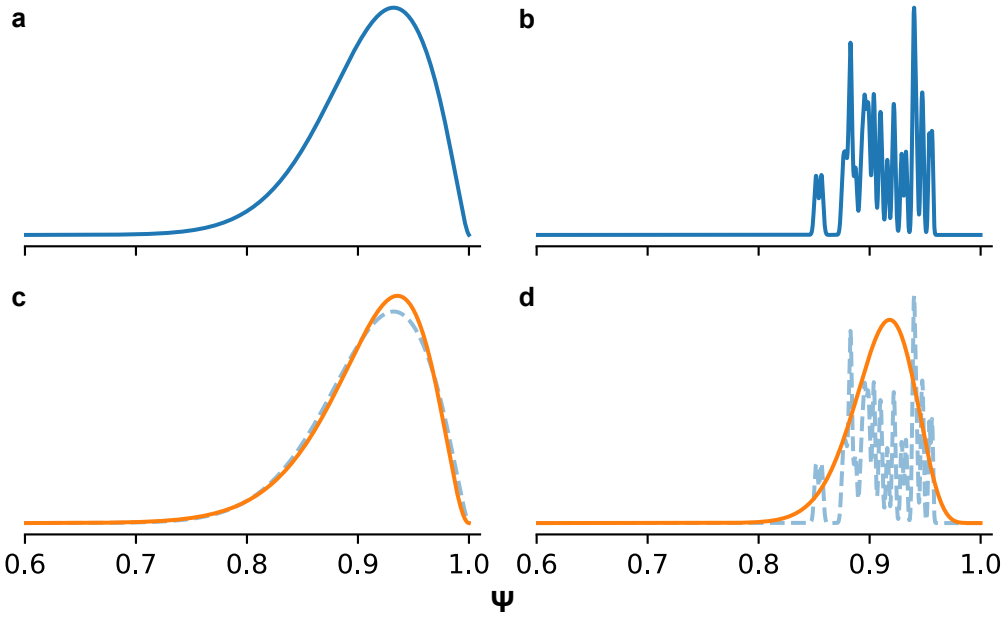


Figure 4.4: MAJIQ v3 uses a smooth approximation to the uniform mixture of bootstrapped posteriors for posterior quantiles and samples. (a) and (b) show the probability density function of a mixture (in blue) of 30 beta distributions, each with the same mean, but corresponding to LSV coverage of 50 vs 50,000 reads. (c) and (d) show MAJIQ v3's smooth approximation of this distribution (in orange) as a beta distribution with matching mean and variance, which is used for posterior samples (HET) and quantiles (CLIN).

4.2.6. MOCCASIN batch correction

MAJIQ v3 includes a rewritten implementation of MOCCASIN for batch correction of coverage over LSVs. The new implementation is vectorized over junctions and bootstrap replicates, resulting in significant performance improvements. It also takes advantage of changes to underlying data structures (Section 4.2.7) to remove no longer necessary steps (e.g., matching indexes for events, which are now consistently ordered) and to enable multithreaded and/or distributed parallelism using Dask.

Similar to changes to MAJIQ build to enable two-pass analysis (Section 4.2.1), MAJIQ v3's implementation of MOCCASIN separates the different modeling and inference steps for both unobserved confounders and coverage. Previously, MOCCASIN assumed that you had coverage for all the LSVs you would ever want to work with (i.e., the coverage produced from v2's builder). So, MOCCASIN would be input with known factors and coverage and output updated coverage. MAJIQ v3's implementation of MOCCASIN is split into four steps:

1. `majiq-moccasin factors-model`: use coverage and matched known factors to a build a model of unobserved confounding factors.

2. `majiq-moccasin factors-infer`: use coverage, matched known factors, and model of unobserved confounding factors (from `factors-model`) to augment set of input factors with inferred unobserved confounding factors. The model requires input coverage to be over the same set of LSVs but works for new, previously unseen experiments.

3. `majiq-moccasin coverage-model`: use coverage and input factors to build a model of coverage vs input factors.

4. `majiq-moccasin coverage-infer`: use coverage, input factors, and model of coverage (from `coverage-model`) to generate updated coverage over LSVs removing variability attributed to confounding factors. The model requires coverage to be over the same set of LSVs but works for new, previously unseen experiments.

4.2.7. Updated underlying data structures

Splicegraph, coverage and VOILA files were rewritten to improve performance, decrease file sizes and enable parallelization with Dask. Redundant information (i.e. gene names, contig names) is stored either by reference or computed as needed to decrease storage and memory usage.

For SJ coverage, per-bin read counts are now kept in storage as a sparse matrix (most junctions only have a few nonzero bins/positions). Bootstrapping is no longer stored and instead computed on the fly as needed. This means that SJ coverage is now deterministic/reproducible for the same input BAM file on different runs of MAJIQ. Splicegraphs exclude read counts, which are split into separate splicegraph coverage files (produced by the `sg-coverage` command). This keeps splicegraphs relatively small, and saves time spent to only assessing coverage when requested rather than automatically for all input samples. Coverage over LSVs is stored redundantly, storing the percentage of reads in the LSV to which a junction or intron was assigned in one array and the total number of reads in the LSV repeated for each junction and intron. This is to enable chunking the data over junctions/retained introns when the number of samples is large such that it can no longer be loaded all at once in memory. LSVs are always stored in the same order.

4.2.8. Comparisons to MAJIQ v2

I randomly selected 20 samples from GTEx for performance comparisons between MAJIQ v2 and MAJIQ v3. I considered three basic tasks: (1) creating SJ files from input BAM files, (2) creating splicegraphs and LSV coverage for analysis from input SJ files, and (3) quantifying PSI from output LSV coverage files. I measured runtime and memory usage with increasing numbers of threads and samples. Performance comparisons were performed on a desktop running Ubuntu 20.04.2 LTS with Intel Xeon Gold 6238R CPU (2.20GHz) and 512GB RAM.

I also retrospectively analyzed runtimes for using SJ files to build a splicegraph, infer coverage over LSVs, and quantifying PSI over the 762 samples from GTEx selected for the analysis of clinically-accessible tissues described in Chapter 3. Build was performed with each

tissue as its own build group, using the simplifier and setting min-experiments to 20%. We compared runtimes building a splicegraph, inferring coverage, and quantifying PSI over each individual sample.

We compared PSI quantifications from MAJIQ v3 vs MAJIQ v2 for a single skeletal muscle sample from the above-described GTEx analysis (SRA accession SRR109847). We ignored LSVs with half exon reference exons. To account for differences in LSV definitions, we compared quantifications for individual junctions and retained introns as part of source vs target events. We matched junctions by gene id and genomic coordinates. We also matched introns by gene id and genomic coordinates, but we permitted matches to only one of the two coordinates to small differences in exon definition. We compared inferred values of PSI, stratifying on whether quantifications were for introns vs junctions and the total number of reads assigned to their LSV in MAJIQ v3. We then explored specific quantifications where the tools gave different results to understand if the changes were improvements.

4.3. Results

MAJIQ v3 introduces new features that further enable incremental analysis with large numbers of samples. By decoupling how the MAJIQ builder produces splicegraphs and coverage over LSVs, MAJIQ v3 enables adding a few samples at a time to a large analysis, analyzing their combined set of splicing events while only needing to requantify events which have changed as two passes for analysis.

We first compared speed and memory usage in generating SJ coverage for 20 RNA-seq samples from GTEx using previously aligned BAM files (Figure 4.5). MAJIQ v3 was always faster than MAJIQ v2 in creating SJ files with the same number of threads. With a single thread, MAJIQ v3 created each SJ file an average of 17 ± 7 seconds faster than MAJIQ v2. This difference increases with the number of threads. MAJIQ v2 runtime plateaus with around 4 threads, achieving a 2.0 ± 0.1 times speed up vs a single thread. In contrast, MAJIQ v3 with 4 threads is 4.5 ± 0.2 times faster than single-threaded v2 (4.0 ± 0.1 times vs single-threaded v3) and does not plateau until around 9 threads (7.8 ± 0.8 times vs v2, 6.9 ± 0.7 times vs v3).

Memory usage remains low in both MAJIQ v2 and v3 with increasing number of threads (less than 1GB).

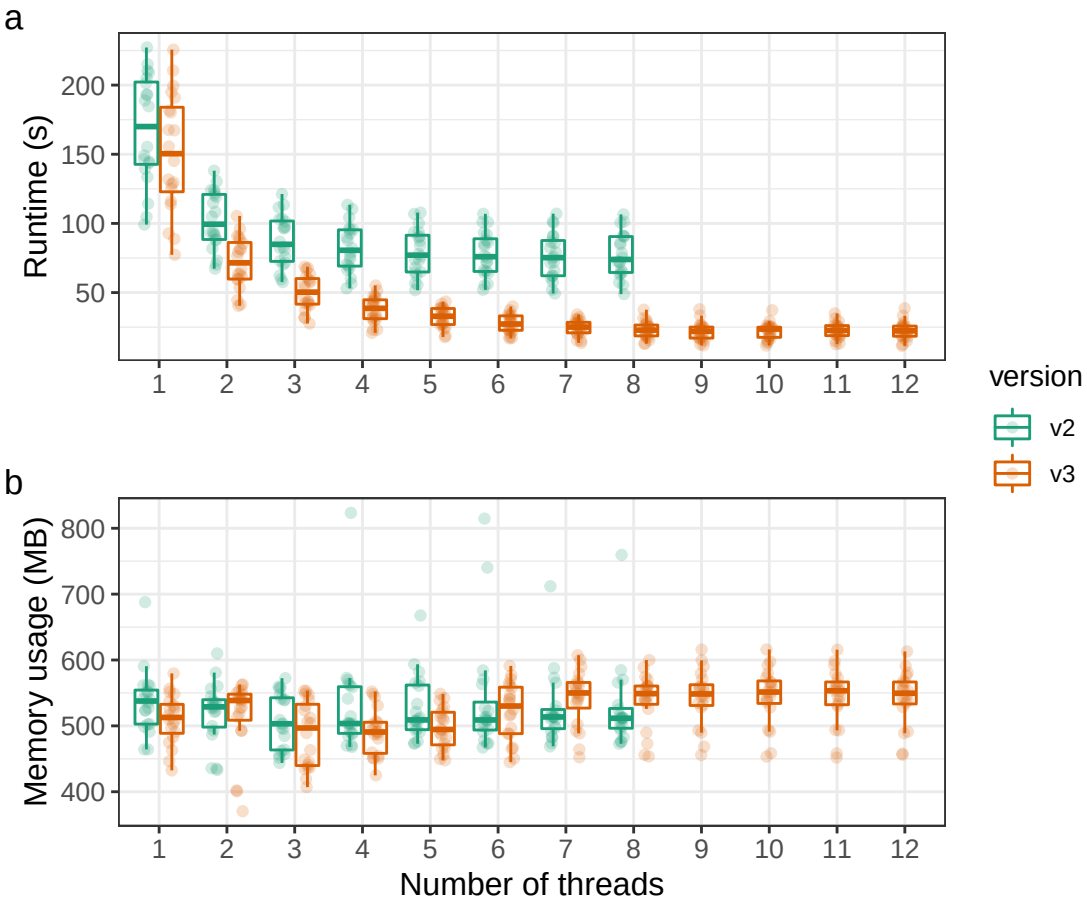


Figure 4.5: MAJIQ v2 vs MAJIQ v3 runtime and memory usage when creating SJ files. (a) Runtime and (b) maximum memory usage by MAJIQ v2 and MAJIQ v3 when creating SJ files with increasing number of threads (horizontal axis). Measurements taken from 20 randomly selected experiments from GTEx.

We then compared speed and memory usage in building splicegraphs and obtaining coverage over LSVs. MAJIQ v2 requires these tasks to be run all together with a single thread. In contrast, MAJIQ v3 supports doing these tasks in multiple steps with multiple threads by first building a splicegraph and subsequently obtaining coverage over LSVs (which can be parallelized over samples on multiple machines). Figure 4.6a-b compares runtime and memory usage for running these steps sequentially with a single thread with increasing number of samples. These

runtimes show a linear trend with increasing sample size. For MAJIQ v2, each experiment contributes an average of 9.8 ± 0.1 seconds to the total runtime. For MAJIQ v3, each experiment contributes an average of 1.78 ± 0.01 seconds. This is on top of baseline runtimes of 34 ± 1 seconds for MAJIQ v2 vs 8.0 ± 0.4 seconds. MAJIQ v3 performs these steps for 20 experiments faster than MAJIQ v2 for a single experiment. If we extrapolated to the sample size of GTEx (17,382 experiments), this would be a difference between 47.4 hours (v2) and 8.6 hours (v3). These performance comparisons assume running these steps sequentially with a single thread. MAJIQ v3 supports running these steps with multiple threads and in parallel. Since obtaining coverage over LSVs is embarassingly parallel, we expect that MAJIQ v3 can complete these tasks even faster (with respect to walltime) if given more resources.

We also compared speed and memory usage in quantifying PSI for individual experiments to TSV files (Figure 4.7). MAJIQ v2 permits using multiple threads/processes for this step. With a single thread, MAJIQ v2 takes an average of 230 ± 40 seconds to quantify each sample. With 8 threads, MAJIQ v2 takes an average of 110 ± 20 seconds. In contrast, MAJIQ v3 with a single thread takes an average of 8.5 ± 0.5 seconds to quantify each sample. Both MAJIQ v2 and MAJIQ v3 require less than 1GB of memory. MAJIQ v3 requires an average of 700 ± 50MB of memory compared to 390 ± 60MB (1 thread) / 845 ± 1MB (8 threads) with MAJIQ v2.

We also retrospectively evaluated runtimes for the MAJIQ builder and MAJIQ PSI given input SJ files for the 762 samples from GTEx described in Chapter 3. For MAJIQ v2, generating the splicegraph over 53 tissues/build groups and generating coverage over LSVs for each sample took place in a single step. MAJIQ v2 does not support running this step in parallel over multiple threads. In contrast, MAJIQ v3 generated the combined splicegraph over 53 build groups by first launching independent jobs to build tissue-specific splicegraphs. Then, MAJIQ (v3) combine was used to generate a combined splicegraph. Afterwards, MAJIQ v3 launched 53 jobs in parallel for each tissue to generate coverage over LSVs, each of which used 2 threads. MAJIQ v2 successfully completed this step in 6 hours and 24 minutes. In contrast, for MAJIQ v3, the cluster walltime (from when first job was launched to when last job was completed) was 2 minutes and 57 seconds.

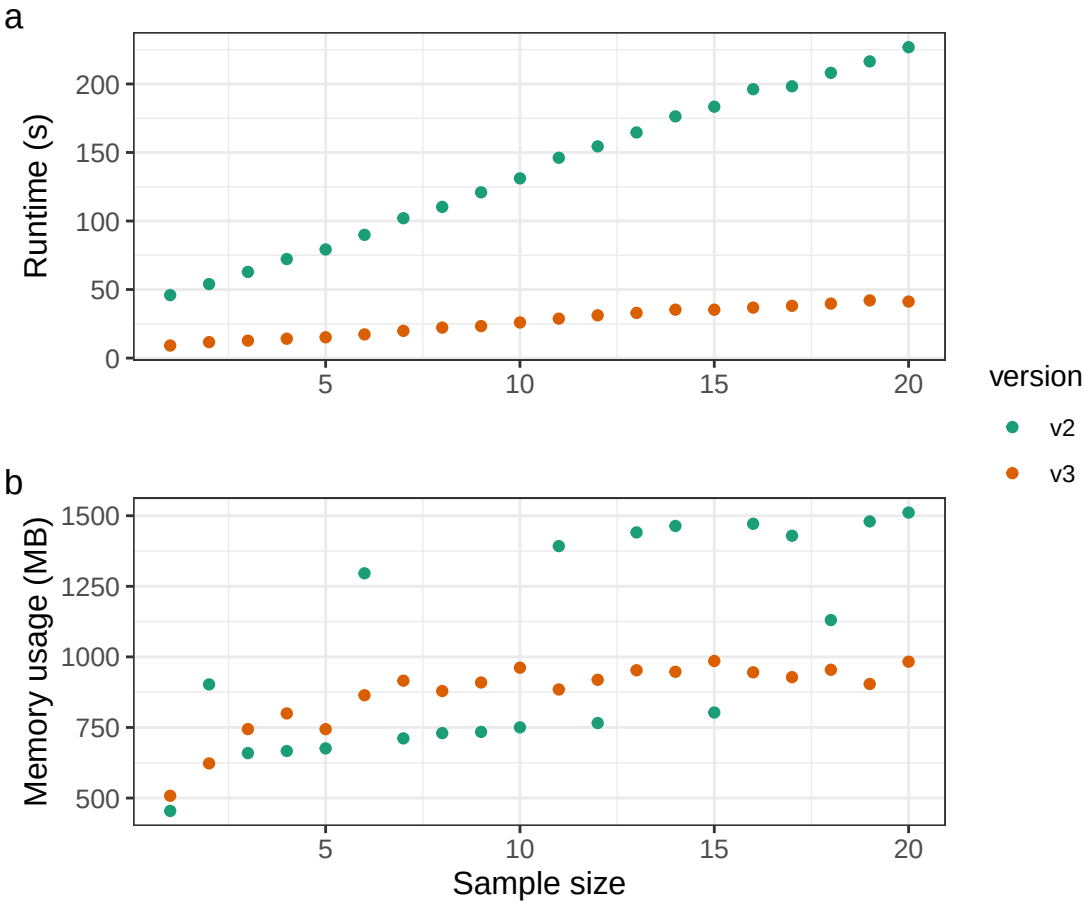


Figure 4.6: MAJIQ v2 vs MAJIQ v3 runtime and memory usage when creating building splice-graphs and LSV coverage. (a) Runtime and (b) maximum memory usage by MAJIQ v2 and MAJIQ v3 when creating SJ files with increasing sample size (horizontal axis). Measurements taken using up to the first 20 randomly selected experiments from GTEx with MAJIQ build default parameters.

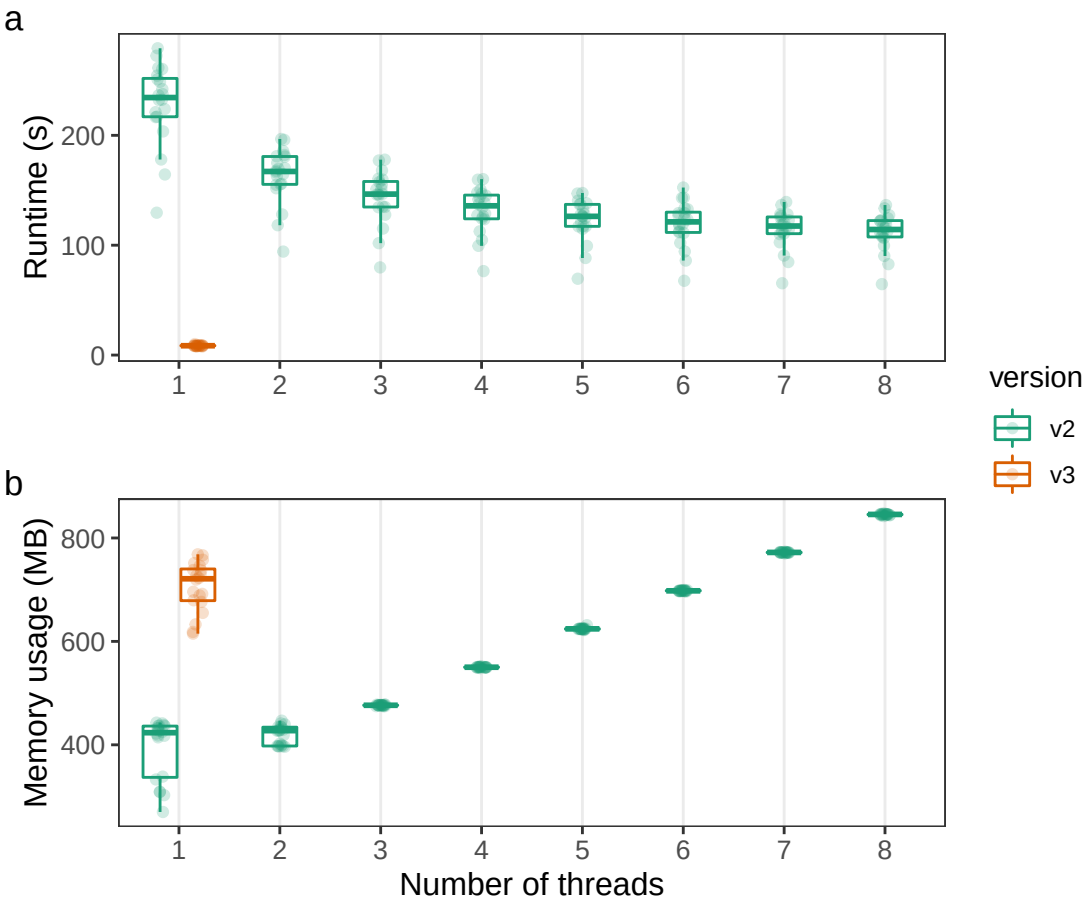


Figure 4.7: MAJIQ v2 vs MAJIQ v3 runtime and memory usage when quantifying PSI. (a) Runtime and (b) maximum memory usage by MAJIQ v2 and MAJIQ v3 when quantifying PSI to TSV files with increasing number of threads (horizontal axis). Measurements taken from 20 randomly selected experiments from GTEx.

The single-node walltime (adding walltimes for each job together) was 37 minutes and 56 seconds. That is, over a single node, we observe a 10 times improvement in speed. Parallelizing over a cluster, we observe a 131 times improvement in speed. Quantification of each independent sample was performed on parallel jobs on the chop cluster for both MAJIQ v2 and MAJIQ v3. For v2, the wall times were 37.5 minutes (cluster) and 2.5 days (single node). For v3, the wall times were 1.5 minutes (cluster) and 14.3 minutes (single node). This amounts to a 25 times improvement on our cluster and a 251 times improvement on a single node for quantifying PSI.

We compared the PSI quantifications from v2 and v3 from one of these GTEx samples (skeletal muscle, accession SRR1098474). MAJIQ v2 quantified 77,271 LSVs, while MAJIQ v3 quantified 60,877 LSVs. From these LSVs, we identified 121,791 matched junction quantifications and 27,042 matched intron quantifications (23,765 matching both coordinates, 3,277 matching only one coordinate). Across these matched quantifications, only 3,893 (2.6%) differered in their quantifications of PSI by more than 5%. Figure 4.8 plots these matched values of PSI, stratifying with respect to total LSV coverage (from v3) and introns vs junctions. Most differences of PSI greater than 5% come from LSVs with low coverage. For quantifications with even more significant differences, we observed that introns tended to have higher values of PSI in MAJIQ v2 and junctions tended to have lower values of PSI. Figure 4.9 provides an example where this difference results from changes to intron retention coverage. Investigating the aligned reads suggests that the higher value of PSI in the gene *ERAP*'s intron comes from the overlapping exons of the gene *CAST*.

4.4. Discussion

We developed MAJIQ v3 to address methodological needs for the clinical pipeline. One of these methodological needs was accurately quantifying uncertainty in the PSI posterior distributions of an individual sample. To address this need, we updated how MAJIQ calculates the variance of PSI posterior distributions to be more accurate and efficient. We further wanted to extract specific quantiles from the the PSI posterior distributions, so we introduced new methodology to make smooth approximations of the posterior distributions on which quantiles

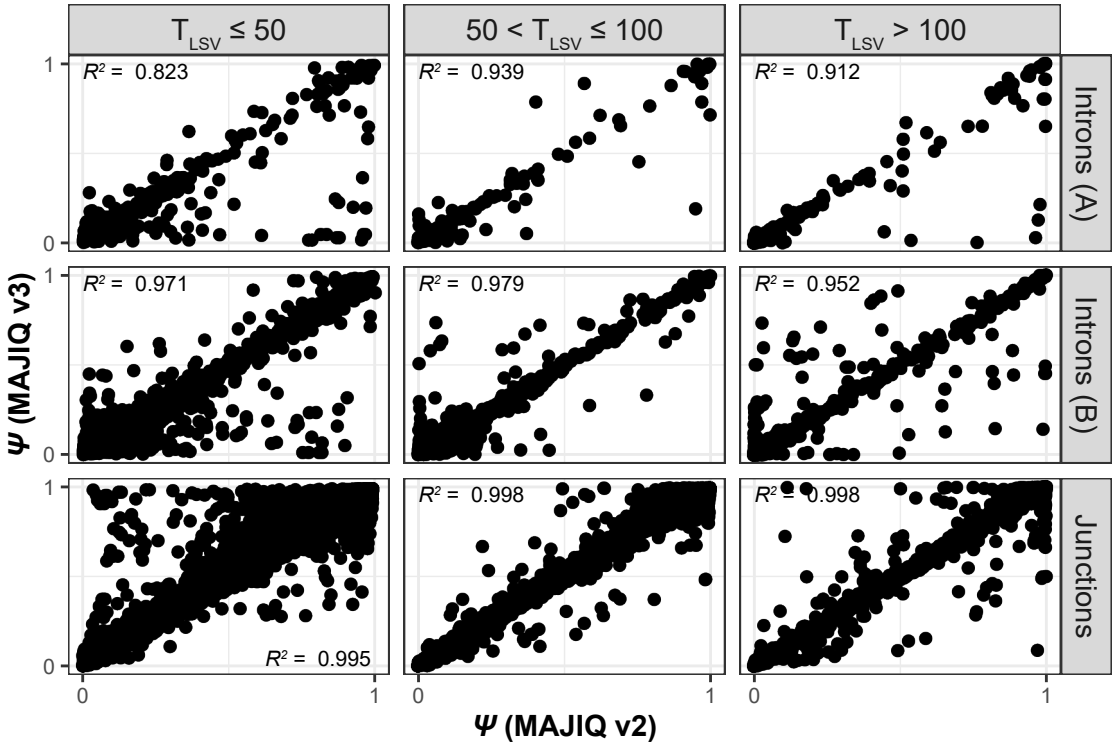


Figure 4.8: MAJIQ v3 vs MAJIQ v2 PSI comparison for introns/junctions and low to high coverage. Points show matched observations of PSI from MAJIQ v2 and MAJIQ v3. The horizontal/vertical axes represent PSI from MAJIQ v2/v3. Points are faceted into columns by total LSV coverage (T_{LSV}) as measured by MAJIQ v3 and rows by whether the observation was from a retained intron or junction. Retained introns are divided into matches with only one coordinate (A) vs with both coordinates (B).

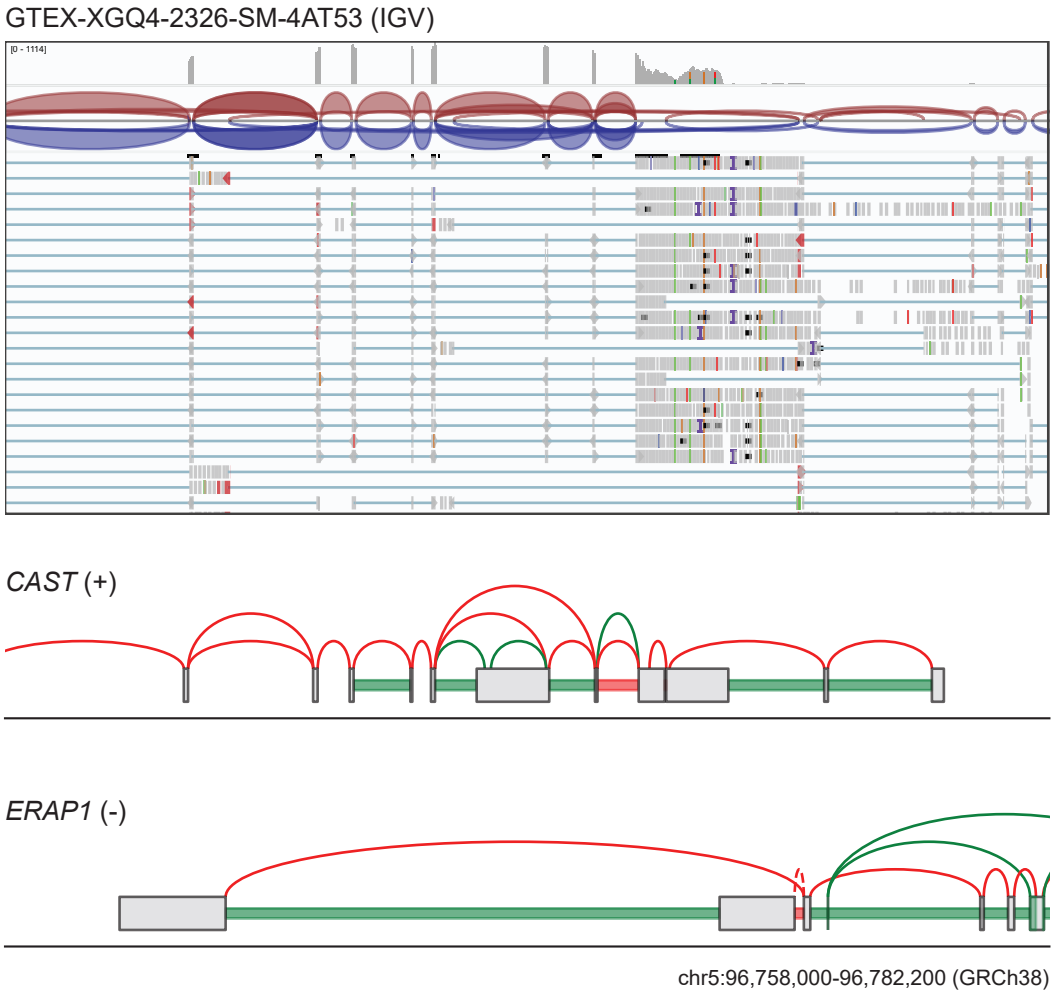


Figure 4.9: MAJIQ v2 artifactually inflates intronic coverage in the gene *ERAP* due to overlapping exons from the gene *CAST*. First panel is IGV view of RNA-seq reads for the GTEx sample GTEX-XGQ4-2326-SM-4AT53 (SRR1098474). Second and third panels show MAJIQ splicegraphs for the overlapping genes *CAST* and *ERAP*. For the left-most intron in *ERAP1*, IGV panel shows coverage aligning with the exons of *CAST*. MAJIQ v2 assigns 49 reads to this intron ($\Psi = 80\%$), while MAJIQ v3 assigns 3 reads to this intron ($\Psi = 23\%$).

could be computed in a numerically stable way. Similarly, although introns in overlapping genes are rare, we did not want our clinical pipeline to enrich for artifactually inflated intron coverage. We addressed this need in v3 by changing how intron coverage is measured and assigned to genes.

MAJIQ v3 is substantially more efficient than MAJIQ v2. We evaluated differences in (1) extracting raw coverage from BAM files, (2) building splicegraphs and extracting coverage over LSVs for analysis, and (3) quantifying PSI from individual experiments. For extracting raw coverage, MAJIQ v3 was always faster than MAJIQ v2 with the same number of threads. MAJIQ v3 also scaled better with increasing numbers of threads. For building splicegraphs and extracting coverage over LSVs, MAJIQ v3 was over 5 times faster than MAJIQ v2 with a single thread. In practice, this step can be performed even faster because MAJIQ v2 only supports this step on a single thread while MAJIQ v3 supports processing LSV coverage in an embarassingly parallel fashion over multiple threads and machines. For quantifying PSI, MAJIQ v3 was 27 times faster than MAJIQ v2. When analyzing the 762 samples described in Chapter 3 on the CHOP cluster, MAJIQ v3 cluster walltimes were 94 times faster (single-node: 76 times faster) than MAJIQ v2. Both MAJIQ v2 and v3 were efficient with memory, typically requiring less than 1GB of memory for each step.

These measured performance improvements do not illustrate how we could reuse analysis results by analyzing data in two passes. While we describe here the methodology of MAJIQ v3's new operations on multiple splicegraphs, their impact in sharing common analyses over controls and identifying novel junctions, retained introns, and events is best seen in the following chapter in which we describe MAJIQ-CLIN.

CHAPTER 5

MAJIQ CLIN identifies splicing aberrations from patient RNA-seq

5.1. Introduction

In this chapter, we use the methodology developed in Chapters 2 and 4 to address the task described in Chapter 1; that is, to use RNA-seq on patient RNA-seq to improve molecular diagnosis, focusing specifically on splice disrupting changes.

Previous studies have approached this problem for identifying splice disrupting changes at the RNA-seq level by comparing patient samples to controls samples from similar tissue. There are two general approaches that have been taken: (1) identifying novel junctions (not found in controls) above some read threshold and (2) producing quantifications of inclusion of all junctions in cases vs controls and identifying instances where cases are outliers relative to controls. With respect to the first approach, Cummings et al. (2017) pioneered the first approach on muscle samples, and Gonorazky et al. (2019) extended it by relaxing the requirement of novel junctions by allowing them to be present in a few samples. With the second approach, different studies have quantified splicing with respect to intron clusters (as from Leafcutter; Li et al. (2018)) or junctions sharing a single 5' or 3' splice site. Kremer et al. (2017) and Jenkinson et al. (2020) (LeafcutterMD) both use Leafcutter's intron clusters for quantification, but identify outliers differently using Dirichlet-multinomial likelihood ratio testing (Kremer et al., 2017) vs beta-binomial tail probabilities (Jenkinson et al., 2020). Meanwhile, Frésard et al. (2019) and Mertes et al. (2021) (FRASER) quantify inclusion with respect to shared 5' or 3' splice sites, but create different statistical frameworks for identifying outliers using Z-scores (Frésard et al., 2019) and beta-binomial tail probabilities (Mertes et al., 2021). These approaches identify a need to correct for covariation between quantifications within each sample and either regress out principal components on the quantifications directly (Frésard et al., 2019) or indirectly on the beta binomial parameters using a linear autoencoder (Mertes et al., 2021). Mertes et al. (2021) also evaluates retained introns by assessing unsplit read coverage at the same splice sites.

Some of these approaches have been made more accessible as open source software packages. The approach described in Jenkinson et al. (2020) was released as an update to Leafcutter, with the new scripts called LeafcutterMD. The approach described in Mertes et al. (2021) was released as the R package FRASER and as part of the DROP pipeline (Yépez et al., 2021).

These packages are indifferent to whether they are structurally novel (as done by Cummings et al. (2017)) and rely solely on quantified differences in splicing. Furthermore, LeafcutterMD provides no way to correct for within-sample covariation. FRASER, while able to perform correction due to unobserved confounders, does not permit specifying observed confounding factors when known.

Here, we introduce MAJIQ-CLIN, a new pipeline to assess splicing in one or more individuals vs controls using the newly updated MAJIQ v3 toolkit for splicing detection, quantification, and visualization. MAJIQ-CLIN uses MAJIQ v3's algebra of splicegraphs to identify retained introns, junctions and events that are structurally novel compared to controls, which can be used to further filter or prioritize identified outliers. MAJIQ-CLIN also integrates MOCCASIN to correct for between-sample covariation in PSI due not only to unobserved confounders but also known confounders between samples. MAJIQ-CLIN implements an efficient two-pass approach using MAJIQ v3's algebra of splicegraphs in order to efficiently process each patient sample against a shared set of external or leave-one-out controls, allowing for the vast majority of splicing variations within controls to be quantified only once, while keeping the complexity of splicing variations analyzed focused on individual patients.

In the remainder of this chapter, we describe how MAJIQ-CLIN works, demonstrate how it compares favorably to existing methods, and apply it to patient data, re-identifying previously-made diagnoses.

5.2. Materials and Methods

MAJIQ-CLIN is a pipeline to identify splicing outliers in patient samples vs some set of controls. First we define splicegraphs over patient cases and controls which allow us to (1) define a set of shared LSVs over all patients that need to only be quantified once, (2) define patient-specific LSVs, and (3) identify novel introns, junctions, and events for each patient. Second, we use MOCCASIN to correct coverage over shared and patient-unique LSVs for observed and unobserved confounders. Afterwards, we summarize population quantiles of PSI over these LSVs in controls, and we compare how posterior quantiles of PSI in patient cases compare in order to identify outlier events. We prioritize these events with respect to structural novelty and the distance between extreme quantiles of the control and case distributions. In the remainder of this section, I elaborate on how these steps are performed and describe how we compare the results of this method to other tools.

5.2.1. Building splicegraphs for analysis

MAJIQ defines the set of junctions, retained introns, and LSVs which can be quantified in terms of a splicegraph. Recall that MAJIQ builds splicegraphs by combining information from transcript annotations and evidence from multiple build groups. For MAJIQ-CLIN, the final set of patient LSVs for quantification is defined by combining a patient-specific build group and build groups over controls. The build groups over controls pass/include retained introns and junctions if they are found in enough experiments in the build group, while patient-specific build groups pass/include retained introns and junctions if they are found in any experiment in the build group.

For example, patient data from UDN (Murdock et al., 2021) includes cases from whole blood and fibroblasts. We can use data from UDN as controls; however, we can further augment the splicegraph with evidence from GTEx. For patients i and j with RNA-seq evidence from one or both tissues, we would define patient-specific splicegraphs with the following build groups:

- Patient splicegraph i (with blood and fibroblast RNA-seq data)

  - patient-specific: patient *i*, blood; patient *i*, fibroblast
  - controls: UDN whole blood
  - controls: UDN fibroblasts
  - controls: GTEx whole blood
  - controls: GTEx fibroblasts
- Patient splicegraph *j* (with only fibroblast RNA-seq data)
  - patient-specific: patient *j*, fibroblast
  - controls: UDN whole blood
  - controls: UDN fibroblasts
  - controls: GTEx whole blood
  - controls: GTEx fibroblasts

Each patient splicegraph has a unique set of LSVs because of their patient-specific build group. However, we can build a single shared splicegraph excluding the patient-specific build groups, which we call the "first-pass splicegraph." When there are many more controls than cases, we expect that this splicegraph shares nearly all the same LSVs as each of the patient splicegraphs, such that there are only a few patient-specific LSVs. Then, analysis can be performed in two passes: (1) once on LSVs from the first-pass splicegraph, and subsequently (2) for each patient, analyze LSVs from the patient-specific splicegraph which were not found in the first-pass splicegraph (and re-use results for the events shared with the first-pass splicegraph).

Frequently, the control build groups will include the patient itself (in the previous example, patient *i* is found in both UDN whole blood and UDN fibroblasts, etc.). When this is the case, we build a second patient-specific splicegraph explicitly excluding the patient from its component build groups. This splicegraph is used to identify retained introns, junctions, and LSVs that are

structurally novel for the patient.

5.2.2. PSI Coverage in two passes

First, we produce uncorrected PSI coverage with respect to the first-pass splicegraph for each patient and the controls to which they will be compared (which may be different than the controls used for building the splicegraph). Then, using this coverage and known confounders, we use MOCCASIN to model unobserved confounders. We plot the percentage of unexplained variance each learned unobserved confounder contributes to determine the number of unknown confounders used in subsequent analysis. Then, using known and selected unobserved confounders, we model and correct the PSI coverage using MOCCASIN. We also use these unobserved factors from the first-pass PSI coverage to model and correct PSI coverage for the LSVs unique to each patient-splicegraph as well.

5.2.3. Identifying and ranking outliers

We quantify PSI in cases and controls to identify and rank outliers using the corrected coverage from Section 5.2.2. We identify events with junctions and retained introns that have a *gap* between the extreme quantiles of the controls (from empirical distribution of PSI posterior means) and each patient (from posterior distribution of PSI). We prioritize these events (and, similarly, genes) with a gap in PSI first by whether they are novel to the patient (i.e., including the patient as its own build group changes the structure of the event) and then by the size of the largest gap between quantiles for the event.

The gaps in PSI between patients and controls are parameterized by a parameter α used similarly to a "significance level" from two-sided null hypothesis tests. Specifically, we calculate the quantiles corresponding to $\alpha/2$ and $1 - \alpha/2$. Over controls, the quantiles are taken over the empirical distribution of PSI posterior means. For each patient, the quantiles are taken over the posterior distribution of PSI. When the intervals defined by these quantiles overlap, there is no gap, otherwise, the gap is quantified as the difference in PSI between the closest quantiles. That

is, if $\Psi_{\{\text{controls,patient}\}}(q)$ is the q-th quantile from the controls or patient distribution, we define

$$\Psi_{\text{gap}}(\alpha) \equiv \begin{cases} \Psi_{\text{controls}}\left(\frac{\alpha}{2}\right) - \Psi_{\text{patient}}\left(1-\frac{\alpha}{2}\right), & \text{if } \Psi_{\text{controls}}\left(\frac{\alpha}{2}\right) > \Psi_{\text{patient}}\left(1-\frac{\alpha}{2}\right), \\ \Psi_{\text{patient}}\left(\frac{\alpha}{2}\right) - \Psi_{\text{controls}}\left(1-\frac{\alpha}{2}\right), & \text{if } \Psi_{\text{patient}}\left(\frac{\alpha}{2}\right) > \Psi_{\text{controls}}\left(1-\frac{\alpha}{2}\right), \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

Figure 5.1 provides a graphical illustration.

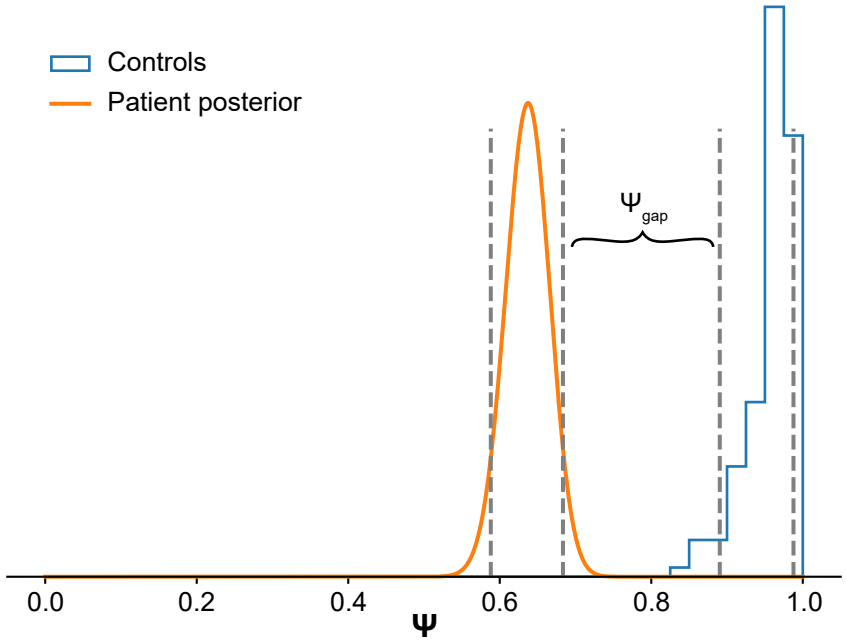


Figure 5.1: MAJIQ outliers have a gap between extreme quantiles of controls and patient distributions. For a retained intron or junction in an LSV, the gap between distributions Ψ_{gap} (defined in Equation (5.1)) is the distance between extreme quantiles of the empirical distribution of PSI posterior means from controls and the posterior distribution of PSI in the patient.

We only identify outliers for events which are quantifiable in the patient and over controls (quantifiable in some minimum number of experiments in controls; default 75%). We also ignore gaps where the difference in Ψ between the median of controls and patient posterior mean is less than some minimum threshold (default 10%).

5.2.4. Patient and control samples

We analyzed patient RNA-seq samples from three different datasets: (1) the Cummings dataset (Cummings et al., 2017), (2) the Baralle dataset (unpublished), and (3) the UDN dataset (Murdock et al., 2021). The Cummings and UDN datasets were downloaded from dbGaP, while the Baralle dataset was shared with us by collaborators (Dr. Diana Baralle's lab).

The Cummings dataset is composed of 53 RNA-seq samples from muscle biopsies of patients with previously genetically undiagnosed rare muscle disorders. From these data and matched exome or genome sequencing, Cummings et al. (2017) identified molecular diagnoses in 25 patients. Of these, we determined that 16 patients were candidates for RNA-first detection as splicing outliers (the others: 4 allele-specific expression, 2 large structural variants, and 3 core-splice site variants identified from genetic sequencing but with insufficient coverage for detection from RNA-seq). All samples were sequenced at the same site, so we treated these data as if there were no confounders.

The Baralle dataset is composed of 55 RNA-seq samples from globin-depleted whole blood from patients with suspected Mendelian disorders with diverse phenotypes. These data were sequenced in three batches with 7, 16, and 32 samples each, so we used batch identity as a confounder in our analyses.

The UDN dataset is composed of 833 RNA-seq samples from fibroblasts ($n = 400$), whole blood ($n = 380$), and other tissues ($n = 53$) from patients with suspected Mendelian disorders with diverse phenotypes and their affected/unaffected relatives. Murdock et al. (2021) describes 14 patients with causative variants which were identified by their RNA-first approach (using DROP/FRASER plus additional filtering). Of these, data from 8 patients (11 samples from blood and fibroblasts) were available as part of the UDN dataset. These data were sequenced with four different sequencing instruments, so we analyzed each tissue separately using sequencing instrument as a known confounder.

While batch correction and detection of outliers was limited to samples from within

each individual dataset, we included samples with RIN score greater than 6 from GTEx v8 as additional build groups for constructing splicegraphs. For the Cummings dataset, we included all skeletal muscle samples as a build group. For the Baralle dataset, we included all GTEx whole blood and EBV-transformed lymphocytes samples as build groups. For the UDN fibroblasts, we included all GTEx cultured fibroblasts as a build group. For the UDN whole blood samples, we included all GTEx whole blood samples as a build group.

5.2.5. Sample read alignment

We aligned RNA-seq reads from input samples following the same procedure as in Chapter 3. That is, we aligned RNA-seq reads from samples to the human genome for splicing analysis with MAJIQ using the following procedure. We performed quality and adapter trimming on each sample using TrimGalore (v0.4.5). We used STAR (v2.5.3a) to perform a two-step gapped alignment of the trimmed reads to the GRCh38 primary assembly with annotations from Ensembl release 94.

5.2.6. Running other tools for splicing outliers

LeafcutterMD

We downloaded Leafcutter (commit hash `63b347a3`) from Github. We followed the standard LeafcutterMD pipeline as described in the online documentation: (1) count excised introns (junctions) using `bam2junc.sh`, (2) identify intron clusters and counts using `leafcutter_cluster.py` (with maximum intron length of 500,000, required minimum reads in a cluster as 50), (3) identify outlier clusters using `leafcutterMD.R` with default parameters.

Neither the LeafcutterMD paper (Jenkinson et al., 2020) nor the online documentation gave any guidance as to how to translate cluster or intron p-values to prioritized clusters or genes besides mentioning FDR correction. So, we adopted the following procedure and cutoffs to consider outliers with respect to genes. We performed FDR adjustment on cluster p-values using the Benjamini–Yekutieli procedure. We then identified outlier clusters using an FDR cutoff of 10%. We prioritized clusters by adjusted p-values. We prioritized genes by mapping clusters back to Ensembl release 94 annotations and scoring by the minimum adjusted p-value.

FRASER

We used FRASER as part of the DROP pipeline. We installed DROP (v1.1.4) using conda. We ran DROP/FRASER with Ensembl release 94 annotations using default parameters (FDR cutoff of 10%, dPSI cutoff of 5%).

5.2.7. Comparison of splicing outlier tools

We evaluated MAJIQ-CLIN as compared to LeafcutterMD and FRASER in the following ways. First, in order to compare the number of genes that would need to be reviewed for a given patient, we compared the overall number of outliers the method produced for each sample. Since clinical analysis is typically restricted to known disease-causing genes, we also filtered to known disease causing genes (using gene lists from Murdock et al. (2021)). Since MAJIQ-CLIN is particularly interested in splicing events which are novel to their patient-specific splicegraph, we also considered gene counts when restricting to these structurally novel outliers. Second, for the samples in the Cummings and UDN datasets that had been previously solved, we evaluated (1) if each method identified the gene as having an outlier, and (2) how it prioritized the gene vs others.

5.3. Results

We ran MAJIQ-CLIN, LeafcutterMD, and FRASER on the Cummings, Baralle, and UDN datasets. We ran FRASER and LeafcutterMD with up to 4 threads, and MAJIQ-CLIN with up to 2 threads on the CHOP cluster. MAJIQ-CLIN was the fastest tool for outlier prioritization in each dataset (Table 5.1). We were unable to successfully run FRASER on the UDN dataset in time for the submission of this dissertation. Running FRASER required allocating a surprising amount of resources. The DROP pipeline repeatedly exceeded over 64GB RAM when counting split and unsplit reads and was only able to finish when given 96GB RAM (twice what is required for alignment). However, we were able to proceed through the pipeline far enough that we know that the lower bound runtime is greater than that for MAJIQ-CLIN.

Without filtering on genes or structural novelty, MAJIQ-CLIN identifies more genes than

Dataset	MAJIQ-CLIN	FRASER	LeafcutterMD
Cummings ($n = 53$)	**4h09m**	8h56m	24h20m
Baralle ($n = 55$)	**8h01m**	1d,19h00m	23h18m
UDN-fibroblasts ($n = 400$)	**3d,20h01m**	4d,12h02m (lower bound)	14d,10h15m

Table 5.1: MAJIQ-CLIN runs efficiently vs FRASER or LeafcutterMD. MAJIQ-CLIN was run with up to 2 threads per job, while LeafcutterMD and FRASER used up to 4 threads. Runtimes with MAJIQ-CLIN excludes processing GTEx alignments for coverage to compare processing same number of alignments (and because they would typically have been preprocessed as part of an incremental build). FRASER runtime for UDN-fibroblasts is a lower bound. We were unable to finish rerunning FRASER in time for the submission of this dissertation due to high memory usage (requiring over 64GB RAM). The reported time is for the counting, PSI calculation, and filtering steps, plus six hours which had been run without completion towards fitting hyperparameters. This excludes fitting the autoencoder, calculating p-values, and extracting results steps. For comparison, these steps took an additional 1.5 hours for the Cummings dataset and 3 hours for the Baralle dataset.

other methods, especially on the Baralle dataset (Figure 5.2). The number of outliers drops significantly when restricted to known disease-causing genes (Figure 5.3). MAJIQ-CLIN still identifies more genes than other methods. However, when MAJIQ-CLIN uses patient splicegraphs to filter on novel LSVs, the number of outliers becomes less or comparable to the other methods.

For all of the previously solved cases from Cummings et al. (2017), Both MAJIQ-CLIN and FRASER successfully identify outliers in the same genes as the 16 previously solved cases from Cummings et al. (2017) amenable to splicing outlier analysis, while LeafcutterMD only identifies the known gene in 6 of the cases (Figure 5.4). Focusing on disease-causing genes, MAJIQ prioritizes the disease-causing gene first in all but two samples. In all samples, MAJIQ identifies each gene as a novel event outlier. MAJIQ gives higher priority to the disease-causing gene in all but one case, while FRASER prioritizes fewer other genes in all but 3 cases.

Looking at previously solved cases from the UDN dataset we can indirectly compare MAJIQ-CLIN to FRASER because findings from Murdock et al. (2021) were made with FRASER. We find that MAJIQ-CLIN identifies a change that FRASER did not. First, Table 2 of Murdock et al. (2021) reports a patient with *AP4M1*-associated spastic paraplegia which was identified by outlier expression but not by splicing. MAJIQ-CLIN lists *AP4M1* as the second highest

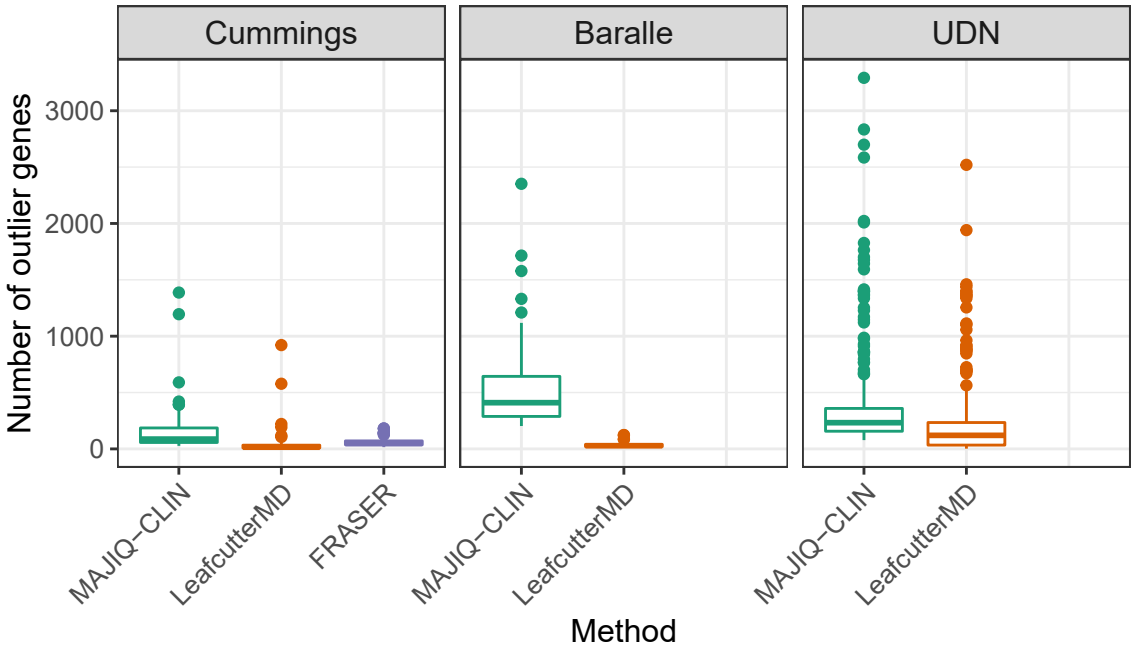


Figure 5.2: Number of outlier genes (all) for each method on each dataset. MAJIQ-CLIN identifies more outlier genes than other methods, especially on the Baralle dataset.

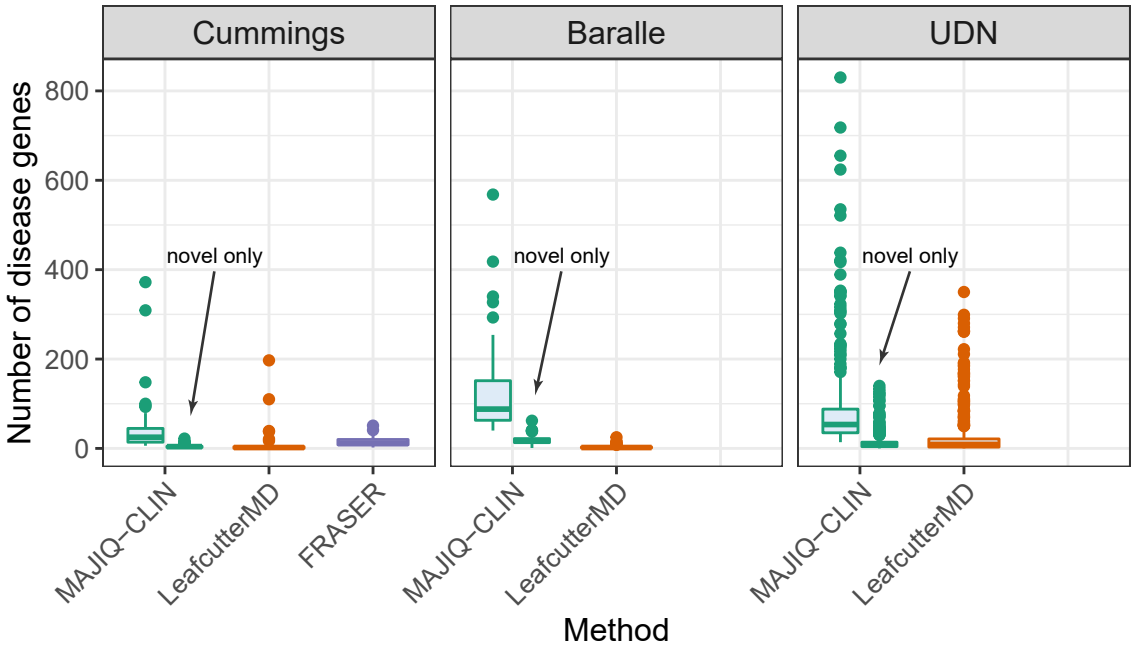


Figure 5.3: Number of outlier disease genes for each method on each dataset. MAJIQ-CLIN outlier disease genes (dodged left) can further be filtered to outlier novel event disease genes (dodged right): that is, genes where at least one of the gaps in PSI was found in an LSV which is structurally different when the patient is excluded from the splicegraph.

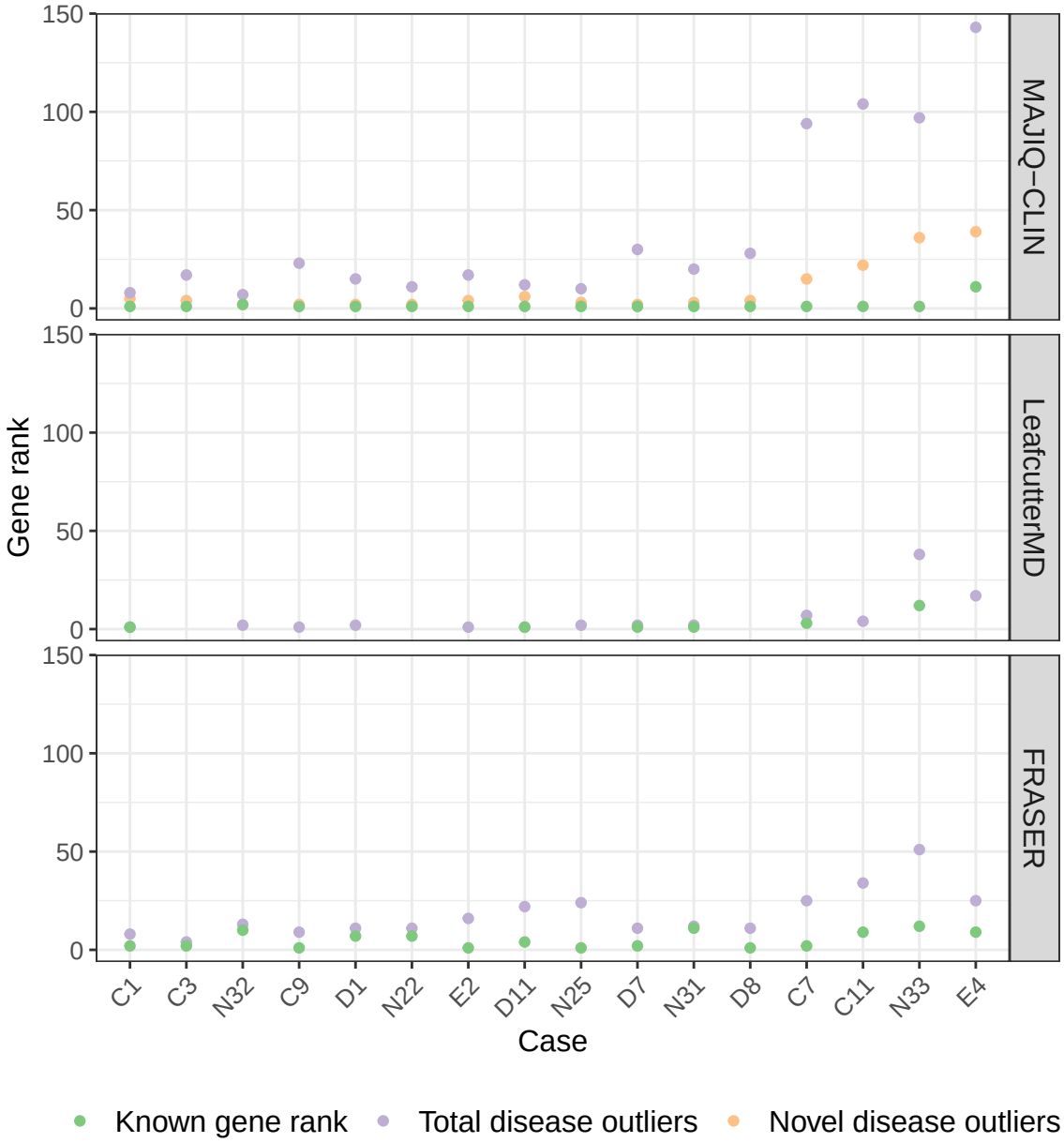


Figure 5.4: MAJIQ-CLIN successfully prioritizes known disease genes from Cummings et al. (2017). MAJIQ-CLIN and FRASER successfully solve all 16 cases from Cummings et al. (2017) amenable to splicer outlier analysis. LeafcutterMD only solves 6 of 16 cases (38%).

priority and specifically highlights the novel 52bp exon extension and intron retention which the previous authors identified after following-up on the expression outlier. MAJIQ-CLIN gives priority to the patients with disease-causing variants in *RPL13* (rank 1), *TBCK* (rank 1), However, MAJIQ-CLIN did not successfully identify an outlier in the patient with Au-Kline syndrome (*HNRNPK*), which the paper reports being identified by both expression and FRASER splicing outlier analysis.

5.4. Discussion

We developed MAJIQ-CLIN as a pipeline to efficiently and accurately detect splicing outliers in RNA-seq from patients with suspected Mendelian disorders. MAJIQ was substantially faster than either LeafcutterMD or FRASER. FRASER was faster than LeafcutterMD on the Cummings dataset but slower on the Baralle dataset. We were unable to finish running FRASER on the UDN fibroblasts in time for submitting this dissertation. However, we were able to run the FRASER pipeline far enough to show that the runtime is greater than that of MAJIQ-CLIN by at least almost a day.

We were surprised by FRASER's high demand for computational resources. Counting split and unsplit reads with FRASER required over 64GB memory for experiments from UDN fibroblasts. This is more than 100 times the memory than the equivalent step with MAJIQ (less than 600MB, see Figure 4.5b).

With current thresholds, MAJIQ-CLIN reports more outliers than either LeafcutterMD or FRASER. This means that without additional changes, clinicians using MAJIQ-CLIN would need to spend more time reviewing these additional genes. For the case of LeafcutterMD, our analysis on the Cummings dataset suggests that LeafcutterMD is too conservative, resulting in missed diagnoses. On the other hand, FRASER manages to solve all the Cummings cases with fewer prioritized genes. However, MAJIQ-CLIN consistently prioritized the known disease gene first or second in all but two of the solved cases. For the Cummings dataset, MAJIQ-CLIN gave higher priority to the disease gene than FRASER in all but one case. More work needs to be done to further strengthen how MAJIQ-CLIN filters outliers, and the high priority given to known disease

causing genes suggests there will be room to do so.

Our analysis is inconclusive about MAJIQ-CLIN vs FRASER. As we were unable to run FRASER on all the other datasets, it is altogether possible that FRASER's number of outliers would be even higher. We note that all known disease genes were identified as novel events, which suggests that we could use that as an additional hard filtering criteria. This would certainly significantly drop the number of outliers to review, but at the potential expense of missing other disease genes. The primary analysis being on the Cummings dataset raises the possibility of selection bias because the original methodology used in Cummings et al. (2017) was dependent entirely on novelty.

Our results on the UDN dataset are challenging to interpret. While identifying the *AP4M1* patient by splicing was encouraging (just as not identifying the *HNRNPK* patient by splicing was the opposite), these differences could arise entirely from annotations (Murdock et al. (2021) used GRCh37, we used GRCh38) or some other small detail. In future work, we will reach out to the authors of DROP/FRASER to resolve our current computational challenges so we can more directly compare MAJIQ-CLIN to FRASER on these data.

MAJIQ-CLIN compares favorably to LeafcutterMD and FRASER. In contrast to MAJIQ-CLIN and FRASER, LeafcutterMD does not quantify intron retention, nor does it correct for the effect of confounding factors. These differences could contribute to LeafcutterMD's missed diagnoses on the Cummings dataset. MAJIQ-CLIN more readily permits analyzing additional samples with its two pass approach. While FRASER and LeafcutterMD can reuse their intermediate coverage files in subsequent analyses, there are several limitations. For LeafcutterMD, new samples require reperforming "intron clustering" and subsequent modeling steps over the entire dataset. For FRASER, the reusable intermediate coverage files only count intron coverage at splice sites from the original dataset. If new samples introduce novel junctions with novel splice sites, FRASER will be unable to quantify intron retention in competition with these new junctions without reprocessing coverage from the original alignments. FRASER also requires performing all modeling steps again from scratch. In contrast, MAJIQ-CLIN can reuse

quantifications of controls and models of unobserved confounders/coverage over the original set of first-pass LSVs. In this case, the only additional processing on controls that MAJIQ-CLIN requires is to extract coverage from SJ files for the small set of second-pass LSVs the new sample introduces, followed by modeling coverage and quantification.

Our analyses mostly focused on known disease-causing genes for the purposes of diagnosing patients with suspected Mendelian disorders. The machinery we developed in this chapter for detecting outliers could be adapted in future work to other settings, including identifying genes of unknown significance or neoantigens for cancer immunotherapy.

Overall, the analyses in this chapter demonstrate that MAJIQ-CLIN is capable of accurately and efficiently detect disease-causing splicing outliers in patients with suspected Mendelian disorders. In doing so, MAJIQ-CLIN can contribute to efforts to use RNA splicing to improve molecular diagnosis rates, thereby helping provide more accurate prognoses and improved clinical care for patients and their families.

CHAPTER 6

Conclusions and future directions

6.1. Conclusions

This dissertation was motivated by the challenge of improving the molecular diagnostic rate in patients with suspected Mendelian disorders. With exome sequencing or even genome sequencing, less than half of these patients are expected to receive a moleclar diagnosis. In order to address this challenge, we focused on methodology for detecting rare splicing aberrations in bulk RNA-seq data from these patients. RNA-seq provides a functional readout of the impact of genetic variation on gene expression and splicing. Thus, clinical RNA-seq provides an orthogonal approach for identifying these rare splicing aberrations which we know are able to cause disease but are challenging to identify with current methods from genetic sequence alone.

In Chapter 2, I describe my contribution as a co-first author to methods in MAJIQ v2 that enable us to accurately quantify splicing in large/heterogeneous populations as found from human RNA-seq (Vaquero-Garcia et al., 2021). In Chapter 3, I used MAJIQ v2 to elucidate the limitations of RNA-seq with clinically accessible tissues for detecting changes due to tissue-specific expression and splicing (Aicher et al., 2020). However, as we began to apply MAJIQ v2 to patient RNA-seq, informed by our analyses in Chapter 3, it became clear that MAJIQ v2 would be insufficient due to challenges with rare artifacts in intron retention quantification and the need to further minimize repeated work in analyzing control samples. Thus, in Chapter 4, I describe how I developed MAJIQ v3 to address these issues. Finally, in Chapter 5, I describe the methodology I built up on the lessons from these previous chapters to identify splicing aberrations from patient RNA-seq samples and show that it compares favorably to previous tools.

6.2. Future directions

While the work presented here in this dissertation contributes to addressing the challenge of improving the molecular diagnostic rate for patients with suspected Mendelian disorders, as well as to how we can model and quantify splicing in general, there is still more work to be done.

In the remainder of this chapter, I describe potential future directions.

This dissertation was focused on observed changes in RNA splicing from bulk RNA-seq as a signal for improving the molecular diagnostic rate. One avenue for future work could be extending the tools developed for splicing in MAJIQ to detect other transcriptomic aberrations. While expression outliers or allele-specific outliers have well-established methods such as OUTRIDER (Brechtmann et al., 2018) and ANEVA-DOT (Mohammadi et al., 2019), the changes they look for are not well-suited for analysis with MAJIQ. However, Cummings et al. (2017) suggests that RNA-seq could also be used to help detect structural variants. They describe a few patients where plotting read coverage over the exons of *DMD* shows a qualitative decrease in coverage at the 3' of the genes consistent with large inversions intersecting the gene which were later confirmed by genome sequencing. Figure 6.1 shows how total LSV coverage measured by MAJIQ can be used to reproduce Fig S8E from Cummings et al. (2017). Future work could develop tests to identify outliers in coverage across larger scale splicing modules or entire genes.

Modeling coverage across the body of a gene, as described in the previous paragraph, relates to more general further improvements to how MAJIQ quantifies changes within a gene. MAJIQ currently focuses on coverage from split reads and coverage aligned to intronic regions. The approach to intron retention coverage introduced in MAJIQ v3 could easily be extended to assess coverage more generally across genes, split by annotated and novel splice sites. This could strengthen the analyses described in the previous paragraph. Furthermore, one weakness of MAJIQ is its difficulty quantifying events with novel exitrons (that is, junctions that start and end in the same exon). Wai et al. (2020) describes a patient case where the putative causal splicing change involved an exitron; as a result, our clinical pipeline would likely fail on this case. MAJIQ's difficulty with exitrons arises precisely because the incremental coverage MAJIQ measures excludes exons, so MAJIQ is unable to comment on intronic coverage underneath the exitron without reprocessing input alignments. Exon coverage could also improve detection and quantification of other relevant changes, including alternative transcript starts/ends, into which MAJIQ currently provides less insight. As a result, expanding how MAJIQ processes read

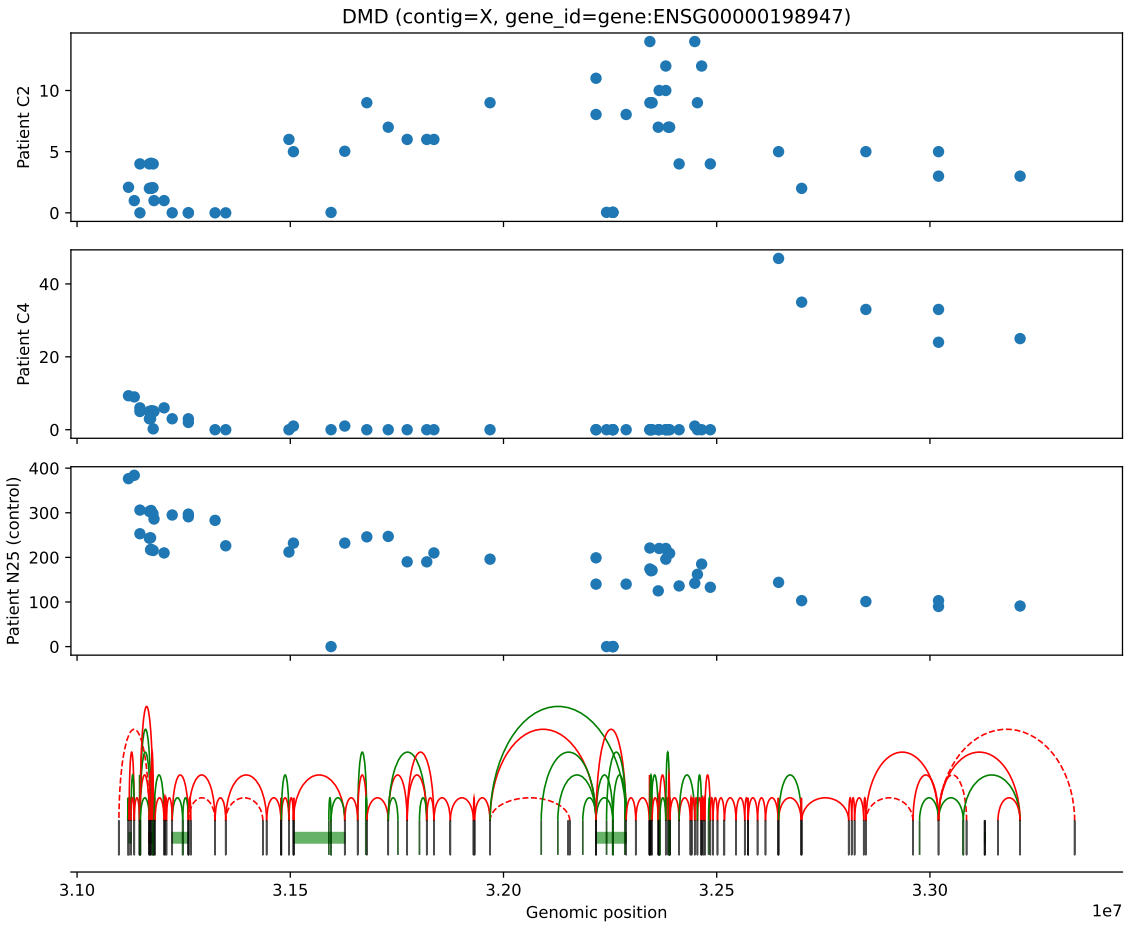


Figure 6.1: MAJIQ could be used to study structural changes in coverage across the body of genes. The plots show total coverage in LSVs across *DMD* in patients where structural variants overlapping *DMD* were identified as causal (C2, C4) and a representative control (N25) (as previously described in Cummings et al. (2017), Fig S8E).

coverage in exons to enable the quantification of additional transcriptomic features is a promising avenue of future work.

Future work could also involve adopting new technologies such as long-read RNA sequencing. Long-read RNA sequencing could be used to observe larger structural variants and phasing of coordinated splicing. However, MAJIQ is specifically designed for analyzing short-read RNA-seq. Future work could be directed towards adapting MAJIQ for modeling and analyzing long-read RNA-seq data, followed by subsequent application of the methodology to patients with suspected Mendelian disorders.

Finally, the analysis in Chapter 3 showed that over 6% of consistently spliced genes in various clinically-inaccessible tissues would not be adequately represented when measuring splicing in clinically-accessible tissues. These genes are a potential blind spot of the clinical pipeline developed in Chapter 5. Future work should be done developing "DNA-first" approaches to predict tissue-specific splicing defects from rare genetic variants to complement the "RNA-first" approaches described in this dissertation. These approaches could use recent advances in machine learning from natural language processing and computer vision to advance on previously published methods for splicing predictions (e.g., Table 1.1). In addition to building prediction models on sequence alone, future models could integrate splicing quantifications from accessible tissues.

BIBLIOGRAPHY

Joseph K. Aicher, Paul Jewell, Jorge Vaquero-Garcia, Yoseph Barash, and Elizabeth J. Bhoj. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, March 2020. ISSN 1530-0366. doi: 10.1038/s41436-020-0780-y.

Ahmed Alfares, Taghrid Aloraini, Lamia Al Subaie, Abdulelah Alissa, Ahmed Al Qudsi, Ahmed Alahmad, Fuad Al Mutairi, Abdulrahman Alswaid, Ali Alothaim, Wafaa Eyaid, Mohammed Albalwi, Saeed Alturki, and Majid Alfadhel. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genetics in Medicine*, 20(11):1328, November 2018. ISSN 1530-0366. doi: 10.1038/gim.2018.41. URL https://proxy.library.upenn.edu:2611/articles/gim201841.

Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):30, February 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1935-5. URL https://doi.org/10.1186/s13059-020-1935-5.

Yoseph Barash, Elinor Dehan, Meir Krupsky, Wilbur Franklin, Marc Geraci, Nir Friedman, and Naftali Kaminski. Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. *Bioinformatics*, 20(6):839–846, 01 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg487. URL https://doi.org/10.1093/bioinformatics/btg487.

Yoseph Barash, Benjamin J. Blencowe, and Brendan J. Frey. Model-based detection of alternative splicing signals. *Bioinformatics (Oxford, England)*, 26(12):i325–333, June 2010a. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq200.

Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010b. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09000. URL http://www.nature.com/articles/nature09000.

Amir Ben-Dor, Nir Friedman, and Zohar Yakhini. Overabundance analysis and class discovery in gene expression data. *Agilent Laboratories, Palo Alto, Tech. Rep*, 2002.

Felix Brechtmann, Christian Mertes, Agnė Matusevičiūtė, Vicente A. Yépez, Žiga Avsec, Maximilian Herzog, Daniel M. Bader, Holger Prokisch, and Julien Gagneur. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *The American Journal of Human Genetics*, 103(6):907–917, December 2018. ISSN 0002-9297. doi: 10.1016/j.ajhg.2018.10.025. URL https://www.sciencedirect.com/science/article/pii/S0002929718304014.

Jun Cheng, Thi Yen Duong Nguyen, Kamil J. Cygan, Muhammed Hasan Çelik, William G.

Fairbrother, žiga Avsec, and Julien Gagneur. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, 20, March 2019. ISSN 1474-7596. doi: 10.1186/s13059-019-1653-z. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6396468/.

Jun Cheng, Muhammed Hasan Çelik, Anshul Kundaje, and Julien Gagneur. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biology*, 22(1):94, March 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02273-7. URL https://doi.org/10.1186/s13059-021-02273-7.

Rocky Cheung, Kimberly D. Insigne, David Yao, Christina P. Burghard, Jeffrey Wang, Yun-Hua E. Hsiao, Eric M. Jones, Daniel B. Goodman, Xinshu Xiao, and Sriram Kosuri. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Molecular Cell*, 73(1):183–194.e8, January 2019. ISSN 1097-4164. doi: 10.1016/j.molcel.2018.10.037.

Michelle M. Clark, Zornitza Stark, Lauge Farnaes, Tiong Y. Tan, Susan M. White, David Dimmock, and Stephen F. Kingsmore. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genomic Medicine*, 3, July 2018. ISSN 2056-7944. doi: 10.1038/s41525-018-0053-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6037748/.

Beryl B. Cummings, Jamie L. Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A. Reghan Foley, Veronique Bolduc, Leigh B. Waddell, Sarah A. Sandaradura, Gina L. O'Grady, Elicia Estrella, Hemakumar M. Reddy, Fengmei Zhao, Ben Weisburd, Konrad J. Karczewski, Anne H. O'Donnell-Luria, Daniel Birnbaum, Anna Sarkozy, Ying Hu, Hernan Gonorazky, Kristl Claeys, Himanshu Joshi, Adam Bournazos, Emily C. Oates, Roula Ghaoui, Mark R. Davis, Nigel G. Laing, Ana Topf, Genotype-Tissue Expression Consortium, Peter B. Kang, Alan H. Beggs, Kathryn N. North, Volker Straub, James J. Dowling, Francesco Muntoni, Nigel F. Clarke, Sandra T. Cooper, Carsten G. Bönnemann, and Daniel G. MacArthur. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine*, 9(386), 2017. ISSN 1946-6242. doi: 10.1126/scitranslmed.aal5209.

Kelly D. Farwell, Layla Shahmirzadi, Dima El-Khechen, Zöe Powis, Elizabeth C. Chao, Brigette Tippin Davis, Ruth M. Baxter, Wenqi Zeng, Cameron Mroske, Melissa C. Parra, Stephanie K. Gandomi, Ira Lu, Xiang Li, Hong Lu, Hsiao-Mei Lu, David Salvador, David Ruble, Monica Lao, Soren Fischbach, Jennifer Wen, Shela Lee, Aaron Elliott, Charles L. M. Dunlop, and Sha Tang. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 17(7):578–586, July 2015. ISSN 1530-0366. doi: 10.1038/gim.2014.154.

Aimee L Fenwick, Jacqueline AC Goos, Julia Rankin, Helen Lord, Tracy Lester, A Jeannette M Hoogeboom, Ans MW van den Ouweland, Steven A Wall, Irene MJ Mathijssen, and Andrew OM Wilkie. Apparently synonymous substitutions in FGFR2 affect splicing and result in mild Crouzon

syndrome. *BMC Medical Genetics*, 15:95, August 2014. ISSN 1471-2350. doi: 10.1186/s12881-014-0095-4. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236556/.

Laure Frésard and Stephen B. Montgomery. Diagnosing rare diseases after the exome. *Molecular Case Studies*, 4(6):a003392, December 2018. ISSN , 2373-2873. doi: 10.1101/mcs.a003392. URL http://molecularcasestudies.cshlp.org/content/4/6/a003392.

Laure Frésard, Craig Smail, Nicole M. Ferraro, Nicole A. Teran, Xin Li, Kevin S. Smith, Devon Bonner, Kristin D. Kernohan, Shruti Marwaha, Zachary Zappala, Brunilda Balliu, Joe R. Davis, Boxiang Liu, Cameron J. Prybol, Jennefer N. Kohler, Diane B. Zastrow, Chloe M. Reuter, Dianna G. Fisk, Megan E. Grove, Jean M. Davidson, Taila Hartley, Ruchi Joshi, Benjamin J. Strober, Sowmithri Utiramerur, Undiagnosed Diseases Network, Care4Rare Canada Consortium, Lars Lind, Erik Ingelsson, Alexis Battle, Gill Bejerano, Jonathan A. Bernstein, Euan A. Ashley, Kym M. Boycott, Jason D. Merker, Matthew T. Wheeler, and Stephen B. Montgomery. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, 25(6):911–919, 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0457-8.

Brian S. Gloss and Marcel E. Dinger. Realizing the significance of noncoding functionality in clinical genomics. *Experimental & Molecular Medicine*, 50(8):97, August 2018. ISSN 2092-6413. doi: 10.1038/s12276-018-0087-0. URL https://www.nature.com/articles/s12276-018-0087-0.

Hernan D. Gonorazky, Sergey Naumenko, Arun K. Ramani, Viswateja Nelakuditi, Pouria Mashouri, Peiqui Wang, Dennis Kao, Krish Ohri, Senthuri Viththiyapaskaran, Mark A. Tarnopolsky, Katherine D. Mathews, Steven A. Moore, Andres N. Osorio, David Villanova, Dwi U. Kemaladewi, Ronald D. Cohn, Michael Brudno, and James J. Dowling. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *The American Journal of Human Genetics*, 104(3):466–483, March 2019. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2019.01.012. URL https://www.cell.com/ajhg/abstract/S0002-9297(19)30012-6.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Alexis Battle, Christopher D. Brown, Barbara E. Engelhardt, and Stephen B. Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017. ISSN 1476-4687. doi: 10.1038/nature24277.

Yi-Fei Huang, Brad Gulko, and Adam Siepel. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics*, 49(4):618–624, April

2017. ISSN 1061-4036. doi: 10.1038/ng.3810. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395419/.

Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D. Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2):214–220, February 2016. ISSN 1546-1718. doi: 10.1038/ng.3477.

Karthik A. Jagadeesh, Aaron M. Wenger, Mark J. Berger, Harendra Guturu, Peter D. Stenson, David N. Cooper, Jonathan A. Bernstein, and Gill Bejerano. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, 48 (12):1581–1586, 2016. ISSN 1546-1718. doi: 10.1038/ng.3703.

Karthik A. Jagadeesh, Joseph M. Paggi, James S. Ye, Peter D. Stenson, David N. Cooper, Jonathan A. Bernstein, and Gill Bejerano. S-CAP extends clinical-grade pathogenicity prediction to genetic variants that affect RNA splicing. *bioRxiv*, page 343749, June 2018. doi: 10.1101/343749. URL https://www.biorxiv.org/content/early/2018/06/20/343749.

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B. Schwartz, Eric D. Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J. Sanders, and Kyle Kai-How Farh. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 0(0), January 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.12.015. URL https://www.cell.com/cell/abstract/S0092-8674(18)31629-5.

Garrett Jenkinson, Yang I Li, Shubham Basu, Margot A Cousin, Gavin R Oliver, and Eric W Klee. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics*, 36 (17):4609–4615, November 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa259. URL https://doi.org/10.1093/bioinformatics/btaa259.

Anupama Jha, Matthew R. Gazzara, and Yoseph Barash. Integrative deep models for alternative splicing. *Bioinformatics (Oxford, England)*, 33(14):i274–i282, July 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx268.

Anupama Jha, Joseph K. Aicher, Matthew R. Gazzara, Deependra Singh, and Yoseph Barash. Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biology*, 21, June 2020. ISSN 1474-7596. doi: 10.1186/s13059-020-02055-7. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7305616/.

Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, March 2014. ISSN 1061-4036. doi: 10.1038/ng.2892. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992975/.

Laura S. Kremer, Daniel M. Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B. Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, Eliška

Koňaříková, Birgit Repp, Gabi Kastenmüller, Jerzy Adamski, Peter Lichtner, Christoph Leonhardt, Benoit Funalot, Alice Donati, Valeria Tiranti, Anne Lombes, Claude Jardel, Dieter Gläser, Robert W. Taylor, Daniele Ghezzi, Johannes A. Mayr, Agnes Rötig, Peter Freisinger, Felix Distelmaier, Tim M. Strom, Thomas Meitinger, Julien Gagneur, and Holger Prokisch. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications*, 8: 15824, 2017. ISSN 2041-1723. doi: 10.1038/ncomms15824.

Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue): D980–D985, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965032/.

Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3): 1752–1779, September 2011. ISSN 1932-6157, 1941-7330. doi: 10.1214/11-AOAS466. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-5/issue-3/Measuring-reproducibility-of-high-throughput-experiments/10.1214/11-AOAS466.full. Publisher: Institute of Mathematical Statistics.

Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, January 2018. ISSN 1546-1718. doi: 10.1038/s41588-017-0004-9. URL https://www.nature.com/articles/s41588-017-0004-9. Number: 1 Publisher: Nature Publishing Group.

Susan J. Lindsay, Yaobo Xu, Steven N. Lisgo, Lauren F. Harkin, Andrew J. Copp, Dianne Gerrelli, Gavin J. Clowry, Aysha Talbot, Michael J. Keogh, Jonathan Coxhead, Mauro Santibanez-Koref, and Patrick F. Chinnery. HDBR Expression: A Unique Resource for Global and Individual Gene Expression Studies during Early Human Brain Development. *Frontiers in Neuroanatomy*, 10, October 2016. ISSN 1662-5129. doi: 10.3389/fnana.2016.00086. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5080337/.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

Christian Mertes, Ines F. Scheller, Vicente A. Yépez, Muhammed H. Çelik, Yingjiqiong Liang, Laura S. Kremer, Mirjana Gusic, Holger Prokisch, and Julien Gagneur. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nature Communications*, 12(1):529, January 2021. ISSN 2041-1723. doi: 10.1038/s41467-020-20573-7. URL https://www.nature.com/articles/s41467-020-20573-7. Number: 1 Publisher: Nature Publishing Group.

Pejman Mohammadi, Stephane E. Castel, Beryl B. Cummings, Jonah Einson, Christina Sousa, Paul Hoffman, Sandra Donkervoort, Zhuoxun Jiang, Payam Mohassel, A. Reghan Foley,

Heather E. Wheeler, Hae Kyung Im, Carsten G. Bonnemann, Daniel G. MacArthur, and Tuuli Lappalainen. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science (New York, N.Y.)*, 366(6463):351–356, October 2019. ISSN 0036-8075. doi: 10.1126/science.aay0256. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6814274/.

David R. Murdock, Hongzheng Dai, Lindsay C. Burrage, Jill A. Rosenfeld, Shamika Ketkar, Michaela F. Müller, Vicente A. Yépez, Julien Gagneur, Pengfei Liu, Shan Chen, Mahim Jain, Gladys Zapata, Carlos A. Bacino, Hsiao-Tuan Chao, Paolo Moretti, William J. Craigen, Neil A. Hanchard, and Brendan Lee. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *The Journal of Clinical Investigation*, 131(1), January 2021. ISSN 0021-9738. doi: 10.1172/JCI141500. URL https://www.jci.org/articles/view/141500. Publisher: American Society for Clinical Investigation.

Scott S. Norton, Jorge Vaquero-Garcia, Nicholas F. Lahens, Gregory R. Grant, and Yoseph Barash. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. *Bioinformatics (Oxford, England)*, 34(9):1488–1497, May 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx790.

Eleftheria Pervolaraki, James Dachtler, Richard A. Anderson, and Arun V. Holden. The developmental transcriptome of the human heart. *Scientific Reports*, 8, October 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-33837-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6194117/.

Towfique Raj, Yang I. Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying-Chih Wang, Bernard Ng, Ishaan Gupta, Vahram Haroutunian, Eric E. Schadt, Tracy Young-Pearse, Sara Mostafavi, Bin Zhang, Pamela Sklar, David A. Bennett, and Philip L. De Jager. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature genetics*, 50(11):1584–1592, November 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0238-1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6354244/.

Kyle Retterer, Jane Juusola, Megan T. Cho, Patrik Vitazka, Francisca Millan, Federica Gibellini, Annette Vertino-Bell, Nizar Smaoui, Julie Neidich, Kristin G. Monaghan, Dianalee McKnight, Renkui Bai, Sharon Suchy, Bethany Friedman, Jackie Tahiliani, Daniel Pineda-Alvarez, Gabriele Richard, Tracy Brandt, Eden Haverfield, Wendy K. Chung, and Sherri Bale. Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 18(7):696–704, 2016. ISSN 1530-0366. doi: 10.1038/gim.2015.148.

Graham R. S. Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of non-coding sequence variants. *Nature methods*, 11(3):294–296, March 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2832. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5015703/.

Alexander B. Rosenberg, Rupali P. Patwardhan, Jay Shendure, and Georg Seelig. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3):

698–711, October 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.09.054.

Jay P. Ross, Patrick A. Dion, and Guy A. Rouleau. Exome sequencing in genetic disease: recent advances and considerations. *F1000Research*, 9:F1000 Faculty Rev–336, May 2020. ISSN 2046-1402. doi: 10.12688/f1000research.19444.1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7205110/.

Marina M. Scotti and Maurice S. Swanson. RNA mis-splicing in disease. *Nature Reviews. Genetics*, 17(1):19–32, January 2016. ISSN 1471-0064. doi: 10.1038/nrg.2015.3.

Michael Seiler, Akihide Yoshimi, Rachel Darman, Betty Chan, Gregg Keaney, Michael Thomas, Anant A. Agrawal, Benjamin Caleb, Alfredo Csibi, Eckley Sean, Peter Fekkes, Craig Karr, Virginia Klimek, George Lai, Linda Lee, Pavan Kumar, Stanley Chun-Wei Lee, Xiang Liu, Crystal Mackenzie, Carol Meeske, Yoshiharu Mizui, Eric Padron, Eunice Park, Ermira Pazolli, Shouyong Peng, Sudeep Prajapati, Justin Taylor, Teng Teng, John Wang, Markus Warmuth, Huilan Yao, Lihua Yu, Ping Zhu, Omar Abdel-Wahab, Peter G. Smith, and Silvia Buonamici. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nature Medicine*, 24(4):497–504, May 2018. ISSN 1546-170X. doi: 10.1038/nm.4493.

Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D. Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51): E5593–E5601, December 2014. doi: 10.1073/pnas.1419161111. URL https://www.pnas.org/doi/10.1073/pnas.1419161111. Publisher: Proceedings of the National Academy of Sciences.

Ravi K. Singh and Thomas A. Cooper. Pre-mRNA splicing in disease and therapeutics. *Trends in molecular medicine*, 18(8):472–482, August 2012. ISSN 1471-4914. doi: 10.1016/j.molmed.2012.06.006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3411911/.

Rahul Sinha, Young Jin Kim, Tomoki Nomakuchi, Kentaro Sahashi, Yimin Hua, Frank Rigo, C. Frank Bennett, and Adrian R. Krainer. Antisense oligonucleotides correct the familial dysautonomia splicing defect in IKBKAP transgenic mice. *Nucleic Acids Research*, 46(10): 4833–4844, June 2018. ISSN 1362-4962. doi: 10.1093/nar/gky249.

Barry Slaff, Caleb M. Radens, Paul Jewell, Anupama Jha, Nicholas F. Lahens, Gregory R. Grant, Andrei Thomas-Tikhonenko, Kristen W. Lynch, and Yoseph Barash. MOCCASIN: a method for correcting for known and unknown confounders in RNA splicing analysis. *Nature Communications*, 12(1):3353, June 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23608-9. URL https://www.nature.com/articles/s41467-021-23608-9. Number: 1 Publisher: Nature Publishing Group.

Elena Sotillo, David M. Barrett, Kathryn L Black, Asen Bagashev, Derek Oldridge, Glendon Wu, Robyn Sussman, Claudia Lanauze, Marco Ruella, Matthew R. Gazzara, Nicole M. Martinez, Colleen T. Harrington, Elaine Y. Chung, Jessica Perazzelli, Ted J. Hofmann, Shannon L. Maude, Pichai Raman, Alejandro Barrera, Saar Gill, Simon F. Lacey, Jan J. Melenhorst, David Allman,

Elad Jacoby, Terry Fry, Crystal Mackall, Yoseph Barash, Kristen W. Lynch, John M. Maris, Stephan A. Grupp, and Andrei Thomas-Tikhonenko. Convergence of Acquired Mutations and Alternative Splicing of CD19 Enables Resistance to CART-19 Immunotherapy. *Cancer discovery*, 5(12):1282–1295, December 2015. ISSN 2159-8274. doi: 10.1158/2159-8290.CD-15-1020. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4670800/.

Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, November 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0150-2. URL https://www.nature.com/articles/s41576-019-0150-2. Number: 11 Publisher: Nature Publishing Group.

Peter D. Stenson, Edward V. Ball, Matthew Mort, Andrew D. Phillips, Jacqueline A. Shiel, Nick S. T. Thomas, Shaun Abeysinghe, Michael Krawczak, and David N. Cooper. Human Gene Mutation Database (HGMD): 2003 update. *Human Mutation*, 21(6):577–581, June 2003. ISSN 1098-1004. doi: 10.1002/humu.10212.

Timothy Sterne-Weiler, Robert J. Weatheritt, Andrew J. Best, Kevin C. H. Ha, and Benjamin J. Blencowe. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell*, 72(1):187–200.e6, October 2018. ISSN 1097-2765. doi: 10.1016/j.molcel.2018.08.018. URL https://www.cell.com/molecular-cell/abstract/S1097-2765(18)30678-6. Publisher: Elsevier.

Jenny C Taylor, Hilary C Martin, Stefano Lise, John Broxholme, Jean-Baptiste Cazier, Andy Rimmer, Alexander Kanapin, Gerton Lunter, Simon Fiddy, Chris Allan, A. Radu Aricescu, Moustafa Attar, Christian Babbs, Jennifer Becq, David Beeson, Celeste Bento, Patricia Bignell, Edward Blair, Veronica J Buckle, Katherine Bull, Ondrej Cais, Holger Cario, Helen Chapel, Richard R Copley, Richard Cornall, Jude Craft, Karin Dahan, Emma E Davenport, Calliope Dendrou, Olivier Devuyst, Aimée L Fenwick, Jonathan Flint, Lars Fugger, Rodney D Gilbert, Anne Goriely, Angie Green, Ingo H. Greger, Russell Grocock, Anja V Gruszczyk, Robert Hastings, Edouard Hatton, Doug Higgs, Adrian Hill, Chris Holmes, Malcolm Howard, Linda Hughes, Peter Humburg, David Johnson, Fredrik Karpe, Zoya Kingsbury, Usha Kini, Julian C Knight, Jonathan Krohn, Sarah Lamble, Craig Langman, Lorne Lonie, Joshua Luck, Davis McCarthy, Simon J McGowan, Mary Frances McMullin, Kerry A Miller, Lisa Murray, Andrea H Németh, M Andrew Nesbit, David Nutt, Elizabeth Ormondroyd, Annette Bang Oturai, Alistair Pagnamenta, Smita Y Patel, Melanie Percy, Nayia Petousi, Paolo Piazza, Sian E Piret, Guadalupe Polanco-Echeverry, Niko Popitsch, Fiona Powrie, Chris Pugh, Lynn Quek, Peter A Robbins, Kathryn Robson, Alexandra Russo, Natasha Sahgal, Pauline A van Schouwenburg, Anna Schuh, Earl Silverman, Alison Simmons, Per Soelberg Sørensen, Elizabeth Sweeney, John Taylor, Rajesh V Thakker, Ian Tomlinson, Amy Trebes, Stephen RF Twigg, Holm H Uhlig, Paresh Vyas, Tim Vyse, Steven A Wall, Hugh Watkins, Michael P Whyte, Lorna Witty, Ben Wright, Chris Yau, David Buck, Sean Humphray, Peter J Ratcliffe, John I Bell, Andrew OM Wilkie, David Bentley, Peter Donnelly, and Gilean McVean. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature genetics*, 47(7):717–726, July 2015. ISSN 1061-4036. doi: 10.1038/ng.3304. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4601524/.

Juan L. Trincado, Juan C. Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J. Elliott, and Eduardo Eyras. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40, March 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1417-1. URL https://doi.org/10.1186/s13059-018-1417-1.

Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R. Gazzara, Juan González-Vallinas, Nicholas F. Lahens, John B. Hogenesch, Kristen W. Lynch, and Yoseph Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:e11752, February 2016. ISSN 2050-084X. doi: 10.7554/eLife.11752.

Jorge Vaquero-Garcia, Scott Norton, and Yoseph Barash. LeafCutter vs. MAJIQ and comparing software in the fast moving field of genomics. *bioRxiv*, page 463927, November 2018. doi: 10.1101/463927. URL https://www.biorxiv.org/content/10.1101/463927v1.

Jorge Vaquero-Garcia, Joseph K. Aicher, Paul Jewell, Matthew R. Gazzara, Caleb M. Radens, Anupama Jha, Christopher J. Green, Scott S. Norton, Nicholas F. Lahens, Gregory R. Grant, and Yoseph Barash. RNA splicing analysis using heterogeneous and large RNA-seq datasets. Technical report, bioRxiv, November 2021. URL https://www.biorxiv.org/content/10.1101/2021.11.03.467086v2. Section: New Results Type: article.

Htoo A. Wai, Jenny Lord, Matthew Lyon, Adam Gunning, Hugh Kelly, Penelope Cibin, Eleanor G. Seaby, Kerry Spiers-Fitzgerald, Jed Lye, Sian Ellard, N. Simon Thomas, David J. Bunyan, Andrew G. L. Douglas, and Diana Baralle. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genetics in Medicine*, 22(6):1005–1014, June 2020. ISSN 1530-0366. doi: 10.1038/s41436-020-0766-9. URL https://www.nature.com/articles/s41436-020-0766-9. Number: 6 Publisher: Nature Publishing Group.

Guey-Shin Wang and Thomas A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews. Genetics*, 8(10):749–761, October 2007. ISSN 1471-0064. doi: 10.1038/nrg2164.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, and Brendan J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218):1254806, January 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1254806. URL http://science.sciencemag.org/content/347/6218/1254806.

Yaping Yang, Donna M. Muzny, Fan Xia, Zhiyv Niu, Richard Person, Yan Ding, Patricia Ward, Alicia Braxton, Min Wang, Christian Buhay, Narayanan Veeraraghavan, Alicia Hawes, Theodore Chiang, Magalie Leduc, Joke Beuten, Jing Zhang, Weimin He, Jennifer Scull, Alecia Willis, Megan Landsverk, William J. Craigen, Mir Reza Bekheirnia, Asbjorg Stray-Pedersen, Pengfei

Liu, Shu Wen, Wendy Alcaraz, Hong Cui, Magdalena Walkiewicz, Jeffrey Reid, Matthew Bainbridge, Ankita Patel, Eric Boerwinkle, Arthur L. Beaudet, James R. Lupski, Sharon E. Plon, Richard A. Gibbs, and Christine M. Eng. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312(18):1870–1879, November 2014. ISSN 1538-3598. doi: 10.1001/jama.2014.14601.

Vicente A. Yépez, Christian Mertes, Michaela F. Müller, Daniela Klaproth-Andrade, Leonhard Wachutka, Laure Frésard, Mirjana Gusic, Ines F. Scheller, Patricia F. Goldberg, Holger Prokisch, and Julien Gagneur. Detection of aberrant gene expression events in RNA sequencing data. *Nature Protocols*, 16(2):1276–1296, February 2021. ISSN 1750-2799. doi: 10.1038/s41596-020-00462-5. URL https://www.nature.com/articles/s41596-020-00462-5. Number: 2 Publisher: Nature Publishing Group.

Zijun Zhang, Zhicheng Pan, Yi Ying, Zhijie Xie, Samir Adhikari, John Phillips, Russ P. Carstens, Douglas L. Black, Yingnian Wu, and Yi Xing. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature Methods*, 16(4):307–310, 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0351-9.