Title

High-Throughput Experimental and Computational Techniques for Assessing Genetic Variants in Microexon Splicing

Author

Burghard, Christina

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

High-Throughput Experimental and Computational Techniques for Assessing Genetic Variants in Microexon Splicing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Christina Phyllis Burghard

2023

ABSTRACT OF THE DISSERTATION

High-Throughput Experimental and Computational Techniques for Assessing Genetic Variants in Microexon Splicing

by

Christina Phyllis Burghard

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2023

Professor Xinshu Xiao, Chair

The identification of causal genetic variants underlying human diseases and traits remains a major challenge in genomics. New progress towards this goal has been made possible by advancing high-throughput technologies in biology and the subsequent collection of large biological datasets. In my thesis work I use both large-scale experimental and computational approaches to identify genetic variants that alter the process of RNA splicing. I also identified a unique opportunity to utilize large-scale RNA-Seq data for variant discovery, particularly focusing on an understudied class of splicing events known as microexons. Microexons are extremely short exons that pose unique challenges for quantification. To address this, I developed an optimized computational pipeline described in Chapter 2. This pipeline was rigorously validated against long-read sequencing data to ensure its reliability. In Chapter 3, the functional importance of microexons is explored

through analysis of multi-tissue RNA-Seq data sourced from the GTEx consortium. My analysis identified thousands of highly conserved microexons. These microexons align well with recent evidence indicating that disruption in microexon splicing is associated with neurological and muscular diseases. Finally in chapter 4, I expand on work using a massively parallel reporter assay (MPRA) to directly test thousands of sequence variants for their effect on splicing. Through a data science-driven approach, this thesis advances our understanding of the intricate landscape of splice-disrupting variants, producing new tools as well as novel insights into the functional importance of microexons in human biology and disease.

The dissertation of Christina Phyllis Burghard is approved.

Sriram Kosuri

Xianghong J. Zhou

Daniel H. Geschwind

Douglas L. Black

Xinshu Xiao, Committee Chair

University of California, Los Angeles

2023

To my husband Neal.

TABLE OF CONTENTS

LIST OF FIGURES

4.1 (Previous page.) (A) Schematic representation of the experimental workflow: An in silico-designed intron-exon-intron library encompassing 9,634 human exons was generated, with each construct containing a variable region consisting of an exon and minimum of 40 bp of upstream and 30 bp of downstream intronic sequence. The splicing reporter constructs were stably integrated into a chromosomal landing pad via Bxb1 integrase to standardize RNA expression levels. Fluorescence-activated cell sorting (FACS) was employed to sort cells based on the reconstituted GFP signal, which serves as an indicator of splicing efficiency. Quantification was achieved by binning cells according to both RFP and GFP fluorescence levels, followed by normalization and calculation of the inclusion ratio based on the ratio of high-to-low GFP bins. (B) Localization of splice-disrupting variants within genes. The proportion of intronic SDVs was slightly higher than that of exonic SDVs. (C) Frequency distribution of SDVs by functional class. Splice site mutations disproportionately caused splicing disruption, but were not the most frequent class of SDVs of all detected. (D) Density distribution of SDVs along the normalized sequence construct length. (E) Scatter plot depicting the relationship between allele frequency and the likelihood of a variant being a splice-disrupting variant, with rare variants showing a higher propensity to be SDVs. . 126

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Professor Grace Xiao and Professor Sri Kosuri, for their guidance, mentorship, and extreme patience through the course of my graduate career. My experiences in both of these labs have shaped my professional and personal development over the last seven years. I want to thank my committee members, Professor Doug Black and Professor Dan Geschwind for their incredible input and enthusiasm over the course of this project. Their comments, guidance, and kindness are tremendously appreciated. Special thanks as well to Professor Jasmine Zhou for her kindness in offering her invaluable feedback on short notice.

I thank the many kind and talented individuals that it has been my privilege to work alongside of in both the Kosuri and Xiao Labs. I especially would like to recognize Rocky Cheung and Jae Hoon Bahn for helping me with the experimental portion of my project, which was very challenging for me, with so much kindness. I am in awe of their incredible talent and experience. I also want to thank Rocky and Kim Insigne for welcoming me onto the great work they had done developing the MFASS project, which inspired much of the work in this thesis.

I would also like to thank Professor Michael Jewett, who opened the door for me to a career in research as a young undergraduate with no experience and set me on the path I have been on for the last decade, and Professor Mitch Guttman for giving me the incredible opportunity to work in his lab and teaching me a lot of bioinformatics.

I'd like to thank M.C., who I have been thankful to know through the entirety of my

graduate career, since my first grad interview. It has been my privilege to grow together over these years and celebrate our biggest personal and professional milestones together. I also thank F.L. who has also been like another sister, who has shared with me the ups and downs of grad school and been a calming presence in my life. I also thank W.L. for the abundance of wisdom, kindness, and acceptance shared with me over the past several years.

Finally I would like to thank my long-time friends A.K., S.C., K.W.J., and A.R. for being there for me over so many years, for most of my favorite memories. Extra thanks to S.C. for being an excellent academic big brother. I thank my parents, Brian and Cindy, for their unconditional and selfless love, their many sacrifices, and for always making me feel like enough. I am also grateful to my incredible and talented siblings Jenny, Paul, Mark, and Kim for always being there for me, as well.

Lastly my biggest thanks to my husband Neal, who has truly lived up to the title of "my better half," without whom this PhD would never have gotten finished, and whose insatiable curiosity and enormous heart are the forces that keep me going- you're the best and I love you very much.

VITA

2016–2023 Ph.D. Candidate, Bioinformatics. University of California, Los Angeles.

2009–2013 BS in Computer Science, McCormick School of Engineering and Applied Math. Northwestern University, Evanston IL.

2016–2023 Bioinformatics Graduate Researcher. Co-Advisors: Dr. Xinshu Grace Xiao, UCLA. Dr. Sriram Kosuri, UCLA

2014-2015 Associate Computational Biologist in the lab of Dr. Mitchell Guttman, California Institute of Technology

2013–2014 Postbaccalaureate Research Fellow. Advisor: Dr. Kevin Briggman, National Institutes for Health

2010–2013 Undergraduate Researcher. Advisor: Dr. Michael Jewett, Northwestern University.

2011–2012 Undergraduate Researcher. Advisor: Dr. Douglas Downey, Northwestern University.

PUBLICATIONS

"Allele-Specific Alternative Splicing and Its Functional Genetic Variants in Human Tissues." Amoah, Kofi, Yun-Hua Esther Hsiao, Jae Hoon Bahn, Yiwei Sun, **Christina**

Burghard, Boon Xin Tan, Ei-Wen Yang, and Xinshu Xiao. 2021. Genome Research 31 (3):359-71.

"A Multiplexed Assay for Exon Recognition Reveals That an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions." Cheung, Rocky, Kimberly D. Insigne, David Yao, **Christina Phyllis Burghard**, Jeffrey Wang, Yun -Hua E. Hsiao, Eric M. Jones, Daniel B. Goodman, Xinshu Xiao, and Sriram Kosuri. 2019. Molecular Cell 73 (1): 183–94.e8.

"The Xist lncRNA Interacts Directly with SHARP to Silence Transcription through HDAC3." McHugh, Colleen A., Chun-Kan Chen, Amy Chow, Christine F. Surka, Christina Tran, Patrick McDonel, Amy Pandya-Jones, Mario Blanco, **Christina Burghard**, Annie Moradian, Michael Sweredoski, Alexander Shishkin, Julia Su, Eric S Lander, Sonja Hess, Kathrin Plath, Mitchell Guttman. 2015.Nature 521 (7551): 232–36.

CHAPTER 1

Background

1.1 Introduction

In 1977, researchers at MIT (Chow et al. (1977), Berget, Moore, and Sharp (1977)) were the first to publish the discovery of RNA splicing. They showed that, instead of being encoded in one continuous sequence, genes are found in DNA as multiple coding segments separated by long non-coding sequences. The full length of the gene is transcribed into RNA, and the non-coding stretches (now called introns) are removed, and the coding segments, or exons, are joined together to form the mature RNA sequence. This process of RNA splicing creates an additional layer of complexity in the processing of genetic information. As many as 95% of all human genes undergo alternative splicing (AS). Through the alternative inclusion or exclusion of individual exons, splicing allows cells to generate diverse proteins from a limited number of genes, enabling the fine-tuning protein function under different cellular conditions, and contributing to the diversity of life on Earth.

Since its discovery, researchers have worked to understand the biochemistry of RNA splicing and its outcomes. RNA splicing is a complex, multi-step reaction orchestrated by a

large macromolecular complex called the spliceosome, made of both protein and RNA components (see Wahl, Will, and Lührmann (2009) for review). Initially, the spliceosome assembles on the pre-mRNA molecule, recognizing specific sequences at the boundaries between exons and introns. These sequences, known as splice sites, consist of a donor site at the 5' end of the intron and an acceptor site at the 3' end. Additional sequence elements are recognized by the splicing machinery, or may be bound by additional protein or RNA interactors. These RNA binding events may cooperate with the spliceosome to enhance splicing or bind antagonistically and hinder the splicing reaction. The interplay of the precursor RNA with these varied components influences the final splicing outcome.

Despite this complexity, RNA splicing proceeds with remarkable fidelity. RNA sequences contain highly conserved elements that preserve fine-tuned splicing events. Researchers have made significant progress in identifying sequence motifs and their corresponding interactors that influence splicing site selection. Attempts to formalize this "splicing code" have been made possible by advances in high-throughput sequencing technologies and computational algorithms (Mortazavi et al. (2008)). Deciphering this code is crucial for predicting and understanding how specific exons are selected or skipped during alternative splicing, which has significant implications for gene regulation and protein diversity. Ultimately, unraveling the RNA splicing code will enhance our ability to accurately predict and manipulate splicing events, advancing our understanding of the molecular processes underpinning human health and disease.

Given that splicing can greatly change the protein sequence and therefore function, AS has been of interest to all areas of cell function. And it has been shown that AS disruptions contribute to many diseases and aspects of human health. To further motivate the work in

this thesis, we discuss here the current understanding of splicing and its contribution to disease. We introduce experimental and computational advances that have shed new light on these topics, and enabled the work contained in this thesis. Finally, we introduce key knowledge gaps in the field that our work addresses.

1.2 Genetic effects on splicing and disease

Determining the full extent of the influence of alternative splicing on protein function and its contribution to diseases is an open and active area of research. Changes to alternative splicing can be caused by genetic variants in or around the pre-mRNA sequence which change the function or availability of splicing factors or other components that participate in the splicing reaction (called trans-factors), or by variation that influences splicing machinery binding by changing the immediate sequence context (commonly referred to as cis-variation). Cis-variation may change the splicing patterns in one gene, but trans-splicing disruption may change the environment and affect splicing of many different genes. Both cis- and trans-splicing disruption can lead to disease states, and their identification poses distinct challenges.

1.2.1 Disruption of splicing trans-factors

The spliceosome is comprised of five snRNPs (small nuclear ribonucleoproteins) and over 150 associated protein factors that come together to catalyze intron excision (Black (2003)). Expression differences in trans-acting splicing factors can lead to different splicing outcomes in different cellular contexts. Mutations disrupting the function of a splicing factor may impact whole networks of splicing events, and some disruptions to the core

splicing machinery are known to be lethal. Two diseases caused by mutations to splicing factors include spinal muscular atrophy (SMA) and retinitis pigmentosa (G.-S. Wang and T. A. Cooper (2007)). In SMA, the SMN1 gene is disrupted, which hinders the formation of snRNP complexes in motor neurons. In some inherited forms of retinitis pigmentosa, progressive retinal degeneration results from snRNP disruption due to mutations to pre-mRNA processing factor genes.

In addition to defects in splicing machinery, another important form of splicing disruption occurs when repeat expansions result in toxic RNA molecules. These expansions create binding sites for splicing factor RBPs, resulting in their sequestration to these transcripts and depletion in the cell. Diseases caused by this type of disruption include Huntington disease, spinocerebellar ataxias, and muscular dystrophies (Ranum and T. A. Cooper, 2006). In myotonic dystrophy type 1, a toxic expansion in the DMPK gene causes the splicing factors MBNL1 and CUGBP levels to be disrupted. Without proper levels of these splicing factors, cells develop splicing patterns similar to embryonic ones in muscle. Splicing patterns are also disrupted in brain and other organs, and individuals suffering from the disease experience a range of symptoms affecting neurological function, motor control, and insulin resistance (Poulos et al. (2011),Ranum and T. A. Cooper (2006)).

While not discussed in detail here, disruptions to splicing are also commonly found in cancer (David et al. (2010) and Venables (2004)) Changes in splicing factors may produce global changes in splicing patterns that assist cancer cells in surviving and expanding (Scotti and Swanson (2016)). In a group of bone marrow cancers called myelodyplastic syndromes (MDS), mutations of a component of the U2 snRNP named SF3B1 is known to produce splicing dysregulation (Scotti and Swanson (2016) and Dolatshad et al. (2015)).

PRPF6, a U5 snRNP is overexpressed in colorectal carcinoma and contributes to the cancer cell's ability to proliferate. In addition, hypoxia characteristic of tumors can create global splicing changes. Targeting the spliceosome has been a recent area of advancement in the treatment of cancers (Bonnal, Vigevani, and Valcárcel (2012) and Hsu et al. (2015)).

1.2.2 Cis-Variation affecting splicing

RNA splicing recognition depends on the main sequence determinants, including both the 5' and 3' splice sites and the branch point/polypirimidine tracts of exons. Splice site choice is determined by the positional assembly of the early spliceosome complex that defines the boundaries of the excised intron. This occurs mainly through base pairing of snRNA (Small nuclear RNA) with the precursor mRNA. Two different spliceosome assemblies exist in humans: the U2-dependent spliceosome, also called the major spliceosome; and the minor U12-dependent spliceosome. The major and minor spliceosomes recognize GT-AG and AT-AC splice sites respectively. Spliceosome recognition of these consensus sequence elements is a critical step for correct splicing, however, these motifs exhibit substantial variation and are not sufficient on their own to dictate splicing. A second layer of control comes from sequence motifs, which may be recognized by varied RNA binding proteins or otherwise affect spliceosome interaction. These sequence elements are referred to as intronic or exonic splicing enhancers/silencers (abbreviated ISE/ISS, ESE/ESS). Generally, SR-proteins and SR-domain containing proteins will bind to ESEs, encouraging exon inclusion, and hnRNPs recognize ESS and ISSs to promote exon skipping. A multitude of SR and hnRNP proteins will bind to many short elements for a single splicing event (Black (2003)) These are not the only RBPs that can influence splicing, and cell- or tissue-specific proteins or isoforms can enable unique splicing patterns in different contexts. For reviews

covering the mechanics of splicing in more detail, see Kelemen et al. (2013), Q. Li, J.-A. Lee, and Black (2007), and Matlin, Clark, and C. W. J. Smith (2005).

Changes to cis-regulatory RNA sequences can affect splicing outcomes. Accurate prediction of variants on splicing remains a key active area of splicing research. DNA variants, including both larger copy number variants and single nucleotide variants (SNVs) can change the RNA properties and therefore splicing. RNA modifications such as RNA editing can also change the sequence in ways that affect splicing (Hsiao et al. (2018)). Changes to the invariant positions of the splice acceptor and splice donor are easily detected, and may result in complete loss of exon recognition. However, the effects of splice region variants are less easy to predict. Alternatively spliced exons tend to have weaker splice sites that less strongly match the consensus. Furthermore, it is possible for variants outside of the splice sites to have effects on splicing behaviors by disrupting the binding of other RNA-binding proteins or RNAs, or by changing the secondary structure of the RNA.

A classic illustration of pathogenic cis-variation is the study of cystic fibrosis variants (Garcia-Blanco, Baraniak, and Lasda, 2004). Patients with the most severe forms of this disease often are homozygous for strong loss-of-function splice variants, but less severe presentations can result from variants of smaller effect size, including a known 3' splice site variant which increases exon 9 skipping in the CFTR gene. Another important example is that of Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency (Nielsen et al. (2007), Anna and Monika (2018)). A pathogenic SNV in an ESE in exon 5 of the MCAD gene causes exon skipping, however, a different silent mutation in this exon was shown to inactivate an upstream ESS and restore exon inclusion in minigene assays. This highlights the complex and combinatorial nature of the splicing code. While most spicing signal is

expected to be contained close to the splice junction, recent evidence of deep intronic variants affecting splicing has also come to light. Deep intronic variants more than 100bp from the nearest exon which contribute to disease have been identified in inherited retinal diseases, ataxia and spastic paraplegia, and in breast/ovarian cancer (Verdura et al. (2020), Moles-Fernández et al. (2021), Sangermano et al. (2019)).

1.2.3 Additional layers of splicing regulation

Beyond splicing factor mutation and changes to their binding sites, other factors are also known to affect splicing, adding to the challenge of predicting splicing behaviors. RNA secondary structure may play an underappreciated role in splicing. For example, RNA structure can "mask" binding sites by forming double-stranded RNA structures (Krüttner et al. (2012), Buratti and F. E. Baralle (2004), Hiller et al. (2007), and Cheah et al. (2007)). Also, splicing can occur co-transcriptionally and is partly dependent on the rate of RNA polymerase. In choosing splice sites, the availability of downstream splice sites that may or may not have been transcribed yet can affect splice site choice. Distal splice sites can also affect splice site choice as splicing can occur recursively. Furthermore, growing evidence supports roles for epigenetic regulation of alternative splicing. DNA methylation may in some cases precludes splice factor binding, and additionally histone modifications and lncRNA binding may also change splicing patterns (X. Zhang et al. (2020)). The sum of effects of all of these changes to the RNA and the cellular environment are challenging to account for in splicing prediction.

1.3 Advances in interrogating splicing variation

It has been the work of over four decades of molecular biologists to uncover a growing repertoire of splicing factors, mechanistic insights, and disease-relevant splicing events. These foundations, along with next-generation sequencing and other technologies, make possible new approaches to generalize and predict splicing behaviors. We look now at improved high-throughput techniques for profiling splicing, as well as bioinformatics approaches for addressing both the resulting technical challenges and subsequent influx of large datasets. These advances allow us to look at splicing patterns across the genome in different contexts, as well as interrogate the effects of sequence variation directly to identify causal variants and their mode of action in disease.

1.3.1 High throughput experimental techniques

While detailed probing of specific genes or splicing factors of interest has been and remains indispensable for understanding the mechanism of splicing and causes of disease, complementary pictures of global splicing patterns have been made possible by the development of high-throughput experimental techniques.

Early high-throughput splicing assays used microarrays to investigate collections of events. Commonly, microarray probes would be designed to target splice junctions corresponding to the included or skipped sequences for an exon. Ratios of RNA hybridization levels measured by fluorescent probes could be used to estimate the abundance of specific splicing isoforms. While these approaches have been insightful, RNA-seq has emerged as the preferred method due to its ability to detect more and novel splicing events, and provide more quantitative measurements (Mortazavi et al. (2008)). RNA-Seq has been

tremendously helpful in elucidating novel isoforms across tissue types and conditions. Long read sequencing and single-cell sequencing technologies are further enabling a more comprehensive and detailed understanding of RNA splicing dynamics. Long read sequencing platforms, such as PacBio and Oxford Nanopore (Rhoads and Au (2015), Lu, Giordano, and Ning (2016), and Jain et al. (2016)) provide the ability to capture entire transcripts. This allows for the direct observation and characterization of full-length alternative splicing isoforms, including novel or rare isoforms that were previously challenging to detect.

Single-cell sequencing techniques, such as single-cell RNA sequencing (scRNA-seq), have unveiled the heterogeneity and cell-to-cell variability in splicing patterns within complex tissues and cell populations (Huang and Sanguinetti, 2017). By profiling individual cells, scRNA-seq enables the identification of cell-specific splicing events, uncovering rare cell populations or states that might exhibit distinct splicing programs. These advancements have significantly advanced our understanding of splicing in developmental processes, cell differentiation, and disease mechanisms.

Combining DNA sequencing and synthesis technologies has enabled innovative methods to test thousands to millions of hypotheses at once. Patwardhan et al., 2012 was the first to develop a massively parallel reporter assay (MPRA), combining custom microarray oligos with RNA-Seq to test single nucleotide variants' effects on promoter activity. Subsequently, this experimental design has been adapted to test sequence-function relationships for a variety of biological problems. Generally, MPRAs start with a large library of sequences which are cleaved and amplified off a custom microarray chip, inserted into a reporter construct, and then functionally assayed *in vivo* or *in vitro*. Some assay designs may

include a unique barcode sequence added to the library member which can be read out and used in quantification.

Several MPRAs have been designed to study splicing. Soemedi et al., 2017 designed an MPRA of transiently transfected minigenes with 200bp of variable sequence between a common upstream and downstream exon in hek293. They used this system to test the effects of splicing of 5,000 disease-causing exonic mutations. They found >10% of these variants caused significant changes in exon inclusion in this assay, suggesting that splicing is a frequent mode of action for pathogenic variants. In Gupta et al., 2019, two million distinct minigenes were generated using random 25-nt sequences at both 3' and 5' locations in a single intron reporter. By leveraging the extreme size of their library, they were able to infer relationships between the sequence elements and splicing outcomes, and were able to use the data to train a machine learning model to classify variants in three known disease genes with accuracy close to 90%. Cheung et al. (2017) used a stably transfected three-exon minigene reporter construct to test variants in >2,700 human exon sequences.

While splicing MPRAs have yielded interesting insights, more work is needed to understand their limits. One limitation of these MPRAs is that they are limited to reporting splicing in the engineered cell type chosen. These and other minigene systems lack the additional sequence context found natively; distal sequence elements outside the length that can be included in the minigenes used can influence splicing. Another limitation in some splicing MPRAs is that microarray DNA synthesis is currently limited to 200-300 nt sequences, so direct sequencing of the spliced products will not be able to identify variants in the excised introns. Some systems trade additional complexity by implementing a barcoding strategy, or stably integrating library members to overcome this

(Cheung et al., 2017). Additionally, different systems' splicing quantification may introduce biases as well, for example PCR drift in RNA-Seq methods or averaging in fluorescent cell sorting counts. Given these limitations, comparison of MPRA readouts to matched multi-tissue RNA- and DNA-Seq datasets could yield useful insights into commonalities and differences in splicing in these systems versus exons' endogenous contexts.

1.3.2 Bioinformatics approaches

The influx of large datasets has necessitated novel bioinformatics approaches to accurately quantify and interpret them. Complementary advances in computational approaches allow us to fully utilize the data, formulating new hypotheses and uncovering overarching biological patterns.

Splicing Quantification

Bioinformatics plays a crucial role in addressing biases and applying statistical methods to RNA-seq data for splicing analyses. Several challenges, such as technical biases, library preparation methods, and sequencing depth, can introduce systematic errors and biases that need to be accounted for. Bioinformaticians develop specialized algorithms and pipelines to preprocess RNA-seq data, including read alignment, transcript abundance estimation, and detection of splicing events.

Splicing quantification methods must be chosen carefully. From short read alignments, it's common to take the ratio of spliced to unspliced supporting reads for a given exon. "Percent spliced in", or PSI, is a common metric. However, this doesn't capture complex splicing patterns, and getting transcript information from short reads is a non-trivial problem. Methods like LeafCutter (Y. I. Li, Knowles, et al., 2018) use an intron-centric

approach to handle more complex splicing patterns statistically, fitting a Dirichlet multinomial distribution to count matrices to identify differential splicing between two groups of samples. Tools like edgeR, DESeq2, and rMATS (Mehmood et al. (2020)), are commonly used to perform differential splicing tests, accounting for factors such as technical variability, sample size, and multiple hypothesis testing.

Detecting Causal Splice Variants

Statistical association studies have been used to identify genetic variants correlated alternative splicing events. These techniques have greatly benefited from the availability of large datasets and have been used successfully to identify functional variation that acts through splicing. Splicing quantitative trait loci (sQTL) analysis focuses on identifying genetic variants that are associated with changes in splicing patterns (Monlong et al. (2014), Garrido-Martın et al. (2021)). By examining the correlation between genetic variants and quantitative measurements of splicing events, sQTL studies aim to uncover regulatory elements and genetic factors that influence alternative splicing. These findings contribute to our understanding of the genetic basis of splicing variation and its implications for phenotypic diversity and disease susceptibility.

Allele-specific alternative splicing (ASAS) analysis focuses on studying the preferential splicing of different alleles in heterozygous individuals (G. Li et al., 2012). This approach allows for the identification of genetic variants that directly influence alternative splicing outcomes. By detecting uneven ratios of alleles associated with splice events, ASAS studies can pinpoint specific variants that alter splice site recognition or regulatory elements.

Splicing variant association studies have also played a crucial role in identifying splicing

variants in rare Mendelian diseases. Large-scale sequencing efforts combined with advanced bioinformatics analyses have enabled the identification of disease-causing variants that disrupt splicing regulation. These studies have revealed splice-altering variants in both exonic and deep intronic regions, expanding our understanding of the genetic mechanisms underlying rare diseases. A notable example is the work by Cummings et al., 2017, which demonstrated the impact of splicing variant association studies in the diagnosis of rare variants. By targeting both exonic and deep intronic regions, they achieved a diagnosis rate of 35% for patients with rare genetic diseases. This highlights the potential of splicing variant association studies in improving the identification and understanding of disease-causing variants.

Splicing Prediction Algorithms

The computational prediction of splicing outcomes from sequence information has been under study for several decades. Position weight matrices created from known 5' and 3' splice sites confirmed observed sequence patterns, but could not be used on their own to establish a cutoff to distinguish between active and inactive splice sites (Jian, Boerwinkle, and Liu, 2014). To predict splice site usage, Yeo and Burge use the maximum entropy distribution (MED) to model splice site sequences. Unlike PWMs, the MED does not assume that positions are independent, and can incorporate pairwise and higher-order dependencies between positions. Many subsequent attempts to computationally predict splicing have incorporated additional sequence elements or additional features such as conservation that affect splicing (reviewed in Jian, Boerwinkle, and Liu (2014), Rowlands, D. Baralle, and Ellingford (2019)). Due to the lack of well-defined splicing rules and the complexity of splicing decisions, machine learning methods have emerged as a popular

approach to splicing prediction.

Initial machine learning approaches to splicing prediction tended to use supervised learning algorithms, such as Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Decision Trees, to classify splicing outcomes based on sequence features such as splice site motifs, exon-intron boundary sequences, and RNA secondary structure information (Buratti and F. E. Baralle (2004)). The models were trained on annotated datasets of known splicing events to learn the relationships between these features and splicing outcomes. Feature selection and dimensionality reduction techniques were often employed to enhance model performance and reduce computational complexity. These machine learning approaches allowed for the prediction of splicing patterns and identification of potential regulatory elements involved in splicing, but their accuracy was limited by the availability of training data and the complexity of splicing regulation.

Deep learning approaches have also been successfully applied to predict RNA splicing events. These methods employ artificial neural networks with multiple layers to learn complex patterns and relationships in RNA sequence data. Different architectures have been proposed based on arguments towards their applicability to splicing decision making. CNNs excel at capturing local sequence motifs and features, while RNNs are well-suited for modeling sequential dependencies in splicing data. SpliceAI used dilated CNNs, which allowed the consideration of much larger input sequences. Interestingly, versions of the SpliceAI model trained on sequences up to 10K away outperformed models using only nearer sequence context (Jaganathan et al. (2019)). While this performance increase may come from biologically relevant features such as splice sites across longer introns, it is difficult to prove. Limited interpretability of these models and the biological relevance of

their descision making has been an ongoing topic of discussion.

Advances in machine learning continue to inspire new approaches to splicing prediction. Recently, language models (LMs) have seen significant advancements in natural language processing (NLP), and the emergence of large language models like GPT has been particularly exciting. These models, using transformer architectures with trillions of parameters, have demonstrated remarkable success in capturing human linguistic patterns. Attempts are now being made to see if the same approach can be effectively applied to the languages of nucleic acids and proteins. One such LM, called SpliceBERT (Chen et al. (2023)), was pre-trained on multi-species RNA and has exhibited the ability to relate splice sites across introns and rank known splicing-disrupting variants (SDVs) higher than non-SDVs in predicting splice disruption. The appeal of using language models for RNA splicing lies in their potential to capture the complex interplay of sequence elements that determine splicing outcomes. With their attention mechanisms and interpretable machine learning, the features used by these models could be more usefully extracted and may lead to new insights on splicing regulation. However, challenges still remain. For example, these models have been known to "hallucinate," generating outputs following expected patterns but that may not be rooted in reality. Therefore, it remains to be seen whether language models can consistently produce accurate and biologically relevant predictions for RNA splicing.

1.4 Microexons

Improvements to computational and statistical analysis have enabled greater insight into splicing mechanisms, and better methodologies allows us to detect and study classes of

splicing events that were previously understudied or difficult to detect. We next turn our focus to one such group, called microexons, which are are considerably shorter than the average exon in humans. Microexons are interesting from a mechanistic perspective, as their length poses challenges to splicing. In addition, mounting evidence points to microexons having functional relevance in tissue development, brain function, and cancer (Ustianenko, Weyn-Vanhentenryck, and C. Zhang (2017a), J.-S. Lee, Lamarche-Vane, and Richard (2022), Parada et al. (2021), Gonatopoulos-Pournatzis and Blencowe (2020)). Microexon quantification may be unreliable using standard approaches for macroexons, so many microexons go undetected in conventional splicing analyses. Thus, their further study is warranted, and we review what is currently known about these exons.

1.4.1 Discovery of Microexons

Typically, exons in humans have a length averaging 137 nt (Berget (1995)). Assays of exon recognition shows a clear disadvantage for exons <50 nt to be included (Dominski and Kole (1991)). It was surprising to find functional exons considerably smaller than this. While known examples of microexons were reported as early as 1985, they were largely considered to be interesting outliers (Beachy, Helfand, and Hogness (1985), T. A. Cooper and Ordahl (1985), and McAllister et al. (1992)). However, in 2003, 223 human microexons were identified by Salzberg et al. by looking for ESTs with alignment gaps and evaluating the surrounding sequence for splicing signals, evidencing that microexons could be more common that was thought (Ustianenko, Weyn-Vanhentenryck, and C. Zhang (2017b)). Later, RNA-Seq studies would detect thousands of microexons in the human genome.

1.4.2 Microexon Quantification

Studies of microexons in RNA-Seq recognized the need for specialized alignment and quantification pipelines for microexons compared to longer conventional exons. This is because traditional short read aligners use seed sequences which often already exceed the size of microexons. One approach to overcome this, a method called oLego (Wu et al. (2013)), used shorter seed lengths to produce multiple alignments. The multiple seed matches are then clustered, and the most statistically likely location is selected (Wu et al. (2013)). Another approach, implemented in VAST-Tools, was to enumerate all possible microexon sequences from known introns, and then to perform ungapped alignment to the database of exon-microexon-exon junctions (Irimia et al. (2014) and Gohr et al. (2022)). Finally, MicroExonator is a recent method which uses a similar two-step approach, with additional filters on exon properties such as canonical U2 3' and 5' splice site strength (Parada et al. (2021)).

The number of microexons reported by these methods varies due to differing selection criteria. VAST-tools only considers microexons 3-15 nt in length, while Y. I. Li, Sanchez-Pulido, et al., 2015 studied exons 6-51nt, and exons 3–27 nt were reported by Irimia et al., 2014. Due to the tissue-specific nature of splicing, the number and type of samples used will also affect the number. Y. I. Li, Sanchez-Pulido, et al., 2015 reported as many as 13,000 microexons from several hundred RNA-Seq samples from major tissue types in humans. However, all three surveys of microexons found striking patterns of functional enrichment.

1.4.3 Microexons Are Highly Conserved and Functional

In Irimia et al. it was shown that brain-specific microexons are among most highly conserved exons in the genome, particularly ones 3-27nt long (Irimia et al. (2014), Barbosa et al. (2022), Merkin et al. (2012), Han et al. (2013)). Microexons also tend to be symmetric, having lengths of multiples of 3, which means they do not disrupt the reading frame. Microexons were also found to be more likely to be found in disordered regions of proteins than expected (Y. I. Li, Sanchez-Pulido, et al. (2015)). Neural microexons significantly overlap or are adjacent to modular domains that function in mediating protein and other ligand interactions (Irimia et al. (2014)). Taken together, these lines of evidence suggest widespread conserved functionality of microexons.

1.4.4 Microexon Splicing Regulation

The short size of microexons is expected to hinder recognition by the spliceosome in two ways. The first is that there is less room for exonic splicing motifs, and the second, is that the close proximity of the 5' and 3' ends of the exon leads to steric hinderance of the spliceosome. However, studies of microexons detected patterns that could compensate for these disadvantages. Li et al. showed shorter, possibly more efficiently defined introns, and found stronger 3' and 5' splice sites for constitutive microexons, while alternatively spliced microexons tended to be flanked by more conserved introns with high densities of splicing enhancers (Y. I. Li, Sanchez-Pulido, et al., 2015). Li et al. also observed high enrichment for Rbfox and PTBP1 binding motifs (Y. I. Li, Sanchez-Pulido, et al. (2015)). Interestingly, Rbfox1 is known to regulate neuronal excitation in mammalian brain, and lack of Rbfox1 leads to neuron hyperexcitation and seizures (Gehman et al. (2011)).

Other noted sequence elements include high CT elements that could bind PTBP, U2AF2, and other splicing factors (Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017b). SRRM4 binds to enhancers that are embedded within the unusually long polypyrimidine tracts present upstream of microexons, thereby compensating for the limited space available for enhancer sequences within the microexon (Sibley, Blazquez, and Ule (2016)).

1.4.5 Microexons in the Brain

Tissue-specific alternative splicing events in disordered regions frequently code for binding motifs affecting protein interactions, and have the capacity to remodel protein–protein interaction networks (Y. I. Li, Sanchez-Pulido, et al. (2015), Buljan et al. (2012), and Ellis et al. (2012)). Indeed, Irimia et al., 2014 proposed to remodel large protein interaction networks in neurons that are disrupted in autism. A surprising finding was that of all identified neural alternative splicing events, microexons made up 1/3 of all events conserved between human and mouse. Neuronal SR-related protein nSR100, also called SRRM4, was identified as a key regulator of these microexons. Over half of the detected microexons were affected in HEK293T kidney cell by overexpression of nSR100 (Irimia et al., 2014). Furthermore, disruption of SRRM4 levels is characteristic in autism, and the authors detected differential microexon splicing in a network of genes related to neurological processes and axonogenesis, and was enriched for known autism genes, including DTNA, ROBO1, SHANK2, and ANK2 (Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017b).

Additionally, transcriptional analyses revealed many microexons regulated by QKI within the Rho GTPase pathway (J. Lee et al., 2020). QKI may also be aberrantly expressed in schizophrenia, Alzheimer's disease (AD), and cancers– conditions in which microexons

splicing have also been implicated (Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017b,J. Lee et al., 2020,Scotti and Swanson, 2016). However, a link between QKI, microexon splicing, and these diseases has not been investigated.

1.4.6 Microexons in Disease

Individual microexons have been linked to schizophrenia (Ovadia and Shifman (2011)), epilepsy (Rusconi et al. (2015)), and L1 syndrome (Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017b). L1 syndrome refers to a series of diseases related to dysfunction of the L1CAM gene, which is important in neuron migration and development. Certain cases of these diseases have been found to be caused by microexon splicing disruption. The neuron-specific microexon 27 of L1CAM codes for a critical four amino acid sequence necessary for proper function of the axonal growth cone in certain neurons (Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017b). Microexon 2 of the same gene is also responsible for L1 protein interactions in nervous system development.

Microexons also play an important role in several different retinopathies (Ciampi et al. (2022)). Ciampi et al. demonstrated that Srrm3-dependent microexons are required for outer segment maintenance and vision. Misregulation of this splicing factor results in severely altered photoreceptors and vision loss similar to zebrafish models of retinopathies (Ciampi et al., 2022). In addition, Vig et al., 2020 discovered a splice variant introducing a stop codon into a retina-specific microexon causing vision loss. Interestingly, this was shown to produce a non-syndromic form of the diesease as the microexon is not present in other cell types. Another 13 nt microexon in Arl6 was found to be required for vision in mammals (Pretorius et al., 2011).

Finally, microexons have been shown to be disrupted in cancers as well. New evidence suggests that lowered microexon inclusion due to silencing of SRRM4 promotes tumor growth across cancers (Head et al., 2021). Microexon splicing was quantified using VAST-Tools in 6,908 samples from 9 cancer types in the Cancer Genome Atlas (TCGA). Surprisingly, it was shown that SRRM4 was the most consistently silenced splicing factor in tumors measured across tissue types. Though SRRM4 is more lowly expressed outside of brain, the authors observed further silencing through promoter methylation of this splicing factor across tissues. Higher levels of SRRM4 were associated with less proliferative, neural-like programs of differentiation, and the work suggests that SRRM4 may act as a "proliferation brake," that when deactivated promotes cancer growth. (Head et al., 2021). Additionally, alternative microexon splicing by Rbfox2 and PTBP1 has been shown to be associated with colorectal cancer metastasis (Mochizuki et al., 2021).

Overall, evidence points to microexons being an exciting and potentially understudied class of alternative splicing events. They are highly conserved and contain a high density of cis regulatory sequences. It is possible that their short length could facilitate the variable inclusion patterns exhibited during cell differentiation and in dynamic cell responses. Many examples of relevant microexons have been found in highly specific tissue and cell types, and can alter modular interaction domains in proteins. The work in this thesis seeks to address knowledge gaps in understanding the contribution of microexon splicing to cellular function and disease, and explores ways in which high-throughput experimental and computational methods can be combined to understand global patterns alternative splicing.

1.5 Knowledge Gaps and Study Aims

1.5.1 Microexon Splicing Quantification

Despite many new examples of functional microexons, an updated global analysis of microexon splicing in human tissues has not been done at scale. Paired RNA and DNA-Seq data from the Genotype-Tissue Expression (GTEx) Consortium would allow us to look at cross individual and cross tissue splicing. We proposed to quantify microexon splicing across over 17,000 samples in 50 different adult tissue types from the v8 release of GTEx. As discussed, microexon quantification requires methods distinct from macroexons. Scaling current methods from hundreds of samples to thousands of samples is a non-trivial task. It is important to evaluate and improve upon current quantification methods for larger datasets. This is the first aim of this work and is outlined in Chapter 2.

1.5.2 Microexon Splicing Across Adult Human Tissues

After development of a suitable computational strategy, additional insight to microexon splicing can be gained by comparing splicing across larger numbers of tissues and individuals. We perform analysis of microexon splicing in the v8 release of GTEx, to our knowledge the largest such study to date. As discussed, reported numbers of microexons vary greatly depending on selection criteria. We first identify robustly spliced microexons and patterns of alternative splicing across 50 adult tissue types. We identify genes containing likely functional microexons, enrichment of RBP sequence motifs, and conservation patterns of microexons in order to assess their contributions to tissue specificity and disease etiology. These and other questions are addressed in Chapter 3.

1.5.3 Contribution of Microexon Splicing to Disease

The number of individuals (n>800) in GTEx allows us to look for genetic variants associated with changes in microexon inclusion. We perform sQTL analysis using our calculated microexon PSIs and identify single- and multi-tissue microexon sQTLs. We perform colocalization with known GWAS variants to assess conditions in which microexon variation might be relevant. While most relevant microexon splicing events have been studied in neurological conditions and function, our data allows us to study microexons in other less studied tissues. Finally, given the overlap of known muscle-specific splicing factors disrupted in myotonic dystrophy type 1 and microexon splicing, including RBFOX and QKI, we asked whether microexon splicing disruption would be detectable in skeletal muscle biopsies from individuals with this disease, and whether this could contribute to disease etiology. We perform microexon quantification

1.5.4 Application of MPRA to Splicing Quantification Across Human Tissues

In the final chapter, we discuss work evaluating the applicability of MPRAs to study alternative splicing. MPRAs, and minigene-based assays in general, suffer from lack of context. Some MPRAs are transiently transfected and the effects of high and varying copy number are not known. splicing because of the abundance of splicing factors A direct comparison of RNA-Seq to MPRA readout would be informative in identifying MPRA use cases and limitations. Finally, even with high-throughput assays, the number of possible variants and the number of possible splicing events is on an order of magnitude that makes it impractical to test them all. Most variants are not expected to affect splicing, so prioritizing variants of interest is an goal. We identify splicing outliers from GTEx that are

compatible with existing splicing MPRAs, including those affecting microexons.

Mechanisms of Alternative Splicing

Title	Authors	Year	Notes
Mechanisms of Alternative Pre-Messenger RNA Splicing	Black	2003	
Mechanisms of Alternative Splicing Regulation:Insights from Molecular and Genomics Approaches	Chen, Manley	2009	Discussion of position-dependent SREs
RNA Structure and the Mechanism of Alternative Splicing	McManus, Gravely	2011	
A Day in the Life of the Spliceosome	Matera, Wang	2014	Human spliceosomal diseases
Alternative Splicing Regulatory Networks:Functions, Mechanisms, and Evolution	Ule, Blencowe	2019	
The Physiology of Alternative Splicing	Marasco, Kornblihtt	2023	Updated review of pathogenic splicing, discussion of splicing noise

Splicing and Disease

Alternative Splicing in Disease and Therapy	Garcia-Blanco, Baranaik, Lasda	2004	
RNA-Mediated Neuromuscular Disorders	Ranum, Cooper	2006	
Splicing in Disease: Disruption of the Splicing Code and the Decoding Machinery	Wang, Cooper	2007	Trans effects on splicing
Function of Alternative Splicing	Kelemen, Convertini et al.	2012	Table of cellular processes changed by alternative splicing
Hallmarks of Alternative Splicing in Cancer	Oltean, Bates	2013	

Roles and Mechanisms of Alternative Splicing in Cancer: Implications for Care	Bonnal, López-Oreja, Valcárcel	2020	
Splicing Regulation			
Alternative pre-mRNA splicing in neurons: growing up and extending its reach	Zheng, Black	2013	
The Neurogenetics of Alternative Splicing	Vuong, Black	2016	
Alternative Splicing During Mammalian Organ Development	Mazin, Khaitovich, Cardoso-Moreira, Kaessmann	2021	
The implications of alternative pre-mRNA splicing in cell signal Transduction	Choi, Cho, Kim	2023	
Splicing Prediction			
In silico tools for splicing defect prediction: a survey from the viewpoint of end users	Jian, Boerwinkle, Liu	2014	Review of pre-deep learning splice predictors, with review of tool comparisons
Machine Learning Approaches for the Prioritization of Genomic Variants Impacting pre-mRNA Splicing	Rowlands, Baralle, Ellingford	2019	Review of non-deep learning ML methods
A Survey on Deep Learning in DNA/RNA Motif Mining	He, Shen, Zhang, Wang, Huang	2020	
Splicing in the Diagnosis of Rare Disease: Advances and Challenges	Lord, Baralle	2021	

Table 1.1: Overview of relevant publications on splicing.

References

Anna, Abramowicz and Gos Monika (Aug. 2018). "Splicing mutations in human genetic disorders: examples, detection, and confirmation". en. In: *J. Appl. Genet.* 59.3, pp. 253–268.

Barbosa, Pedro et al. (Aug. 2022). "Clinical significance of genetic variation in hypertrophic cardiomyopathy: comparison of computational tools to prioritize missense variants". en. In: *Front Cardiovasc Med* 9, p. 975478.

Beachy, Philip A, Stephen L Helfand, and David S Hogness (1985). "Segmental distribution of bithorax complex proteins during Drosophila development". In: *Nature* 313.6003, pp. 545–551.

Berget, Susan M (1995). "Exon Recognition in Vertebrate Splicing (*)". In: *Journal of biological Chemistry* 270.6, pp. 2411–2414.

Berget, Susan M, Claire Moore, and Phillip A Sharp (Aug. 1977). "Spliced segments at the 5 terminus of adenovirus 2 late mRNA". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.8, pp. 3171–3175.

Black, Douglas L (2003). "Mechanisms of alternative pre-messenger RNA splicing". In: *Annual review of biochemistry* 72.1, pp. 291–336.

Bonnal, Sophie, Luisa Vigevani, and Juan Valcárcel (2012). "The spliceosome as a target of novel antitumour drugs". In: *Nature reviews Drug discovery* 11.11, pp. 847–859.

Buljan, Marija et al. (2012). "Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks". In: *Molecular cell* 46.6, pp. 871–883.

Buratti, Emanuele and Francisco E Baralle (2004). "Influence of RNA secondary structure on the pre-mRNA splicing process". In: *Molecular and cellular biology* 24.24, pp. 10505–10514.

Cheah, Ming T et al. (2007). "Control of alternative RNA splicing and gene expression by eukaryotic riboswitches". In: *Nature* 447.7143, pp. 497–500.

Chen, Ken et al. (May 2023). "Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction". en.

Cheung, Rocky et al. (2017). "Many rare genetic variants have unrecognized large-effect disruptions to exon recognition". In: *bioRxiv*, p. 199927.

Chow, L T et al. (Sept. 1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA". en. In: *Cell* 12.1, pp. 1–8.

Ciampi, Ludovica et al. (July 2022). "Specialization of the photoreceptor transcriptome by Srrm3-dependent microexons is required for outer segment maintenance and vision". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 119.29, e2117090119.

Cooper, Thomas A and CP Ordahl (1985). "A single cardiac troponin T gene generates embryonic and adult isoforms via developmentally regulated alternate splicing." In: *Journal of Biological Chemistry* 260.20, pp. 11140–11148.

Cummings, Beryl B et al. (2017). "Improving genetic diagnosis in Mendelian disease with transcriptome sequencing". In: *Science translational medicine* 9.386, eaal5209.

David, Charles J et al. (2010). "HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer". In: *Nature* 463.7279, pp. 364–368.

Dolatshad, Hamid et al. (2015). "Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells". In: *Leukemia* 29.5, pp. 1092–1103.

Dominski, ZBIGNIEW and RYSZARD Kole (1991). "Selection of splice sites in pre-mRNAs with short internal exons". In: *Molecular and cellular biology* 11.12, pp. 6075–6083.

Ellis, Jonathan D et al. (2012). "Tissue-specific alternative splicing remodels protein-protein interaction networks". In: *Molecular cell* 46.6, pp. 884–892.

Garcia-Blanco, Mariano A, Andrew P Baraniak, and Erika L Lasda (May 2004). "Alternative splicing in disease and therapy". en. In: *Nat. Biotechnol.* 22.5, pp. 535–546.

Garrido-Martın, Diego et al. (Feb. 2021). "Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome". en. In: *Nat. Commun.* 12.1, p. 727.

Gehman, Lauren T et al. (2011). "The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain". In: *Nature genetics* 43.7, pp. 706–711.

Gohr, André et al. (2022). "Computational Analysis of Alternative Splicing Using VAST-TOOLS and the VastDB Framework". In: *Alternative Splicing: Methods and Protocols*. Springer, pp. 97–128.

Gonatopoulos-Pournatzis, Thomas and Benjamin J Blencowe (Dec. 2020). "Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder". en. In: *Curr. Opin. Genet. Dev.* 65, pp. 22–33.

Gupta, Abhishek K et al. (2019). "Degenerate minigene library analysis enables identification of altered branch point utilization by mutant splicing factor 3B1 (SF3B1)". In: *Nucleic acids research* 47.2, pp. 970–980.

Han, Hong et al. (June 2013). "MBNL proteins repress ES-cell-specific alternative splicing and reprogramming". en. In: *Nature* 498.7453, pp. 241–245.

Head, Sarah A et al. (Feb. 2021). "Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers". en. In: *PLoS Biol.* 19.2, e3001138.

Hiller, Michael et al. (2007). “Pre-mRNA secondary structures influence exon recognition”. In: *PLoS genetics* 3.11, e204.

Hsiao, Yun-Hua Esther et al. (2018). “RNA editing in nascent RNA affects pre-mRNA splicing”. In: *Genome research* 28.6, pp. 812–823.

Hsu, Tiffany Y-T et al. (2015). “The spliceosome is a therapeutic vulnerability in MYC-driven cancer”. In: *Nature* 525.7569, pp. 384–388.

Huang, Yuanhua and Guido Sanguinetti (Apr. 2017). “Transcriptome-wide splicing quantification in single cells”. en.

Irimia, Manuel et al. (Dec. 2014). “A highly conserved program of neuronal microexons is misregulated in autistic brains”. en. In: *Cell* 159.7, pp. 1511–1523.

Jaganathan, Kishore et al. (Jan. 2019). “Predicting Splicing from Primary Sequence with Deep Learning”. en. In: *Cell* 176.3, 535–548.e24.

Jain, Miten et al. (2016). “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. In: *Genome biology* 17, pp. 1–11.

Jian, Xueqiu, Eric Boerwinkle, and Xiaoming Liu (Dec. 2014). “In silico prediction of splice-altering single nucleotide variants in the human genome”. en. In: *Nucleic Acids Res.* 42.22, pp. 13534–13544.

Kelemen, Olga et al. (Feb. 2013). “Function of alternative splicing”. en. In: *Gene* 514.1, pp. 1–30.

Krüttner, Sebastian et al. (2012). “Drosophila CPEB Orb2A mediates memory independent of Its RNA-binding domain”. In: *Neuron* 76.2, pp. 383–395.

Lee, Jee-San, Nathalie Lamarche-Vane, and Stéphane Richard (Mar. 2022). “Microexon alternative splicing of small GTPase regulators: Implication in central nervous system diseases”. en. In: *Wiley Interdiscip. Rev. RNA* 13.2, e1678.

Lee, Jeesan et al. (Dec. 2020). "QUAKING Regulates Microexon Alternative Splicing of the Rho GTPase Pathway and Controls Microglia Homeostasis". en. In: *Cell Rep.* 33.13, p. 108560.

Li, Gang et al. (July 2012). "Identification of allele-specific alternative mRNA processing via transcriptome sequencing". en. In: *Nucleic Acids Res.* 40.13, e104.

Li, Qin, Ji-Ann Lee, and Douglas L Black (Nov. 2007). "Neuronal regulation of alternative pre-mRNA splicing". en. In: *Nat. Rev. Neurosci.* 8.11, pp. 819–831.

Li, Yang I, David A Knowles, et al. (Jan. 2018). "Annotation-free quantification of RNA splicing using LeafCutter". en. In: *Nat. Genet.* 50.1, pp. 151–158.

Li, Yang I, Luis Sanchez-Pulido, et al. (Jan. 2015). "RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts". en. In: *Genome Res.* 25.1, pp. 1–13.

Lu, Hengyun, Francesca Giordano, and Zemin Ning (2016). "Oxford Nanopore MinION sequencing and genome assembly". In: *Genomics, proteomics & bioinformatics* 14.5, pp. 265–279.

Matlin, Arianne J, Francis Clark, and Christopher W J Smith (May 2005). "Understanding alternative splicing: towards a cellular code". en. In: *Nat. Rev. Mol. Cell Biol.* 6.5, pp. 386–398.

McAllister, Linda et al. (1992). "Alternative splicing of micro-exons creates multiple forms of the insect cell adhesion molecule fasciclin I". In: *Journal of Neuroscience* 12.3, pp. 895–905.

Mehmood, Arfa et al. (Dec. 2020). "Systematic evaluation of differential splicing tools for RNA-seq studies". en. In: *Brief. Bioinform.* 21.6, pp. 2052–2065.

Merkin, Jason et al. (Dec. 2012). "Evolutionary dynamics of gene and isoform regulation in Mammalian tissues". en. In: *Science* 338.6114, pp. 1593–1599.

Mochizuki, Yasushi et al. (Nov. 2021). "Alternative microexon splicing by RBFOX2 and PTBP1 is associated with metastasis in colorectal cancer". en. In: *Int. J. Cancer* 149.10, pp. 1787–1800.

Moles-Fernández, Alejandro et al. (July 2021). "Role of Splicing Regulatory Elements and In Silico Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes". en. In: *Cancers* 13.13.

Monlong, Jean et al. (Aug. 2014). "Identification of genetic variants associated with alternative splicing using sQTLseekeR". en. In: *Nat. Commun.* 5.1, p. 4698.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature methods* 5.7, pp. 621–628.

Nielsen, Karsten Bork et al. (Mar. 2007). "Seemingly neutral polymorphic variants may confer immunity to splicing-inactivating mutations: a synonymous SNP in exon 5 of MCAD protects from deleterious mutations in a flanking exonic splicing enhancer". en. In: *Am. J. Hum. Genet.* 80.3, pp. 416–432.

Ovadia, Galit and Sagiv Shifman (2011). "The genetic variation of RELN expression in schizophrenia and bipolar disorder". In: *PloS one* 6.5, e19955.

Parada, Guillermo E et al. (Jan. 2021). "MicroExonator enables systematic discovery and quantification of microexons across mouse embryonic development". en. In: *Genome Biol.* 22.1, p. 43.

Patwardhan, Rupali P et al. (2012). "Massively parallel functional dissection of mammalian enhancers in vivo". In: *Nature biotechnology* 30.3, pp. 265–270.

Poulos, Michael G et al. (Jan. 2011). "Developments in RNA splicing and disease". en. In: *Cold Spring Harb. Perspect. Biol.* 3.1, a000778.

Pretorius, Pamela R et al. (Apr. 2011). "Functional analysis of BBS3 A89V that results in non-syndromic retinal degeneration". en. In: *Hum. Mol. Genet.* 20.8, pp. 1625–1632.

Ranum, Laura P W and Thomas A Cooper (2006). "RNA-mediated neuromuscular disorders". en. In: *Annu. Rev. Neurosci.* 29, pp. 259–277.

Rhoads, Anthony and Kin Fai Au (2015). "PacBio sequencing and its applications". In: *Genomics, proteomics & bioinformatics* 13.5, pp. 278–289.

Rowlands, Charlie F, Diana Baralle, and Jamie M Ellingford (Nov. 2019). "Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing". en. In: *Cells* 8.12.

Rusconi, Francesco et al. (2015). "LSD1 neurospecific alternative splicing controls neuronal excitability in mouse models of epilepsy". In: *Cerebral cortex* 25.9, pp. 2729–2740.

Sangermano, Riccardo et al. (Aug. 2019). "Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides". en. In: *Genet. Med.* 21.8, pp. 1751–1760.

Scotti, Marina M and Maurice S Swanson (2016). "RNA mis-splicing in disease". In: *Nature Reviews Genetics* 17.1, pp. 19–32.

Sibley, Christopher R, Lorea Blazquez, and Jernej Ule (2016). "Lessons from non-canonical splicing". In: *Nature Reviews Genetics* 17.7, pp. 407–421.

Soemedi, Rachel et al. (Apr. 2017). "Pathogenic variants that alter protein code often disrupt splicing". en. In: *Nat. Genet.* 49.6, pp. 848–855.

Ustianenko, Dmytro, Sebastien M Weyn-Vanhentenryck, and Chaolin Zhang (July 2017a). "Microexons: discovery, regulation, and function". en. In: *Wiley Interdiscip. Rev. RNA* 8.4, e1418.

— (2017b). "Microexons: discovery, regulation, and function". In: *Wiley Interdisciplinary Reviews: RNA* 8.4, e1418.

Venables, Julian P (2004). "Aberrant and alternative splicing in cancer". In: *Cancer research* 64.21, pp. 7647–7654.

Verdura, Edgard et al. (Jan. 2020). "A deep intronic splice variant advises reexamination of presumably dominant SPG7 Cases". en. In: *Ann Clin Transl Neurol* 7.1, pp. 105–111.

Vig, Anjali et al. (Dec. 2020). "DYNC2H1 hypomorphic or retina-predominant variants cause nonsyndromic retinal degeneration". en. In: *Genet. Med.* 22.12, pp. 2041–2051.

Wahl, Markus C, Cindy L Will, and Reinhard Lührmann (2009). "The spliceosome: design principles of a dynamic RNP machine". In: *cell* 136.4, pp. 701–718.

Wang, Guey-Shin and Thomas A Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery". In: *Nature Reviews Genetics* 8.10, pp. 749–761.

Wu, Jie et al. (2013). "OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds". In: *Nucleic acids research* 41.10, pp. 5149–5163.

Zhang, Xuexiang et al. (Dec. 2020). "Unraveling the mechanobiology of immune cells". en. In: *Curr. Opin. Biotechnol.* 66, pp. 236–245.

CHAPTER 2

A Pipeline for Scalable and Accurate Microexon Discovery and Quantification

2.1 Introduction

Microexons (here defined as exons 27nt or shorter) are frequently highly conserved across species, and exhibit tissue-specific splicing patterns. While only a limited number of genome-wide microexon studies exist, evidence points to their importance to neurological and muscular development and function, autism spectrum disorders, and cancer. The short length of microexons makes their detection challenging. Typical short read aligners may miss microexons, primarily because they rely on ungapped alignment of seed sequences which are often longer than the microexon. This can lead to microexon sequences producing mis-alignments, failing to align, or being ignored as errors (T. D. Wu and Watanabe, 2005; J. Wu et al., 2013; Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017a; Yu et al., 2022). Thus, dedicated alignment strategies are required for efficient and accurate microexon quantification. Towards this end, several recent tools have been developed (J. Wu et al., 2013; Irimia et al., 2014; Parada et al., 2021). These tools generally take one of two main approaches to microexon alignment, either using shorter

seeds in a multiple seed-and-extend strategy, or by enumerating and mapping to sequence tags representing microexon-containing spliced sequences.

The splice-aware aligner OLego (J. Wu et al., 2013) employs a multiple seed-and-extend approach, similar to strategies used in aligners like MapSplice and later releases of TopHat (H. Li and Durbin, 2009; Wang et al., 2010). While designed as a general-purpose spliced read aligner, its improved ability to detect microexons was highlighted. OLego uses short (12-14nt) seeds to better locate splice junctions. Though seeds at this size produce multiple hits to the genome, OLego is able to cluster hits from multiple seeds to determine the statistically most likely source of the read. Once the most likely location has been determined, the algorithm considers which individual exons the read spans. Seed matches within each exon are extended to a splice junction, and a final spliced read alignment is constructed. If an unmappable insertion is encountered, the intronic region between the mapped regions is searched for a microexon sequence matching the insertion. The splice site strengths and resulting intron sizes of potential microexons are also considered and sequences with a low likelihood of being true splicing events are filtered. OLego demonstrated a significant increase in sensitivity over existing aligners of the time in detecting microexons of size 9-15nt (J. Wu et al., 2013).

As an alternative to the multiple seed-and-extend approach, some tools use transcriptome alignments using known splice junctions to identify microexons. One such method, called ATMap, used alignments to known splice junctions to find insertions signifying likely microexons. These sequences were checked against the genome for intron matches flanked by canonical splice sites. ATMap was tested by Y. I. Li et al. (2015) and performed comparably to OLego on synthetic data (Ustianenko, Weyn-Vanhentenryck, and C. Zhang,

2017b). Employing a similar strategy, *vast-tools* is another tool which has been widely used to quantify alternative splicing (AS) events (Irimia et al., 2014). For microexons, reads are mapped to sequence tags from a database of known exon-microexon-exon junctions. *vast-tools* then compares the counts of sequence tag matches to determine AS events between samples. *vast-tools* provides a valuable database of high-confidence AS events, however, the publicly available *vast-tools* only measures microexons of length 3-15nt which are already included in VASTDB (Tapial et al., 2017). To remedy this, the MicroExonator (Parada et al., 2021) pipeline was developed to allow for *de novo* microexon discovery as well as fast quantification of known microexons. Similar to ATMap, MicroExonator's discovery routine first performs a gapped alignment to known splice junctions, and the resulting alignments are checked for insertions. An insertion is considered a possible microexon if its sequence has a match in the intron sequence, is supported by sufficient reads, and has a high combined splice site strength for the 3' and 5' splice sites. By decoupling the more computationally intensive discovery step from the relatively faster quantification steps, MicroExonator is better able to handle large datasets compared to previous methods (Parada et al., 2021).

While differing in their approaches, these methods together have been used to uncover thousands of microexons in the human genome, as well as in mouse and fly, which are highly enriched for functional signals (reviewed in Ustianenko, Weyn-Vanhentenryck, and C. Zhang, 2017b; Lee, Lamarche-Vane, and Richard, 2022). For example, microexons are strongly enriched for lengths of multiples of three, a common feature of alternatively spliced exons, which preserves the reading frame. Additionally, microexons are often highly conserved across species, with some sequences conserved over 600 million years

(Torres-Méndez et al., 2019). One example of note is the neuron-specific CPEB4 exon 4. Increased skipping of this exon has been observed in autism (Irimia et al., 2014; Parras et al., 2018, and more recently in schizophrenia (Ollà et al., 2023.)

Beyond being conserved, microexon splicing also appears to be highly tissue-specific. Alternatively spliced microexons are most common in tissues with highly tissue-specific splicing patterns, including brain, skeletal muscle, heart, and testis (Y. I. Li et al., 2015). Alternatively spliced microexons had intron sequences which were enriched for sequence motifs which bind tissue-specific splicing factors (Y. I. Li et al., 2015; Irimia et al., 2014). A study by Irimia et al. (2014) applied *vast-tools* to find approximately 2500 neural-regulated AS events by comparing splicing across 50 tissue types, and from this they concluded that an identified set of 308 alternative microexons represented the most highly conserved component of developmental AS regulation as yet identified. Furthermore, many of these microexons were shown to be enriched in shared functional pathways. Subsequent gene ontology (GO) analysis showed that among these, the AS events predicted to generate alternative protein isoforms formed networks based on functions associated with neuronal biology, signaling pathways, structural components of the cytoskeleton, and the plasma membrane (Irimia et al., 2014). The authors proposed that microexons may operate in a switch-like manner, disrupting surface protein domains regulating protein-protein interactions in these pathways.

These studies may point to an underappreciated role of microexons in tissue-specific function, but despite this, uncertainty about the prevalence and function of microexons remains, largely due to the challenges of validation. True microexon splicing events are more difficult to distinguish from alignment artifacts than longer exons, and the number of

experimentally validated microexons is small compared to the numbers reported in bioinformatics analyses. Due to the lack of a suitable standard, comparison between methods has largely been on synthetic data. Based on the performance of popular splice-aware aligners such as TopHat (J. Wu et al., 2013) Hisat2, and STAR (Parada et al., 2021) on synthetic data, it seems likely that the full picture of microexon splicing is not captured in standard RNA-Seq analyses. Additionally, the reported number of microexons in real data varies due to differing conditions, selection criteria, and sample sizes, which make it difficult to directly compare methods. As microexons are highly tissue-specific, the limited number of tissue types for which microexon discovery has been performed may not include all microexons of functional importance.

With these issues in mind, we set out to more definitively measure the prevalence of microexons in the genome and their functional relevance. In order to capture the splicing variation of microexons between individuals across tissues, we sought to leverage over 17,000 RNA-Seq samples representing 50 tissue types from the GTEx consortium GTEx Consortium, 2020. Challenges we faced associated with data at this scale pointed to a need for an improved pipeline. Namely, we carefully consider sources of error in microexon detection, some of which are more apparent at scale when larger numbers of observed sequences yields large numbers of putative microexons. We identified sources of false positives and implemented additional filtering criteria at the individual alignment level as well as at the microexon level. Additionally, we implemented an improved computational workflow for for microexon detection with better scalability to sample numbers in the thousands. Our pipeline is implemented in the Workflow Description Language (WDL), which allows us to execute our analysis in a variety of contexts, e.g. local machines, remote

high-performance computing (HPC) systems, and in the cloud. Importantly, we validate our microexon discovery using long read data and report a high level of support for our microexon calls, including novel microexons. Our microexons detected in GTEx, as well as our microexon detection pipeline will be released publicly at the time of publication. These resources should pave the way for comprehensive investigations into the molecular mechanisms underlying tissue-specific regulation of microexons and a deeper understanding of their functional significance in biological processes and disease.

2.2 Methods

Overview of Microexon Detection Pipeline

The complete workflow used for microexon detection is outlined in Figure 2.1. This section will describe the capabilities of each section of the pipeline, in order of their appearance.

Pre-processing of RNA-Seq Reads

For input data, the pipeline can accept BAM files, SAM files, or unaligned FASTA/FASTQ files. Paired-end and single-end reads are both supported. If using inputs which have already been aligned to the genome, genome subtraction can optionally be performed. This step removes reads that were found to have an ungapped alignment with 1 or fewer mismatches to the genome. For these reads, a subsequent match to a microexon splice junction would be ambiguous, so we conservatively prioritize the genomic alignment and remove the read from consideration. Genome subtraction is recommended to substantially improve runtime. For paired-end data, we use PEAR to combine overlapping paired-end reads to avoid double counting during re-alignment (J. Zhang et al., 2014).

De novo Microexon Discovery

After pre-processing of input data, *de novo* microexon discovery may be performed, or, users can provide a previously generated set of microexon annotations and proceed directly to microexon quantification. We implement the *de novo* microexon discovery strategy outlined by Parada et al. in their method MicroExonator (Parada et al., 2021). This method expands on the approach developed in VAST-Tools to generate spliced sequence tags corresponding to all possible microexon splicing events from a given set of annotations. First, all splice junctions are retrieved from a user-provided GTF file of transcript annotations. Sequence tags for each splice junction are generated, consisting of 100nt upstream and downstream of the flanking exonic sequences (Parada et al., 2021). BWA-MEM is used to perform gapped alignment to the generated sequence tags (H. Li, 2013). From this alignment, insertions found at exon-exon junctions are reported. For each potential microexon insertion, the corresponding intron is scanned for the insertion sequence. An insertion is reported as a likely microexon if the sequence is present in the intron and flanked by canonical (AG-GT, optionally AG-GC) splice sites.

Combining Sequence Tag Lists

Microexonator concatenates the filtered splice junction insertion reads from each file in the run, and considers all the reads together to generate a combined list of putatitve microexons. Theoretically, this should help discover low-abundance microexons by combining reads across samples. However, we find that this is impractical for sample sizes of more than a few hundred. We opt to count the insertion reads in each sample separately, retaining any putative microexon with at least 5 supporting reads. Putative microexon sequence tags from all samples are combined. The read threshold can be changed based on

user requirements. The final sequence tag list generated in this step can be used for microexon quantification in subsequent samples without running the discovery portion of the pipeline.

Read Alignment to Sequence Tags

After generation of sequence tags, our workflow differs from existing methods in several key ways. First, ungapped alignment to sequence tags is performed jointly with re-alignment to the genome. In tests with STAR-aligned reads from GTEx (GTEx Consortium, 2020), we found that an ungapped alignment to genomic sequences was reported for up to 20% of reads after the initial genome subtraction step in pre-processing. This policy also aligns with bowtie documentation, which recommends genomic and sequence tag alignments be done jointly Langmead and Salzberg, 2012. In addition, it is computationally more efficient to build a joint index and align once, rather than perform two bowtie runs per sample and combine the results. With cloud-supported analysis of very large datasets in mind, we also reduced file I/O to improve efficiency for a more cost-effective workflow.

Alignment Selection and Prioritization

During bowtie alignment, we report all alignments that tie for the best alignment score using the "strata" option. *vast-tools* and MicroExonator both use the default bowtie settings with allowances for up to two mismatches. To correct for ambiguous microexons with near-matches to flanking exon sequences, we consider all alignments with the highest score, and then examine the types of sequences that aligned. We conservatively prioritize any alignments to genomic sequences, then to annotated splice junctions, and lastly report alignments to novel microexon sequence tags if no other equally likely match exists.

Tag Coordinate Conversion

After finalizing the alignments to the sequence tags, we optionally convert the tag alignments to genomic alignments based on the genomic positions of origin of the sequence tags. This is highly convenient for visualizing the re-alignments in a genome browser and allows them to be directly compared to other alignments. Additionally, other quantification tools or algorithms can be run on the genomic alignments that could not be applied using only the tag counts output by existing methods.

Calculating PSI and Splice Junction Usage

Finally, we calculate the Percent Spliced In (PSI) from the genomic alignments (Schafer et al. (2015)). This measurement has been used to summarize the efficiency of splicing of a specific exon. The score is calculated at the exon-level and is independent of the underlying distribution of the different transcripts which contain the exons. PSI is calculated by counting the number of inclusion reads (IR) from transcripts supporting the exon's inclusion, that is all reads which map to positions within the exon, and counting the number of exclusion reads (ER) from transcripts that support exon skipping- split reads that map to positions upstream or downstream of the exon but not the exon itself. The percent included is calculated as the IR total divided by the sum of IR and ER counts (Schafer et al. (2015)).

We note that PSI and exon-centric measurements of splicing have inherent limitations capturing transcript-level variance. Specifically, PSI does not capture all the relevant information, as it only reports the ratio of inclusion reads and exclusion reads for a given splice junction, and it does not provide information on how frequently that splice junction is used. This means a given microexon may have a high PSI, but may only be present in a

very small percentage of transcripts for that gene. We use samtools (Danecek et al. (2021)) to calculate the total number of gapped and ungapped read alignments at the center coordinate of the microexon, and estimate splice junction usage as the number of splice junction reads over this estimate of the gene coverage.

Splice Site Strength Scoring

Splice site strength is scored by calculating a combined 3' and 5' splice site strength score as in Parada et al. (2021). Splice sites are scored against the canonical splicing motif as defined by the U2-type intron position frequency matrices. This score is then normalized to a range from 1 to 100, and referred to as the U2 score.

Microexon Filtering

Once alignment is complete, we can filter on properties of the quantified microexon sequences. Filtering steps include 1) microexons which are supported by at least 10 reads in a single sample, 2) microexons present in a minimum of 75% of samples of at least one tissue, 3) microexon whose sequence does not match the flanking exon sequences, or any reported insertions given a set of genome calls in a VCF file. 4) we filter microexons to have a minimum estimated splice junction usage of 5%. We applied filters to select for the most widely used microexons that have robust inclusion. We require a minimum read coverage (default 10 reads) per microexon to ensure a reliable PSI, otherwise no PSI is reported for that sample. Additionally, we use our splice junction usage estimate to filter microexons which only occur in rare splice junctions. As previously mentioned, a high PSI can be reported if a microexon occurs in a rare splice junction, giving the appearance of a constitutive microexon. We also filter for microexons that reach the minimum read coverage in a minimum percentage of samples in at least one condition or tissue. These

filter parameters were chosen to survey the most widely used microexons in the human genome, but for use cases requiring higher sensitivity, they can be relaxed.

In addition to the microexon-level sequence checks, our method also summarizes alignment-level statistics for each microexon sequence. The average number of mismatches per read, and the average number of alignments per read are reported. Based on the distribution of average mismatches per read, we choose a cutoff of an average of 0.9 or fewer mismatches per sequence, as shown in Figure 2.4. Because mismatches are discrete, an average mismatch of less than 1 requires some fraction of mapping reads to be perfect. If a microexon annotation has an average mismatch greater than 1, it is more likely to be a misalignment, and thus should be filtered out. Similarly, we require an average number of alignments $\leq$1.9. This cutoff necessitates that for a given microexon, some fraction of reads are mapped uniquely, which filters out microexons with ambiguous or repeated sequences in the genome.

2.3 Results

2.3.1 Challenges in Microexon Detection and Alignment

One of the fundamental challenges in microexon detection is the high likelihood of short sequences occurring by chance within long, non-coding intronic regions. There are far more potential exon sequences than there are true exons, and the likelihood of observing a given microexon sequences gets progressively larger with decreasing size. Using sequence tag alignment-based approaches, the number of potential microexon sequences detected is highly influenced by the depth of coverage and the sample size. This observation is further

complicated by the potential for sequencing errors and spurious or rare splicing events to additionally increase sequence tag numbers. This may capture more splicing events, but we observe that it can also lead to misalignment. Due to the high number of reads and samples, misalignments could be present in high enough levels to exceed default thresholds. Additionally, microexon sequence tag generation requires as input a list of all annotated splice junctions to search for potential microexons. We observed that when using more comprehensive splice junction annotation sets and running on hundreds to thousands of RNA-Seq samples, the number of putative microexon sequences is significantly expanded. To illustrate this, from discovery on 100 samples of 50 tissues each, our combined putative sequence list was over 500,000 tags. This exceeds the number of known unique exons in the human genome (roughly 200,000), and of these, the expected number of microexons is only roughly 1-5% Y. I. Li et al., 2015. This excessive putative sequence count is due to many highly similar sequences representing many alternative 5' and 3' splice sites in the same gene, some of which may be rarely used.

The ambiguity that arises from closely spaced alternative 5' or 3' splice sites is an important source of error in microexon detection. The sequence between two alternative splice sites may only be a few bases, and may be found in the adjacent intron by chance. In this case the RNA sequences derived from alternative splice site usage and splicing of the intronic sequence as a casette microexon would be identical. Parada et al. shows that statistically, this can be a considerable source of spurious exons at smaller sizes, and include a filter to check microexons against flanking sequences (Parada et al., 2021). When filtering for intron matches, we found 833/5281 (15.8%) of our microexon sequences were indistinguishable from alternative splice site usage. Importantly, we observe that this

situation is further complicated by variants or sequencing errors in these flanking exon sequences, in which case exact matching will fail. Such an altered flanking sequence is equally likely to be found in the intron by chance, but will not be filtered out by exact matching. This becomes an important source of error when attempting to identify variants that affect microexon splicing. In our initial analysis of GTEx data, we identified many ambiguous matches of this type. These misalignments were identified by examining the read sequences, which were found to lack the microexon sequence they aligned with and better matched the flanking exon. Figure 2.2 shows an example misalignment from this data. Accounting for this required adjusting the tag alignment parameters.

For any two splice sites separated by a few base pairs, there is a chance the in-between sequence will appear in the intron by chance. For example, in the illustration in figure 2.2, alternative 3' splice sites A and B are 5nt apart. This flanking exonic sequence, plus the four base pairs needed to create the canonical splice sites, is statistically likely to occur in a long exon by chance, and will be spuriously tagged as a microexon. This problem is greatly exacerbated by the default allowance of 2 errors per sequence tag alignment, which MicroExonator (and many short read alignment applications) use to account for sequencing errors. From combined reads of 5 GTEx samples, we find that over 30% of reads aligned to more than one sequence tag when allowing for up to two mismatches. However, the majority of reads had only one top-scoring sequence when using the bowtie strata flag. This means a non-optimal alignment would be reported by MicroExonator at least half the time for these sequences, or up to 15% of all reads. Correcting for this by reporting all alignments reduces error and does not appreciably increase runtime.

Additional filters can be set to manage the number and stringency of microexon matches

reported on a user-defined basis. To focus on high-confidence microexons that were well captured in the tissues in this study, we applied filters to microexon coverage. This was defined as having at least 10 supporting reads in ≥75% of all samples in a tissue for at least one tissue Following PSI calculation, we found 17,765 microexons supported by at least 10 supporting splice junction reads in at least one sample, as shown in Figure 2.3. Our final list after filtering contained 3,343 microexons with sufficient coverage in at least one tissue. We next attempted to compare these microexons with previous findings to assess the accuracy of our pipeline.

2.3.2 Discovered Microexons are Enriched for Functional Signals

Microexons have been shown to be strongly enriched for lengths that are multiples of three, as shown in Figure 2.5A. These symmetric exons will not disrupt the reading frame if alternatively spliced. Some microexons, like those identified in Irimia et al., 2014, 2014 are highly conserved. Conservation is a common proxy for functionality as sequence element intolerant to random genetic changes are expected to be deleterious if disrupted. Figure 2.5B shows the conservation before and after filtering, which indicates that pipeline filtering accurately captures conserved microexons.

In Parada et al., a useful metric was proposed to measure the combined splice site strength score of the 5' and 3' splice regions based on the major spliceosome splice site motifs, referred to as the U2 splice-site score (Parada et al., 2021). Importantly, the distribution of this metric for candidate microexon sequences from MicroExonator resembles a bimodal distribution, as shown in Figure 2.5C, which is reproduced fromParada et al., 2021. This is expected to be due to one distribution of higher scores coming from real microexons, and

one distribution of lower scores coming from false positives. They use this metric to filter sequences by fitting a Gaussian mixture model, and keeping microexons which are assigned to the former group. Because there is substantial overlap between these two distributions, an optional "rescue" step can be performed to retain microexons showing high conservation but have lower splice-site scores. Alternatively spliced microexons are known to have lower splice site strengths than constitutively spliced exons, so this step is important to prevent omission of microexons of interest.

In Figure 2.5D, we apply this U2 score calculation from the MicroExonator discovery step onto all putative microexon sequences (pink), the filtered sequences from MicroExonator (dark blue), and the re-aligned and filtered microexons (light blue). The shift in U2 score between pink and dark blue indicates good filtering for conserved microexon sequences, but the dark blue curve is bimodal, similar to 2.5C. The additional filtering steps in our method produces a microexons with a much more unimodal distribution of U2 score, indicating that the pipeline has likely already filtered most spurious microexon sequences, eliminating the need for the Gaussian assumption used by Parada et al., 2021. We compare the distribution of U2 scores before and after applying our microexon-level filters and show a further decrease in frequency of lower-scoring microexons.

Through careful consideration of alignment errors, the improvements in read alignment and microexon filtering outlined here have a substantial effect on microexon detection, as evidenced by the distribution of splice-site strengths. Our method does not make any assumptions about splice-site strength or conservation to detect similar numbers of likely microexon sequences, while keeping false positives low.

2.3.3 Microexon Discovery Captures Known Microexon Events

Based on splice site strength distributions and the previous findings of Parada et al., 2021 on synthetic data, our results suggest that our method has high specificity for detecting microexon sequences from RNA-Seq. However, we also wanted to estimate sensitivity in detecting functional microexons of interest. While the complete landscape of microexon splicing in adult tissues is not known, 308 exceptionally conserved, neurally-regulated microexons were curated by Irimia et al., 2014, so we asked whether our method was able to detect these same microexons. Encouragingly, 302/308 microexons were found in our data, and 287 (93%) passed our high coverage and sample count filters. Our PSI calculations also reveal distinctive tissue-specific splicing patterns expected of these microexons, as shown in Figure 2.6. As previously reported, these microexons have highest inclusion in the cerebellum. We see distinct patterns across brain tissues and tissue-specific splicing in pituitary, heart, and muscle tissues as well. This indicates that we have good sensitivity on an important subset of tissue-specific microexons.

2.3.4 Long Read Sequencing Data Supports Detected Microexons

The alignment of short reads poses a challenge due to potential ambiguity from multiple alignments, a concern further exacerbated when analyzing spliced reads containing diverse combinations of RNA segments. In response to this challenge, we leveraged genome-aligned long read sequencing data to directly investigate microexons within splice junctions initially detected using short read data. The utilization of long reads spanning entire transcripts provided a higher level of confidence in identifying splicing events and allowed for the validation of transcript assemblies derived from short reads. The specific identification of

microexons from long read data represents a non-trivial alignment problem, which has been addressed by specialized long read aligners, such as IsoQuant. However, the confirmation of the presence of a particular microexon sequence within a given long read is straightforward. This involved the following steps: 1. For each aligned long read, we cataloged all sequence tags overlapping with the gene of origin. 2. Subsequently, we conducted a local alignment of each sequence tag to the long read sequence, thereby pinpointing microexon splice junctions. 3. We quantified the number of long reads that unambiguously contained the splice junction and matched the skipped or included microexon sequence tag. 4. These counts were then used to compute a long read Percent Spliced In (PSI) metric, which was subsequently compared to our short read microexon measurements.

This procedure was performed on raw sequencing reads from nine long read sequencing datasets from the ENCODE project, specifically from the dorsolateral prefrontal cortex. Within these datasets, we identified microexon sequences that were supported by at least three long reads in a given sample. The count of microexon sequences that could be tested varied across the samples and is detailed in Figure 2.7A. The majority of microexon sequences identified in long read data were present in at least two of the nine samples, as illustrated in Figure 2.7B.

We examined the distribution of U2 scores and conservation, which have been previously proposed as indicators of microexon functionality. Our analysis revealed that some microexons with low conservation were still observed in the long read exon data. Furthermore, the splice site strength score exhibited a wide range of values (Figure 2.7C,D). Stringent filtering based on these values could potentially lead to the exclusion of alternatively spliced microexons, which may be evolutionarily younger and

exhibit lower splice site strength due to their alternative splicing characteristics.

We compared the number of annotated microexons to the number of novel microexons that were validated through our approach. Additionally, we observed a notable increase in the number of microexons with a size less than 5 nucleotides (Figure 2.7E). Given that long read data frequently includes short insertions or deletions as the primary type of sequencing error, we conducted a thorough analysis to eliminate sequencing errors as potential sources of the observed microexons. Our findings suggested that these sequences could not be attributed to genomic variants, sequencing errors, proximity to splice sites, or other transcripts. As a result, they emerged as strong candidates for validation through wet lab experiments.

Finally, we compared the PSI values to the ratio of transcripts present in the long read data and observed a high degree of agreement (r=) (Figure 2.8). While we anticipated some discrepancies due to differences in tissue samples, brain tissue cell composition heterogenicity, coverage differences, and other factors, we observed a relatively small number of microexons that were included in the GTEx samples but not detected in the long read data. Thus, the long read data provides further support the microexons detected with our method in short read RNA-Seq.

2.4 Discussion

Due to their short length, alignment to microexon sequences requires a specialized approach. Existing approaches perform well when performed on ten to hundreds of samples, and produce lists of likely microexon splicing events that are highly enriched for

signals indicating functionality. However, when quantifying microexon splicing in a large (17K RNA-Seq samples) dataset, we identified several areas for improvement for better alignment accuracy and higher computational efficiency and ease of use when applying to large datasets. We identified several sources of error in the alignment process. Spurious microexons can be reported due to 1) close sequence matches in introns with flanking exon sequences, 2) DNA variants which produce a different alignment from the reference sequence, 3) highly similar sequence tags or genomic sequences, and 4) microexon sequences in rare splice junctions only occurring in a small proportion of transcripts. We designed an improved alignment process that considers all potential tag matches to pick the most appropriate alignment and accounts for these sources of error. We demonstrate using previously developed metrics, including conservation and splice site strength distributions, that our changes further enrich for likely functional microexons. We note that these improvements come from changes to alignment strategy and do not require any filters or assumptions about functional properties of microexons.

We successfully ran our workflow on over 17,000 RNA-Seq samples from GTEx to produce a global picture of microexon splicing in adult humans across individuals and across 50 tissue types, and we discuss these findings in the next chapter. There is a growing need for scalable and reproducible workflows in bioinformatics that can be applied to large datasets in a variety of environments. We chose to implement our workflow using WDL and Docker, which may be run on cloud, remote, and local environments using the Cromwell engine with minimal adjustments (Voss, Auwera, and Gentry, 2017). Due to our workflow's scalability and improved ability to control alignment errors in large datasets, our method is better suited to studying the effects of sequence variation on microexon splicing over

existing methods.

2.5 Accessibility of Data and Materials

WDL pipelines will be publicly available at `https://portal.firecloud.org/#methods`. Individual python scripts will be made available and accessed by the workflow through the Xiao Lab GitHub. `https://github.com/gxiaolab`. Accessed Docker images are stored in the Google Container Registry. Data analyses, and visualizations from this and the following chapter will be available as a Terra workspace. `https://app.terra.bio`.

References

Danecek, Petr et al. (Feb. 2021). "Twelve years of SAMtools and BCFtools". en. In: *Gigascience* 10.2.

GTEx Consortium (Sept. 2020). "The GTEx Consortium atlas of genetic regulatory effects across human tissues". en. In: *Science* 369.6509, pp. 1318–1330.

Irimia, Manuel et al. (Dec. 2014). "A highly conserved program of neuronal microexons is misregulated in autistic brains". en. In: *Cell* 159.7, pp. 1511–1523.

Langmead, Ben and Steven L Salzberg (Mar. 2012). "Fast gapped-read alignment with Bowtie 2". en. In: *Nat. Methods* 9.4, pp. 357–359.

Lee, Jee-San, Nathalie Lamarche-Vane, and Stéphane Richard (Mar. 2022). "Microexon alternative splicing of small GTPase regulators: Implication in central nervous system diseases". en. In: *Wiley Interdiscip. Rev. RNA* 13.2, e1678.

Li, Heng (Mar. 2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: arXiv: `1303.3997 [q-bio.GN]`.

Li, Heng and Richard Durbin (July 2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". en. In: *Bioinformatics* 25.14, pp. 1754–1760.

Li, Yang I et al. (Jan. 2015). "RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts". en. In: *Genome Res.* 25.1, pp. 1–13.

Ollà, Ivana et al. (Aug. 2023). "Pathogenic Mis-splicing of CPEB4 in Schizophrenia". en. In: *Biol. Psychiatry* 94.4, pp. 341–351.

Parada, Guillermo E et al. (Jan. 2021). "MicroExonator enables systematic discovery and quantification of microexons across mouse embryonic development". en. In: *Genome Biol.* 22.1, p. 43.

Parras, Alberto et al. (Aug. 2018). "Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing". en. In: *Nature* 560.7719, pp. 441–446.

Schafer, Sebastian et al. (Oct. 2015). "Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI)". en. In: *Curr. Protoc. Hum. Genet.* 87, pp. 11.16.1–11.16.14.

Tapial, Javier et al. (Oct. 2017). "An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms". en. In: *Genome Res.* 27.10, pp. 1759–1768.

Torres-Méndez, Antonio et al. (Apr. 2019). "A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons". en. In: *Nat. Ecol. Evol.* 3.4, pp. 691–701.

Ustianenko, Dmytro, Sebastien M Weyn-Vanhentenryck, and Chaolin Zhang (July 2017a). "Microexons: discovery, regulation, and function". en. In: *Wiley Interdiscip. Rev. RNA* 8.4, e1418.

— (2017b). "Microexons: discovery, regulation, and function". In: *Wiley Interdisciplinary Reviews: RNA* 8.4, e1418.

Voss, Kate, Geraldine Van Der Auwera, and Jeff Gentry (2017). *Full-stack genomics pipelining with GATK4 + WDL + Cromwell [version 1; not peer reviewed]*. slides. DOI: 10.7490/f1000research.1114634.1. URL: https://f1000research.com/slides/6-1381.

Wang, Kai et al. (Oct. 2010). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". en. In: *Nucleic Acids Res.* 38.18, e178.

Wu, Jie et al. (2013). "OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds". In: *Nucleic acids research* 41.10, pp. 5149–5163.

Wu, Thomas D and Colin K Watanabe (May 2005). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences". en. In: *Bioinformatics* 21.9, pp. 1859–1875.

Yu, Huihui et al. (Feb. 2022). "Pervasive misannotation of microexons that are evolutionarily conserved and crucial for gene function in plants". en. In: *Nat. Commun.* 13.1, p. 820.

Zhang, Jiajie et al. (Mar. 2014). "PEAR: a fast and accurate Illumina Paired-End reAd mergeR". en. In: *Bioinformatics* 30.5, pp. 614–620.

Table 2.1: ENCODE Long Read Sequencing Sample Descriptions

Accession	Biosample summary
ENCSR690QHM	Homo sapiens with Alzheimer's disease; dorsolateral prefrontal cortex tissue female adult (90 or above years)
ENCSR316ZTD	Homo sapiens with Alzheimer's disease; dorsolateral prefrontal cortex tissue female adult (90 or above years)
ENCSR462COR	Homo sapiens with Alzheimer's disease; dorsolateral prefrontal cortex tissue female adult (90 or above years)
ENCSR257YUB	Homo sapiens with Alzheimer's disease; dorsolateral prefrontal cortex tissue male adult (90 or above years)
ENCSR094NFM	Homo sapiens dorsolateral prefrontal cortex tissue female adult (90 or above years)
ENCSR463IDK	Homo sapiens dorsolateral prefrontal cortex tissue female adult (79 years)
ENCSR697ASE	Homo sapiens with Alzheimer's disease; dorsolateral prefrontal cortex tissue female adult (90 or above years)
ENCSR205QMF	Homo sapiens dorsolateral prefrontal cortex tissue female adult (88 years)
ENCSR169YNI	Homo sapiens dorsolateral prefrontal cortex tissue female adult (90 or above years)

Table 2.2: ENCODE Long Read Sequencing Sample Identifiers

Accession	Assay name	Biosample term name	Dbxrefs	Description	Lab
ENCSR690QHM	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219641	RUSH brain E6341028	Ali Mortazavi, UCI
ENCSR316ZTD	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219675	RUSH brain E6207143	Ali Mortazavi, UCI
ENCSR462COR	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219871	RUSH brain E3348003	Ali Mortazavi, UCI
ENCSR257YUB	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219645	RUSH brain E4368365	Ali Mortazavi, UCI
ENCSR094NFM	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219452	RUSH brain E4368365	Ali Mortazavi, UCI
ENCSR463IDK	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219875	RUSH brain E5194210	Ali Mortazavi, UCI
ENCSR697ASE	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219655	RUSH brain E7948794	Ali Mortazavi, UCI
ENCSR205QMF	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219574	RUSH brain E5115600	Ali Mortazavi, UCI
ENCSR169YNI	long read RNA-seq	dorsolateral prefrontal cortex	GEO:GSE219545	RUSH brain E7461192	Ali Mortazavi, UCI

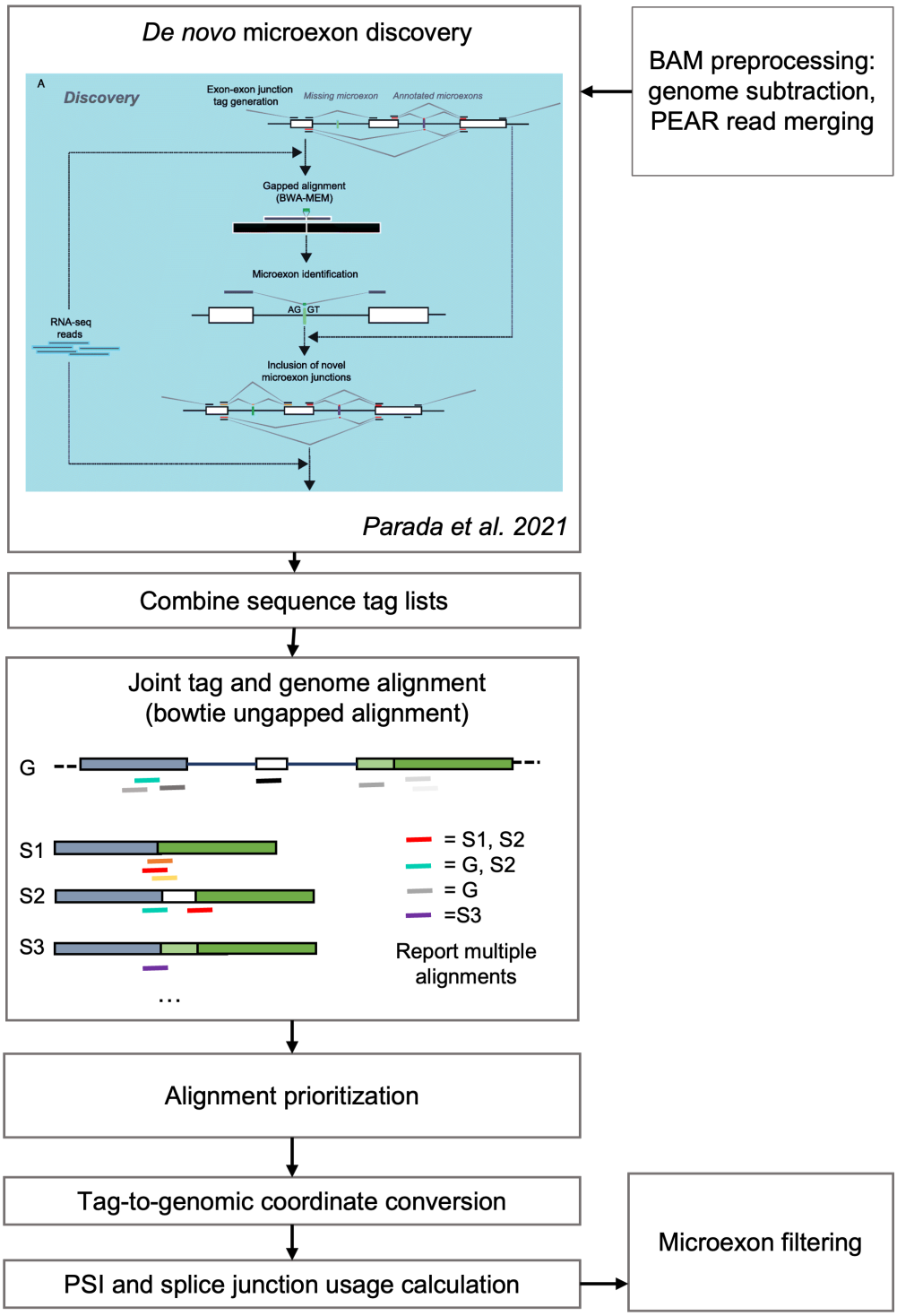


Figure 2.1: An improved pipeline for microexon discovery and quantification in large datasets: Processing steps in proposed pipeline. The region highlighted in blue indicates the MicroExonator tool pacakge developed by Parada et al., 2021.

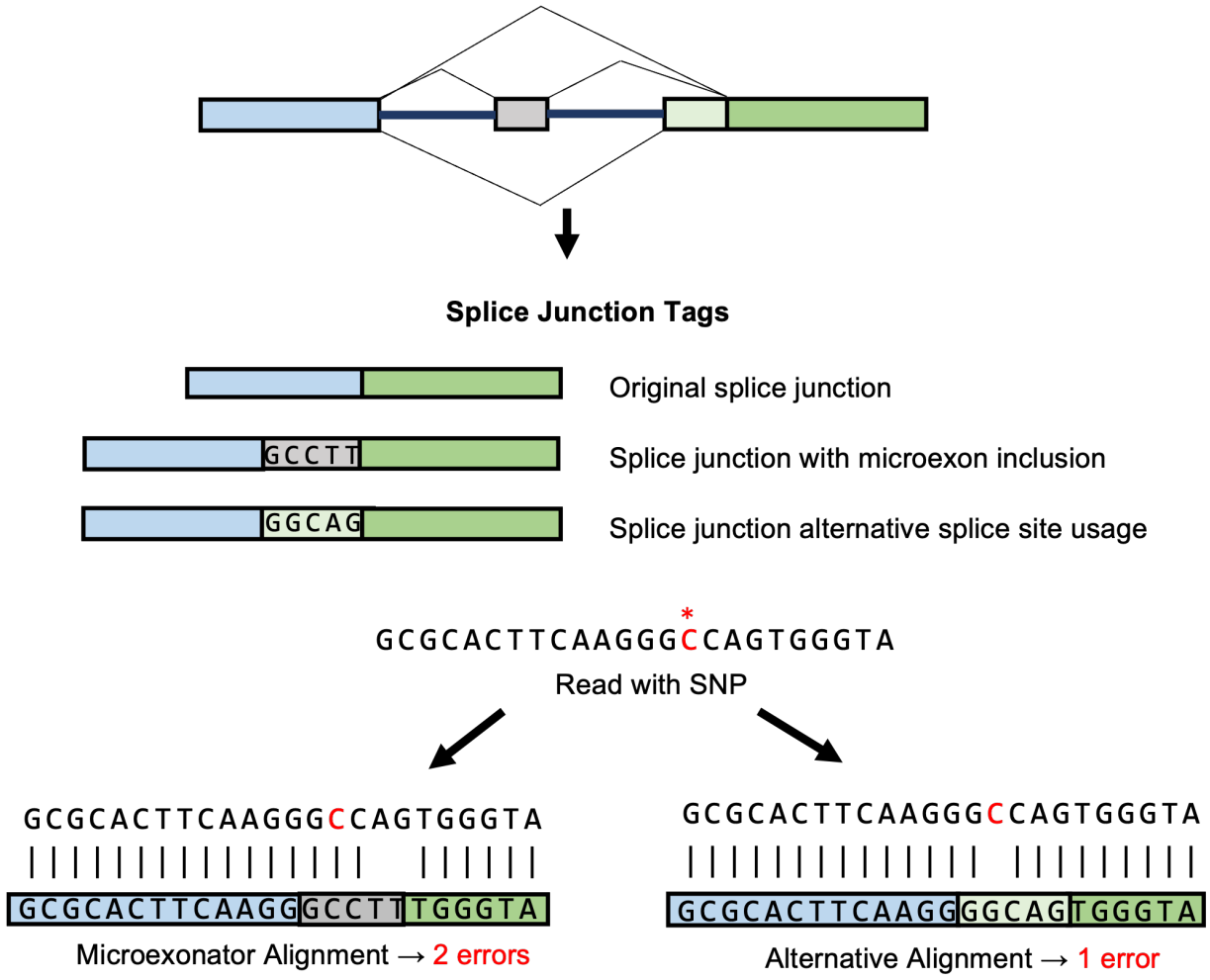


Figure 2.2: Variants and sequencing errors can create spurious microexon alignments: Example of a type of misalignment error that is captured by newly added filtering steps.

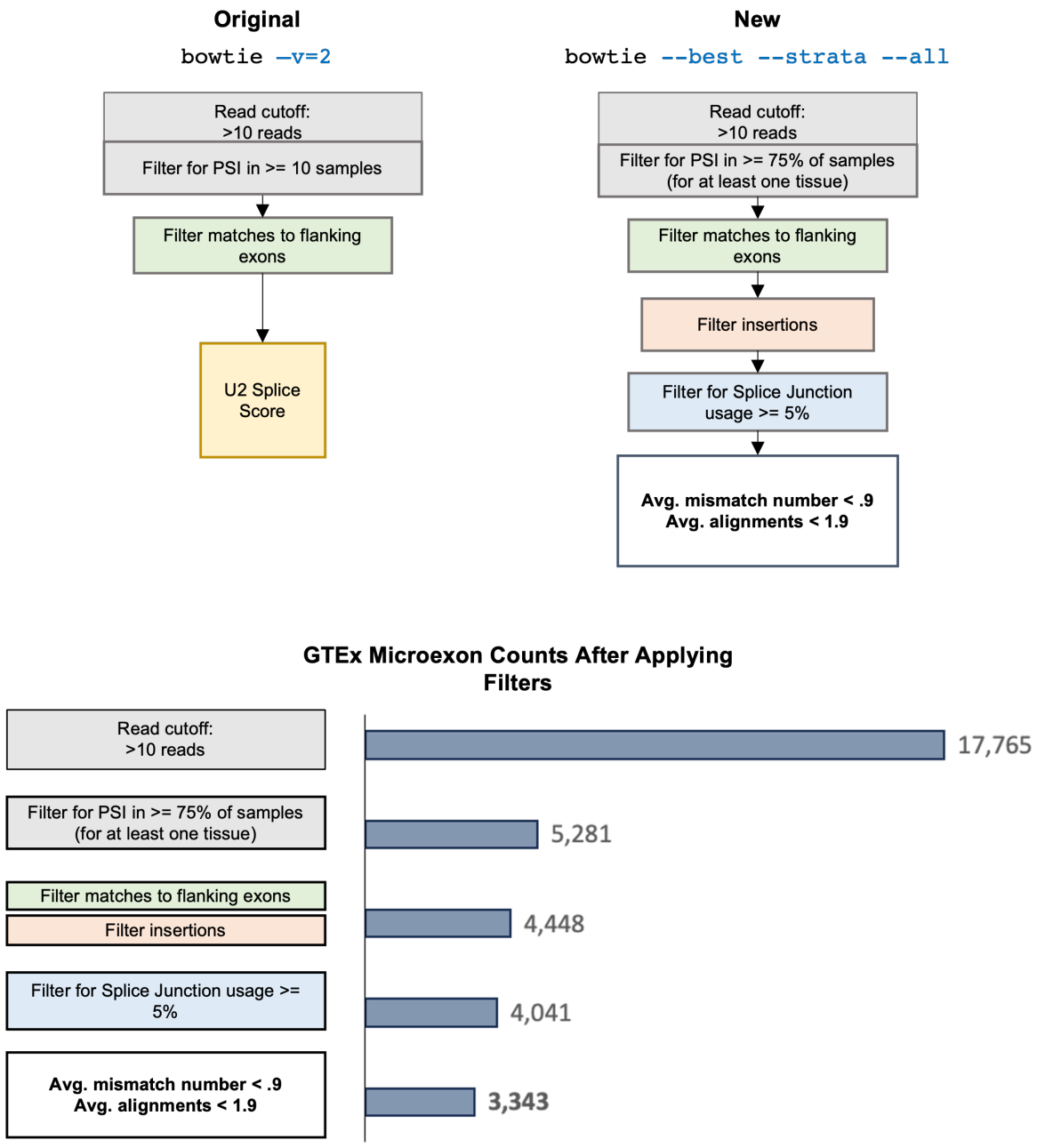
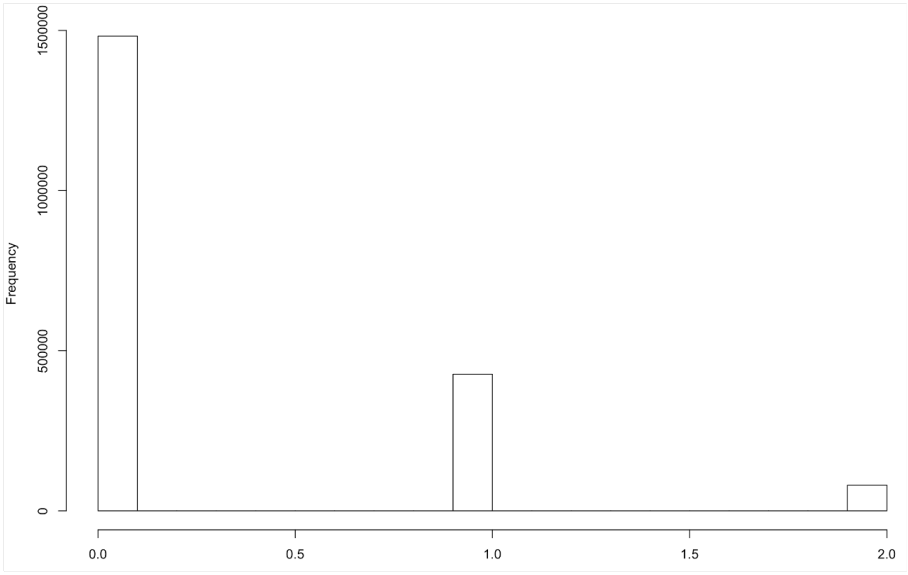


Figure 2.3: MICROEXON QUALITY FILTERS: Flowchart highlighting the differences in the filtering process in this work, as well as chart of the number of exons remaining at each filtering step.



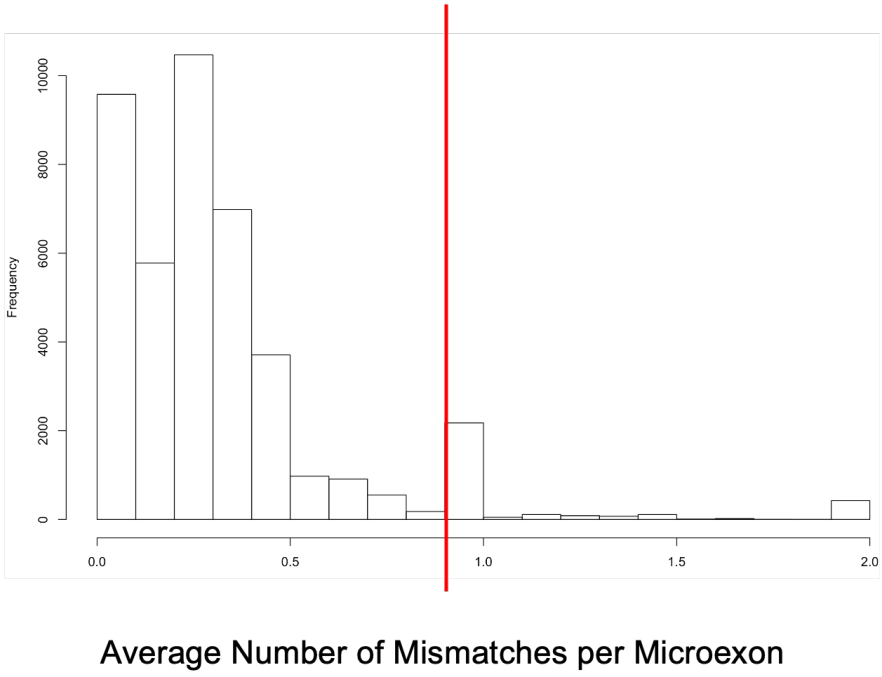


Figure 2.4: Alignment Average Mismatches for Microexon Reads: Histograms showing mismatches per alignment and average mismatches per microexon, along with the cutoff used for filtering.

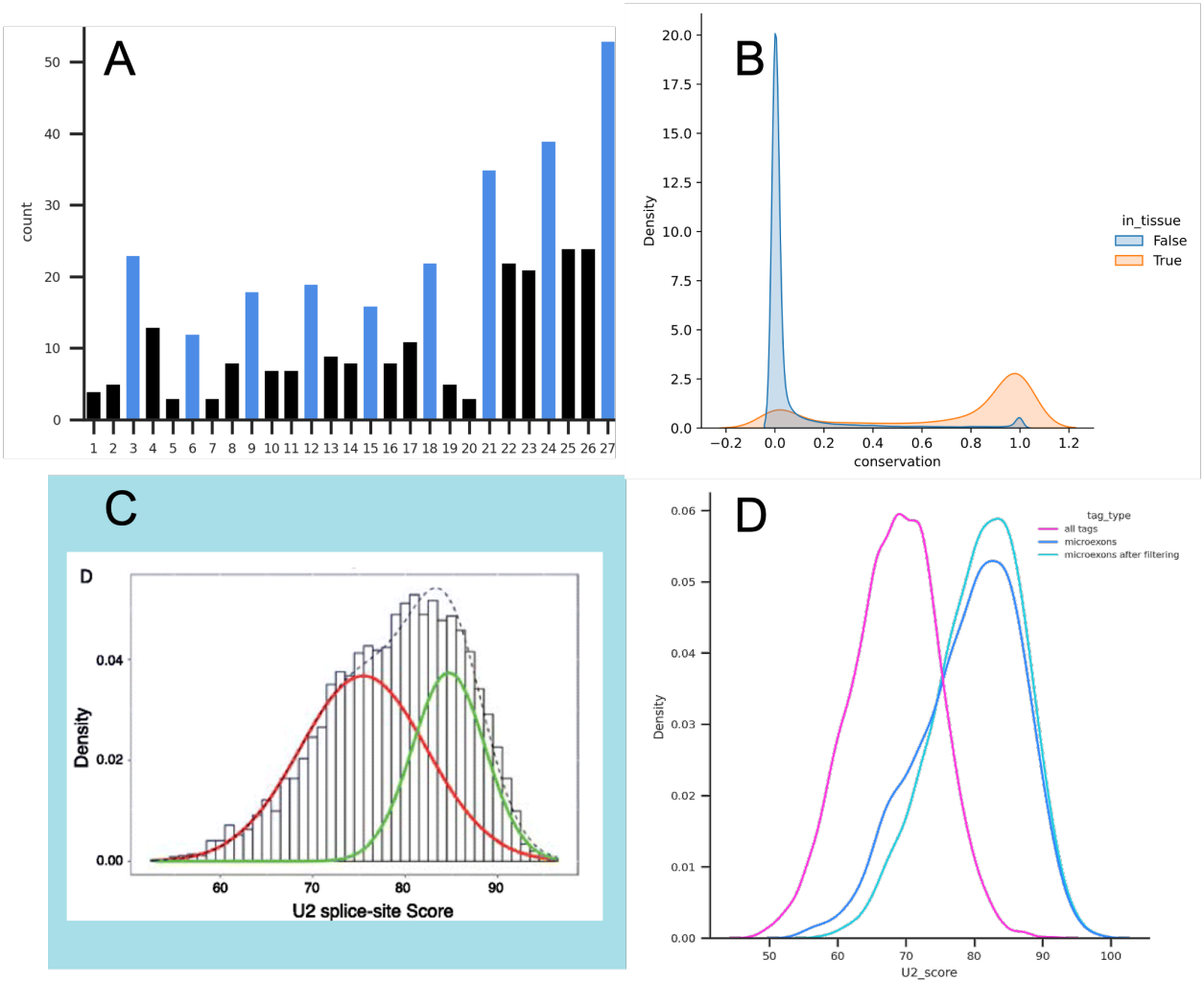


Figure 2.5: Detected microexons are enriched for functional signals: A) Histogram of microexon length showing enrichment for lengths that are multiples of three. B) Conservation plot showing that filtering captures conserved microexons. C) U2 Score plot from Parada et al., 2021 showing highly overlapping distributions for microexon reads and misalignments. D) Comparative plot showing how filtering in this work removes overlapping distributions.

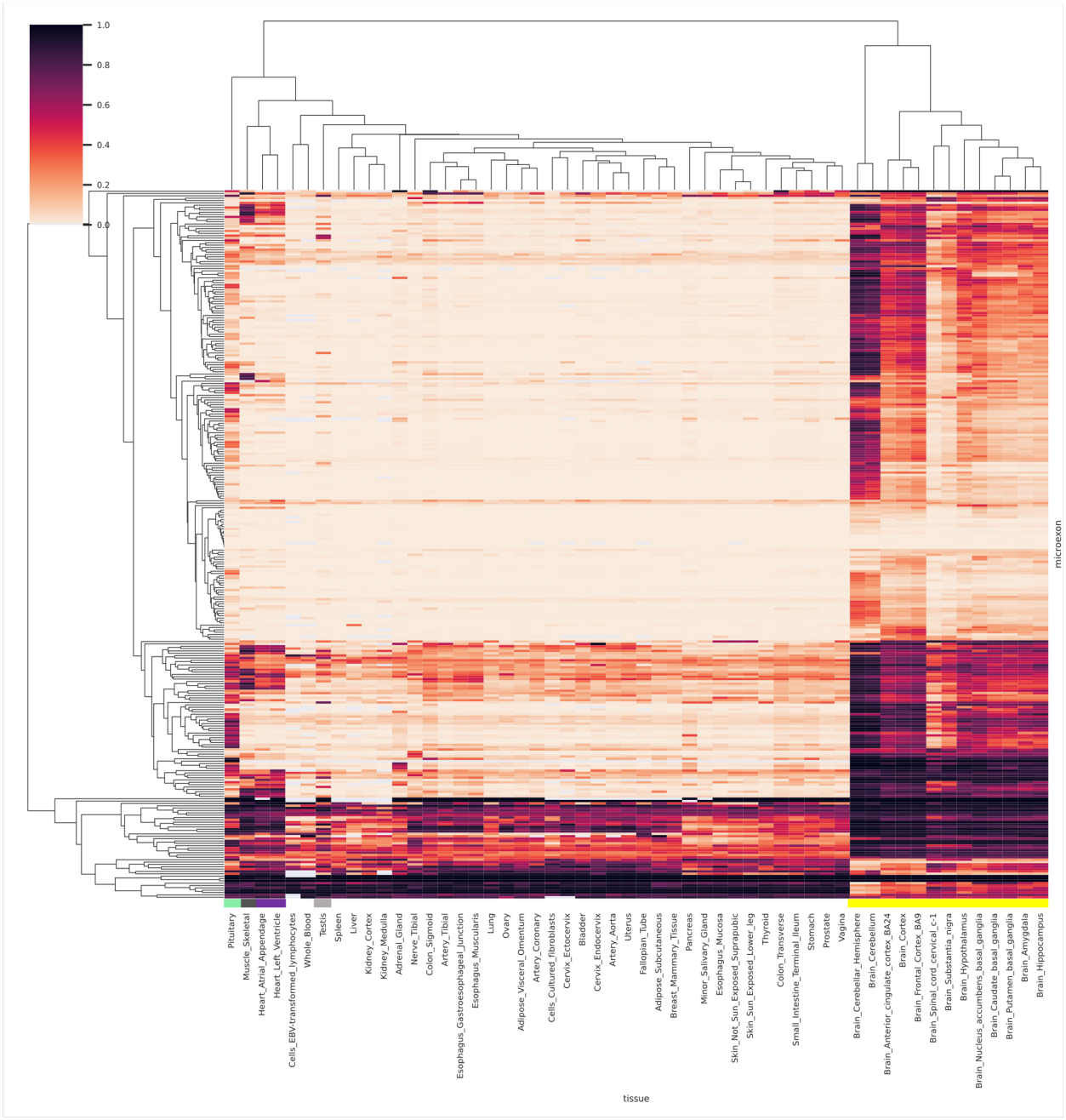


Figure 2.6: Splicing of Conserved Neuronal Microexons: Heatmap of PSI calculations showing distinctive tissue-specific splicing patterns for 308 curated microexons in Irimia et al.

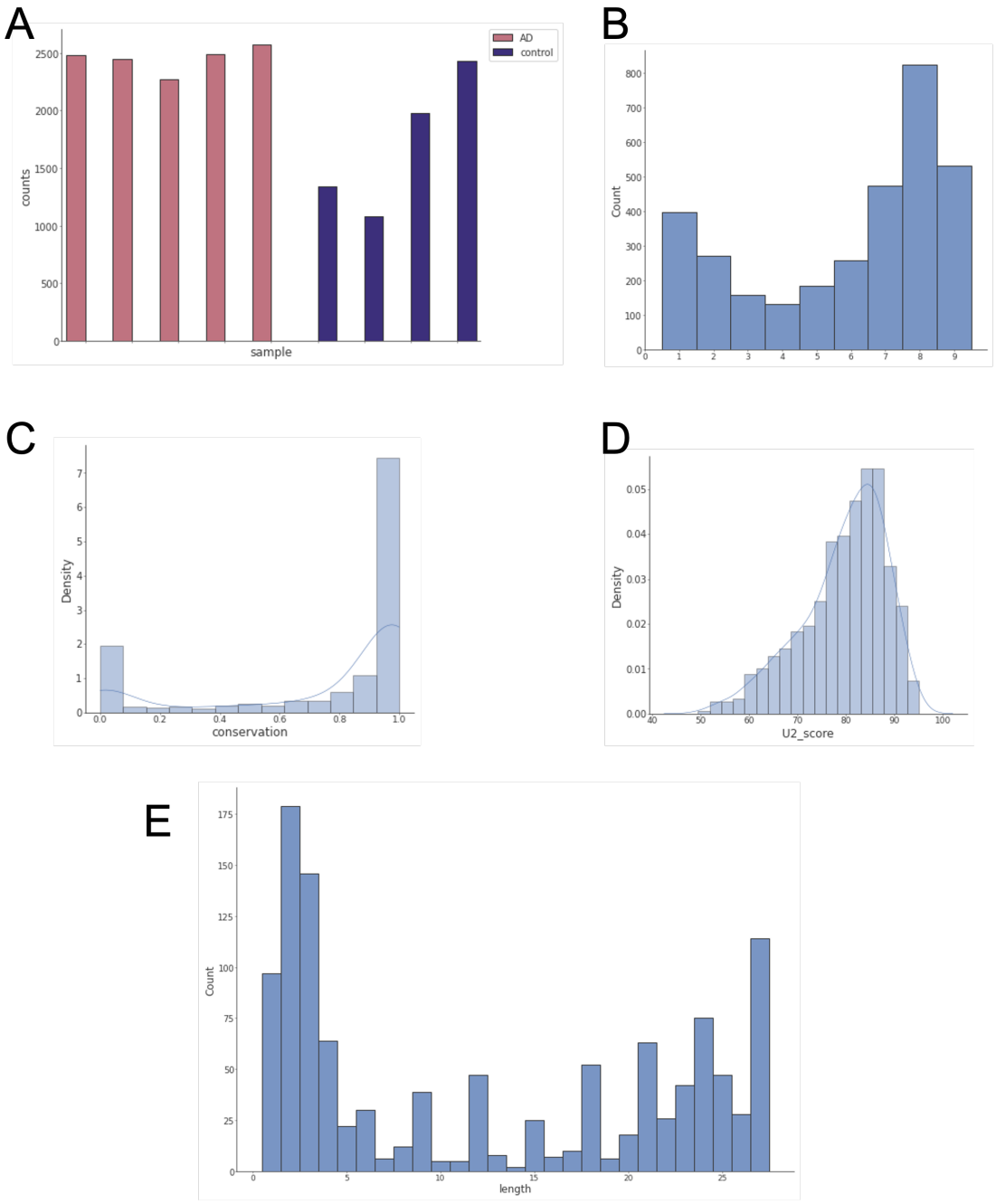


Figure 2.7: Long Read Validation of Microexons found in GTEx. A) Number of microexon splice junctions with a minimum of 3 reads in each long read sample. B) number of samples each microexon is detected in. C) and D) histograms of PhastCons average conservation and combined splice site strength score (U2 score) for microexon sequences which validated in microexon. E) length distribution of microexons found in long read RNA-Seq

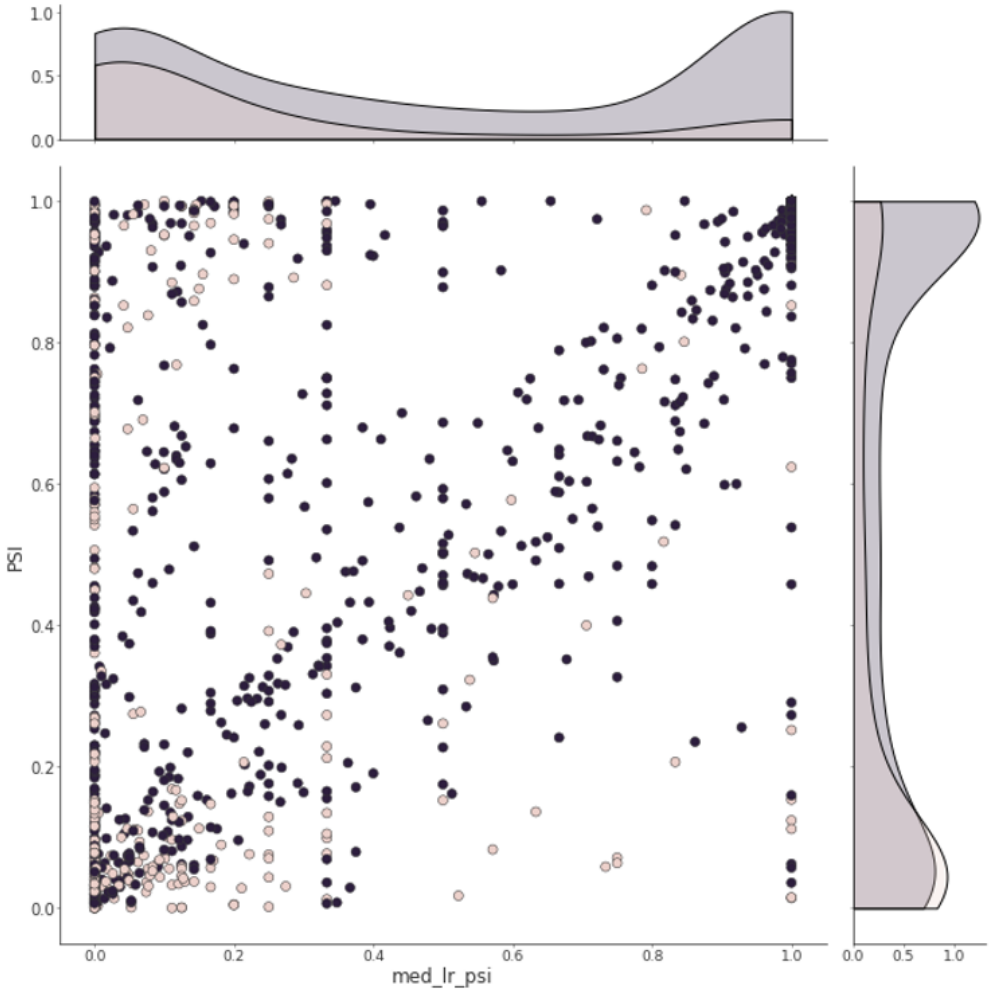


Figure 2.8: Scatterplot showing the median long read PSI versus microexon PSI calculated in GTEx. Light colored dots indicate ambiguous microexon sequences that were filtered from the final reported microexon list.

CHAPTER 3

Microexon Discovery and Quantification in Human Adult Tissues

3.1 Introduction

RNA splicing is the process by which stretches of non-coding sequence are removed from pre-mRNA molecules, and the remaining exons are joined together. Splice site usage is heavily influenced by cooperative and competing interactions between components the spliceosome, RNA-binding proteins and RNA molecules. Differences in RNA sequence, RNA structure, and cellular contexts result in different interaction patterns and therefore different splicing outcomes. This alternative splicing may result in functionally distinct isoforms from the same gene under different conditions. In humans, exons are typically much shorter than introns, and are on average around 140nt, with the majority being between 50 and 200nt. This size preference is thought to facilitate the assembly of the spliceosome across an exon, but remain long enough to contain enough information for splicing and not create steric hinderance of the spliceosome. Despite this size preference, "microexons" that are considerably shorter (previously defined as $\leq$27nt or $\leq$51nt) have been observed.

Despite their small size, microexons have been shown to have significant functional importance. Microexon splicing has been shown to be most common in tissues with highly tissue-specific splicing patterns including the brain, pituitary gland, heart, muscle and testis, suggesting that microexon splicing may create functionally important isoforms in these tissues. Microexons have also been shown to be important in a variety of contexts, including neural splicing and autism. In addition, microexons have been implicated in cancer and muscular dystrophies. Interestingly, microexons are some of the most conserved exons known, indicating that they may have played a crucial role in the evolution of complex neural systems. Despite their importance, a comprehensive survey of microexon splicing in adult human tissues has not been conducted to date. Quantification of microexon splicing requires special consideration, as the short length of these exons makes them challenging to align and quantify accurately.

In this study, we performed a comprehensive survey of microexons in adult human tissues. We employed an improved microexon alignment pipeline to produce a list of high-quality microexons and examine the types of splicing events found in these exons. We also explore the genetic regulation of microexon splicing changes using splicing Qualitative Trait Loci (sQTLs). Our analysis reveals that microexons exhibit tissue-specific splicing patterns and that individual microexons have tissue-specific variation. We identify microexons that are key to neuromuscular junction function and show that pathogenic intronic variants from ClinVar may be due to cryptic microexon splicing. Overall, our study provides a comprehensive characterization of microexon splicing in human tissues and sheds light on the functional importance of these unique exons. Our findings suggest that microexons play a crucial role in regulating splicing events in tissues with complex and dynamic

splicing patterns, and highlight the need for further research into the role of microexons in human biology and disease.

3.2 Results

3.2.1 Quantifying Microexon Splicing In Adult Human Tissues

In order to characterize the landscape of microexon splicing in adult human tissues, we performed microexon discovery and quantification on 17,121 RNA-Seq samples in 49 unique tissue types from the V8 release of the Genotype-Tissue Expression (GTEx) Project (GTEx Consortium et al., 2017). Microexon discovery was performed within all annotated splice junctions in the GENCODE comprehensive v39 annotation (Frankish et al., 2021), resulting in over 1M putative microexon sequences of length 27nt or shorter. We retained PSI values for all microexons having at least 10 reads mapping to the splice junction in a given sample. In order to identify the most commonly included microexons, we filtered for microexons present in 75% of samples in at least one tissue, with an average PSI of at least 5%. After filtering, we retained 3,343 microexons in 2,242 genes. Figure 3.1A summarizes the filtering steps (discussed in detail in the previous chapter) and the number of microexon sequences passing each step.

In our survey of microexons across tissues, we first asked whether the microexons detected in our data were captured by existing annotations. Interestingly, we identified only moderate overlap with annotations from existing databases. Of 16,071 microexons annotated in at least one of GENCODE comprehensive v39(Frankish et al., 2021), VASTDB, or RefSeq, we observed that 1,547 passed the above filtering criteria. Note that

approximately half of all microexon annotations in VASTDB(Tapial et al., 2017) and Refseq [VERSION](O'Leary et al., 2016) found in just one annotation set, possibly due to incomplete annotations (Figure 3.1B). As many microexons have been shown to be specific to tissue type, cell type, and developmental stage (Gonatopoulos-Pournatzis et al., 2018; Irimia et al., 2014; Parada et al., 2021; Li et al., 2015), it is expected that the large panel of adult tissues included in our project may identify novel microexons unobserved in previous studies. The differences in microexon numbers across annotation sets and analyses demonstrates the utility of performing microexon discovery in different biological contexts.

We also looked for evidence of functional microexons from their length and location. As previously reported, the microexons we observed are enriched for having lengths that are multiples of three (Figure 3.1C). Alternative splicing of these symmetric microexons does not disrupt the reading frame. We also found that microexons were predominantly found in protein-coding genes, with a small fraction in lncRNAs and pseudogenes (Figure 3.1D).Additionally, one would expect most constitutive microexons to have been previously annotated in various databases as they are less tissue-specific than alternative microexons. Indeed, we found that 535/597 (89.6%) of our detected constitutive microexons were present in at least one annotation set. Similarly, only 3.5% (62/1789) of novel microexons were constitutive. (Fig 1E).

Next, we investigated the tissue specificity of microexon splicing patterns. We used tSNE to cluster the PSI of the 3,343 microexons across 17,121 GTEx RNA-Seq samples (Fig 1F). To mitigate the effects of gene expression on the t-SNE, we imputed missing PSI values due to low splice junction coverage with the cross-tissue means. Surprisingly, samples clustered tightly by tissue for all tissue types, despite previous literature suggesting that

strong patterns of tissue-specific splicing of microexons mainly observed in the brain, muscle, heart, and testis (Li et al., 2015). In contrast, our results suggest that tissue-specific microexon alternative splicing occurs beyond the above well-studied tissues. We also carried out sample clustering based on the gene expression of the microexon-containing genes and observed a similar tissue-specific pattern (supp fig 3.) This observation suggests that microexons occur both in highly tissue-specific genes and also have tissue-specific exon inclusion.

Additionally, we measured microexons which showed a high degree of tissue-specificity by calculating a tissue-specificity score tau, which has been previously used to measure degree of tissue specificity (see Methods). We reported the number of tissue-specific microexons (tau $\geq$.7) with a PSI of $\geq$.1 in each tissue type (Figure 3.1G) (see methods for tissue-specific calculation.) As previously reported, brain tissues had the highest number of included microexons, followed by skeletal muscle, pituitary, heart, and testis. All tissues had at least some included tissue-specific microexons. After identifying microexons across tissues, we asked what types of genes contained microexons.

3.2.2 Microexons are Enriched in Genes Involved in Mechanotransduction

To examine the functional relevance of genes that contain alternatively spliced microexons (1,367 in total), we analyzed the enrichment of GO Molecular Function terms of these genes using enrichR (Kuleshov et al., 2016) (Figure 3.2). EnrichR uses UMAP to perform unsupervised clustering of significantly enriched GO terms based on the gene IDs associated with each term. GO terms with more overlapping genes should appear closer together on the plot. We found 457 significantly enriched GO terms ($p<.05$), and observed

several distinct clusters of GO terms.

We noted that the majority of enriched terms relate to various aspects of mechanobiology, including cytoskeleton proteins and their organization, signalling mediators and pathways including ATPase and GTPase activity and Ras signalling, as well as sensor proteins and adaptors like calcium ion channels, and cadherin. Many terms also relate to brain- muscle- and heart-specific signaling, such as sarcoplasmic calcium ion transport, muscle contraction and response to muscle stretch, and cortical actin cytoskeleton organization. Biomechanical processes including cell-matrix adhesion, cell projection, stress fiber assembly, focal adhesion assembly, basement membrane assembly, and wound healing are also enriched. We note that related terms are often found in different clusters, indicating multiple distinct sets of genes related to these processes. We also find that the number of genes overlapping each cluster's terms are in the hundreds (Figure 3.2). This suggests a large number of microexon-containing genes overlap these processes, as opposed to a small number of genes driving enrichment of multiple related terms.

The top ten enriched terms also identify individual microexons in several voltage-dependent calcium channels, including CACNA1A, B1, B2, and B3, which are long proteins with complex, brain-tissue specific splicing patterns. Among these channels, CACNB1 exon 7 of 13 was a conserved 20nt microexon that had been previously reported in Irimia et al., 2014. Recent studies have also validated microexon inclusion and novel isoforms in the psychiatric risk factor CACNA1C using long-read sequencing and RT-PCR (Clark et al., 2020).

3.2.3 Widespread Evidence of Conserved, Tissue-Specific Regulation of Microexon Splicing

Having identified microexons in many functional and disease-relevant genes, we next looked for evidence that the alternative splicing of these microexons is functional.

We next looked for evidence of regulation by RNA binding proteins by performing enrichment analysis for RBP motifs in the flanking intronic sequences of microexons using SEA (Figure 3.3A) (Bailey and Grant (2021)) Analysis on all alternatively spliced microexons with high brain or muscle inclusion showed enrichment for PTBP1, which is a key microexon splicing factor reported in (Li et al. (2015)). Interestingly, we also found strong enrichment of UGCU, which is known regulate MBNL-1 mediated splicing (Klinck et al., 2014). Additionally, a motif matching PCBP1/2 or QKI was also found. Positionally enriched long U-rich and G-rich motifs were also observed.

We also asked whether microexons in functionally related genes would have enrichment for different splicing factors (Figure 3.3A). We used the lists of overlapping genes from our GO Term clustering, as distinct clusters should have minimally overlapping genes with different functional enrichment. Clusters tended to have unique RBP enrichments. Enriched RBPs included known tissue-specific splicing factors. Cluster 9 contains terms related to muscle contraction and function, and is enriched for FXR1, which is known to be important for muscle development (Smith et al., 2020). We see that the clusters tend to have distinct enrichments for different RBP motifs, suggesting that these sequences are under control of distinct regulatory mechanisms. SRSF1 was enriched in several clusters. Interestingly, Matrin3 was strongly enriched in the 5' intron sequences of cluster 2 genes enriched for

cytoskeleton protein organization terms. It is an RNA- and DNA-binding nuclear matrix protein found to be associated with neural and muscular degenerative diseases. These results suggest microexons may be regulated by highly-tissue specific splicing factors.

We measured the average PhastCons score of microexons and 100 bp of their flanking intronic sequences of microexons grouped by the tissue in which they are most highly expressed, and compare to constitutive microexons (Figure 3.3B) All tissue-specific microexons had higher average conservation scores for their intronic sequences than constitutive ones, with the 3' intron of brain-specific microexons especially conserved. Brain-specific microexons have been shown to be among the most conserved sequences in the human genome, with some being conserved across 600 million years and in species as distant as zebrafish (Irimia et al. (2014), Torres-Méndez et al. (2019)). The high averages suggest that many microexons have highly conserved intronic sequence elements and have tissue-specific splicing patterns.

3.2.4 Microexons may affect transcription as well as protein structure

We next sought to determine whether microexon sequences were predicted to affect protein function. Previously, it has been shown that microexons are often found in intrinsically disordered protein regions and have the potential to alter protein-protein interactions (Li et al., 2015, Irimia et al., 2014). We additionally show that microexons are significantly enriched for lycine residues over the average in human proteins and in longer exons in microexon-containing genes (Figure 3.3C). Lycine makes up of 5.3% of all protein residues, but made up 8.5% of residues in annotated protein-coding microexons. Lycine residues are positively charged, and can also impact protein function by being ubiquinated.

Interestingly, lycine residues are also are involved in collagen crosslinking and actin acytylation, which can change cytoskeleton protein binding affinities (A et al., 2020). It would be interesting to see if microexon sequences could be targets of these modifications.

In addition to protein changes, we also observed over 600 annotated and novel microexons in gene 5' UTRs, which to our knowledge has not been previously reported (Figure 3.3D). While coding microexons were strongly enriched for multiples of three, we observed that many of our frameshifting microexons were found in the 5' UTR of genes (Figure 3.3E). Approximately 35% of genes have introns in their 5' UTRs, which can be a few hundred or larger nucleotides. 5' UTRs impact RNA translation and degradation rates, and can bind localizing proteins in tissue-specific manner. We found that over half of our UTR microexons for both 5' and 3' UTRs exhibited high conservation (Figure 3.3F). The alternative splicing of 5' UTR microexons represents a novel potential mechanism for microexon splicing to impact protein abundance and localization.

3.2.5 Identification of Microexon splicing QTLs across GTEx Tissues

Next, we sought to leverage the large sample size of GTEx to look for genetic variants that can affect microexon splicing. In order to investigate the effect of common genetic variants on microexon splicing, we performed sQTL analysis using tensorQTL (Taylor-Weiner et al., 2019). Our approach involved identifying genetic variants associated with changes in the measured microexon PSI values across samples for each tissue. We utilized the genotype covariates calculated in GTEx 2017, along with explicit covariates for sex, PCR, and sequencing platforms. All variants in a window of 1 Mb around each microexon were tested, and PEER was used to calculate hidden covariates for each tissue based on the

sample size, as recommended by GTEx 2017. A permutation-based analysis was conducted using TensorQTL, and all significant sQTLs were reported at a false-discovery rate of 0.05.

Our analysis identified a total of 2,113 variants associated with 957 microexons, representing 5,560 single-tissue sQTL pairs, with 497 of them being multi-tissue sQTLs. The number of sQTLs per tissue varied linearly depending on the sample size, as expected (Figure 3.4A). We found that more significant microexon sQTLs tended to be located closer to the exon, which agrees with our knowledge of splicing sequence information density (Figure 3.4B). We also observe that sQTL microexons tend to have lower conservation (Figure 3.4C). One possible explanation of this is that since they may be more variably included due to common variants, they are more likely to keep the protein in-frame.

After identifying numerous sQTLs associated with microexon splicing changes, we asked whether any of these sQTLs could be identified as associated with common disease. To this end, we conducted colocalization analysis on 86 harmonized GWAS datasets from Barbeira et al. (2018). The summaray statistics of each of the 86 traits were tested against those of the sQTL analyses in each tissue. We tested 373,784 sQTL-trait-tissue combinations. Overall, we found 320 significant colocalizations, involving 23 different microexon regions and 39 unique GWAS traits. Of the traits with sQTL colocalizations, we found 13 related to blood cell counts, 8 related to body composition traits, 8 relating to autoimmune conditions, including asthma and Crohn's disease. The trait categories are given in Figure 3.4D. We identified a top colocalization for GasderminB, which was also identified by Garrido-Martın et al., 2021 in sQTL analysis of STAR aligned GTEx RNA-Seq. This variant is associated with asthma risk and likely causes a multi-exon skipping event from exons 5-9, with exon 7 being a microexon. Our data showed that the skipping event is

exons 4-9. Additionally, we found a multi-tissue sQTL for an annotated 21nt microexon in INO8E, a gene involved in spontaneous calcium-dependent neurotransmitter release and chromatin remodeling (Poli, Gasser, and Papamichos-Chronakis, 2017), which was previously identified as a top gene for follow-up in GWAS for schizophrenia (Lago and Bahn, 2022). This sQTL colocalized with esinophil counts and red blood cell count phenotypes as well. We also demonstrated the change in PSI based on the number of alleles for the associated SNP (rs) in Figure 3.4E, and show representative alignments for individuals from each genotype to demonstrate the effect of the variant.

3.2.6 Microexons Splice Sites are Enriched for Pathogenic Rare Disease Variants

Since GTEx is not expected to have disease state individuals, we considered annotated disease variants from CLINVAR which overlapped with microexon annotations. We overlapped 3 million CLINVAR variants with our microexon annotations and found that 2,555 of them fell within the splice regions of microexons. Of these, 8% were classified as Pathogenic by CLINVAR, which was on par with the proportion of long exon splice region variants with the same classification, and substantially higher than the baseline frequency of Pathogenic variants in CLINVAR (5%) (Figure 3.5A). Interestingly, we noted that for novel microexon annotations, the percentage of CLINVAR Pathogenic variants was higher, with 10% of variants in novel splice sites labeled Pathogenic. This suggests that the alteration of splicing of some of the novel microexons could explain the pathogenicity of some variants currently labeled as deep intronic variants. To further investigate this, we looked at variants labeled by CLINVAR as intronic, which were not labeled as splice region variants. For microexon-overlapping intronic CLINVAR variants, we observed rates of

pathogenicity on par with those overlapping long exon splice regions, and higher than that for other intronic variants (Figure 3.5B). The microexon-overlapping Pathogenic variants were frequently associated with diseases which related to muscular and neurological phenotypes, in line with the gene sets which are enriched for microexons. These microexon variants may warrant follow-up as potential mechanistic explanations for some disease-causing variants.

3.2.7 Microexon Splicing May be Dysregulated in Myotonic Dystrophy Type I

We identified functional microexons in genes related to muscle function, calcium ion signalling, and neuromuscular junctions (Figure 3.2). We also identified enrichment for several muscle-specific splicing factors, as well as RBFOX1 and QKI, which have previously been shown to be key regulators of microexons (Li et al., 2015, Lee et al., 2020). Recently, it was shown that RBFOX1 cooperates with MBNL1 to control splicing events in muscle, including events dysregulated in Myotonic Dystrophy Type I (DM1) (Klinck et al., 2014). DM1 is the most common adult-onset muscular dystrophy, affecting 1 in 8000 individuals. DM1 is primarily caused by unstable repeat expansion in the DMPK gene, resulting in sequestration of MBNL1 and other splicing factors by toxic RNA (Taneja et al., 1995). Dysregulation of splicing factors is thought to be a main driver of the disease, which may also include tau protein misfolding and microRNA dysregulation (Caillet-Boudin et al., 2014), with missplicing present in both muscle and brain (Degener et al., 2022). RBFOX1 and QKI, both important for microexon splicing, have also been shown to be dysregulated in DM1 (Klinck et al., 2014). These findings prompted us to investigate whether microexon missplicing could contribute to DM1 pathology.

We quantified the identified microexons in publicly available RNA-Seq samples obtained from skeletal muscle biopsies of 16 DM1 adults, 6 adult controls, 9 congenital myotonic dystrophy, and 9 pediatric controls. We used rMATS to test for differential exon inclusion between the adult samples and controls based on the resulting PSI values. We found 97 differentially spliced exons, with almost half being <51nt long, and 14 being <27nt long (Figure 3.6A). To further investigate the potential disease relevance of these microexons, we performed t-SNE analysis of the samples and found that microexon splicing patterns alone could largely differentiate between disease and control samples (Figure 3.6C. Additionally, clustering on the differentially expressed microexon PSIs showed that control samples tended to cluster together and disease samples mostly clustered together, indicating the potential involvement of these microexons in disease pathogenesis (Figure 3.6B). Although the overall effect size is modest, small changes in highly disease-relevant splicing events could contribute to the observed phenotypes. The longer mis-splicing events included MBNL1 and QKI, which are known to be dysregulated in DM1. We input the genes from all disrupted microexons <51nt into Cytoscape gene network using STRINGDB and found that these differentially spliced microexons were involved in closely interacting genes associated with muscle contraction (Figure 3.6D). Ryr1 and Bin1 are known to have strongly disrupted microexons, with 15nt and 42nt exons being differentially spliced and validated by RT-PCR. Furthermore, we observed signs of mis-splicing correlated with the disease for ACTG1, MYL6, and other genes, although the specific exons were not identified.

In our analysis, we observed a significant overlap between the muscle- and brain-specific microexons we identified and CLINVAR rare disease variants associated with myopathies, CMT disease, and other related phenotypes. This suggests that the splicing events

involving these microexons may play a critical role in gene function. Finally, we investigated the potential involvement of specific splicing factors in these microexons and found an enrichment of SRSF10 binding sites in the upstream intron compared to scrambled background sequences. We propose that these microexons and splice junctions should be studied further in the context of DM1, myopathies, and other related diseases to better understand their potential contribution to disease pathology.

3.3 Discussion

Our study presents the largest survey of short exon splicing in humans to date. Our analyses of microexon sequence, conservation, and enrichment for pathogenic variants suggests that these sequences are highly functional. Our focus on microexons with sequences that can be unambiguously mapped and supported by higher numbers of reads has revealed hundreds of highly conserved microexon sequences. However, we identified more than 12,000 additional microexons that passed quality control but not the PSI filter, which could be individual-specific, cell-type specific, or cryptic/weak splice sites. Our findings raise the question of how frequently to expect such cases of splicing events. We acknowledge that our study did not examine these microexons, but we believe that the scalability and selectivity of our method could enable further investigations.

The importance of a limited number of microexons has been previously reported, including exons associated with autism, cancer, and development in embryonic stages. Recent studies have argued that rho-gtpase microexons play a role in neurological phenotypes. However, our results suggest that these are not the only functionally relevant microexons. We observed highly regulated microexon inclusion in hundreds of critical genes with

tissue-specific roles, particularly in muscle, heart, and brain, containing some of the most specialized tissue and cell types. We argue that microexons preserved in the genome could be a strong sign of functionality, and that microexons may

Since our microexon detection and quantification is limited by sequence complexity and transcript abundance, it is possible we are limited in detecting microexons to highly tissue-specific and abundant genes, and cannot draw strong conclusions about the overall prevalence of microexon splicing or the proportion of functional short exon splicing events. Nevertheless, our study showed that microexons are present in super important genes, and our subsequent analyses suggest that they are susceptible to disruption, warranting attention.

The surprising abundance of functional short exons leads us to consider why this is the case, given the disadvantages of their smaller size. Previous studies have proposed switch-like behavior enabled by the small size of microexons. In the evolution of neural-specific microexons, researchers have discussed their role in promoting neural complexity.

We speculate that microexons could be splicing events with a significant impact on gene function, such that major disruption by a splice site-destroying variant causes Mendelian disease. Moreover, their collective mis-splicing due to trans factors disrupted in DM1 impairs muscle function and contributes to degeneration. Our study suggests that the disruption and importance of microexons may have been overlooked before, perhaps due to their lower effect size. However, their high interconnectivity and functionality may make them have an additive effect that affects disease. Previous studies have reported thousands

of splicing events that are widely disrupted, but microexons are difficult to find if one is not actively looking for them. The shortest ones may be missed using traditional aligners, and some studies have used sequence capture arrays that would not be able to detect microexons.

We acknowledge that there has been a lack of agreement about the causes of DM1, DM2, and other muscle disorders, and microexons could be pleiotropic in brain, heart, and muscle disorders. Future studies could investigate the protein domains of specific microexons, co-splicing of microexons, and isoform-level splicing of microexon-containing genes.

Our study highlights the complexity of the alternative splicing landscape in humans and the potential utility of simple exon inclusion ratios for very short exons, which represent an underappreciated class of highly functional splicing events. These findings underscore the importance of considering microexons in studies of splicing events and their role in disease.

3.4 Methods

3.4.1 Microexon Filtering

3.4.2 Unsupervised Clustering of GTEx RNA-Seq Samples

We used the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm with our microexon PSIs for high-dimensional data visualization. Since we wanted to cluster on splicing patterns, we imputed missing data which indicated that a transcript containing a microexon was not expressed in that sample to mitigate the effects of gene expression. Percent Spliced In (PSI) values were imputed using the tissue mean by the 'simpleimpute' Python package to handle missing splicing events. Dimensionality reduction was then

performed using the scikit-learn implementation of t-SNE in Python with a perplexity value of 30 and 1000 The resulting two-dimensional t-SNE plot was generated and subsequently used for visualization and interpretation of the complex splicing landscape across the diverse set of RNA-Seq samples by labeling each sample with the tissue color designation assigned by the GTEx consortium.

3.4.3 Calculation of Tissue-Specificity Metric Tau

To quantify the tissue specificity of gene expression across multiple tissue samples, the Tau metric was employed as described by Yanai et al. (2005). Tau values range from 0 to 1, with higher values indicating greater tissue specificity. For each gene g, its expression level E_{gt} was measured across T different tissues. The Tau value for each gene was then calculated using the formula

$$Tau(g) = 1 - \frac{\sum_{t=1}^{T} \left(\frac{E_{gt}}{Max(E_g)}\right)^{\beta}}{T}$$

, where $Max(E_g)$ is the maximum expression level of gene g across all tissues, and β is a user-defined exponent set to 1 in this study. The calculations were implemented in Python using the NumPy and Pandas libraries. Genes were categorized as tissue-specific if their Tau value exceeded a predetermined threshold, which we set to .7, as the data contained several groups of related tissues.

3.4.4 enrichR GO Enrichment of Molecular Function Terms

To perform functional enrichment analysis on the list of differentially expressed genes (DEGs), we utilized Enrichr (Chen 2013, Kuleshov 2016, Xie 2021), a comprehensive online

platform offering various precomputed gene-set libraries. Specifically, we focused on Gene Ontology (GO) Molecular Function terms to identify the biological functions potentially affected by the identified microexons. The list of alternatively spliced microexons was uploaded to the Enrichr web application, and the GO Molecular Function 2021 library was selected. Enrichr then generated an enrichment score for each term, considering the overlap between the input gene list and the gene sets in the selected library. To visualize the relationships among enriched GO Molecular Function terms, we used the interactive appyter module provided to perform the Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce the dimensionality of the data. Shared gene membership of the GO terms was used to calculate the distances in the UMAP.

3.4.5 RPB Binding Motif Enrichment

To identify RNA-binding protein (RBP) motifs enriched in our RNA sequences, we employed the MEME Suite, a toolkit tailored for motif discovery and analysis from sequence lists. Specifically, the MEME algorithm within the suite was used to uncover statistically significant, novel motifs within the provided RNA sequence dataset. RNA sequences were first preprocessed to remove repeated microexon sequences and then input into the MEME software with default parameters unless otherwise specified. Following motif discovery, the Tomtom tool, also part of the MEME Suite, was leveraged to compare identified motifs against a curated database of known RBP motifs. This enabled the identification of potential RNA-binding proteins that might recognize and bind to the discovered motifs in our sequences. Results were filtered based on recommended E-value thresholds. We performed this procedure for genes belonging to terms in each UMAP identified cluster of GO terms, reasoning that microexons involved in similar processes

might be under regulation by the same splicing factor. We reported the known Tomtom motifs which were significantly enriched for each cluster.

3.4.6 Positional Conservation Scoring

To assess the conservation of flanking intron sequences surrounding our annotated exons, we utilized PhastCons scores, which quantify the level of evolutionary conservation across multiple species. The exon-intron boundaries were identified based on existing genome annotations, and PhastCons scores were retrieved for each position in the intronic regions flanking the exons. For each exon, PhastCons scores were averaged over 100 nucleotides of the intronic sequence both upstream and downstream of the exon boundaries. These scores were extracted from the precomputed PhastCons files, available from the UCSC Genome Browser, using a custom Python script that employed the PyBigWig library for efficient querying of bigWig files. This approach allowed for a quantitative assessment of the conservation level of intronic regions adjacent to the exons of interest.

3.4.7 Frequency of Amino Acids Coded by Microexons

To quantify the frequency of coding exons within a given set of genomic annotations, we leveraged the Ensembl BioMart database, an online resource offering genome annotation data. Subsequent analyses were conducted using custom Python scripts that employed the Pandas library for data manipulation. The genomic intervals for the coding exons were retrieved and the frame start and end for the exon according to Ensembl was used to translate the nucleotide sequence into the expected protein sequence.

3.4.8 UTR Microexon Locations

To determine whether a given exon is located in the untranslated region (UTR) of its corresponding transcript, we first retrieved the transcript annotation data, which included the start and end coordinates for each exon, as well as the coordinates delineating the coding sequence (CDS) within the transcript. This information was obtained from the Ensembl database via the BioMart interface, and downloaded in a tab-delimited format for further analysis. Subsequent computational operations were performed using custom Python scripts, employing the Pandas library for data manipulation. For each exon in question, its genomic coordinates were compared with the start and end coordinates of the CDS of its corresponding transcript. An exon was classified as being part of the 5' UTR if its end coordinate was upstream of the CDS start coordinate. Similarly, it was classified as part of the 3' UTR if its start coordinate was downstream of the CDS end coordinate. Exons that overlapped with the CDS were designated as coding exons. This approach enabled us to systematically categorize each exon within its transcript context, thus elucidating its functional role as either coding or untranslated.

3.4.9 Microexon sQTL Detection

We performed sQTL analysis using our calculated microexon PSI values as the quantitative trait. Where possible, we followed the protocol outlined for the sQTL analyses performed in GTEx release v8. The genotype data used for sQTL analyses in release V8 was based on WGS from 838 donors, which all had RNA-seq data available in V8. We retained all variants with MAF 5%.

QTL Mapping using FastQTL cis-sQTL mapping was performed using tensorQTL, a

GPU-enabled implemetation of FastQTL (Taylor-Weiner, Aguet et al. 2019., Ongen et al. 2016). We tested all SNPs within a cis-region of 1Mb upstream and downstream of the microexon center coordinates. The GTEx consortium provided precomputed genotype principle components. As in GTEx, the adaptive permutations mode was used with the setting "–permute 1000 10000". To control for hidden confounders, we used PEER (Probabilistic Estimation of Expression Residuals). PEER was run on the PSI values to generate a set of covariates that capture unknown sources of variance in the phenotype data. This method allows for a more accurate model by accounting for latent variables that may influence PSI but are not captured by the provided covariates such as age and sex. These were combined with the first 15 principle components calculated for the GTEx genotype data as were used in the previous study. Nominal p-values were generated for each variant-gene pair by testing the alternative hypothesis that the slope of a linear regression model between genotype and expression deviates from 0. Beta distribution-adjusted empirical p-values from FastQTL were used to calculate q-values (Storey Tibshirani, PNAS, 2003), and a false discovery rate (FDR) threshold of 0.05 was applied to identify significant microexons.

3.4.10 sQTL Colocalization Analysis

To perform colocalization analysis of our identified sQTLs with summary statistics from 86 different Genome-Wide Association Studies (GWAS), we utilized the coloc R package. This package employs a Bayesian framework to assess the probability that two sets of association signals arise from the same causal variant. The sQTL SNP effect sizes, standard errors, and p-values from fastQTL were input with the GWAS summary statistics from each of the 86 traits. The GWAS statistics were preprocessed to ensure SNP

identifiers and genomic coordinates matched those in our sQTL dataset. We then executed the 'coloc.abf' function for each pair of sQTL, tissue and GWAS datasets to calculate the posterior probabilities for five different hypotheses capturing whether the p-values indicate a significant SNP for the sQTLs, for the GWAS SNPs, for both in different loci, for both in the same loci, or neither. The key outcome measured is the posterior probability for a shared causal variant (PP4). A threshold of PP4 > 0.8 was used to define significant colocalization between sQTLs and GWAS hits.

3.4.11 Differential Splicing of Microexons in DM1 Muscle

In order to identify differentially spliced exons between two sets of samples, we employed rMATS (replicate Multivariate Analysis of Transcript Splicing), a robust tool specifically designed for the detection of alternative splicing events from RNA-Seq data. First, we obtained RNA-Seq data from skeletal muscle biopsy samples of individuals with Myotonic Dystrophy Type 1 (DM1), Congenital Myotonic Dystrophy (CDM) or adult and pediatric controls (GEO accession GSE201255). Fastq reads from this study were run through our microexon detection pipeline (see chapter 2). For the two comparative groups within rMATS, we used only the adult disease and control samples. We configured the tool to detect five primary types of alternative splicing events: skipped exons (SE), mutually exclusive exons (MXE), alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), and retained introns (RI). The program then quantified the inclusion levels of exons in each sample and performed statistical tests to discern differences in splicing patterns between the two groups. Differential splicing events were deemed significant based on a user-defined false discovery rate (FDR) threshold, which we set at 0.05 for this analysis. We also performed unsupervised clustering of the microexon PSI values from all four sample groups

using the scikit-learn implementation of tSNE.

References

A, Mu et al. (Dec. 2020). "Lysine acetylation of cytoskeletal proteins: Emergence of an actin code". en. In: *J. Cell Biol.* 219.12.

Bailey, Timothy L and Charles E Grant (2021). "SEA: Simple enrichment analysis of motifs". In: *BioRxiv*, pp. 2021–08.

Barbeira, Alvaro N et al. (May 2018). "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics". en. In: *Nat. Commun.* 9.1, p. 1825.

Caillet-Boudin, Marie-Laure et al. (Jan. 2014). "Brain pathology in myotonic dystrophy: when tauopathy meets spliceopathy and RNAopathy". en. In: *Front. Mol. Neurosci.* 6, p. 57.

Clark, Michael B et al. (Jan. 2020). "Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain". en. In: *Mol. Psychiatry* 25.1, pp. 37–47.

Degener, Max J F et al. (Mar. 2022). "A comprehensive atlas of fetal splicing patterns in the brain of adult myotonic dystrophy type 1 patients". en. In: *NAR Genom Bioinform* 4.1, lqac016.

Frankish, Adam et al. (Jan. 2021). "GENCODE 2021". en. In: *Nucleic Acids Res.* 49.D1, pp. D916–D923.

Garrido-Martın, Diego et al. (Feb. 2021). "Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome". en. In: *Nat. Commun.* 12.1, p. 727.

Gonatopoulos-Pournatzis, Thomas et al. (Nov. 2018). "Genome-wide CRISPR-Cas9 interrogation of splicing networks reveals a mechanism for recognition of autism-misregulated neuronal microexons". en. In: *Mol. Cell* 72.3, 510–524.e12.

GTEx Consortium et al. (Oct. 2017). "Genetic effects on gene expression across human tissues". en. In: *Nature* 550.7675, pp. 204–213.

Irimia, Manuel et al. (Dec. 2014). "A highly conserved program of neuronal microexons is misregulated in autistic brains". en. In: *Cell* 159.7, pp. 1511–1523.

Klinck, Roscoe et al. (Sept. 2014). "RBFOX1 cooperates with MBNL1 to control splicing in muscle, including events altered in myotonic dystrophy type 1". en. In: *PLoS One* 9.9, e107324.

Kuleshov, Maxim V et al. (July 2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". en. In: *Nucleic Acids Res.* 44.W1, W90–7.

Lago, Santiago G and Sabine Bahn (Mar. 2022). "The druggable schizophrenia genome: from repurposing opportunities to unexplored drug targets". en. In: *NPJ Genom Med* 7.1, p. 25.

Lee, Jeesan et al. (Dec. 2020). "QUAKING Regulates Microexon Alternative Splicing of the Rho GTPase Pathway and Controls Microglia Homeostasis". en. In: *Cell Rep.* 33.13, p. 108560.

Li, Yang I et al. (Jan. 2015). "RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts". en. In: *Genome Res.* 25.1, pp. 1–13.

O'Leary, Nuala A et al. (Jan. 2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". en. In: *Nucleic Acids Res.* 44.D1, pp. D733–45.

Parada, Guillermo E et al. (Jan. 2021). "MicroExonator enables systematic discovery and quantification of microexons across mouse embryonic development". en. In: *Genome Biol.* 22.1, p. 43.

Poli, Jérôme, Susan M Gasser, and Manolis Papamichos-Chronakis (Oct. 2017). "The INO80 remodeller in transcription, replication and repair". en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372.1731.

Smith, Jean A et al. (Apr. 2020). "FXR1 splicing is important for muscle development and biomolecular condensates in muscle cells". en. In: *J. Cell Biol.* 219.4.

Taneja, K L et al. (Mar. 1995). "Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues". en. In: *J. Cell Biol.* 128.6, pp. 995–1002.

Tapial, Javier et al. (Oct. 2017). "An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms". en. In: *Genome Res.* 27.10, pp. 1759–1768.

Taylor-Weiner, Amaro et al. (2019). "Scaling computational genomics to millions of individuals with GPUs". In: *Genome biology* 20.1, pp. 1–5.

Torres-Méndez, Antonio et al. (Apr. 2019). "A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons". en. In: *Nat. Ecol. Evol.* 3.4, pp. 691–701.

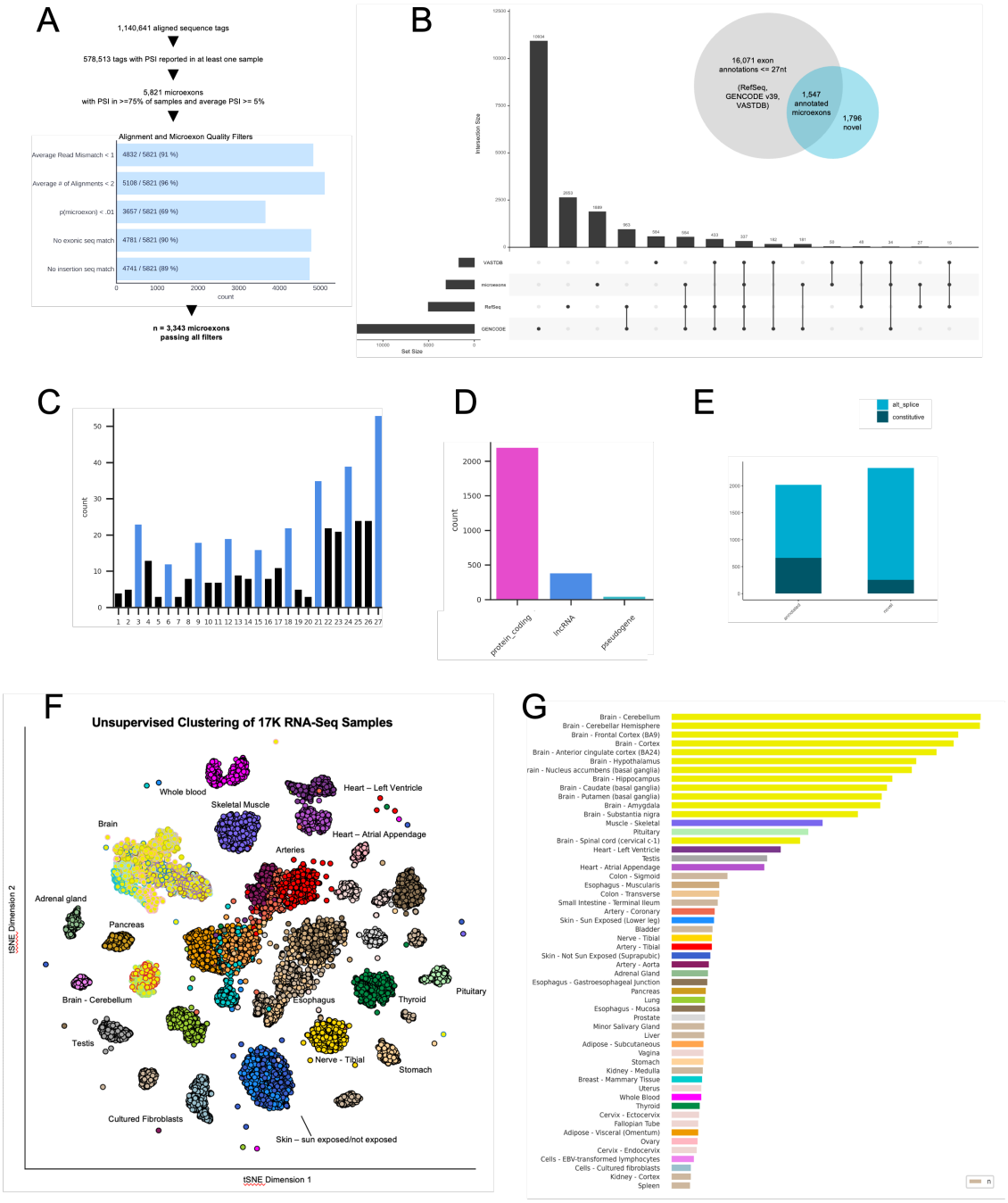


Figure 3.1: An Inventory of Microexons in 50 Adult Tissue Types: A) Filtering steps and final numbers for microexon quantification across GTEx samples. B) Comparison of microexons detected to microexon annotations in GENCODE, RefSeq, and VASTDB. C) Distribution of detected microexon lengths. D) Histogram of protein-coding vs. non-coding microexon-containing transcripts. E) t-SNE plot of RNA-Seq samples based on microexon PSI measurements. F) Number of tissue-specific microexons with inclusion >.75 in each GTEx tissue.

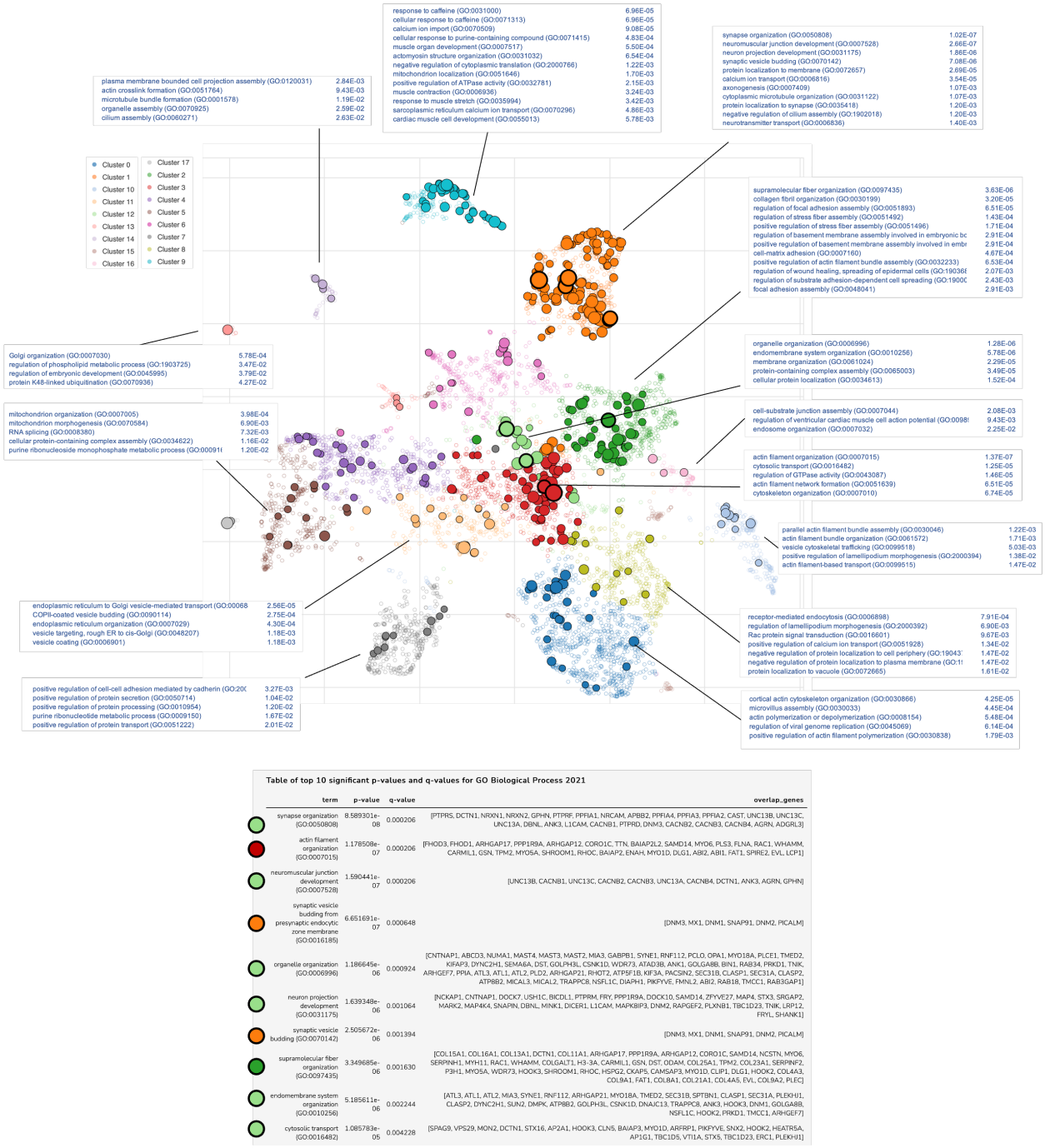


Table of top 10 significant p-values and q-values for GO Biological Process 2021

term	p-value	q-value	overlap_genes
synapse organization (GO:0050808)	8.589301e-08	0.000206	[PTPRS, DCTN1, NRXN1, NRXN2, GPHN, PTPRF, PPFIA1, NRCAM, APBB2, PPFIA4, PPFIA3, PPFIA2, CAST, UNC13B, UNC13C, UNC13A, DBNL, ANK3, L1CAM, CACNB1, PTPRD, DNM3, CACNB2, CACNB3, CACNB4, AGRN, ADGRL3]
actin filament organization (GO:0007015)	1.178509e-07	0.000206	[FHOD3, FHOD1, ARHGAP17, PPP1R9A, ARHGAP12, CORO1C, TTN, BAIAP2L2, SAMD14, MYO6, PLS3, FLNA, RAC1, WHAMM, CARMIL1, GSN, TPM2, MYO5A, SHROOM1, RHOC, BAIAP2, ENAH, MYO1D, DLG1, ABI2, ABI1, FAT1, SPIRE2, EVL, LCP1]
neuromuscular junction development (GO:0007528)	1.590441e-07	0.000206	[UNC13B, CACNB1, UNC13C, CACNB2, CACNB3, UNC13A, CACNB4, DCTN1, ANK3, AGRN, GPHN]
synaptic vesicle budding from presynaptic endocytic zone membrane (GO:0016185)	6.651691e-07	0.000648	[DNM3, MX1, DNM1, SNAP91, DNM2, PICALM]
organelle organization (GO:0006996)	1.186645e-06	0.000924	[CNTNAP1, ABCD3, NUMA1, MAST4, MAST3, MAST2, MIA3, GABPB1, SYNE1, RNF112, PCLO, OPA1, MYO18A, PLCE1, TMED2, KIFAP3, DYNC2H1, SEMA6A, DST, GOLPH3L, CSNK1D, WDR73, ATAD3B, ANK1, GOLGA8B, BIN1, RAB34, PRKD1, TNIK, ARHGEF7, PPIA, ATL3, ATL1, ATL2, PLD2, ARHGAP21, RHOT2, ATP5F1B, KIF3A, PACSIN2, SEC31B, CLASP1, SEC31A, CLASP2, ATP8B2, MICAL3, MICAL2, TRAPPC8, NSFL1C, DIAPH1, PIKFYVE, FMNL2, ABI2, RAB18, TMCC1, RAB3GAP1]
neuron projection development (GO:0031175)	1.639348e-06	0.001064	[NCKAP1, CNTNAP1, DOCK7, USH1C, BICDL1, PTPRM, FRY, PPP1R9A, DOCK10, SAMD14, ZFYVE27, MAP4, STX3, SRGAP2, MARK2, MAP4K4, SNAPIN, DBNL, MINK1, DICER1, L1CAM, MAPKBIP3, DNM2, RAPGEF2, PLXNB1, TBC1D23, TNIK, LRP12, FRYL, SHANK1]
synaptic vesicle budding (GO:0070142)	2.505672e-06	0.001394	[DNM3, MX1, DNM1, SNAP91, DNM2, PICALM]
supramolecular fiber organization (GO:0097435)	3.349685e-06	0.001630	[COL15A1, COL16A1, COL13A1, DCTN1, COL11A1, ARHGAP17, PPP1R9A, ARHGAP12, CORO1C, SAMD14, NCSTN, MYO6, SERPINH1, MYH11, RAC1, WHAMM, COLGALT1, H3-3A, CARMIL1, GSN, DST, ODAM, COL25A1, TPM2, COL23A1, SERPINF2, P3H1, MYO5A, WDR73, HOOK3, SHROOM1, RHOC, HSPG2, CKAP5, CAMSAP3, MYO1D, CLIP1, DLG1, HOOK2, COL4A3, COL9A1, FAT1, COL8A1, COL21A1, COL4A5, EVL, COL9A2, PLEC]
endomembrane system organization (GO:0010256)	5.185611e-06	0.002244	[ATL3, ATL1, ATL2, MIA3, SYNE1, RNF112, ARHGAP21, MYO18A, TMED2, SEC31B, SPTBN1, CLASP1, SEC31A, PLEKHJ1, CLASP2, DYNC2H1, SUN2, DMPK, ATP8B2, GOLPH3L, CSNK1D, DNAJC13, TRAPPC8, ANK3, HOOK3, DNM1, GOLGA8B, NSFL1C, HOOK2, PRKD1, TMCC1, ARHGEF7]
cytosolic transport (GO:0016482)	1.085783e-05	0.004228	[SPAG9, VPS29, MON2, DCTN1, STX16, AP2A1, HOOK3, CLN5, BAIAP3, MYO1D, ARFRP1, PIKFYVE, SNX2, HOOK2, HEATR5A, AP1G1, TBC1D5, VTI1A, STX5, TBC1D23, ERC1, PLEKHJ1]

Figure 3.2: GO Molecular Function Enriched Terms for 1,367 Alternatively Spliced Microexons: UMAP of GO Molecular Function 2021 terms. Significantly enriched terms are solid colored. Top ten significant terms have a bold border. Selected terms and p-values for each cluster are listed.

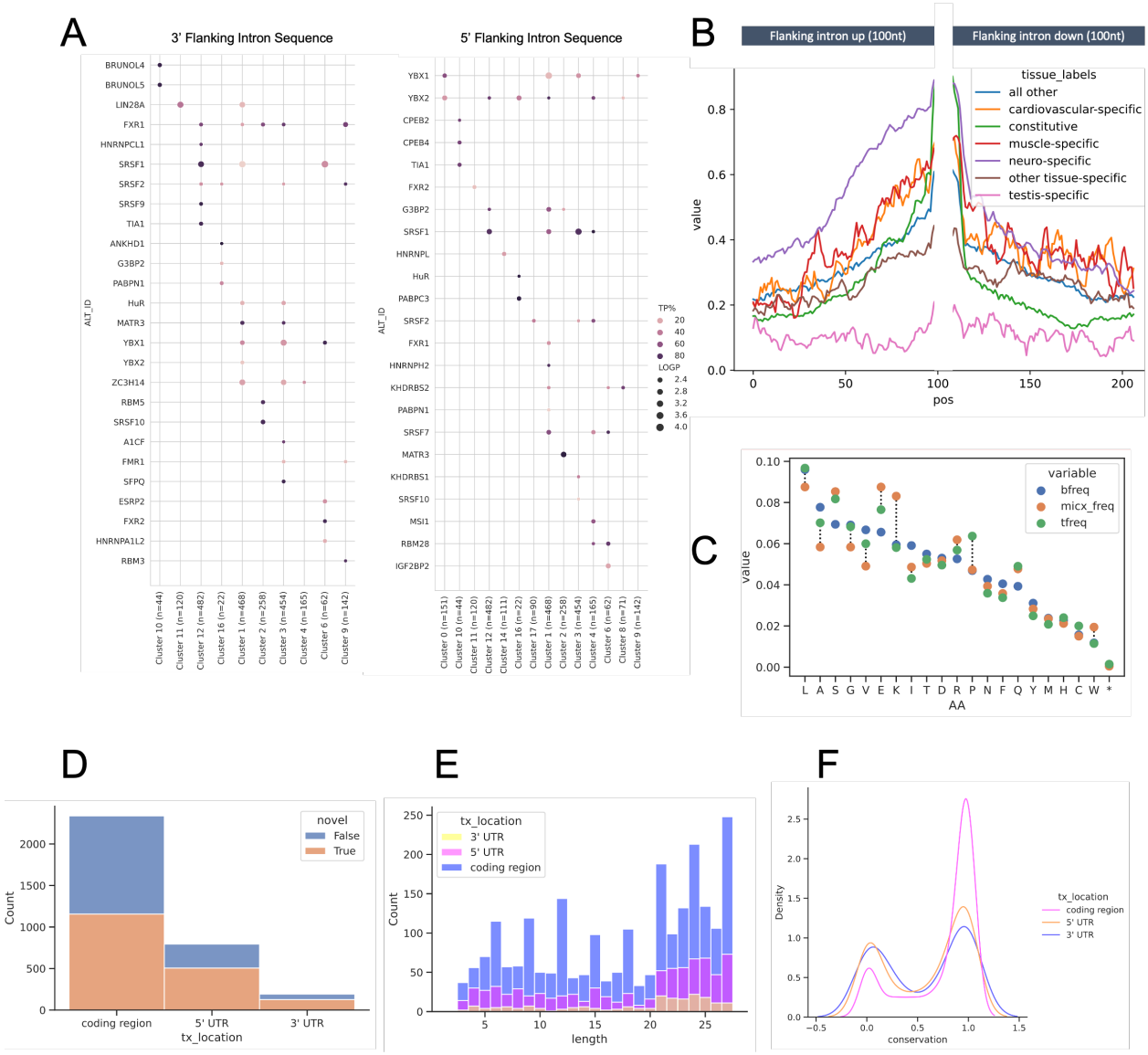


Figure 3.3: Microexons may affect RNA stability as well as protein sequence: A) Upstream and Downstream intronic sequences for microexons from each GO term cluster. SEA motif enrichment was performed for known RBP motifs. Color represents the percentage of microexon genes in that cluster containing a positive match for the motif. Size of the dot represents -logp value. B) PhastCons conservation scores averaged over flanking intronic sequences for microexon sequences. C) Histogram of protein coding and UTR microexons. D) UTR microexons do not exhibit a strong bias toward symmetric lengths. E) More than half of UTR microexons are highly conserved.

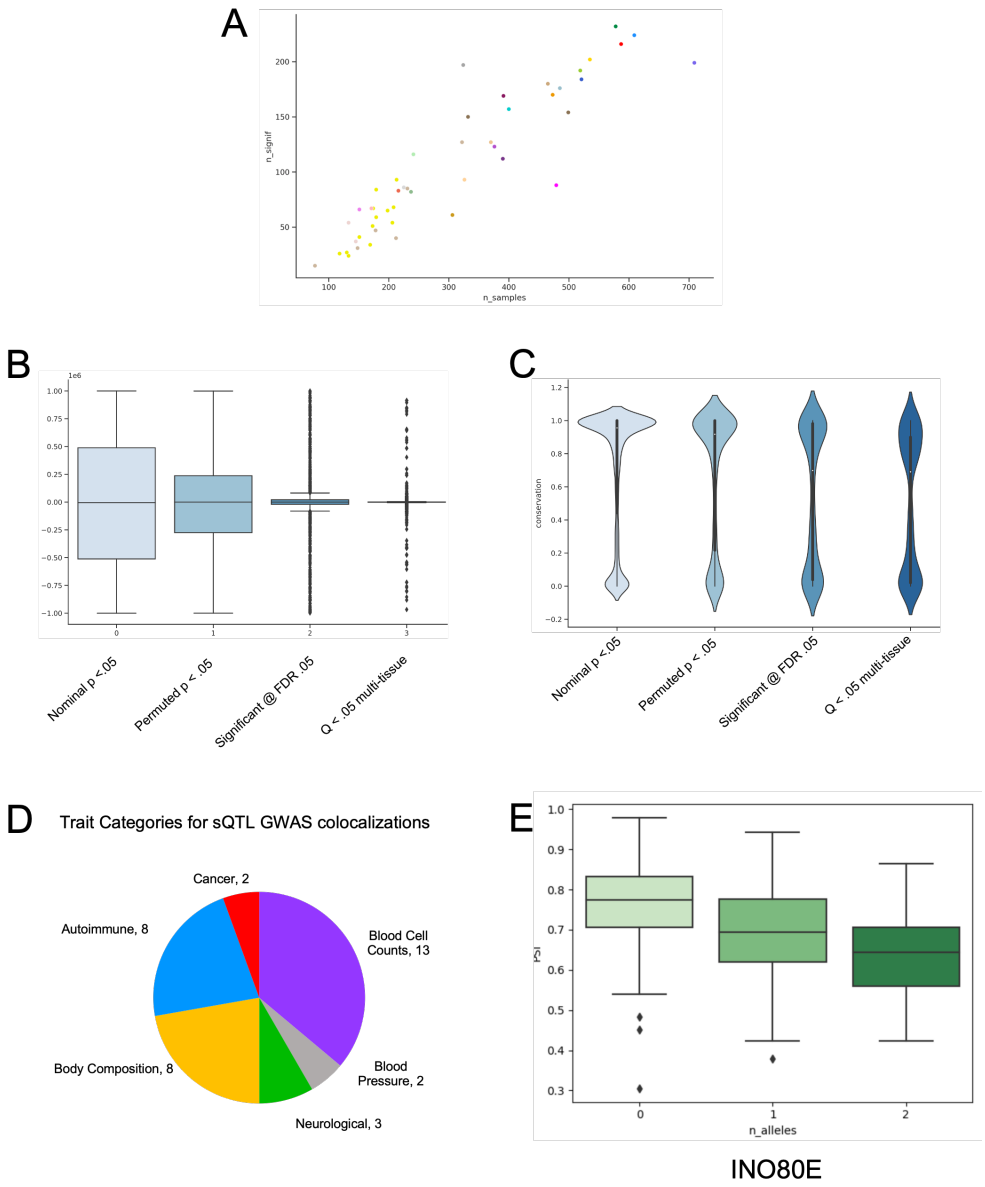


Figure 3.4: sQTL analysis and colocalization of diseases: A) plot of number of detected sQTLs vs sample number for each GTEx tissue. B) Plot of distance of sQTL to center of microexon for sQTLs of increasing significance score. c) Plot of conservation scores compared to increasing significance. D) Disease colocalizations for microexon sQTLs with harmonized GWAS summary statistics. E) sQTL colocalization for INO8E gene

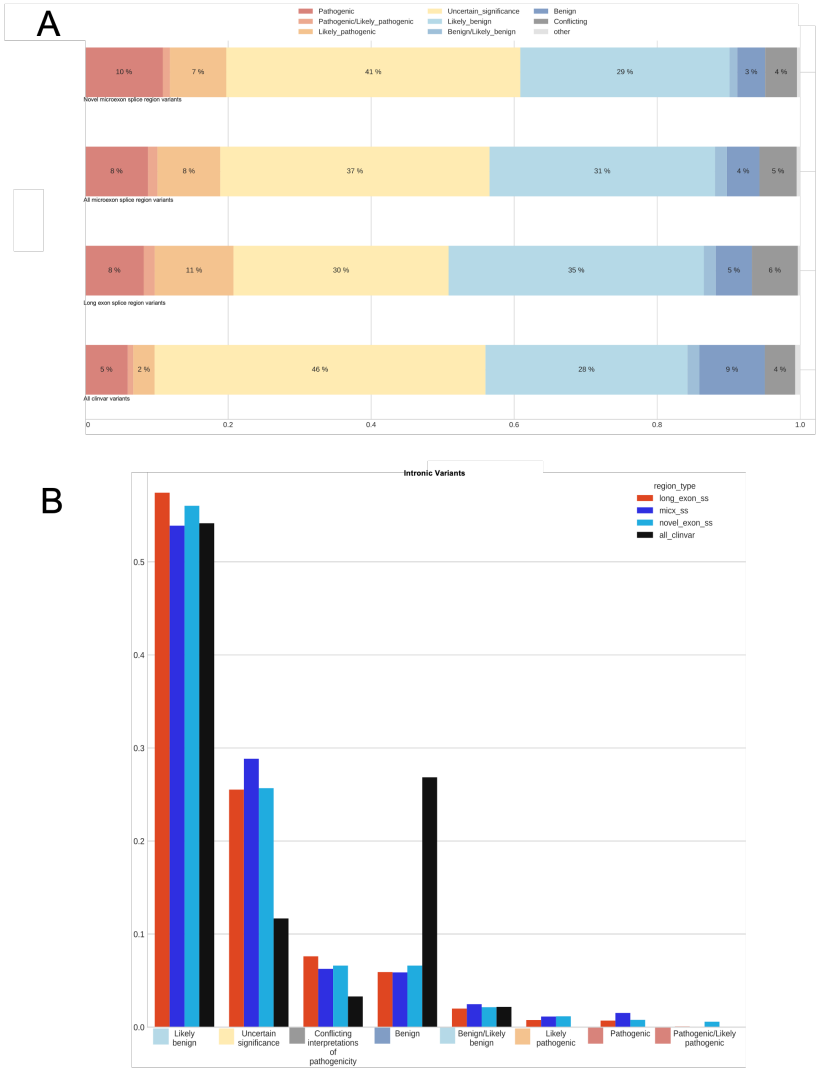


Figure 3.5: Microexon splice sites overlap with pathogenic rare mendelian disease variants : A) distribution of CLINVAR pathogenicity scores for microexons and other sequences. B) Histogram of each CLINVAR category for intronic variants.

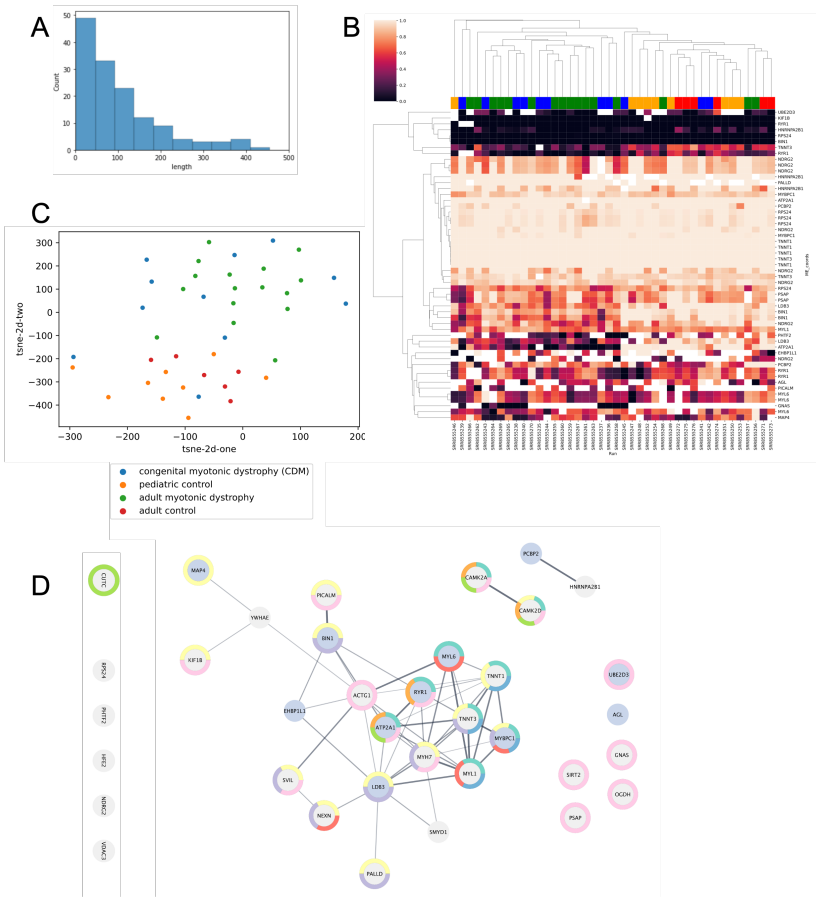


Figure 3.6: Microexon splicing is disrupted in Myotonic Dystrophy Type 1: A) Plot of lengths of differentially spliced exons from rMATS. B) t-SNE of DM1 and CDM samples and controls based on microexon inclusion. C) heatmap of microexon events differentially spliced in DM1. D) stringDB network of protein interactions and associations for differentially spliced microexons

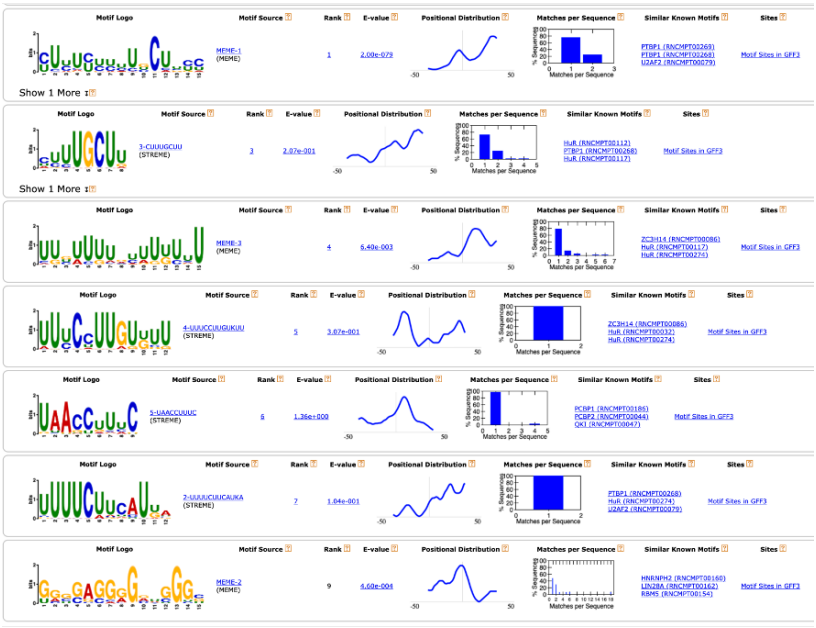


Figure 3.7: Motif enrichment of neuromuscular microexons: discovered enriched motifs for alternatively spliced exons in brain and muscle.

CHAPTER 4

Testing Splicing Variants Using a Massively Parallel Reporter Assay: Constructing A Reference Library of Human Exon Sequences

Collaboration Note:

This chapter discusses the work published in Cheung, Insigne, Yao, Burghard, Jones, et al. (2017), on which I am a co-author, with some results and figures reproduced for background. Rocky Cheung led the development of the MFASS experiments and Kim Insigne developed the computational methods. As a rotation student, I assisted with the computational methods, and performed a follow-up analysis to measure splicing of SDVs in GTEx RNA-Seq. The subsequent work in this chapter was performed by me, with guidance from Rocky Cheung in the Kosuri Lab and Jae Hoon Bahn in the Xiao Lab.

4.1 Background

Recent explosive growth in human populations has resulted in an abundance of rare variation which is expected to contribute to disease risk (Keinan and Clark (2012)). While most of the genetic variants in a given individual's genome are common, 1-4% of variants

will have a minor allele frequency less than or equal to .5% (1000 Genomes Project Consortium et al., 2015). Taken at the population level, however, the vast majority of variation is rare due to the contribution of unique alleles from a large number of individuals. Diagnosing individual rare variants and assessing their overall impact remains a key challenge in human genetics (Kremer et al., 2017; Frésard, Gamazon, Garrido-Martın, et al., 2017; Cummings et al., 2017; H. Lee et al., 2020).

Mendelian disease is often caused by variants that are very rare or absent in the general population, as natural selection limits the frequency of deleterious variants (Altshuler, Daly, and Lander, 2008; Cummings et al., 2017; H. Lee et al., 2020). Rare variation also plays a role in many complex diseases, though to what extent is largely unknown (Bomba, Walter, and Soranzo (2017), Ganna et al. (2018), and Wainschtein et al. (2022)). In the past, most studies have excluded variants below a minor allele frequency (MAF) of .01, as association tests require very large sample sizes to evaluate variants below this threshold (Mancuso et al. (2016)). However recent work highlights the importance of ultra-low frequency and *de novo* variation (X. Li, Kim, Tsang, Davis, Damani, Chiang, Hess, Zappala, Strober, Scott, et al., 2017b; Hernandez et al., 2019; Wainschtein et al., 2022. Whole-exome sequencing of 900 individuals demonstrated that nonsense and splice disrupting de novo variants are associated with autism and can have large individual effect size (Takata et al. (2016)). Rare variants also explain a substantial amount of heritability in prostate cancer (Mancuso et al. (2016)). Even in traits expected to be driven by common variation, rare variants of large effect have been observed. In schizophrenia, common variants with small effect sizes and rare variants with larger effect sizes have been identified (Arslan (2018) and Purcell et al. (2014)). In human height, a well studied, highly

polygenic trait predominantly driven by common variation, recent studies have shown rare variants can still have a substantial effect (Marouli et al. (2017)), (Yang et al. (2015)).

Until recently, studies attempting to quantify the collective contribution of rare variation have been limited, but decreased cost of sequencing and improved methodology has provided new insights (Bomba, Walter, and Soranzo (2017), Manolio et al. (2009), Cirulli and Goldstein (2010)). Large scale, multi-tissue RNA-Seq reveals that gene expression outliers are enriched for rare variants in all functional categories, with 58% and 28% of under- and overexpression outliers having a nearby conserved rare variant (X. Li, Kim, Tsang, Davis, Damani, Chiang, Hess, Zappala, Strober, Scott, et al. (2017a)). Additionally, population genetics simulations suggest that extremely rare variation could contribute more substantially to the heritability of gene expression than previously appreciated (Hernandez et al., 2019; Wainschtein et al., 2022). In studies of both collective and individual pathogenicity, it is of increasing importance to diagnose rare variation and its mechanisms.

One major mode through which single nucleotide variation can be deleterious is by altering splicing. One study estimates that one third of disease causing mutations disrupt splicing (Lim et al. (2011)). Additionally, it is estimated that the contribution of splicing-disrupting variation to disease risk is comparable to that of variation affecting gene expression (Y. I. Li et al. (2016)). Rare variants may be an important source of splicing disruption. In a study of Startgart disease (a type of juvenile macular degeneration), 913 variants from over 4,000 affected individuals were evaluated in a minigene assay of the ABCA4 gene, which is responsible for the disease. 44 of 47 rare variants caused significant loss of the correct splice isoform (Sangermano et al. (2018)).

Despite their importance, rare variation remains difficult to assess at scale. While protein-truncating variants or missense variants may be easier to detect, splice-disrupting variants (SDVs), especially those outside of the splice site region, remain challenging to predict due to the complexity of the splicing code; canonical splice site motifs alone predict only 50-60% of splicing (Yeo and Burge, 2004; Jian, Boerwinkle, and Liu, 2014). Various computational predictors have had some success, but rely heavily on conservation. Traditional minigene assays are low-throughput and cannot keep pace with the discovery rate of rare variants. One potential strategy has been the development of large-scale functional assays to test large numbers of potential splicing-disrupting variants *in vivo*. These methods allow us to directly query natural human variation, as well as ask questions about the sequence-function relationship of splicing.

Cheung, Insigne, Yao, Burghard, Jones, et al. (2017) developed, a massively parallel reporter assay that can directly test the effect of tens of thousands of variants on exon inclusion. We designed a splicing reporter construct where an exon and its immediate intronic sequence is positioned between two halves of a fluorescent reporter. Correct exon recognition results in inclusion of the exon, keeping the two halves of the reporter separated, while an exon skipping event results in reconstitution of the fluorescent molecule. This construct is stably integrated at a specific site in HEK-293T cells, resulting on one library member per cell. In this way, we are able to use flow-sorting and high throughput sequencing to determine the inclusion of tens of thousands of designed sequences. Inclusion ratios are calculated as weighted averages of sequence counts from each of the sorted bins.

While this is not the first splicing MPRA (Rosenberg et al., 2015; Adamson, Zhan, and Graveley, 2018; Soemedi et al., 2017a, our system has several advantages over existing

systems. Our amplicon sequencing technique allows us to sequence the full library sequence and thus screen both exonic and intronic mutations, unlike RNA-Seq based methods (Soemedi et al. (2017b)). Our reporter construct tests library sequences in a longer intron background, where short introns have been shown to be less robust. Finally, stable integration improves over transient assays more prone to overexpression effects (Smith and Lynch (2014)).

Our findings challenge previous expectations of splicing variants. Unexpectedly, we find that many (3.8%) of tested human variants cause large-scale splicing disruptions in our assay. We believe this points to an underappreciated role for rare variation to affect human trait diversity through splicing. We also find that large effect disruptions are caused by changes spread widely across exons and introns, rather than concentrated in the splice region. Nearly all of the 1,050 identified splice disrupting variants were rare or ultra-rare, with the overall percent of SDVs increasing with rarity. This represents a significant number of mis-splicing variants that likely are currently missed by variant effect prediction tools and would typically be excluded from disease studies. This supports the ongoing shift in paradigm to focus on the effects of rare variation in complex disease.

We now set out to test a much larger set of human exons and explore the effect of mutation on a broader set of human exon contexts. By identifying large numbers of validated large-effect rare variants, we can better assess the extent to which rare variation affects complex disease through mis-splicing. In addition, the identified SDVs may be useful to train better splicing predictors, inform rare variation association studies, and provide new insight into patterns of splicing disruption. We will assay over 80,000 human exon sequences (roughly one third of all human exon sequences) representing all internal, coding

exon sequences under 118bp. Compared with other published methods, this is the largest screen of variants tested by an order of magnitude, and we expect to be able to scale up even further in future experiments (Soemedi et al. (2017b)), (Adamson, Zhan, and Graveley (2018)).

Once we determine which exon sequences are included, we will test human SNVs from 500 individuals sequenced through the Genotype-Tissue Expression (GTEx) project. This will be the largest functional screen of human splicing variants to date and, importantly, we will be able to make in-depth comparisons to splicing in GTEx multi-tissue RNA-Seq data. This will pave the way to applying large functional screens like MFASS to detect splicing variants in disease cohorts with higher confidence, without the need for large RNA-Seq validation datasets.

We developed MFASS as a high-throughput system to test the effect of sequence variation on exon recognition. Next steps include expanding our investigation of rare variation by testing many more variants in an expanded set of human exon contexts, and performing more in-depth validation of SDVs using available RNA-Seq data. We will generate a large set of human variants from matched DNA- and multi-tissue RNA-Seq datasets and compare splicing disruption across tissues. We will also use this data to better understand how well our experimental model for exon recognition simulates splicing in endogenous contexts. This is important for interpreting results in MFASS and in all minigene reporter systems for splicing.

4.2 Results

4.2.1 MFASS: Rare Variation often disrupts splicing

In our recent published work, (Cheung, Insigne, Yao, Burghard, Wang, et al. (2019b)) we perform a screen of over 27,000 human variants sampled from the Exome Aggregation Consortion (ExAC) (Frankish et al. (2019)). Our results point to an underappreciated contribution of rare variation to human trait diversity by acting through splicing. We also demonstrate that due to the amount of human variation outside of the conserved splice sites, SDVs occur frequently outside of the splice region. In order to compare the behavior of SDVs in MFASS to their endogenous behavior, we look for variants present in individuals with transcriptome sequencing in GTEx. These results motivate further investigation of rare variants which affect splicing and our ability to detect them using MFASS.

We employed MFASS to test single nucleotide variants present in ExAC. We tested variants from 2,198 exon backgrounds which were measured to have an inclusion ratio of at least .8 in our system, with replicability in two different intronic backbones. From these backgrounds, we were able to test 27,733 variants, which represents roughly half (27,733/52,965) of all ExAC SNVs in these regions. Variants with inclusion of $\leq .2$ were classified as SDVs if there was agreement between two replicates (replicate IRs within .2.) Surprisingly, 3.8% of these variants caused nearly complete loss of exon recognition in MFASS. As expected, the percent of variants disrupting splicing is dependent on functional class (Figure 4.1C). Splice site and splice region are most sensitive to disruption, with 76% of splice site variants classified as SDVs (Figure 4.1C). This is considerably higher than other classes of variant. However, when we look at the location of SDVs detected overall

(Figure 4.1C), we see that the majority of variants fall outside of splice region (83%.) This leads to the important observation that due to the larger amount of human variation outside of the conserved splice sites, most SDVs fall outside of the canonical splice region. Splice Disrupting Variants are largely rare variants. We compare the allele frequency of SDVs and find that 98.8% of our SDVs are very rare (MAF $< .005$), and that the rate of SDVs increases with decreasing allele frequency, with extremely rare and singleton variation having the greatest percentage of SDVs (Figure 4.1E). This is consistent with population genetics theory which expects the frequency of deleterious variants to be constrained by negative selection. SDVs also occur in regions of significantly stronger evolutionary conservation compared to non-SDVs, and this enrichment is especially strong in introns where overall conservation is low. While conservation is an indicator of function, it is not sufficient to predict splice disruption due to most human variation occurring outside of the most highly conserved sites.

We compared our SDV calls with those from several modern variant effect predictors designed to evaluate non-coding variation, including CADD, phastCons, SPANR, and others. These computational tools were applied to our list of SDVs. We generally saw low agreement between predictors, with no tool calling more than a small percentage of SDVs deleterious. SPANR, a method specifically designed to evaluate splicing variants, performed best, calling 11.8% of SDVs with 44.5% precision (Cheung, Insigne, Yao, Burghard, Jones, et al., 2017). Our results indicate that current variant effect predictors are unlikely to reliably detect splicing variants outside of the splice region, highlighting the need for functional testing of variants.

4.2.2 RNA-Seq validation reveals differences in splicing patterns in MFASS and native contexts.

We asked whether SDVs in MFASS caused splice disruption in their endogenous context. In order to evaluate this, we overlapped our 1,050 SDVs with variants from individuals sequenced through the Genotype-Tissue Expression project. As most of the SDVs are rare, there was limited overlap, but we were able to evaluate the effects of 28 SDVs in multi-tissue RNA Seq. 7 of these 28 were significant when PSI values for the containing exon were compared in individuals with and without the SDV by Mann-Whitney U test (at an FDR of 5%) including 3 strong loss of function variants across all tissues (Figure 4.2A-G). We also see some evidence of tissue-specific loss of splicing (Figure 4.2H-I), with two SDVs which had a marked effect in relevant tissue types. Low concordance with GTEx RNA-Seq could be due to a number of factors. As we can only validate variants which occur in both our ExAC sample and GTEx, we may be enriching for common variants less likely to have a large effect size. Our three strongest validated SDVs are present in only 1-2 individuals in GTEx and are of the rarest tested. Additionally, PSI values may not capture alternative splicing events occurring endogenously such as alternative splice site usage. Another source of underestimation is that PSI calculated from GTEx RNA-Seq may lack coverage in many cases to detect smaller effect sizes observed in vivo. On the other hand, our assay may overestimate the number of splice disrupting variants as minigene reporters may be less robust to splice disruptions than endogenous sequences due to the shorter intron context and non-native intron sequence of the reporter construct. Trans splicing factors also contribute to highly cell-type specific splicing patterns differing from the HEK-293 cells we used. These results

motivate further exploration of splice disrupting variants in varying tissue types.

4.2.3 A Reference Library of Human Exon Sequences for MFASS

In order to test the effects of rare variation on splicing, we expanded our study to a broader range of exon contexts. We designed a library to test the inclusion of all testable human exon sequences in our system. We tested 82,000 internal, coding exon sequences under 108bp in length, which represents roughly a third of all human exons (Figure 4.3 3). We can then select GTEx variants for exons we know are included in our system, and have supporting RNA-Seq data with which to compare. We would like to enrich for splice disrupting variants as the identification will be useful for determining underlying mechanisms and important individual instances of splicing disruption. These experiments can help to understand how well our experimental model for exon recognition simulates endogenous RNA-Seq datasets. This will allow for better interpretation of mutations in varied cellular contexts.

We generated a library of all human exon sequences from the GENCODE human reference annotation which are testable in MFASS (internal, coding exon sequences under 108bp long) (Consortium et al. (2015)). In designing the library, we first took all GENCODE exons, filtered for protein-coding transcripts, and for exons small enough to include enough intronic sequence context when synthesized on the microarray. This ultimately left us with 106,424 unique sequences for 250-mer designed sequences (Figure 4.3C). We applied MFASS to these sequences, and generated a list of exons which are fully included in our system. As in Cheung et al., this is done by sorting cells into three bins of high, intermediate, and low GFP fluorescence, and calculating the inclusion ratios as described

(Cheung, Insigne, Yao, Burghard, Wang, et al. (2019b)). We then can look for candidate SNVs to test which overlap these exons.

The library was roughly broken down into four subpools to be processed in parallel. This creates smaller libraries of more manageable size for transfection, as well as allows us to check for differences in exon inclusion between symmetric (exons which have a length of a multiple of 3) and asymmetric exons. The start and end phase of library members is given in Figure 4.3B. We used this information to create four subpools of roughly even size, with in-phase and out-of-phase exon groups split in two again by strand (Figure 4.3). After cloning, transfection, sorting, and sequencing of our pilot library, we found that we had captured roughly 40% of the designed sequences from our original library, though the proportions of captured sequences from each subpool varied. This could be due to inconsistencies in handling the subpools. We noticed a higher than expected level of cell death after lipofectamine treatment of subpool 3, which may have led to a bottleneck and loss of sequence diversity.

4.2.4 Comparing Reference Library to Original MFASS Results

In order to assess the reproducibility of MFASS, we compare our results from the reference library with sequences that were chosen in the first MFASS SDV reference library. We found several differences between the results obtained for the reference library and the original SRE and SDV libraries in the MFASS paper. First, we noted that we had many more exons with intermediate splicing measurements, whereas the sequences were more bimodal in MFASS (Figure 4.4C). This could be due to a number of reasons, but we note that we used a titration to choose the level of antibiotic selection and observe more

mCherry negative cells in the sorts. We also performed one less passage per sample than was done in the original paper. This might have led to capturing more intermediate members that have lower frequencies in the pools.

In comparison with MFASS results, we found that 1,036 exons were measured in both the original and new libraries. The first MFASS library only tested in-phase exons, so all these sequences are in-phase. We find that the inclusion ratios of the same sequences performed in the two experiments had a Pearson's r of .51, and a tetrachoric correlation .77. This is compared to r=.89, and tetrachoric of .97 between MFASS replicates. The slight differences in handling and passaging may have led to lower agreement with these two different runs, but overall we see good agreement between the two batches.

4.2.5 The Relationship Between Length and Exon Inclusion in MFASS

While this reference library was primarily intended to be used as a jumping off point to provide valid exons for which outlier variants could be tested, we can also examine trends based on the properties of these natural sequences. Having an interest in microexon splicing, we wanted to know whether any short exons could be recognized and spliced in the MFASS system. Remarkably, we see two strong trends in the length vs. inclusion plot, shown in Figure 4.5 First, we see that the proportion of exons that are recognized at each length seems to increase linearly up until about 50 nt long, and then stabilizes at 40-60%. We also see that symmetric exons are included roughly 20% more often. This agrees with previous results that showed that exons under 50 nt are less likely to be included. The asymmetric exons may be more frequently subject to because frameshifts can lead to added stop codons leading to nonsense mediated decay (NMD).

4.3 Discussion

The development of massively parallel reporter assays (MPRAs) has expanded our capacity to probe diverse biological systems at unprecedented scale, empowering researchers to dissect complex regulatory frameworks like splicing. Our MFASS platform has yielded an unexpected number of splice-disrupting variants (SDVs), a finding that is both fascinating and challenging. On the one hand, the identification of numerous SDVs opens up new avenues for understanding the mechanism of splicing and its impact on cellular function. On the other hand, the large number of detected SDVs raises concerns about their true biological relevance, as well as the replicability of these findings across different assays and biological contexts. The landscape of our SDVs highlights the potential of less-contrained intronic mutations, which are present in greater numbers, to disrupt splicing in individuals. These variants are likely missed in most studies.

Our follow-up study provides a comprehensive assessment of naturally occurring human exon sequences, significantly broadening the scope of exon contexts previously investigated. We designed and tested a library of around a third of all human internal, coding exon sequences shorter than 108 bp. This approach not only sets the stage for a more nuanced understanding of splice-disrupting variants but also helps elucidate underlying mechanisms that govern splicing disruptions. Our library brings attention to important natural sequence properties affecting splicing. Notably, our findings corroborate previous studies suggesting that shorter exons, especially those under 50 nt, are less likely to be included in the final transcript. This linear increase in exon recognition up to around 50 nt could reflect a natural boundary condition in the spliceosome machinery, and further studies may help to elucidate the biochemical basis of this phenomenon. Additionally, the observation

that symmetric exons are more frequently included raises questions about the consequences of exon symmetry on splicing efficiency. The decreased inclusion rates for asymmetric exons could be related to a higher susceptibility to nonsense-mediated decay (NMD), possibly due to frameshifts leading to premature stop codons.

However, our study is not without limitations. The obvious being that our library, while extensive, only covers the shortest one-third of all human exons, constrained by the 108 bp length cut-off. We may miss additional complexity of splicing events that occur in untranslated regions or non-coding RNA. One of the fundamental caveats is that splicing is highly context-dependent, both at the cellular and tissue levels. As such, the extent to which our MPRA findings can be generalized to other biological settings remains uncertain. This leads us to crucial questions for interpreting splice variants: if a variant disrupts splicing in our assay, can we expect the same in an endogenous context? How does this behavior translate across different life stages, such as embryonic versus adult, or across different cell types like neurons and muscles, which are known to have tissue-specific splicing patterns? Expanded libraries of variants corresponding to splicing outliers in RNA-Seq data may prove useful in understanding the applicibility of such assays to rare variant studies. Furthermore, additional cell lines from tissues with distinct splicing patterns may also be useful. In conclusion, our expanded exon library provides a greatly expanded landscape for studying the impact of rare genetic variations on splicing. Our library serves as a solid foundation for future work targeting outlier variants, furthering our understanding of splicing mechanisms and their disruptions in human disease.

4.4 Methods

4.4.1 Human Exon Reference Library Design

The reference human exon library was designed to represent as many annotated exonic sequences from the human genome as possible. We chose to use the the GENCODE Basic/CCDS annotation as a representative and non-repetitive collection of exons. The Consensus Coding Sequence Project (CCDS) is a list of transcript-coding genomic sequence regions that are annotated in both RefSeq and Ensembl/GENCODE, and which are subject to extensive manual review. GENCODE Basic "Identifies a subset of representative transcripts for each gene; prioritises full-length protein coding transcripts over partial or non-protein coding transcripts within the same gene, and intends to highlight those transcripts that will be useful to the majority of users" (Frankish et al., 2019). Exon sequences were retrieved using the Bioconductor BiomaRt package in R (Durinck, Moreau, et al., 2005; Durinck, Spellman, et al., 2009). Sequences were obtained from GENCODE version 30 in March 2018.

Sequences were additionally filtered to be compatible with the reporter construct design used in Cheung, Insigne, Yao, Burghard, Wang, et al., 2019a. Based on this design, we selected all human exons that were small enough to allow for at least 40bp upstream 30bp downstream of intronic context centered on the exon. This is based on size contraints of the length of commerically available oligonucleotide pools, and the density of intronic splicing sequence elements. For shorter exons, length was increase symmetrically. Our sequence library was ordered as 230bp microarray-derived oligonucleotides from Agilent Technologies. This allowed us to include exon sequences that were a maximum of 118nt

long, to allow for the minimum length flanking intronic context, as well as the addition of amplification primer sequences, and restriction sites. We also required that the exon sequences be internal, coding exons. We filtered exons partially or completely within UTR regions of transcripts based on their Ensembl biomart annotations of start and end frame (-1 indicating a non-coding start or end) (Cunningham et al., 2022).

We also filtered sequences that contained the restriction sites used in our reporter, namely AgeI and NheI. We separated the library into four subpools to be processed independently. This was mainly to ensure that adequate coverage of library members could be achieved for estimated numbers of cells cloned and transfected. As was noted in Cheung, Insigne et al., nonsense mediated decay of frameshifted sequences can not be directly measured from our assay, so we decided to separate the sequences into subpools based on exon sequences that were in-frame, and those which would cause a frameshift. We then divided these groups again based on positive and negative gene strand (which is not expected to affect splicing) to create four roughly evenly sized subpools. We included 800 control sequences to act as negative controls. These sequences were generated as in Cheung, Insigne et al. by substituting the canonical splice site nucleotides (AG/GT) such that splicing would be predicted to be completely disrupted. We also selected 200 exon sequences to be included in every subpool to allow for comparison of splicing across subpools. 50 sequences each were selected randomly from each subpool to be included.

4.4.2 MFASS Reporter System

Splicing Reporter Construct

We used the splicing reporter design developed in Chueng, Insigne et al. For a more

complete description of its design, please see the methods of that paper. Briefly, a three-exon, two-intron reporter construct is used. The middle exon and its surrounding intronic context are varied, while the remaining introns and flanking exons are fixed. The flanking exons are the emerald GFP (emGFP) coding sequence which has been split. The intron sequences are longer intron sequences which have been utilized in splicing studies. Of the two constant intron backbones tested in Cheung, Insigne et al., we chose to use the SMN1 intron backbone as it is a human sequence, and the splicing results from SMN1 and DHFR introns from textitC. griseus were noted to have shown high agreement. We used the SMN1 intron backbone plasmid RCP125 for cloning of the human reference exon library. This plasmid was used with engineered landing pad HEK293T cell line RCA7 (Cheung, Insigne et al.) to stably integrate the reporter construct at the AAVS1 locus.

Library Amplification and Cloning

The library was synthesized by Agilent Technologies and was resuspended in 100 uL of TE, which was then diluted in a 1:25 ratio. The diluted library was amplified using Q5 High-Fidelity 2x Master Mix in 50 uL reactions, and amplified with subpool-specific primers- subpool 1 with oRC280, oRC284; subpool 2 with oRC281, oRC285; subpool3 with oRC279, oRC283; and subpool 4 with oRC282, oRC286. The reaction and cycling conditions are: 98C for 2 min, 5 cycles of 95C for 3 s, 50C for 20 s, 60C for 10 s, 15 cycles of 95C for 3 s, 60C for 30 s, followed by an extension of 60C for 5 min.

After amplification the library members and RCP125 vector were digested overnight with ageI-HF and nheI-HF (New England Biolabs). Digested PCR products were run on a 1% TAE agarose gel and approximately 180bp amplicons were extracted using a Zymoclean Gel DNA Recovery Kit. We noted that these sticky ends appeared to be highly heat-sensitive,

and isolating fragments using gel electrophoresis needed to be performed at 70V or lower on a 1% gel for successful ligation. The vector and library were ligated with T4 DNA ligase (NEB) overnight at 37C. Cloning was performed with 10-beta electrocompetent textitE. coli, and plated on LB + kanamycin. Following overnight growth, cells were scraped and resuspended in liquid LB medium, and then roughly 800 million cells were inoculated in 450mL LB + kan in a 1L flask and grown overnight before plasmid purification.

HEK293T Cell Lines and Transfection

The RCA7 HEK293T chromosomal landing pad cells from Cheung, Insigne et al. were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 0.1 mg/mL streptomycin. For each subpool we transfected at 80% confluency with plasmid and Bxb1 integrase using Lipofectamine 3000 (Thermo Fisher Scientific). 10ug of DNA, 2.5ug BxbI plasmid, and 24uL of Lipofectamine 3000 reagent were added. Vortex, let stand 3 min and add 24uL DNA-lipid complex, rest 30 min. The reaction was added to each of 10 10cm plates per subpool, containing 500uL Opti-MEM (Thermo Fisher Scientific) media. Cells were transfected for 72 hours, and then selected with puromycin. Passages were performed serially for 10 days before cell sorting.

Fluorescence-activated cell sorting and DNA-Seq

We measured cell samples for GFP and mCherry fluorescence intensities by flow cytometry. As in the previous experiments, cells which met a threshold level of mCherry flourescence (indicating expression of the reporter) were binned based on high- intermediate- or low-GFP fluorescence. We sorted on a FACSAria III (BD Biosciences) machine into three bins based on GFP fluorescence. Binning profiles were based on saved profiles from original

MFASS runs, as well as mCherry and GFP single color controls. We sorted approximately 1 million cells each for the reference library high- and low- GFP bins, and .25 million cells for the intermediate min. Based on the stability of fluorescence observed in the original MFASS libraries, we opted not to perform a second sort.

DNA-Seq of FACS-sorted libraries

Sorted cells were grown again to 80% confluency. We extracted genomic DNA using the blood and cell culture DNA midi kit from Qiagen and amplified each subpool using the forward primer oCB214 and reverse primer 1r in a number of 50uL proportional to each bin size. We found that Taq DNA polymerase (Thermo Fisher Scientific) produced more specific gDNA target amplification over Q5. Per bin, we amplified library variants using 5 ug of gDNA under the following conditions: 98°C for 45 seconds, followed by 30 cycles using: 98°C for 15 seconds, 62°C for 30 seconds, 72°C for 30 seconds, followed by an extension of 72°C for 1 minute.

4.4.3 Splicing Quantification

Reference exon library data was generated from Illumina HiSeq 150-bp paired-end sequencing runs. Reads were trimmed and aligned to perfect and 1-off sequence tags for all library members using BBMerge. See Cheung, Insigne et al. for more detailed descriptions of the scripts used to calculate the inclusion ratios of library members. In keeping with the original paper, we require at least 5 reads across bins to calculate an inclusion ratio for a given library member. A normalized read count is calculated to account for the percentage of reads sorted and the reads per million for each library member is calculated by . The inclusion ratio for the given bin counts is calculated according to the formula .

References

1000 Genomes Project Consortium et al. (Oct. 2015). "A global reference for human genetic variation". en. In: *Nature* 526.7571, pp. 68–74.

Adamson, Scott I, Lijun Zhan, and Brenton R Graveley (2018). "Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency". In: *Genome biology* 19, pp. 1–12.

Altshuler, David, Mark J Daly, and Eric S Lander (2008). "Genetic mapping in human disease". In: *science* 322.5903, pp. 881–888.

Arslan, Ayla (2018). "Imaging genetics of schizophrenia in the post-GWAS era". In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 80, pp. 155–165.

Bomba, Lorenzo, Klaudia Walter, and Nicole Soranzo (2017). "The impact of rare and low-frequency genetic variants in common disease". In: *Genome biology* 18.1, pp. 1–17.

Cheung, Rocky, Kimberly D Insigne, David Yao, Christina P Burghard, Eric M Jones, et al. (2017). "Many rare genetic variants have unrecognized large-effect disruptions to exon recognition". In: *bioRxiv*, p. 199927.

Cheung, Rocky, Kimberly D Insigne, David Yao, Christina P Burghard, Jeffrey Wang, et al. (Jan. 2019a). "A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions". en. In: *Mol. Cell* 73.1, 183–194.e8.

— (2019b). "A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions". In: *Molecular cell* 73.1, pp. 183–194.

Cirulli, Elizabeth T and David B Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing". In: *Nature Reviews Genetics* 11.6, pp. 415–425.

Consortium, GTEx et al. (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans". In: *Science* 348.6235, pp. 648–660.

Cummings, Beryl B et al. (2017). "Improving genetic diagnosis in Mendelian disease with transcriptome sequencing". In: *Science translational medicine* 9.386, eaal5209.

Cunningham, Fiona et al. (Jan. 2022). "Ensembl 2022". en. In: *Nucleic Acids Res.* 50.D1, pp. D988–D995.

Durinck, Steffen, Yves Moreau, et al. (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". In: *Bioinformatics* 21.16, pp. 3439–3440.

Durinck, Steffen, Paul T Spellman, et al. (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". In: *Nature protocols* 4.8, pp. 1184–1191.

Frankish, Adam et al. (2019). "GENCODE reference annotation for the human and mouse genomes". In: *Nucleic acids research* 47.D1, pp. D766–D773.

Frésard, L, E R Gamazon, D Garrido-Martın, et al. (2017). "Co-expression networks reveal the tissue-specific regulation of transcription and splicing". In: *Genome*.

Ganna, Andrea et al. (June 2018). "Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum". en. In: *Am. J. Hum. Genet.* 102.6, pp. 1204–1211.

Hernandez, Ryan D et al. (Sept. 2019). "Ultrarare variants drive substantial cis heritability of human gene expression". en. In: *Nat. Genet.* 51.9, pp. 1349–1355.

Jian, Xueqiu, Eric Boerwinkle, and Xiaoming Liu (Dec. 2014). "In silico prediction of splice-altering single nucleotide variants in the human genome". en. In: *Nucleic Acids Res.* 42.22, pp. 13534–13544.

Keinan, Alon and Andrew G Clark (2012). "Recent explosive human population growth has resulted in an excess of rare genetic variants". In: *science* 336.6082, pp. 740–743.

Kremer, Laura S et al. (June 2017). "Genetic diagnosis of Mendelian disorders via RNA sequencing". en. In: *Nat. Commun.* 8, p. 15824.

Lee, Hane et al. (Mar. 2020). "Diagnostic utility of transcriptome sequencing for rare Mendelian diseases". en. In: *Genet. Med.* 22.3, pp. 490–499.

Li, Xin, Yungil Kim, Emily K Tsang, Joe R Davis, Farhan N Damani, Colby Chiang, Gaelen T Hess, Zachary Zappala, Benjamin J Strober, Alexandra J Scott, et al. (2017a). "The impact of rare variation on gene expression across tissues". In: *Nature* 550.7675, pp. 239–243.

Li, Xin, Yungil Kim, Emily K Tsang, Joe R Davis, Farhan N Damani, Colby Chiang, Gaelen T Hess, Zachary Zappala, Benjamin J Strober, Alexandra J Scott, et al. (Oct. 2017b). "The impact of rare variation on gene expression across tissues". en. In: *Nature* 550.7675, pp. 239–243.

Li, Yang I et al. (2016). "RNA splicing is a primary link between genetic variation and disease". In: *Science* 352.6285, pp. 600–604.

Lim, Kian Huat et al. (2011). "Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes". In: *Proceedings of the National Academy of Sciences* 108.27, pp. 11093–11098.

Mancuso, Nicholas et al. (2016). "The contribution of rare variation to prostate cancer heritability". In: *Nature genetics* 48.1, pp. 30–35.

Manolio, Teri A et al. (2009). “Finding the missing heritability of complex diseases”. In: *Nature* 461.7265, pp. 747–753.

Marouli, Eirini et al. (2017). “Rare and low-frequency coding variants alter human adult height”. In: *Nature* 542.7640, pp. 186–190.

Purcell, Shaun M et al. (Feb. 2014). “A polygenic burden of rare disruptive mutations in schizophrenia”. en. In: *Nature* 506.7487, pp. 185–190.

Rosenberg, Alexander B et al. (Oct. 2015). “Learning the sequence determinants of alternative splicing from millions of random sequences”. en. In: *Cell* 163.3, pp. 698–711.

Sangermano, Riccardo et al. (2018). “ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease”. In: *Genome Research* 28.1, pp. 100–110.

Smith, Sarah A and Kristen W Lynch (2014). “Cell-based splicing of minigenes”. In: *Spliceosomal pre-mRNA splicing: methods and protocols*, pp. 243–255.

Soemedi, Rachel et al. (Apr. 2017a). “Pathogenic variants that alter protein code often disrupt splicing”. en. In: *Nat. Genet.* 49.6, pp. 848–855.

— (2017b). “Pathogenic variants that alter protein code often disrupt splicing”. In: *Nature genetics* 49.6, pp. 848–855.

Takata, Atsushi et al. (2016). “De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia”. In: *Neuron* 89.5, pp. 940–947.

Wainschtein, Pierrick et al. (Mar. 2022). “Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data”. en. In: *Nat. Genet.* 54.3, pp. 263–273.

Yang, Jian et al. (2015). "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index". In: *Nature genetics* 47.10, pp. 1114–1120.

Yeo, Gene and Christopher B Burge (2004). "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals". en. In: *J. Comput. Biol.* 11.2-3, pp. 377–394.

Table 4.1: Primer Sequences Used in This Study

Name	Sequence
ODY88	CATCCATTCCAAGTGTTATGCCAG
ODY89	GCCGGGCTATTTACTTTTGTAAAAC
ORC279	GGGTCACGCGTAGGA
ORC280	CGCGTCGAGTAGGGT
ORC281	CGATCGCCCTTGGTG
ORC282	GGTCGAGCCGGAACT
ORC283	GTTCCGCAGCCACAC
ORC284	GCCGTGTGAAGCTGG
ORC285	CACGCCGGCTAAACC
ORC286	GGATGCGCACCCAGA
OCB214	AGCTATCTATGTCTACCGGT
1R	TAAATAGCCCGGCTGGAAC

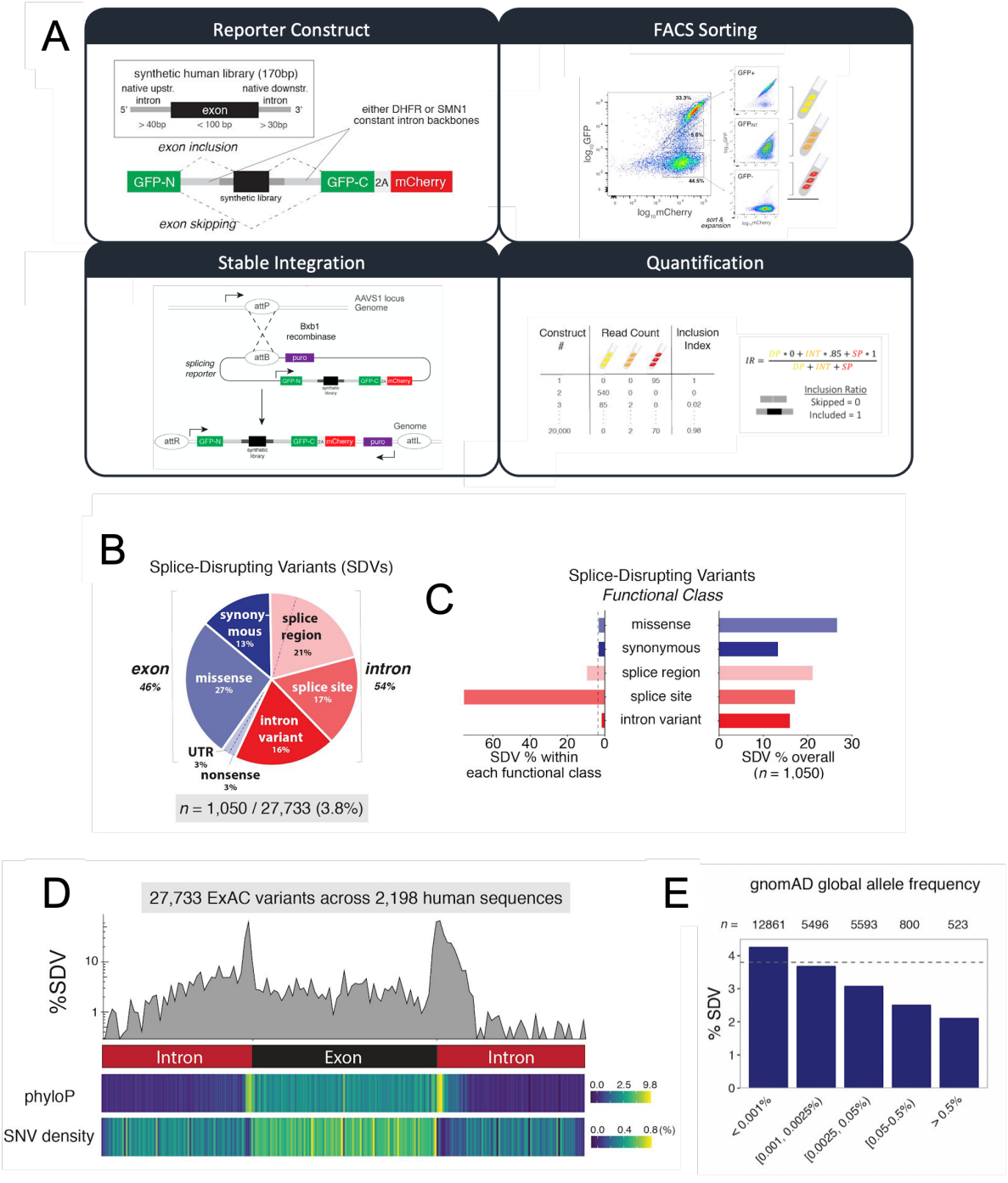


Figure 4.1: MPRA Design and Results: A) **Highlights of the MFASS Reporter Assay (Cheung, Insigne et al.)** Workflow and Analyses of Splice-Disrupting Variants (SDVs) using a Multi-Parallel Reporter Assay (MPRA).

Figure 4.1: (Previous page.) (A) Schematic representation of the experimental workflow: An in silico-designed intron-exon-intron library encompassing 9,634 human exons was generated, with each construct containing a variable region consisting of an exon and minimum of 40 bp of upstream and 30 bp of downstream intronic sequence. The splicing reporter constructs were stably integrated into a chromosomal landing pad via Bxb1 integrase to standardize RNA expression levels. Fluorescence-activated cell sorting (FACS) was employed to sort cells based on the reconstituted GFP signal, which serves as an indicator of splicing efficiency. Quantification was achieved by binning cells according to both RFP and GFP fluorescence levels, followed by normalization and calculation of the inclusion ratio based on the ratio of high-to-low GFP bins. (B) Localization of splice-disrupting variants within genes. The proportion of intronic SDVs was slightly higher than that of exonic SDVs. (C) Frequency distribution of SDVs by functional class. Splice site mutations disproportionately caused splicing disruption, but were not the most frequent class of SDVs of all detected. (D) Density distribution of SDVs along the normalized sequence construct length. (E) Scatter plot depicting the relationship between allele frequency and the likelihood of a variant being a splice-disrupting variant, with rare variants showing a higher propensity to be SDVs.

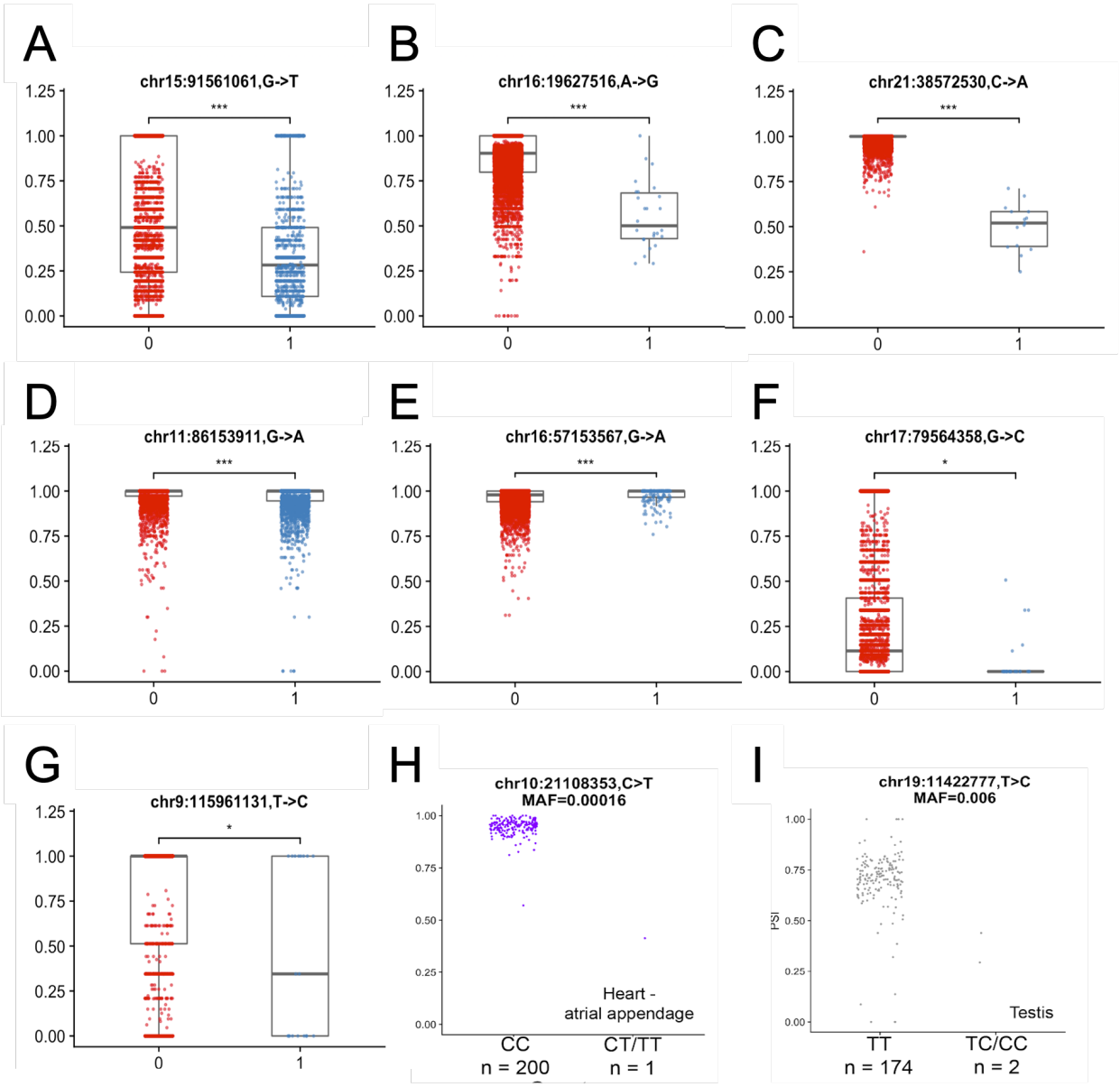


Figure 4.2: Comparison of MFASS SDVs with RNA-Seq Multi-Tissue Splicing: A-G) Splicing of global SDVs measured in all GTEx samples from individuals having at least one copy of the SDV. H and I) comparison of splicing in individuals with and without SDVs in exons that are only expressed in one tissue type.

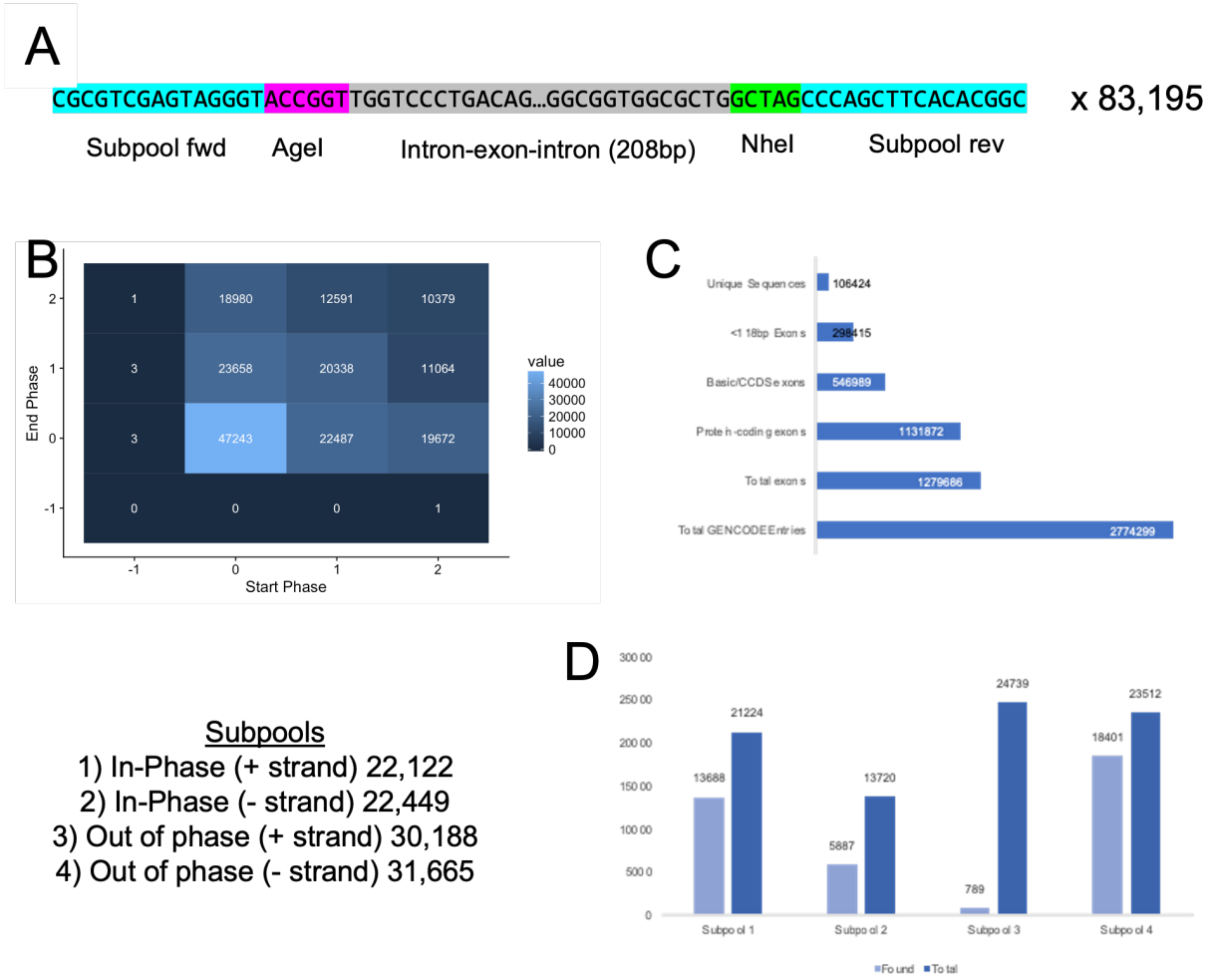


Figure 4.3: Design of a Library of Reference Human Exon Sequences: A) Library insert design. B) Heatmap of start and end phases of library sequences. C) Exon sequence filtering from GENCODE to library construction criteria. D) Subpool sequence membership and counts. E) Final counts for sequences measured after performing MFASS experiment on new subpools compared to library total counts.

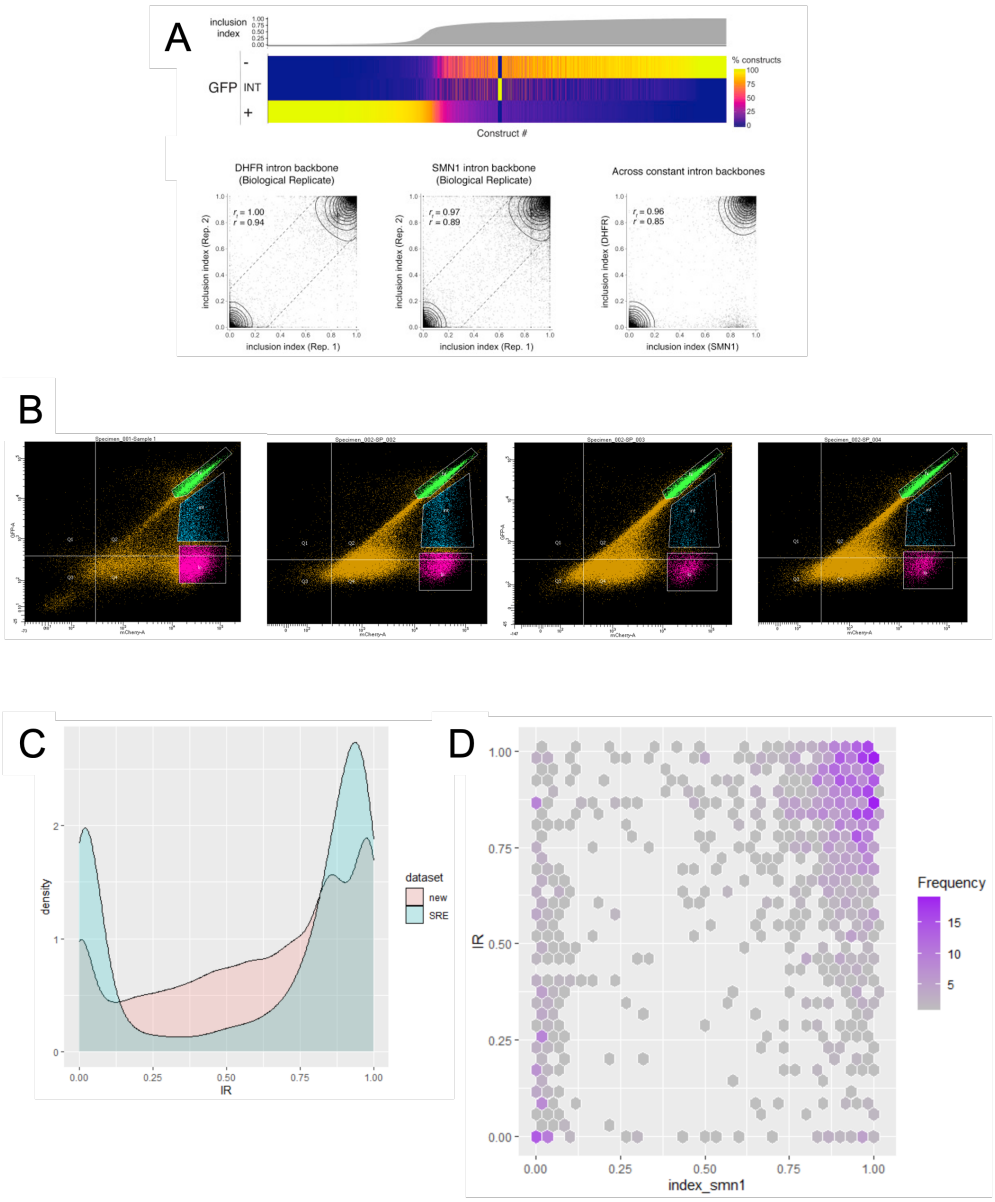


Figure 4.4: Comparison of Reference Library to MFASS Sequence Inclusion: We compared exon inclusion measurements from the original paper to the new reference library. A) (Reproduced from Cheung, Insigne et al.) MFASS showed a low number of sequences with intermediate values. Agreement between biological replicates was high. B) FACS sorting for each of the four subpools of reference library. C) Reference library inclusion levels contained many more intermediate values than the original MFASS SRE library. D) Plot of IR values for sequences present in both libraries.

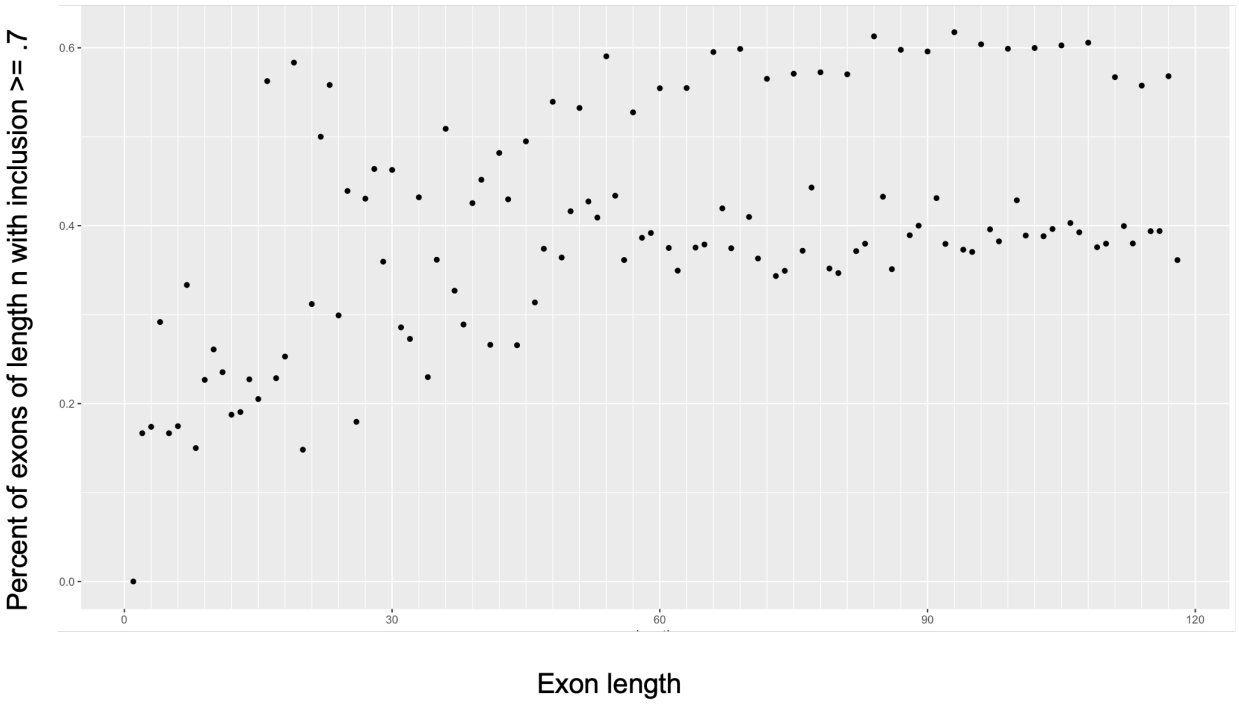


Figure 4.5: The Relationship Between Exon Length and Inclusion in MFASS: For exons of each given nucleotide length, the proportion of exons with an inclusion ratio (IR) $\geq$.7 is plotted. The plot shows a strong relationship between the length of the exon and rate of inclusion, and the phase of the exon and inclusion.

CHAPTER 5

Conclusion

The work in this thesis represents novel contributions that contribute to our understanding of the role of splicing in disease and our ability to assess genetic variants' capacities for altering splicing. We employ and refine high-throughput computational and experimental techniques in the study of splice variants and microexons. The completed aims of these projects specifically addressed 1) improved microexon quantification and discovery in large RNA-Seq datasets, 2) performing a global survey of microexon splicing across adult human tissues, and 3) evaluating splicing of human exon sequences in a massively parallel reporter assay and generating a reference of compatible exon sequences.

In chapter 2 we showed that current methods for microexon discovery and quantification could be improved for large datasets. When quantifying microexon splicing in a large (17K RNA-Seq samples) dataset, we identified several areas for improvement for better accuracy, computational efficiency and ease of use. We identified several sources of error that affected a substantial number of alignments and designed an improved alignment process that considers all potential tag matches to pick the most appropriate alignment and accounts for these sources of error. We demonstrate using previously developed metrics, including conservation and splice site strength distributions, that our changes further enrich for likely

functional microexons. We note that these improvements come from changes to alignment strategy and do not require any filters or assumptions about functional properties of microexons. We successfully ran our workflow on over 17,000 RNA-Seq samples from GTEx to produce a global picture of microexon splicing in adult humans across individuals and across 50 tissue types. This addresses the need for a scalable and reproducible workflow for microexon discovery. We chose to implement our workflow using WDL and Docker, which may be run on cloud, remote, and local environments using the Cromwell engine with minimal adjustments. Due to our workflow's scalability and improved ability to control alignment errors in large datasets, our method is better suited to studying the effects of sequence variation on microexon splicing over existing methods.

We also used long read RNA-Seq data to provide additional support for the microexons detected by our pipeline. This data showed that the microexon splice junction sequences detected in short read data were also supported by the sequences from long read data. The presence of these microexons was not accounted for by nearby splice sites, sequencing error, or genomic variants. Surprisingly, many of the validatated sequences were extremely short (¡5nt) microexons with high read coverage. Extremely short microexons have historically been very difficult to detect from short read RNA-Seq and filtered even from most microexon studies. However, experimental evidence for microexons as short as 1nt exists. The microexons identified here are good candidates for experimental follow-up and may indicate that microexon splicing is more prevalent in humans than previously appreciated.

In chapter 3, we perform a global survey of microexon splicing using our improved quantification workflow. We measure microexon splicing in over 17,000 RNA-Seq datasets from over 800 individuals and 50 tissue types from GTEx. To our knowledge this is the

biggest such study to date. We find evidence of thousands of conserved alternatively spliced and constitutive microexons that have the capacity to alter protein function in many cellular processes. Our results suggest that microexons are common in genes related to mechanical signal transduction in cells. We find strong enrichment for microexons in components of the cell cytoskeleton and interacting proteins, as well as in diverse mechanical processes such as axonogenesis and muscle cell contraction. We find strong evidence of tissue-specific splicing regulation, and find enrichment for many known splicing factor binding sites. We also find sQTLs and overlapping mendelian disease variants suggesting a role for microexon missplicing in neuromuscular disease. We quantified microexon splicing in myotonic dystrophy type 1 and show that many microexons are found to be differentially included. These microexon events were found to be functionally highly connected and involved in muscle contraction, and included several known DM1 disease markers. We also note the overlap with these microexons' splice sites and pathogenic rare CLINVAR variants. Our work suggests that further study of microexons, their potential role in mechanical signal transduction, and link to neuromuscular diseases is warranted.

In our final chapter, we evaluated a high-throughput system to test the effect of sequence variation on exon recognition. The reference library exon sequences shown to be recognized in this system provide a useful jumping off point for future studies of variants by providing a list of compatible exons. In addition, we noted interesting relationships between exon length and inclusion that could be informative in splice variant studies. Next steps include expanding our investigation of rare variation by testing many more variants in an expanded set of human exon contexts, and performing more in-depth validation of SDVs using available RNA-Seq data. The experimental model can also be compared to splicing in

endogenous contexts. This is important for interpreting results in MFASS and in all minigene reporter systems for splicing.

To conclude, exciting advances in experimental and computational techniques have opened the door to global studies of splicing patterns that may lead to better understandings of splicing's role in disease and better splicing prediction. We provide useful results in quantifying microexons, and show that this class of exon may have an underappreciated role in many important cellular functions and may be involved in myotonic dystrophy type 1, as well as other diseases. We hope our results motivate additional work on short exons which may lead to new insights in neuromuscular function and disease.