Title
Functional Roles of Single Nucleotide Variants in Alternative Splicing

Author
Hsiao, Yun-Hua

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Functional Roles of Single Nucleotide Variants in Alternative Splicing

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

by

Yun-Hua Hsiao

2018

ABSTRACT OF THE DISSERTATION

Functional Roles of Single Nucleotide Variants
in Alternative Splicing

by

Yun-Hua Hsiao
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2018
Professor Xinshu Xiao, Chair

Next-generation sequencing has greatly facilitated large-scale analyses of the human genome. However, the function and biological significance of many single nucleotide variants, including single nucleotide polymorphisms (SNPs) and RNA editing sites, are still poorly understood and remain to be uncovered. The effects of single nucleotide variants on transcriptional regulation of gene expression are relatively well studied, but their impact on co- and post-transcriptional control of gene expression is less appreciated. In this project, the overarching goal was to investigate the functional roles of single nucleotide variants in RNA splicing regulation.

To tackle this question, we first developed a method to identify intronic tag SNPs that can modulate alternative splicing. The cell fractionation RNA-seq datasets from the ENCODE Project made identification of intronic SNPs possible in our method. Over 600 intronic and exonic splicing-relevant SNPs were predicted in this study, and our comprehensive analyses

revealed the genomic, evolutionary and regulatory features of these genetically modulated alternative splicing (GMAS) events.

Next, we performed a genome-wide study to uncover the interplay between individual steps of RNA processing, an essential aspect of gene regulation. In eukaryotes, nascent RNA transcripts undergo an intricate series of RNA processing steps to achieve mRNA maturation. RNA editing and alternative splicing are two major RNA processing steps that can introduce significant modifications to the final gene products. Here, we aimed to determine RNA editing sites' impact on RNA splicing. RNA-seq datasets from different subcellular fractions enabled tracking of RNA editing sites over the course of transcription. About 500 editing sites were observed to reside in 3' acceptor sites, which could abolish normal splicing of the associated exons. In addition, we also explored the mechanism that editing sites affected splicing by RNA secondary structure remodeling and identified a subset of editing sites belonged to this category.

Lastly, the analyses of GMAS events in GTEx datasets provided insights into the characteristics of splicing-related SNPs and the associated exons. We observed that GMAS exons were highly tissue-specific and individual-specific, and rare GMAS events demonstrated outstanding associations with disease annotations compared to common events. In addition, we developed a new bioinformatic method to predict causal SNPs that alter splicing. This novel method identified over 600 causal SNPs for over 300 GMAS exons from the GTEx samples. Further analyses brought to light several splicing factors that were responsible for the GMAS patterns by interacting with the putative causal SNPs.

The dissertation of Yun-Hua Hsiao is approved.

Jingyi Li

Michael Alan Teitell

Matteo Pellegrini

Xinshu Xiao, Committee Chair

University of California, Los Angeles

2018

To God and my family

TABLE OF CONTENTS

1 Background 1

1.1 Introduction 1

1.2 Identification and Validation of Alternative Splicing Events 2

1.2.1 Identification of Alternative Splicing Events 2

1.2.2 Validation of Alternative Splicing Events 8

1.3 Methodologies for Studies of Splicing Regulation 9

1.3.1 Cis-Regulation of Alternative Splicing 10

1.3.2 Genetic Variants Associated with Splicing 13

1.3.3 Trans-Acting Regulators of Alternative Splicing 16

1.3.4 Splicing Code 18

1.3.5 Useful Databases 19

1.4 Ongoing Questions 19

1.4.1 Gene Expression Kinetics and Co-Transcriptional Splicing 19

1.4.2 RNA Modifications 21

1.4.3 Splicing Generates Other RNA Species 21

1.4.4 Global Misregulation of Splicing in Disease 22

1.5 Splicing as A Therapeutic Target 22

1.6 Figures 25

1.7 Tables 28

1.8 References 33

2 Alternative Splicing Modulated by Genetic Variants Demonstrates Accelerated Evolution Regulated by Highly Conserved Proteins 51

2.1 INTRODUCTION 51

2.2 RESULTS 53

2.2.1 Compartment-specific RNA-seq enhances coverage of intronic SNVs 53

2.2.2 Identification of intronic tag SNVs for GMAS 54

2.2.3 iGMAS events identified in seven ENCODE cell lines 56

2.2.4 Experimental validation of iGMAS predictions 57

2.2.5 A compendium of GMAS events 58

2.2.6 Cis-regulatory elements are primary drivers of splicing variation in GMAS 59

2.2.7 GMAS exons demonstrate accelerated sequence evolution in primate lineages 59

2.2.8 GMAS-related genes, exons, and SNVs undergo positive or balancing selection 60

2.2.9 GMAS events are likely regulated by conserved splicing factors 62

2.2.10 Alteration of splicing factor binding by GMAS SNVs 63

2.2.11 Allele-specific binding of SRSF1 in enabling GMAS regulation 64

2.2.12 Functional analyses of GMAS genes and exons 66

2.2.13 More than 100 GMAS SNVs are in LD with GWAS SNPs 68

2.3 DISCUSSION 70

2.4 METHODS 73

2.4.1 RNA-seq data and mapping 73

2.4.2 Prediction of iGMAS events 74

2.4.3 Splicing reporter assays 76

2.5 ACKNOWLEDGEMENTS 76

2.6 FIGURES 77

2.7 REFERENCES 88

3 RNA Editing in Nascent RNA Affects Pre-mRNA Splicing 94

3.1 INTRODUCTION 94

3.2 RESULTS 96

3.2.1 Subcellular RNA-seq and RNA editing 96

3.2.2 More than 95% of A-to-I editing occurs co-transcriptionally prior to polyadenylation 98

3.2.3 Comparative analyses of RNA editing sites across subcellular fractions 101

3.2.4 The impact of editing on exons with different splicing kinetics 102

3.2.5 Editing sites located in the splice site regions often disrupt 3' splice sites 103

3.2.6 Exons with A-to-I editing in the 3'ss tend to be highly conserved 106

3.2.7 Experimental verification of editing-dependent alternative splicing 107

3.2.8 Regulation of splicing by ADAR acting on dsRNA structures 109

3.3 DISCUSSION 111

3.4 METHODS 114

3.4.1 Cell fractionation and RNA-seq library construction 114

3.4.2 Detection of RNA editing sites 115

3.4.3 Definition of RNA editing kinetic groups 116

3.4.4 Prediction of novel exons and PSI calculation 118

3.4.5 Prediction of RNA secondary structures 119

3.4.6 Minigene constructs 120

3.5 DATA ACCESS 120

3.6 ACKNOWLEDGEMENTS 120

3.7 FIGURES 122

3.8 REFERENCES 132

4 Landscape of Genetic Modulation of Alternative Splicing and Its Causality in Human Tissues 140

4.1 INTRODUCTION 140

4.2 RESULTS 143

4.2.1 Overview of genetic modulation of alternative splicing events in GTEx data 143

4.2.2 GMAS variation across tissues and individuals 144

4.2.3 Comparative analysis of GMAS patterns across tissues 145

4.2.4 Comparative analysis of GMAS patterns across individuals 146

4.2.5 Rare GMAS SNPs exhibiting higher degreed of disease relevance 147

4.2.6 Rare GMAS exons are enriched with disease-related genes 150

4.2.7 Prediction of causal SNPs for GMAS 151

4.2.8 Causal SNPs for GMAS in GTEx 152

4.2.9 Experimental validation of causal SNPs for GMAS 154

4.2.10 RBPs regulating GMAS events by differential binding to causal SNPs 155

4.3 DISCUSSION 157

4.4 METHODS 160

4.4.1 Preprocessing of GTEx RNA-seq data and identification of GMAS events 160

4.4.2 Estimation of tissue vs. individual contributions to GMAS pattern variations 161

4.4.3 Tissue-specificity quantified by Jaccard index and GMAS frequency 162

4.4.4 Prevalence of a GMAS exon 163

4.4.5 Enrichment of functional SNPs among scarce vs. ubiquitous GMAS SNPs *163*

4.4.6 Prediction of causal SNPs for GMAS *164*

4.4.7 Power analysis for predicting causal SNPs for GMAS *166*

4.4.8 Calculation for the PSI of GMAS exons *166*

4.5 ACKNOWLEDGEMENTS 167

4.6 FIGURES 168

4.7 REFERENCES 177

5 Concluding Remarks 183

5.1 SUMMARY 183

5.2 CONCLUSIONS 185

5.3 REFERENCES 189

LIST OF FIGURES

1.1 Overview of previous studies in alternative splicing regulation..... 25

1.2 High-throughput experimental approaches for splicing detection..... 27

1.3 Co-transcriptional splicing and related RNA products 28

1.4 Therapeutic approaches to modulate splicing..... 29

2.1 Compartment-specific RNA-seq enables coverage of intronic SNVs 77

2.2 Evaluation of the iGMAS method..... 79

2.3 Splicing alteration by genetic variants is highly cell type-independent..... 81

2.4 GMAS events undergo accelerated evolution..... 82

2.5 GMAS SNVs alter the binding of splicing factors..... 84

2.6 Functional relevance of GMAS events..... 86

3.1 Kinetic groups of RNA editing sites identified via subcellular RNA-seq.....121

3.2 Characteristics of editing sites in different kinetic groups 123

3.3 RNA editing sites located in splice site regions.....125

3.4 Experimental validation of splicing changes induced by 3'ss editing 127

3.5 Regulation of splicing by ADAR acting on dsRNA structures 129

4.1 Landscape of GMAS exons in human tissues 167

4.2 Tissue-specificity of GMAS exons in GTEx population 169

4.3 Prevalence of GMAS exons across human tissues 170

4.4 Functional implications of GMAS events 171

4.5 Causal SNP predictions for GMAS events.....173

4.6 Splicing factors that regulate GMAS events through predicted causal SNPs 175

LIST OF TABLES

1.1 Tools for analysis of alternative splicing using RNA-seq data 30

1.2 Databases (DB) of alternative splicing events and software programs (PR) for studies of splicing regulation..... 31

ACKNOWLEDGEMENTS

First, I would like to thank my thesis advisor, Dr. Xinshu Xiao, for her guidance for the past six years. I am thankful for her courage to take me into the lab even after knowing that I did not have a strong bioinformatics background coming in. She set up a really great example for me to learn how to be a successful researcher, and I was inspired by her passion and dedication to research. Whenever I faced an obstacle in research, her office door was always open for discussion, and she would give very insightful and constructive feedback to help me overcome different challenges. I really appreciated the time and patience she spent in training me to be a real scientist, and I would not be able to achieve what I have accomplished today without her mentorship during my doctoral training.

Second, I would like to thank the members of the Xiao Lab for providing a positive learning environment for me. I have had the privilege to be in a lab with both dry lab and wet lab components and was able to learn from both areas, which broadened the breadth of my scientific understanding. In addition, our lab has people with various strengths and expertise, and everyone is willing to share his or her experiences and help one another to succeed. Besides providing great support scientifically, my lab mates are also fun and nice people to hang out with. I have been enjoying our chat time during lunch and lab socials, and my Ph.D. life definitely becomes much more colorful because of each and every one of the Xiao Lab members.

Third, I would like to thank all of my committee members: Dr. Jingyi Li, Dr. Michael Teitell, Dr. Matteo Pallegrini, and Dr. Xinshu Xiao. I am grateful that they always set aside time

from their busy schedules and responsibilities to meet with me. I have been benefited a lot from the critical and constructive suggestions they provided during the committee meetings over the years. Their advice and encouragements meant a lot to me.

I also want to thank the ENCODE Consortium for generating precious data that facilitated a significant portion of this study. I had the privilege to be part of the ENCODE Computational Analysis group and to participate the annual consortium meetings for three years. Those conferences were eye-opening experiences for me as I could learn the latest cutting-edge technology and research from the best laboratories in the genomics field. By presenting my own work at the consortium meetings, I was able to gain valuable comments and critiques from the experts, and the quality of my study was enhanced a lot from these discussions. I am very thankful for this opportunity and proud of being a member of the ENCODE family.

Next, I want to express my sincere thanks to my parents, Mr. Kang Le Hsiao and Ms. Hung En Chen, who gave me the opportunity to study in the US starting from high school and have been supporting me both financially and emotionally ever since. Although my family is not very wealthy, my parents worked very hard to provide good education for me in the US. I am very fortunate that I do not need to worry about tuition as an international student since undergrad. Without them, I would not be able to pursuit such wonderful education in world famous universities (UCSD for undergrad and UCLA for Ph.D.). Furthermore, I would like to thank my grandmother, Ms. Hsiu-Yuan Cheng Lin, and the rest of my family member in Taiwan and in the US for their constant prayers, support, and encouragement for my daily life. I also

want to thank my brother, Yun Peng Wesley Hsiao, for all the great accompany in the US for the past couple years.

In addition, I would also like to show my gratitude to my church pastor, Rev. Daniel Fan, and his wife, Ms. Grace Fan. They are like my parents in the US for the past 10 years or so when my real parents were in Taiwan. They are very caring not only for my physical needs but also for my emotionally and spiritual needs. They always give me strong support and encouragement during the emotional lows of my Ph.D. journey. I enjoy every talk we have had, which always fills me with positive energy after the conversations. On May 7th, 2018, just one month before my defense, Rev. Daniel Fan passed away and is now home in Heaven with our Heavenly Father. I cannot use words to express how much I miss him and my appreciation for him. He made a great impact on my life, and this is what keeps him alive in my heart and in my mind.

Last but not least, I want to thank all my close friends from Mandarin Baptist Church of Temple City, Kairos Christian Church in San Diego, and my schools in Taiwan and in the US. The friendship they share with me means a great deal for me. They keep me accountable and stand by my side when struggles come, and I am very thankful for having such strong and wonderful network of support in my life.

This work was partially supported by the Eureka, Hyde, and Bioengineering departmental fellowships of UCLA. The datasets analyzed in this work came from the ENCODE project or were generated by our lab with the assistance of the UCLA sequencing core.

Chapter 1 was published in *Translational Bioinformatics in Transcriptomics and Gene Regulation*, Springer Netherlands, 2016. Y.H.E. Hsiao, A.A. Cass, J.H. Bahn, X. Lin, and X. Xiao. "Global analyses of alternative splicing and its regulation – Recent advances and open questions." The dissertation author was the primary author of this book chapter.

Chapter 2 was published in *Genome Research*, 2016. Y.H.E. Hsiao, J.H. Bahn, X. Lin, T.M. Chan, R. Wang, and X. Xiao. "Alternative Splicing Modulated by Genetic Variants Demonstrates Accelerated Evolution Regulated by Highly Conserved Proteins." The dissertation author was the primary author of this paper.

Chapter 3 was published in *Genome Research*, 2018. Y.H.E. Hsiao, J.H. Bahn, Y. Yang, X. Lin, S. Tran, E.W. Yang, G. Quinones-Valdez, X. Xiao. "RNA editing in nascent RNA affects pre-mRNA splicing." The dissertation author was the primary author of this paper.

Chapter 4 is currently prepared for submission. Y.H.E. Hsiao, Y. Sun, J.H. Bahn, S. Tran, X. Xiao. "Landscape of genetic modulation of alternative splicing and its causality in human tissues." The dissertation author was the primary author of this paper.

Chapter 5.2 is partially a subsection of the published chapter in *Translational Bioinformatics in Transcriptomics and Gene Regulation*, Springer Netherlands, 2016. Y.H.E. Hsiao, A.A. Cass, J.H. Bahn, X. Lin, and X. Xiao. "Global analyses of alternative splicing and its regulation – Recent advances and open questions." The dissertation author was the primary author of this book chapter.

VITA

2008-2012	B.S. in Bioengineering: Bioinformatics, University of California, San Diego, La Jolla, CA
Winter 2010	Teaching Assistant, Department of Computer Science, University of California, San Diego, La Jolla, CA
Spring 2010	Teaching Assistant, Department of Computer Science, University of California, San Diego, La Jolla, CA
2011-2012	Lab Assistant, Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, CA
2012-2018	Graduate Student Researcher, Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA
Spring 2014	Teaching Assistant, Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA
Winter 2016	Teaching Assistant, Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA

Honors and Awards

May 2015	Hyde Fellowship, University of California, Los Angeles, CA
June 2016	Honorable Mentions Best Poster Award (2nd place), ENCODE Annual Consortium Meeting, La Jolla, CA
June 2016	Eureka Scholarship, University of California, Los Angeles, CA
September 2017	Selected Abstract for UCLA QCBio Annual Retreat, Los Angeles, CA
October 2017	Bioengineering Supplemental Fellowship, University of California, Los Angeles, CA

Publications

1. **Y.H.E. Hsiao**, A.A. Cass, J.H. Bahn, X. Lin, and X. Xiao. "Global analyses of alternative splicing and its regulation – Recent advances and open questions." *Translational Bioinformatics in Transcriptomics and Gene Regulation*. Springer Netherlands, 2016; 37-71.
2. **Y.H.E. Hsiao**, J.H. Bahn, X. Lin, T.M. Chan, R. Wang, and X. Xiao. Alternative Splicing Modulated by Genetic Variants Demonstrates Accelerated Evolution Regulated by Highly Conserved Proteins. *Genome Research*, 2016.
3. A.A. Cass, J.H. Bahn, J.H. Lee, C. Greer, X. Lin, Y. Kim, **Y.H.E. Hsiao**, and X. Xiao. Global analyses of endonucleolytic cleavage in mammals reveal expanded repertoires of cleavage-inducing small RNAs and their targets. *Nucleic Acids Research*, 2016.
4. **Y.H.E. Hsiao**, J.H. Bahn, Y. Yang, X. Lin, S. Tran, E.W. Yang, G. Quinones-Valdez, and X. Xiao. RNA editing in nascent RNA affects pre-mRNA splicing. *Genome Research*, 2018.
5. **Y.H.E. Hsiao**, Y. Sun, J.H. Bahn, B.X. Tan, S. Tran, and X. Xiao. Causality inference of genetic modulation of alternative splicing in human tissues. (In prep)
6. ENCODE Consortium. ENCODE Phase III: Building an Encyclopedia of Regulatory Elements for Human and Mouse. Nature. (In revision)
7. S. Tran, H.I. Jun, J.H. Bahn, A. Azghadi, G. Ramaswami, E.L. Van Nostrand, T.B. Nguyen, **Y.H.E. Hsiao**, C. Lee, G.A. Pratt, G.W. Yeo, D.H. Geschwind, and X. Xiao. Widespread RNA editing dysregulation in Autism Spectrum Disorders. Nature Neuroscience. (In revision)
8. E.W. Yang, **Y.H.E. Hsiao**, J.H. Bahn, B.X. Tan, E.L. Van Nostrand, B. Zhou, G. Quinones-Valdez, G.A. Pratt, P. Freese, X. Wei, A.E. Urban, B. Graveley, C.B. Burge, G.W. Yeo, and X. Xiao. Identifying allele-specific binding sites of RNA-binding proteins from eCLIP-Seq data. (In prep)
9. Y. Sun, **Y.H.E. Hsiao**, G. Quinones Valdez, and X. Xiao. GMAS DB: Genetic Modulation of Alternative Splicing DataBase. (In prep)

CHAPTER 1

Background

1.1 Introduction

First discovered nearly 40 years ago (Berget et al. 1977; Chow et al. 1977), pre-mRNA splicing consists of a series of biochemical reactions that function to remove introns and ligate flanking exons. Exon-intron boundaries are defined by highly conserved consensus sequences including the 5' splice site (5'ss, or donor site), 3' splice site (3'ss, or acceptor site), and branch point sequences (BPS) (Figure 1.1). These sequences are recognized by the spliceosome, a dynamic multi-ribonucleoprotein complex composed of small nuclear ribonucleoproteins (snRNPs) (refer to (Wahl et al. 2009) for detailed reviews). The spliceosome is the basic machinery that carries out splicing reactions.

In recent years, it was estimated that more than 90% of human genes are processed through alternative splicing where multiple spliced isoforms are generated from a single gene, thus significantly increasing transcriptome diversity (Wang et al. 2008; Pan et al. 2008; Nilsen TW and Graveley BR 2010). The most extreme case of alternative splicing is the *Drosophila* Down syndrome cell adhesion molecule gene (*Dscam*) which includes 48 exons and can theoretically produce 38,016 alternative transcripts from a single gene (Schmucker et al. 2000). Different types of alternative splicing exist with the most common ones being exon skipping, alternative 5'ss usage, alternative 3'ss usage, mutually exclusive exons and intron retention (Matlin et al. 2005).

It is now well established that alternative splicing contributes to a wide spectrum of cellular functions (Kalsotra and Cooper 2011). Disruption of normal splicing was reported for a large number of human diseases, which has been reviewed extensively (Wang and Cooper 2007; Cooper et al. 2009; Poulos et al. 2011). As a functionally critical process, alternative splicing is regulated by a myriad of *cis*-elements and *trans*-acting factors (Figure 1.1). Splicing regulatory elements reside in exons or introns and function to either enhance or silence splicing. These *cis*-elements are thus named accordingly as: exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs), and intronic splicing silencers (ISSs). These *cis*-elements interact with many *trans*-acting factors (i.e., splicing factors), including, for example, serine/arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) (Black 2003). RNA secondary structures also affect alternative splicing, likely by facilitating or blocking accessibility of splicing factors to their cognate RNA (Buratti and Baralle 2004).

Understanding the regulatory mechanisms of alternative splicing in health and disease is an essential topic of gene regulation. Recent advances in high-throughput technologies and related bioinformatic methodologies are enabling exciting discoveries in this area. Here, we first focus on global approaches for splicing identification, followed by an in-depth review of methodologies to study splicing regulatory mechanisms.

1.2 Identification and Validation of Alternative Splicing Events

1.2.1 Identification of Alternative Splicing Events

1.2.1.1 High-throughput Experimental Approaches

The first high-throughput method developed to detect and quantify alternative splicing events was customized microarrays (Lee and Roy 2004; Cuperlovic-Culf et al. 2006; Blencowe 2006). An initial study by Hu *et al.* (Hu et al. 2001) used multi-probe design of Affymetrix arrays to detect splicing variants, demonstrating the utility of microarrays for splicing analyses. Later studies (Johnson et al. 2003; Pan et al. 2004) developed different techniques to improve the microarray probe design and successfully profiled alternative splicing events and their expression on the genome-wide scale. Johnson *et al*. (Johnson et al. 2003) used splice junction arrays to probe around 10,000 human multi-exon genes across 52 tissues. Besides the known alternative splicing events, they were also able to discover novel spliced isoforms of many genes. Pan *et al*. (Pan et al. 2004) took the focused probe design approach (see review (Cuperlovic-Culf et al. 2006)) with three exon body probes and three spliced junction probes for each known alternative splicing event to achieve more sensitive expression quantification. In this study, they were able to globally determine the tissue-specificity of alternative splicing events in mouse tissues. Many recent studies adopted different probe designs and microarray platforms to investigate splicing profiles and splicing levels in healthy and disease samples (reviewed in (Lee and Roy 2004; Cuperlovic-Culf et al. 2006; Blencowe 2006)).

Since the advent of next-generation sequencing (NGS), RNA sequencing (RNA-Seq) became an essential technology for global studies of alternative splicing (Figure 1.2A). It provides a means to directly or indirectly sequence the RNA molecules in a high-throughput manner. At present, often-used RNA-Seq methods first convert the RNA sample of interest into cDNAs, which are then made into a sequencing library that consists of short DNA fragments

(corresponding to the RNA of interest) flanked by pre-designed adapter oligos. The DNA library is then sequenced from one end (single-end sequencing, or SE) or both ends (paired-end sequencing, or PE) to yield final RNA-Seq reads (Wang et al. 2009). The resulting RNA-Seq reads correspond to a snapshot of RNA expression in the respective cellular sample.

RNA-Seq is advantageous in several ways. First, it can detect novel isoforms and alternative splicing events that are not yet annotated (Sultan et al. 2008; Lee et al. 2011). Second, RNA-Seq is not affected by the cross-hybridization problem that confounds many microarray-based studies (Wang et al. 2009). Third, RNA-Seq data can provide relatively accurate quantification of levels of gene expression and splicing (Wang et al. 2009; Mortazavi et al. 2008). Lastly, RNA-Seq provides single-nucleotide information that enables studies of genetic variants (Li et al. 2012a; Majewski and Pastinen 2011) and RNA editing sites (Wulff et al. 2011; Bahn et al. 2012; Lee et al. 2013), in addition to gene or exon expression. Using RNA-Seq, a large number of alternative splicing events was identified in human and mouse tissues (Wang et al. 2008; Pan et al. 2008).

Although RNA-Seq has dramatically improved our knowledge on alternative splicing, there are still remaining challenges to be addressed (Wang et al. 2009; Kratz and Carninci 2014). RNA-Seq library construction is the first critical step. Different library preparation protocols were developed to study various biological questions and thus have their own merits and limitations (van Dijk et al. 2014; Head et al. 2014). In addition, RNA-Seq library generation protocols often need optimization for specific RNA samples based on sample quality, concentration and other variables. In large-scale experiments, batch effects in RNA-Seq data

may be a critical problem to consider (Leek et al. 2010), which may mislead study conclusions if not properly accounted for. Finally, RNA-Seq experiments are still costly, especially for studies of alternative splicing. In such applications, reads covering spliced junctions are examined closely to guide the identification and quantification of alternative splicing. Thus, it is highly desirable to have a relatively large number of spliced reads. Often-used settings of RNA-Seq in splicing studies favor PE reads, long read length (e.g., > 75bp) and high sequencing depth (≥ 100 million PE reads for human samples) (Liu et al. 2013; Katz et al. 2010).

Alternative approaches were developed to address some of the above challenges in RNA-Seq. For example, RNA-mediated oligonucleotide Annealing, Selection, and Ligation with next-generation Sequencing (RASL-Seq) allows for RNA sequencing of a limited set of exons in hundreds or thousands of biological samples (Li et al. 2012b) (Figure 1.2B). Thus, it is ideal for large-scale analysis of up to 500 exons in complex networks or pathways (Li et al. 2012b). The main difference between RNA-Seq and RASL-Seq is the use of oligonucleotides that recognize a specific spliced junction in the latter method. After ligating the pairs of oligos, these specific RNAs are then isolated with biotinylated oligo-dTs and pulled down with streptavidin-coated magnetic beads. A unique barcode for each sample is incorporated during PCR, allowing for pooled sequencing of >1,500 samples per lane (Li et al. 2012b). Analyzing expression of a limited number of genes in many samples has clinical applications, such as screening for drugs that inhibit splicing events implicated in cancer (Li et al. 2012c). One factor of consideration in RASL-Seq is the efficiency and specificity of ligation; Rnl2 was shown to have higher efficiency than T4 ligase (Larman et al. 2014).

A limitation common to all sequencing-based methods is the sequencing read length, which is typically much shorter than the full-length isoform of long transcripts. Full-length isoforms are thus reconstructed computationally using overlapping reads, though there is always a degree of uncertainty (Garber et al. 2011; Steijger et al. 2013). To overcome this limitation, a new method SeqZip was recently developed (Roy et al. 2015) (Figure 1.2C). It uses ~40-60nt DNA "ligamers" that recognize the 5' and 3' ends of single or multiple alternatively spliced exons that may be thousands of nucleotides apart, causing the intermediate sequence to loop out (Roy et al. 2015). Multiple ligamers hybridized to the same transcript are then ligated together, thereby connecting distant exons in the same transcript. Assessing the length and sequence of the DNA ligamers allows for deduction of the full-length isoform. Thus, SeqZip greatly improves the ability to sequence long transcripts.

Aside from the above-mentioned RNA-based approaches, protein-based approaches may be used to identify changes in protein expression resulting from alternative splicing events. Mass spectrometry has been used to identify alternative splicing events in breast cancer (Zhang et al. 2013). Still, RNA-based approaches are far more commonly used for alternative splicing identification. The choice of experimental method depends on the experimental goal. As sequencing technology improves, so will the ability to identify alternative splicing events.

1.2.1.2 Bioinformatic Algorithms for Analyzing Alternative Splicing Using RNA-Seq

Current bioinformatic methods for analyzing alternative splicing in RNA-Seq can be largely classified into two categories: exon-centric and isoform-centric. Exon-centric approaches directly estimate the splicing level of each exon typically by calculating its percent spliced-in (PSI)

(Wang et al. 2008), a measure of the frequency of exon inclusion among all mature mRNA molecules of the gene (also see reviews (Chen 2013; Hooper 2014)). In contrast, isoform-centric methods aim to quantify the abundance of each alternative isoform of the gene, which can be followed by further comparisons to determine differential splicing (Patro et al. 2014; Zhang and Wang 2014; Aschoff et al. 2013).

The benefit of using exon-centric splicing detection is that the type and PSI of each alternatively spliced exon are directly interrogated. Such single-exon information is useful in designing experiments to validate and further examine these events (Katz et al. 2010; Shen et al. 2014). PSI can be calculated in different ways. First, abundance of reads aligned directly to alternative exon junctions is used, with the exon body reads optionally included (Katz et al. 2010; Shen et al. 2014). However, it is difficult to precisely estimate the PSI value in cases of complex alternative splicing. To overcome this problem, other tools, such as SplAdder and DiffSplice (Kahles et al. 2015; Hu et al. 2013), adopt a splicing graph strategy to capture the complexity of alternative splicing by building a graph of spliced isoforms where nodes represent exons and edges represent spliced introns. Input RNA-Seq data is used to update the alternative path in the graph. The challenge in these approaches is that the splicing graph can be complicated by poorly supported events, so post-filtering is necessary to reduce false positives. In general, exon-centric methods alone do not support identification of novel alternative splicing events due to their requirement of gene annotation.

Instead of focusing on specific splicing events, isoform-centric methods use RNA-Seq to construct isoforms and estimate their expression levels (Guttman et al. 2010; Trapnell et al.

2010; Grabherr et al. 2011; Li and Dewey 2011). Most tools also utilize the reference genome to guide isoform reconstruction, but others perform *de novo* transcriptome assembly without relying on the reference genome. The latter type is particularly helpful for alternative splicing analyses in species with poorly annotated genomes. Early isoform-centric methods were developed under the assumption that the read distribution is uniform, though this is rarely the case. New methods are now available to account for RNA-Seq read non-uniformity (Suo et al. 2014; Hu et al. 2014). Another recent development for isoform-centric analysis is the alignment-free approach, which bypasses the time-consuming alignment step by building a hash index from the reference transcripts using sequence k-mers as keys and applying an expectation maximization algorithm to estimate isoform abundance (Patro et al. 2014; Zhang and Wang 2014). This approach speeds up the computational time considerably while maintaining prediction accuracy. However, it remains to be evaluated whether such methods perform well in the presence of sample-specific genetic variants.

Once alternative splicing is identified, both classes of methods provide a means to detect differentially spliced events (Table 1.1). The outcome from exon-centric analyses is a list of differentially spliced events that can be directly used for further analysis (e.g., experimental validation, functional interpretation, regulatory studies). On the other hand, isoform-centric analysis captures the splicing complexity of a series of related events within the same isoform, but further steps are often needed to pinpoint individual splicing events of interest. In Table 1.1, we summarize often-used tools for splicing analysis.

1.2.2 Validation of Alternative Splicing Events

In silico tools that detect alternative splicing events based on RNA-Seq data usually generate a large number of candidates. A subset of these events should be experimentally validated *in vivo* or *in vitro*. Verification experiments for alternative splicing events are readily carried out by reverse transcription followed by PCR (RT-PCR) using primers that target flanking constitutive exons (Wang et al. 2003). This strategy works well for alternative splicing events in genes with intermediate or high expression levels. In order to verify lowly expressed events, *in vitro* minigene expression analysis by RT-PCR can be utilized (Wang et al. 2004; Xiao et al. 2007, 2009). Compared with *in vivo* assays, the minigene system is able to validate events regardless of their endogenous expression level. However, since only a limited region flanking the exon of interest can be cloned into the minigene vector, this *in vitro* approach may not faithfully reproduce *in vivo* splicing patterns. It should be noted that both types of experiments are considered low-throughput and labor intensive, thus only validating a relatively small number of events.

High-throughput methods for validation of alternative splicing events are in great demand and several such approaches are in the horizon. For example, RT-PCR experiments may be scaled up when used in conjunction with microfluidic devices (Mark et al. 2010). In addition, recent methods, such as the "designer exons" approach (Arias et al. 2015), may be further developed for this purpose. With the rapid technology development in synthetic biology and genome editing, it is likely that high-throughput splicing validation will soon become a reality.

1.3 Methodologies for Studies of Splicing Regulation

Pre-mRNA splicing is regulated by a large number of *cis*-elements and *trans*-acting factors. In this section, we will review the bioinformatic and experimental approaches for the identification and analysis of splicing regulatory mechanisms.

1.3.1 Cis-regulation of Alternative Splicing

1.3.1.1 Splice Site Consensus Sequence

Splice site sequences are among the best-characterized *cis*-elements in splicing regulation, owing to the simplicity of their identification. Each internal exon is flanked by a 5'ss and a 3'ss. Thus, splice site sequences can be easily collected based on gene annotation. The majority of human exons are flanked by the GU-AG canonical sequences. However, the splice site signals normally involve a much longer sequence motif, which confers specificity and a dynamic range of splice site strength. Using known splice site sequences as training data, many algorithms were developed to predict splice site strength (see reviews (Jian et al. 2014; Desmet et al. 2012)). The most intuitive model is the position weight matrix (PWM), which is straightforward to implement but fails to consider the positional dependency between nucleotides in the splice site (Desmet et al. 2009). Other algorithms adopt more sophisticated probabilistic models such as neural networks or maximum entropy to more accurately estimate the splice site scores (REESE et al. 1997; Yeo and Burge 2004).

1.3.1.2 Branch Point Sequences (BPSs)

The prediction of BPS is challenging because its location in the intron can be highly variable. For example, a BPS may be close to the 3'ss (~40nt upstream) or 100-400nt upstream of the 3'ss

in the AG Exclusion Zone (AGEZ) (Gooding et al. 2006). Additionally, the BPS motif is highly degenerate (Gao et al. 2008) and multiple potentially functional BPSs may exist in a particular intron. A number of bioinformatic methods were developed to identify BPS and evaluate their strength. Human Splice Finder (Desmet et al. 2009) uses PWMs and the algorithm proposed by Gooding *et al.* (Gooding et al. 2006) to search for BPS candidates in a limited region. Another predictive approach makes use of sequence conservation and partial sequence complementarity of U2 snRNA to the BPS (Plass et al. 2008; Schwartz et al. 2008). A recent study showed that using machine learning methods such as support vector machines together with polypyrimidine and other sequence information could increase accuracy in BPS prediction (Corvelo et al. 2010). Pastuszak *et al*. took advantage of the fact that Splicing Factor 1 (SF1) recognizes BPSs and restricted their motif analysis to sites with high SF1 binding affinity to predict BPS with relatively high accuracy (Pastuszak et al. 2011).

Recently, a few studies used the NGS technology to identify BPS globally. In the RNA-Seq data, a minority of reads may derive from the junction of the 5'ss-branch point of the intron lariat. A search for such reads has led to successful identification of hundreds of BPS in human RNA-Seq data sets (Bitton et al. 2014; Taggart et al. 2012). The advantage of these approaches is that they do not require prior knowledge about the BPS locations or sequences. However, one drawback is that lariat reads are very rare among those generated from standard RNA-Seq libraries. Thus, very deeply sequenced data sets are needed to obtain adequate lariat read coverage. Another NGS-based method, called CaptureSeq (Mercer et al. 2012), was applied recently to identify BPS (Mercer et al. 2015). In this method, tiling arrays were designed that contain oligonucleotide probes to target the 5'ss-branch point junctions (Mercer et al. 2015).

cDNAs from the RNA samples of interest were then hybridized, eluted and sequenced. As a complementary approach, RNase R digestion was applied to enrich for reads containing BPS without requiring pre-designed arrays. This study identified >50,000 human BPS in >10,000 genes, which enabled further investigation of global features of this class of splicing regulatory signal (Mercer et al. 2015).

1.3.1.3 Splicing Regulatory Elements

Besides the core splicing signals, a large number of motifs in the exons or introns can also regulate splicing (Figure 1.1) (Matlin et al. 2005). Identification and characterization of these splicing regulatory elements (SREs) are instrumental to the understanding of splicing regulatory mechanisms. In general, genome-wide experimental or bioinformatic screens have been designed to identify SREs. Wang et al. developed the first large-scale screen of ESSs using splicing reporter assays in cultured cells(Wang et al. 2004). This effort successfully identified hundreds of ESS sequences and shed light on the global properties of these elements. Later, a number of other experimental screens were carried out to identify different types of SREs (Culler et al. 2010; Ke et al. 2011; Wang et al. 2012, 2013). These studies greatly expanded the catalog of known or predicted SREs without the associated *trans*-factors necessarily identified. Other experimental methods that pinpoint SREs for known splicing factors will be discussed later.

In addition to experimental approaches, bioinformatic methods are also essential to SRE studies. Fairbrother *et al*. developed a motif comparison approach, RESCUE-ESE, to identify ESEs by evaluating motif enrichment correlated with different features of splicing (Fairbrother et al. 2002). Similar principles were applied later to identify other types of SREs (Zhang and Chasin 2004; Yeo et al. 2004). A myriad of other bioinformatic methods were also developed for

this purpose, such as those based on comparative genomics (Venkatesh and Yap 2005), PWMs (Cartegni et al. 2003), or machine learning techniques (Zhang et al. 2003; Stadler et al. 2006; Zhang et al. 2012; Badr and Heath 2014).

With the increasing number of SREs, a great deal of effort was dedicated to understand the functional interaction among different elements and their context-dependent roles in splicing regulation. For examples, Bayesian networks were used to study coevolutionary relationships of SREs in eukaryotes that reflect functional interaction (Xiao et al. 2007). Bioinformatic and statistical methods, combined with experimental approaches, were used to infer combinatorial function of different types of SREs (Friedman et al. 2008; Yu et al. 2008; Ke and Chasin 2010). The function of individual motifs (corresponding to one splicing factor) was studied in detail via bioinformatic modeling and analysis to reveal their context-dependent function globally (Xiao et al. 2009; Weyn-Vanhentenryck et al. 2014; Zhang et al. 2010; Han et al. 2014) (refer to (Fu and Ares 2014) for a detailed review of this topic).

1.3.2 Genetic Variants Associated with Splicing

Genetic variants (such as mutations or single-nucleotide polymorphisms (SNPs)) play important roles in gene regulation because they can potentially alter *cis*-regulatory motifs. Previous studies estimated that 15-60% of point mutations that result in human genetic diseases disrupt splicing (Wang and Cooper 2007; Krawczak et al. 1992; Ars et al. 1999; Teraoka et al. 1999; Lopez-Bigas et al. 2005). In recent years, exciting progress has been made in analyzing the involvement of genetic variants in modulating alternative splicing, which is reviewed in this section.

1.3.2.1 Splicing QTLs

Splicing quantitative trait loci (sQTL) analysis is an often-used method to identify SNPs associated with splicing phenotypes. In this method, the correlation between SNP genotypes and exon inclusion levels is examined using different means, ranging from simple linear correlation to model-based analysis (Kwan et al. 2008; Zhao et al. 2013; Monlong et al. 2014). Early sQTL studies used microarrays to detect isoform or exon expression levels, which is rapidly replaced by RNA-Seq-based analysis. However, this method requires a large number of samples to achieve adequate statistical power. In addition, sQTL analyses only deduce correlative relationships, without the capability of pinpointing the causal SNP for splicing alteration.

1.3.2.2 Machine Learning-based Methods

In contrast to sQTLs, methods based on machine learning principles aim to predict the functional (causal) SNP that modulates alternative splicing. Different types of machine learning or statistical methods were adopted for this purpose (Mort et al. 2014; Sterne-Weiler et al. 2011; Xiong et al. 2014). One study used a random forest-based strategy and predicted exonic splicing-altering variants(Mort et al. 2014). Another study developed a splicing code where "code quality" was optimized using information theory on a large number of features (Barash et al. 2010). This splicing code was applied to predict genetic variants that may alter splicing (Xiong et al. 2014; Barash et al. 2010, 2013). One common challenge to such approaches is the limited availability of training data sets that should include experimentally validated SNPs with confirmed function in splicing and those that are known to have no influence on splicing. To overcome this problem, previous studies used disease-causing exonic mutations from existing databases as positive training data set and common SNPs in the general population as negative

data set (assuming they do not affect splicing) (Mort et al. 2014; Sterne-Weiler et al. 2011). In contrast, the splicing code-based studies used human RNA-Seq data of different tissues to derive the code, without the need of direct model training using splicing-related variants (Xiong et al. 2014; Barash et al. 2010, 2013).

1.3.2.3 Allele Specific Alternative Splicing

To infer genetic regulation of alternative splicing, another powerful approach is built upon allele-specific expression (ASE) of genetic variants. ASE refers to the biased expression of the two alleles of a variant in diploid cells. RNA-Seq data provide single-nucleotide information that is appropriate for ASE studies. One advantage of ASE analysis is that the two alleles of a variant serve as within-sample controls of each other, which naturally eliminates the environmental and *trans*-acting effects that might alter splicing patterns or introduce variance in the data (Pastinen 2010). Nevertheless, one challenge in using RNA-Seq for ASE analysis lies in the step of read mapping. It is now clear that standard mapping methods induce a mapping bias that favors the reference allele of the genetic variant because the reference genome is utilized in mapping (Degner et al. 2009; Heap et al. 2010). Various strategies were developed to reduce this type of bias (Wulff et al. 2011; Bahn et al. 2012). Once ASE patterns are identified, they can be further analyzed to detect allele-specific alternative splicing events, as proposed in (Li et al. 2012a). While sQTL studies and machine learning methods necessitate many data points for correlative analysis or model training, the ASE-based approach can predict splicing-associated genetic variants using RNA-Seq data of a single individual. Thus, it is both cost-effective and computationally inexpensive.

1.3.3 Trans-acting Regulators of Alternative Splicing

1.3.3.1 Methods for Identification of Splicing Factors

Recently, an increasing number of RNA binding proteins (RBPs) have been identified as regulators of splicing (Fu and Ares 2014). However, the associated splicing factors are not yet known for a large number of SREs identified using the experimental or bioinformatic methods described above. To this end, a modified RNA affinity purification method was used to identify *trans*-factors for known SREs (Wang et al. 2012, 2013; Wang and Wang 2014b). In addition, *in vivo* siRNA screens targeting known splicing factors were also used to reveal the *trans*-factor for specific SREs (Culler et al. 2010; Izquierdo et al. 2005; Underwood et al. 2005; Huelga et al. 2012).

Previous efforts were also dedicated to predict or validate proteins with splicing regulatory activity (Fu and Ares 2014). For example, a computational pipeline was designed to search for proteins with splicing factor-like properties, which led to discovery of an SR-related protein with important function in neuronal tissues (Calarco et al. 2009). Given a pool of RBPs, a previous study screened for splicing-related ones by examining the correlation of their expression with changes in levels of alternative splicing (Ray et al. 2013). Combined with motif analysis, the authors successfully identified known and novel splicing factors.

1.3.3.2 Methods for Identifying Binding Motifs of Splicing Factors

Given a splicing factor or RBP, a number of experimental methods were developed to identify their binding motifs globally. These methods can be largely categorized into two classes depending on their *in vitro* or *in vivo* nature. The Systematic Evolution of Ligands by EXponential enrichment (SELEX) approach is one of the *in vitro* methods (Tuerk and Gold 1990). SELEX was applied to identify ESEs and other SREs in several studies (Lee and Rio 2015). Recently, this method was combined with microarray assays to increase the throughput (Reid et al. 2009). Another *in vitro* method called RNAcompete uses *in vitro* transcribed RNA (structured or unstructured) for pulldown with an RBP of interest, followed by microarray analysis of the bound RNA (Ray et al. 2009). Binding motifs of over 200 RBPs were determined by this method (Ray et al. 2013). More recently, a new *in vitro* method called RNA Bind-n-Seq (RBNS) was developed to improve quantification of the sequence and structural specificity of RBPs (Lambert et al. 2014). Besides canonical motifs, RBNS identified additional near-optimal binding motifs, which were shown to be functional *in vivo* (Lambert et al. 2014).

To identify global *in vivo* binding sites of RBPs, the most-widely used method is UV cross-linking and immunoprecipitation (CLIP) followed by sequencing (CLIP-Seq) (Ule et al. 2005). Variations of this method are also used for different applications, including high-throughput sequencing of RNAs isolated by CLIP (HITS-CLIP) (Licatalosi et al. 2008), photoactivatable-ribonucleoside-enhanced cross-linking and precipitation (PAR-CLIP) (Benhalevy et al. 2016), and individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) (Konig et al. 2010). Detailed discussions of these methods are provided by previous reviews (McHugh et al. 2014). Briefly, CLIP-based methods have relatively high sensitivity and specificity compared to RNA Immunoprecipitation alone.

However, the cross-linking efficiency is generally limited in regular CLIP, which is improved in PAR-CLIP via the usage of 4-thiouridine, a photo-activated nucleotide. Deletions, substitutions, or insertions usually occur near the cross-linking sites in CLIP-Seq/HITS-CLIP (Zhang and Darnell 2011), whereas T-to-C substitutions are observed near the cross-linking sites in PAR-CLIP. These mutations can serve as diagnostic features to pinpoint binding sites. Nonetheless, accurate read mapping tolerating such mutations is challenging. Currently, bioinformatic tools are designed to handle read mapping, cluster calling, and motif enrichment. In the future, development of tools that integrate these basic analyses with RNA secondary structure, evolutionary conservation, and *in vitro* binding data will tremendously facilitate a systematic understanding of protein-RNA interaction.

Notably, the ENCODE Consortium has devoted great efforts to generate CLIP-Seq data of about 200 RBPs. In addition, shRNA knockdown experiments of each RBP are carried out followed by RNA-Seq in cultured cells (K562 and HepG2). These data sets will facilitate identification of splicing regulatory motifs, analysis of splicing factor functions, and generation of global regulatory maps of these RBPs.

1.3.4 Splicing Code

While most existing methods focus mainly on one or a few aspects of splicing regulation, Barash *et al*. took a step further to assemble a "splicing code" by integrating hundreds of RNA features and the alternative splicing patterns of a wide panel of tissues (Barash et al. 2010). This model takes as input exon sequences of interest and their flanking introns, and recursively selects for features and parameters that maximize the "code quality." The code was later improved using

Bayesian Neural Networks on an expanded list of RNA features (Xiong et al. 2011; Barbosa-Morais et al. 2012) and applied to predict splicing-altering disease mutations (Xiong et al. 2014). The above work mainly focused on analysis of alternatively skipped exons. A more recent splicing code was designed to identify RNA sequence features that categorizes several major classes of alternative splicing, including exon skipping, alternative 5'ss and alternative 3'ss exons (Busch and Hertel 2015). This work demonstrated that RNA sequence features (splice sites, conservation levels, and exon/intron architecture) confer strong discriminatory contributions to classify different types of splicing.

Current versions of the splicing code are not able to predict absolute levels of exon inclusion, but rather focus on predictions of relative changes in splicing across tissues or in the presence of genetic mutations. Future development of splicing code could be empowered by consideration of regulatory networks of multiple splicing factors, epigenetic influence and kinetic aspects of splicing, some of which are discussed below.

1.3.5 Useful Databases

Over the years, the splicing community has built many databases and web-resources to include data on global profiling of alternative splicing and systematic analysis of splicing regulatory mechanisms. Table 1.2 summarizes some of these resources ranging from catalogs of alternative splicing events to disease-related mutations that affect splicing.

1.4 Ongoing Questions

1.4.1 Gene Expression Kinetics and Co-Transcriptional Splicing

With the advent of RNA-Seq and related methodologies described previously in this chapter, it is now possible to study kinetics of gene expression and splicing on the global scale. It was recently shown that several steps in RNA processing often but not always occur co-transcriptionally, including capping, splicing, and polyadenylation, allowing for efficient and accurate pre-mRNA maturation (reviewed in (Lee and Rio 2015; de Klerk et al. 2015; Bentley 2014)). In particular, co-transcriptional splicing depends on the rate of RNA Pol II elongation with the idea that slower elongation allows more time for splicing of pre-mRNAs. Pol II elongation can be affected by nucleosome positioning, DNA methylation, histone modifications, and chromatin remodeling (de Almeida and Carmo-Fonseca 2014; Zhou et al. 2014) (Figure 1.3). Additionally, the C-terminal domain of RNA Pol II can be post-translationally and reversibly modified to direct interactions with different proteins involved in RNA processing. Thus, chromatin modifications, transcription, and splicing are all interconnected processes (Bentley 2014; de Almeida and Carmo-Fonseca 2014).

To study dynamic regulation of gene expression and/or co-transcriptional splicing, nascent RNA must be captured. Modified RNA sequencing methods such as genomic run-on sequencing (GRO-Seq) or sequencing of 4-thiouridine labeled RNA may be analyzed in conjunction with RNA-Seq (Rabani et al. 2014; Davis-Turak et al. 2014; de Pretis et al. 2015). Additionally, cell fractionation and selection of non-polyadenylated RNA in the chromatin fraction may be used. Recently, a native elongating transcript sequencing (NET-Seq) approach was used by two groups to identify spliceosome-mediated cleavage, Pol II dynamics related to splicing, and antisense transcription (Nojima et al. 2015; Mayer et al. 2015). Figure 1.3

illustrates co-transcriptional splicing and other events described below including RNA editing, mirtron biogenesis, and circRNA biogenesis.

1.4.2 RNA Modifications

RNA modifications such as methylation (primarily N^6-methyladenosine, or m6A) and RNA editing were not extensively studied until recently. m6A, originally identified in tRNAs, rRNAs, and snoRNAs, was recently shown to be widespread in mRNAs with potential impact on splicing, mRNA degradation, and RNA secondary structures (Chandola et al. 2015; Liu et al. 2015). The most prevalent form of RNA editing is the conversion of adenosine to inosine (A-to-I) via deamination, typically in double stranded RNA (dsRNA) regions by adenosine deaminases acting on RNA (ADARs) (Figure 1.3). In order for editing to affect splicing, it is expected to occur before splicing is completed. Indeed, several lines of evidence suggest editing precedes splicing (reviewed in (Rieder and Reenan 2012)), although exceptions do exist. These findings are only the beginning of a new era of functional and mechanistic studies of RNA modifications.

1.4.3 Splicing Generates Other RNA Species

Though introns are typically degraded after removal, they can also generate other RNA species. For example, biogenesis pathways of snoRNAs, mirtrons, and simtrons rely on intron splicing (reviewed in (Hube et al. 2015)). Whereas canonical miRNA biogenesis depends on the microprocessor (DGCR8 and DROSHA), mirtrons depend on lariat debranching (Figure 1.3) and simtrons depend on U1 snRNP. Another RNA species underappreciated until recently are circular RNAs (circRNAs) (reviewed in (Jeck and Sharpless 2014; Lasda and Parker 2014)). It was shown that biogenesis of certain circRNAs depends on intronic sequence content (Liang and

Wilusz 2014; Ivanov et al. 2014; Wang and Wang 2014a), which may compete with pre-mRNA splicing (Ashwal-Fluss et al. 2014). Additionally, circRNAs can contain both exons and introns, and two of these were shown to regulate gene expression (Li et al. 2015). The splicing factor QKI was shown to regulate production of many circRNAs (Figure 1.3) (Conn et al. 2015). The biogenesis and functions of circRNAs are currently under active investigation.

1.4.4 Global Misregulation of Splicing in Disease

Since splicing is required for RNA maturation, misregulation of splicing may lead to disease states (Cooper et al. 2009). In addition to well-known splicing diseases, such as myotonic dystrophy (Philips et al. 1998), there are several examples of point mutations in specific genes that cause splicing misregulation (reviewed in (Lee and Rio 2015; Zhang and Manley 2013)). Furthermore, global splicing misregulation also characterizes some diseases such as cancer. The Cancer Genome Atlas (TCGA, www.cancergenome.nih.gov) provides a wealth of genomic data from cancer patients, allowing for the study of global splicing alterations within and across cancer types (Brooks et al. 2014; Dorman et al. 2014). Splicing abnormalities were also shown in autistic brains (Irimia et al. 2014). Although splicing alterations in cancer are well established, it is difficult to identify the cause and functional significance of these events, especially considering that up to hundreds of RNA binding proteins may be involved in the regulation of thousands of alternative splicing events in both normal and disease states (Lee and Rio 2015; Zhang and Manley 2013). In the future, an understanding of the causes and functional consequences may lead to splicing-targeted therapeutics.

1.5 Splicing as a Therapeutic Target

Given the critical roles of splicing misregulation in disease, a number of strategies are under development to therapeutically correct aberrant splicing events. First, small molecules can be used to directly modulate the activity of splicing factors (Ohe and Hagiwara 2015). The advantage of this method is the ease of delivery and the potential for individual-specific dosage control. As examples, small molecule inhibitors were examined that target SR protein kinases (SRPKs), CDC2-like kinases (CLKs), or protein phosphatase-1 (PP1), which can then modulate phosphorylation of SR proteins (Figure 1.3). However, such inhibitors often have off-target effects and affect splicing of many genes.

A more targeted approach involves usage of antisense oligonucleotides (ASO), reverse complementary sequences that bind to target mRNA sequences. Because ASOs are sequence-specific, they can block binding of splicing factors at specific loci and modulate alternative splicing. For example, aberrant splicing events caused by an intronic mutation in the human β-globin gene were corrected by ASO treatment in a β-thalassemia mouse model (Svasti et al. 2009). In addition, clinical trials are underway for ASO-based therapy of Duchenne muscular dystrophy and spinal muscular atrophy (Arechavala-Gomeza et al. 2014). Although ASO therapy overcomes the non-specificity issue of small molecules, their delivery is relatively difficult. Another method, trans-splicing, is an effective strategy for repairing multiple mutations in exons or transcripts. Also referred to as Spliceosomal-Mediated RNA trans-splicing (SMaRT) (Wally et al. 2012), this method can replace the entire mRNA sequence 5’ or 3’ of a target splice site by trans-splicing between an ASO and the endogenous RNA (Havens et al. 2013). This approach was proposed as a therapy for β-thalassemia to replace the first exon of the β-globin gene resulting from aberrant splicing (Kierlin-Duncan and Sullenger 2007). However, the delivery of

trans-splicing therapy is also challenging as it necessitates incorporation of DNA vectors to cells (Wang and Cooper 2007).

1.6 Figures

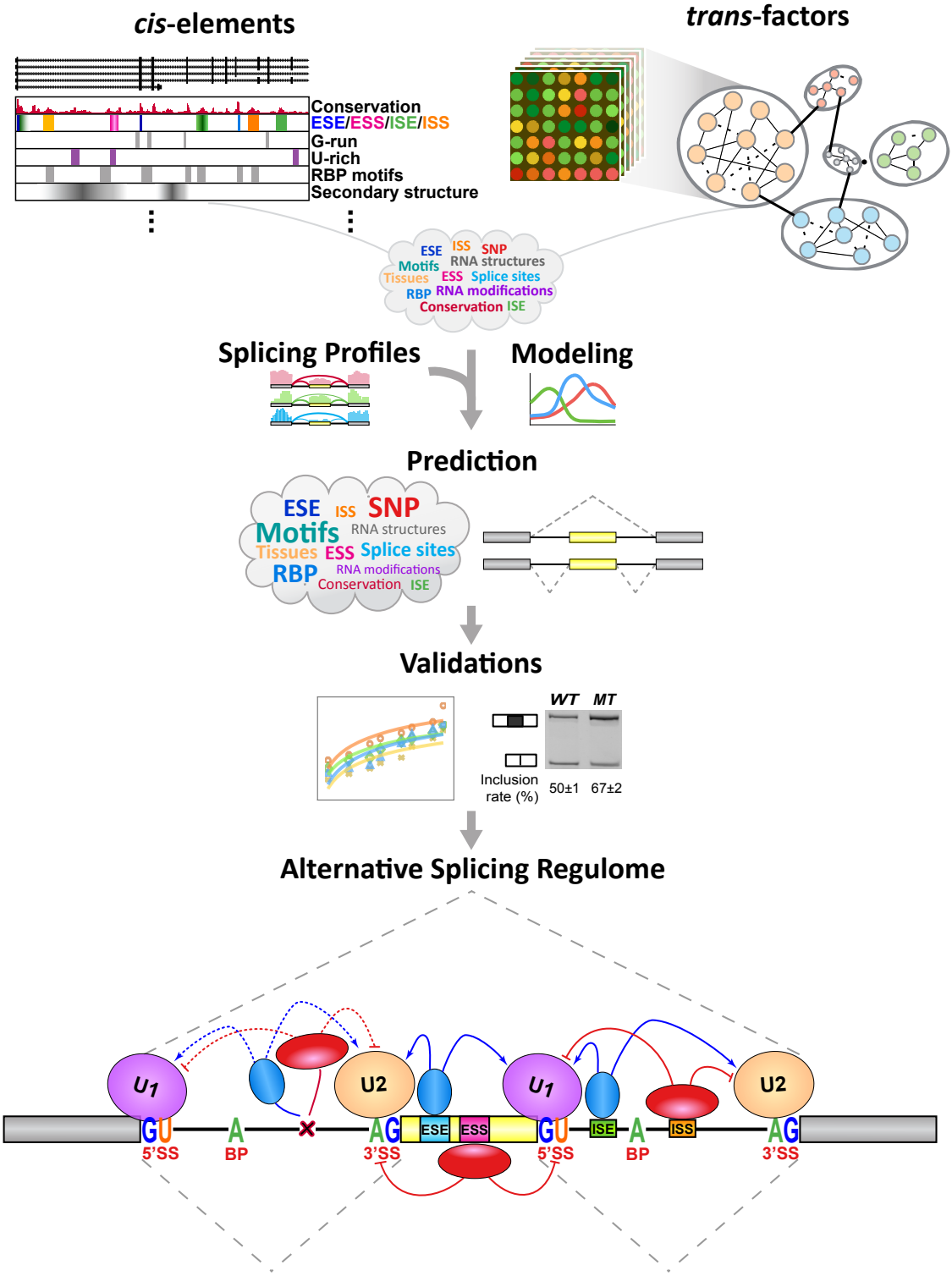

cis-elements
trans-factors
Conservation
ESE/ESS/ISE/ISS
G-run
U-rich
RBP motifs
Secondary structure
ESE ISS SNP
Motifs RNA structures
Tissues ESS Splice sites
RBP RNA modifications
Conservation ISE
Splicing Profiles
Modeling
Prediction
ESE ISS SNP
Motifs RNA structures
Tissues ESS Splice sites
RBP RNA modifications
Conservation ISE
Validations
WT MT
Inclusion rate (%)
50±1 67±2
Alternative Splicing Regulome
U1
U2
U1
U2
GU
A
AG
ESE
ESS
GU
ISE
A
ISS
AG
5'SS
BP
3'SS
5'SS
BP
3'SS

Figure 1.1 Overview of previous studies in alternative splicing regulation. *Cis*-regulatory elements and *trans*-acting factors are key components in the splicing regulatory networks (alternative splicing regulome), which have been actively examined. Combined with global profiles of alternative splicing patterns, bioinformatic models were developed to predict the relative impacts of different regulators and splicing outcome of a given exon. Experimental validations are critical steps to evaluate the accuracy of the predicted splicing regulation. The bottom diagram illustrates well-known components of splicing regulation. The yellow box represents an alternatively skipped exon, which has ESE and ESS motifs that can be recognized by splicing factors. The flanking introns of this exon harbor ISE and ISS motifs. Interactions between the splicing factors and the core splicing machinery (U1, U2 snRNPs etc.) are illustrated. Splicing enhancers (ESEs, ISEs) normally promote exon inclusion, which is represented by the arcs with arrowheads, whereas splicing silencers (ESSs, ISSs) repress exon inclusion, which is represented by flat-headed arcs. Genetic variants may disrupt splicing motifs and alter the binding strength of splicing factors (illustrated by the x). Other mechanisms such as RNA modifications or RNA secondary structures may also affect alternative splicing, which are not illustrated in this diagram. ESE: exonic splicing enhancer; ESS: exonic splicing silencer; ISE: intronic splicing enhancer; ISS: intronic splicing silencer; 5'ss: 5' splice site; 3'ss: 3' splice site; BP: branch point.

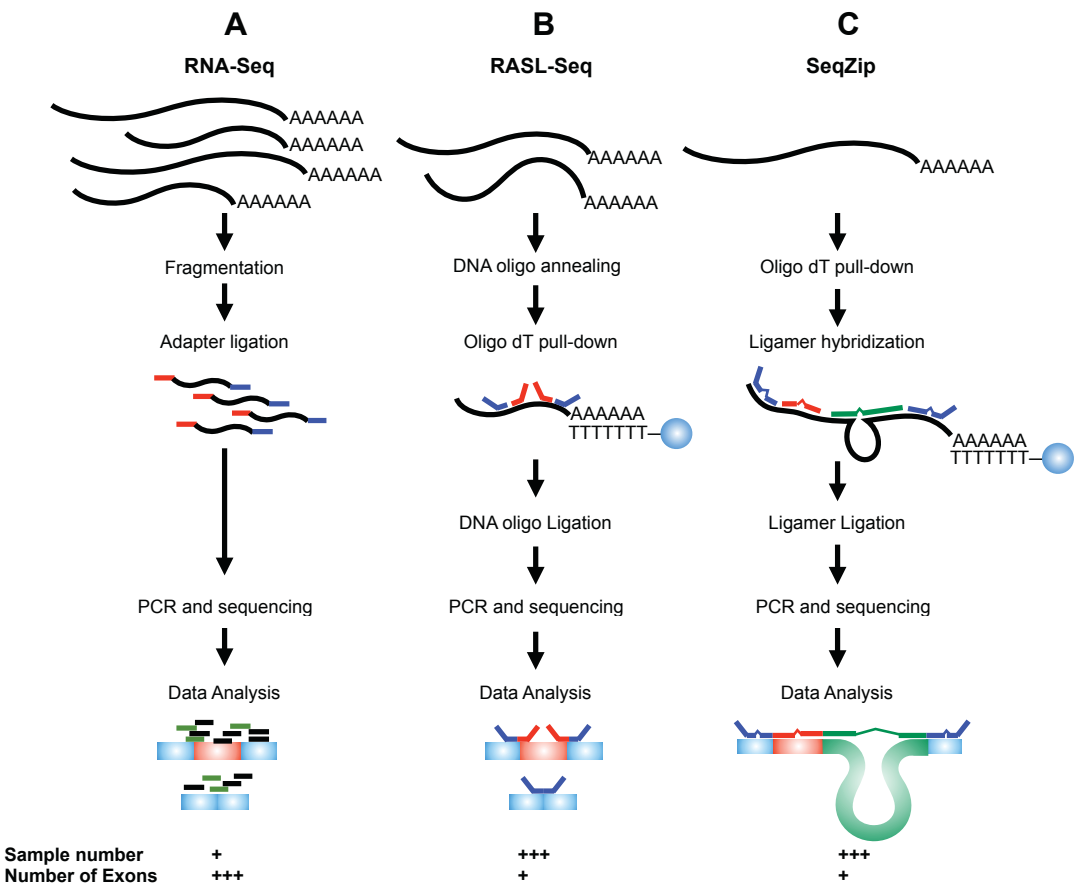


Figure 1.2 High-throughput experimental approaches for splicing detection. (A) RNA-Seq, the most popular method for splicing analysis, begins with creating cDNA libraries of fragmented RNA. Then, sequencing adapters are added to make a sequencing library, followed by PCR amplification and sequencing. In data analysis, reads that span spliced exon junctions and those that are located within exon bodies are identified bioinformatically to detect and quantify alternative splicing. This method can provide data for many expressed exons (+++) in the sample of interest. The cost of RNA-Seq is relatively high, which may limit the number of samples (+) that can be analyzed in a specific study. (B) RASL-Seq requires a pair of pre-designed oligonucleotides that recognize specific splice junctions of intact (i.e. unfragmented) mRNA. Biotinylated oligo-dTs with streptavidin-coated magnetic beads are then used to pull

down the RNA. Barcode incorporation during PCR allows for pooled sequencing of ~1,500 samples per sequencing lane. Compared to RNA-seq, RASL-Seq is ideal for few (up to 500) exons (+) in hundreds or thousands of samples (+++). (C) SeqZip uses DNA "ligamers" to directly sequence long transcript isoforms, causing intermediate regions to loop out. Compared to RNA-Seq and RASL-Seq, SeqZip is specialized for targeting long transcripts.

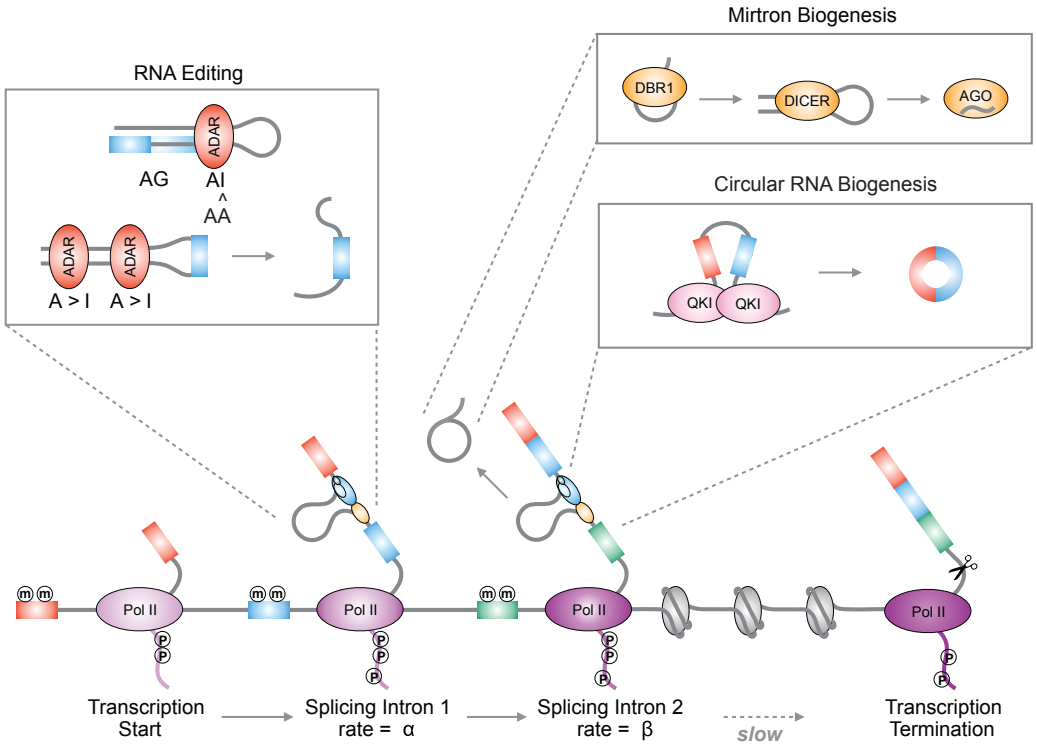


Figure 1.3 Co-transcriptional splicing and related RNA products. Co-transcriptional splicing of two introns with splicing rates α and β is shown. The following epigenetic factors are illustrated: DNA methylation (m) enrichment in exons (Zhou et al. 2014), dynamic phosphorylation (P) of the C-terminal domain of RNA Pol II (de Almeida and Carmo-Fonseca 2014), and nucleosomes slowing down Pol II transcription. Splicing coupled with RNA editing and the biogenesis of mirtrons and circRNAs are shown in the insets. RNA editing can generate

new splice sites (e.g., changing A to I may create a new AG 3'ss, RNA Editing inset, top (Rieder and Reenan 2012)) and prevent circRNA biogenesis (RNA Editing inset, bottom (Ivanov et al. 2014)). Mirtrons are derived from lariats that are de-branched by DBR1 and processed by DICER (Mirtron Biogenesis inset (Hube et al. 2015)). QKI regulates the production of a subset of circRNAs (circRNA Biogenesis inset (Conn et al. 2015)).

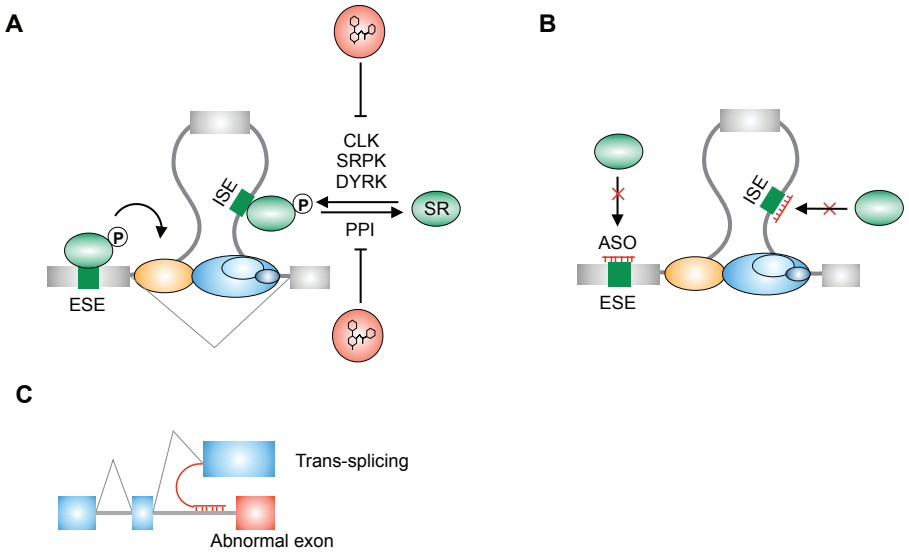


Figure 1.4 Therapeutic approaches to modulate splicing. (A) Small molecule therapy. Phosphorylation or dephosphorylation of SR proteins are regulated by CDC2-like kinase (CLK), dual-specificity tyrosine-(Y)-phosphorylation-regulated kinase (DYRK), SR protein kinase (SRPK) and protein phosphatase-1 (PP1). Inhibitors of these kinases and phosphatase affect the associated splicing events. (B) Antisense oligonucleotides therapy. SR protein or other splicing factor binding sites can be blocked by ASO to achieve specific alternation of splicing. (C) Trans-splicing therapy. ASO linked to a restoring normal exon can rescue an abnormal splicing event that may result due to multiple mutations.

1.7 Tables

Table 1.1. Tools for analysis of alternative splicing using RNA-seq data. AS: alternative splicing; DAS: differential alternative splicing; Spliced Mapper: read aligners specialized in mapping junction reads.

Function	Category	Program	URL	Input	Gene annotation
AS prediction	Isoform-centric	Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/cufflinks	SAM or BAM files	✗
AS prediction	Isoform-centric	eXpress	http://bio.math.berkeley.edu/eXpress	SAM or BAM files	✗
AS prediction	Isoform-centric	Trinity	http://trinityrnaseq.github.io	FASTQ files	✓
AS prediction	Isoform-centric	Trans-ABySS	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	FASTQ files	✓
AS prediction	Isoform-centric	Scripture	http://www.broadinstitute.org/software/scripture	SAM or BAM files	✗
AS prediction	Isoform-centric	RSEM	http://deweylab.biostat.wisc.edu/rsem	FASTA or FASTQ files, transcript expression	✓
AS prediction	Isoform-centric	PennSeq	http://sourceforge.net/projects/pennseq	UCSC Genome Browser gene annotation, SAM files	✓
AS prediction	Isoform-centric	RNA-Skim	http://www.csbio.unc.edu/rs	Specialized transcriptome FASTA, RNA-Seq FASTQ files	✓
AS prediction	Isoform-centric	Sailfish	http://www.cs.cmu.edu/~ckingsf/software/sailfish	*k*-mer size, RNA-Seq in FASTA or FASTQ, transcriptome in GTF	✓
AS prediction	Isoform-centric	Sequgio	http://fafner.meb.ki.se/biostatwiki/sequgio	BAM files	✓
AS prediction	Exon-centric	SplAdder	https://github.com/ratschlab/spladder	GFF annotation, BAM files	✓
AS prediction	Exon-centric	SpliceTrap	http://rulai.cshl.edu/splicetrap/doc/help.html	Gene annotation in BED or GTF format, TXdb exon isoform database, FASTA or FASTQ files	✓
AS prediction	Exon-centric	ESFinder	http://mlg.hit.edu.cn/ybai/ES/ESFinder.html	GTF annotation, UCSC Genome Browser AS events, BAM files	✓
AS prediction	Exon-centric	SplicePie	https://github.com/pulyakhina/splicing_analysis_pipeline	GTF annotation, BAM and BAM index files	✓
AS prediction	Both	MISO	https://miso.readthedocs.org/en/fastmiso	GFF annotation, BAM files	✓
DAS	Isoform-centric	CuffDiff 2	http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html	GFF/GTF annotation, SAM files	✓
DAS	Isoform-centric	IUTA	http://www.niehs.nih.gov/research/resources/software/biostatistics/iuta/index.cfm	GFF annotation, BAM files	✓
DAS	Isoform-centric	SplicingCompass	http://www.ichip.de/software/SplicingCompass.html	Read coverage in GFF, BED files	✓
DAS	Isoform-centric	rSeqDiff	http://www-personal.umich.edu/~jianghui/rseqdiff	".sampling_rates" files from rSeq (Jiang and Wong 2009), gene expression files	✓
DAS	Isoform-centric	FDM	http://csbio-linux001.cs.unc.edu/nextgen/software/FDM	GTF annotation, SAM files	✓
DAS	Isoform-centric	rDiff	http://bioweb.me/rdiff	GFF annotation, BAM files	✓

DAS	Exon-centric	DEXSeq	http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html	GFF/GTF annotation, SAM files	✓
DAS	Exon-centric	DSGSeq	http://bioinfo.au.tsinghua.edu.cn/software/DSGseq	BAM and BED files	✓
DAS	Exon-centric	DiffSplice	http://www.netlab.uky.edu/p/bioinfo/DiffSplice	Parsed SAM files, program configuration files	✗
DAS	Exon-centric	dSpliceType	http://dsplicetype.sourceforge.net	GFF annotation, read coverage in bedgraph format, junction files in BED format	✓
DAS	Exon-centric	MATS/rMATS	http://rnaseq-mats.sourceforge.net	GFF annotation, FASTQ or BAM files, Bowtie index	✓
DAS	Exon-centric	rSeqNP	http://www-personal.umich.edu/~jianghui/rseqnp	Transcript expression files	✗
DAS	Both	MISO	https://miso.readthedocs.org/en/fastmiso	GFF annotation, BAM files	✓
DAS	Both	SUPPA	https://bitbucket.org/regulatorygenomicsupf/suppa	GTF annotation, transcript expression files	✓
Spliced Mapper	***NA***	HMMSplicer	http://derisilab.ucsf.edu/software/hmmsplicer	FASTQ files, Bowtie index	✓
Spliced Mapper	***NA***	PASSion	https://trac.nbic.nl/passion	FASTQ files, SMALT index	✓
Spliced Mapper	***NA***	PASTA	http://www.biotech.ufl.edu/cores/bioinformatics/dibig/dibig-software/pasta	FASTQ files, Bowtie index	✓
Spliced Mapper	***NA***	OLego	http://zhanglab.c2b2.columbia.edu/index.php/OLego	FASTA or FASTQ files, BWT index	✓
Spliced Mapper	***NA***	TrueSight	http://bioen-compbio.bioen.illinois.edu/TrueSight/	FASTQ files, Bowtie index	✓
Spliced Mapper	***NA***	UnSplicer	http://opal.biology.gatech.edu/paul/unsplicer/index.htm	FASTQ files, Bowtie index	✓
Spliced Mapper	***NA***	JAGuaR	http://www.bcgsc.ca/platform/bioinfo/software/jaguar	FASTQ files, BWA index	✓
Spliced Mapper	***NA***	Rail-RNA	https://github.com/nellore/rail	FASTQ files, Bowtie index	✓

Table 1.2. Databases (DB) of alternative splicing events and software programs (PR) for studies of splicing regulation.

Type	Category	Database	URL	Function	Input	Notes
DB	AS Events	DBASS	http://www.dbass.org.uk	Catalogs splice site mutations and diseases	Chosen from user drop-down menu	Includes DBASS3 and DBASS5
DB	Data	HGMD	http://www.hgmd.cf.ac.uk/ac/index.php	Provides disease-associated variants	-	Requires license to access the full database
DB	Data	ENCODE	https://www.encodeproject.org	Provides various NGS datasets	Chosen from user drop-down menu	-
DB	Data	RBPDB	http://rbpdb.ccbr.utoronto.ca	Provides RBP PWMs	-	-
DB	Data	CisBP-RNA	http://cisbp-rna.ccbr.utoronto.ca	Provides RBP PWMs	-	-
DB	Data	CLIPdb	http://lulab.life.tsinghua.edu.cn/clipdb	Catalogs references for RBP studies and CLIP datasets	RBP names, cell types, species, or technology types	Also supports database browsing without requiring input

DB	Data	CLIPZ	http://www.clipz.unibas.ch	Provides CLIP datasets with target predictions of RBPs and miRNAs	Chosen from user drop-down menu	Require account sign-up; reviewc(Reyes-Herrera and Ficarra 2014)
DB	Data	doRINA 2.0	http://dorina.mdc-berlin.de	Provides CLIP datasets with RBP sites annotations	Chosen from user drop-down menu	Review (Reyes-Herrera and Ficarra 2014)
DB	Data	StarBase V2.0	http://starbase.sysu.edu.cn/index.php	Provides CLIP datasets with annotations	Chosen from user drop-down menu	Review (Reyes-Herrera and Ficarra 2014)
DB	Tools	OMICtools	http://omictools.com	Databases of genomic, transcriptomic, proteomic, and metabolomic tools	Tool names, analysis type, or web-browsing	-
DB	Regulation	RegulomeDB	http://regulomedb.org	Displays various regulatory tracks for the input sequences	dbSNP IDs, genomic coordinates in BED files, VCF files, or GFF3 files	Links regulation to GWAS: http://regulomedb.org/GWAS/index.html
DB	Regulation	RegRNA2.0	http://regrna2.mbc.nctu.edu.tw	Displays various regulatory tracks for the input sequences	RNA sequences in FASTA format	-
DB	Regulation	Brain RNA-Seq	http://web.stanford.edu/group/barres_lab/brain_rnaseq.html	Provides FPKM data in various brain cells; compare gene enrichment across available cell types	Gene name, or cell types	Done in mouse; also provides data browsing
DB	Regulation	rSNPBase	http://rsnp.psych.ac.cn	Provides regulatory annotation for SNPs	SNP ID or gene names	-
PR	Regulation	CRYP-SKIP	http://cryp-skip.img.cas.cz	Predicts splice site mutations	Nucleotide sequence in FASTA format	-
PR	Regulation	EX-SKIP	http://ex-skip.img.cas.cz	Compares a SNV's impact on ESE/ESS to induce exon skipping	Two exonic sequences in FASTA format	Maximum 4000nt per submission
PR	Regulation	Human Splicing Finder	http://www.umd.be/HSF	Combines 12 different algorithms to predict mutations' impact on *cis*-regulatory elements	Chosen from user drop-down menu	-
PR	Regulation	MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html; http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html	Predicts splice site strength	Splice site sequences in FASTA format	Web-based or command line-based analyses available
PR	Regulation	WASP	http://genes.toronto.edu/wasp	Predicts AS exons and the potential regulation codes	RNA sequence in FASTA format or genomic coordinates in BED format	Maximum 10 input exons per query
PR	Regulation	AVISPA	http://avispa.biociphers.org	Predicts AS exons and the potential regulation codes	FASTA or BED files containing a single putative AS exon or cassette exon triplet	-
PR	Regulation	SPANR	http://tools.genes.toronto.edu	Predicts SNVs effects on cassette exons	Maximum 40 SNVs per file in tab-delimited VCF format	This tool was designed for detecting exon-skipping events, so it may or may not work for other AS types.

1.8 References

Arechavala-Gomeza V, Khoo B, Aartsma-Rus A. 2014. Splicing modulation therapy in the treatment of genetic diseases. *Appl Clin Genet* **7**: 245–252.

Arias MA, Lubkin A, Chasin LA. 2015. Splicing of designer exons informs a biophysical model for exon definition. *RNA* **21**: 213–229.

Ars E, Kruyer H, Gaona A, Serra E, Lazaro C, Estivill X. 1999. Prenatal diagnosis of sporadic neurofibromatosis type 1 (NF1) by RNA and DNA analysis of a splicing mutation. *Prenat Diagn* **19**: 739–742.

Aschoff M, Hotz-Wagenblatt A, Glatting K-H, Fischer M, Eils R, König R. 2013. SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* **29**: 1141–8.

Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. 2014. circRNA Biogenesis Competes with Pre-mRNA Splicing. *Mol Cell* **56**: 55–66.

Badr E, Heath LS. 2014. Identifying splicing regulatory elements with de Bruijn graphs. *J Comput Biol* **21**: 880–97.

Bahn JH, Lee J-HH, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–50.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**: 53–9.

Barash Y, Vaquero-Garcia J, González-Vallinas J, Xiong HY, Gao W, Lee LJ, Frey BJ, Gonzalez-Vallinas J, Xiong HY, Gao W, et al. 2013. AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biol* **14**: R114.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587–93.

Benhalevy D, McFarland HL, Sarshad AA, Hafner M. 2016. PAR-CLIP and streamlined small RNA cDNA library preparation protocol for the identification of RNA binding protein target sites. *Methods*.

Bentley DL. 2014. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* **15**: 163–75.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**: 3171–3175.

Bitton DA, Rallis C, Jeffares DC, Smith GC, Chen YYC, Codlin S, Marguerat S, Bahler J, Bähler J. 2014. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res* **24**: 1169–79.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291–336.

Blencowe BJ. 2006. Alternative splicing: new insights from global analyses. *Cell* **126**: 37–47.

Brooks AN, Choi PS, de Waal L, Sharifnia T, Imielinski M, Saksena G, Pedamallu CS, Sivachenko A, Rosenberg M, Chmielecki J, et al. 2014. A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS One* **9**: e87361.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol* **24**: 10505–14.

Busch A, Hertel KJ. 2015. Splicing predictions reliably classify different types of alternative

splicing. *RNA* **21**: 813–23.

Calarco JA, Superina S, O'Hanlon D, Gabut M, Raj B, Pan Q, Skalska U, Clarke L, Gelinas D, van der Kooy D, et al. 2009. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138**: 898–910.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. 2003. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31**: 3568–3571.

Chandola U, Das R, Panda B. 2015. Role of the N6-methyladenosine RNA mark in gene regulation and its implications on development and disease. *Br Funct Genomics* **14**: 169–179.

Chen L. 2013. Statistical and Computational Methods for High-Throughput Sequencing Data Analysis of Alternative Splicing. *Stat Biosci* **5**: 138–155.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**: 1–8.

Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, Roslan S, Schreiber AW, Gregory PA, Goodall GJ. 2015. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160**: 1125–1134.

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136**: 777–793.

Corvelo A, Hallegger M, Smith CWJ, Eyras E. 2010. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* **6**: e1001016.

Culler SJ, Hoff KG, Voelker RB, Berglund JA, Smolke CD. 2010. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Res* **38**: 5152–5165.

Cuperlovic-Culf M, Belacel N, Culf AS, Ouellette RJ. 2006. Microarray analysis of alternative

splicing. *OMICS* **10**: 344–57.

Davis-Turak JC, Allison K, Shokhirev MN, Ponomarenko P, Tsimring LS, Glass CK, Johnson TL, Hoffmann A. 2014. Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. *Nucleic Acids Res* **43**: 699–707.

de Almeida SF, Carmo-Fonseca M. 2014. Reciprocal regulatory links between cotranscriptional splicing and chromatin. *Semin Cell Dev Biol* **32**: 2–10.

de Klerk E, 't Hoen PAC, Djebali S, Al. E, Neph S, Al. E, Sanyal A, Al. E, (DGT) FC and RP and C, Goossens S, et al. 2015. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet* **31**: 128–139.

de Pretis S, Kress T, Morelli MJ, Melloni GE, Riva L, Amati B, Pelizzola M. 2015. INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics*.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207–3212.

Desmet F-OO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C, Collod-Béroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* **37**: e67.

Desmet FO, Beroud C, Béroud C. 2012. Bioinformatics and mutations leading to exon skipping ed. A. Aartsma-Rus. *Methods Mol Biol* **867**: 17–35.

Dorman SN, Viner C, Rogan PK. 2014. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci Rep* **4**: 7063.

Fairbrother WG, Yeh R-FF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**: 1007–13.

Friedman BA, Stadler MB, Shomron N, Ding Y, Burge CB. 2008. Ab initio identification of functionally interacting pairs of cis-regulatory elements. *Genome Res* **18**: 1643–1651.

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

Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* **36**: 2257–2267.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**: 469–477.

Gooding C, Clark F, Wollerton MC, Grellscheid S-NN, Groom H, Smith CWJ. 2006. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol* **7**: R1.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503–510.

Han A, Stoilov P, Linares AJ, Zhou Y, Fu X-DD, Black DL. 2014. De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Comput Biol* **10**: e1003442.

Havens MA, Duelli DM, Hastings ML. 2013. Targeting RNA splicing for disease therapy. *Wiley Interdiscip Rev RNA* **4**: 247–266.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P. 2014. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56**: 61–4, 66, 68, passim.

Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, et al. 2010. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet* **19**: 122–134.

Hooper JE. 2014. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum Genomics* **8**: 3.

Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, Thomas J, Wang Y. 2001. Predicting splice variant from DNA chip expression data. *Genome Res* **11**: 1237–1245.

Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan P-FF, Hammond SM, Makowski L, et al. 2013. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res* **41**: e39.

Hu Y, Liu Y, Mao X, Jia C, Ferguson JF, Xue C, Reilly MP, Li H, Li M. 2014. PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. *Nucleic Acids Res* **42**: e20.

Hube F, Francastel C, Hubé F, Francastel C. 2015. Mammalian introns: when the junk generates molecular diversity. *Int J Mol Sci* **16**: 4429–4452.

Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, et al. 2012. Integrative genome-wide analysis reveals cooperative regulation of

alternative splicing by hnRNP proteins. *Cell Rep* **1**: 167–78.

Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallieres M, Tapial J, Raj B, O'Hanlon D, et al. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**: 1511–1523.

Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, et al. 2014. Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals. *Cell Rep* **10**: 170–7.

Izquierdo JM, Majos N, Bonnal S, Martinez C, Castelo R, Guigo R, Bilbao D, Valcarcel J. 2005. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell* **19**: 475–484.

Jeck WR, Sharpless NE. 2014. Detecting and characterizing circular RNAs. *Nat Biotechnol* **32**: 453–461.

Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* **42**: 13534–44.

Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**: 1026–1032.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**: 2141–4.

Kahles A, Ong CS, Rätsch G, Kahles A Ratsch G OCS. 2015. SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data. *bioRxiv* 017095.

Kalsotra A, Cooper TA. 2011. Functional consequences of developmentally regulated alternative

splicing. *Nat Rev Genet* **12**: 715–29.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–15.

Ke S, Chasin LA. 2010. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biol* **11**: R84.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* **21**: 1360–1374.

Kierlin-Duncan MN, Sullenger BA. 2007. Using 5'-PTMs to repair mutant beta-globin transcripts. *RNA* **13**: 1317–1327.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909–915.

Kratz A, Carninci P. 2014. The devil in the details of RNA-seq. *Nat Biotechnol* **32**: 882–884.

Krawczak M, Reiss J, Cooper DN. 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* **90**: 41–54.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40**: 225–31.

Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. 2014. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol Cell* **54**: 887–900.

Larman HB, Scott ER, Wogan M, Oliveira G, Torkamani A, Schultz PG. 2014. Sensitive, multiplex and direct quantification of RNA sequences using a modified RASL assay. *Nucleic Acids Res* **42**: 9146–9157.

Lasda E, Parker R. 2014. Circular RNAs: diversity of form and function. *RNA* **20**: 1829–1842.

Lee C, Roy M. 2004. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol* **5**: 231.

Lee J-H, Gao C, Peng G, Greer C, Ren S, Wang Y, Xiao X. 2011. Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ Res* **109**: 1332–41.

Lee J-HH, Ang JK, Xiao X. 2013. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA* **19**: 725–32.

Lee Y, Rio DC. 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem* **84**: 291–323.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**: 733–739.

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

Li G, Bahn JH, Lee J-HH, Peng G, Chen Z, Nelson SF, Xiao X. 2012a. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res* **40**: e104.

Li H, Qiu J, Fu X-DD. 2012b. RASL-seq for massively parallel and quantitative analysis of gene expression. *Curr Protoc Mol Biol* **Chapter 4**: Unit 4 13 1-9.

Li H, Zhou H, Wang D, Qiu J, Zhou Y, Li X, Rosenfeld MG, Ding S, Fu XD. 2012c. Versatile pathway-centric approach based on high-throughput sequencing to anticancer drug discovery. *Proc Natl Acad Sci U S A* **109**: 4609–4614.

Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al. 2015. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* **22**: 256–264.

Liang D, Wilusz JE. 2014. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev* **28**: 2233–2247.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**: 464–469.

Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. 2015. N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* **518**: 560–564.

Liu Y, Ferguson JF, Xue C, Silverman IM, Gregory B, Reilly MP, Li M. 2013. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLoS One* **8**: e66883.

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

Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**: 72–9.

Mark D, Haeberle S, Roth G, von Stetten F, Zengerle R. 2010. Microfluidic lab-on-a-chip platforms: requirements, characteristics and applications. *Chem Soc Rev* **39**: 1153–1182.

Matlin AJ, Clark F, Smith CWJ. 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**: 386–398.

Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS. 2015. Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell* **161**: 541–554.

McHugh CA, Russell P, Guttman M. 2014. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol* **15**: 203.

Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**: gr.182899.114.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. 2012. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30**: 99–104.

Monlong J, Calvo M, Ferreira PG, Guigó R, Guigo R. 2014. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun* **5**: 4698.

Mort M, Sterne-Weiler T, Li B, Ball E V, Cooper DN, Radivojac P, Sanford JR, Mooney SD. 2014. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol* **15**: R19.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–63.

Nojima T, Gomes T, Grosso ARFRF, Kimura H, Dye MJJ, Dhir S, Carmo-Fonseca M,

Proudfoot NJJ. 2015. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**: 526–540.

Ohe K, Hagiwara M. 2015. Modulation of alternative splicing with chemical compounds in new therapeutics for human diseases. *ACS Chem Biol* **10**: 914–924.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–1415.

Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, et al. 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* **16**: 929–41.

Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**: 533–8.

Pastuszak AW, Joachimiak MP, Blanchette M, Rio DC, Brenner SE, Frankel AD. 2011. An SF1 affinity model to identify branch point sequences in human introns. *Nucleic Acids Res* **39**: 2344–56.

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32**: 462–4.

Philips A V, Timchenko LT, Cooper TA. 1998. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science (80-)* **280**: 737–741.

Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet* **24**: 590–4.

Poulos MG, Batra R, Charizanis K, Swanson MS. 2011. Developments in RNA splicing and disease. *Cold Spring Harb Perspect Biol* **3**: a000778.

Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJJ, Pauli A, Hacohen N, Schier AFF, Blackshear PJJ, Friedman N, et al. 2014. High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. *Cell* **159**: 1698–1710.

Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR, Pena Castillo L, et al. 2009. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* **27**: 667–70.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–7.

REESE MG, EECKMAN FH, KULP D, HAUSSLER D. 1997. Improved Splice Site Detection in Genie. *J Comput Biol* **4**: 311–323.

Reid DC, Chang BL, Gunderson SI, Alpert L, Thompson WA, Fairbrother WG. 2009. Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA* **15**: 2385–2397.

Reyes-Herrera PH, Ficarra E. 2014. Computational Methods for CLIP-seq Data Processing. *Bioinform Biol Insights* **8**: 199–207.

Rieder LE, Reenan RA. 2012. The intricate relationship between RNA structure, editing, and splicing. *Semin Cell Dev Biol* **23**: 281–288.

Roy CK, Olson S, Graveley BR, Zamore PD, Moore MJ. 2015. Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNA ligation. *Elife* **4**.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671–684.

Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* **18**: 88–103.

Shen S, Park JW, Lu Z-X, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**: E5593-601.

Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. 2006. Inference of splicing regulatory activities by sequence neighborhood analysis ed. Y. Hayashizaki. *PLoS Genet* **2**: e191.

Steijger T, Abril JF, Engström PG, Kokocinski F, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P, et al. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**: 1177–1184.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res* **21**: 1563–71.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**: 956–60.

Suo C, Calza S, Salim A, Pawitan Y. 2014. Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data. *Bioinformatics* **30**: 506–13.

Svasti S, Suwanmanee T, Fucharoen S, Moulton HM, Nelson MH, Maeda N, Smithies O, Kole R. 2009. RNA repair restores hemoglobin expression in IVS2-654 thalassemic mice. *Proc Natl Acad Sci U S A* **106**: 1205–1210.

Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG. 2012. Large-scale mapping

of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* **19**: 719–721.

Teraoka SN, Telatar M, Becker-Catania S, Liang T, Onengut S, Tolun A, Chessa L, Sanal O, Bernatowska E, Gatti RA, et al. 1999. Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am J Hum Genet* **64**: 1617–1631.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (80-)* **249**: 505–510.

Ule J, Jensen K, Mele A, Darnell RB. 2005. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**: 376–86.

Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL. 2005. Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. *Mol Cell Biol* **25**: 10005–10016.

van Dijk EL, Jaszczyszyn Y, Thermes C. 2014. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* **322**: 12–20.

Venkatesh B, Yap WH. 2005. Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *Bioessays* **27**: 100–107.

Wahl MC, Will CL, Luhrmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701–718.

Wally V, Murauer EM, Bauer JW. 2012. Spliceosome-mediated trans-splicing: the therapeutic cut and paste. *J Invest Dermatol* **132**: 1959–1966.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.

Wang G-SS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**: 749–61.

Wang Y, Ma M, Xiao X, Wang Z. 2012. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol* **19**: 1044–1053.

Wang Y, Wang Z. 2014a. Efficient backsplicing produces translatable circular mRNAs. *RNA* **21**: 172–9.

Wang Y, Wang Z. 2014b. Systematical identification of splicing regulatory cis-elements and cognate trans-factors. *Methods* **65**: 350–358.

Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z. 2013. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol* **20**: 36–45.

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

Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, Lee MP. 2003. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res* **63**: 655–657.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**: 831–45.

Weyn-Vanhentenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, Xue C, Herre M, Silver PA, Zhang MQ, et al. 2014. HITS-CLIP and integrative modeling define the Rbfox

splicing-regulatory network linked to brain development and autism. *Cell Rep* **6**: 1139–1152.

Wulff BE, Sakurai M, Nishikura K. 2011. Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. *Nat Rev Genet* **12**: 81–85.

Xiao X, Wang Z, Jang M, Burge CB. 2007. Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci* **104**: 18583–18588.

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. 2009. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat Struct Mol Biol* **16**: 1094–100.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2014. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, December 18.

Xiong HY, Barash Y, Frey BJ. 2011. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**: 2554–62.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377–94.

Yeo G, Hoon S, Venkatesh B, Burge CB. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* **101**: 15700–15705.

Yu Y, Maroney PA, Denker JA, Zhang XH-F, Dybkov O, Lührmann R, Jankowsky E, Chasin LA, Nilsen TW, Luhrmann R, et al. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell* **135**: 1224–36.

Zhang C, Darnell RB. 2011. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* **29**: 607–614.

Zhang C, Frias MA, Mele A, Ruggiu M, Eom T, Marney CB, Wang H, Licatalosi DD, Fak JJ, Darnell RB. 2010. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science (80-)* **329**: 439–443.

Zhang F, Wang M, Michael T, Drabier R. 2013. Novel alternative splicing isoform biomarkers identification from high-throughput plasma proteomics profiling of breast cancer. *BMC Syst Biol* **7 Suppl 5**: S8.

Zhang J, Kuo C-CJC, Chen L. 2012. VERSE: a varying effect regression for splicing elements discovery. *J Comput Biol* **19**: 855–865.

Zhang J, Manley JL. 2013. Misregulation of pre-mRNA alternative splicing in cancer. *Cancer Discov* **3**: 1228–37.

Zhang XH-F, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**: 1241–1250.

Zhang XH-F, Heller KA, Hefter I, Leslie CS, Chasin LA. 2003. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res* **13**: 2637–2650.

Zhang Z, Wang W. 2014. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics* **30**: i283–i292.

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

Zhou H-LL, Luo G, Wise JA, Lou H. 2014. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res* **42**: 701–13.

CHAPTER 2

Alternative Splicing Modulated by Genetic Variants Demonstrates Accelerated Evolution Regulated by Highly Conserved Proteins

2.1 Introduction

Since the completion of the human genome project, major efforts have been devoted to genome-wide association studies (GWAS) with the ultimate goal of elucidating the genetic underpinnings of human diseases or phenotypic traits (Evangelou and Ioannidis 2013). It is now clear that many disease-associated genetic variants are located in non-coding regions whose functional relevance is much harder to interpret than coding variants. Thus, prediction and validation of functional variants are imperative tasks of the post-genomic era. Recent work integrating genomic and bioinformatic analyses made significant progress in prediction and prioritization for causal genetic variants, but mostly focusing on their involvement in transcriptional control (reviewed in (Li et al. 2014)). In addition to transcriptional regulation, genetic variants located in exons or introns may affect alternative splicing (Wang and Cooper 2007; Xiao and Lee 2010), an aspect that is not yet widely appreciated.

It is known that both *cis*- and *trans*-genetic variants can affect alternative splicing, which may contribute to the etiology, susceptibility or progression of diseases (Wang and Cooper 2007; Sterne-Weiler and Sanford 2014). Previous studies reported that 15-60% of disease-causing mutations may alter splicing, an estimate mainly based on examination of splice site mutations (Wang and Cooper 2007). Recently, a wide spectrum of *cis*-regulatory elements of splicing was

identified, known as splicing enhancers and silencers (Wang and Burge 2008), and proved essential in predicting splicing levels (Barash et al. 2010; Busch and Hertel 2015). Thus, the prevalence of splicing-altering genetic variants that disrupt *cis*-regulatory elements could be much higher than previously appreciated.

To identify such genetic variants, a number of studies examined splicing quantitative trait loci (sQTL) in cell lines derived from human populations (Zhao et al. 2013; Monlong et al. 2014; Kwan et al. 2008). This approach evaluates statistical associations between genotypes and splicing phenotypes (or RNA isoforms) that necessitates a large number of parallel data sets from individuals of diverse genetic background. We previously developed a second type of approach that examines allele-specific expression patterns of genetic variants to identify splicing events under regulation by these variants (Li et al. 2012). Applicable to a single RNA sequencing (RNA-seq) data set, this method identifies *cis*-regulatory variants independent of *trans*-acting effects. Machine learning-based techniques have also been applied to score genetic variants that affect splicing (Mort et al. 2014; Xiong et al. 2014; Sterne-Weiler et al. 2011). In general, methodology development is still the focus of most studies in this area. The number of experimentally validated functional variants in alternative splicing remains relatively small. Consequently, little is known about the genomic, evolutionary and regulatory features of genetically modulated alternative splicing (GMAS).

To this end, we conducted a systematic analysis of both intronic and exonic genetic variants involved in splicing modulation capitalizing on the cellular compartment-specific human RNA-seq data generated by the ENCODE Project (Djebali et al. 2012). We examined the allele-

specific expression patterns of single nucleotide variants (SNVs) and analyzed nuclear and cytosolic RNA contents comparatively. Our analyses identified more than 600 GMAS exons and associated SNVs, which enabled a detailed examination of the global features of these events. We observed that *cis*-regulatory variation is the major driving force of splicing variation in GMAS, often rendering cell type-independent splicing phenotypes. Interestingly, GMAS-associated genes, exons and SNVs demonstrated significant bias reflecting positive or relaxed evolutionary selection in human and other primates. In contrast, GMAS events are likely regulated by highly conserved splicing factors with strong consensus motifs that are susceptible to SNV-disruption. We also analyzed the specific nucleotide positions of splicing factor motifs disrupted by GMAS SNVs, which yielded important insights that have both functional and evolutionary implications. Using gel shift assays and CLIP-seq data, we confirmed the allele-specific binding features of one splicing factor *SRSF1*, a major regulator of GMAS events. More than 100 (18%) GMAS SNVs were in linkage disequilibrium (LD) with GWAS SNPs, accounting for the possible function of 99 GWAS SNPs residing deep in the introns. Our study reports a large number of GMAS events and enables a better understanding of the evolutionary and regulatory features of this phenomenon.

2.2 Results

2.2.1 Compartment-specific RNA-seq enhances coverage of intronic SNVs

Compartment-specific RNA-seq data sets derived from nuclear and cytosolic RNA of a number of cell lines were obtained by the ENCODE Project (Djebali et al. 2012). In addition, polyadenylated RNA (polyA+) and RNA without polyA tails (polyA-) were processed into

separate RNA-seq libraries. Thus, for each cell line, a total of four types of data are available: nuclear polyA+ (NA+), nuclear polyA- (NA-), cytosolic polyA+ (CA+) and cytosolic polyA- (CA-) (Supplemental Table 1). Our study focused on the first three types of data. We hypothesized that the nuclear-specific data sets represent enriched RNA content compared to that captured by traditional polyA+ RNA-seq, with a specific advantage of enhanced coverage of intronic RNA. As shown in Figure 1A, read distribution of the RNA-seq data showed that intronic RNA was the most enriched in the NA- data and the least abundant in the CA+ data set, consistent with expectations. We next examined the expression of SNVs in the RNA-seq data by analyzing their allele-specific expression (ASE) patterns (Li et al. 2012). As expected, the NA- data set yielded the highest percentage of intronic SNVs with ASE (Fig. 1B). Since the NA- RNA fraction is enriched with nascent RNA prior to completion of splicing and spliced introns to be degraded or undergoing degradation (Fig. 1C), this fraction represents additional information that is not normally captured in standard polyA+ RNA-seq.

2.2.2 Identification of intronic tag SNVs for genetic modulation of alternative splicing

We analyzed the above data sets to identify intronic tags for genetic modulation of alternative splicing (iGMAS). In general, if an alternative splicing event is regulated by genetic variants, the exon demonstrates allele-specific splicing patterns depending on the alleles of the causal variant. If other exonic or intronic variants exist in LD with the causal variant, they also exhibit corresponding ASE profiles. Furthermore, if the causal or LD variants reside in the introns, the spliced-out product (i.e., spliced introns) is expected to display allele-specific bias. In the NA- data set, such spliced-out products may be captured to identify allele-specific alternative splicing

events. Thus, we focused on read pairs that may represent spliced introns of alternative splicing to inform a search for iGMAS SNVs, given the paired-end NA- data.

Specifically, we anchored the search on heterozygous intronic SNVs covered by reads whose corresponding read pairs mapped to the neighboring exon or the other flanking intron of the exon (Fig. 1C, red arrows), and asked whether these reads were enriched with one of the two alleles of the intronic SNVs. Since these read pairs might have originated from spliced products where the exon was alternatively skipped, such an allelic bias may reflect existence of allele-specific alternative splicing. However, the above read pairs may also be generated from nascent RNA transcripts prior to splicing completion, in which case the allelic bias of the intronic SNVs was resulted from ASE of the entire gene. We thus examined the allelic expression ratios of the transcripts and excluded those with predicted gene-level ASE. Following this step, we tested the null hypothesis that intronic SNVs covered by read pairs as described above were not expressed in an allele-specific manner in the NA- data (Methods). The iGMAS SNVs were then defined as those associated with a rejected null hypothesis. Note that another possibility that leads to the above observed allelic bias is allele-specific intron degradation. Since this phenomenon was barely reported in the literature, we hypothesize that it is a rare mechanism and that GMAS accounts for the majority of the observed allelic bias (which is tested by experimental validation using splicing reporters below).

As a proof of principle, we first applied this method to the ENCODE data sets derived from the GM12878 cell line. In the NA- data set, a total of 3476 heterozygous intronic SNVs, not residing in genes with whole gene-level ASE, were identified to be associated with read pairs

covering the SNVs and their respective flanking exons or introns. Among these SNVs, 35 had significant allelic bias and were identified as iGMAS SNVs, an example of which is shown in Figure 2A. The small fraction of iGMAS SNVs among all testable intronic SNVs may partly reflect the stringency in defining significance in our method to enhance accuracy.

To evaluate the accuracy of this method, we applied a read-randomization procedure to estimate the empirical FDR (Methods). Based on this simulation, the predicted iGMAS SNVs had an FDR of < 3% (Fig. 2B). As an alternative test of performance, we focused on the specific type of iGMAS events where there exists a heterozygous exonic SNV in addition to the intronic one. For a true iGMAS event, we expect to observe "opposite allelic ratios" for the exonic SNV in the NA- data compared to the CA+ RNA-seq data (Fig. 1C). That is, the allele enriched in the NA- data (among read pairs covering the intronic and exonic SNVs) should be relatively depleted in the CA+ RNA-seq data compared to its counterpart allele. Indeed, all five such exonic SNVs associated with predicted iGMAS events demonstrated opposite allelic ratios between NA- and CA+ fractions, attesting to the validity of this method.

2.2.3 iGMAS events identified in seven ENCODE cell lines

We applied the method to data derived from six additional ENCODE human cell lines: K562, HeLa, HepG2, HUVEC, NHEK, and H1-hESC. Combined with those yielded from the GM12878 data, a total of 174 unique iGMAS SNVs were identified, associated with 190 AS exons (Supplemental Fig. 1, Supplemental Table 2). The number of predicted iGMAS events differs for different data sets partly because of the variation in the amount of mapped reads across data sets (Supplemental Fig. 1, Supplemental Table 1). In addition, the genetic

background of each cell line also contributes to this difference because our method inherently requires presence of intronic SNVs close to exons. Another variable is the insert size of the libraries (Supplemental Fig. 2) since pairs of reads that overlap the intronic SNVs and the neighboring exonic regions were used whose abundance is restrained by the insert size of RNA-seq libraries. Among all predicted iGMAS SNVs and related genes, one SNP and five genes were present in two or more cell lines. This low level of overlap possibly reflects genetic diversity across the cell lines and/or the existence of a much larger number of genetically regulated splicing events not yet identified in this analysis.

2.2.4 Experimental validation of iGMAS predictions

We randomly picked 10 predicted iGMAS events from the results of three cell lines for validation using a minigene system (Fig. 2C, Supplemental Table 3). The iGMAS exon and about 450 bases of the flanking introns were cloned into a region flanked by two other exons (encoding GFP) and related intronic sequences (Xiao et al. 2009). For each iGMAS exon, two versions of the minigene were constructed, each carrying one of the two alleles of the intronic SNV. The minigenes were transfected into HeLa cells. Splicing of the iGMAS exon was analyzed using RT-PCR with primers targeting the two flanking GFP exons. Among the 10 candidates, eight were confirmed to have allele-specific splicing with the direction of allelic bias consistent with our RNA-seq analysis (Fig. 2C). Two other candidates (*ECT2* and *MPDZ*) did not show the predicted difference of exon inclusion levels between the alternative alleles. Failed validation could be due to the fact that the causal SNV is located outside of the limited intronic region cloned into the minigene system, or alternatively, the responsible *trans*-factor is not available or functional in HeLa. Overall, the high validation rate (80%) supports the accuracy of

the iGMAS method and the effectiveness of this method in capturing causal SNVs for splicing alteration. The results also suggest that it is unlikely that the observed GMAS events were mainly resulted from allele-specific intron degradation.

2.2.5 A compendium of genetically modulated alternative splicing events

We previously developed a method to identify GMAS events leveraging the information in RNA-seq data to reveal allelic association of exonic SNVs with splicing patterns (Li et al. 2012). The current method focusing on intronic SNVs complements the previous one (which we now call eGMAS for exonic tags of GMAS), which together generates an expanded catalog of genetically modulated exons. Combining the results of iGMAS and eGMAS on the above ENCODE data sets, we identified a total of 630 GMAS-related SNVs in 538 genes associated with splicing change of 622 exons (Supplemental Table 2, Supplemental Fig. 3A). Among all GMAS SNVs, 34% (215 out of 630) were previously reported in large-scale splicing QTL studies (’t Hoen et al. 2013; Lappalainen et al. 2013), a much larger overlap than expected ($P < 0.0001$, Supplemental Fig. 3B). It should be noted that eGMAS SNVs were more often shared across cell lines compared to iGMAS SNVs, with 73 eGMAS SNVs present in more than one cell line. This observation may be partly explained by the enhanced stringency imposed in iGMAS identification. We applied high stringency requirement considering that relatively few splicing-altering intronic SNVs were known in the literature compared to exonic ones. In addition, we observed that intronic genetic background in general is more diverse across cell lines than exonic sequences, another factor contributing to the small degree of overlap of iGMAS events across cell lines.

2.2.6 *Cis*-regulatory elements are primary drivers of splicing variation in GMAS

Among all GMAS exons, 90 were predicted in at least two cell lines, with 73% (66 out of 90 exons) having the same associated SNV. Motivated by this observation, we next addressed the question whether splicing change caused by genetic variants is often cell type-specific or shared across cell types. To this end, we collected SNVs with adequate statistical power common to two cell lines and asked whether those identified as GMAS SNVs in one cell line were often predicted as GMAS SNVs in the other cell line. In all pair-wise comparisons, we observed significantly higher numbers of shared GMAS SNVs across two cell lines than expected by chance (hypergeometric test, $P < 2.2e-16$, Fig. 3A). These data suggest that genetic variants often affect splicing in a cell type-independent manner. Thus, *cis*-regulatory elements may be the primary determinants of splicing variation in GMAS.

We tested the above hypothesis experimentally by expressing the human splicing reporters (Fig. 2C) in a mouse cell line (3T3) and measuring the splicing levels. Among the 10 iGMAS events tested in Figure 2C, seven were validated in 3T3 cells with the allelic bias consistent with iGMAS prediction (Fig. 3B). These results strongly suggest that *cis*-regulatory elements are key factors determining GMAS phenotypes. Notably, the *MPDZ* gene that failed validation in HeLa cells was successfully validated in 3T3 cells. Only one event (in *ECT2*) consistently failed in both cell types, suggesting that the accuracy of iGMAS prediction could be as high as 90% if multiple cell types were used in experimental validation.

2.2.7 GMAS exons demonstrated accelerated sequence evolution in primate lineages

With the large set of GMAS events, we examined their evolutionary characteristics from multiple perspectives. First, we compared the sequence conservation level of GMAS exons and associated introns to that of control exons and introns (Supplemental Methods). PhastCons scores that were derived using the genomes of 46 vertebrates spanning primates to fish were used for this purpose (Siepel et al. 2005). Interestingly, we found that GMAS exons (both coding and non-coding) were less conserved than the controls that were randomly chosen alternatively spliced (AS) exons (Fig. 4A). This observation indicates that GMAS exons may be evolving faster compared to AS exons in general.

To better understand the evolutionary pattern of GMAS exons, we examined their sequence divergence between human and other species in pair-wise comparisons (Fig. 4B). As expected, the sequence conservation level decreased as the evolutionary distance of the considered species increased relative to human. Interestingly, GMAS exons were similarly conserved as random control exons (Supplemental Methods) when comparing human and other primate sequences (chimpanzee and rhesus macaque). However, the lower sequence conservation of GMAS exons (compared to controls) became evident in mouse and other species with longer evolutionary distance from human. Based on the parsimony model of evolution, these data suggest that GMAS exons experienced faster evolution in recent evolutionary history, which is common to primates, but occurred after the speciation event leading to primate and rodent lineages (between 25 and 80 mya, Fig. 4B).

2.2.8 GMAS-related genes, exons and SNVs undergo positive or balancing selection

The primate-specific nature of accelerated evolution of GMAS exons prompted us to examine whether they are subject to positive selection. We first asked whether the genes harboring GMAS exons were enriched with those undergoing positive selection. About 46% (246 out of 538) GMAS genes were categorized as positively selected genes in the Selectome database (Moretti et al. 2014). This fraction is significantly higher than that among all known human genes (10%) and among genes undergoing AS (30%) based on the ENCODE RNA-seq data (Fig. 4C).

We next examined the amino acid selection pressure (d_N/d_S) of protein-coding GMAS exons. As controls, we randomly selected AS exons that are also protein-coding and have similar exon inclusion levels as the GMAS exons (Supplemental Methods). To quantify the exon inclusion level, we used previously published tissue-specific human RNA-seq data to calculate the percent-spliced-in (PSI) values of exons in brain, heart, and liver, respectively (Barbosa-Morais et al. 2012). Compared between human and mouse, GMAS exons demonstrated significantly higher d_N/d_S values than control exons in all three tissues (Fig. 4D, Supplemental Fig. 4). This difference in d_N/d_S is mainly due to a higher value of d_N for GMAS exons compared to control AS exons (Supplemental Fig. 4). The d_N/d_S of GMAS exons is not significantly different from that for control exons between human and other primates (chimpanzee and rhesus macaque). These results suggest that GMAS exons are under relaxed amino acid selection pressure in recent primate evolution, which is in line with the above observation of accelerated sequence evolution in primate lineages.

A related question is whether GMAS SNVs demonstrate any signs of accelerated evolution. To address this question, we calculated three population genetics measures: Tajima's *D* (Tajima 1989), fixation index (F_{ST}) (Weir and Cockerham 1984), and integrated haplotype score (iHS) (Voight et al. 2006). Using data from the 1000 Genomes project (Abecasis et al. 2012), we observed that most GMAS SNVs had positive Tajima's *D* values that are often statistically greater than controls in multiple populations (CEU, FIN, GBR, and TSI) (Fig. 4E, Supplemental Fig. 5A). Positive Tajima's *D* is considered as a signature of balancing selection where there is an excess of common variants compared to the neutral expectation within a population. Alternatively, positive values of Tajima's *D* may also reflect recent population contraction, which is unlikely true here since the European populations are known to have experienced recent population expansion (Tennessen et al. 2012; Keinan and Clark 2012; Nelson et al. 2012). Values of F_{ST} and iHS of the GMAS SNVs are not significantly different from those of controls (Supplemental Fig. 5). Nevertheless, GMAS SNVs demonstrated a trend of higher F_{ST} than control SNPs ($P = 0.06$). Higher F_{ST} values suggest higher population differentiation, that is, lower level of shared alleles across populations, which is considered as a sign of positive selection. The selection signals on SNVs are generally weaker than those of GMAS genes or exons, which may be due to the existence of some non-causal SNVs among GMAS SNVs.

Altogether, the above evolutionary and populational analyses converge to the same conclusion that GMAS events are associated with a signature of accelerated evolution or positive selection in primates.

2.2.9 GMAS events are likely regulated by conserved splicing factors

A natural hypothesis about the regulatory mechanism of GMAS SNVs is that the alternative alleles of a SNV change the binding strength of a splicing factor, which then alters the outcome of splicing. Based on this hypothesis, we searched for known motifs of a large compendium of RNA binding proteins (RBPs) (Cook et al. 2011; Ray et al. 2013) flanking each GMAS SNV and selected one putative RBP whose motif score was altered most significantly by the SNV (Supplemental Methods). Random intronic SNVs that were at least 5000 bases away from exons were used as controls to compare with GMAS SNVs. Supplemental Figure 6 shows the difference in RBP binding scores between the two alternative alleles of each SNV. GMAS SNVs alter the binding scores of RBPs more significantly than control SNVs, supporting the expected relationship between RBPs and GMAS SNVs and the expectation that GMAS SNVs are enriched with causal SNVs.

Next, we focused on 15 splicing factors that were predicted to target at least five GMAS SNVs (Fig. 5A). An immediate question is whether these splicing factors also evolve faster than expected, similarly as their GMAS exon targets. Interestingly, we observed that both the protein sequences and the RNA-binding domains of these factors are significantly more conserved than control splicing factors that were not predicted to target any GMAS SNV (Fig. 5B and C, Supplemental Methods). Thus, splicing of the fast-evolving GMAS exons is likely regulated by conserved splicing factors. This finding again supports our earlier conclusion that alterations in *cis*-elements consist of the primary driving force of splicing evolution.

2.2.10 Alteration of splicing factor binding by GMAS SNVs

To better understand the relationship between GMAS SNVs and the predicted splicing factors, we next examined the general motif strength of splicing factors predicted to bind to GMAS SNVs. Intuitively, only splicing factors with strong consensus motifs can be readily disrupted by a single SNV in their binding sites. Other splicing factors with highly degenerate motifs should be relatively resistant to perturbations by SNVs. Consistent with this expectation, we observed that the average information content of the 15 splicing factors predicted to target ≥ 5 GMAS SNVs is significantly higher than that of the control splicing factors (all known splicing factors that are not predicted to regulate GMAS exons) (Fig. 5D).

We next examined the individual nucleotides corresponding to GMAS SNVs within the binding motif of each splicing factor. Consistent with the expectation that these SNVs should disrupt motif positions with strong consensus, we observed that the information content of the SNV-overlapping nucleotides is generally higher than that of random nucleotides within the motif of the same splicing factor (Fig. 5E). Intriguingly, the SNV-overlapping nucleotides had lower information content than those with the maximum information content of each motif, suggesting that the GMAS SNV does not generally target the strongest consensus position. This observation may have important evolutionary implications (see Discussion).

2.2.11 Allele-specific binding of *SRSF1* in enabling GMAS regulation

We next focused on GMAS regulation by the protein *SRSF1*, the splicing factor with the largest number of GMAS targets predicted by motif analysis (Fig. 5A). To further confirm the predicted GMAS exons as *SRSF1* targets, we analyzed the recent ENCODE RNA-seq data sets of *SRSF1* knockdown (and controls) and eCLIP-seq data of this protein (in HepG2 and K562 cells)

(Supplemental Methods). The majority (70%) of predicted *SRSF1*-targeted GMAS exons showed a change of exon inclusion level of at least 10% upon knockdown of the protein, and many (67%) had at least one eCLIP-seq peak within the exon or less than 200bp in the vicinity in at least one cell line (Supplemental Table 4). Thus, a large fraction of the predicted *SRSF1* GMAS targets are confirmed endogenously. Those that did not show splicing changes upon *SRSF1* knockdown could be under complex regulation where other splicing factors may compensate for the loss of *SRSF1*. The regulatory impact of *SRSF1* on splicing is known to be complex depending on its interaction with other splicing factors (Pandit et al. 2013; Anczuków et al. 2015), which may also explain the opposite direction of splicing changes for many exons between HepG2 and K562 cells (Supplemental Table 4).

To provide direct experimental support that GMAS SNVs alter the binding of *SRSF1* to RNA, we carried out electrophoretic mobility shift assay (EMSA, or gel shift) on a panel of randomly selected GMAS targets of *SRSF1*. In addition to the wild-type protein, we also cloned and expressed an RNA binding mutant of *SRSF1* (FF-DD mutant) that was shown to have significantly reduced RNA binding capacity (Cho et al. 2011) (Supplemental Fig. 7, Supplemental Table 5). Among all predicted *SRSF1*-bound GMAS SNVs, the majority overlap with the 5th position in the consensus motif, with the second most often targeted position being the 2nd position that has the strongest consensus (Fig. 5F). Thus, we tested three GMAS SNVs targeting the 5th position and one GMAS SNV targeting the 2nd position.

As shown in Figure 5F, the binding of *SRSF1* to target RNAs was stronger with increasing protein input. The RNA-protein interaction was very weak or abolished when the FF-

DD mutant of *SRSF1* was used (Supplemental Fig. 7). Thus, these observations confirm a direct binding of *SRSF1* to the target RNAs. To confirm that GMAS SNVs alter the binding of *SRSF1*, two versions of each target RNA were synthesized harboring the alternative alleles of the SNV. Strong signals of differential binding to the alternative alleles of GMAS SNVs was observed for three of the four RNA targets (*RMND5B*, *TACC2*, *PIK3C3*), with stronger binding to the allele dominant in the consensus motif (Fig. 5F). The fourth target, *HSPD1*, demonstrated a small degree of allelic difference (with the G allele having slightly stronger signal than the A allele at high *SRSF1* concentrations 1.5 and 3.0μM), possibly due to the fact that this *SRSF1* binding site is overall relatively weak.

Importantly, *SRSF1* demonstrated global allele-specific binding to heterozygous SNPs based on the eCLIP-seq data. In the eCLIP reads, almost all observed heterozygous SNPs (covered by ≥ 20 reads) had significant allelic bias (Supplemental Table 6, Supplemental Methods). Among GMAS exons predicted as regulated by *SRSF1*, many did not have heterozygous SNPs in the HepG2 or K562 cells or adequate read coverage overlapping the heterozygous SNPs. However, one exon had a heterozygous SNP with a read coverage > 20 in the HepG2 eCLIP data. This SNP had a significant allelic bias (G>A, $p < 0.001$), confirming our prediction. These results provide strong evidence that *SRSF1* is an important regulator of GMAS through allele-specific binding.

2.2.12 Functional analyses of GMAS genes and exons

To gain a better understanding of the functional relevance of GMAS, we first conducted gene ontology (GO) analysis of the related genes. Using a previously published method (Lee et al.

2011), we observed 15 enriched GO terms among the GMAS genes (Fig. 6A, Supplemental Table 7). Interestingly, a few mitochondria-relevant GO terms were observed, e.g., respiratory chain complex IV assembly, mitochondrial inner membrane, and mitochondrial transport. Two major mitochondria-related gene families, cytochrome c oxidases (*COX*) and mitochondrial transporter family *SLC25*, host GMAS exons accounting for these GO categories. Both families of proteins have critical functions related to mitochondrial diseases (Zee and Glerum 2006) and metabolic processes (Gutiérrez-Aguilar and Baines 2013).

For coding GMAS exons, we further examined whether they are located in disordered protein regions (Supplemental Methods). These regions, defined as those that lack a stable protein structure in solution, are known to play important roles in protein-protein interaction and cellular signaling (Fu and Ares 2014). Our analysis showed significant enrichment of GMAS exons in disordered regions than corresponding controls (AS exons picked randomly that match the exon length and GC content of GMAS exons) (Fig. 6B). Thus, GMAS exons display features of functionally important exons.

Next, we examined the protein domains encoded (possibly partly) by GMAS exons (Supplemental Methods). Many domains were identified in this analysis (Supplemental Table 8), suggesting that GMAS exons may have diversified functional impacts. Notably, three domains were significantly enriched among GMAS exons compared with controls (AS exons picked randomly that match the exon length and GC content of GMAS exons; Bonferroni-corrected $P \leq 0.05$ and present in ≥ 2 distinct exons) (Supplemental Table 8). The mitochondrial carrier protein domain was encoded partly by GMAS exons in the *SLC25* gene family (*SLC25A26* and

SLC25A28), again attesting to the association of GMAS with mitochondrial function. Another significant protein domain is the folate carrier domain encoded by *SLC19A1*. We identified three SNPs associated with three different GMAS exons in *SLC19A1*, all of which overlap with the folate carrier domain in the gene. The other GMAS exon-encoded domain, Mib-Herc2 domain, has a critical role in the activation of the notch signaling pathway (Itoh et al. 2003). In the above examples, the same type of protein domain (in the same gene or different genes) was repeatedly observed to be associated with distinct GMAS SNVs, suggesting possible existence of selection pressure for genetic modulation of their splicing.

2.2.13 More than 100 GMAS SNVs are in LD with GWAS SNPs

GWAS analyses have identified thousands of genetic loci that are associated with a diverse set of diseases or phenotypic traits (Welter et al. 2014). However, it has been a major challenge to elucidate the molecular function of the vast majority of GWAS hits that may not affect protein-coding sequences. The large catalog of GMAS SNVs allowed us to examine whether a GWAS SNP may be associated with splicing alteration. We observed that 116 (18%) GMAS SNVs are in LD with GWAS SNPs (and within 200kb in distance) (Supplemental Table 9), associated with a total of 110 distinct GMAS exons. For convenience, we refer to these GMAS SNVs as GMAS-GWAS SNVs henceforth.

Next, we analyzed the motif strength of nucleotides overlapping the GMAS-GWAS SNVs based on predicted splicing factor binding, similarly as described above. As shown in Figure 5E, the information content of these nucleotides is at a similar level as maximum individual-nucleotide information content within the corresponding motifs. This observation is in

stark contrast to that of overall GMAS SNVs whose information content was lower than the maximum values of their respective motifs, although higher than those of random nucleotides within the motifs (Fig. 5E). Because GMAS-GWAS SNVs are associated with phenotypic traits or diseases, they represent the subset of GMAS SNVs that may have direct functional/biological implications. Therefore, the observed higher information content associated with GMAS-GWAS SNVs is consistent with the functional impact of these SNVs, which likely represent strong perturbations to splicing factor binding.

Since most GWAS observations are not supported by molecular mechanisms, splicing-altering GMAS SNVs help to shed light on the potential functional mechanisms underlying many GWAS observations. Figure 6C demonstrates an analysis of GWAS SNPs that are in LD with (and within 200kb in distance) GMAS SNVs. About two thirds of these GWAS SNPs are located in or close to the same genes as GMAS SNVs in LD. Overall, the vast majority (71%) of GWAS SNPs in Figure 6C are in introns, whose functional relevance was previously elusive, but may now be explained by the splicing-altering GMAS SNVs. In addition, a small fraction (15%) of these GWAS SNPs resides in coding regions, some of which were annotated as synonymous SNPs in the GWAS catalog. Thus, these seemingly nonfunctional synonymous SNPs could indeed cause splicing alteration. Only about 8% of the GWAS SNPs were predicted to cause missense changes. Among the 90 coding GMAS-GWAS exons, 40 were in-frame with length dividable by 3 (none harboring stop codons). Many of the 90 coding exons overlap known protein domains (Supplemental Table 9). Taken together, the GMAS results suggest that alteration of splicing may be part of the molecular basis of many GWAS observations.

Finally, we asked whether certain diseases or traits were enriched with GMAS-GWAS SNVs. For a number of GWAS traits, such as metabolic traits, cardiovascular disease risk factors and attention deficit hyperactivity disorder (ADHD), multiple GMAS SNVs were observed in the associations, significantly more often than expected by chance (Supplemental Table 10, Supplemental Methods). Association of multiple GMAS genes with the same trait strongly indicates that splicing alterations are part of the functional pathways linked to the trait. Supplemental Figure 8 lists a few example genes with important biological functions that harbor GMAS events that were either supported by previous literature (low-density lipoprotein receptor (*LDLR*)) or by our experimental validations.

2.3 Discussion

We report a comprehensive study of alternative splicing events regulated by genetic variants. Using ENCODE RNA-seq data sets with high sequencing depth and cellular compartment-specific features, we identified a large number of GMAS exons that showed allele-specific splicing patterns. Expression analysis of SNVs in the allele-specific manner can effectively capture *cis*-acting regulatory variants since the relative expression of the alternative alleles of a heterozygous SNV is measured in the same cellular context, eliminating *trans*-acting or environmental influences on gene expression. Previous allele-specific splicing studies were mainly restricted to exonic SNVs since mature mRNAs without introns were normally interrogated in RNA-seq. Our iGMAS method fills in this gap by comparatively analyzing nuclear and cytosolic RNA contents to examine intronic SNVs and their allele-specific association with exon expression.

Our study identified more than 600 alternative splicing events that are likely regulated by genetic variants, which facilitated the first global study of genomic, evolutionary and regulatory characteristics of GMAS events. One interesting observation is that the impact of genetic variants on splicing is largely cell type-independent (Fig. 3), suggesting that *cis*-regulatory elements are the primary determinants of the splicing phenotype in GMAS exons. This result is in line with a previous study, which reported that species-specific alternative splicing is primarily driven by *cis*-regulatory elements (Barbosa-Morais et al. 2012). Our observation has an important implication that splicing-altering mutations are likely effective in multiple cell types or tissues. Thus, their functional impacts could be widespread or largely ubiquitous across tissues.

We observed that GMAS exons are under selection for accelerated sequence evolution in primate genomes (Fig. 4). This finding is in line with a previous study reporting that primates demonstrated faster accumulation of alternative exons than other mammalian lineages (Merkin et al. 2012). Lower levels of sequence conservation were often interpreted as reduced functional significance of the relevant genomic regions. However, the lower conservation of GMAS sequences is unlikely resulted from neutral or random mutations. Instead, evolutionary selection exists as reflected by the enrichment of positively selected GMAS genes and accelerated evolution of GMAS exons and SNVs (Fig. 4). Accelerated evolution may have profound functional relevance. For example, it is now well known that positive selection affects many genes of human and other primates possibly to define species-specific phenotypes or to enable biological adaptation to the environments. Positively selected genes in human play important roles in many aspects of biology, such as brain development and function, cognition, behavior, vocalization, reproduction, dietary adaptation, metabolism, physical appearance and host-

pathogen interactions (Nielsen et al. 2007). The fact that GMAS-related genes, and alternatively spliced genes in general, demonstrated strong enrichment of positive selection pressure (Fig. 4C) indicates that alternative splicing may be a mechanism to introduce adaptive changes to gene expression during primate evolution.

It is important to note that positive or balancing selection is arising as an evolutionary signature of many mutations contributing to complex diseases. The accelerated evolution is often driven by selection for certain beneficial traits associated with the mutations although these mutations may cause other diseases (Nielsen et al. 2007). A well-known example is the sickle cell mutation in the hemoglobin beta (*HBB*) gene that is positively selected due to its properties rendering malaria resistance despite its role in causing sickle cell disease (Currat et al. 2002). Thus, it is possible that disease-related splicing variations are under positive selection due to functional advantages associated with these variations for certain biological processes, a hypothesis that needs further investigation.

Another aspect of our study revealed that the molecular mechanisms of many GMAS SNVs likely lie in their disruption of splicing factor binding (Fig. 5). It is interesting to note that GMAS-associated splicing factors are often more conserved than expected, in contrast to the accelerated evolution of GMAS exons or SNV-flanking sequences. This observation again suggests that alterations in *cis*-elements, rather than *trans*-factors, are the main driving forces of splicing evolution. The evolutionary cost of a mutation in a *cis*-regulatory motif is much smaller than that in a *trans*-factor, since the latter may impact hundreds to thousands of splicing events. Thus, *cis*-elements-driven splicing evolution is cost-effective.

An important observation is that GMAS SNVs often disrupt motif positions that had stronger consensus nucleotides than expected by chance. However, they do not generally correspond to the strongest consensus nucleotide of the respective motifs. In contrast, the subset of GMAS SNVs that presumably has close biological relevance (i.e., those in LD with GWAS SNPs) often overlap with the strongest consensus nucleotides. These results indicate that in general GMAS SNVs can alter splicing factor binding, but many of them may only cause moderate splicing changes. Moderate tuning on splicing serves as a mechanism to introduce novel gene expression products for evolutionary selection without severely affecting biological or functional pathways. During evolution, new SNVs may occur at the strongest consensus nucleotide of a splicing factor binding site, which may be selected against due to strong adverse effects. However, some of these SNVs may survive evolutionary selection (such as the GMAS-GWAS SNVs) as a result of, for example, advantageous function or adaptive response to environment changes.

2.4 Methods

2.4.1 RNA-seq data and mapping

Paired-end RNA-seq data (2x76nt) from seven human cell lines (GM12878, K562, HeLa, HepG2, HUVEC, NHEK, and H1-hESC) were downloaded from the ENCODE data repository (www.encodeproject.org) under the ENCODE Data Coordination Center accession number ENCSR037HRJ (The ENCODE Project Consortium 2012), or NCBI GEO accession number GSE30567, where two biological replicates are available for all cell lines except H1-hESC. The

reads were mapped using a stringent mapping method described in our previous work (Bahn et al. 2012). Further details of read mapping are given in Supplemental Methods.

2.4.2 Prediction of iGMAS events

The goal of the iGMAS analysis is to capture intronic SNVs associated with allele-specific splicing. To this end, we used the RNA-seq data of nuclear RNA without polyA tails (NA-) to capitalize on the enrichment of intronic RNA in these data sets. Specifically, we aim to determine whether the alternative alleles of an intronic SNV are present in a biased manner in the spliced-out products. Spliced-out products considered here only include those that contain an alternatively skipped exon and its flanking introns, where the intronic SNV could reside in either intron. Thus, our analysis is limited to alternative exon skipping, the most prevalent type of alternative splicing.

To achieve the above goal, we focused on read pairs where one read was mapped to either flanking intron with the other read mapped to the exon, or the two reads were mapped to the two flanking introns respectively (Fig. 1C). In addition, at least one intronic read must harbor an SNV. Thus, these read pairs (referred to as "linked reads" below) were likely originated from the spliced-out products that contain skipped exons. Alternatively, these reads could have come from nascent unspliced RNA that is also present in the NA- data. In this scenario, intronic SNVs may have imbalanced allelic expression if the nascent RNA is transcribed in an allele-specific manner, which leads to ASE of the entire gene. Since our goal is to enrich for reads that represent spliced-out products, we applied two exclusion filters to remove reads that may have arisen from ASE of the gene. First, we identified genes that showed significant ASE patterns on

the whole-gene level using our previous method and CA+ RNA-seq data (Li et al. 2012). Reads mapped to these genes were excluded from further analysis. Second, we estimated the gene-level allelic expression ratio (r_{est}) for each gene that had multiple SNVs using CA+ data, as follows:

$$r_{est} = \frac{\sum_{i=1}^{n} N_{ref_i} + a}{\sum_{i=1}^{n} (N_{ref_i} + N_{\mathrm{var}_i}) + b}$$

where N_{ref} and N_{var} represent the read counts for the reference or variant allele of a SNV respectively, assuming there are n SNVs in a gene. Variables a and b are priors which were set to be one since the reference and variant alleles were known in our problem. Since the haplotypes of SNVs are unknown, this estimated allelic ratio is not always accurate, but can only serve as an exclusion filter. We removed reads mapped to genes with an estimated allelic ratio outside the range of (0.4, 0.6), which indicates possible allelic bias on the gene level.

For all candidate intronic SNVs that passed the above exclusion filters, we next determined whether they demonstrate significant allelic bias deviating from the expected allelic ratio of 0.5 (with allelic ratio calculated as $N_{ref}/(N_{ref}+N_{var})$). In these calculations, only the linked reads (as defined above) were used. A two-sided binomial test P-value was calculated as:

$$P = 2 \times \left(1 - \sum_{i=0}^{N_{ref}} \binom{N}{i} p^i (1-p)^{N-i}\right)$$

where $p = 0.5$ in this case; $N = N_{ref} + N_{var}$.

To achieve adequate statistical power, we required that the number of linked reads for each intronic SNV (after combining biological replicates of RNA-seq data) was at least 19

according to the power estimation conducted in our previous work (Li et al. 2012). We required the Binomial *P*-value to be less than 0.01 to call significant allelic bias. Finally, to enhance stringency, we required that the deviation of the allelic ratio ($N_{ref}/(N_{ref}+N_{var})$) from 0.5 was at least 0.2. Intronic SNVs that satisfied all the above requirements were categorized as iGMAS SNVs.

To estimate the FDR of the iGMAS method, we randomly distributed reads covering intronic SNVs to their alternative alleles, maintaining the total number of reads for each SNV. This randomization was carried out assuming an expected allelic ratio of 0.5. This procedure controls for the read coverage of each gene and SNV and maintains the read distribution in alternative and constitutively processed regions. The same iGMAS identification framework as described above was applied to the randomized data and an FDR was estimated.

2.4.3 Splicing reporter assays

Minigenes containing iGMAS exons and flanking introns were constructed, each harboring one alternative allele of the targeted SNV. Further details of the splicing assay are described in Supplemental Methods.

2.5 Acknowledgements

We are grateful for the insightful comments provided by the anonymous reviewers. We thank members of the Xiao laboratory for helpful discussions and comments on this work. We thank the ENCODE Project Consortium for production of the RNA-seq and eCLIP-seq data. This work

was supported in part by grants from the National Institute of Health (R01HG006264 and U01HG007013), National Science Foundation (1262134), Alfred P. Sloan Foundation and the University of California Cancer Research Coordinating Committee to X.X.

2.6 Figures

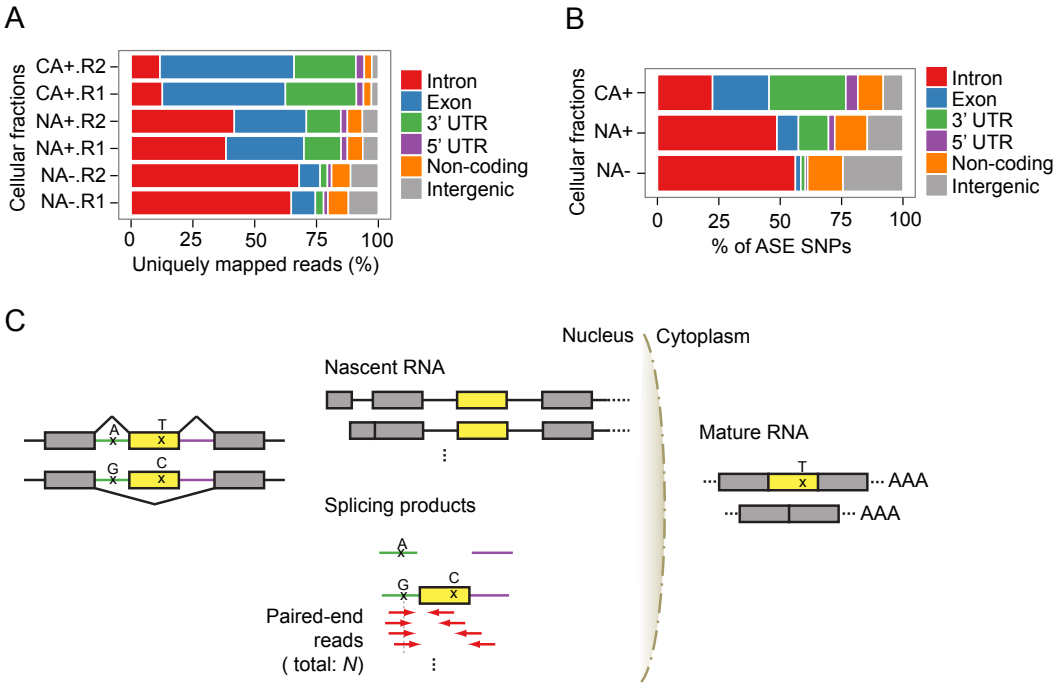


Figure 2.1 Compartment-specific RNA-seq enables coverage of intronic SNVs. (A) The genomic context of uniquely mapped reads of GM12878 CA+, NA+, and NA- RNA-seq datasets. Two biological replicates (R1 and R2) were analyzed. Non-coding refers to non-coding transcripts or genes. Reads that mapped to regions with multiple annotation categories were classified into one genomic context according to priorities given as follows: coding exon > 5' UTR > 3' UTR > intron > non-coding > intergenic. (B) Similar as (A), the genomic context of SNVs with ASE patterns in GM12878 CA+, NA+, and NA- datasets. Biological replicates were

combined. (C) The biological principle underlying iGMAS detection. In this hypothetical example, the yellow exon is alternatively spliced depending on the allele of the intronic SNV, with the *A* allele associated with exon inclusion and *G* allele associated with exon skipping. In the nucleus, NA- RNA-seq reads could originate from nascent RNA or spliced-out products. In the spliced-out products, the *A* and *G* alleles will reside, respectively, in the single intron and the intron-exon-intron molecule. In the nascent RNA, the *A* and *G* alleles are also present, which is not shown in the diagram. RNA-seq reads (red arrows) originating from spliced-out products covering the intronic SNV and neighboring exon or intron are enriched with the *G* allele, which can be analyzed to infer allele-specific regulation of splicing. An exonic SNV is also illustrated, whose *T* allele (the one in the same haplotype as the intronic *A* allele) is expected to be enriched in RNA-seq data of the cytoplasmic RNA (CA+). Note that exonic SNVs are not always present in iGMAS exons.

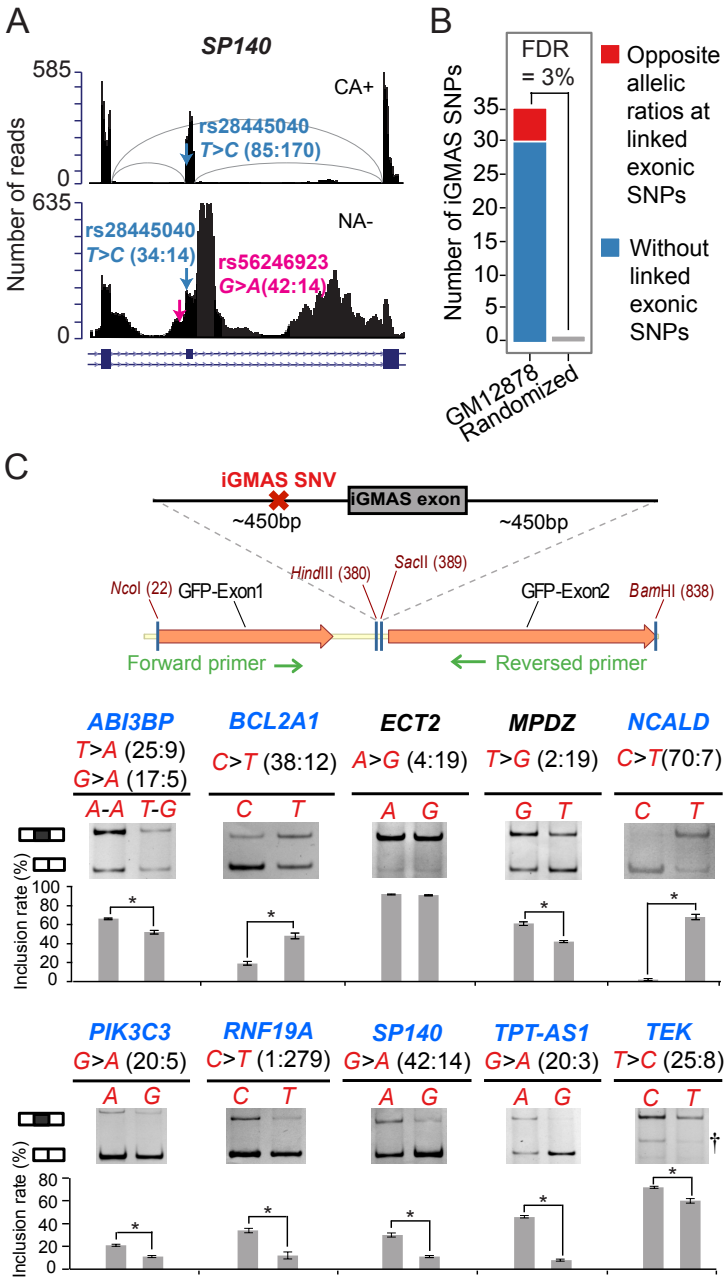


Figure 2.2 Evaluation of the iGMAS method. (A) An example iGMAS event. Read distributions of the region in the gene *SP140* around an iGMAS event identified in GM12878 are illustrated using the CA+ (upper) and NA- (lower) data, respectively. Exon-intron structures (RefSeq annotation) are shown at the bottom. Pink arrows and texts denote iGMAS SNP location,

allele types and number of reads harboring each allele in NA- data. This iGMAS exon also has an exonic SNP whose information is provided in blue (haplotypes: *T*-*G*, *C*-*A*). As expected for an authentic iGMAS event, the exonic SNP has opposite allelic bias in CA+ and NA- data (i.e., allele *C* enriched in CA+, but under-represented in NA-). Arcs in light gray represent existence of spliced junction reads across exons. (B) The number of iGMAS events identified in GM12878 cells (blue and red) where five events (red) had corresponding exonic SNVs (all of which showed opposite allelic ratios comparing their allelic coverage in the CA+ and NA- fractions). Only one iGMAS event was identified in the randomized data (Methods), yielding an estimated false discovery rate (FDR) of 3%. (C) The minigene system used for experimental validation of iGMAS events is illustrated (Supplemental Methods). Validation results in HeLa cells of 10 randomly picked iGMAS events are shown. Alternative alleles of iGMAS SNVs are shown together with their read counts in NA- data. All events but one (*ABI3BP*) had only one iGMAS SNV. Mean and SD of exon inclusion levels based on three biological replicates are shown. As expected for successful validation, the more enriched allele in NA- RNA-seq data should be associated with a smaller exon inclusion level in the splicing assay. Among the 10 iGMAS events, eight (gene names in blue) were successfully validated in HeLa cells. (*$P \leq 0.05$, Wilcoxon Rank Sum test; †non-specific bands)

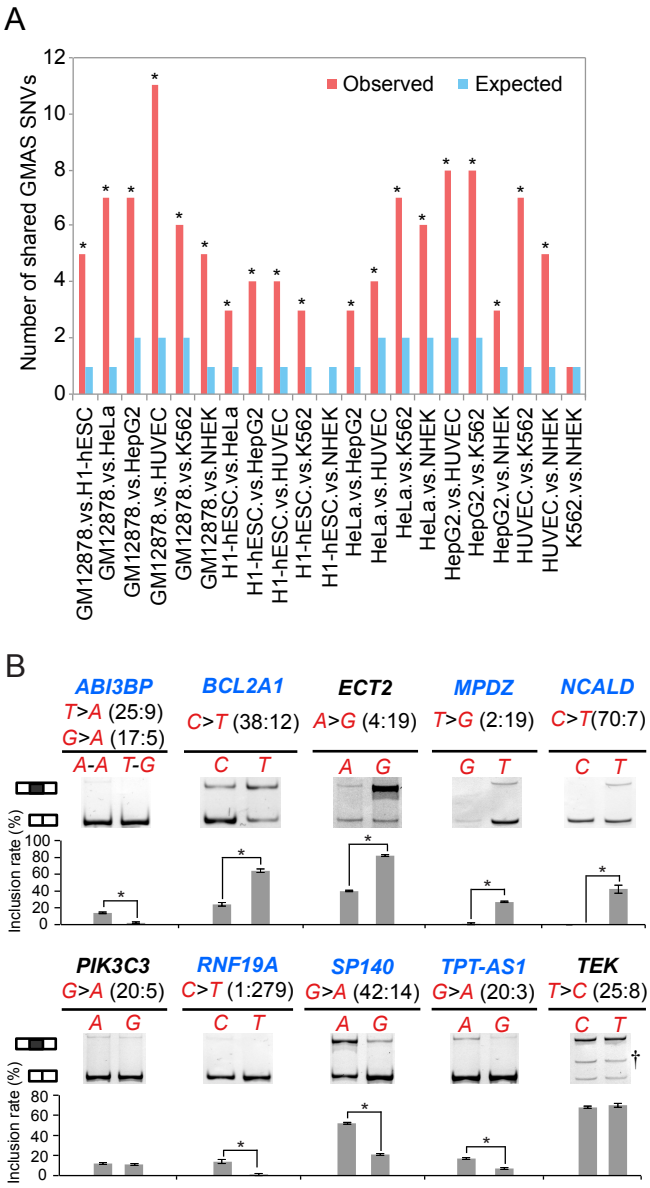


Figure 2.3 Splicing alteration by genetic variants is highly cell type-independent. (A) The number of common GMAS SNVs between each pair of samples is shown (red bars). The expected number of shared GMAS SNVs for each pair of samples (blue bars) was calculated assuming independent occurrence of each event in each sample. Enrichment of common GMAS SNVs between cell types was evaluated using hypergeometric test by comparing the observed

and expected occurrences. (*$P < 0.05$) (B) The same splicing assays as shown in Figure 2C were repeated by transfecting the minigenes (containing human iGMAS exons) into mouse 3T3 cells. The results are illustrated in the same way as in Figure 2C. Successfully validated cases are illustrated with the gene names in blue.

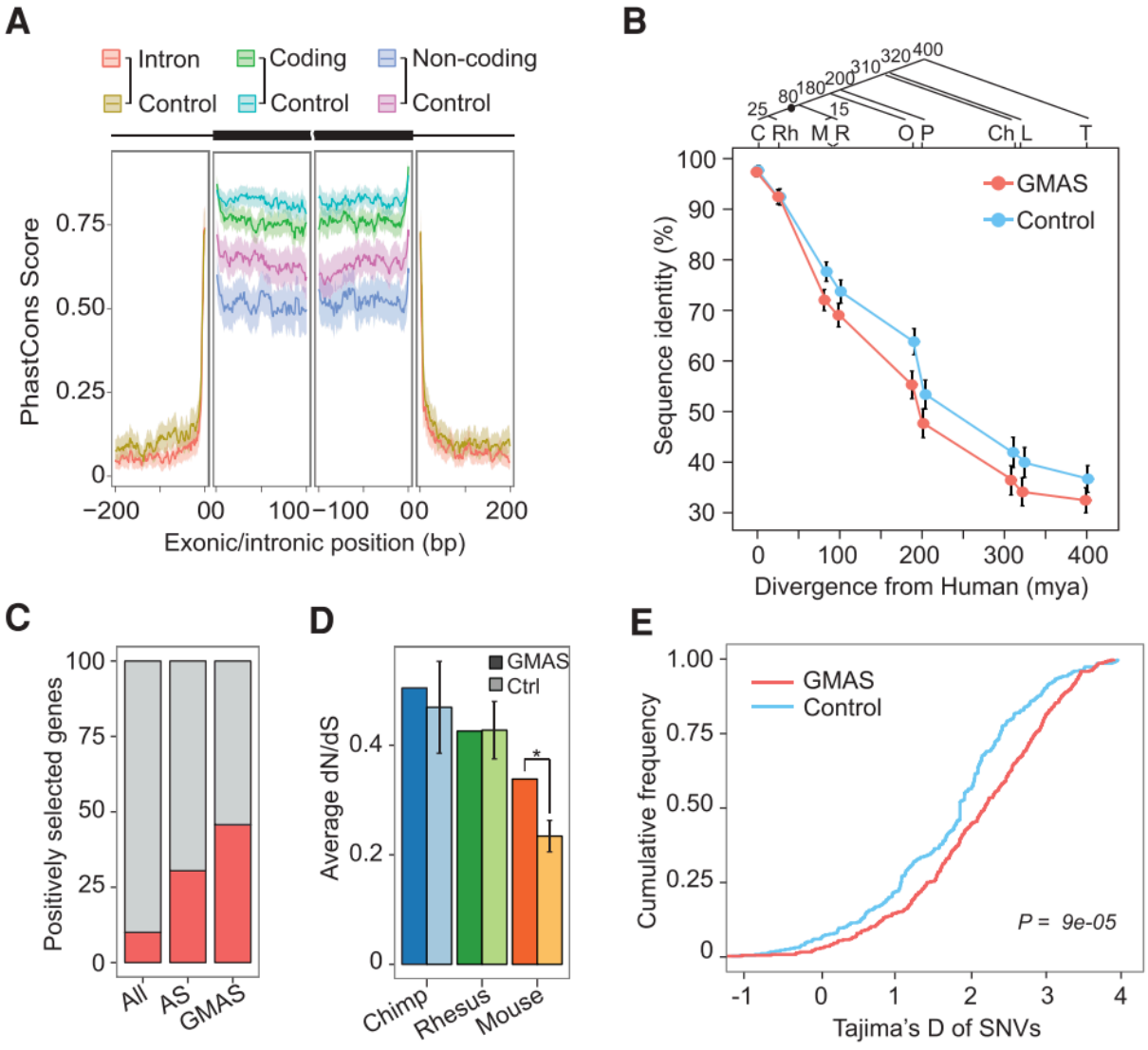


Figure 2.4 GMAS events undergo accelerated evolution. (A) Conservation profile (mean and 95% confidence interval of 46-way PhastCons scores) of GMAS exons and flanking introns. GMAS exons were separated into coding and non-coding groups with corresponding control

exons of the same type. (B) Percent sequence identity (mean and 95% confidence interval) of GMAS exons (red) and control exons (blue) between human and each of the organisms shown in the graph. Estimated divergence times of each species from human (in million years (mya)) are shown along the *x*-axis and in the phylogenetic tree. (C: chimpanzee; Rh: rhesus macaque; M: mouse; R: rat; O: opossum; P: platypus; Ch: chicken; L: lizard, T: tetraodon). (C) Percentage of genes that are under positive selection according to the Selectome database. The analysis was conducted for all known human genes, alternatively spliced (AS) genes according to ENCODE RNA-seq data used in this study, and GMAS genes. GMAS genes are more often positively selected ($P < 2.2e\text{-}16$ compared to all human genes, $P = 3.4e\text{-}13$ compared to AS genes, Fisher's exact test). (D) d_N/d_S values (mean and SD) of coding GMAS exons comparing human vs. chimpanzee, human vs. rhesus macaque, or human vs. mouse. Results for GMAS exons and control exons are shown in darker and lighter colors respectively. Control exons used in this graph were randomly picked AS exons with similar PSI values as GMAS exons as calculated using brain RNA-seq data (Supplemental Methods). The average d_N/d_S ratio of GMAS coding exons between human and mouse is significantly greater than that of control exons ($^*P < 0.001$ based on empirical distribution of 1000 control sets). (E) Tajima's *D* values of GMAS (red) and control SNPs (blue) in the CEU population. The *P*-value was obtained from the Kolmogorov–Smirnov test.

A

SF	# GMAS SNV
SRSF1	31
SNRPA	15
SRSF2	14
SNRPB2	14
ELAVL2	12
PCBP4	9
PCBP3	9
RBM45	8
RBM4	6
ELAVL3	6
RBM4B	6
QKI	6
CELF3	5
HNRNPA3	5
HNRNPA1	5

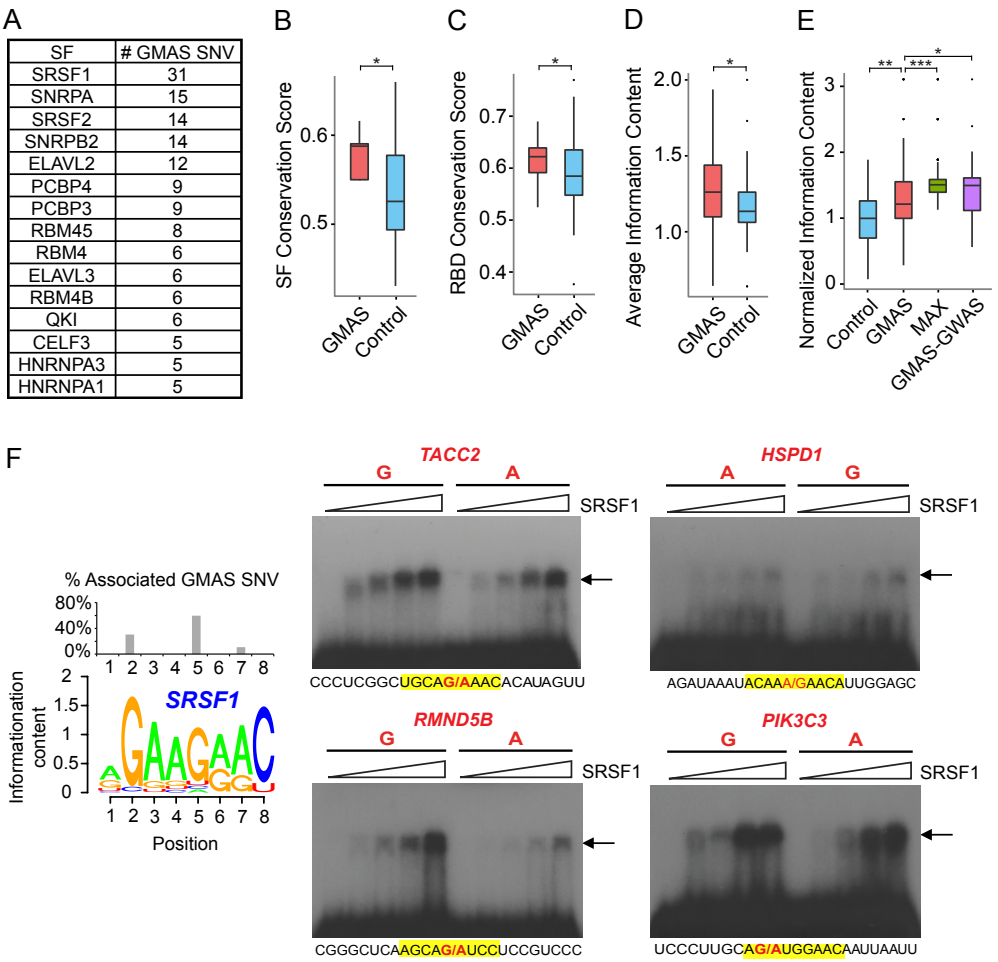


Figure 2.5 GMAS SNVs alter the binding of splicing factors. (A) Splicing factors predicted to target at least five GMAS SNVs (ranked by the number of their GMAS targets). SF: splicing factor. (B) Amino acid conservation scores of the GMAS-associated splicing factors (GMAS) and control splicing factors (Control). The 15 splicing factors listed in (A) were included in this analysis. Controls were defined as all non-GMAS-targeting splicing factors. Known splicing factors were obtained from pervious literature (Han et al. 2013). (*$P = 0.004$, Wilcoxon Rank Sum test). (C) Amino acid conservation scores of the RNA-binding domains (RBDs) of the splicing factors. The same splicing factors and controls as described in (B) were analyzed (*$P =$ 0.012, Wilcoxon Rank Sum test). (D) Average information content of sequence motifs of

splicing factors. The same splicing factors and controls as described in (B) were analyzed (*P = 0.006, Wilcoxon Rank Sum test). (E) Normalized information content of specific nucleotide positions in the sequence motifs of the 15 splicing factors in (A) (Supplemental Methods). Normalization was carried out against the average information content of all nucleotides within each motif. Control: a random nucleotide in a sequence motif. GMAS: the nucleotides disrupted by GMAS SNVs. MAX: the strongest consensus positions of each motif. GMAS-GWAS: the nucleotides disrupted by GMAS SNVs that are in LD with GWAS SNPs. (*P = 0.004, **P = 2.26e-8, ***P = 6.30e-12, Wilcoxon Rank Sum test) (F) Left: The sequence logo of *SRSF1* binding motif generated from the PWM provided in (Ray et al. 2013). The percent of GMAS SNVs targeting specific nucleotide positions of the *SRSF1* motif is shown as a bar plot above the sequence logo. Middle and Right: EMSA results of *SRSF1* binding to predicted GMAS targets. Alternative alleles of the GMAS SNVs were synthesized, as labeled above the gel images. The sequences of the synthetic RNA fragments are shown below each gel image, where the *SRSF1* sequence motif is highlighted in yellow and the two alleles of GMAS SNVs are written in red. The arrow indicates RNA-protein complex. Increasing concentrations of *SRSF1* were used in different lanes of the gel image (from left to right: 0, 0.37, 0.75, 1.5, and 3.0 μM).

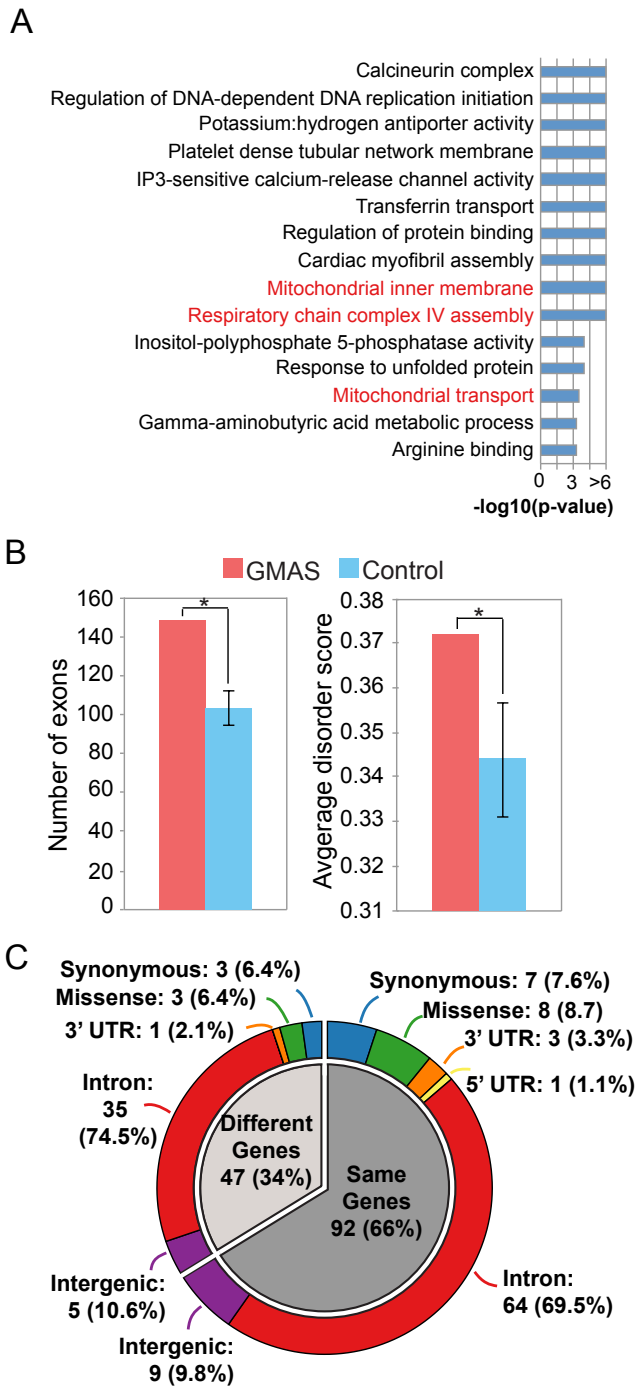


Figure 2.6 Functional relevance of GMAS events. (A) Enriched GO terms among GMAS genes. The GO terms are ranked by *P*-values. Mitochondrial-related terms are colored in red. (B) Intrinsically disordered regions (IDRs) overlapping GMAS exons. Left: number of GMAS and

number of control exons (mean and SD of 1000 sets of controls) overlapping IDRs. GMAS exons are more often located in IDRs than control exons. Right: average disorder scores of amino acids in GMAS and control exons (mean and SD). (* $P < 0.05$, based on the empirical distribution of 1000 control sets) (C) Genomic context of GWAS SNPs (outer ring) located in the same genes as GMAS SNVs (inner pie: darker gray) or different genes as GMAS SNVs (inner pie: light gray). Most GWAS SNPs are located in introns, whose function was elusive.

2.7 References

't Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, Laros JFJ, Buermans HPJ, Karlberg O, Brännvall M, et al. 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol* **31**: 1015–22.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, Raimer A, Sun S, Jensen MA, Hua Y, et al. 2015. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol Cell* **60**: 105–117.

Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–50.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**: 53–9.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587–93.

Busch A, Hertel KJ. 2015. Splicing predictions reliably classify different types of alternative splicing. *RNA* **21**: 813–23.

Cho S, Hoang A, Sinha R, Zhong X-Y, Fu X-D, Krainer AR, Ghosh G. 2011. Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein

determines early spliceosome assembly. *Proc Natl Acad Sci U S A* **108**: 8233–8.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**: D301–8.

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Langaney A, Excoffier L. 2002. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. *Am J Hum Genet* **70**: 207–23.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Evangelou E, Ioannidis JPA. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**: 379–89.

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

Gutiérrez-Aguilar M, Baines CP. 2013. Physiological and pathological roles of mitochondrial SLC25 carriers. *Biochem J* **454**: 371–86.

Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**: 241–5.

Itoh M, Kim C-H, Palardy G, Oda T, Jiang Y-J, Maust D, Yeo S-Y, Lorick K, Wright GJ, Ariza-McNaughton L, et al. 2003. Mind bomb is a ubiquitin ligase that is essential for efficient activation of Notch signaling by Delta. *Dev Cell* **4**: 67–82.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**: 740–3.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40**: 225–31.

Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–11.

Lee J-H, Gao C, Peng G, Greer C, Ren S, Wang Y, Xiao X. 2011. Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ Res* **109**: 1332–41.

Li G, Bahn JH, Lee J-H, Peng G, Chen Z, Nelson SF, Xiao X. 2012. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res* **40**: e104.

Li MJ, Yan B, Sham PC, Wang J. 2014. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief Bioinform*.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**: 1593–9.

Monlong J, Calvo M, Ferreira PG, Guigó R. 2014. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun* **5**: 4698.

Moretti S, Laurenczy B, Gharib WH, Castella B, Kuzniar A, Schabauer H, Studer RA, Valle M,

Salamin N, Stockinger H, et al. 2014. Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res* **42**: D917–21.

Mort M, Sterne-Weiler T, Li B, Ball E V, Cooper DN, Radivojac P, Sanford JR, Mooney SD. 2014. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol* **15**: R19.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**: 100–4.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* **8**: 857–68.

Pandit S, Zhou Y, Shiue L, Coutinho-Mansfield G, Li H, Qiu J, Huang J, Yeo GW, Ares M, Fu X-D. 2013. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell* **50**: 223–35.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–7.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–50.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res* **21**: 1563–71.

Sterne-Weiler T, Sanford JR. 2014. Exon identity crisis: disease-causing mutations that disrupt

the splicing code. *Genome Biol* **15**: 201.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64–9.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. ed. L. Hurst. *PLoS Biol* **4**: e72.

Wang G-S, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**: 749–61.

Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**: 802–13.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution (N Y)* **38**: 1358–1370.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001–6.

Xiao X, Lee J-H. 2010. Systems analysis of alternative splicing and its regulation. *Wiley Interdiscip Rev Syst Biol Med* **2**: 550–65.

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. 2009. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat Struct Mol Biol* **16**: 1094–100.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2014. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, December 18.

Zee JM, Glerum DM. 2006. Defects in cytochrome oxidase assembly in humans: lessons from yeast. *Biochem Cell Biol* **84**: 859–69.

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

CHAPTER 3

RNA Editing in Nascent RNA Affects Pre-mRNA Splicing

3.1 Introduction

The maturation of eukaryotic mRNAs involves an intricate series of molecular processes that modify newly synthesized RNA molecules, such as 5' capping, splicing, RNA editing and polyadenylation (Bentley 2014). These RNA processing steps play essential roles in eukaryotic gene expression (Nishikura 2015; Martins and Fåhraeus 2017). Recent studies have enabled substantial progress in understanding the regulation and function of each RNA processing step. However, it is expected that extensive crosstalk exists between these molecular processes (Mitchell and Parker 2014; Hsiao et al. 2016b), which remains poorly understood.

Among the various RNA processing steps, RNA editing and RNA splicing share the largest overlap in time and space during mRNA maturation (Bentley 2014). Catalyzed by the Adenosine Deaminases Acting on RNA (ADARs), adenosine-to-inosine (A-to-I) editing is the most prevalent type of RNA editing in mammals (Nishikura 2010; Bazak et al. 2014). Inosine is recognized as guanosine (G) by most cellular machineries. As a result, A-to-I editing could alter regulatory motifs, RNA-protein interactions or RNA secondary structures, all of which play important roles in modulating RNA processing (Raitskin et al. 2001; Rieder and Reenan 2012; Washburn and Hundley 2016). Indeed, there are a handful of well-documented examples where A-to-I editing affects alternative splicing (Higuchi et al. 1993; Rueter et al. 1999; Higuchi et al. 2000; Reenan et al. 2000; Beghini et al. 2000; Bratt and Ohman 2003; Flomen et al. 2004;

Laurencikiene et al. 2006; Lev-Maor et al. 2007; Schoft et al. 2007; Agranat et al. 2010; Penn et al. 2013). However, the specific mechanisms for some of these events remain unclear.

Creation or elimination of splice sites and branch points by RNA editing was shown to explain several editing-dependent splicing events (Rueter et al. 1999; Beghini et al. 2000; Flomen et al. 2004; Laurencikiene et al. 2006; Lev-Maor et al. 2007; Schoft et al. 2007). Previous studies also demonstrated that A-to-I editing of *Alu* repeats enabled exonization of the noncoding sequences (e.g., via creation of a 3' splice site), suggesting that RNA editing may be an important means for exon evolution (Athanasiadis et al. 2004; Lev-Maor et al. 2007; Möller-Krull et al. 2008). Another mechanism through which RNA editing may influence splicing is by alteration of RNA secondary structures (Rieder and Reenan 2012). Since ADAR proteins recognize double-stranded RNAs (dsRNAs), substitution of adenosines by inosines may alter the stability of the RNA structures and influence the accessibility of splice sites (Morse et al. 2002; Blow et al. 2004; Levanon et al. 2004; Mazloomian and Meyer 2015). In general, ADAR and splicing machineries may act on the same dsRNA substrates, leading to the interplay between these two processes (Reenan et al. 2000; Rieder et al. 2013). However, structure-mediated splicing modulation by RNA editing or ADAR binding is relatively poorly understood.

In addition to editing-dependent splicing events, the reverse could also be true where splicing affects the outcomes of RNA editing. For example, intron removal could determine the availability of editing complementary sequences (ECSs) that are required to form dsRNA substrates for editing (Levanon et al. 2005). Thus, RNA splicing efficiency could play a key role in regulating RNA editing levels (Licht et al. 2016). In addition, the interaction between ADARs

and splicing factors may also fine-tune editing efficiency (Tariq et al. 2013). These studies have further demonstrated the importance of deciphering the interplay between RNA editing and splicing to achieve a better understanding of the regulation of both processes.

On the global scale, the prevalence of RNA editing events that affect alternative splicing and the associated mechanisms remain unclear. If this functional pathway is widespread, RNA editing must occur co-transcriptionally and prior to the completion of most splicing events. Previous studies provided evidence of co-transcriptional editing by showing that the C-terminal domain of Pol II facilitated site-specific editing and that editing occurs prior to splicing in several genes (Laurencikiene et al. 2006; Ryman et al. 2007). In addition, a recent global study showed that A-to-I editing modifies nascent RNAs co-transcriptionally in *Drosophila* (Rodriguez et al. 2012). Moreover, reduction of ADAR expression in both *Drosophila* and human cells induced global splicing changes (Agrawal and Stormo 2005; St Laurent et al. 2013; Solomon et al. 2013) although the molecular mechanisms are unknown.

Despite the above progress, the kinetic timing of RNA editing during mRNA maturation in human cells remains to be characterized. Furthermore, a global understanding of the influence of RNA editing on splicing is still lacking. Our study aims to address these questions. We first elucidated the timing of RNA editing during mRNA maturation. The finding that more than 95% of A-to-I RNA editing occurred co-transcriptionally in nascent RNAs motivated the hypothesis that the impact of RNA editing on splicing is more widespread than previously appreciated. We then focused on investigating two potential mechanisms: (1) splicing alteration by RNA editing in splice site sequences; (2) splicing modulation by ADAR binding and editing of dsRNA

structures. Supported by experimental validations, our study expands the repertoire of RNA editing sites that influence alternative splicing and provides new insights regarding the interplay between these two processes.

3.2 Results

3.2.1 Subcellular RNA-seq and RNA editing

To examine the timing of RNA editing occurrence during mRNA maturation, we performed cell fractionation (Bhatt et al. 2012) to extract RNA in the chromatin-associated (Ch), nucleoplasmic (Np) and cytoplasmic (Cp) fractions of a human glioblastoma cell line (U87MG). The separation quality of these subcellular fractions was confirmed by Western blot (Supplemental Fig. S1). We obtained RNA-seq data in triplicates, with two types of RNA (polyadenylated and non-polyadenylated) in the Ch fraction and polyadenylated RNA in the Np and Cp fractions (referred to as ChA–, ChA+, NpA+, and CpA+ respectively, Supplemental Table S1). Read mapping was then carried out using our previously published method, which has superior performance in handling single nucleotide variants in the reads (Ahn and Xiao 2015).

We observed that the ChA– fraction yielded a large amount of intronic reads, much more abundant than those in the other fractions (Supplemental Fig. S2). This result is consistent with the expectation that ChA– RNA mainly consists of nascent RNA molecules that are not yet fully spliced. The percentage of intronic reads in ChA+ was much larger than those in later stages such as NpA+ and CpA+, indicating incomplete splicing in the ChA+ fraction, consistent with previous literature (Bhatt et al. 2012). In contrast, the percentage of reads mapped to exons increased in the order of ChA–, ChA+, NpA+, and CpA+. These observations are consistent with

the expectation that the four subcellular fractions capture RNA molecules at different stages of their life cycle. Biological replicates showed highly similar results (Supplemental Fig. S3) and were thus combined for the RNA editing analysis described below.

Next, we identified RNA editing sites using our previously published methods (Bahn et al. 2012; Lee et al. 2013; Zhang and Xiao 2015). In addition, to capture A-to-I editing sites that cluster in close proximity, we implemented a pipeline (similarly as in (Porath et al. 2014)) to identify editing sites in hyper-edited regions (Methods). Altogether, we identified ~80,000 to ~367,000 editing sites per fraction, with an average of 83% comprising A-to-I and C-to-U types (Supplemental Table S2). It should be noted that ChA− had the most non-A-to-G or non-C-to-T types, which may reflect existence of unknown biology (Wang et al. 2014). Among all the A-to-G sites identified (440,821 in total), 64% and 75% were included in the RADAR (Ramaswami and Li 2014) and REDIportal (Picardi et al. 2017) databases, respectively. Hereafter, we will limit all analyses to the predicted A-to-G editing sites, given that our primary focus is on A-to-I editing.

3.2.2 More than 95% of A-to-I editing occurs co-transcriptionally prior to polyadenylation

The subcellular fractionation data allowed us to examine where editing was first observed, thus shedding light on the kinetic timing of RNA editing. We took the union of all editing sites observed in at least one fraction (440,821 in total). Hereafter, we refer to this collection of sites as “editing sites” in general. It is important to note that these editing sites could be edited,

unedited or undetermined (without adequate read coverage) in a specific fraction (Fig. 1A, Methods).

To obtain a detailed view of editing events across subcellular fractions, we enumerated all possible scenarios of an editing site being "edited", "unedited" or "undetermined" in each fraction (Supplemental Table S3). We further classified the editing sites into four groups based on their editing status in each fraction (Fig. 1A, Methods). The groups (1 through 4) included editing sites that were first observed with edited reads in ChA–, ChA+, NpA+ and CpA+, respectively. This grouping represents an approximate kinetic timing of editing occurrence. For example, the ChA– group consists of early editing events occurring in nascent RNAs, whereas the CpA+ group reflects "late" editing that were first observed in the later stage of an RNA's life cycle. A total of 232,976 A-to-I editing sites (53% of all) were categorized into one of the four groups (Fig. 1B, Supplemental Fig. S4). The sites that could not be unambiguously classified into these groups were excluded (Supplemental Table S3), which could exist due to two main reasons. The first reason is inadequate read coverage in at least one of the cell fractions. In general, if an editing site was "undetermined" due to low read coverage in any fractions prior to the fraction where it was observed with edited reads, this editing site cannot be categorized unambiguously. A total of 170,268 such editing sites were identified and excluded from the grouping. Second, an editing site was excluded if it was "unedited" (with adequate read coverage, but no edited reads) in a later fraction but "edited" in an earlier fraction. A total of 37,577 such sites were identified. These editing sites may be associated with complicated regulatory or functional mechanisms, which should be investigated in future studies.

Among all categorized editing sites, more than 226,000 (97%) were included in group 1 (ChA–). This observation suggests that most editing events occur co-transcriptionally in nascent RNA prior to polyadenylation. Since group 4 (CpA+-specific) consists of a very small number of editing sites (84 sites total, which is 0.04% of all sites), we excluded this group from further analyses.

To corroborate the above observed kinetic timing of RNA editing using independent data sets, we analyzed the cell fractionation RNA-seq data from the ENCODE Project in which nuclear poly(A)– (NA–), nuclear poly(A)+ (NA+), and cytosolic poly(A)+ (CA+) data were obtained from human cell lines (HepG2 and K562) (Djebali et al. 2012), and another ENCODE data set of chromatin-associated (ribo-depleted) and nucleoplasmic (ribo-depleted) RNA-seq of K562 cells (Djebali et al. 2012). These analyses showed that more than 95% of editing sites occurred in the NA–or chromatin fraction that largely consists of nascent RNAs or spliced introns (Supplemental Fig. S5, Methods). Furthermore, we analyzed additional RNA-seq data that captured nascent RNA using 4sU labeling, a method that is completely different from cell fractionation (Rabani et al. 2014). These data were derived from mouse dendritic cells after lipopolysaccharide (LPS) stimulation. The results demonstrated again that the majority of RNA editing sites occurred in nascent RNAs shortly after their transcription (Supplemental Fig. S6). Together, our analyses showed widespread co-transcriptional RNA editing in human and mouse cells, which is consistent with the previous finding in *Drosophila* (Rodriguez et al. 2012). Furthermore, our data suggest that the vast majority of RNA editing not only occurs co-transcriptionally, but also precedes polyadenylation.

3.2.3 Comparative analyses of RNA editing sites across subcellular fractions

Next, we carried out a comparative analysis of A-to-I editing in different subcellular fractions. For this purpose, we included editing sites with a total read coverage of at least 5 in all fractions. The levels of RNA editing differ greatly for different editing sites (Fig. 1C). In each subcellular fraction, editing levels varied from 0.01 to 1, with the average being 0.21 to 0.31 (Fig. 1D). We observed that editing levels in the later fractions (NpA+ and CpA+) were generally higher than those from fractions of more nascent RNAs (ChA– and ChA+). To exclude the possibility that the observed editing change is due to loss of low-level editing sites in introns in later fractions, we carried out the same analysis but excluding intronic editing sites and observed the same pattern of higher editing ratios in the later fractions (Supplemental Fig. S7). Indeed, a "monotonicity Z-score" (MZ score) (Wang et al. 2015) calculation showed that the majority of editing sites in the kinetic groups 1 and 2 had positive MZ scores, which reflected an increasing trend in their editing levels across fractions (Fig. 1E). Consistent with this trend, in the CpA+ fraction, mostly composed of mature mRNAs, group 1 editing sites had higher final editing levels than the other groups (Fig. 1F). The increasing trend of editing levels of groups 1 and 2 indicates that additional editing may have occurred for these sites as their associated RNA traveled through the nucleoplasm to the cytoplasm. Group 1 sites may have subjected to the most additional editing, which explains their highest average editing levels in CpA+ among all 3 groups. The above observations still hold if higher read coverage thresholds were imposed for the editing sites (Supplemental Fig. S8). Among the known recoding sites in the literature (Pinto et al. 2014), five were identified as editing sites in our data set, four of which were categorized as group 1 events (Supplemental Table S4). All of these recoding sites had positive MZ scores.

For all kinetic groups, the majority of editing sites were located in the intronic regions (Fig. 2A). The length of the introns harboring the editing sites of different groups demonstrated significant difference, with introns of group 1 being the longest (Fig. 2B). The same trend was observed by randomly sampling 100 introns from each group for this analysis, excluding the possibility that the sample size difference accounted for the observed trend (Supplemental Fig. S9). Thus, the "earlier" editing sites, those that are edited early in nascent RNA, tend to occur in longer introns.

All kinetic groups were mostly composed of editing sites located in *Alu* regions (Fig. 2C). Notably, group 3 editing sites, those first observed in the NpA+ fraction, were more often enriched in non-repetitive regions than the other two groups. Next, we calculated the distance between two closest editing sites. It should be noted that this distance calculation was not restricted to sites within the same group. The result suggested that groups 1 and 2 editing sites were in tighter clusters of editing compared to group 3 editing sites (Fig. 2D), consistent with the relatively smaller enrichment of group 3 sites in *Alu* regions (Fig. 2C). This observation indicates that early editing sites are more often promiscuously edited than late editing sites. In addition, dsRNA predictions suggested that early editing sites were more enriched in dsRNA regions compared to late editing sites (Fig. 2E). Evolutionary conservation analysis showed that early editing sites resided in relatively less conserved regions (Fig. 2F, Supplemental Methods).

Sequence analysis showed that the three groups were associated with similar sequence biases in the –1 and +1 positions around the editing sites (Fig. 2G), resembling the typical UAG signature of ADAR substrates (Lehmann and Bass 2000). Nonetheless, group 3 editing sites

demonstrated additional consensus sequences at the –2 and +2 positions, indicating likely existence of unknown regulatory mechanisms. In an examination of these editing sites upon *ADAR1* knockdown (KD) in U87MG cells (Bahn et al. 2012), we observed that editing in both groups 1 and 2 was lost upon loss of *ADAR1* (Fig. 2H), confirming their dependence on ADAR1. Group 3 was not detected in these data sets possibly due to low editing levels in control cells and inadequate sequencing depth to detect these editing sites.

Together, the above analyses support that all editing kinetic groups (1 to 3) consist of ADAR-regulated editing events. Yet, the differences among these groups indicate that additional regulatory mechanisms may exist to impact RNA editing at different stages of RNA maturation, a topic of future investigation.

3.2.4 The impact of editing on exons with different splicing kinetics

Since our data suggested that most editing sites occurred soon after the RNA was transcribed and likely before some introns were spliced out, we hypothesized that RNA editing affects post-transcriptionally spliced exons more than co-transcriptionally spliced ones. Note that this hypothesis does not preclude the possibility that editing may affect co-transcriptionally spliced exons, depending on the relative kinetics of these two processes on specific exons. To test this hypothesis, we adopted the "completed splicing index" (coSI) metric that quantifies the splicing completeness around an internal exon (Tilgner et al. 2012). We calculated the coSI values to identify co-transcriptionally spliced (coTS) exons and post-transcriptionally spliced (postTS) exons in the U87MG RNA-seq. Similarly as in (Tilgner et al. 2012), we defined postTS exons as those with coSI ≤ 0.75 in the ChA+ fraction and ≥ 0.95 in the NpA+ fraction, and coTS exons as

those with coSI ≥ 0.95 in the ChA– fraction. In total, we identified 6,104 coTS exons and 3,304 postTS exons. We then compared the splicing difference between these two groups of exons upon *ADAR1* KD in U87MG cells. The results showed that the change in exon inclusion in postTS exons was larger than that in coTS exons (Supplemental Fig. S10), suggesting that RNA editing imposes a larger impact on exons undergoing post-transcriptional splicing.

3.2.5 Editing sites located in the splice site regions often disrupt 3' splice sites

Next, we hypothesized that A-to-I editing may affect pre-mRNA splicing by altering splice site sequences. To this end, we searched for editing sites identified in the ChA– fraction that were located in the splice site consensus. Specifically, a 9-mer sequence was defined as the 5' splice site (5'ss) with 3nt in the exon and 6nt in the intron. A 23-mer sequence was defined as the 3'ss with 3nt in the exon and 20nt in the intron, based on previous literature (Yeo and Burge 2004). To carry out a comprehensive search, we included both annotated exons (GENCODE v24lift37) and novel exons identified from *ADAR1* KD RNA-seq data sets (Bahn et al. 2012) (see Methods). A total of 492 editing sites were found to reside in splice site sequences, which were predicted to alter the splice site strength to different degrees (Yeo and Burge 2004) (Fig. 3A, Supplemental Table S5). We observed that most (78%, 385) of these editing sites were located in the 3'ss, altering the splice site sequences from AG to GG (3'ss: –2 location). For these sites, A-to-I editing resulted in a dramatic reduction of the predicted splice site strength (Fig. 3A), which is expected since the canonical dinucleotide splice site sequences were altered. We thus focused on this group of exons in the analyses below.

We next examined the splicing patterns of these exons in *ADAR1* KD RNA-seq data (Bahn et al. 2012). A total of 209 such exons were identified with adequate read coverage to calculate their percent spliced-in (PSI) values (Methods). Based on the RNA-seq data, 158 of the 209 exons were categorized into a single category of alternative splicing: skipped exons (SE), alterative 3'ss (A3SS) or retained introns (RI). The remaining exons had complex splicing patterns (e.g., both SE and A3SS). For all exons, we calculated their PSI values in *ADAR1* KD and control data sets. Among the 158 exons with a single alternative splicing type, 137 (87%) had PSI changes in the same direction as expected based on the alteration in the 3'ss sequence by the A-to-G change (Fig. 3B). The expected direction of PSI changes for exons with complex splicing is unknown. It should be noted that the level of PSI change is expected to be low because the editing levels of the 3'ss-altering editing sites are relatively low (mean: 0.42, median: 0.36) and the change from A to I may not completely abolish splicing activity.

As an independent analysis, we searched for splice site-altering editing sites in the NA– samples of ENCODE cell fractionation RNA-seq in HepG2 and K562 cells. Similar to the U87MG data, the majority of splice site-disrupting editing sites are located in the 3'ss and alter AG to GG (Supplemental Fig. S11A), resulting in a total of 498 such unique editing sites if combining data from the three cell lines. To verify their impact on splicing, we analyzed the splicing patterns of the associated exons as described above, using the ENCODE *ADAR1* KD RNA-seq of the respective cell lines (Nostrand et al. 2017). Among the 35 exons with a single alternative splicing type, 29 (83%) showed the expected direction of PSI changes upon *ADAR1* KD (Supplemental Fig. S11B). Altogether, our analyses of multiple data sets led to identification

of an unexpectedly large number of editing sites located in splice site regions that can potentially alter splicing.

3.2.6 Exons with A-to-I editing in the 3'ss tend to be highly conserved

To characterize the exons that harbor A-to-I editing sites in their 3'ss, we next examined the overlap of these exons with annotated *Alu* repeats. About 25% of these exons overlapped *Alu* sequences, a much higher percentage than what was observed for control exons that do not have known editing sites in the 3'ss (Fig. 3C). Similarly, the flanking intronic regions of these exons (within 500nt from the exons) were also enriched with *Alu* repeats (Fig. 3C). This observation is consistent with the fact that A-to-I editing sites often reside in *Alu* elements.

The above *Alu*-containing exons had editing sites that weaken the 3'ss strength (from AG to GG). This alteration in 3'ss is opposite to the previously reported example where A-to-I editing creates a 3'ss (changing AA to AG) and enables *Alu* exonization (Lev-Maor et al. 2007). Thus, it is unlikely that the exons in our study represent *Alu* exonization cases resulting from RNA editing. In contrast, since RNA editing appears to destroy the 3'ss, we asked whether these exons are under relaxed selection pressure for their preservation as exons. Because *Alus* are highly primate-specific, we examined the density of human single-nucleotide polymorphisms (SNPs) in *Alu*-containing exons as a proxy for sequence conservation. We observed that *Alu*-containing exons with 3'ss editing had significantly lower SNP density than control exons (that were also *Alu*-containing, but without known 3'ss editing) (Fig. 3D, Supplemental Methods). The Fay and Wu's H values (Pybus et al. 2014) were not significantly different between *Alu*-containing exons with 3'ss editing and the control exons (Supplemental Fig. S12). Thus, it is

likely that edited *Alu*-containing exons are under selection to preserve their sequences more than random controls.

Next, we examined the sequence conservation of non-*Alu*-containing exons with 3'ss editing sites across 46 vertebrates spanning primates to fish (Siepel et al. 2005). Since these exons demonstrated multiple types of alternative splicing in the RNA-seq data (Fig. 3B) and because different types of alternative splicing may have distinct conservation patterns, we focused on 96 exons that were alternatively skipped, the largest category of alternative splicing. Compared to control skipped exons from annotation (Katz et al. 2010), the 3'ss-edited exons had significantly higher conserved exonic and splice site sequences (Fig. 3E, Supplemental Methods). Together, both *Alu*-containing and non-*Alu* exons with 3'ss editing showed evidence of higher conservation than their corresponding controls.

Based on Gene Ontology (GO) analysis (Supplemental Methods), the genes that contain 3'ss-edited exons (*Alu* or non-*Alu*) were enriched in functional categories related to protein trafficking, cellular metabolism (protein or energy-related), RNA processing or cytoskeletal structures (Supplemental Fig. S13). Thus, consistent with the highly conserved nature of the exons, these genes are potentially implicated in critical cellular functions.

3.2.7 Experimental verification of editing-dependent alternative splicing

To experimentally confirm the impact of RNA editing on splicing, we randomly picked five 3'ss-edited exons and tested whether their splicing patterns were altered upon *ADAR1* KD in U87MG cells (Fig. 4A, Supplemental Methods). Note that the endogenous editing levels are

generally low, and these five candidates had an average editing level of 0.36. Thus, the endogenous splicing changes of these exons upon *ADAR1* KD are not expected to be high. Nevertheless, all five exons demonstrated splicing changes in the expected direction (although some did not reach statistical significance) (Fig. 4B). Two exons (in genes *FBXL4* and *DENND4A*) were alternatively skipped exons (SEs), with increased exon inclusion levels in *ADAR1* KD cells. This observation is consistent with the expectation that the edited version of the 3'ss (GG) weakens the splicing signal and causes exon skipping. The other three exons each had an alternative 3'ss (A3SS). In these cases, given the editing event, the original 3'ss was partially or completely abolished. Instead, an alternative 3'ss in the downstream (*PARP4*, *PDE4DIP*) or upstream (*RAP1GDS1*) region was used, which created a shorter or longer form of the original exon, respectively.

As a complementary experiment, we performed minigene reporter assays on the above five exons and two additional examples, all of which had 3'ss editing (Methods, Supplemental Table S6). We created two versions of the constructs for each target, one harboring the unedited (A) and the other for the edited (G) alleles in the 3'ss sequence. For the SE events (in genes *FBXL4*, *DENND4A*, and *PDE4DIP*), the exons had significantly reduced inclusion levels in the G version of the construct (Fig. 4C, top panels). Consistent with the endogenous results described above, these data confirm that RNA editing reduces the 3'ss strength and causes exon skipping. In addition, the four A3SS exons also demonstrated expected splicing differences given the A and G alleles (Fig. 4C, bottom panels), where an alternative 3'ss in the downstream (*PARP4*, *PDE4DIP*) or upstream (*RAP1GDS1*, *ENTPD1AS1*) region was used in the presence of

the G allele. Taken together, our experimental results support that RNA editing could alter splicing by modifying the splice site sequences.

3.2.8 Regulation of splicing by ADAR acting on dsRNA structures

Since the majority of RNA editing sites resides outside of canonical splice sites, we sought to investigate whether RNA editing may affect splicing through other mechanisms. One possible mechanism for these editing sites to modulate splicing is through alteration of splicing regulatory motifs (Fu and Ares 2014). However, a motif analysis similar to that in our previous study (Hsiao et al. 2016a) did not yield significant enrichment of any class of motifs. Hence, we examined a second hypothesis that considers the involvement of RNA secondary structures in splicing regulation.

To this end, we first identified a set of exons (GENCODE v24lift37) that showed a PSI change (in either direction) of at least 10% upon *ADAR1* KD in U87MG, HepG2 or K562 cells as reflected in the respective RNA-seq data (Bahn et al. 2012)(Nostrand et al. 2017). The splicing of these exons could be directly or indirectly affected by ADAR1. To identify direct targets of ADAR1, we further required existence of exonic editing sites or intronic editing sites within 500nt of the exon boundaries. A total of 555 exons satisfied these criteria. For these exons, we carried out BLASTN analysis to identify those located in dsRNA regions (Altschul et al. 1990) (Methods). Thirty-three exons (27 of which overlapped *Alu* elements) passed a stringent criterion for dsRNA formation (Fig. 5A). Therefore, these exons are likely located in ADAR1-binding substrates, whose splicing is directly affected by ADAR1.

To validate the above hypothesis, we examined our previously published CLIP-seq data set (Bahn et al. 2015). Among the 33 exons, 82% were located in dsRNA structures that were associated with ADAR1 CLIP peaks, providing strong support for ADAR1 binding. Furthermore, since the initial requirement of exon PSI change (≥10%) upon *ADAR1* KD was not direction-specific, we asked whether the 33 predicted ADAR1 direct targets demonstrated splicing change in a particular direction. The data showed that the majority (73%) of the 33 exons had increased PSI values upon *ADAR1* KD (Fig. 5B). These data suggest that ADAR1 represses splicing of these exons that reside in dsRNA regions in the pre-mRNA.

To provide experimental validation of the ADAR1-dependent splicing changes, we tested the splicing patterns of four exons in U87MG cells of *ADAR1* KD. These exons were randomly chosen from those predicted to be repressed by ADAR1 and associated with long dsRNA structures in the above analyses, the structures of which were also confirmed by mfold (Zuker et al. 1999) (Supplemental Fig. S14). Three of the four exons had CLIP-seq binding sites close to the associated exon-intron boundaries (Fig. 5C, Supplemental Fig. S14). Endogenous splicing inclusion levels of these exons were estimated via RT-PCR followed by gel electrophoresis. Three exons demonstrated a trend of higher splicing levels upon *ADAR1* KD, consistent with the predictions via RNA-seq (Fig. 5C). Note that the magnitudes of these splicing changes are generally small, which may explain the lack of statistical significance for two of the three exons. The fourth exon (in the gene *C17orf67*) failed this experiment due to existence of multiple PCR amplification products indicating likely unspecific primers or complex splicing patterns (Supplemental Fig. S15).

The above results suggest that ADAR1 may bind to dsRNA structures and repress splicing of the neighboring exons. Since we have shown that RNA editing occurs early in the nascent pre-mRNA, and it is known that splicing is also co-transcriptional (Bhatt et al. 2012), this observation likely reflects the competition between ADAR1 and the spliceosomal complex to bind to and/or process the pre-mRNA.

3.3 Discussion

RNA editing and splicing are both RNA processing steps that may significantly diversify gene expression. An increasing number of studies have examined the intricate crosstalk between these processes (Solomon et al. 2013; Mazloomian and Meyer 2015; Licht et al. 2016). In general, there are two broad categories of mechanisms via which RNA editing may influence alternative splicing. In the first category, the RNA editing sites themselves impact splicing directly, by altering *cis*-regulatory sequences or by changing the stability of dsRNA structures, which may affect the interaction of splicing machineries with RNA. In the second category, the editing machinery (ADAR proteins), instead of editing sites, influences splicing either by competitively binding to dsRNA regions and precluding access of splicing machineries to the RNA or by recruiting *trans*-acting factors that regulate splicing.

The impact of RNA editing on splicing has been demonstrated in a number of example genes although the underlying mechanisms sometimes remain unknown (Higuchi et al. 2000; Agrawal and Stormo 2005; St Laurent et al. 2013; Solomon et al. 2013; Goldberg et al. 2017). Despite the increasing recognition of the complex interplay between various post-transcriptional gene regulation steps, a global understanding of the influence of RNA editing on splicing is still

lacking. Here, we fill in this gap by performing a genome-wide study that leverages subcellular fractionation RNA-seq data. We first showed that the vast majority of RNA editing occurs co-transcriptionally prior to polyadenylation in human cells. This observation provided a foundation for the hypothesis that many RNA editing events (i.e., the resulted editing sites or the ADAR binding events) may alter splicing. Indeed, using a combination of different types of RNA-seq data sets, we identified more than 500 editing events that could alter splicing by changing splice site sequences or through ADAR's interaction with dsRNA structures.

The usage of chromatin-associated non-polyadenylated (i.e., ChA−) RNA-seq is essential to this study. In identifying RNA editing sites using standard polyadenylated RNA-seq data, it is a general practice to exclude RNA-DNA mismatches located in the splice site regions from the list of predicted editing sites (Lee et al. 2013). This practice is due to the concern that these sites are likely false positives resulting from reads erroneously mapped to the exon-intron boundaries because standard RNA-seq should be depleted of such "unspliced" reads. Thus, previous studies likely missed many RNA editing sites in the splice site regions. In contrast, in ChA− RNA, intronic reads are very abundant (Supplemental Fig. S2), reflecting incomplete splicing (Bhatt et al. 2012). Thus, these data afforded a unique opportunity to capture RNA editing sites in the 5'ss and 3'ss sequences.

In this study, we identified exons whose splicing was likely affected by ADAR1's binding to dsRNA structures. These exons, 33 in total, were required to form stable dsRNA structures with their neighboring intronic regions. The exons were also required to demonstrate at least 10% change in PSI values upon *ADAR1* KD, in either direction. Nonetheless, we

observed that the majority of these exons had positive PSI changes upon *ADAR1* KD, suggesting that ADAR1 plays a repressive role in their splicing. It should be noted that it is unlikely that the editing sites themselves alter the secondary structures, since the predicted free energy of these structures barely changed when A's were substituted by G's. Thus, the most likely model to explain these data is that binding of ADAR1 to dsRNA prevented the spliceosome or splicing regulators from accessing the RNA or enabling splicing of the related exons. Similar examples were reported in the literature (Reenan et al. 2000; Rieder et al. 2013), and our study expanded the repertoire of exons regulated by this mechanism.

It should be noted that the endogenous splicing changes imposed by editing or ADAR1 binding to dsRNAs are generally small, which is likely due to the relatively low level of RNA editing for most sites. We applied a series of stringent criteria to identify exons whose splicing is likely affected by ADAR1 binding. The small number of exons resulted from this analysis represent highly confident predictions, which may constitute only a fraction of all exons influenced by this mechanism. Although the splicing change of each exon is small, the collective changes exerted by many events together may confer substantial impacts on cellular function. In addition, such functional impact of RNA editing may be highly important under certain conditions, such as environmental stress or diseases, a subject for future investigations.

Lastly, we found that exons with editing sites in the 3'ss AG sequences were highly conserved compared to controls. This observation indicates that these exons and their host genes may have important functional roles. In addition, RNA editing, by weakening a canonical 3'ss, introduces novel splicing patterns to these exons. These novel gene expression products,

although often having low endogenous levels, may diversify the transcriptome and provide evolutionary substrates for selection without perturbing cellular function significantly.

3.4 Methods

3.4.1 Cell fractionation and RNA-seq library construction

U87MG cells were fractionated as previously described (Bhatt et al. 2012) with some modifications. Briefly, 5×10^6 U87MG cells were treated with the plasma membrane lysis buffer (10 mM Tris-HCl, pH 7.5, 0.1% - 0.15% Nonidet P-40, 150 mM NaCl) on ice for 4 min after homogenization by flicking. The sample was centrifuged for 10 min, at 4°C, 15,000g after loading the lysate onto 24% sucrose cushion (24% RNase-free sucrose in plasma membrane lysis buffer) using large orifice tips. The supernatant was extracted as the cytoplasmic fraction. After washing with 1× PBS/1 mM EDTA, the pellet was resuspended in 100 μL of prechilled glycerol buffer (20 mM Tris-HCl [pH 7.9], 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 0.125 mM PMSF, 50% glycerol) by gentle flicking of the tube. An equal volume (100 μL) of nuclei lysis buffer (10 mM HEPES, pH 7.6, 1 mM DTT, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 0.3 M NaCl, 1 M Urea, 1% Nonidet P-40) was added. The sample was then vortexed vigorously (highest setting on vortex) for 2×2 seconds. After incubation for 10 min on ice, the nucleoplasm and chromatin fraction were then separated by centrifugation at 4°C, 15,000g for 2 min. Fractionation efficiency was validated by Western Blotting using antibodies specific to marker proteins for each fraction: β-tubulin (Sigma, Cat#T8328) for cytoplasm, U1-70k (a kind gift from Dr. Douglas Black) for nucleoplasm, and Histone 3 (Abcam, Cat#ab1791) for chromatin.

Total RNA was extracted from each fractionated sample using TRIzol LS (Thermo Fisher Scientific, Cat# 10296028), and ribosomal RNA was depleted using the RiboMinus™ Transcriptome Isolation Kit (Thermo Fisher Scientific, Cat# K1550-02) according to manufacturer's instructions. Polyadenylated RNA was then selected by Oligo $(dT)_{25}$ Dynabeads (Thermo Fisher Scientific, Cat# 61002) following the instructions provided by manufacturer. The flow through was collected, and the RNA was extracted using TRIzol LS as non-polyadenylated RNA. Three biological replicates (independent cell culture) were obtained for each fraction.

To prepare for RNA-seq libraries, we started with 50ng of RNA extracted from each sample. Paired-end strand-specific RNA-seq was performed with the NEBNext Ultra Directional RNA Library Prep Kit (NEB, Cat# E7420) with the following adjustments to the manufacturer's instructions: fragmentation was performed for 12 minutes at 94°C and PCR cycling was set to 12 cycles. Unique index was ligated to each sample RNA library construct using the NEBNext® Multiplex Oligos for Illumina® Index Primers Sets 1&2 (NEB, Cat# E7335S & E7500S).

3.4.2 Detection of RNA editing sites

The RNA-seq datasets were mapped to the hg19 genome and Ensembl (Release 75) transcriptome (Yates et al. 2016) using a stringent mapping method (named RASER) described in our previous work (Ahn and Xiao 2015). RASER parameters were specified as m = 0.05 and b = 0.03. Uniquely mapped reads were retained for further analyses. We realigned a subset of reads to the GRCh38 (hg38) assembly to test whether the version of human genome assembly will lead to a significant difference in the alignment results. We observed highly similar results using hg19

and hg38-based alignments, with an overlap rate of >96%, suggesting that the version of genome assembly does not affect our conclusions significantly. To capture hyper-edited reads, we took an approach similar to that described in (Porath et al. 2014). Specifically, we collected all unmapped read pairs and converted all As to Gs in the reads, thus creating read sequences with C, G and T nucleotides only. Likewise, we created a three-nucleotide version of the hg19 genome consisting of only C, G, and T bases. The converted reads were aligned against the converted hg19 genome, and uniquely aligned read pairs were retained. After mapping, the original A nucleotides in the reads were reinstated (Supplemental Material). Reads mapped in the hyper-editing alignment step were pooled with those uniquely mapped in the first round of alignment. Triplicates of the same experiment were combined for RNA editing detection to increase read coverage. Mismatches identified in the final mapped reads were examined. Final RNA editing sites were identified after removing likely false positives due to sequencing errors, inaccurate mapping to repetitive sequences or spliced junctions, or technical biases (such as strand bias of reads) (Bahn et al. 2012; Lee et al. 2013; Zhang and Xiao 2015). In addition, a final predicted editing site was required to be supported by at least 2 reads with the edited G nucleotide, and covered by at least 5 reads in total. It should be noted that predicted editing sites located close to exon-intron junctions were not excluded from those identified in the ChA− fraction, but were excluded from the list of editing sites identified in the other fractions.

3.4.3 Definition of RNA editing kinetic groups

We used ChA–, ChA+, NpA+, and CpA+ samples to represent the approximate kinetic order of RNA maturation. RNA editing sites identified from all fractions were pooled together to obtain the overall set of editing sites considered in this study. For each editing site, we then examined

whether it was edited, unedited or undetermined in each fraction as follows: (1) a site was deemed "edited" in one fraction if it had ≥ 1 read with the edited G nucleotide, regardless of the total number of reads; (2) a site was deemed "unedited" in one fraction if its total read coverage was at least *k* (see below), but no read contained the edited G; (3) a site was deemed "undetermined" if its read coverage was inadequate (< *k*), and no read contained the edited G. Given these three possibilities, an editing site may fall into one of 65 possible categories, considering its status in the four fractions (Supplemental Table S3). In this study, we focused on the categories that have unambiguous interpretation regarding the kinetic timing of an editing site's occurrence and further organized them into 4 groups (Fig. 1A).

To unambiguously determine whether an editing site was "unedited" in a specific fraction, the parameter *k* was set to be 20 in order to achieve adequate statistical power to capture edited reads (Li et al. 2012). One exception was the *k* value for group 1 editing sites. A group 1 editing site was required to be "edited" in ChA−, and either "edited" or "undetermined" in the other three fractions (Fig. 1A). For this group, we used $k = 5$, a relatively low value that increases the chance of calling "unedited" sites in the other fractions and being excluded from group 1. This cutoff thus results in a conservative estimate of the number of group 1 editing site. For example, if *k* was set to be 10 or 20, a total of 77,169 and 81,248 group 1 editing sites were identified respectively, a much larger number than the 68,539 sites resulted from $k = 5$. Thus, the estimated proportion of group 1 (co-transcriptional) editing sites in this study is relatively conservative.

From the ENCODE Consortium, we obtained RNA-seq data of subcellular fractions of HepG2 and K562 cells. Each cell line had data sets derived from nuclear poly(A)− (NA−), nuclear poly(A)+ (NA+), and cytoplasmic poly(A)+ (CA+) RNA (Djebali et al. 2012). Using these data, the editing kinetic groups of A-to-I editing sites were defined similarly as for the U87MG data. Specifically, group 1 editing sites included those edited in NA−, and either edited or undetermined in NA+ and CA+. Group 2 editing sites included those that were unedited in NA−, but edited in NA+, and either edited or undetermined in CA+. Group 3 consisted of sites that were unedited in neither NA− nor NA+, but edited in CA+. For groups 1, 2 and 3, the *k* parameters were set to be 5, 20, and 20, respectively, similarly as for the U87MG data.

The editing kinetic groups of A-to-I editing sites in mouse dendritic cells were derived from published time-course 4sU-labeled RNA-seq data sets (Rabani et al. 2014). 4sU was used to label newly synthesized RNA at different time points following LPS stimulation. RNA editing sites in 4sU labeled RNA were compared to those derived from total RNA collected at 180 minutes after LPS stimulation (Total-180min), which was the last time point available in this data set. Early edited sites were defined as those with editing ratio (ER) > 0 in both 4sU-RNA and Total-180min samples, whose ER did not change significantly between these two samples (Fisher's exact test, $P > 0.05$). Intermediate group consisted of editing sites with ER > 0 in both 4sU-RNA and Total-180min, but significant increase of ER was observed in Total-180min relative to 4sU-RNA (Fisher's exact test, $P < 0.05$). Late editing refers to those editing sites with ER = 0 in 4sU-RNA, and ER > 0 in Total-180min RNA.

3.4.4 Prediction of novel exons and PSI calculation

We used our previously developed method (Lee et al. 2011) and the GENCODE v24lift37 transcriptome (Harrow et al. 2012) as reference annotation to identify novel exons in RNA-seq data of *ADAR1* KD and control samples. This method examines spliced junction reads and read coverage in putative exonic and intronic regions to call exons. These putative exons were compared with those in the reference annotation to identify novel exons. We required a novel exon to have a read density of ≥ 10 (normalized by read length and exon length) and ≥ 2 exon junction reads supporting each exon boundaries. The PSI values of annotated and novel exons were calculated following a previously developed protocol (Schafer et al. 2015). Each exon was required to have ≥ 10 reads in total (corresponding to exon inclusion and exclusion) or ≥ 2 exon exclusion reads in the KD or control data set.

3.4.5 Prediction of RNA secondary structures

To test whether the predicted A-to-I editing sites are located in or near dsRNA regions (Fig. 2E), we used a previously published approach (Bahn et al. 2012). We extracted 4,001bp genomic sequences with the editing sites located in the middle. This sequence was then reverse-complemented and aligned against the immediate neighborhood (200bp flanking each side) of the editing sites using BLASTN (Altschul et al. 1990). Excluding the case of self-alignment, if the second-best alignment has an alignment length ≥ 50 and total identity ≥ 80% (parameters chosen to include most known A-I editing with strong dsRNA structures), we concluded that the editing site resides in dsRNA structures.

For the structure analysis presented in Fig. 5A, we collected a list of exons with at least 10% PSI changes upon *ADAR1* KD (called "AS exons" henceforth) based on the RNA-seq data

of three human cell lines (U87MG, K562 and HepG2). We focused on the AS exons with exonic or intronic editing sites within 500nt. AS exons without editing sites in these regions were used as controls. To identify candidate dsRNA structures encompassing the AS exons, we ran BLASTN (Altschul et al. 1990) to align the AS exonic sequences and the reverse-complement sequences of the corresponding flanking introns. Structure with a percent identity > 70% and alignment E-values ≤ 0.001 were retained as putative dsRNA structures associated with the AS exons.

3.4.6 Minigene constructs

Genomic regions encompassing the candidate exon and ~500nt upstream and downstream flanking introns (including the editing site) were amplified using genomic DNA from HEK293 cell. Primers used in this study are listed in Supplemental Table S6. After double digestion by *Hin*dIII and *Sac*II or *Eco*RI and *Sac*II, the DNA fragment was sub-cloned into a splicing reporter used in a previous study (Xiao et al. 2009).

3.5 Data Access

RNA-seq data derived from subcellular fractions of the U87MG cells from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE105773. Custom scripts are available in Supplemental_Code.zip.

3.6 Acknowledgements

We thank Dr. Douglas Black for helpful discussions and for the kind gift of the U1-70k antibody. We thank members of the Xiao laboratory for helpful discussions and comments on this work. We also acknowledge the UCLA Broad Stem Cell Research Center High-Throughput Sequencing Core for assistance in sequencing. We thank the ENCODE Project Consortium for generating the *ADAR1* knockdown RNA-seq data sets (Dr. Brenton Graveley group) and the nuclear and cytoplasmic RNA-seq data sets (Dr. Thomas Gingeras group). This work was supported in part by grants from the National Institutes of Health (U01HG009417, R01HG006264, R01AG056476) to XX. YHH was supported by the Eureka and Hyde fellowships of UCLA.

3.7 Figures

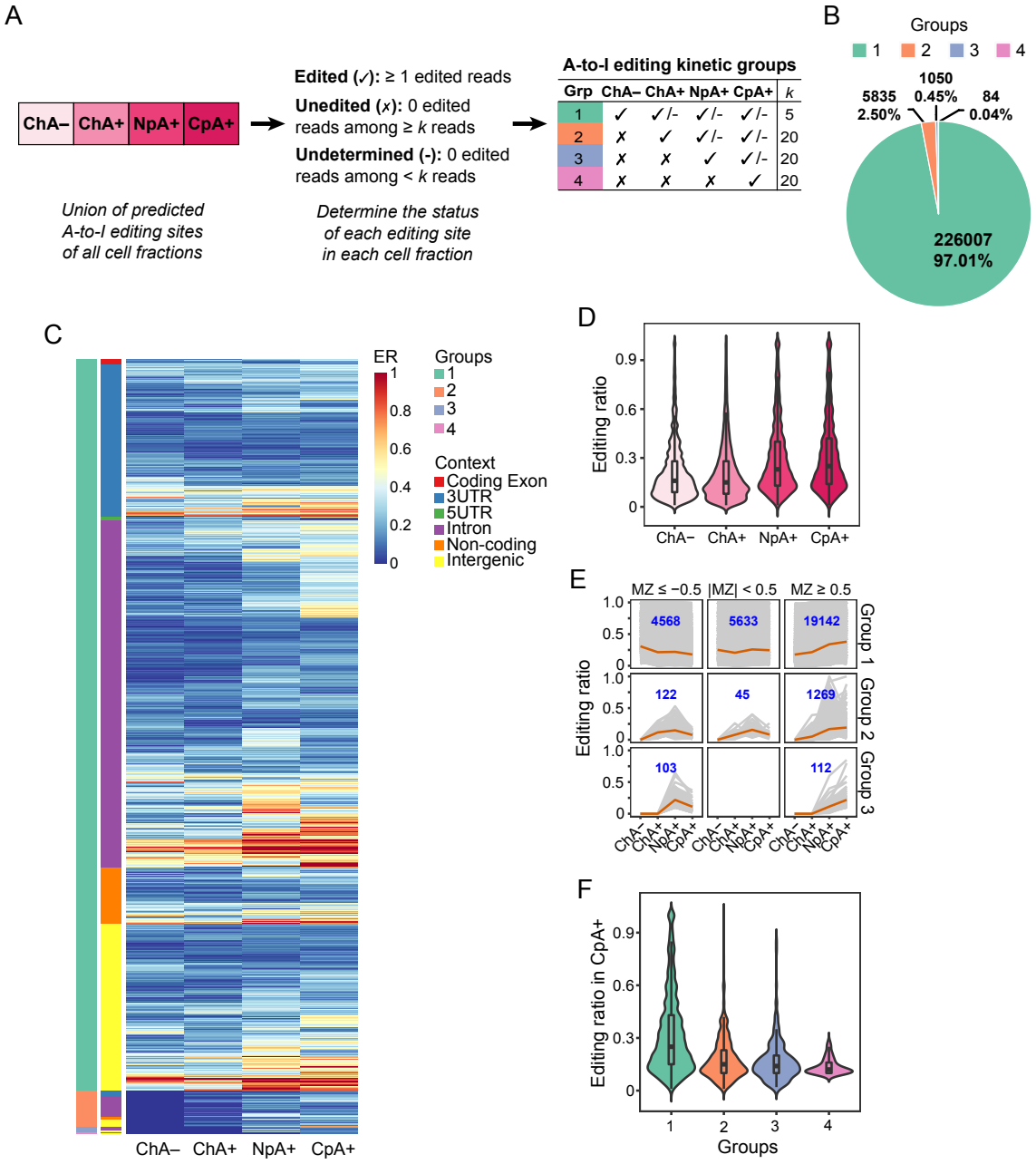


Figure 3.1 Kinetic groups of RNA editing sites identified via subcellular RNA-seq. (A) Categorization of RNA editing sites into four kinetic groups using cell fractionation RNA-seq. Editing sites in each group are labeled as edited (✓), unedited (✗) or undetermined (-) in each

subcellular fraction. The read coverage cutoffs of the editing sites (*k*) are illustrated in the table (see Methods for details). (B) Number and percentage of A-to-I RNA editing sites in each editing kinetics group. (C) The editing ratios (ER) of each editing site in the four groups. The genomic location of an editing site is shown in color codes. All editing sites included in this analysis were required to have at least 5 total reads in each fraction (same for panels D, E and F). A total of 29,345, 1,437, 220 and 84 sites were included for group 1, 2, 3 and 4 respectively, with group 1 editing sites accounting for 94% of all. (D) Editing ratio distribution of the editing sites in (C) (union of all four kinetic groups) in each subcellular fraction. Note that for each fraction, editing ratios of the union of all groups are shown. (E) Editing ratios in different subcellular fractions of editing sites in (C) with different ranges of MZ scores. Each gray line represents one editing site. The average editing ratios are highlighted in orange. The number of editing sites in each panel is shown in blue. It should be noted that MZ scores cannot be calculated for sites with constant editing ratios across two consecutive fractions, thus were excluded from this analysis. (F) Editing ratios of each kinetics group observed in mature mRNAs (CpA+).

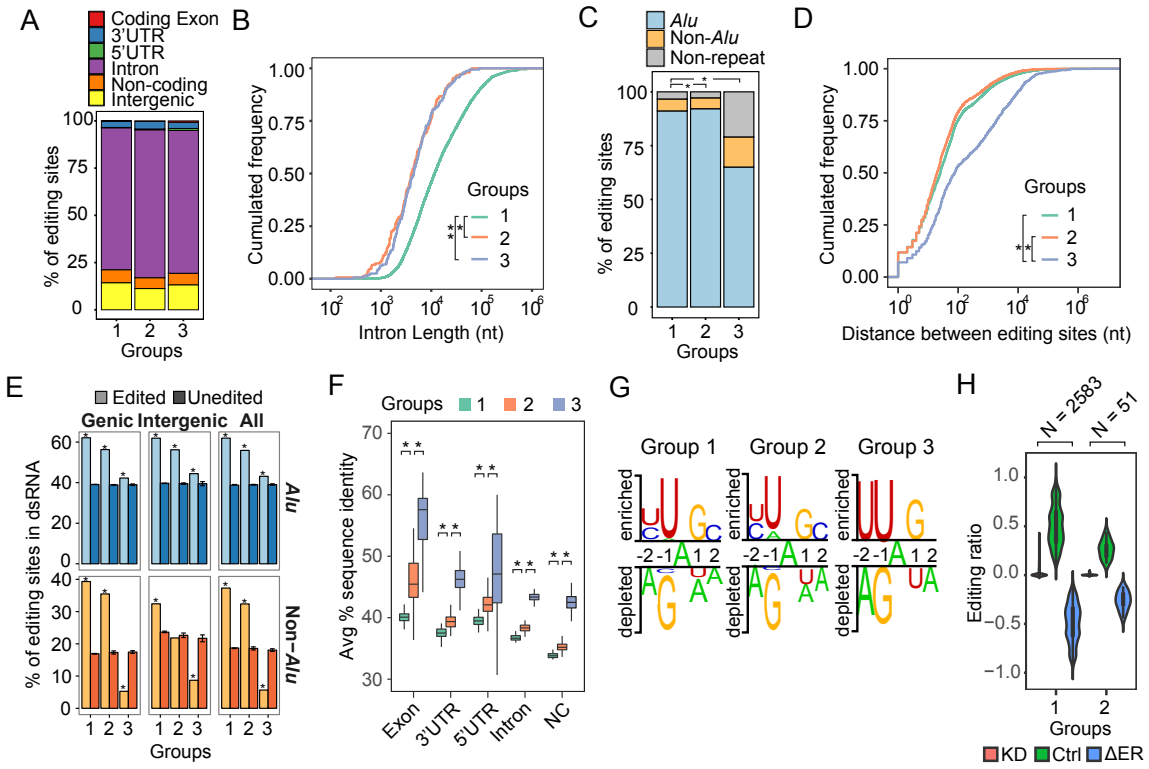


Figure 3.2 Characteristics of editing sites in different kinetic groups. (A) Genomic context of the editing sites. (B) Length of introns harboring editing sites. *: $P = 1.75 \times 10^{-4}$ for group 1 vs. 2, **: $P = 3.02 \times 10^{-4}$ for group 1 vs. 3 (Kolmogorov–Smirnov test). Groups 2 and 3 were not significantly different. (C) Presence of editing sites in *Alus*, non-*Alu* repeats and non-repeats. *: $P < 2.2 \times 10^{-16}$ for group 1 vs. 3 and group 2 vs. 3 (Fisher's exact test). Groups 1 and 2 were not significantly different. (D) Closest distance (in nucleotides) between editing sites. For each editing site in a specific group, its closest neighboring editing site (regardless of group) was used to calculate this distance. *: $P = < 2.2 \times 10^{-16}$ for group 1 vs. 3 and group 2 vs. 3 (Kolmogorov–Smirnov test). Groups 1 and 2 were not significantly different. (E) Percentage of editing sites located in dsRNA structures predicted using BLASTN. Unedited: random A's in the same type of genomic region as editing sites (see Methods). Error bars: 95% confidence interval based on 100 random trials. *: $P = < 2.2 \times 10^{-16}$ (Wilcoxon Rank Sum test). (F) Average percent sequence

identity across primates (Human, Chimp, Gorilla, Orangutan, Rhesus, Baboon, Marmoset, Tarsier, Mouse lemur, Bushbaby and Tree shrew) calculated using the 46-way multiz alignments (Kent et al. 2002). NC: non-coding transcripts. *: $P < 2.2 \times 10^{-16}$ (Wilcoxon Rank Sum test). (G) Sequence consensus around the editing sites of each group, with the editing sites at position 0. Sequence motifs were identified via the two-sample logo program (Vacic et al. 2006). Random A's that were in the same type of genomic region as editing sites were used as negative controls. (H) Editing ratios of group 1 and 2 editing sites in U87MG *ADAR1* knockdown (KD) and control (Ctrl) RNA-seq data (Bahn et al. 2012). The differences in editing ratios (ΔER) were calculated as (KD – Ctrl). The number of editing sites identifiable in the RNA-seq data is shown for each group (N).

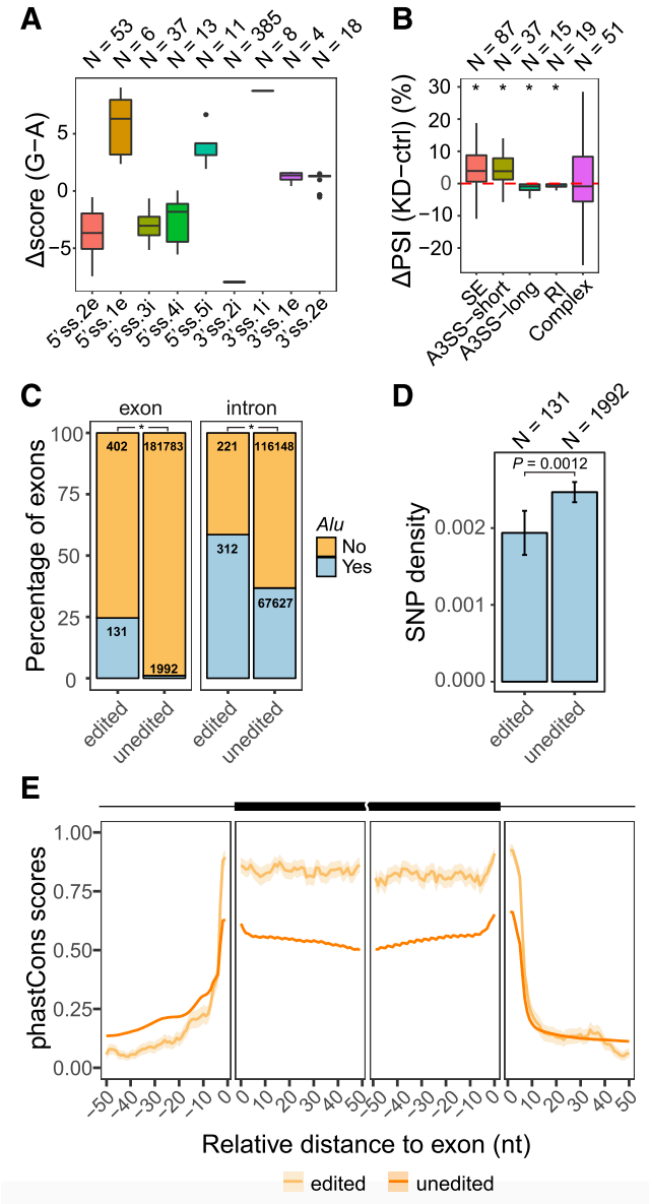


Figure 3.3 RNA editing sites located in splice site regions. (A) Boxplot of the difference in splice site strength caused by A-to-I editing. Splice site scores were calculated using the maxEnt method (Yeo and Burge 2004). 5'ss: 5' splice site; 3'ss: 3' splice site; *i*: intron; *e*: exon. The number before *i* or *e* represents the distance (in nucleotides) from the exon-intron boundaries. N: number of editing sites. Nucleotide positions without any editing sites in our data were excluded from this plot. (B) Box plot of PSI changes (ΔPSI) upon *ADAR1* KD in U87MG cells measured via RNA-seq for exons with editing sites at the −2 intronic position of 3'ss (3'ss.2*i*), altering the

canonical 3' acceptor site from AG to GG. SE: skipped exons; A3SS-short: alternative 3'ss generating shorter exons given the edited G in the 3'ss; A3SS-long: alternative 3'ss generating longer exons given the edited G in the 3'ss; RI: retained introns; Complex: complex splicing patterns. N: number of exons. *: $P < 0.05$ for $\Delta PSI \neq 0$ (Wilcoxon Rank Sum test). (C) Percentage of exons or their upstream introns that overlap *Alu* repeats. "Edited": exons or introns with editing sites in the 3'ss AG sequence. "Unedited": all internal exons or introns without editing sites in the 3'ss. The number of exons or introns in each category is shown. The edited groups are significantly more enriched with *Alus* than unedited groups for both panels. *: $P < 2.2 \times 10^{-16}$ (Fisher's exact test). (D) Density of SNPs (common SNPs in dbSNP version 150) in *Alu*-overlapping exons as described in (C). Error bars represent standard errors. Edited exons have smaller SNP density than unedited ones ($P = 0.0012$, Wilcoxon Rank Sum test). (E) Average PhastCons scores of non-*Alu* overlapping skipped exons (described in (C)) and their flanking intronic regions. Exon-intron boundaries are positioned at 0. The shaded area represents standard errors.

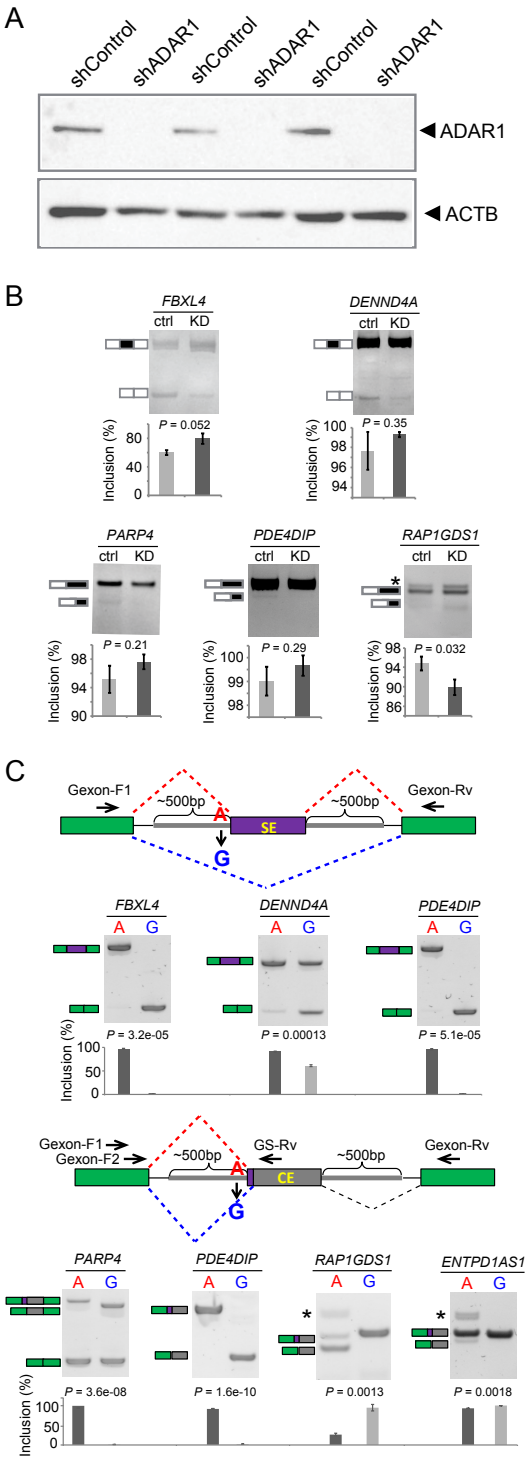


Figure 3.4 Experimental validation of splicing changes induced by 3'ss editing. (A) Western blot of ADAR1 in U87MG cells stably expressing control shRNA (shControl) or *ADAR1* shRNA (shADAR1). Three biological replicates are shown. (B) RT-PCR results of endogenous splicing

levels of five randomly chosen exons with 3'ss editing. U87MG cells in (A) were used. Ctrl: shControl. KD: shADAR1. Exon inclusion levels (%) were calculated based on triplicated experiments. One example gel image is shown. Error bars represent standard deviation. *P* values were calculated by Student's *t*-test. (C) Seven randomly chosen editing-dependent alternative splicing events were validated in minigene systems. The minigene designs to validate exon skipping (top panel) and alternative splice site events (bottom panel) are shown respectively. Primers used to amplify the splicing products are shown as arrowheads. In the bottom diagram, Gexon-F1 and Gexon-Rv represent the forward and reverse primers used for *PARP4*. Gexon-F2 and GS-Rv refer to the forward and gene-specific reverse primers for the other three exons. Exon inclusion levels (%) were calculated based on triplicated experiments. One example gel image is shown. Error bars represent standard deviation. *P* values were calculated by Student's *t*-test.

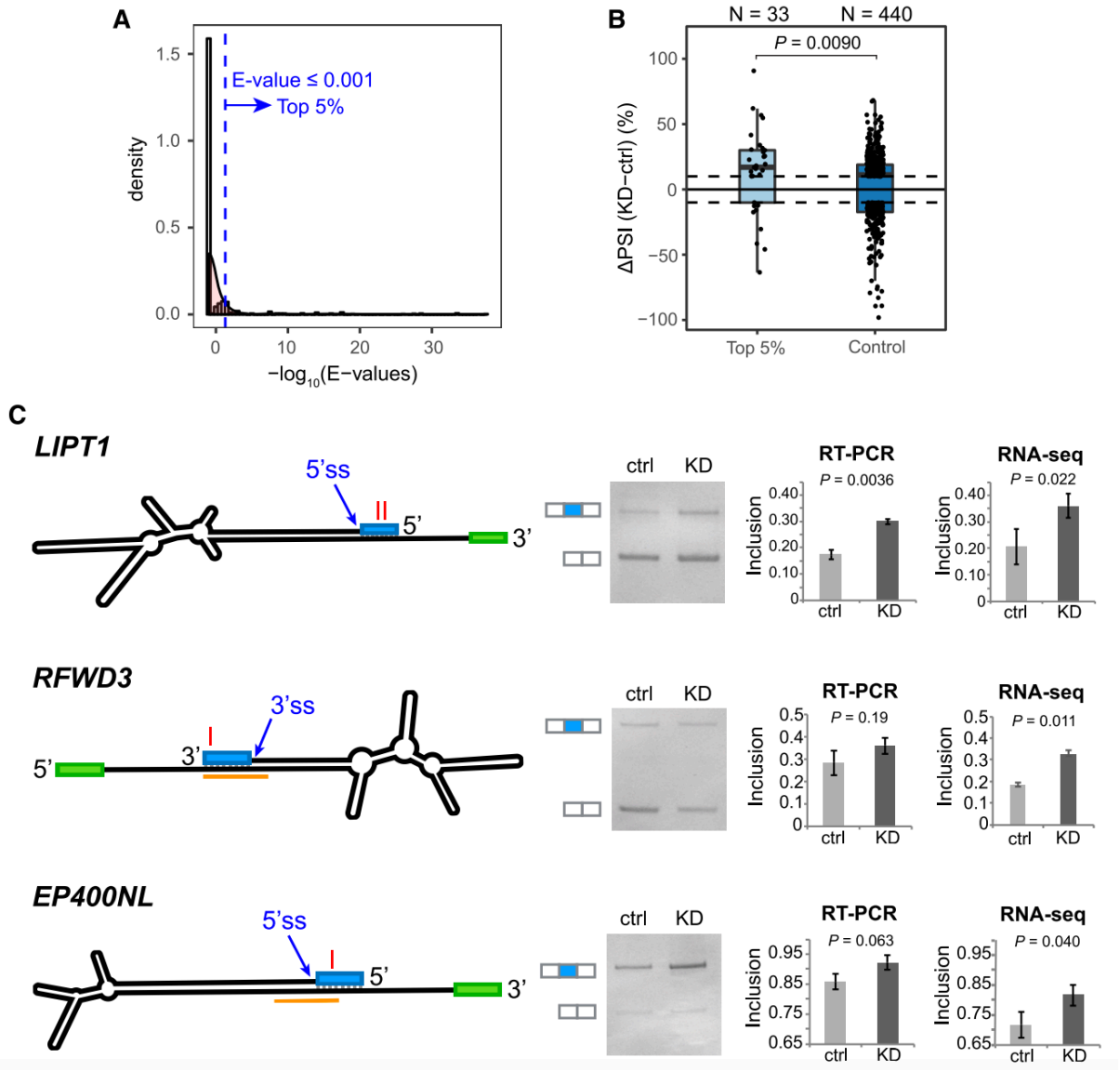


Figure 3.5 Regulation of splicing by *ADAR* acting on dsRNA structures. (A) E-value distribution of BLASTN alignments between exon and intron sequences, for exons with ≥ 10% PSI change upon *ADAR1* KD. A total of 33 exons (top 5% of all) had E-value ≤ 0.001 (blue dashed line). (B) Boxplot of PSI changes (ΔPSI) upon *ADAR1* KD, for the 33 exons in (A), and control exons with E-value >10 in the BLASTN analysis. Each dot represents an exon. The dashed lines mark ΔPSI = ±10%. The number of exons in each group is shown. $P = 0.0090$ comparing ΔPSI values of the two groups (Wilcoxon Rank Sum test). (C) Experimental testing

of endogenous splicing changes by RT-PCR in *ADAR1* KD and control U87MG cells. The mfold-predicted RNA secondary structures are depicted (not to scale). Blue box: candidate exon under regulation. Green box: a neighboring exon. ADAR1 CLIP peaks are highlighted in orange. The red letter "I" denotes the position of editing sites. Example gel images are shown using RT-PCR products generated by primers on the two flanking exons (white boxes next to the gel images). Mean and standard deviation (SD) of exon inclusion levels based on three biological replicates of RT-PCR are shown. Mean and SD of PSI values derived from two biological replicates of RNA-seq are also shown. *P* values were calculated by Student's *t*-test.

3.8 References

Agranat L, Sperling J, Sperling R. 2010. A novel tissue-specific alternatively spliced form of the A-to-I RNA editing enzyme ADAR2. *RNA Biol* **7**: 253–62.

Agrawal R, Stormo GD. 2005. Editing efficiency of a Drosophila gene correlates with a distant splice site selection. *RNA* **11**: 563–6.

Ahn J, Xiao X. 2015. RASER: Reads Aligner for SNPs and Editing sites of RNA. *Bioinformatics* **31**: 3906–3913.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* **2**: e391.

Bahn JH, Ahn J, Lin X, Zhang Q, Lee J-H, Civelek M, Xiao X. 2015. Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nat Commun* **6**: 6355.

Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–50.

Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, et al. 2014. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24**: 365–76.

Beghini A, Ripamonti CB, Peterlongo P, Roversi G, Cairoli R, Morra E, Larizza L. 2000. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Hum Mol Genet* **9**: 2297–2304.

Bentley DL. 2014. Coupling mRNA processing with transcription in time and space. *Nat Rev*

Genet **15**: 163–75.

Bhatt DM, Pandya-Jones A, Tong A-J, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST. 2012. Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions. *Cell* **150**: 279–290.

Blow M, Futreal PA, Wooster R, Stratton MR, Futreal AP, Wooster R, Stratton MR. 2004. A survey of RNA editing in human brain. *Genome Res* **14**: 2379–2387.

Bratt E, Ohman M. 2003. Coordination of editing and splicing of glutamate receptor pre-mRNA. *RNA* **9**: 309–18.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Flomen R, Knight J, Sham P, Kerwin R, Makoff A. 2004. Evidence that RNA editing modulates splice site selection in the 5-HT2C receptor gene. *Nucleic Acids Res* **32**: 2113–2122.

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

Goldberg L, Abutbul-Amitai M, Paret G, Nevo-Caspi Y. 2017. Alternative Splicing of STAT3 Is Affected by RNA Editing. *DNA Cell Biol* dna.2016.3575.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N, Feldmeyer D, Sprengel R, Seeburg PH. 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**: 78–81.

Higuchi M, Single FN, Köhler M, Sommer B, Sprengel R, Seeburg PH. 1993. RNA editing of AMPA receptor subunit GluR-B: A base-paired intron-exon structure determines position and efficiency. *Cell* **75**: 1361–1370.

Hsiao Y-HE, Bahn JH, Lin X, Chan T-M, Wang R, Xiao X. 2016a. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome Res*.

Hsiao Y-HE, Cass AA, Bahn JH, Lin X, Xiao XX. 2016b. Global Approaches to Alternative Splicing and Its Regulation—Recent Advances and Open Questions. In *Transcriptomics and Gene Regulation* (ed. J. Wu), Vol. 9 of *Translational Bioinformatics*, pp. 37–71, Springer Netherlands, Dordrecht.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–15.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Laurencikiene J, Källman AM, Fong N, Bentley DL, Ohman M. 2006. RNA editing and alternative splicing: the importance of co-transcriptional coordination. *EMBO Rep* **7**: 303–7.

Lee J-H, Ang JK, Xiao X. 2013. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA* **19**: 725–32.

Lee J-H, Gao C, Peng G, Greer C, Ren S, Wang Y, Xiao X. 2011. Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ Res* **109**: 1332–41.

Lehmann KA, Bass BL. 2000. Double-Stranded RNA Adenosine Deaminases ADAR1 and ADAR2 Have Overlapping Specificities. *Biochemistry* **39**: 12875–12884.

Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G. 2007. RNA-editing-mediated exon evolution. *Genome Biol* **8**: R29.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* **22**: 1001–5.

Levanon EY, Hallegger M, Kinar Y, Shemesh R, Djinovic-Carugo K, Rechavi G, Jantsch MF, Eisenberg E. 2005. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res* **33**: 1162–1168.

Li G, Bahn JH, Lee J-H, Peng G, Chen Z, Nelson SF, Xiao X. 2012. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res* **40**: e104.

Licht K, Kapoor U, Mayrhofer E, Jantsch MF. 2016. Adenosine to Inosine editing frequency controlled by splicing efficiency. *Nucleic Acids Res*.

Martins RP, Fåhraeus R. 2017. A matter of maturity: The impact of pre-mRNA processing in gene expression and antigen presentation. *International Journal of Biochemistry and Cell Biology*, May 24.

Mazloomian A, Meyer IM. 2015. Genome-wide identification and characterization of tissue-specific RNA editing events in D. melanogaster and their potential role in regulating alternative splicing. *RNA Biol* **12**: 1391–401.

Mitchell SF, Parker R. 2014. Principles and Properties of Eukaryotic mRNPs. *Mol Cell* **54**: 547–558.

Möller-Krull M, Zemann A, Roos C, Brosius J, Schmitz J. 2008. Beyond DNA: RNA Editing and Steps Toward Alu Exonization in Primates. *J Mol Biol* **382**: 601–609.

Morse DP, Aruscavage PJ, Bass BL. 2002. RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. *Proc Natl Acad Sci U S A* **99**: 7906–7911.

Nishikura K. 2015. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* **17**: 83–96.

Nishikura K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**: 321–49.

Nostrand EL Van, Freese P, Pratt GA, Wang X, Wei X, Blue SM, Dominguez D, Cody NAL, Olson S, Sundararaman B, et al. 2017. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* 179648.

Penn AC, Balik A, Greger IH. 2013. Steric antisense inhibition of AMPA receptor Q/R editing reveals tight coupling to intronic editing sites and splicing. *Nucleic Acids Res* **41**: 1113–1123.

Picardi E, D'Erchia AM, Lo Giudice C, Pesole G, Giudice C Lo, Pesole G. 2017. REDIportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res* **45**: D750–D757.

Pinto Y, Cohen HY, Levanon EY. 2014. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome Biol* **15**: R5.

Porath HT, Carmi S, Levanon EY. 2014. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat Commun* **5**: 4726.

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2014. 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res* **42**: D903-

9.

Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, Hacohen N, Schier AF, Blackshear PJ, Friedman N, et al. 2014. High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. *Cell* **159**: 1698–1710.

Raitskin O, Cho D-SS, Sperling J, Nishikura K, Sperling R. 2001. RNA editing activity is associated with splicing factors in lnRNP particles: The nuclear pre-mRNA processing machinery. *Proc Natl Acad Sci U S A* **98**: 6571–6.

Ramaswami G, Li JB. 2014. RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* **42**: D109–D113.

Reenan RA, Hanrahan CJ, Ganetzky B. 2000. The mle(napts) RNA helicase mutation in drosophila results in a splicing catastrophe of the para Na+ channel transcript in a region of RNA editing. *Neuron* **25**: 139–149.

Rieder LE, Reenan RA. 2012. The intricate relationship between RNA structure, editing, and splicing. *Semin Cell Dev Biol* **23**: 281–288.

Rieder LE, Staber CJ, Hoopengardner B, Reenan RA. 2013. Tertiary structural elements determine the extent and specificity of messenger RNA editing. *Nat Commun* **4**: 2232.

Rodriguez J, Menet JS, Rosbash M. 2012. Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. *Mol Cell* **47**: 27–37.

Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399**: 75–80.

Ryman K, Fong N, Bratt E, Bentley DL, Ohman M. 2007. The C-terminal domain of RNA Pol II helps ensure that editing precedes splicing of the GluR-B transcript. *RNA* **13**: 1071–8.

Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. 2015. Alternative Splicing

Signatures in RNA-seq Data: Percent Spliced in (PSI). *Curr Protoc Hum Genet* **87**: 11.16.1-11.16.14.

Schoft VK, Schopoff S, Jantsch MF. 2007. Regulation of glutamate receptor B pre-mRNA splicing by RNA editing. *Nucleic Acids Res* **35**: 3723–32.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–50.

Solomon O, Oren S, Safran M, Deshet-Unger N, Akiva P, Jacob-Hirsch J, Cesarkas K, Kabesa R, Amariglio N, Unger R, et al. 2013. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *RNA* **19**: 591–604.

St Laurent G, Tackett MR, Nechkin S, Shtokalo D, Antonets D, Savva YA, Maloney R, Kapranov P, Lawrence CE, Reenan RA. 2013. Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. *Nat Struct Mol Biol* **advance on**: 1333–1339.

Tariq A, Garncarz W, Handl C, Balik A, Pusch O, Jantsch MF. 2013. RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation. *Nucleic Acids Res* **41**: 2581–2593.

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**: 1616–25.

Vacic V, Iakoucheva LM, Radivojac P. 2006. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**: 1536–7.

Wang ET, Ward AJ, Cherone JM, Giudice J, Wang TT, Treacy DJ, Lambert NJ, Freese P, Saxena T, Cooper TA, et al. 2015. Antagonistic Regulation of mRNA Expression and Splicing by CELF and MBNL Proteins. *Genome Res* **25**: 858–71.

Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, et al. 2014. RNA-DNA Differences Are Generated in Human Cells within Seconds after RNA Exits Polymerase II. *Cell Rep* **6**: 906–915.

Washburn MC, Hundley HA. 2016. Controlling the Editor: The Many Roles of RNA-Binding Proteins in Regulating A-to-I RNA Editing. *Adv Exp Med Biol* **907**: 189–213.

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. 2009. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat Struct Mol Biol* **16**: 1094–100.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44**: D710–D716.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377–94.

Zhang Q, Xiao X. 2015. Genome sequence-independent identification of RNA editing sites. *Nat Methods* **12**: 347–350.

Zuker M, Mathews DH, Turner DH. 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In *RNA Biochemistry and Biotechnology*, pp. 11–43, Springer Netherlands, Dordrecht.

CHAPTER 4

Landscape of Genetic Modulation of Alternative Splicing and Its Causality in Human Tissues

4.1 Introduction

Alternative splicing (AS) is a universal RNA processing step that is conserved in eukaryotes. It is controlled by the interactions between *trans*-factors and *cis*-regulatory elements and plays an important role in regulating species complexity, phenotype, and disease (Hsiao et al. 2016b). Previous high-throughput studies have estimated more than 90% of human genes undergo AS (Pan et al. 2008; Wang et al. 2008) and suggested that almost 50% of the RNA isoforms are differentially expressed in at least one tissue (Castle et al. 2008). These findings indicate that tissue-specific alternative splicing is a widespread process. However, the key regulators of such process were yet to be discovered. AS exons with different inclusion rates across tissues are referred to as tissue-specific exons. Mouse tissue microarray analysis showed that the length of tissue-specific exons are usually in multiples of three (Xing and Lee 2005), suggesting that their differential inclusion often preserves the open reading frame, and such exons could impact the protein levels. In addition, many tissue-specific exons have essential roles in governing organ development and maintaining tissue and cell type identity (Baralle and Giudice 2017; Witten and Ule 2011; Kalsotra and Cooper 2011). For example, a large-scale study using splicing microarrays discovered a postnatal switch of AS reprogramming during murine heart development, and this phenomenon was highly prevalent and conserved between mammalian

and avian species (Kalsotra et al. 2008). Despite the significance of tissue-specific AS, its regulation and physiological implications are still open questions.

Since the genetic background of a healthy individual is identical for all different tissue types, tissue-specific AS is more often explained by tissue-specific expression of the splicing factors and their mRNA regulation (Grosso et al. 2008; Castle et al. 2008; David and Manley 2008; Ohno et al. 2008). For example, brain and testis are the two tissues with the highest levels of AS events and the largest number of differentially expressed splicing factor genes (Grosso et al. 2008; Baralle and Giudice 2017). Splicing patterns in the brain are complex and conserved, and aberrant AS in the brain could lead to neurological disorders (Quesnel-Vallières et al. 2016). Several studies have identified numerous tissues-specific AS events and their responsible splicing factors by correlating the exon inclusion level with the gene expression of splicing factors. In addition, it has been shown that constitutive splicing factors such as SR proteins could act as tissue-specific AS regulators in murine heart (Xu et al. 2005; Ding et al. 2004). It is conceivable that genetic variants may affect splicing in a tissue-specific manner by modulating the binding or gene expression of the tissue-specific *trans*-factors. However, few studies have considered the impact of genetic variants in the context of tissue-specificity and compared the differences in genetic regulation of splicing across multiple tissues. The first study that focused on the genetic control of tissue-specific splicing used an splicing quantitative trait locus (sQTL) approach to correlate genotype with exon-level microarray data to identify splicing-associated SNPs in brain and blood (Heinzen et al. 2008). However, although sQTL is a straightforward method to detect variants with splicing relevance, it is not the most accurate or computationally efficient method because it requires large datasets to reach adequate statistical power for genetic

association inference (Hsiao et al. 2016b). After the launch of the Genotype-Tissue Expression (GTEx) project, large number of genotype and tissue-specific RNA-seq datasets from a wide panel of human subjects became available and facilitated large-scale analyses on tissue-specific control of gene expression and splicing. In the GTEx pilot study, sQTL analysis was performed for nine tissues, and a high degree of tissue sharing for sQTLs was observed. However, the study did not go in depth to incorporate the information from splicing factors, and thus leaving the mechanisms of how the predicted SNPs mediate tissue-specific AS and the physiological relevance largely unknown. In addition, it is also challenging to derive a clear causal relationship between the genetic variant and the splicing phenotypes using sQTLs (Hsiao et al. 2016b).

In this study, we aim to investigate genetic modulation of alternative splicing (GMAS) across large panel of human tissues via analyses of allele-specific expression (ASE) of the heterozygous SNPs in an AS region. When a splicing factor has differential binding preference at a SNP locus, we expect to observe a GMAS event. The identified GMAS SNPs could serve as tags for the causal SNPs and help pinpoint the splicing factors associated with the corresponding genetic regulation. These findings can drive future in-depth regulatory and mechanistic studies. This ASE approach has been used in previous studies (Li et al. 2012; Hsiao et al. 2016a) and demonstrated good performance in predicting high quality GMAS events. First, we took advantage of the large-scale genotype and tissue RNA-seq data from the GTEx project (Lonsdale et al. 2013) and identified >8,000 GMAS events. We observed that the GMAS exons are highly tissue-specific and individual-specific, and both tissue and individual variances had similar contribution to the allelic ratio bias observed at the GMAS SNPs. Functional analyses of these GMAS events revealed disease implications for the splicing-modulating variants, thus

encouraging future studies to understand the mechanisms potentially underlying these phenotypes. Since the GMAS SNPs were predicted based solely on the association between the allelic bias of a SNP and the splicing pattern of an alternative region, a predicted GMAS SNP was not guaranteed to be causal. Leveraging the datasets from a large population, we further developed a method to detect causal SNPs for GMAS events. More than 600 causal SNPs were predicted for over 300 GMAS exons. Further downstream analyses on RBP motifs and ENCODE eCLIP and RBP knockdown RNA-seq data systematically elucidated how genetic variants and splicing factors could coordinate to impact AS in different tissues.

4.2 Results

4.2.1 Overview of genetic modulation of alternative splicing events in GTEx data

We analyzed 7822 RNA-seq samples in total, across 47 tissues and 515 donors from the GTEx project (v6p release). We excluded samples with fewer than 25 million uniquely mapped read pairs by HISAT2 (Kim et al. 2015). Around 90% of the samples satisfied this criterion and were kept for further analyses. We excluded eight tissues that had fewer than 50 samples with genotype information. For each sample, we collected a list of individual-specific heterozygous SNPs by combining the GTEx genotype information and expressed allelic counts in the RNA-seq reads (Method). Subsequently, we identified genetically modulated alternative splicing (GMAS) events using our previously published method (Li et al. 2012). This effort identified a total of 8286 high-quality GMAS events, involving 7404 SNPs and 4941 exons.

On average, we identified 10% of all testable exons exhibiting GMAS pattern in a given tissue (Fig. 1A, Methods). GMAS exons were most abundant in whole blood (16%), which is

consistent with the GTEx pilot study that showed that whole blood had the highest number of sQTLs among nine tissues analyzed in that study (Ardlie et al. 2015). Pancreas, in which splicing regulation is not well understood (Baralle and Giudice 2017), had the least fraction of exons demonstrating GMAS pattern (8%). This lower percentage could be partly explained by the low sample read depths relative to other tissues (Supplemental Fig. S1). Around 80% of the GMAS were skipped exon (SE) events, and this pattern was consistent in all the tissues we analyzed (Supplemental Fig. S2A). The global distribution of percent spliced-in (PSI) of GMAS events was similar for different tissues (Supplemental Fig. S2B). The PSI of AS exons (SE, alternative 5' splice site (A5SS), and alternative 3' splice site (A3SS)) showed bimodal patterns, consistent with previous findings for AS exons in general (Shalek et al. 2013; Song et al. 2017). In addition, SE events were more included than A5SS and A3SS exons. Retained intron (RI) events, which comprised around 10% of all GMAS events, had lower PSI than AS exons in general across tissues.

4.2.2 GMAS variation across tissues and individuals

A previous GTEx study reported that both gene expression and splicing contributed more to tissue variability than to individual variability (Melé et al. 2015). To investigate the tissue variability vs. individual variability attributable to the GMAS phenomenon, we used a linear mixed model similar to the previously published approach (Melé et al. 2015) to model the allelic imbalance of GMAS events by tissues and individuals while taking age, ethnicity, and gender into account (Methods). In contrast to the results from gene expression and splicing, we did not see a significant difference in the variability from tissues vs. from individuals for the GMAS

exons (Wilcoxon rank sum test, $P = 0.29$) (Fig. 1B), suggesting that genetic background and tissue-specificity had comparable effects on the occurrence of GMAS pattern of an exon.

To investigate the types of genes that characterize tissue and individual variability of GMAS events, we examined genes containing GMAS exons showing high tissue variance (tissue variance contributing ≥ 50% of GMAS variability), high individual variance (individual variance contributing ≥ 50% of GMAS variability), or low variance (tissue variance and individual variance each contributing ≤ 10% of GMAS variability) (Fig. 1B, scatter plot). We performed Gene Ontology (GO) analyses, as described previously (Hsiao et al. 2016a). Genes with high tissue variance GMAS exons were enriched for GO terms associated with sarcomere organization, cardiac muscle development, and cytoskeleton organization, all of which were related to heart or skeletal muscle functions (Fig. 1B, green text). This finding is consistent with the literature, as extensive tissue-specific alternative splicing occurs and contributes to increased protein diversity in striated muscle (heart and skeletal muscle) (Baralle and Giudice 2017). Genes hosting high individual variance GMAS exons were often related to immune response and signaling pathways (Fig. 1B, magenta text). This GO enrichment for genes with high individual variability is reasonable, as individuals vary dramatically in how they react to stress and pathogens (Schirmer et al. 2016; Ter Horst et al. 2016; Aguirre-Gamboa et al. 2016). For genes with low GMAS exon variability, the most significant GO terms comprised positive or negative regulation of various essential processes, including immunity, cell movement, and RNA metabolism (Supplemental Fig. S3). The absence of a clear overarching theme for these significant GO terms may suggest that these functions were universally essential across tissues in the population.

4.2.3 Comparative analysis of GMAS patterns across tissues

Leveraging the large number of tissues available from the GTEx project, we were able to explore the tissue-specificity of GMAS exons. We first investigated the profile of tissue-sharing of GMAS exons by calculating the Jaccard index for each possible pair of tissues in this study (Methods). We observed that related tissues, such as brain regions, heart and skeletal muscles, skin tissues, and reproductive tissues (ovary and vagina), formed clear clusters and exhibited higher Jaccard index values, suggesting a higher degree of similarity among these tissues (Fig. 2A). Most brain regions shared about 25-43% of GMAS exons with one another, with the exception of cerebellum and cerebellar hemisphere, which were previously reported as outliers with distinctly higher splicing factor expression (Melé et al. 2015). The difference in splicing factor expression could also explain the observation that cerebellum and cerebellar hemisphere shared the most GMAS exons with each other, and that only a few of the shared GMAS exons were found in other brain tissues with low prevalence (Supplemental Fig. S4).

To evaluate tissue sharing across many tissues, we focused on 26 representative tissues, determined by highest percentages of GMAS exons, and we required exons to be testable in at least 10 out of the 26 tissues in a single individual. We calculated a GMAS frequency, defined as the fraction of tissues demonstrating GMAS pattern among all testable tissues from an individual (Methods). We observed that most GMAS exons occurred only in fewer than 25% of the tissues in an individual. To ensure this observation did not arise from the stringent cutoffs used when calling a GMAS event, we constructed a tissue-by-individual matrix for each exon, where each element in the matrix represented whether the exon was GMAS, testable but not GMAS, or non-

testable in the corresponding tissue and individual. To generate randomized datasets as negative controls, we randomized the labels of both tissues and individuals in this matrix for each exon and calculated GMAS frequency. In contrast to random controls, GMAS exons were highly tissue-specific (Fig. 2B). Comparing the GMAS frequency of an exon between individuals, we observed a trend of higher standard deviation of GMAS frequency than in random controls (Fig. 2C). This finding suggested that the difference in genetic background is another player contributing to the tissue diversity of GMAS exons, in line with the findings mentioned earlier (Fig. 1B).

4.2.4 Comparative analysis of GMAS patterns across individuals

To assess the prevalence of each GMAS exon among individuals in a tissue, we calculated the fraction of testable individuals who had the exon showing GMAS signature in each tissue (Methods). Our results showed that 75% of GMAS exons were observed in less than 50% of the individuals (Fig. 3A). We also observed that the prevalence of GMAS exons was usually lower than the random controls, which were simulated from permuted matrices, as in the GMAS frequency analysis (Fig. 3A). This pattern was consistent across all tissues, indicating that the GMAS exons were highly individual-specific. The vast majority of GMAS exons exhibited consistently low prevalence across different tissues (Fig. 1B), while a smaller fraction of GMAS exons with mid-range prevalence showed a higher standard deviation among different tissues than the random controls, suggesting that they could be mediated in a tissue-specific manner in addition to genetic modulation.

The diverse genetic background in the GTEx population and the complex splicing regulatory networks could explain the observation of high individual-specificity of GMAS exons. In addition, the predicted GMAS SNPs were not necessarily causal because the method ensures only that the GMAS SNPs were linked with the real causal SNPs. Therefore, individuals with the same tag SNPs did not all demonstrate the same splicing outcome. If an individual had homozygous alleles at the causal SNP position, there would be no splicing difference and thus no allelic bias at the tag SNPs in the affected exon. This caveat highlighted the importance of causal SNP detection for GMAS events, which will be discussed in a later section.

4.2.5 Rare GMAS SNPs exhibiting higher degreed of disease relevance

Many large-scale efforts have been devoted to understanding the functional relevance of SNPs in the human genome. To date, the genome-wide association study (GWAS) catalog has documented hundreds of thousands of phenotype-associated SNPs from over 3000 publications (MacArthur et al. 2017). Yet, many traits were found to associate with non-coding or intergenic SNPs that do not alter the protein sequences, which makes GWAS interpretation challenging. Recent population-level transcriptional studies have found many splicing variants and sQTLs in linkage disequilibrium (LD) with GWAS SNPs, shedding light to the molecular mechanisms underlying GWAS observations. A previous study also suggested the importance of broad tissue sampling for better GWAS interpretations (GTEx Consortium 2017).

Here, our large compendium of GMAS events could facilitate a better understanding of the impact of splicing-modulating SNPs on GWAS traits. We observed that 3206 (43%) of the GMAS SNPs were in LD with GWAS SNPs (and within 200kb away, Fig. 4B). Around 28% of

these GMAS SNPs were in the same genes as the GWAS SNPs (Fig. 4A). Of the GWAS SNPs that are in LD with GMAS, the vast majority (57%) were located in introns, alluding to the regulatory role of such seemingly non-functional intronic variants to affect phenotypic traits through splicing. We found that 6% and 3% of all GMAS-related GWAS SNPs were annotated to cause missense and synonymous changes, respectively. By GMAS analysis, the molecular mechanism underlying these GWAS observations was more likely to be splicing-related instead of introducing protein alterations. Furthermore, these synonymous variants could have important impact on splicing although the corresponding protein sequences remained the same.

Increasing lines of evidence have suggested that rare variants are a major contributor to genetic risk of complex diseases (Tennessen et al. 2012; Nelson et al. 2012; UK10K Consortium et al. 2015). To assess the abundance of rare variants among the GMAS SNPs, we focused on the subset of GMAS SNPs cataloged in the Exome Aggregation Consortium (ExAC) (Lek et al. 2016) and used global minor allelic frequency (MAF) of <0.1%, the same cutoff as previously proposed, to define rare variants. Overall, 759 (~10%) of the GMAS SNPs were rare, and over 60% of the rare GMAS SNPs were in LD with GWAS SNPs (Fig. 4B), a much greater percentage than that among all the GMAS SNPs (Fisher's exact test, $P < 2.2 \times 10^{-16}$). We further asked whether GMAS SNPs overlap disease-related SNPs from public databases (Forbes et al. 2017; Griffith et al. 2017; Landrum et al. 2014). Consistent with the previous results, we observed a higher percentage of diseased SNPs overlapping with the rare GMAS SNPs compared to that of all GMAS SNPs (Fig. 4C). Together, these findings support the inference that rare SNPs are more likely to be functionally relevant to phenotypes than common SNPs.

4.2.6 Rare GMAS exons are enriched with disease-related genes

Next, we aimed to investigate how the prevalence and tissue specificity of GMAS patterns relate to disease regulation. To this end, we extracted genes containing exons showing either ubiquitous or scarce GMAS patterns across individuals and tissues. We defined an exon as a ubiquitous GMAS exon if it exhibited the GMAS pattern in over 60% of all testable individuals, and additionally, 60% of the tissues showed the GMAS pattern for one individual. On the other hand, an exon was defined as a scarce GMAS exon if its GMAS pattern occurred in fewer than 10% of the testable individuals, and at most 10% of the tissues demonstrated the GMAS pattern for one individual. In total, we identified 60 ubiquitous GMAS exons and 1172 scarce GMAS exons. Of the genes hosting ubiquitous GMAS exons, only 6 (10%) and 8 (13%) were mutation intolerant genes (Lek et al. 2016) or genes with phenotype-causing mutations in OMIM database (Hamosh et al. 2000) (Fig. 4D). Of the genes containing scarce GMAS exons, a much higher fraction (233 genes, 24%) were mutation intolerant genes, and this fraction was significantly higher than those of genes with ubiquitous GMAS exons (Fisher's exact test, $P = 0.017$) (Fig. 4D). We also found 233 genes (24%) in OMIM database with scarce GMAS exons, but this fraction was not statistically significant compared to genes with ubiquitous GMAS exons (Fisher's exact test, $P = 0.06$). Nevertheless, genes with scarce GMAS exons demonstrated a trend of enrichment for disease implications. This observation is in parallel with the expectation that detrimental splicing variants that are widespread in various tissues tend to be removed by purifying selection, and thus, disease-related mutations are relatively scarce in the population. This finding was also observed in previous eQTL functional analyses on GTEx data (GTEx Consortium 2017).

4.2.7 Prediction of candidate causal SNPs for GMAS

By definition, GMAS SNPs were only tag SNPs for GMAS events. To pinpoint the candidate causal SNPs of GMAS events, we developed a new method that leverages the rich population data from the GTEx project. We employed the rationale that concordance should be observed between the genotype of a causal SNP and the splicing patterns of a GMAS exon across a large number of individuals (Fig. 5A, Methods). Briefly, we first calculated a concordance score (S_i) between a candidate SNP and the GMAS event in an individual by adapting equations proposed previously for SNPs regulating gene expression (Lappalainen et al. 2013). In contrast to that previous approach for gene expression, we used the allelic imbalance at the tag SNP as the representation of the splicing pattern of the GMAS exon and called putative causal SNPs for splicing in a statistically rigorous way (Methods). Specifically, for each candidate SNP, we constructed a distribution of S_i over all individuals with the candidate SNP and GMAS event. In theory, if the candidate SNP was causal and one allele completely abolished the exon inclusion, the S_i distribution would display a prominent peak close to $S_i = 1$. Depending on the splicing difference induced by the two alternating alleles, this peak could shift away from $S_i = 1$ (Supplemental Fig. S5A). We detected specific patterns in the S_i distributions to predict candidate causal SNPs (Method, Supplemental Fig. S5A). To determine whether a peak was significant, we randomly shuffled the S_i data points into evenly spaced bins to generate a background distribution for S_i and required the observed peak to be significantly greater than the random background after FDR correction (Methods). We tested all heterozygous SNPs (in the population, not necessarily within an individual) in an alternative exon plus 1kb flanking introns or in a retained intron as candidate SNPs for each GMAS event (Supplemental Fig. S5B).

To evaluate how many individuals are needed for our method, we simulated 100 hypothetical GMAS exons with causal SNPs and tested method performance given different fractions of the testing cohort having the heterozygous genotype at the causal SNP (Method). Our results showed that greater predictive power is achieved when given more individuals with the GMAS event. Another important factor that affects predictive power was the proportion of testing individuals who had heterozygous alleles at the candidate SNP, referred to as "heterozygous ratio" henceforth. The performance improved when there were more heterozygous individuals available for the tested case. To reach 75% predictive power, around 30 individuals with 80% heterozygous ratio at the SNP are required (Supplemental Fig. S6).

4.2.8 Candidate causal SNPs for GMAS in GTEx

Applying our novel method, we predicted candidate causal SNPs for each tissue. In addition, since only 2% of the testable SNPs had a heterozygous ratio above 80%, we pooled all the GTEx samples together to increase predictive power. We calculated the maximum S_i score per event for each individual among all tissues. The candidate causal SNPs predicted from all tissues combined could include SNPs that affect binding motifs of pervasive splicing factors. To control the number of false discoveries, we applied a stringent FDR cutoff of 5%. Altogether, 628 candidate causal SNPs were predicted for 317 GMAS events (Fig. 5B), 50 of which were intron retention cases. We found that on average, ~18% of the putative causal SNPs were sQTLs previously reported by GTEx (Ardlie et al. 2015). Since all internal exons were subject to GMAS prediction, it was reasonable to see a higher percentage of overlap (around 32%) with GTEx sQTLs predicted by Altrans, which had the ability to detect *de novo* associations (Supplemental Fig. S7A). Additionally, we overlapped our putative causal SNPs with the sQTLs predicted in

the GEUVADIS dataset from the 1000 Genomes Project (’t Hoen et al. 2013; Lappalainen et al. 2013) and found a high overlap of 46% (Supplemental Fig. S7B).

Over 65% of the putative causal SNPs were located inside alternative regions (either in exons or in introns) (Fig. 5C), and more than 58% of them were the exact GMAS tag SNPs. Of the putative causal SNPs, 15 (2.4%) were in the 5’ splice sites (5’ss), and 24 (3.8%) were in the 3’ss. We observed a larger alteration in splice site strength by the candidate causal SNPs at the 5’ss (Supplemental Fig. S8), consistent with the fact that the 5’ss has a less degenerate sequence motif compared to the 3’ss. In general, the putative causal SNPs were enriched near the splice sites, and this density declined as the distance to the exon-intron boundaries increased (Fig. 5D), further supporting the expectation that SNPs with regulatory prospective on splicing should reside in close proximity to the targeted exons.

Next, we revisited the prevalence of the subset of GMAS exons with putative causal SNPs. We observed that this subset of GMAS exons was more prevalent than all GMAS exons (with or without causal SNPs predicted), as we hypothesized (Supplemental Fig. S9A). Furthermore, we focused only on individuals with heterozygous alleles at the predicted candidate causal SNPs and assessed the prevalence of GMAS patterns among these heterozygous individuals. As expected, there was an even more profound trend of high GMAS prevalence for these exons (Supplemental Fig. S9A), which provided further evidence that GMAS exon prevalence depended on the genotype of the causal SNP. In theory, 100% of individuals with heterozygous alleles at the causal SNPs should demonstrate GMAS patterns for these target exons. There are a couple of reasons why we did not see the ideal pattern for each GMAS event.

First, the genetic background was different within the population, and there could be individual-specific splicing regulatory elements (SREs) involved in splicing regulation. Second, the expression of splicing factors could vary among individuals, and thus the splicing patterns could be individual-specific, even within the same tissue. Third, our predictions could include false positives. However, our putative causal SNPs showed significant enrichment near splice sites, demonstrating high fidelity of the predictions (Fig. 5D).

We also investigated the tissue-specificity of the GMAS exons with putative causal SNPs. The results showed that these GMAS exons had higher GMAS frequency (Supplemental Fig. S9B). Focusing on only the individuals with heterozygous alleles at the candidate causal SNPs, we observed a bimodal-like distribution of GMAS frequencies (Supplemental Fig. S9B), which reflected splicing regulation by tissue-specific as well as ubiquitously expressed splicing factors.

4.2.9 Experimental validation of candidate causal SNPs for GMAS

To validate our computational method that predicts candidate causal SNPs for GMAS events, we performed minigene reporter experiments using a splicing reporter from a previous study (Xiao et al. 2009). The experiments were carried out in HeLa cells following the protocol described in a previous paper (Hsiao et al. 2016a). Briefly, we created two versions of the minigene construct, one with the reference allele, and the other with the variant allele, at the candidate causal SNP locations. We expect to see a significant inclusion differences between the two alleles if the SNP is causal for GMAS. Using this system, we validated the three SE events (*PDE4DIP*, *MAP2K3*, and *UGT2B17*) and two RI events (*SEPT4* and *ATHL1*), all of which showed that the alternative

regions demonstrated allele-specific splicing depending on the alleles at the predicted causal SNPs (Fig. 5E). These results strongly support our method in predicting candidate causal SNPs for GMAS events.

4.2.10 RBPs regulating GMAS events by differential binding to candidate causal SNPs

To investigate the splicing factors that were responsible for regulating GMAS events through predicted causal SNPs, we carried out *de novo* motif analysis, following our previously published approach (Hsiao et al. 2016a). Briefly, for each candidate causal SNP, we generated all possible 7mer genomic sequences covering the causal SNP and created two versions for each sequence: one for the reference allele and the other for the variant allele at the candidate causal SNP position. As controls, we randomly selected 10,000 intronic SNPs that were at least 5kb away from exons on both sides and similarly generated 7mer sequences around these intronic SNPs. We scored each sequence by the position weight matrices (PWMs) of each known splicing factor (Han et al. 2013) from public databases (Cook et al. 2011; Ray et al. 2013). A splicing factor was called a putative regulator if its motif binding was dramatically altered by the candidate causal SNP relative to random controls. In total, we predicted 374 (60%) of the candidate causal SNPs to significantly alter RNA binding strength of at least one splicing factor, via creating or disrupting the binding motif, and 23 splicing factors were predicted to have binding affected by at least 10 candidate causal SNPs each (Fig. 6A). In addition, we expanded our analysis by taking advantage of the large ENCODE eCLIP datasets in two human cell lines, HepG2 and K562 (Van Nostrand et al. 2017), which provide evidence for *in vivo* binding of splicing factors. Taking the union of results from *de novo* motif analysis and causal SNPs

overlapping eCLIP peaks, we identified a total of 36 splicing factors whose RNA binding showed evidence of being affected by at least 10 candidate causal SNPs (Fig. 6A).

Around 43% of candidate causal SNPs were found in at least one eCLIP peak of a splicing factor (Fig. 6B). The remaining 57% of candidate causal SNPs likely did not overlap with eCLIP peaks because eCLIP experiments might not capture all *in vivo* binding events of a splicing factor, and the eCLIP data are from human cell lines, which could have *trans*-factor profiles different from human tissues. Nonetheless, we observed the most abundant eCLIP peak overlap, with more than 150 putative causal SNPs, for BUD13 (Fig. 6A), which suggested BUD13 is a strong regulator of GMAS events. Indeed, among the BUD13 eCLIP-seq reads at the putative causal SNPs, we observed a significant allelic bias (Fig. 6B), signifying that BUD13 has a strong preference for one of the alleles. Among the splicing factors with supporting data from both *de novo* motif and eCLIP datasets, SRSF1 was predicted to be the strongest regulator of GMAS events (Fig. 6A), and this observation is consistent with formerly reported findings (Hsiao et al. 2016a).

To assess the impact of predicted splicing factors on splicing, we analyzed ENCODE RBP knockdown (KD) RNA-seq (Van Nostrand et al. 2017) and calculated the differences in percent spliced-in (PSI) between KD and control samples for the corresponding target GMAS exons. To evaluate whether the PSI change (ΔPSI) of GMAS exons was larger than expected, we used all internal exons with sufficient read coverage for PSI calculation (Methods) as controls. The results showed that upon KD of the predicted splicing factors, the ΔPSI of the regulated GMAS exons was significantly higher than that of the controls (KS test, $P = 0.0038$) (Fig. 6C).

Thus, our findings altogether support the functional roles of these splicing factors in regulating GMAS patterns of their targets.

4.3 Discussion

Alternative splicing has immense influence on tissue development and cellular diversity, and yet, the contributors to splicing regulation remain elusive. To date, only a handful of genetic variants have been known to affect splicing outcomes. To fill in this gap, we conducted a large-scale study to understand genetic impact on splicing in a diverse panel of human tissues from the GTEx Consortium. In total, 8,286 high-quality GMAS events, involving 7,404 SNPs and 4,941 exons, were identified from 7,822 RNA-seq samples, across 47 tissues from 515 individuals. We found that more than 40% of GMAS SNPs were in LD with GWAS hits, and most of these GWAS SNPs were intronic with poorly understood functional relevance (Fig. 4A). Based on our results, the biological impact of these non-coding GWAS SNPs may now be explained by GMAS associations. Public access to personal genomic datasets for many individuals also facilitated analyses of rare variants. We observed higher enrichment of rare GMAS events in known diseased databases (Fig. 4B and 4C). These results underscore the greater disease relevance of rare GMAS SNPs and genes hosting exons with scarce GMAS patterns across individuals and tissues compared to ubiquitous ones. These findings echoed a recent study reporting that most serve splice-disrupting variants inducing are rare (Cheung et al. 2017).

In addition, this multi-dimensional dataset from GTEx allowed us to investigate the tissue-specificity and individual-specificity of GMAS events. Unlike previous results reported for gene expression and splicing (Melé et al. 2015), the proportion of GMAS variability

explained by individuals was comparable to that of tissues (Fig. 1B), indicating that personal genomes are a significant determining factor for GMAS events. We expected this observation, as we aimed to detect genetic influence on splicing.

When we surveyed the GMAS pattern of each individual, we noticed a trend of high tissue-specificity (Fig. 2). This observation could be explained by the regulation of tissue-specific splicing factors. Interestingly, when we compared GMAS tissue-specificity across individuals, we observed that for a subset of exons, the degree of GMAS tissue-specificity varied drastically, sometimes even more so than random controls (Fig. 2C). This finding suggested that the genetic background of an individual could affect the expression of tissue-specific splicing factors, as well as their subsequent regulation.

Indeed, when we looked at the prevalence of GMAS exons in the GTEx population, we found that a large majority of exons showed individual-specific GMAS patterns (Fig. 3). We reasoned that only individuals with heterozygous alleles at the causal SNP for splicing would induce the GMAS pattern of an exon (Supplemental Fig. S9A). Thus, we were motivated to develop a method to predict causal SNPs for GMAS based on this premise.

Exploiting the genotype information and GMAS events from the large GTEx population, we developed a novel method to computationally predict candidate causal SNPs that regulate splicing. Specifically, our method appraises the concordance between the allelic bias of a candidate SNP and the splicing pattern of an alternatively spliced region, as represented by the ASE signature of the tag GMAS SNP. The key factor that determines the performance of our

method is the "heterozygous ratio" of a candidate causal SNP among the testing cohort. Our method demonstrates high predictive power when many individuals have heterozygous alleles at the candidate SNP locus. Since the heterozygous ratios of the vast majority of SNPs were low in the GTEx cohort, we had limited predictive power to detect causal SNPs in this dataset. Nevertheless, we were able to predict over 600 candidate causal SNPs for over 300 GMAS exons, and the quality of our predictions was evident by the enrichment of SNP density near the splice site, a popular metric used to examine the splicing relevance of a SNP. This method can be applied to other datasets encompassing large populations, such as 1000 Genomes Project, to expand the repertoire of candidate causal SNPs for GMAS.

With candidate causal SNPs at hand, we investigated the mechanisms of how these SNPs regulate splicing levels. One obvious model is through altering RNA binding of splicing factors. To assess this model, we carried out a comprehensive motif search for splicing factors whose binding preference significantly changed in an allele-specific manner. In addition, we complemented the motif results with eCLIP analysis using ENCODE data. Overlap between a candidate causal SNP and an eCLIP peak revealed in vivo protein binding to that SNP locus. Almost half of the candidate causal SNPs were bound by a splicing factor or shown to alter a binding motif. Some of the strongest GMAS regulators included BUD13, SRSF1, and PTBP1. Moreover, dramatic ΔPSI's were observed for the affected GMAS exons upon KD of the responsible splicing factors, further supporting our model. Moving forward, we plan to first carry out splicing reporter assays using minigene systems to validate the GMAS effects of the predicted causal SNPs. Next, we will perform Electrophoretic Mobility Shift Assay (EMSA) to validate allele-specific binding of BUD13, SRSF1, and PTBP1 to their respective targeted

candidate causal SNPs. Based on the results of these experiments altogether, we will evaluate the accuracy of our method to call candidate causal SNPs for GMAS and demonstrate the feasibility of the model that splicing factors regulate GMAS exons through differential binding to candidate causal SNPs.

4.4 Methods

4.4.1 Preprocessing of GTEx RNA-seq data and identification of GMAS events

FASTQ files from individuals with genotype information (from whole genome sequencing, whole exome sequence, or Illumina SNP Arrays) were downloaded from the GTEx database (v6p release). The adaptors were trimmed by Cutadapt (Martin 2011). We aligned the reads to the hg19 genome and transcriptome using HISAT2 (Kim et al. 2015) with parameters --mp 6,4 --no-softclip --no-mixed --no-discordant, keeping only the uniquely mapped read pairs for the following analyses. Samples with fewer than 25 million uniquely aligned read pairs were considered as insufficient read coverage for detecting GMAS events and thus discarded. We focused on the tissues that have at least 50 samples with sufficient read coverage. In total, 7822 RNA-seq samples across 47 tissues from 515 donors were kept for the GMAS analysis.

We collected a list of high-quality SNPs from whole genome sequencing (quality filter: GQ $\geq$ 20), whole exome sequence (quality filter: GQ $\geq$ 20), and Illumina SNP Arrays (quality filter: IGC $\geq$ 0.2) provided by GTEx. In addition to the genotyped SNPs, we included all dbSNPs (version 146) that showed evidence of being heterozygous in at least one GTEx individual as potential candidates for the GMAS analysis. To determine which dbSNPs were heterozygous, we used the RNA-seq reads covering the candidate dbSNP position and defined the SNP to be

heterozygous if it had at least 3 reads for either alleles. Additionally, for the SNPs that passed the coverage filter, we further filtered out those with extreme allelic ratio (AR), i.e., AR < 0.1 or AR > 0.9, to avoid potential amplification biases or sequencing errors.

We applied our published pipeline (Li et al. 2012) to predict GMAS events using this list of confident SNPs (genotyped SNPs and filtered heterozygous dbSNPs with at least 20 read coverage) and the uniquely aligned reads with the following modifications. First, instead of focusing solely on annotated AS exons from GENCODE comprehensive annotation (v24lift37), we tested all internal exons for potential GMAS events. Second, we replaced the normalized expression value (NEV) by PSI calculated by the method described in (Schafer et al. 2015), only keeping exons with ≥ 15 total reads (inclusion reads + exclusion reads) or ≥ 2 exclusion reads. An exon is testable if it passes the read coverage requirements for PSI calculation and has a powerful non-ASE SNP in another exon of the same gene. A SNP is defined as powerful if it has ≥ 20 reads (Li et al. 2012). To avoid false positives, we only focused on GMAS events that were called in at least three samples out of the total 7822 samples we analyzed.

4.4.2 Estimation of tissue vs. individual contributions to GMAS pattern variations

We used the lmer function from the lme4 package in R to model the allelic imbalance for each GMAS exon as the following:

Allelic imbalance ~ (1|individual) + (1|tissue) + age + ethnicity + sex

The allelic imbalance was calculated as the absolute difference of allelic ratio to 0.5. The fixed effects (age, ethnicity, and sex) were chosen based on the previous literature (Melé et al. 2015). The allelic imbalance variations contributed by tissues and by individuals were estimated from the above model.

4.4.3 Tissue-specificity quantified by Jaccard index and GMAS frequency

We used Jaccard index to quantify the extent of sharing of GMAS pattern for an exon *e* between tissues *i* and *j* (s_{eij}). Specifically, we calculated the Jaccard index for each GMAS exon by $s_{eij} = \frac{N_{ei} \cap N_{ej}}{N_{ei} \cup N_{ej}}$, where N_{ei} and N_{ej} are the number of individuals with *e* showing GMAS pattern in tissues *i* and *j* respectively ($i \neq j$). To more accurately estimate a reliable value for s_{eij}, we required $N_{ei} \cup N_{ej} \geq 10$. The final GMAS pattern sharing between tissues *i* and *j* (s_{ij}) was calculated as $s_{ij} = \frac{\sum_{e=1}^{E} s_{eij}}{E}$, where *E* is the total number of exons with reliable s_{eij} for tissues *i* and *j*.

To quantify the extend of tissue-specificity of a GMAS exon, we looked at its GMAS frequency among the 26 representative tissues defined by the highest fraction of exons exhibiting GMAS patterns among similar tissues. The GMAS frequency of a GMAS exon *e* in individual *I* (f_e^I) was calculated by $f_e^I = \frac{GMAS_e^I}{tissues_e^I}$, where $GMAS_e^I$ is the number of tissues showing *e* as a GMAS exon in individual *I*, and $tissues_e^I$ is the total number of tissues with *e* as a testable exon for GMAS in individual *I*. Similar to the prevalence calculation, we required $tissues_e^I \geq 10$ to calculate a reliable GMAS frequency. The GMAS frequency is within (0,1], where values closer

to 0 means higher tissue-specificity, and the value of 1 means the GMAS pattern is widespread across tissues.

4.4.4 Prevalence of a GMAS exon

The prevalence of a GMAS exon *e* in tissue *T* (prv_e^T) was calculated by $prv_e^T = \frac{GMAS_e^T}{individuals_e^T}$, where $GMAS_e^T$ is the number of individuals having *e* as a GMAS exon in tissue *T*, and $individuals_e^T$ is the total number of individuals with *e* as a testable exon for GMAS in tissue *T*. Testable exons are those with sufficient read coverage for PSI calculation and a powerful control SNP in the same gene (Li et al. 2012). To more accurately estimate the prevalence, we required $individuals_e^T \geq 10$. The minimum value for $GMAS_e^T$ should be 1 (one individual with GMAS exon *e*), and prv_e^T is within (0, 1].

4.4.5 Enrichment of functional SNPs among scarce vs. ubiquitous GMAS SNPs

The scarce GMAS exons were those that occurred in fewer than 10% of the testable individuals and fewer than 10% of the tissues of an individual with the GMAS pattern. The ubiquitous GMAS exons were those that appeared in at least 60% of the testable individuals and 60% of the tissues of an individual with the GMAS pattern. We obtained lists of genes with phenotypic implications from OMIM (Hamosh et al. 2000) and ExAC (Lek et al. 2016) databases and assess how many of genes with the ubiquitous or scarce GMAS exons (GMAS genes) were in these databases. The enrichment was calculated as the fraction of GMAS genes in a database among all GMAS genes.

4.4.6 Prediction of candidate causal SNPs for GMAS

The basic rationale for our method is that a real causal SNP for GMAS should show concordant relationship between its genotype and the splicing pattern of the target exon across a large number of individuals. In the toy example illustrated in Figure 5A, we first calculated the allelic ratio of the tag SNP and its difference from 0.5, which is the expected allelic ratio for an unbiased SNP. For candidate SNPs with homozygous genotype, the allelic ratio of the tag SNP would be approximately 0.5, and candidate SNPs with heterozygous genotype would associate with tag SNP of allelic ratio either 0 or 1 depending on the haplotype combination of the individual. We used the previously proposed formula (Lappalainen et al. 2013) shown below to calculate the concordance score (S_i) for an event (candidate causal SNP and target exon pair) in individual *i* to measure the consistent pattern between the genotype and the splicing pattern.

$$S_i = \begin{cases} \dfrac{{d_i}^2}{0.5^2} & \text{if candidate is heterozygous} \\ 1 - \dfrac{{d_i}^2}{0.5^2} & \text{if candidate is homozygous} \end{cases}$$

We compiled a list of candidate causal SNPs for each individual. The processing of SNPs is described in the previous section. In addition, we made sure that the SNP genotype inferred from RNA-seq was supported by at least two tissues for further analysis. SNPs that are homozygous in all individuals we analyzed were filtered out. We used the GMAS events identified from each tissue as the tag SNPs and the target exons and calculated the concordance score of every possible candidate SNP-tag SNP pair (event) within a specific region shown in Supplemental Fig. S5B.

For each event, we generated a score distribution from the individuals who have the GMAS event. There are two major patterns of the distribution that indicate the candidate SNP as potentially causal. First, if the two alleles show 100% splicing difference, i.e., one allele completely abolishes the exon inclusion, we expect to see a major peak near the score of 1. On the other hand, if one allele only has partial impact on splicing, we would expect to observe two peaks: one at score of 1, which is from the homozygous individuals, and the other at scores > 0, which the exact value depends on the splicing difference the two alleles have, and this peak is contributed by the heterozygous individuals (Supplemental Fig. S5A). In both scenarios, the score of 0 comes from the homozygous individuals, and it indicates the candidate SNP is not causal. If there are more heterozygous individuals in the testing cohort, we anticipate seeing the dominant peaks at score > 0, determined by the allele-specific splicing change. We fitted a Gaussian Mixture Model (GMM) to the score distribution to identify a major peak whose mean was close to the score of 1 or away from 0 based on the *P* value calculated from z-scores. The number of GMM components fitted was determined by Bayesian information criterion (BIC). We kept the events that were significantly different from score of 0 after FDR correction (FDR ≤ 0.1). For a true splicing causal event, the S_i score should be consistent regardless of the genotype of the causal SNP. To avoid potential false positive due to small fraction of individuals with a different genotype at the candidate SNP, we carried out Fisher's exact test to evaluate whether a failed score always associated with the individuals with the different genotype. Events that showed significant FDR corrected *P* value ≤ 0.1 indicated the failed scores were linked to a particular genotype and thus were eliminated. To ensure the magnitude of the peak was significant, we binned the x-axis (S_i scores) into 100 bins and randomized the data points evenly

across the x-axis to generate a background distribution. This process was repeated for 500 times to estimate an average background level and its standard deviation. We compared the peak height to the background in the same bin and defined significant peaks by z-score > 2.58, which corresponds to $P < 0.01$.

4.4.7 Power analysis for predicting candidate causal SNPs for GMAS

To assess how many individuals our method needed to predict candidate causal SNPs for GMAS, we simulated 100 causal SNPs with two alternating alleles inducing 75% difference in PSI. This allelic-specific inclusion difference would be reflected in the allelic ratios. The total read counts of a SNP were simulated from a negative binomial distribution using parameters estimated from a real GTEx RNA-seq sample. We required all simulated SNPs to have at least 20 read coverage. The allelic ratios of the simulated SNPs were generated from a binomial distribution.

We simulated six groups of 200 individuals. Each group has a specific heterozygous frequency (Supplemental Fig. S6), which is defined as the fraction of individuals with heterozygous alleles at the candidate causal SNP position in a group. We ran our causal SNP prediction pipeline on these 100 SNPs providing different number of individuals while maintaining the heterozygous frequency for prediction.

4.4.8 Calculation for the PSI of GMAS exons

The PSI of GMAS exons were calculated as described in the previous section. Replicates of the ENCODE RBP KD RNA-seq were combined for this calculation. When calculating the PSI

differences between RBP KD and control samples, we required the event had sufficient read coverage as previously mentioned in both samples. We applied Fisher's exact test to evaluate the significance in PSI changes.

4.5 Acknowledgements

We thank members of the Xiao laboratory for helpful discussions and comments on this work. We thank the GTEx Consortium for generating the genotype and RNA-seq data. We thank the ENCODE Consortium for generating the eCLIP and RBP KD RNA-seq data. This work was supported in part by grants U01HG009417 and R01AG056476 to XX. YHH was supported by the Bioengineering supplemental fellowship of UCLA.

4.6 Figures

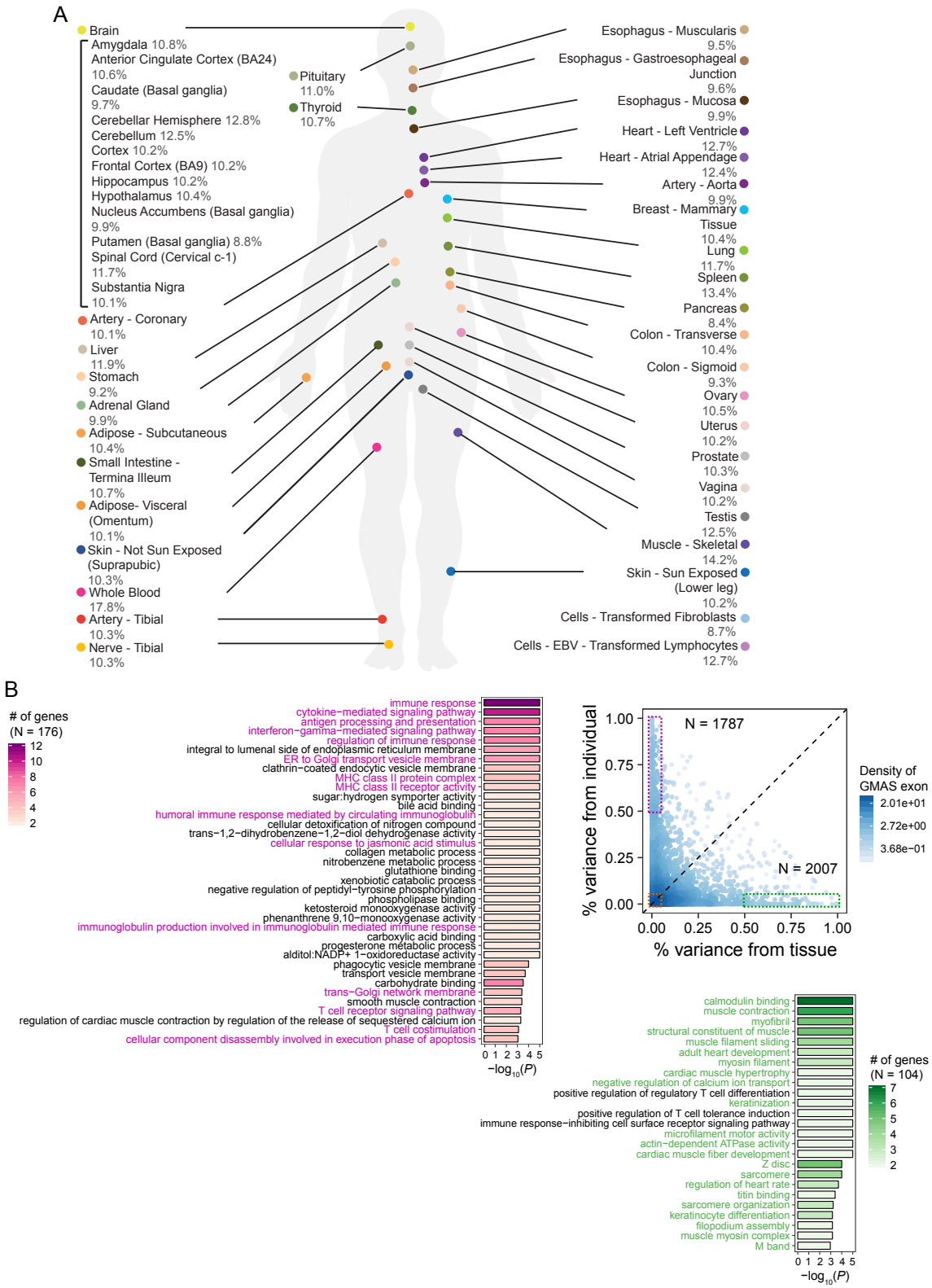

A
Brain
Amygdala 10.8%
Anterior Cingulate Cortex (BA24) 10.6%
Caudate (Basal ganglia) 9.7%
Cerebellar Hemisphere 12.8%
Cerebellum 12.5%
Cortex 10.2%
Frontal Cortex (BA9) 10.2%
Hippocampus 10.2%
Hypothalamus 10.4%
Nucleus Accumbens (Basal ganglia) 9.9%
Putamen (Basal ganglia) 8.8%
Spinal Cord (Cervical c-1) 11.7%
Substantia Nigra 10.1%
Artery - Coronary 10.1%
Liver 11.9%
Stomach 9.2%
Adrenal Gland 9.9%
Adipose - Subcutaneous 10.4%
Small Intestine - Termina Illeum 10.7%
Adipose- Visceral (Omentum) 10.1%
Skin - Not Sun Exposed (Suprapubic) 10.3%
Whole Blood 17.8%
Artery - Tibial 10.3%
Nerve - Tibial 10.3%
Pituitary 11.0%
Thyroid 10.7%
Esophagus - Muscularis 9.5%
Esophagus - Gastroesophageal Junction 9.6%
Esophagus - Mucosa 9.9%
Heart - Left Ventricle 12.7%
Heart - Atrial Appendage 12.4%
Artery - Aorta 9.9%
Breast - Mammary Tissue 10.4%
Lung 11.7%
Spleen 13.4%
Pancreas 8.4%
Colon - Transverse 10.4%
Colon - Sigmoid 9.3%
Ovary 10.5%
Uterus 10.2%
Prostate 10.3%
Vagina 10.2%
Testis 12.5%
Muscle - Skeletal 14.2%
Skin - Sun Exposed (Lower leg) 10.2%
Cells - Transformed Fibroblasts 8.7%
Cells - EBV - Transformed Lymphocytes 12.7%
B
# of genes (N = 176)
12
10
8
6
4
2
immune response
cytokine−mediated signaling pathway
antigen processing and presentation
interferon−gamma−mediated signaling pathway
regulation of immune response
integral to lumenal side of endoplasmic reticulum membrane
ER to Golgi transport vesicle membrane
clathrin−coated endocytic vesicle membrane
MHC class II protein complex
MHC class II receptor activity
sugar:hydrogen symporter activity
bile acid binding
humoral immune response mediated by circulating immunoglobulin
cellular detoxification of nitrogen compound
trans−1,2−dihydrobenzene−1,2−diol dehydrogenase activity
cellular response to jasmonic acid stimulus
collagen metabolic process
nitrobenzene metabolic process
glutathione binding
xenobiotic catabolic process
negative regulation of peptidyl−tyrosine phosphorylation
phospholipase binding
ketosteroid monooxygenase activity
phenanthrene 9,10−monooxygenase activity
immunoglobulin production involved in immunoglobulin mediated immune response
carboxylic acid binding
progesterone metabolic process
alditol:NADP+ 1−oxidoreductase activity
phagocytic vesicle membrane
transport vesicle membrane
carbohydrate binding
trans−Golgi network membrane
smooth muscle contraction
T cell receptor signaling pathway
regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion
T cell costimulation
cellular component disassembly involved in execution phase of apoptosis
0 1 2 3 4 5
−log10(P)
N = 1787
N = 2007
% variance from individual
% variance from tissue
0.00
0.25
0.50
0.75
1.00
Density of GMAS exon
2.01e+01
2.72e+00
3.68e−01
calmodulin binding
muscle contraction
myofibril
structural constituent of muscle
muscle filament sliding
adult heart development
myosin filament
cardiac muscle hypertrophy
negative regulation of calcium ion transport
positive regulation of regulatory T cell differentiation
keratinization
positive regulation of T cell tolerance induction
immune response−inhibiting cell surface receptor signaling pathway
microfilament motor activity
actin−dependent ATPase activity
cardiac muscle fiber development
Z disc
sarcomere
regulation of heart rate
titin binding
sarcomere organization
keratinocyte differentiation
filopodium assembly
muscle myosin complex
M band
0 1 2 3 4 5
−log10(P)
# of genes (N = 104)
7
6
5
4
3
2

Figure 4.1 Landscape of GMAS exons in human tissues. (A) The average fraction of GMAS exon among all testable exons in each tissue. (B) Scatter plot: contribution of tissue variability (Var^T) and individual variability (Var^I) to the GMAS pattern of an exon. Each dot represents an exon, and the colors represent the density of overlapping dots. The diagonal dashed line denotes $y = x$ (equal values for Var^T and Var^I). The N's below and above the diagonal line show the number of GMAS exons with $Var^T > Var^I$ and $Var^I > Var^T$, respectively. Green, magenta, and orange dotted boxes outline the GMAS exons with high Var^T and high Var^I, respectively. GO analysis results for each of the mentioned group is shown in separate bar plots with the same color for the corresponding group. The total number of genes inputted for the analyses are shown in the parentheses. Color intensity represents the number of genes associated with each significant GO term. The *P* values were estimated based on 10,000 randomizations controlling for gene length and GC content (Hsiao et al. 2016a). The empirical *P* values were corrected by Bonferroni corrections, and the significant cutoff was set to be 1/total GO terms considered.

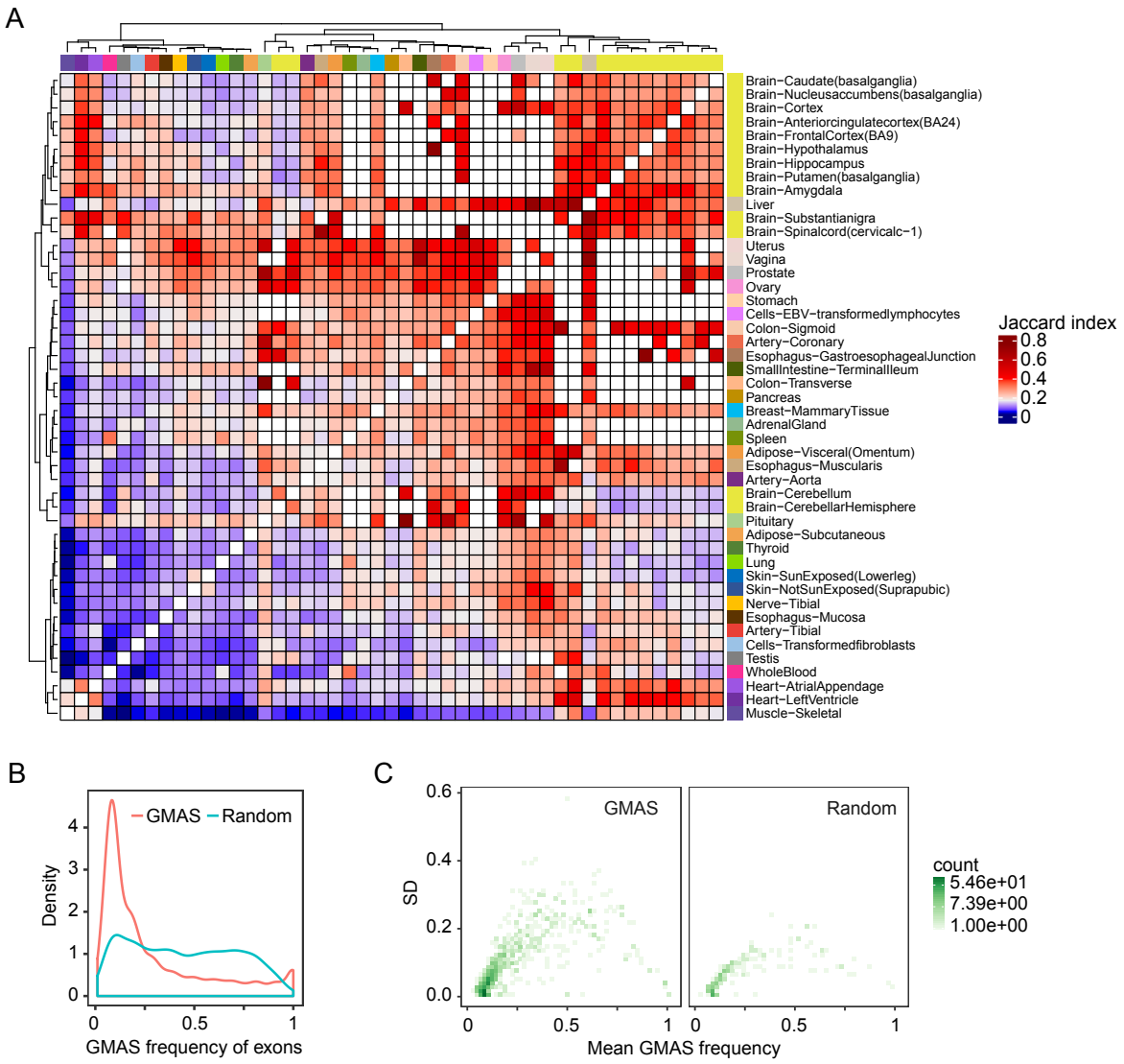


Figure 4.2 Tissue-specificity of GMAS exons in GTEx population. (A) Heatmap of Jaccard indices of GMAS exons between each pair of tissues. White boxes represent the tissue pairs with < 10 common testable exons. (B) GMAS frequency distribution of GMAS exons in an individual. One GMAS exon is plotted multiple times for different individuals with this exon. The green curve represents the GMAS frequency calculated from the randomized control matrices described in the main text. (C) Density plot of the mean of GMAS frequency vs. the standard deviation (SD) of GMAS frequency between individuals. The color intensity represents the count of GMAS exons.

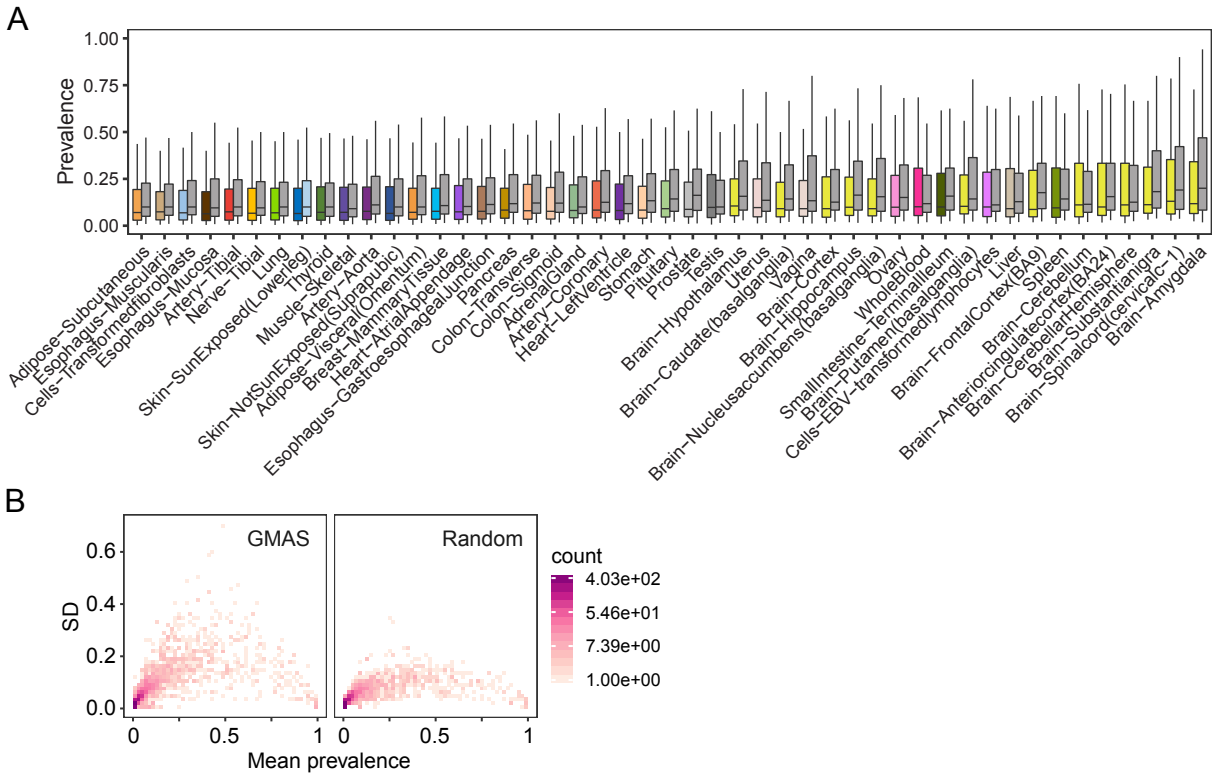


Figure 4.3 Prevalence of GMAS exons across human tissues. (A) Distribution of GMAS exon prevalence per tissue ranked by the 75th percentile. The colored boxes represent the GMAS exons, and the gray boxes are the randomized controls. (B) Density plot of the mean of prevalence vs. the standard deviation (SD) of prevalence between tissues. The color intensity represents the count of GMAS exons.

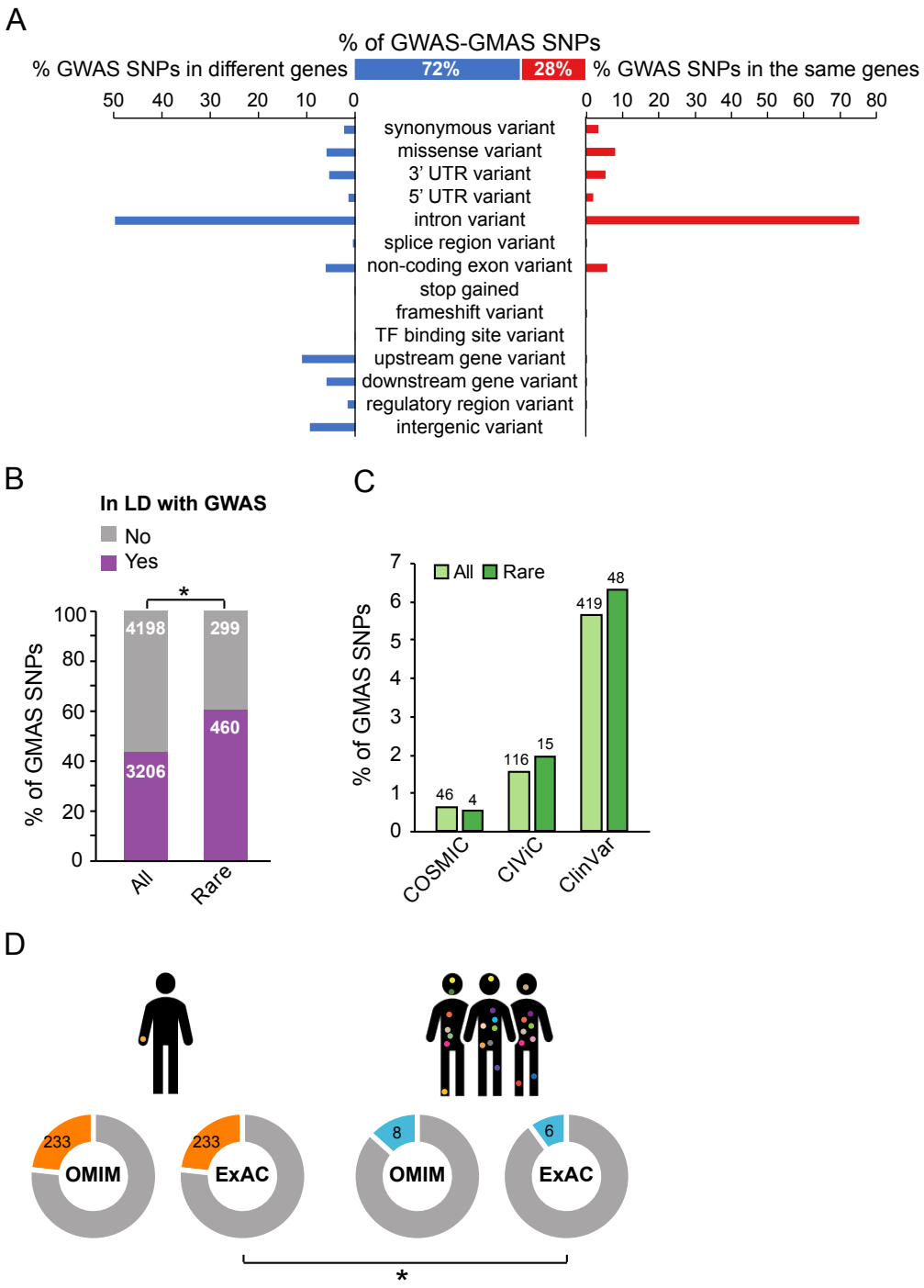
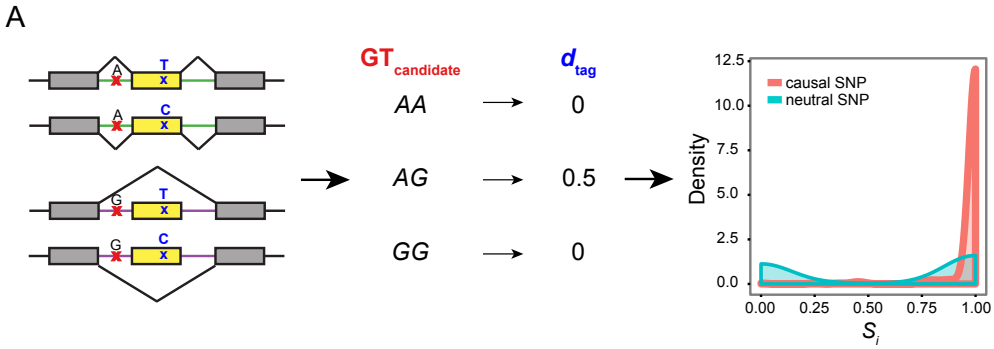
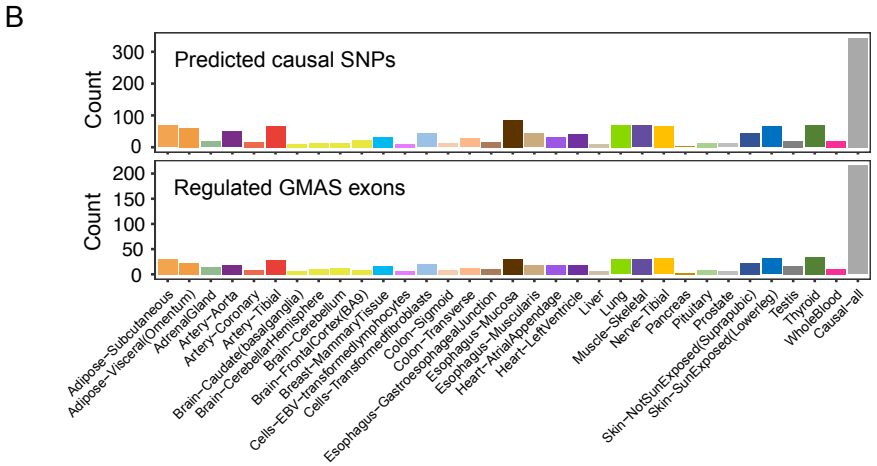
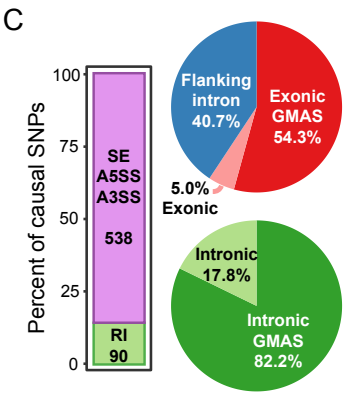


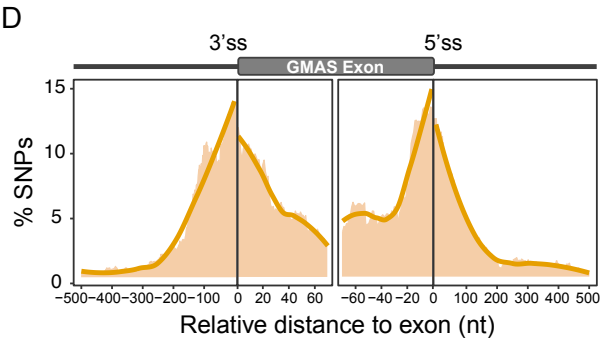
Figure 4.4 Functional implications of GMAS events. (A) The stacked bar illustrates the fraction of GMAS-GWAS SNPs in the same gene (red) or in different genes (blue) as the linked GWAS genes. The bar plots below demonstrate the context distribution of the GWAS SNPs that are in LD with GMAS SNPs in either group. (B) The fraction of all vs. rare GMAS SNPs that are

also GMAS-GWAS SNPs. *: Fisher's exact test, $P < 2.2 \times 10^{-16}$. (C) The fraction of all vs. rare GMAS SNPs that are found in each disease databases. (D) Enrichment of disease-associated genes in genes hosting exons with scarce (left) or ubiquitous (right) GMAS patterns. The colored portions of the ring charts represent the number of genes found in each database. The colorful dots on the people icons symbolize the tissues where GMAS patterns were observed. OMIM: genes with phenotype-causing mutations. ExAC: loss-of-function-intolerant gene sets (pLI > 0.9) from ExAC database. *: Fisher's exact test, $P < 0.05$.


A
GT candidate
d tag
AA
AG
GG
0
0.5
0
causal SNP
neutral SNP
Density
S i


B
Predicted causal SNPs
Regulated GMAS exons
Count
Adipose−Subcutaneous
Adipose−Visceral(Omentum)
AdrenalGland
Artery−Aorta
Artery−Coronary
Artery−Tibial
Brain−Caudate(basalganglia)
Brain−CerebellarHemisphere
Brain−Cerebellum
Brain−FrontalCortex(BA9)
Breast−MammaryTissue
Cells−EBV−transformedlymphocytes
Cells−Transformedfibroblasts
Colon−Sigmoid
Colon−Transverse
Esophagus−GastroesophagealJunction
Esophagus−Mucosa
Esophagus−Muscularis
HeartAtrialAppendage
Heart−LeftVentricle
Liver
Lung
Muscle−Skeletal
Nerve−Tibial
Pancreas
Pituitary
Prostate
Skin−NotSunExposed(Suprapubic)
Skin−SunExposed(Lowerleg)
Testis
Thyroid
WholeBlood
Causal−all


C
Percent of causal SNPs
SE
A5SS
A3SS
538
RI
90
Flanking intron 40.7%
Exonic GMAS 54.3%
5.0% Exonic
Intronic 17.8%
Intronic GMAS 82.2%


D
3'ss
5'ss
GMAS Exon
% SNPs
Relative distance to exon (nt)

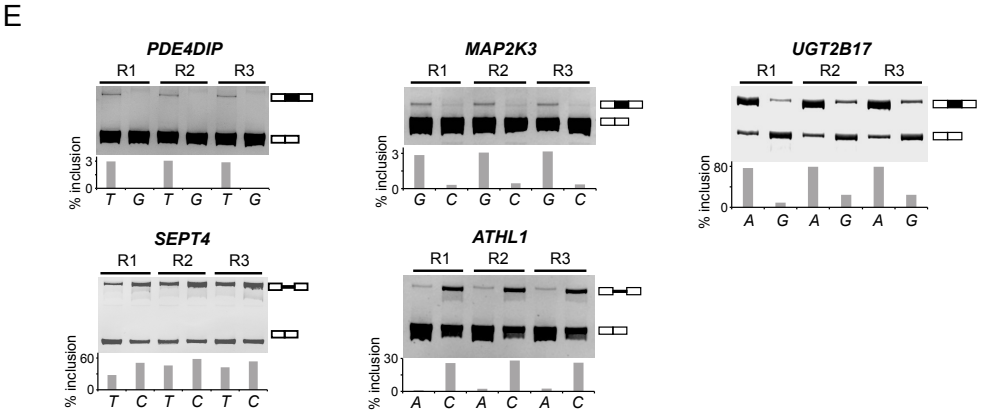

E
PDE4DIP
MAP2K3
UGT2B17
SEPT4
ATHL1
R1
R2
R3
% inclusion

Figure 4.5 Causal SNP predictions for GMAS events. (A) Diagram illustrating the methodology to predict causal SNPs for GMAS. In the toy example shown in the left, the yellow rectangles represent the alternative exon whose splicing is regulated by the upstream intronic SNP (A>G) shown in red. $GT_{candidate}$: The genotype of the candidate causal SNP, which is the red intronic SNP (A>G) in this case. The tag GMAS SNP here is the blue exonic SNP (T>C), and d_{tag} represents the allelic ratio imbalance of the tag SNP. On the right panel, we show the ideal S_i distribution for a causal SNP vs. a neutral SNP. In this example, since the candidate SNP is the causal one, we expect to see a pronounced peak at $S_i = 1$. (B) Top panel: number of predicted causal SNPs per tissue. Bottom panel: number of affected GMAS exons by the predicted causal SNPs per tissue. Causal-all: predictions made by pooling samples from all tissues together. (C) The types of GMAS event and the context distribution of predicted causal SNPs. Stacked bar: fraction of GMAS exons vs. GMAS introns. SE: skipped exon. A5SS: alternative 5' ss. A3SS: alternative 3'ss. Top pie: context distribution of the 538 GMAS exons. Bottom pie: context distribution of the 90 GMAS introns. Exonic GMAS: exonic causal SNPs that are also GMAS SNPs. Flanking: intronic causal SNPs in the flanking regions of the AS exons. Intronic GMAS: intronic causal SNPs that are also GMAS SNPs. (D) Density of the predicted SNPs near the exon-intron boundaries of their target GMAS exons. The number of causal SNPs was normalized by the total number of testable SNPs at each nucleotide position. Orange curve is the fitted trend line of the shaded orange area that represents the SNP density at single nucleotide resolution. (E) Minigene experiments validating predicted causal SNPs for GMAS in triplicates (Rep 1 to Rep 3). The inclusion rates were estimated from the band intensities of the PAGE gel.

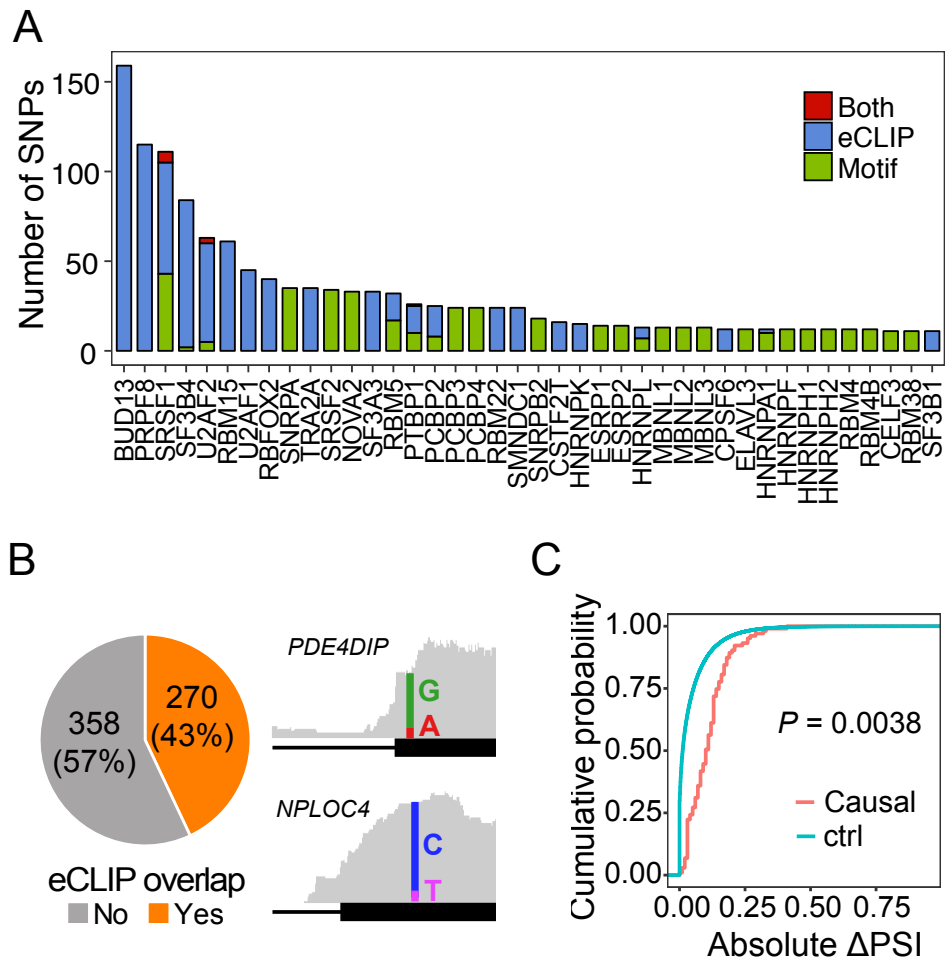


Figure 4.6 Splicing factors that regulate GMAS events through predicted causal SNPs. (A) Number of causal SNPs predicted to alter RNA binding motifs of splicing factors by *de novo* motif and/or eCLIP overlap analyses. Both: causal SNPs predicted by both methods. (B) Left: Proportion of putative causal SNPs that overlap with at least one eCLIP peak. Right: Two examples from BUD13 eCLIP data showing allele-specific binding preference at the causal SNPs. The read distribution shown in gray depicts the eCLIP peaks along the gene structure. The fraction of reads supporting either allele at the causal SNP position is labeled by the vertical stacked bars with different colors.

4.7 Reference

't Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, Laros JFJ, Buermans HPJ, Karlberg O, Brännvall M, et al. 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol* **31**: 1015–22.

Aguirre-Gamboa R, Joosten I, Urbano PCM, van der Molen RG, van Rijssen E, van Cranenbroek B, Oosting M, Smeekens S, Jaeger M, Zorro M, et al. 2016. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. *Cell Rep* **17**: 2474–2487.

Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-)* **348**: 648–660.

Baralle FE, Giudice J. 2017. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol* **18**: 437–451.

Castle JC, Zhang C, Shah JK, Kulkarni A V, Kalsotra A, Cooper TA, Johnson JM. 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet* **40**: 1416–25.

Cheung R, Insigne KD, Yao D, Burghard CP, Jones EM, Goodman DB, Kosuri S. 2017. Large-scale screening of rare genetic variants in humans reveals frequent splicing disruptions. *bioRxiv* 199927.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**: D301-8.

David CJ, Manley JL. 2008. The search for alternative splicing regulators: new approaches offer a path to a splicing code. *Genes Dev* **22**: 279–85.

Ding J-H, Xu X, Yang D, Chu P-H, Dalton ND, Ye Z, Yeakley JM, Cheng H, Xiao R-P, Ross J, et al. 2004. Dilated cardiomyopathy caused by tissue-specific ablation of SC35 in the heart. *EMBO J* **23**: 885–96.

Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. 2017. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**: D777–D783.

Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, et al. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* **49**: 170–174.

Grosso AR, Gomes AQ, Barbosa-Morais NL, Caldeira S, Thorne NP, Grech G, von Lindern M, Carmo-Fonseca M. 2008. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res* **36**: 4823–32.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213.

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* **15**: 57–61.

Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**: 241–5.

Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna K V, Gabriel WN, Welsh-Bohmer KA, Hulette CM, Denny TN, Goldstein DB. 2008. Tissue-specific genetic control of splicing: implications for the study of complex traits. ed. E. Liu. *PLoS Biol* **6**: e1.

Hsiao Y-HE, Bahn JH, Lin X, Chan T-M, Wang R, Xiao X. 2016a. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome Res* **26**: 440–50.

Hsiao Y-HE, Cass AA, Bahn JH, Lin X, Xiao XX. 2016b. Global Approaches to Alternative Splicing and Its Regulation—Recent Advances and Open Questions. In *Transcriptomics and Gene Regulation* (ed. J. Wu), Vol. 9 of *Translational Bioinformatics*, pp. 37–71, Springer Netherlands, Dordrecht.

Kalsotra A, Cooper TA. 2011. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet* **12**: 715–29.

Kalsotra A, Xiao X, Ward AJ, Castle JC, Johnson JM, Burge CB, Cooper TA. 2008. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad Sci* **105**: 20333–20338.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–60.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**: D980-5.

Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–11.

Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, O’Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Li G, Bahn JH, Lee J-HH, Peng G, Chen Z, Nelson SF, Xiao X. 2012. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res* **40**: e104.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–5.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896–D901.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10.

Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**: 660–5.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**: 100–4.

Ohno G, Hagiwara M, Kuroyanagi H. 2008. STAR family RNA-binding protein ASD-2 regulates developmental switching of mutually exclusive alternative splicing in vivo. *Genes Dev* **22**: 360–74.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–1415.

Quesnel-Vallières M, Dargaei Z, Irimia M, Gonatopoulos-Pournatzis T, Ip JY, Wu M, Sterne-Weiler T, Nakagawa S, Woodin MA, Blencowe BJ, et al. 2016. Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. *Mol Cell* **64**.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–7.

Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. 2015. Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). *Curr Protoc Hum Genet* **87**: 11.16.1-11.16.14.

Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, Ter Horst R, Jansen T, Jacobs L, Bonder MJ, et al. 2016. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**: 1125–1136.e8.

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**: 236–40.

Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, Yeo GW. 2017. Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Molecular Cell*, July 6.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64–9.

Ter Horst R, Jaeger M, Smeekens SP, Oosting M, Swertz MA, Li Y, Kumar V, Diavatopoulos

DA, Jansen AFM, Lemmers H, et al. 2016. Host and Environmental Factors Influencing Individual Human Cytokine Responses. *Cell* **167**: 1111–1124.e13.

UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JRB, Xu C, Futema M, et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82–90.

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Blue SM, Dominguez D, Cody NAL, Olson S, Sundararaman B, et al. 2017. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* 179648.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.

Witten JT, Ule J. 2011. Understanding splicing regulation through RNA splicing maps. *Trends Genet* **27**: 89–97.

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. 2009. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat Struct Mol Biol* **16**: 1094–100.

Xing Y, Lee CJ. 2005. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. ed. G. Gibson. *PLoS Genet* **1**: e34.

Xu X, Yang D, Ding J-H, Wang W, Chu P-H, Dalton ND, Wang H-Y, Bermingham JR, Ye Z, Liu F, et al. 2005. ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell* **120**: 59–72.

CHAPTER 5

Concluding Remarks

5.1 Summary

Identifying functional RNA variants and elucidating their regulatory mechanisms represent significant challenges of the post-genomic era. A poorly understood topic is the involvement of single nucleotide variants (SNPs and RNA editing sites) in mediating alternative splicing. In this work, we carried out three in-depth studies to gain a better understanding of the functional roles of single nucleotide variants in alternative splicing.

In Chapter 2, to understand the functional significance of GMAS, we systematically identified intronic tag variants for GMAS using ENCODE RNA-seq data specific to cellular compartments. Combined with our previous method that identifies exonic tags for GMAS, this study yielded 622 GMAS exons. We investigated the functional significance of GMAS events through various downstream analyses. Our results showed that GMAS variants were cell-type-independent, implying they could have function across cell types (among the 7 cell lines included in this study), and GMAS patterns demonstrated positive selection or accelerated evolution specific to primates. In addition, many GMAS variants were in LD with GWAS SNPs, suggesting the functional significance of GMAS. To uncover the *trans*-factors responsible for GMAS events, we carried out *de novo* motif analysis and predicted that GMAS SNVs often alter binding of splicing factors. Demonstrating global allele-specific binding, SRSF1 regulated the most GMAS events. This work has been published in *Genome Research*, 2016. Y.H.E. Hsiao,

J.H. Bahn, X. Lin, T.M. Chan, R. Wang, and X. Xiao. "Alternative Splicing Modulated by Genetic Variants Demonstrates Accelerated Evolution Regulated by Highly Conserved Proteins."

In Chapter 3, we performed a comprehensive investigation on the genome-wide impact of A-to-I editing on pre-mRNA splicing. To do so, we generated a unique fractionation RNA-seq dataset by separately sequencing RNAs from chromatin, nucleoplasm, and cytoplasm. The RNA-seq of different subcellular fractions gave us the opportunity to examine the timing of A-to-I RNA editing during transcription. Over 95% of A-to-I RNA editing events occurred in the chromatin-associated RNA prior to polyadenylation. About 500 editing sites that altered splicing of the associated exons were located in 3' acceptor sequences. Furthermore, the combination of bioinformatic and experimental analyses revealed a second category of exons whose splicing is likely modulated by RNA secondary structures recognized by ADAR. This work has been published in *Genome Research*, 2018. Y.H.E. Hsiao, J.H. Bahn, Y. Yang, X. Lin, S. Tran, E.W. Yang, G. Quinones-Valdez, X. Xiao. "RNA editing in nascent RNA affects pre-mRNA splicing."

In Chapter 4, we investigated the GMAS phenomenon in a large panel of human tissues from GTEx. We identified over 8000 GMAS events and showed that these GMAS exons are highly tissue- and individual-specific. To understand the functional relevance of GMAS events, we performed GO analysis and enrichment analyses by utilizing LD with GWAS SNPs and gene overlap with various disease databases. Our results showed that lowly prevalent GMAS events are more often disease-relevant than ubiquitous GMAS events. Furthermore, we developed a novel method to predict causal SNPs for GMAS using population-level data. A total of 628 putative causal SNPs for 317 GMAS exons emerged from this genome-wide prediction. By motif

and eCLIP analyses of RBPs, we additionally predicted a few key splicing factors regulating GMAS exons through putative causal SNPs. This work is in preparation for submission.

5.2 Conclusions

In recent years, technological advances have brought a fundamental shift in our approaches to splicing-related questions. Global analyses that combine high-throughput experimental assays and bioinformatic methods have become indispensable. As a result, numerous novel insights have been revealed regarding the landscape of alternative splicing and the regulatory mechanisms of splicing in various cell types and tissues. Our efforts on developing robust methodologies and generating high-quality RNA-seq datasets contributed to the understanding of the roles of RNA SNVs (both SNPs and RNA editing sites) in *cis*-regulation of alternative splicing. In each of the studies presented in Chapters 2 to 4, we further explored and validated different models of regulatory mechanisms including allele-specific binding by RNA binding proteins (RBPs) and RNA accessibility modulated by RNA secondary structures. These accomplishments facilitated better interpretations of the synergy between *cis*-elements and *trans*-factors in order to regulate splicing. These global discoveries constitute a foundation for further mechanistic and functional studies in model systems and translational research. However, many challenges remain in handling high-throughput experiments and data analyses.

First, aligners still suffer from low accuracy when mapping reads to spliced junctions. This issue is critical since PSI estimation primarily depends on read alignment. It is difficult to determine whether a *de novo* junction read that skips many exons arises from real biological signal or technical artifacts. In addition, accurate PSI estimation is a long-standing challenge in

the field. Many different approaches have been proposed for PSI calculation. Methods that utilize reads from both the exon body and spliced junctions have power to estimate an inclusion level for every exon that is expressed in the sample. However, the PSI estimation can be easily skewed by misaligned reads. On the other hand, some methods use only junction reads that resemble spliced junctions. The benefit of these junction-based methods is that the splicing pattern is directly supported by the data. However, the problem is that these methods are limited to datasets that are deeply sequenced in order to provide sufficient read coverage at the exon-exon boundaries. In addition, these junction-based methods cannot be applied to lowly expressed exons without junction reads in the sample. Another caveat is the lack of agreement of the results from different methods, even when their calculations are based on the same RNA-seq data (Kakaradov et al. 2012). Thus, there is still a significant need in developing a sensitive and accurate PSI calculation method.

Secondly, most splicing studies so far rely on short-read RNA-seq datasets, ranging from 32bp to 101bp, to detect splicing patterns. It is often unfeasible to reconstruct full-length isoforms or identify complex splicing patterns with short reads. Third-generation sequencing technologies are promising in detecting full-length RNAs. Yet, long-read RNA-seq is limited by its high error rate and low throughput. Thus, transcript quantification based on long reads is still an arduous task. One approach for dealing with these tradeoffs is to take the merits of both short-read and long-read RNA-seq by combining the two data types (Saldi et al. 2016). This hybrid method can provide a comprehensive view of isoform structure, as well as isoform quantification. Future data analysis strategies and improvements in sequencing will contribute to more sophisticated transcriptome characterization.

Furthermore, we should also strive to generate datasets from more diverse population cohorts. Although we have many valuable large-scale datasets publicly available, such as those from ENCODE and GTEx, most samples in these datasets come from the Caucasian population, while other ethnic groups are underrepresented. As ethnicity appears to be an important factor in many diseases and traits, datasets from a broader range of populations as well as more phenotypic conditions will dramatically empower prospective translational research on the functional relevance of GMAS. Another important source of data for GMAS studies is based on the eCLIP and knockdown RNA-seq data of RBPs, including splicing factors. The ENCODE Consortium has devoted great effort to generate data capturing *in vivo* and *in vitro* RBP binding (Van Nostrand et al. 2017). These datasets are invaluable to investigating the mechanisms of GMAS observations. We anticipate great advances in the splicing field upon generation of additional datasets for more RBPs.

Additionally, in Chapter 3, we demonstrated an underappreciated mode of splicing regulation, which is the impact of RNA structure on splicing. Thus far, we still lack the ability to not only computationally predict but also experimentally capture RNA secondary structures accurately. Recent emerging high throughput sequencing technology allows us to obtain a better picture of RNA structures in a much larger scale than before (Graveley 2016). Nonetheless, substantial improvements in experimental techniques and data analysis methods await to be made in order to uncover functional and regulatory insights into RNA secondary structures.

Moreover, there is not yet a database cataloging experimentally validated GMAS events from the literature. As shown in Chapters 2 and 4, GMAS likely explains the biological functions of the vast number of intronic or intergenic GWAS SNPs, which do not induce amino acid changes and are thus very hard to decipher. In Chapter 4, we presented the first attempt to predict causal SNPs for GMAS events. During our investigation, we realized that a large fraction of causal SNP candidates turned out to be homozygous in the GTEx cohort. We envisage to achieve more predictions when applying our causality method to other large-scale datasets obtained from individuals with more diverse genetic backgrounds, such as those from the 1000 Genomes Project. Our studies provide a way to prioritize candidates for wet lab verification, and the validated events can then serve as positive controls or training data for future development of bioinformatic algorithms that predict functional SNPs in alternative splicing. Such a database will be an important resource to the genomics and clinical communities.

Last but not least, future efforts can be made to develop machine learning methods that incorporate both SNPs and RNA editing sites as causal candidates for AS events. This advancement will open intriguing opportunities to study the interaction between RNA variants in splicing regulation. In sum, we expect that addressing these challenges and gaps, via data generation and coordination in addition to methodology development and standardization, will catalyze exciting discoveries in splicing research.

5.3 References

Graveley BR. 2016. RNA Matchmaking: Finding Cellular Pairing Partners. *Mol Cell* **63**: 186–189.

Kakaradov B, Xiong HY, Lee LJ, Jojic N, Frey BJ. 2012. Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. *BMC Beioinformatics* **13 Suppl 6**: S11.

Saldi T, Cortazar MA, Sheridan RM, Bentley DL. 2016. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *Journal of Molecular Biology*.

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Blue SM, Dominguez D, Cody NAL, Olson S, Sundararaman B, et al. 2017. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* 179648.