COMPUTATIONAL MODELING OF SPLICING REGULATION

Hai Lin

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

July 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Huanmei Wu, PhD, Chair

Doctoral Committee

Sarath Chandra Janga, PhD

April 20, 2017

Xiaowen Liu, PhD

Yunlong Liu, PhD

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor Professor Dr. Yunlong Liu, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I would also like to thank my committee members, professor Huanmei Wu, professor Sarath Chandra Janga, and professor Xiaowen Liu for serving as my committee members. I also want to thank you for your brilliant comments and suggestions, thanks to you. I would also like to thank all the faculty members of School of Informatics and Computing. Thanks for their tremendous effect in design and delivering excellent course material.

A special thanks to my family. Words cannot express how grateful I am to my parents for all of the sacrifices that they have made on my behalf. I would also like to thank all of my friends who supported me in writing, and incented me to strive towards my goal. At the end, I would like express appreciation to my beloved wife who was always my support in the moments I need most.

Hai Lin

COMPUTATIONAL MODELING OF SPLICING REGULATION

Alternative splicing is one of the most important post-transcriptional modification in cell. It increases the coding capacity of the genome by enable one gene encoding multiple proteins. The majority of human protein-coding genes undergo alternative splicing. And mis-splicing of those genes are known to be associated with many human diseases. Therefore, it is important to study and understand the splicing regulatory machinery. The splicing regulation consists of two components: trans-acting regulators and cis-acting elements. In this dissertation, we explored these two aspects of splicing regulation. First, we investigate the relationship of three key trans-acting regulators: hnRNP A1, SRSF1 and U2AF with transcriptome-wide individual-nucleotide resolution cross-linking and immunoprecipitation (iCLIP) data. Our result revealed the competition relationship between hnRNP A1 and SRSF1 on 3' splicing sites, and the inhabitation effects on U2AF recruitment after hnRNP A1 overexpression. We also discovered that Alu elements may serve as cis-acting elements and compete with authentic exons for the binding of U2AF. Second, we developed a machine learning algorithm to prioritize the disease-causing probability of intronic single-nucleotide variants (iSNVs) by evaluating their cis-acting impact on both alternative splicing and protein structure. The resulting predictive model can predict pathogenic iSNVs with high accuracy and outperform popular algorithms such as splicing-based analysis of variants (SPANR) and

combined annotation–dependent depletion (CADD). This suggests that protein structure features can provide additional layer of information in prioritizing pathogenic iSNVs. In conclusion, our studies provide remarkable insights on alternative splicing regarding both trans-acting regulation and cis-acting regulation. The discoveries of our research on trans-acting regulators are valuable for understanding splicing regulatory machinery. The algorithm we developed can be used to prioritize pathogenic iSNVs without needing to test them all in expensive and laborious assays.

Huanmei Wu, PhD, Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
CADD	Combined Annotation–dependent Depletion
cDNA	Complementary Deoxyribonucleic Acid
CLIP	Cross-linking Immunoprecipitation
DNA	Deoxyribonucleic Acid
ESE	Exonic Splicing Enhancer
ESS	Exonic Splicing Silencer
GTEx	Genotype-Tissue Expression Project
HGMD	Human Gene Mutation Database
hnRNP A1	Heterogeneous Nuclear Ribonucleoprotein A1
iCLIP	Individual-nucleotide Resolution Cross-linking and Immunoprecipitation
INDEL	Insertion/Deletion
ISE	Intronic Splicing Enhancer
ISS	Intronic Splicing Silencer
MCC	Matthews Correlation Coefficient
NOVA	Neuro-oncological Ventral Antigen
PAR-CLIP	Photoactivatable Ribonucleoside–enhanced Crosslinking and Immunoprecipitation
PSI	Percentage of Spliced In
PTB	Polypyrimidine Tract-binding Protein
RIP	RNP Immunoprecipitation

RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RRM	RNA Recognition Motif
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SF1	Splicing Factor 1
SNP	Single Nucleotide Polymorphism
snRNP	Small Nuclear Ribonucleoproteins
SNV	Single Nucleotide Variant
SPANR	Splicing-based Analysis of Variants
SRE	Splicing Regulatory Elements
SRSF1	Serine/Arginine-rich Splicing Factor 1
SVM	Support Vector Machine
U2AF2	U2 Auxiliary Factor 2
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

Chapter 1 Introduction

1.1 Background

Alternative splicing is a post-transcriptional process that allows one gene to code multiple proteins. Recent high-throughput sequencing studies estimate that more than 95% of human multi-exon genes undergo alternative splicing [1]. The malfunction of splicing can cause a lot of human diseases. It has been estimated that 60% of mutations cause disease by disrupting splicing [2]. Thus, the study of splicing regulation is very important.

Splicing regulation consist of two parts. The first part is trans-acting regulations. Trans-acting regulation is performed by RNA binding proteins which can bind to different regulatory elements and act as splicing repressors or activators. The majority of splicing repressors are heterogeneous nuclear ribonucleoproteins (hnRNPs) such as hnRNP A1. On the other hand, most of the splicing activators are members of the serine/arginine-rich protein family (SR proteins) such as SRSF1 (also known as SF2) [3]. The splicing repressors and activators can inhibit or promote the binding of splicing factors such as U2 auxiliary factor (U2AF), assist the assembly of spliceosome, and further affect the alternative splicing [4]. Investigating of key trans-acting regulators can help us understand the splicing machinery.

The second component of splicing regulation is cis-acting elements. Such regulatory elements consist of exon and intron splicing enhancers (ESEs and

ISEs), as well as silencers (ESSs and ISSs) [3]. Those elements locate at different genomic regions can promote or suppress the inclusion of exons. The same regulatory element can have different effects on splicing regulation when bound by different RNA binding proteins. Hence, the trans-acting regulators and cis-acting elements work together as a complex system to regulate the alternative splicing.

The function of cis-acting elements can be disrupted by genetic mutations. Intronic regions are a common source of such genetic mutations. Many of the mutations in intronic regions affect pre-mRNA splicing. Alternative splicing occurs in the majority of genes and frequently alters the function of the resulting gene. Since there are a large number of intronic variants but only a small fraction of them cause altered biological functions, it is important to be able to predict the impact of the variants without needing to test them all in expensive and laborious assays.

In this dissertation, our goal is to: 1.) investigate trans-acting regulatory mechanism of several key splicing-related RBPs; and 2.) predict the cis-acting impact of intronic variants on splicing outcome and protein function.

1.2 Trans-acting Regulatory Mechanism of RNA Splicing

1.2.1 Significance

Trans-acting regulators compose a key component of splicing regulation. While there are hundreds of regulators involve in the mRNA splicing, only a small proportion of them are well studied. Moreover, the relationship between regulators

and how they contribute to the global splicing regulation network need to be thoroughly investigated. Regulators like U2AF, hnRNP A1, and SRSF1 interact with each other and contribute in the initial process of spliceosome assembly. Analysis of their binding patterns with high-throughput sequencing technology in the whole transcriptome can help understand their role in splicing regulation network.

An Alu element is a kind of DNA repetitive element only exists in primate [5]. With over one million copies, it is known to be the most abundant transposable element in the human genome [6]. While cumulative evidence suggests it plays a critical role in nonsense-mediated mRNA decay and primate evolution, its role in splicing regulation remains less understood [7-9].

1.2.2 Critical Barrier

As mentioned above, the splicing regulation is a very complex system involves hundreds of RBPs. Although individual-nucleotide resolution cross-linking and immunoprecipitation (iCLIP) and other high-throughput analysis methods allow the identification of RBP binding in whole transcriptome scale, large effort is still needed to understand the relationship between different RBPs. The same RNA binding protein may have different effects on other proteins and alternative splicing when binding to different genomic regions. The function of RNA binding proteins relate to both genomic context and other RNA binding proteins. Previous low-throughput studies show the competition relationship between hnRNP A1 and

SRSF1 on recruiting splicing factor U2AF. However, transcriptome-wide analysis is needed to understand function and relationship of these key splicing regulators.

Although the exonization of Alu is reported to play a role in nonsense-mediated mRNA decay, the mechanism of different RBP binding on Alu elements is still unclear [10]. The function of Alu in splicing regulation is also unknown. A systematically investigation need to be performed to understand the RBP such as U2AF binding on Alu elements based on different cellular context.

1.2.3 Innovation

This study collected the high-throughput sequencing data of hnRNP A1, U2AF and SRSF1. Our collaborators constructed hnRNP A1 over expressed cell line, which allows us to integrate iCLIP data of hnRNP A1, U2AF and SRSF1 under both control and hnRN PA1 over expression conditions. Different RNA binding proteins are analyzed as a whole system rather than separated pieces. The relationship among those proteins will be analyzed systematically based on different genomic context. It can help people to understand the function of those RNA binding proteins in alternative splicing regulation.

The U2AF binding change after hnRNP A1 over expression on Alu elements has never been reported before. How this binding change affects alternative splicing is unknown. The mechanism of how those proteins cooperate or compete around Alu elements is also unclear. This study not only try to interpret the mechanism of

U2AF binding change on Alu elements, but also further analyze the effects on alternative splicing and nonsense-mediated mRNA decay.

1.3 Cis-acting Impact of Intronic Variants on Splicing and Protein Function

1.3.1 Significance

The fundamental goal of genetics is to identify functional variants that can potentially cause human diseases. Although the functional importance of intornic single nucleotide variants (iSNVs) are not well studied; accumulated evidence suggests that iSNVs often affect pre-mRNA splicing and have very important phenotypic impacts [11-13]. In the Human Gene Mutation Database (HGMD), thousands of SNVs have been documented to be functionally associated with human diseases. Among those iSNVs, a large proportion can affect the regulation of alternative pre-mRNA splicing [14, 15]. With the current development of high-throughput genomic technologies, more and more disease-related iSNVs are continuing to be identified.

1.3.2 Critical Barrier

Currently, there are several bioinformatics algorithms that are available to predict the functional impact of DNA variants on RNA splicing. However, they lack the ability to predict the effect of iSNVs on protein function and further disease phenotypes. To effectively study thousands of iSNVs for further functional and clinical researches, an effective bioinformatics algorithm is required for prioritizing the disease-causing probability of iSNVs based on their impacts on splicing

regulation and protein function. So far, there is only one algorithm, SPANR (Splicing-based Analysis of Variants), can predict the effect of iSNVs on splicing. However, it cannot predict the disease-causing probability and does not consider the impact of the alternatively spliced exons on the protein function [16]. Our studies on small insertions/deletions (INDELs) and synonymous SNVs suggest that simply including or excluding a sequence of amino acids does not necessarily alter the protein function [17-19]. This is also confirmed by previous reports that splicing-related variations can be “passenger” variants, and thus unimportant to the phenotype [20]. Therefore, to predict the functional importance and disease-causing probability of iSNVs, it is important to integrate the protein structure-related features that determine the impacts of alternatively spliced exons on protein structure and function.

1.3.3 Innovation

One major challenge of predict the functional importance of iSNVs is the lack of training data sets. In this study, we collected a high quality training and testing dataset on functional iSNVs by integrating the data from several proprietary and public data sources. First, we collaborated with the Human Gene Mutation Database (HGMD), and extracted all the iSNVs that are manually curated to cause human diseases by altering RNA splicing. The current HGMD professional database contains > 3,000 intronic disease-causing iSNVs that affect RNA splicing. Additionally, the iSNVs reported in the ClinVar database, a public database that contains potential human pathogenic variants, are used as an independent test set

[21]. To our knowledge, this is the most comprehensive and high quality dataset to study the function impacts of iSNVs in human diseases.

Besides genomic features that potentially disrupt binding affinities of RNA-binding proteins (RBP), our predictive algorithm innovatively integrates protein structure related features which can characterize the impacts of alternatively spliced exons on protein function. As shown in our previous INDEL (insertion/deletion) study, the inclusion and exclusion of amino acid sequences does not necessarily alter the function of affected proteins. In this study, our result demonstrates that including protein structure related features significantly improves prediction power when prioritizing pathogenic iSNVs. This represents the major innovation of the developed algorithm since the only published algorithm that can be used for iSNVs prioritization, SPANR, only predicts how an iSNV can affect splicing outcome and does not consider if the affected exon can alter protein function [16].

Finally, the bioinformatics algorithm we developed in this study can effectively combines multi-omic datasets form the public domain. This includes data form the 1,000 Genome Projects, Human Gene Mutation Database, ClinVar, GTEx (Genotype-Tissue Expression) project, CLIPdb (CLIP-seq dataset repository), and RNAcompete (in vitro RNA binding protein – RNA interaction map) [22-24]. Additionally, several of bioinformatics tools that are developed for RNA biology and protein structure prediction such as SPINE-D and SPINE-X are also integrated into the model [25, 26].

1.4 Objectives

The main objective of this dissertation is to explore the splicing regulation machinery through two aspects. First, we investigated the relationship of three key splicing regulators: hnRNP A1, SRSF1, and U2AF. Second, we developed a machine learning model to predict the cis-acting impact of intronic single-nucleotide variants on splicing outcome and protein function.

Chapter 2 reviews the function and importance of RNA splicing, as well as its association with diseases. It introduces the two components of splicing regulation: trans-acting regulators and cis-acting elements. It also summarizes the experimental assays and computational approaches to investigate the trans/cis-acting splicing regulation. Current algorithms used to predict the functional impacts of genetic variants are also discussed in this chapter.

Chapter 3 describes the major findings we got from the iCLIP data on hnRNP A1, SRSF1, and U2AF. We investigated the competition relationship between hnRNP A1 and SRSF1 in U2AF recruitment in the whole transcriptome. We also discovered the absorptive ability of Alu elements on U2AF triggered by hnRNP A1 overexpression. The iCLIP data in both control and hnRNP A1 overexpression HEK293 cell lines are provided by Jonathan Howard and Jeremy Sanford from Molecular, Cellular and Developmental Biology Department, University of California Santa Cruz.

Chapter 4 describes the machine learning model we developed to prioritize the disease-causing probability of intronic single-nucleotide variants (iSNVs). We evaluated not only the influence of iSNVs on alternative splicing, but also the impact of altered exons on protein function. By combining the splicing, conservation, and protein structural features, our method can prioritize pathogenic iSNVs with high accuracy.

Chapter 5 concludes that our research is valuable for understanding the splicing regulation. The predictive model we proposed can pre-screen large scale of intronic variants for further experimental validation. This chapter also discusses the further direction of our study.

Chapter 2 Literature Review

2.1 Alternative Splicing

Gene splicing is one of the most important post-transcriptional modification in which primary transcript RNA is converted into mature RNA. This modification is important for the correct translation of eukaryotic genome since the precursor messenger RNA (pre-mRNA) consists of both exons and introns. During the splicing process, introns are removed from the pre-mRNA and exons are ligated to re-form the mature messenger RNA (mRNA) molecule. The resulting mRNA is then further decoded by ribosomes to generate proteins.

The RNA splicing process can generate different mature mRNA molecules from the pre-mRNA transcribed from one gene by altering the exon composition of the RNA molecule. This is known as alternative splicing. Alternative splicing can happen because of different types of splicing events. Exons can be included or excluded. Introns can be removed or retained. Exons can also have alternative 3' splice sites (acceptor sites) or 5' splice sites (donor sites). With numerous combination of splicing events, alternative splicing allows one gene to greatly increase its coding capacity. Hence, alternative splicing is a critical cause of proteomic diversity.

2.1.1 The Function and Importance of Splicing

Recent researches suggest that human genome may contain as few as 19,000 protein-coding genes [27]. In contrast, it is estimated that there are above 200,000

proteins in human genome [28]. Alternative splicing is one of the major reasons why such a large number of proteins can be coded in a much smaller number of genes. Current high-throughput sequencing studies show that more than 90% of human genes are alternatively spliced [1, 29, 30]. As a result, the coding potential of human genes are significantly boosted by alternative splicing. Consequently, proteins can have varies functions due to the regulation of alternative splicing.

Alternative splicing has a great potential to regulate the function of proteins by alter the exon composition of protein-coding transcripts. A lot of studies show that alternative splicing have significant biological effects in different species, despite the fraction of alternative splicing has such impact is hard to estimate [31]. Although the function change caused by a single splicing event usually is small, changes often happen at a bunch of related genes regulated by a small set of splicing regulatory proteins, which can have strong biological effects. Such effects can include the regulation of the protein-protein interaction, protein-nucleic acid interaction and the binding of membrane proteins. Alternative splicing can also control the protein localization, enzyme activity as well as their interaction with ligands [32]. These impacts suggest that alternative splicing play critical role in the physiological process.

2.1.2 The Relationship between Alternative Splicing and Disease

The functional importance and complexity of alternative splicing make it vulnerable to genetic and environmental changes [33]. The malfunction of splicing is known

to associate with many human diseases [16, 34]. Studies show that 10% of pathological mutations locate at splice sites [35, 36]. Further, it has been estimated that 60% of mutations cause disease by disrupting RNA splicing [2, 37].

Such a huge number of disease-associated alternative splicing events can be results of different mechanisms [34, 36]. First, mutations directly happen at splicing sites of functional important proteins can cause diseases. Studies show that in familial dysautonomia (FD), 99.5% of cases are caused by the intronic mutation near 5'-splice site of exon 20 in IKBKAP gene [38]. This mutation can prevent the binding of U1 and cause the skipping of exon 20. Such exon skipping can introduce the premature stop codon and result in the malfunction of IKBKAP [39]. Second, splicing regulatory elements can be affected by pathogenic mutations. Evidence show that in spinal muscular atrophy (SMA), the mutation happens at splicing regulatory element of exon 7 in SMN2 converts an exonic splicing enhancer (ESE) to an exonic splicing silencer (ESS) by allowing a binding site for hnRNPA1 [40, 41]. This causes the exon skipping of exon 7 in SMN2 and produces nonfunctional SMN2 protein [42]. Third, the dysfunction of splicing factors can also disrupt the alternative splicing of critical proteins. In frontotemporal lobar degeneration (FTLD) and amyotrophic lateral sclerosis (ALS), splicing factor TDP43 is cleaved by caspase-3 [43, 44]. Loss of TDP43 expression induces anomalous upregulation of CDK6, which causes the cell death in FTLD [45].

Significant number of alternative splicing changes are also found in many cancers. Evidence suggests that the changes of splicing regulators such as hnRNPs and SR proteins are related to cancer development. For instance, the expression level and activity of YB1 and ASF/SF2 have been reported to change in ovarian and breast cancer [46, 47]. In addition, several studies demonstrate that splicing alterations in cancer-associated genes can be critical in cancer progression. One example is that the mutation near the splicing site of exon 2 in tumor suppressor KLF6 can induce the expression of KLF6-SV1 isoform. The upregulation of such isoform promotes the cell proliferation and results in the acceleration of prostate cancer [48-50]. Another example is the mutations at donor site of intron 4 in adenomatous polyposis coli (APC) can cause exon skipping and lead to the familial adenomatous polyposis (FAP) [51]. All those examples illustrate that the alteration of splicing play important role in cancer development and progression. Study of splicing regulation is critical in cancer diagnosis and treatment.

2.2 Splicing Regulation

RNA splicing is regulated by the combination of both trans-acting regulators and cis-acting elements. The detail function of those two components will be discussed in the following sessions. With the assistance of those two, a series of phosphodiester transfer reactions are performed by a large RNA and protein complex known as the spliceosome. The spliceosome is composed of more than 100 proteins and five small nuclear ribonucleoproteins (snRNPs) [4]. There are two types of spliceosomes: the major spliceosome which involves in the canonical

splicing; and the minor spliceosome which involves in the non-canonical splicing [52].

The major spliceosome, which performs canonical splicing, contains snRNP U1, U2, U4, U5, and U6 [4, 53]. During the splicing process, snRNP U1 recognizes and binds to the GU sequence at the 5’ splicing site of an intron [54]. Other RNA-binding proteins such as splicing factor 1 (SF1) binds to the branch site of the intron [55]. U2 auxiliary factor 1 and 2 (U2AF1/U2AF2) then binds to the 3' splice site and polypyrimidine tract respectively [56]. In this stage, a complex known as E complex is assembled. Next, snRNP U2 replaces SF1 and binds to the branch site with assistance of U2AF [57]. This forms a pre-spliceosome A complex. Formation of A complex is considered playing a critical role in determining the exon-intron boundary during alternative splicing [3]. Further, a U4/U6.U5 tri-snRNP is recruited with U5 binding to the 5’ splicing site of exon region and U6 binding to snRNP U2 [58]. The intermediate product is known as pre-catalytic spliceosome B complex. It is then converted to B* complex by releasing snRNP U1, shifting U5 from exon to intron region, and attaching U6 to 5’ splicing site [59]. Finally, snRNP U4 is released and catalytic spliceosome C complex is generated with a series of conformation changes. The C complex then performs two transesterification reactions. The first one cleaves the 5'-end of the intron and ligates it to the branch site to form a lariat through a 2',5'-phosphodiester linkage [60]. The second one cleaves 3'-end of the intron, and ligates the two exons together through ATP

hydrolysis. The lariat is then released and degraded [61]. This whole process is known as canonical splicing or lariat pathway.

While canonical splicing explaining more than 99% of splicing, non-canonical splicing happens much less frequently [62]. The non-canonical splicing is carried out by minor spliceosome which consists of a different set of snRNPs comparing to the major spliceosome. The snRNPs U1, U2, U4, and U6 in major spliceosome are replaced with U11, U12, U4atac, and U6atac respectively in minor spliceosome [63]. The snRNP U5 is shared by both major spliceosome and minor spliceosome. Consequently, minor spliceosome recognizes non-canonical splicing site rather than the canonical one (GU-AG) [62]. Also, it locates near the nuclear membrane outside the nucleus [64].

2.3 Trans-acting Splicing Factors and Their Role in Splicing Regulation

Besides the core proteins of spliceosome, splicing relies on the regulation of numerous trans-acting regulators. Trans-acting regulators are RNA-binding proteins which can promote or suppress the splicing process. Based on their function, they can be classified as activators and repressors. However, the function of those trans-acting regulators usually depend on the location of binding [65]. By binding towards different regions of RNA, a regulator can switch its role between activator and repressor. This special phenomenon introduces additional complexity into the splicing regulation. Despites such complexity, tremendous efforts have been put into the research of key trans-acting regulators such as

serine/arginine-rich proteins (SR proteins), heterogeneous nuclear ribonucleoprotein (hnRNP), and U2 auxiliary factor (U2AF).

SR proteins constitute a family of proteins which involve in splicing regulation [66]. All members of the family contain two protein domains: the RNA recognition motif (RRM) region and the RS domain with enriched serine (S) and arginine (R) residues [67]. Most studies suggest that SR proteins serve as activators in splicing process [68, 69]. During spliceosome assembly, SR proteins bind to the exonic splicing enhancers (ESEs) and recruit snRNP U1 and U2AF to the 3' splicing site and 5' splicing site respectively [70, 71]. This process helps to stabilize the formation of E complex. Later, SR proteins also assist snRNP U2 recognizing and binding to the branch site, which converts E complex to A complex [72, 73]. Finally, SR proteins can recruit U4/U6.U5 tri-snRNP and help to assemble B complex [74, 75]. As stated above, trans-acting regulators may play as different role when binding to different positions. SR proteins serve as activators when binding to exonic regions, but act as repressors when binding to intronic regions [70, 76]. These mechanisms show that SR proteins play critical roles in both constitutive splicing and alternative splicing regulation.

hnRNPs are protein complexes of RBPs and heterogeneous nuclear RNAs (hnRNAs). They typically contain at least one of the three RNA-binding motifs: the RNP-CS-RBD domains; the RGG (Arg-Gly-Gly) domains; and the K-homology (KH) domains [77]. Although the mechanism of how hnRNPs regulate splicing is unclear,

they are most known as repressors in splicing regulation [78]. Reports show hnRNPs can compete with SR proteins and suppress the exon inclusion. During this process, hnRNPs may bind to the exonic splicing silencers (ESS) and prevent the binding of SR proteins to the adjacent ESE [79, 80]. On the other hand, similar to SR proteins, hnRNPs can also perform as splicing activators. Studies show that hnRNPs can help recruit U2AF to assemble spliceosome by proofreading 3' splicing site AG dinucleotides [81].

U2 auxiliary factor (U2AF) is a heterodimer which consists of two subunits: U2AF65 (65-kDa) and U2AF35 (35-kDa) [82, 83]. U2AF65 contains four RNA-binding domain and binds to the pyrimidine tract (Py tract) upstream of 3' splicing sites, while U2AF35 associates with the AG dinucleotide at 3' splicing sites [84-87]. Binding of U2AF is required for the recruitment of snRNP U2 and the assembly of spliceosome [88]. Multiple studies show that U2AF can help recognize and define 3' splicing sites, as well as regulate alternative splicing [89-91]. At the meantime, the binding of U2AF is also regulated by other RNA-binding proteins including hnRNPs and SR proteins. It has been shown that U2AF has direct competition relationship with certain hnRNPs such as hnRNP C [10]. Contrarily, the binding of SR proteins can help recruit U2AF during spliceosome assembly [70].

2.4 Technologies for Detecting RNA-protein Interaction

Since RBPs play critical roles in RNA splicing, investigating RNA-protein interaction can help to understand the mechanism of splicing regulation and determine the effects of RBPs on the transcriptome. Due to the importance and complexity of RBPs, several methods have been developed to study the interaction between RNA and proteins. Traditionally, such interactions were analyzed in a low-throughput manner. In recent years, high-throughput techniques such as microarray and high-throughput sequencing allow researchers to investigate the function of RBPs systematically.

Systematic evolution of ligands by exponential enrichment (SELEX) is an effective approach to identify the binding motifs of RBPs [92]. The process first synthesizes all the possible short oligonucleotides randomly as potential targets. The random oligonucleotides are then exposed to interested RBPs. After washing away un-bound oligonucleotides, the bound targets are amplified with RT-PCR [93]. This method is wildly used to determine high affinity RBP binding motifs. However, SELEX usually can only detect motifs with the highest affinity and fail in identifying moderate and low affinity motifs [94, 95]. Moreover, traditional SELEX analysis lacks the ability to investigate the RBP binding sites across the whole transcriptome *in vivo*.

To systematically investigate the RNA-protein binding within a cellular context, RNP immunoprecipitation (RIP) followed by microarray (RIP-Chip) or high-

throughput sequencing (RIP-Seq) is developed [96, 97]. The main process relies on immunoprecipitation of RBPs with specific antibodies. Any RNA fragments that bound by the RBP of interest pulled down. The isolated RNAs then can be analyzed with microarray or high-throughput sequencing technology [98]. This method allows the measure of dynamic RBP binding status. It can not only capture the RNA-protein interaction but also keep the information of RNA-RNA interaction [99]. However, this approach can only be used on ribonucleoprotein particles (RNPs) and the low resolution of the data limits the accuracy of RBP binding sites identification [100, 101].

In order to address the drawbacks of RIP, cross-linking immunoprecipitation (CLIP) is developed to investigate the RNA-protein interaction with high resolution and specificity [102-104]. This approach utilizes ultraviolet (UV) light to cross-link RNA and proteins *in vivo*. The UV light will induce the formation of covalent bonds between proteins and nucleic acids, which can stabilize the RNA-protein complex [105]. After isolating the RNA-protein complex with antibodies, the protein will be digested with Proteinase K. The remaining RNA molecules can be enriched with RT-PCR for further analysis [102]. Originally, the downstream analysis is performed by Sanger sequencing. After the invention of high-throughput sequencing, several modified approaches are developed to improve the scale and accuracy of detection.

High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP), also known as CLIP-Seq, combines CLIP with high-throughput sequencing technology to detect RNA-protein interaction across the whole transcriptome [106]. Since the UV cross-linking is irreversible, the peptides are left at the cross-link sites after Proteinase K digestion. These peptides can affect reverse transcriptase during RT-PCR step and produce truncated cDNA molecules. Even if the reverse transcriptase can read through those cross-link sites, higher error rate will be induced during nucleotide base paring [102]. Such cross-linking induced mutation sites (CIMS) can provide additional information for identifying RNA-protein binding sites with high resolution [104].

Photoactivatable ribonucleoside–enhanced crosslinking and immunoprecipitation (PAR-CLIP) is developed to increase the resolution of detecting RNA-protein interaction. Due to the low efficiency of 254 nm UV used in the HITS-CLIP and the low percentage of CIMS in read-through cDNA fragments, the ability of identifying exact RNA-protein binding sites is limited [107]. PAR-CLIP solves this issue by adding photoreactive ribonucleoside analogs into the RNA of cultured cells. 365 nm UV light is then used to cross link the photoreactive ribonucleoside-labeled transcripts with RNA-binding proteins. Introducing of photoreactive ribonucleoside analogs, such as 4-thiouridine (4-SU) and 6-thioguanosine (6-SG), can cause thymidine to cytidine, and guanosine to adenosine mutations during cross-linking process respectively [108]. Such process can effectively produce CIMS for high-resolution RNA-protein binding site detection. However, PAR-CLIP can only be

applied on cultured cells, which limits its application on tissue-based assays [101]. Also, studies show that the incorporation ribonucleoside analogs may inhibit the synthesis of ribosomal RNA and trigger a nucleolar stress response [109].

Individual-nucleotide resolution cross-linking and immunoprecipitation (iCLIP) allows the identification of RBP binding sites at single nucleotide resolution. It improves HITS-CLIP with cDNA circularization after reverse transcription. This modification enables both truncated and read-through cDNAs to be sequenced effectively [110]. Mapping of truncated sites back to genome can allow identification of cross-link sites at nucleotide level [111].

2.5 Cis-acting Elements and Their Influence on Splicing Regulation

Splicing regulatory elements (SREs) are short cis-acting regulatory sequences locate in transcripts. They can recruit trans-acting regulators to promote or inhibit the splicing process. Based on the function and location of those elements, they can be classified as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) [4].

Splicing Enhancers are usually bound by activators such as SR proteins. SR proteins which bind to ESEs through their RRM domains can further recruit U2AF65 with RS domains for spliceosome assembly [112]. ESEs can also promote splicing through bridging interactions with the help of SRm160 and SRm300 [113,

114]. During this process, those splicing co-activators can mediate the interactions between SR proteins and snRNPs. In addition, ESEs can assist the binding between RS domains and branch sites, which may promote the formation of A complex [72, 115]. ISEs activate splicing through different mechanisms. One example is that the binding of hnRNP L to the CA repeats can recruit snRNP U1 to weak 5' splice sites [116].

Splicing Silencers typically interact with repressors such as hnRNPs. They inhibit splicing through two mechanisms [3]. First, silencers can block the access of activators to the splicing sites by recruiting high-affinity repressors. Polypyrimidine tract-binding protein (PTB), as a member of hnRNP family, can bind to ESSs and compete with U2AF65 [117]. This process can prevent the interaction between snRNP U1 and U2, which inhibits the assembly of A complex [118]. Second, splicing silencers can bind to either side of exons and loop out the exon in between, suppressing the splicing reaction. The binding of hnRNP A1 to ESSs can block the recruitment of snRNPs and SR proteins, which can induce the “looping out” of exon [79, 119]. Similarly, ISSs can also be bound by hnRNPs such as PTB, hnRNP A1 and hnRNP L to inhibit the splicing process [116, 120].

The function of SREs can also depend on the context. Elements with similar nucleotide sequences can function as enhancers or silencers based on the relative positions within transcripts [121]. For instance, poly-G usually serve as splicing enhancers when located in intronic regions, but they can suppress the splicing

when located in exonic regions [122, 123]. In contrast, the YCAY binding motifs of neuro-oncological ventral antigen (NOVA) act as splicing enhancers or silencers when located in exons or introns, respectively [124]. Such “context dependence” of SREs exhibits the flexibility and complexity of splicing regulation.

2.6 Methods for Modeling Cis-acting Splicing Regulation

With the utilization of high-throughput sequencing in RNA-protein interaction studies, a large number of cis-acting elements and associated trans-acting regulators are identified. It dramatically extends our knowledgebase on splicing regulation. Numerous attempts have been made to predict the splicing pattern based on known regulatory elements and splicing machinery.

ExonScan achieves this goal by simulating the splicing process based on cis-acting regulatory elements located in transcripts [125]. The algorithm scans the pre-mRNA and looks for potential splicing sites with maximum entropy splice site models [126]. A score is then assigned to each candidate exon based on enrichment of regulatory elements in the nearby region. Enhancers and silencers contribute to positive and negative scores respectively. The score cutoff is trained based on a set of 1,820 human genes to achieve high accuracy. The results give the prediction of splicing outcome of given primary transcripts.

Algorithm like ExonScan, however, does not take into account the tissue specificity of alternative splicing. The splicing outcome is co-regulated by both cis-acting

elements and trans-acting regulators, while the expression and activity of trans-acting regulators is tissue-dependent. Yoseph et al. addressed this issue by combining hundreds of RNA features to predict tissue-specific alternative splicing changes [127]. They collected both the genomic features around exons of interest and splicing profiles across different tissue types. A splicing code was inferred with selected features to predict the splicing pattern in particular tissue types using information theory [128]. This method can be used to predict alternative splicing in a tissue-specific manner. And the result of the prediction can be evaluated to help better understand the splicing regulation machinery.

Michael et al. further improved such prediction framework by adopting deep learning architecture [129]. They extended the feature set to over a thousand RNA features. Those features, together with the tissue-specific splicing profiles, were used to build a multi-layer deep neural network [130]. Such model can predict not only the absolute percentage of spliced in (PSI) of exons for each tissue types, but also the splicing change (delta PSI) between different tissues. The result demonstrates the regulatory function of both cis-acting elements and trans-acting regulators. The various model performance across different tissues, however, suggests the complexity of splicing regulation.

2.7 Algorithms for Prioritizing Functional Effect of Variants

Single nucleotide polymorphisms (SNPs) are the most common genetic variants in the human genome, and they are believed to involve in many diseases. Much

effort has been put into the prediction of functional effects of SNPs. Most early algorithms focus on exonic non-synonymous SNPs (nsSNPs) due to their simplicity and data availability. Two of the most widely used algorithms are SIFT and PolyPhen-2 [131]. SIFT predicts the impacts of nsSNPs on protein function based on sequence homology information gathered by PSI-BLAST, as well as the physical properties of amino acids [132]. PolyPhen-2 collects not only sequence and phylogenetic related information but also structure-based features. A naive Bayes classifier is then trained to predict the functional effects of SNPs [133].

As increasing number of prediction algorithms being developed, researchers realized the limitation of using single predictive method. Ensemble learning framework like combined annotation–dependent depletion (CADD) has been developed to overcome such limitation [134]. CADD integrated the genomic features such as GERP, phastCons, phyloP, as well as prediction scores from other algorithms like SIFT and PolyPhen-2. It then built a support vector machine (SVM) on top of those to predict the deleteriousness of SNPs. Although the prediction power is improved by such method, the performance is still restrained by the individual model and features used in each sub-models.

Since the price of whole-genome sequencing drops below $1,000, more and more intronic SNPs data are available. Intronic SNPs started to gather its attention on functional prediction. Most of tools focus on predicting the impacts of intronic SNPs on splicing outcome. SPANR is based on a neural network trained on RNA-Seq

data of 16 tissue types with more than a thousand genomic features [16]. It can predict the delta percentage of inclusion (dPSI) of exons caused by nearby intronic SNPs. However, such method fail to consider how those altered exons will affect the function of transcribed protein products. Our previous studies on small insertions/deletions (INDELs) and synonymous SNVs demonstrate that simply altering the inclusion of an exon may not have any impact on protein function [17].

Chapter 3 Alu Elements and hnRNPA1 Facilitate Transcriptome-wide Redistribution of Splicing Factor U2AF2

3.1 Introduction

Alternative splicing drives transcriptome and proteome diversification. While splicing regulatory proteins govern this process, their global mechanisms remain enigmatic. Splicing factors orchestrate the assembly of alternative messenger RNA transcripts from precursor mRNA. This process diversifies genetic information by expanding the coding capacity of the genome [121]. hnRNPA1 is a splicing silencer that is overexpressed in many human cancers [135-139]. Biochemical experiments suggest hnRNPA1 regulates exon inclusion through several mechanisms [79, 80, 140-144]. One commonality may be antagonization of splicing factors that recognize the 3' splices site. hnRNPA1 modulates interactions between U2AF2 and the polypyrimidine track [81]. Likewise, hnRNPA1 may also influence the association of SRSF1 with exonic splicing enhancer sequences [79, 140] (Figure 1). Despite recent genome wide protein-RNA interaction studies the generality of these models remains enigmatic [145-147]. Our collaborator generated high resolution transcriptome-wide protein-RNA interaction maps to determine how the splicing repressor hnRNPA1 influences the global association of spliceosome assembly factor U2AF2 and SRSF1 with pre-mRNA. We observed changes in the distribution of U2AF2 crosslinking sites relative to the 3’ splice sites of cassette exons but not constitutive exons upon hnRNPA1 overexpression. By contrast, SRSF1 crosslinking patterns relative to splice sites are independent of hnRNPA1 expression levels. We also observed an

hnRNPA1-dependent increase in U2AF2 but not SRSF1 crosslinking to exon proximal antisense Alu elements. Thus Alu elements can serve as splicing factor-responsive sinks for U2AF2. These results not only demonstrate a novel mechanism for alternative splicing regulation but also implicate retrotransposon-derived sequences in the evolution of species-specific alternative splicing.

All the PCR, sequencing, and western blot experiment are performed by Jonathan Howard and Jeremy Sanford from Molecular, Cellular and Developmental Biology Department, University of California Santa Cruz.

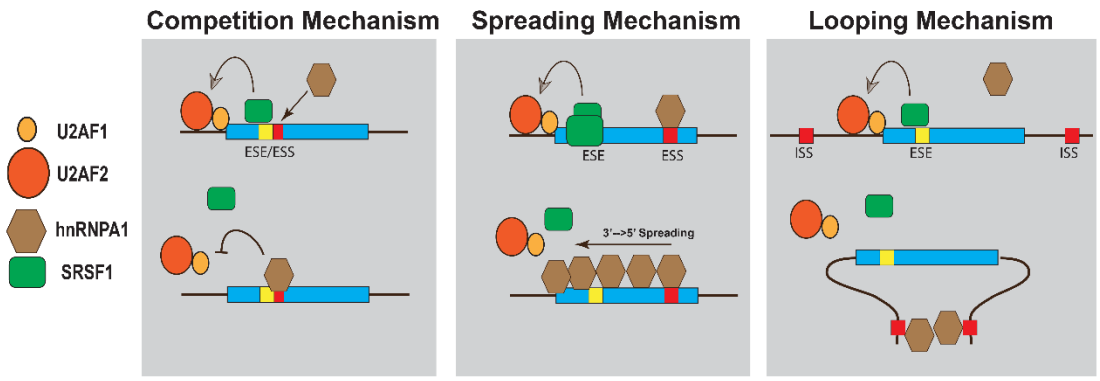

Figure 1 hnRNPA1 regulatory mechanisms. Potential models by which hnRNPA1 is thought to regulate binding of additional splicing factors to promote exon skipping. These models universally suggest hnRNPA1 act primarily as an antagonizer of splicing enhancers through different, but not necessarily mutually exclusive, mechanisms**.**

3.2 Materials and Methods

3.2.1 iCLIP method

iCLIP was performed as previously described [111, 148]. Briefly, TREX FLP-in HEK293T cells (Invitrogen) lacking or containing a stable, inducible T7-tagged version of hnRNP A1. Cells were treated with tetracyclin for 24 hr and then irradiated with UV-C light to form irreversible covalent cross-link between proteins and nucleic acids *in vivo*. After cell lysis, RNA was partially fragmented using low concentrations of Micrococcal nuclease, and U2AF65-, SRSF1-, or hnRNP A1–RNA complexes were immunopurified with α-U2AF65, (MC3;SCBT), α-SRSF1 (96;SCBT), and α-hnRNP A1 (4B10;SCBT) antibodies immobilized on protein A–coated magnetic beads (Life Technologies), respectively. After stringent washing and dephosphorylation (Fast AP, Fermentas), RNAs were ligated at their 3′ ends with a pre-adenylated RNA adaptor (Bioo Scientific) and radioactively labeled to allow visualization. Samples were run using MOPS-based protein gel electrophoresis (in-house recipe) and transferred to a nitrocellulose membrane. Protein-RNA complexes migrating 15 -80 kDa above free protein were cut from the membrane, and RNA was recovered from the membrane by proteinase K digestion under denaturing (3.5 M Urea) conditions. The oligonucleotides for reverse transcription contained two inversely oriented adaptor regions adapted from the Bioo NEXTflex small RNA library preparation kit (Bioo Scientific), separated by a BamHI restriction site as well as a barcode region at their 5′ end containing a 4-nt experiment-specific barcode within a 5-nt random barcode to mark individual cDNA molecules. cDNA molecules were size-purified using denaturing PAGE gel

electrophoresis, circularized by CircLigase II (Epicenter), annealed to an oligonucleotide complementary to the restriction site and cut using BamHI (NEB). Linearized cDNAs were then PCR-amplified using (Immomix PCR Master Mix, Bioline) with primers (Bioo) complementary to the adaptor regions and were subjected to high-throughput sequencing using Illumina HiSeq. A more detailed description of the iCLIP protocol has been published [111].

3.2.2 Mapping and Analysis of iCLIP sequencing data

Single end reads generated by Illumina HiSeq were inspected for the presence of adaptor sequences. Reads containing sequences corresponding to the 3'RNA adaptor were retained if they were at least 30bp long after the adaptor sequence was trimmed off. The first 9bp in each read from the iCLIP library preparation, containing an internal barcode comprising 4bp for replicate identification and 5bp of random nucleotides for use in duplicate mapping removal, were also removed before mapping. Trimmed reads were checked for mapping to a repeat filter comprising RepeatMasker elements in the human genome using Bowtie2 [149]. Reads that passed the repeat filter were mapped to the transcriptome and genome with Tophat [150]. If reads mapped equally well to multiple loci, a single mapping was selected randomly by Tophat. Duplicate mappings from each replicate were reduced to one per position if they had the same genomic end points and if they originated from reads with the same set of random 5bp nucleotides. Following mapping and duplicate removal, individual reads were truncated to their 5' ends to represent the site of crosslinking consistent with the iCLIP methodology. For all

samples only such crosslinking sites found to have non-zero mapping counts in two out of three replicates (or two out of two duplicates where applicable) were considered to be biologically reproducible candidates for further analysis. The counts at such reproducible crosslinking sites were summed over all replicates to create an aggregated data set for each cell condition and CLIP. To determine background from the iCLIP data sets, the two cell conditions (control and hnRNPA1-overexpressing) were temporarily further aggregated for each CLIP (U2AF, SF2, A1) and those binding sites that had non-zero counts in all three temporary aggregate data sets were determined. A 41 nucleotide mask was created by extending 20nt upstream and 20nt downstream from each such 3-way common binding site. The aggregated data set of binding sites for each cell condition and CLIP was then filtered using this mask, keeping only sites outside the mask that also had a mapping count of at least 3 in the aggregate data. These aggregated and filtered data were used for downstream analyses. This aggregation and filtering strategy was adapted from previously described iCLIP analysis pipelines [151, 152]. For use as input to CLIPPER [153], the filtered (single nucleotide) binding sites were expanded by 15nt upstream and 15nt downstream.

CLIPper (CLIP-seq peak enrichment; https://github.com/YeoLab/clipper), was used to determine genomic distribution of RNA crosslinking peaks as well as identify clusters representing binding sites for hnRNP A1, U2AF2, and SRSF1 for each condition as previously described [153].

3.2.3 RBP Binding Analysis

40,769 cassette exons are extracted from MISO human genome (hg19) alternative events annotation version 2. 200,880 constitutive exons are extracted from RefSeq gene annotation by excluding the exons that overlap with cassette exons. Gene differential expression analysis is done by edgeR. 40,952 constitutive exons that are not significantly differential expressed (FDR > 0.05) are used in further analysis. For each RNA binding protein in each cell line, the iCLIP reads of all the replicates are merged together. The start position of reads are considered as crosslinking sites. The number of reads near 3' splice site (100bp into the intron, 50bp into the exon) of each exon is calculated based on a 10bp window. The raw read counts is normalized by the total library size. Exons with more than 20 reads in the 150bp region are shown in the plot.

The binding changes of U2AF and SF2 near 3' splice sites are further analyzed with edgeR. Read counts are calculated for 200bp intron regions near 3' splice sites of cassette exons. For each RBP, the regions with more than one count per million (CPM) in at least half of the replicates in either of the cell line are used for binding change analysis.

3.2.4 Relationship between U2AF Binding and Splicing Change

Among 267 splicing changed skipped exon events, with U2AF binding, the PSI (Percent Spliced In) of 74 events are decreased after hnRNPA1 over expression,

and the PSI of 9 events are increased. On the other hand, without U2AF binding, the PSI of 134 events are decreased and the PSI of 50 events are increased.

3.2.5 Motif Analysis

For each condition, the iCLIP data of replicates are merged. Binding peaks are called with CLIPper. The peaks are divided into different categories based on genomic regions including CDS, intron and UTR. Each category is further divided based on whether overlap with Alu elements. 50bp sequences of peak region (crosslinking site ± 25bp) are extracted. A strand-specific MEME-ChIP analysis is performed to find the enriched motifs with width between 6bp to 10bp.

3.2.6 RBP Binding Near Alu Elements

315,974 anti-sense Alu elements are extracted from RepeatMasker. The merged iCLIP data for each condition is down-sampled to 1M reads. The total number of sense strand reads are calculated for Alu and nearby regions (250bp from Alu boundary). For each cassette exon events (cassette exon + up/downstream introns + up/downstream exons), the number of reads in anti-sense Alu elements and the total number of reads in the whole event are calculated separately. The proportion of reads fall into anti-sense Alu elements for each event is used to represent the RBP binding change in Alu regions.

3.2.7 mRNA-Seq of Control or hnRNPA1 Over Expressing HEK293 cells

RNA was isolated from whole cell lysates of control and hnRNPA1-overexpressing TREX Flp-IN HEK 293T cells using TRI-Reagent LS (Sigma). Poly-A+ sequencing libraries were prepared using the TrueSeq RNA library prep kit (Illumina, San Diego, CA). We analyzed each condition in duplicate using the HiSeq2000.

3.2.8 Quantification of Alternative Splicing by RNA-Seq

Poly-A+ transcriptome sequencing reads were mapped to human reference genome (hg19) with TopHat2. Mapped reads of duplicates were merged together for splicing analysis. Splicing change was analyzed with MISO [154]. The MISO result was filtered with parameters: --num-inc 1 --num-exc 1 --num-sum-exc 10 --delta-psi 0.20 --bayes-factor 10. After filtering, 267 skipped exon events were left for further analysis.

3.3 Results

3.3.1 hnRNPA1 Overexpression Induces Genome-wide Redistribution of SRSF1 and U2AF2

To determine how hnRNPA1 influences 3' splice site recognition genome-wide, we perturbed its expression in HEK293 cells and assayed SRSF1, U2AF2 and hnRNPA1 protein-RNA interactions using individual nucleotide resolution crosslinking immunopreciptitation and high throughput sequencing (iCLIP-seq) [148]. Induction of hnRNPA1 results in an approximately 2-3 fold increase compared to the endogenous protein (and relative to EWSR1) and has no

appreciable effect on SRSF1 or U2AF2 steady state protein levels (Figure 2). Additionally, the expression and localization of hnRNPC are unaffected by hnRNPA1 overexpression. We used iCLIP to purify hnRNPA1-, SRSF1-, and U2AF2-RNA complexes, from control and hnRNPA1 over-expressing cells (Figure 4). As expected, the immunoprecipitated material was both UV- and antibody-dependent, nuclease sensitive and produced robust sequencing libraries (Table 1, Table 2, Table 3 and Figure 3). After identification of peaks using CLIPper [153] (Table 4) the distribution of hnRNPA1 peaks between different gene regions is largely unchanged upon hnRNPA1 overexpression overexpression (Figure 5). However, we observed differences between control and hnRNPA1 overexpression cell lines for both U2AF2 and SRSF1 peaks (Figure 5). Most notably, the proportion of U2AF2 peaks located near coding exons or in exon-proximal intronic regions was reduced whereas intronic peaks located more than 500 nt from exons (distal intron) increased. A similar trend was observed for SRSF1 peaks, where a reduction in CDS and concomitant increase in distal intron peaks was evident upon hnRNPA1 overexpression (Figure 5). To determine if hnRNPA1 influences the RNA binding specificity of U2AF2 and SRSF1 we searched for over represented RNA sequences within the binding site peaks (Figure 6). In control cells, U2AF2 peaks are characterized by a pyrimidine-AG motif, closely resembling authentic 3'splice sites. By contrast, a more pyrimidine-rich motif lacking the AG dinucleotide is observed in peaks from hnRNPA1 overexpression cells. By contrast, SRSF1 and hnRNPA1 motifs are similar between control and hnRNPA1 overexpression

cells. These data suggest that enforced expression of hnRNPA1 alters U2AF2 RNA binding specificity.

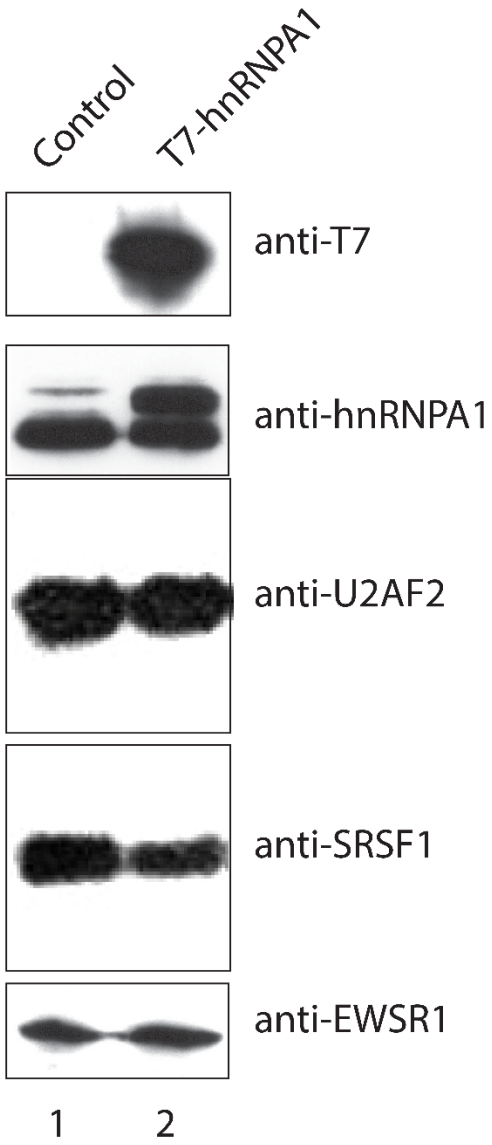

Figure 2 Western blot analysis of control and hnRNPA1 overexpression cell lines. Nuclear extracts were subjected to SDS-PAGE and transferred to nitrocellulose blotting paper. Samples were interrogated with anitbodies for α-T7 peptide tag (with which overexpressed hnRNPA1 has been tagged), α-hnRNPA1 (4B10; Santa Cruz Biotechnology), α-SRSF1 (96;Santa Cruz Biotechnology), α-U2AF2 (MC3;Santa Cruz Biotechnology), and α-EWSR1 (C-9; Santa Cruz Biotechnology) as positive control.

Experiment is performed by Jonathan Howard from Molecular, Cellular and Developmental Biology Department, University of California Santa Cruz.

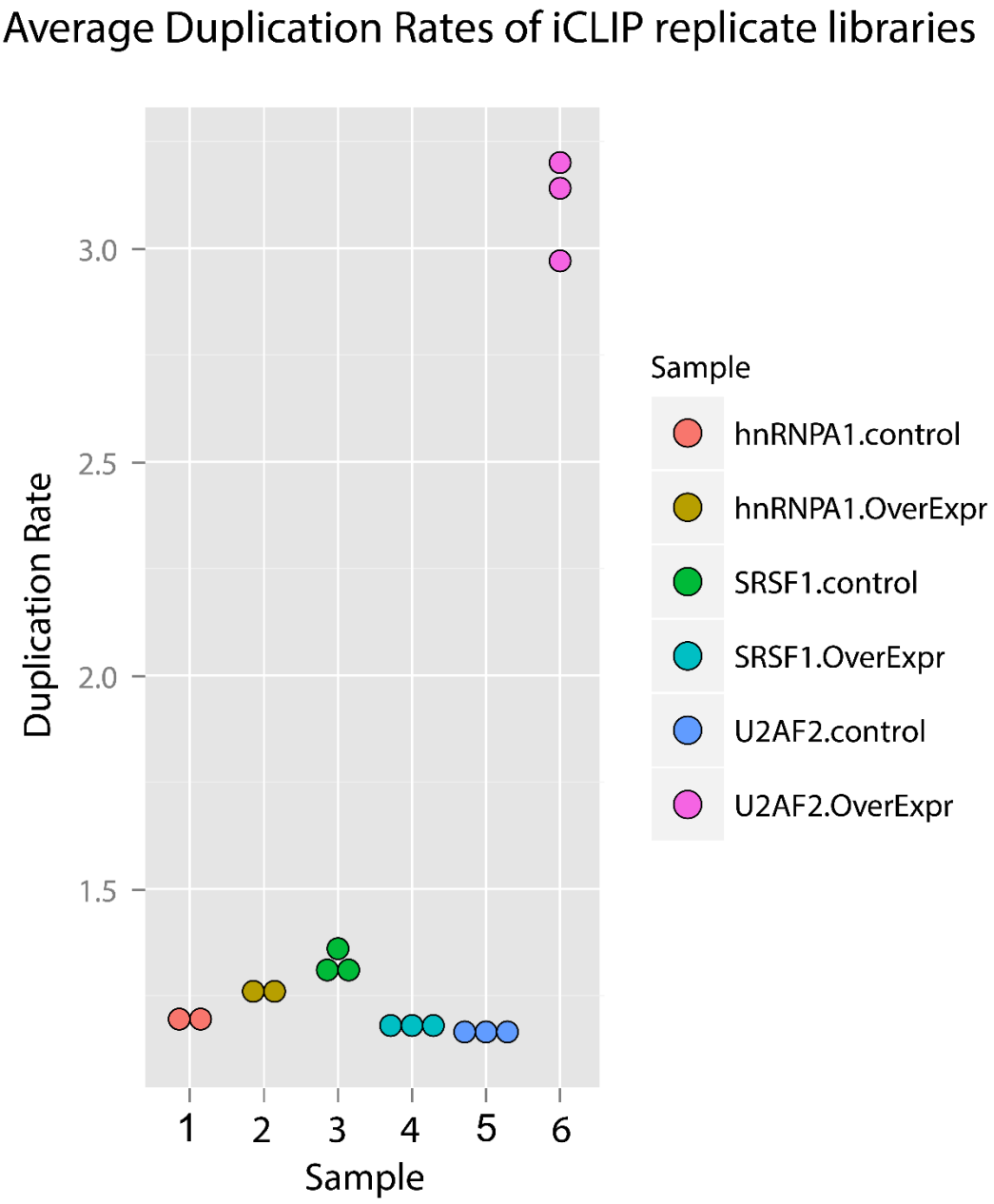

Figure 3 Average Duplication Rates of iCLIP replicate libraries. Dot plot showing the PCR duplication rate for each sequencing library replicate for each iCLIP experiment. “Control” refers to control HEK293 cell lines and “OverExpr” refers to cell lines in which hnRNP A1 overexpression has been induced.

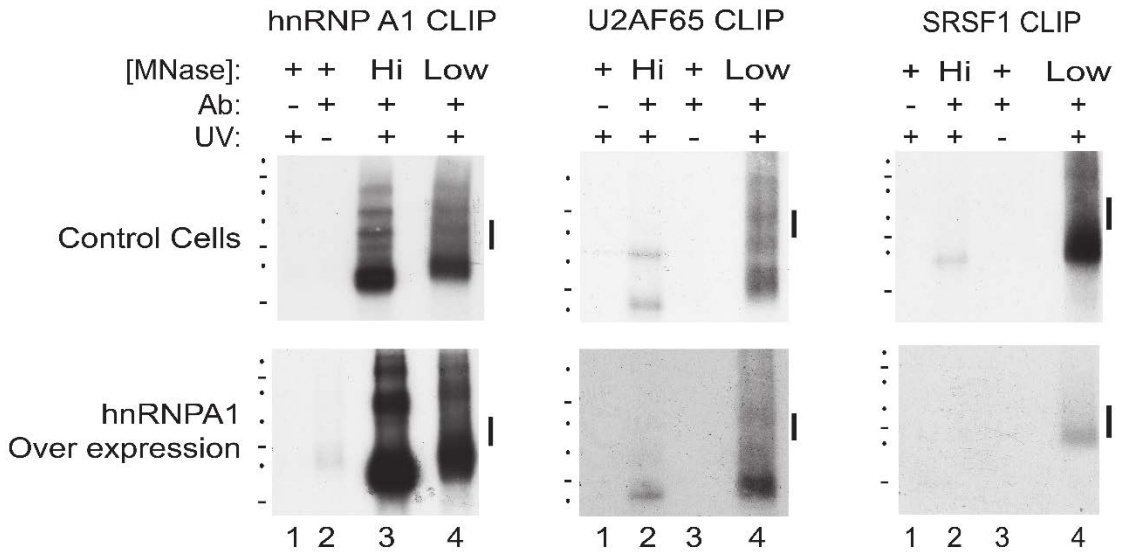

Figure 4 Examples of iCLIP autoradiographs for each protein under control and overexpression of hnRNPA1. Protein-RNA complex shifts are UV-, antibody- and Micrococcal nuclease-sensitive. Bars denote region of nitrocellulose blot excised for RNA isolation for iCLIP library preparation.

Experiment is performed by Jonathan Howard from Molecular, Cellular and Developmental Biology Department, University of California Santa Cruz.

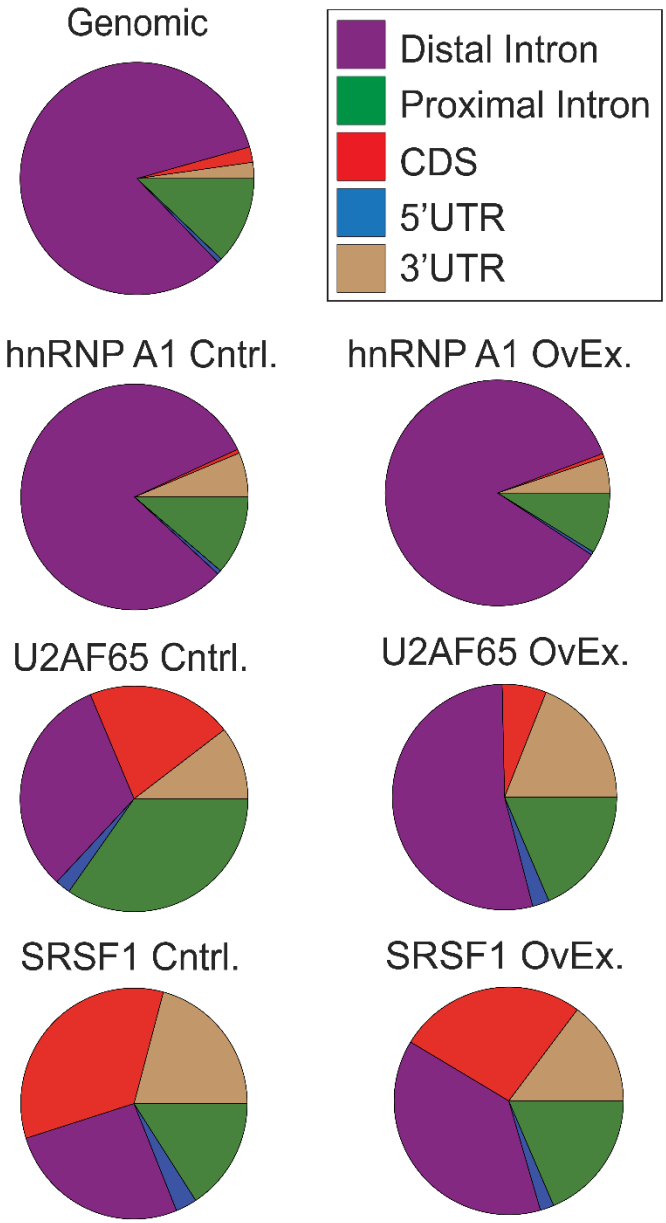

Figure 5 Peak distribution across the genome. CLIPper analysis of iCLIP RNA distribution for hnRNPA1, SRSF1, and U2AF2 for control and hnRNPA1 overexpression conditions.

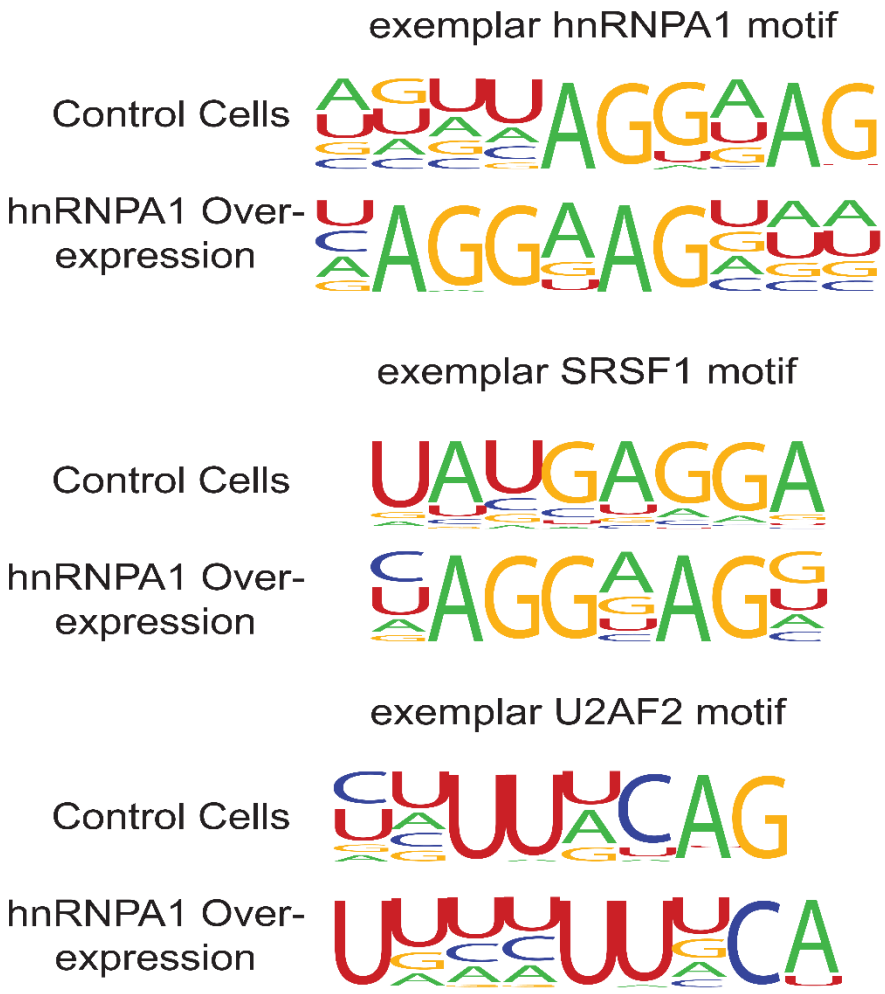

Figure 6 Top HOMER consensus binding motifs for hnRNPA1, SRSF1, and U2AF2 for control and hnRNPA1 overexpression conditions.

Table 1 iCLIP mapping *statistics.*

Sample	Total count	Mapped count	Mapped percentage	Unique count	Unique percentage
hektrx_hna1_jh103b	28.55M	27.59M	96.60%	25.77M	93.40%
hektrx_hna1_jh103c	24.43M	23.60M	96.60%	21.98M	93.10%
hekha1_hna1_jh104a	8.27M	7.91M	95.70%	7.35M	92.80%
hekha1_hna1_jh104b	9.10M	8.74M	96.00%	8.08M	92.50%
hektrx_u2af_jh105a	12.54M	11.75M	93.70%	11.03M	93.90%
hektrx_u2af_jh105b	12.24M	11.79M	96.30%	11.08M	94.00%
hektrx_u2af_jh105c	17.38M	16.40M	94.40%	15.38M	93.70%
hekha1_u2af_jh106a	28.23M	25.68M	91.00%	21.18M	82.50%
hekha1_u2af_jh106b	30.72M	28.12M	91.50%	24.01M	85.40%
hekha1_u2af_jh106c	37.20M	33.33M	89.60%	27.73M	83.20%
hektrx_sf2x_jh107a	3.20M	3.04M	95.10%	2.70M	88.80%
hektrx_sf2x_jh107b	2.63M	2.47M	94.10%	2.16M	87.30%
hektrx_sf2x_jh107c	2.95M	2.85M	96.50%	2.55M	89.30%
hekha1_sf2x_jh108a	3.71M	3.56M	96.10%	3.25M	91.30%
hekha1_sf2x_jh108b	6.09M	5.90M	96.90%	5.45M	92.40%
hekha1_sf2x_jh108c	5.18M	5.01M	96.80%	4.61M	91.90%
hektrx_sf2x_jh107a	17.61M	16.78M	95.30%	14.63M	87.10%
hektrx_sf2x_jh107b	13.06M	12.36M	94.60%	10.44M	84.50%
hektrx_sf2x_jh107c	13.55M	13.12M	96.80%	11.53M	87.90%
hekha1_sf2x_jh108a	20.95M	20.14M	96.10%	18.31M	90.90%
hekha1_sf2x_jh108b	34.03M	33.00M	97.00%	30.44M	92.20%
hekha1_sf2x_jh108c	26.69M	25.87M	96.90%	23.72M	91.70%

Table 2 Comparison of library complexity before and after PCR duplication removal using random indexes. A: before PCR duplication removal. B: after PCR duplication removal.

Sample	Mapped count A	Mapped count B	Ratio AB
hektrx_hna1_jh103b	27.59M	23.25M	1.19
hektrx_hna1_jh103c	23.60M	19.70M	1.2
hekha1_hna1_jh104a	7.91M	6.28M	1.26
hekha1_hna1_jh104b	8.74M	6.91M	1.26
hektrx_u2af_jh105a	11.75M	10.16M	1.16
hektrx_u2af_jh105b	11.79M	10.15M	1.16
hektrx_u2af_jh105c	16.40M	14.02M	1.17
hekha1_u2af_jh106a	25.68M	8.64M	2.97
hekha1_u2af_jh106b	28.12M	8.80M	3.2
hekha1_u2af_jh106c	33.33M	10.60M	3.14
hektrx_sf2x_jh107a	3.04M	2.36M	1.29
hektrx_sf2x_jh107b	2.47M	1.82M	1.36
hektrx_sf2x_jh107c	2.85M	2.14M	1.33
hekha1_sf2x_jh108a	3.56M	3.04M	1.17
hekha1_sf2x_jh108b	5.90M	4.99M	1.18
hekha1_sf2x_jh108c	5.01M	4.21M	1.19
hektrx_sf2x_jh107a	16.78M	3.88M	4.32
hektrx_sf2x_jh107b	12.36M	2.57M	4.81
hektrx_sf2x_jh107c	13.12M	3.02M	4.34
hekha1_sf2x_jh108a	20.14M	14.64M	1.38
hekha1_sf2x_jh108b	33.00M	23.86M	1.38
hekha1_sf2x_jh108c	25.87M	18.74M	1.38

Table 3 Summary of aggregated crosslink data from replicate iCLIP experiments.

	SRSF1 CTL	SRSF1 CTL	hnRNPA1 OE	hnRNPA1 CTL	U2AF OE	U2AF CTL
Total coverage	29.45M	3.99M	1.65M	9.82M	16.54M	6.84M
CDS coverage	3.11M	0.34M	0.02M	0.12M	0.42M	0.25M
CDS percentage	10.57%	8.54%	0.97%	1.25%	2.51%	3.63%
UTR3 coverage	3.76M	0.66M	0.13M	1.01M	3.70M	0.68M
UTR3 percentage	12.78%	16.50%	7.61%	10.30%	22.37%	10.00%
UTR5 coverage	0.60M	0.10M	0.02M	0.11M	0.43M	0.11M
UTR5 percentage	2.05%	2.55%	0.98%	1.11%	2.57%	1.67%
Non-coding coverage	4.53M	0.93M	0.25M	0.96M	1.60M	0.99M
Non-coding percentage	15.38%	23.39%	15.02%	9.80%	9.67%	14.44%
Exon coverage	12.01M	2.03M	0.41M	2.21M	6.14M	2.03M
Exon percentage	40.78%	50.98%	24.58%	22.46%	37.13%	29.73%
Intron coverage	13.59M	1.18M	1.07M	6.36M	5.28M	4.15M
Intron percentage	46.16%	29.52%	64.96%	64.73%	31.93%	60.65%
Intergenic coverage	3.84M	0.78M	0.17M	1.26M	5.12M	0.66M
Intergenic percentage	13.06%	19.51%	10.46%	12.81%	30.94%	9.63%

Table 4 CLIPper output statistics from "pre-mRNA" analysis for aggregated iCLIP data.

	SRSF1 OE	SRSF1 CTL	hnRNPA1 OE	hnRNPA1 CTL	U2AF OE	U2AF CTL
Mapped reads	19.29M	2.41M	0.68M	4.73M	12.92M	2.60M
Number of peaks	202,511	22,393	5,789	59,778	125,588	23,918
Median tags	22	23	19	20	26	20
Average tags	37.1	39.6	26.1	27.6	54.3	29.5
Median length	29bp	34bp	33bp	35bp	14bp	35bp
Average length	28bp	32bp	29bp	33bp	20bp	33bp

3.3.2 hnRNPA1 Influences Binding of SRSF1 and U2AF2 Near 3' Splice Sites

Previous work demonstrated that hnRNPA1 influences binding of SRSF1 and U2AF2 near 3' splice sites [81, 155, 156]. To test this hypothesis, we determined how titration of hnRNPA1 affected the distribution of SRSF1- and U2AF2-RNA crosslinks relative to 3' splice sites of constitutive or alternative cassette exons. As suggested by the peak analysis (Figure 5) there are no differences in hnRNPA1 crosslinking site between control and overexpression cells (Figure 7 and Figure 8 left panels, blue and red lines respectively). SRSF1 crosslinking to exonic sequences was modestly reduced in the hnRNPA1 overexpression cells compared to control, but the positional distribution of the SRSF1 sites relative to the 3'ss was largely unchanged for constitutive and skipped exons (Figure 7 and Figure 8, right panels). U2AF2 crosslinking distribution relative to the 3'ss was substantially altered in hnRNPA1 over expressing cells compared to the control, where a characteristic peak is observed over the 3'ss of both constitutive and skipped exons (Figure 7 and Figure 8, bottom panels). By contrast, in cells overexpressing hnRNPA1, U2AF2 crosslinking density near alternative exons shifts downstream of the 3'ss and the peak is substantially reduced (Figure 8 bottom panel, red line).

To determine if there is a direct relationship between hnRNPA1 binding and changes in U2AF2 or SRSF1 association with transcripts, we examined regions flanking the 3'ss of skipped exons with hnRNPA1 crosslinking in either conditions (Figure 9). In regions with no detectable hnRNPA1 crosslinks, the change in

U2AF2 crosslinking exhibits a bimodal distribution, which corresponds to regions flanking the 3’ss that show either increased or decreased U2AF2 crosslinking in hnRNPA1 overexpression cells relative to control cells (Figure 9 top panel, blue). By contrast, U2AF2 crosslinking to the vicinity of the 3'ss is significantly reduced when direct association of hnRNPA1 is also evident (Figure 9 top panel, pink). Changes in SRSF1 crosslinking appears to be independent of direct hnRNPA1-RNA interactions (Figure 9 bottom panel). To determine if changes in U2AF2 crosslinking correlated with hnRNPA1-dependent splicing regulation we sequenced polyA+ selected RNA libraries from control and hnRNPA1 overexpression cells (Table 5). Of the 267 hnRNPA1-regulated cassette exons, the majority exhibited increased levels of exon skipping upon hnRNPA1 overexpression (Figure 10). A total of 44 hnRNPA1-regulated splicing events also exhibited changes <2-fold change in U2AF2 intronic crosslinking (Table 6). 54% of hnRNPA1-dependent exon skipping events also exhibit depletion of U2AF2 crosslinking within 200bp of the 3’ss, whereas 46% show hnRNPA1-dependent increases in U2AF2 crosslinking in the same region. For example, hnRNPA1-dependent U2AF2 redistribution and alternative splicing COG4 (Conserved oligomeric Golgi complex subunit 4; Figure 11) and SRSF6 (serine/arginine-rich splicing factor 6 or SRp55; Figure 12). In both cases, U2AF2 crosslinking near the 3’ splice sites is reduced in the cell lines over expressing hnRNPA1 (Figure 11 and Figure 12).

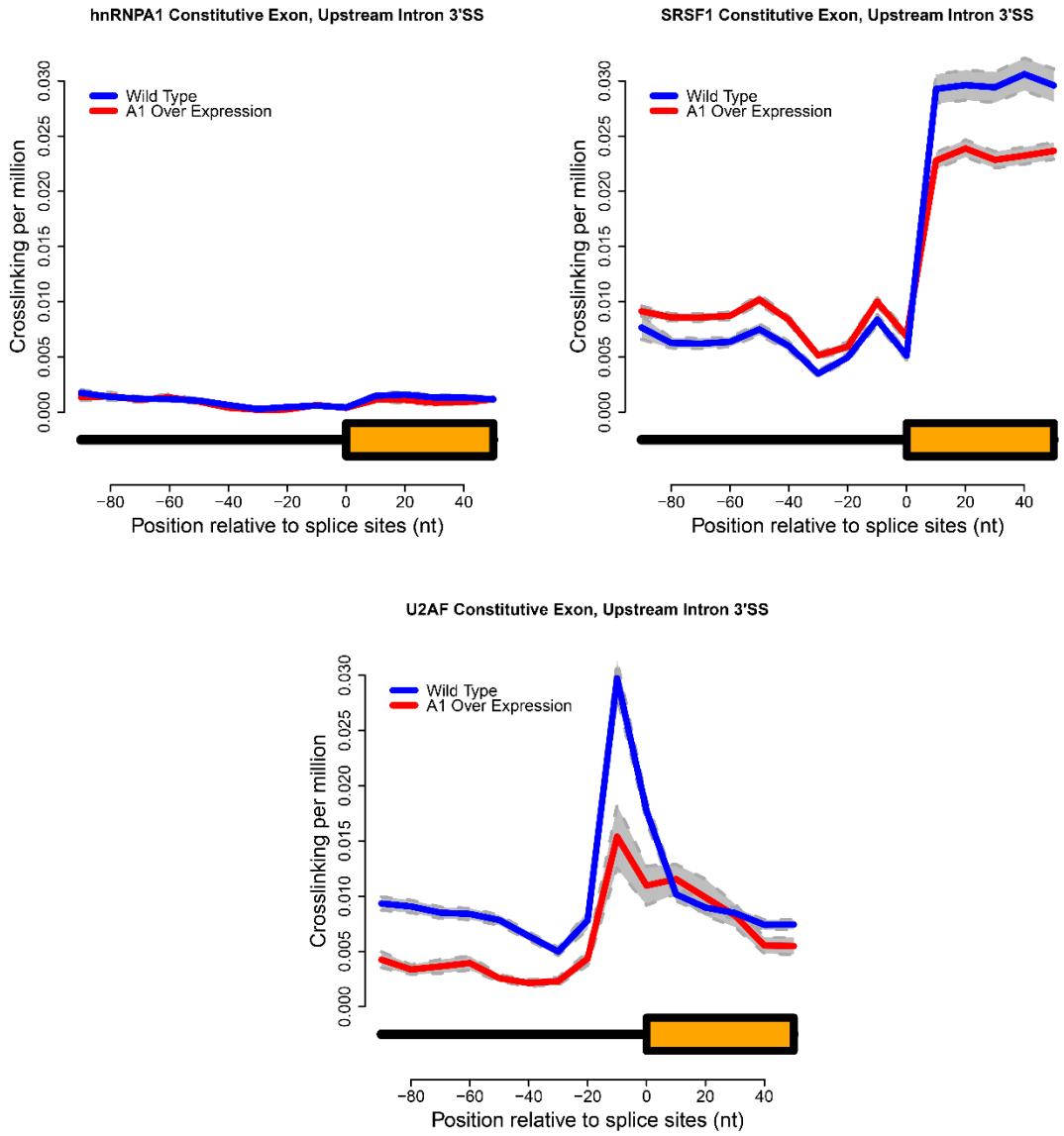

Figure 7 Crosslinking near 3' splice sites of constitutive exons. Normalized crosslinking distribution for hnRNPA1 (left panel), SRSF1 (right panel) and U2AF2 (bottom panel) in wild type (blue line) and hnRNPA1 overexpression cell lines (red line) with 95% confidence interval (grey area).

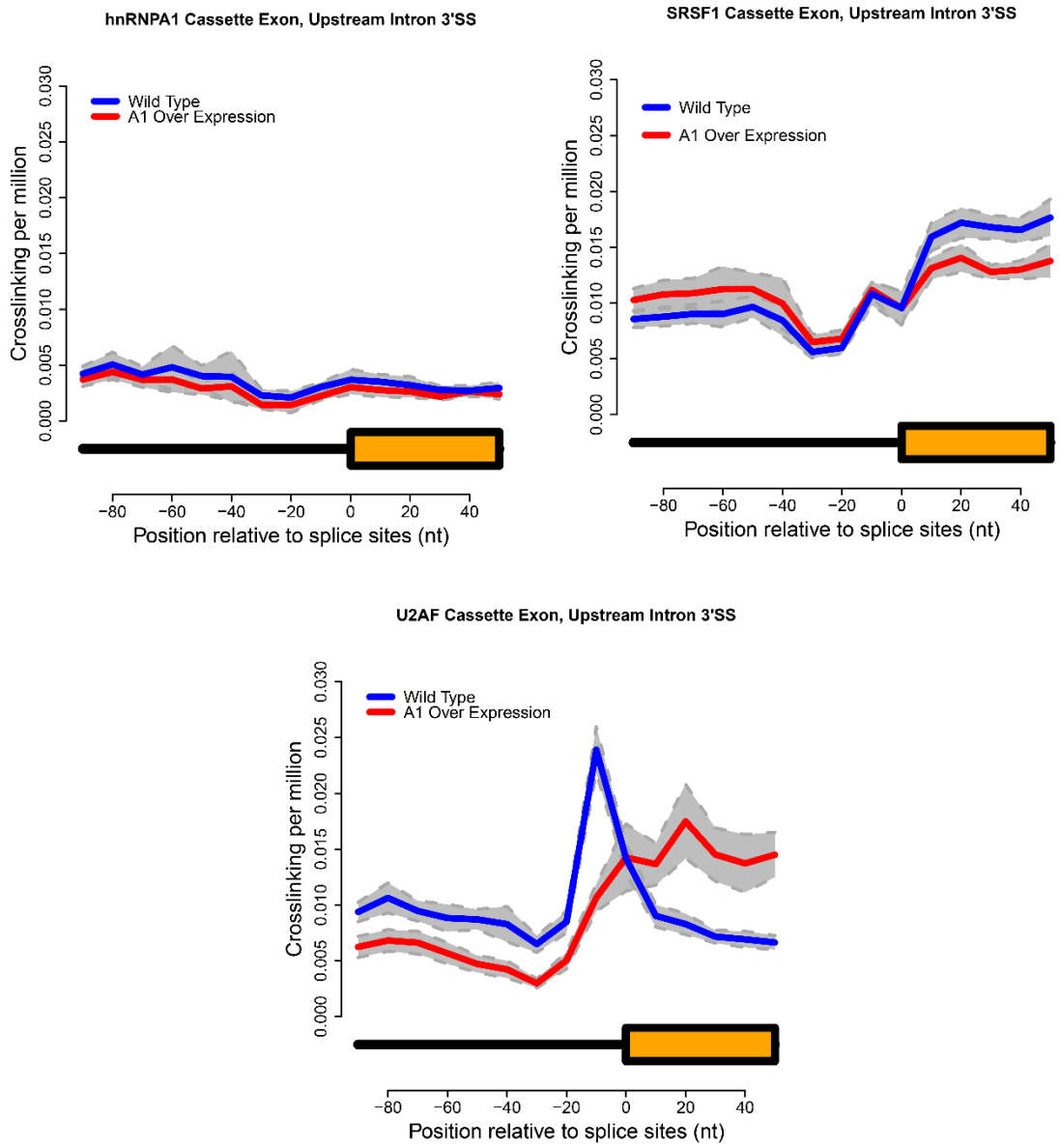

Figure 8 Crosslinking near 3' splice sites of cassette exons. Normalized crosslinking distribution for hnRNPA1 (left panel), SRSF1 (right panel) and U2AF2 (bottom panel) in wild type (blue line) and hnRNPA1 overexpression cell lines (red line) with 95% confidence interval (grey area).

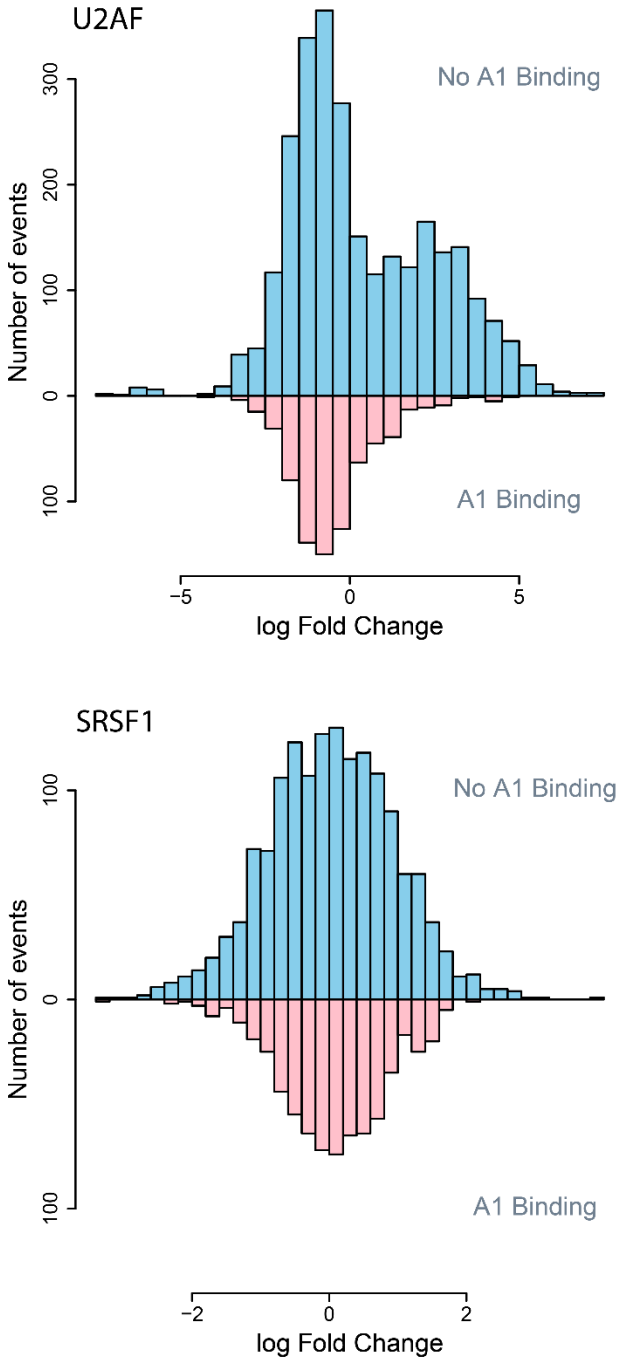

Figure 9 hnRNPA1 induced redistribution of U2AF2 crosslinking near 3' splice sites. Nature log fold change distribution of U2AF and SRSF1 within 200bp intron regions near 3' splice sites of cassette exons. Blue bars corresponds to annotated alternative splicing events with no evidence of hnRNPA1 crosslinking in either condition and pink represents annotated events with detectable hnRNPA1 crosslinking.

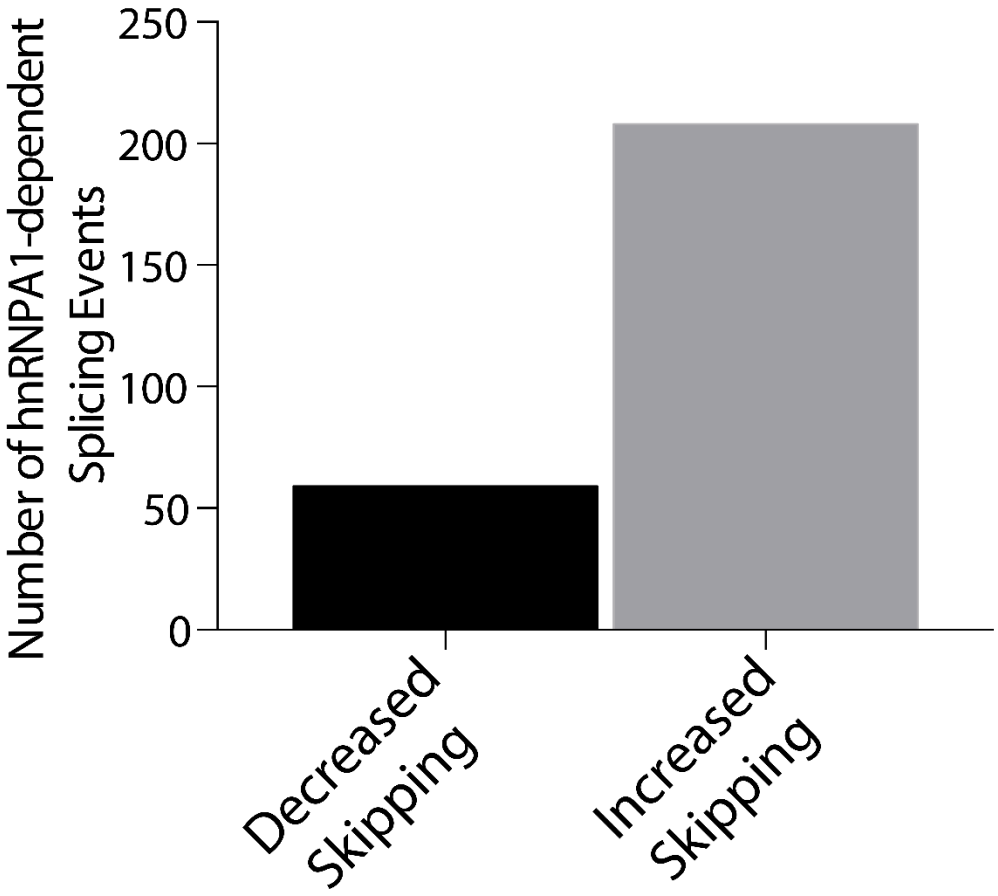

Figure 10 Bar graph depicting the number of alternative cassette exons differentially expressed upon hnRNPA1 overexpression.

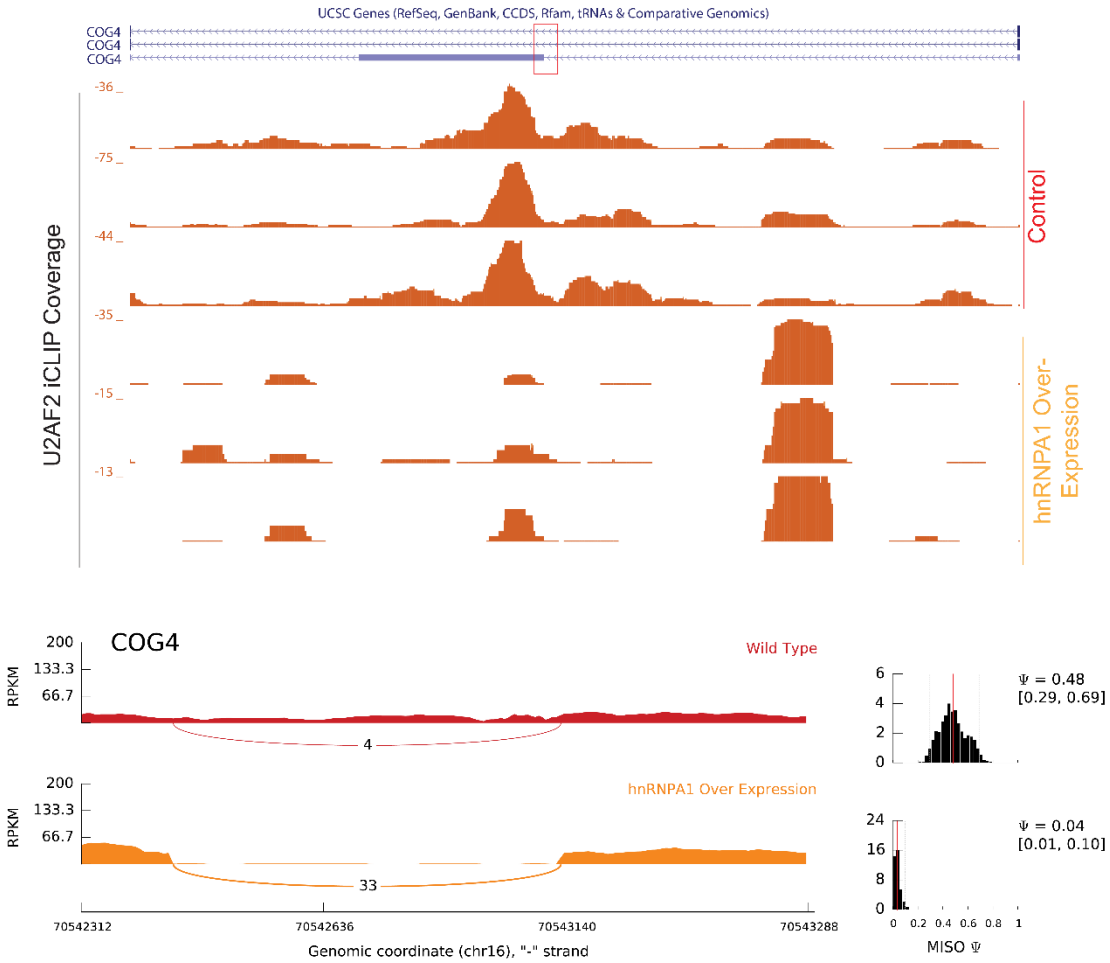

Figure 11 Example of hnRNPA1-dependent modulation of U2AF2 crosslinking and alternative splicing on COG4. UCSC genome browser example of COG4 and CLIP read coverage data for U2AF2 under control and hnRNPA1 overexpression. The sashimi plot representing MISO analysis of RNA sequencing data from samples used for iCLIP. It represents splicing data corresponding to the exon show in the USCS genome browser snapshots.

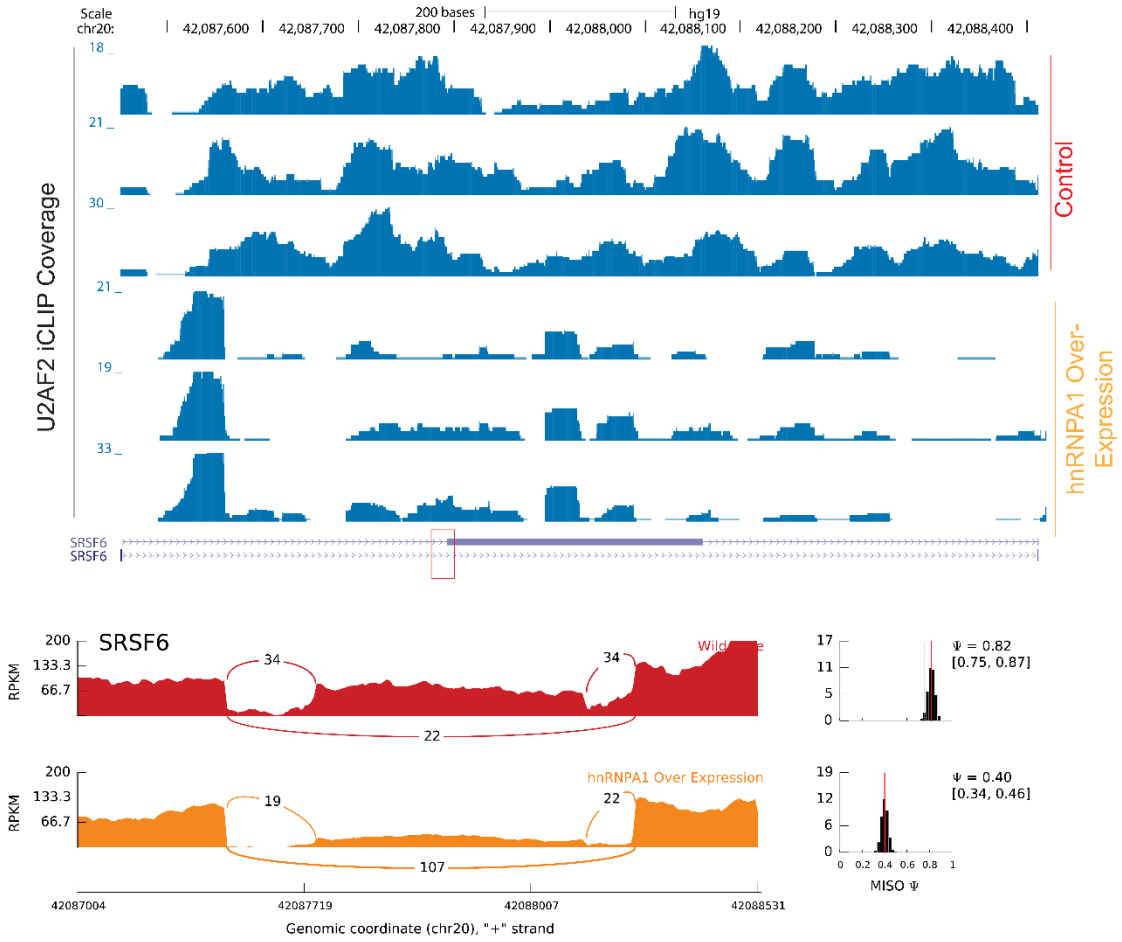

Figure 12 Example of hnRNPA1-dependent modulation of U2AF2 crosslinking and alternative splicing on SRSF6. UCSC genome browser example of SRSF6 and CLIP read coverage data for U2AF2 under control and hnRNPA1 overexpression. The sashimi plot representing MISO analysis of RNA sequencing data from samples used for iCLIP. It represents splicing data corresponding to the exon show in the USCS genome browser snapshots.

Table 5 mRNA-seq mapping statistics from hnRNPA1 over expressing or control cells.

	hnRNPA1 OE1	hnRNPA1 OE2	hnRNPA1 CTL1	hnRNPA1 CTL2
Total count	17.48M	20.98M	20.06M	17.35M
Mapped count	1.96M	15.07M	13.84M	12.56M
Mapped percentage	11.20%	71.80%	69.00%	72.40%
Unique count	1.57M	14.62M	13.49M	12.27M
Unique percentage	80.00%	97.00%	97.50%	97.70%
Junction count	0.39M	5.35M	5.82M	5.31M
Junction percentage	24.80%	36.60%	43.10%	43.30%

Table 6 summary hnRNPA1-dependent changes in exon skipping and U2AF2 positioning.

		hnRNPA1-dependent change in U2AF2 crosslinking near 3'ss		
	Total # of Events	Increased U2AF2	Decreased U2AF2	Redistirbution to Alu
hnRNPA1-dependent exon inclusion	3	2	1	2
hnRNPA1-depednent exon skipping	41	19	22	17

3.3.3 Overexpression of hnRNPA1 Triggers U2AF2 Re-localization to Alu Elements

Previous work by Zarnack et al. demonstrated that hnRNP proteins, such as hnRNPC, can antagonize binding of U2AF2 to Alu element, to repress their exonization [10]. We asked if hnRNPA1 similarly repressed U2AF2 crosslinking to Alu elements by measuring global distribution of crosslinks for each protein overlapping of antisense Alu elements throughout intronic regions with titration of hnRNPA1 levels. Surprisingly upon overexpression of hnRNPA1, we detected a dramatic increase in U2AF2 crosslinking to antisense Alu-containing RNA transcripts compared to control cells (Figure 13). Conversely, hnRNPA1 crosslinking globally decreases over Alu elements with overexpression. By contrast to U2AF2, crosslinking of SRSF1 to antisense Alu elements shows no appreciable changes, suggesting that the effect of hnRNPA1 is specific to U2AF2.

These results suggest that Alu elements with increased U2AF2 crosslinking are located in *cis*- relative to skipped exons (Alu elements upstream or downstream of a splicing event) (Figure 14). To test this hypothesis we compared the proportion of U2AF2 crosslinks within Alu-elements relative to flanking sequences across individual exon skipping events in control or hnRNPA1 overexpression cells. The scatter plot shown in Figure 15, demonstrates that the proportion of U2AF2 crosslinks present in Alu elements increases significantly across virtually all exon skipping events, upon hnRNPA1 overexpression, whereas the proportion of hnRNPA1 crosslinks are decreased (Figure 15). By contrast, the proportion of

SRSF1 crosslinks to Alu elements are refractory to changes in hnRNPA1 expression levels (Figure 15). These data demonstrate a global change in U2AF2-Alu association and refute the hypothesis that a few spurious Alu-elements are responsible for the signal observed in Figure 13. We found that 41% of hnRNPA1-dependent exon skipping events exhibited redistribution of U2AF2 to adjacent Alu elements (Table 6). An example from the PIEZO1 gene is shown in Figure 16. Taken together, these data demonstrate that overexpression of hnRNPA1 triggers U2AF2 relocalization to Alu elements.

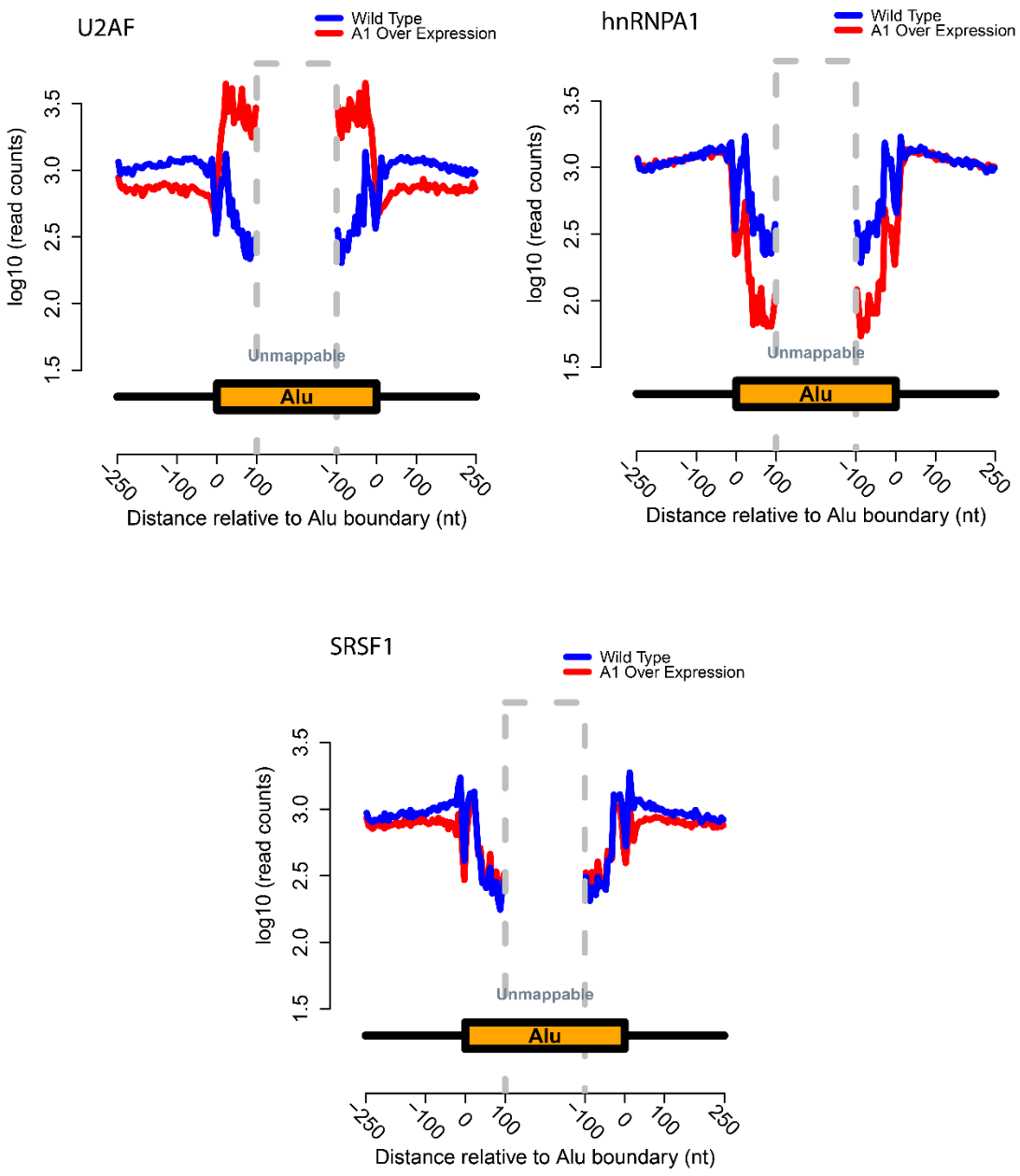

Figure 13 Aggregated read counts on Alu elements and nearby regions. Aggregated read counts on Alu elements and nearby regions for U2AF2 (left panel), hnRNPA1 (right panel) and SRSF1 (bottom panel). Blue represents wild-type binding of the given RNA binding protein and red represents hnRNPA1 overexpression of the log10 number of iCLIP read counts across all antisense-Alu elements.

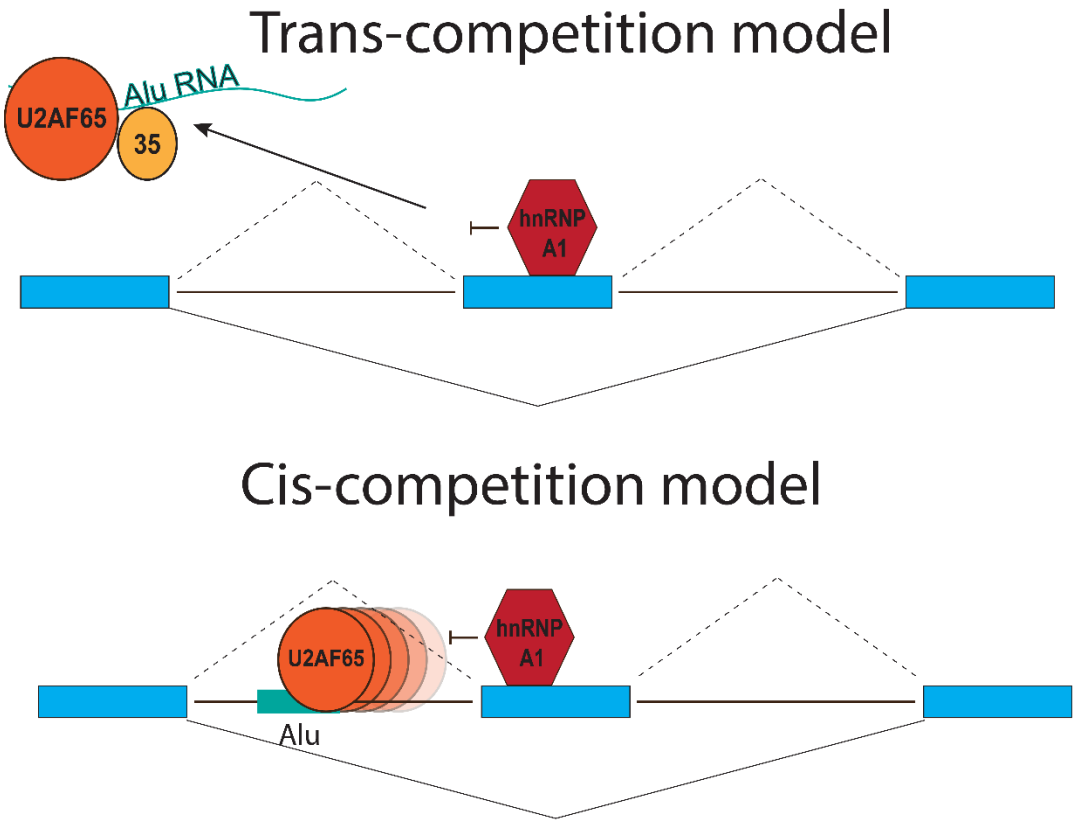

Figure 14 Potential regulatory mechanisms of Alu element. Model representing two potential modes by which U2AF2 may associate with Alu RNA: trans-competition suggests U2AF2 binds to Alu elements on other RNAs, while a cis-competition suggest U2AF2 binds to Alu elements within the same RNAs that a particular exon is associated.

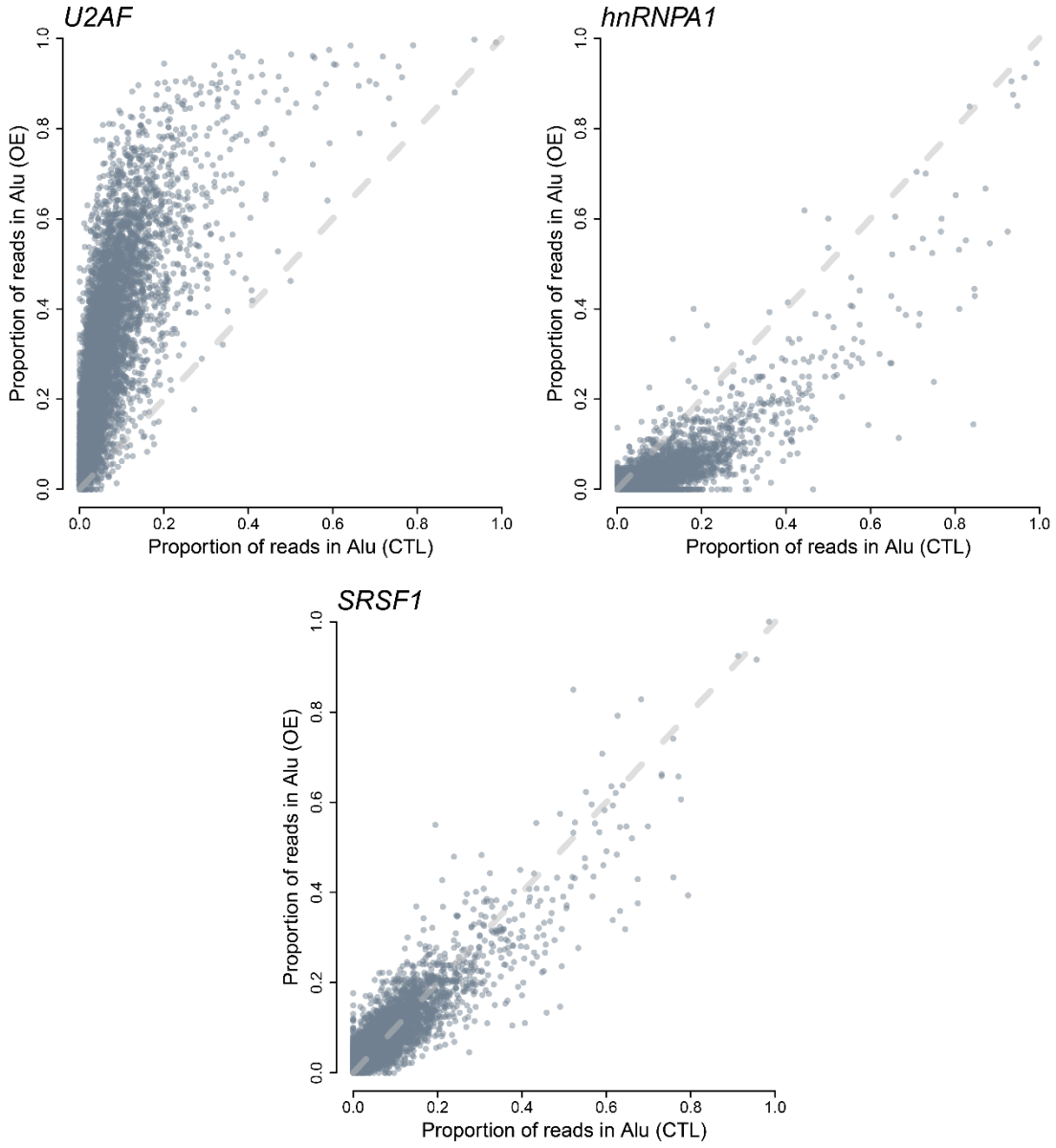

Figure 15 hnRNPA1 overexpression correlates with global redistribution of U2AF2 signal to Alu RNA elements. Scatter plot of all human cassette exons measuring the proportion of U2AF2 iCLIP crosslinks found within Alu elements within the cassette exon event over the total number of crosslinks found within the event. Proportions from control and hnRNPA1 overexpression samples are compared for each individual cassette exon event.

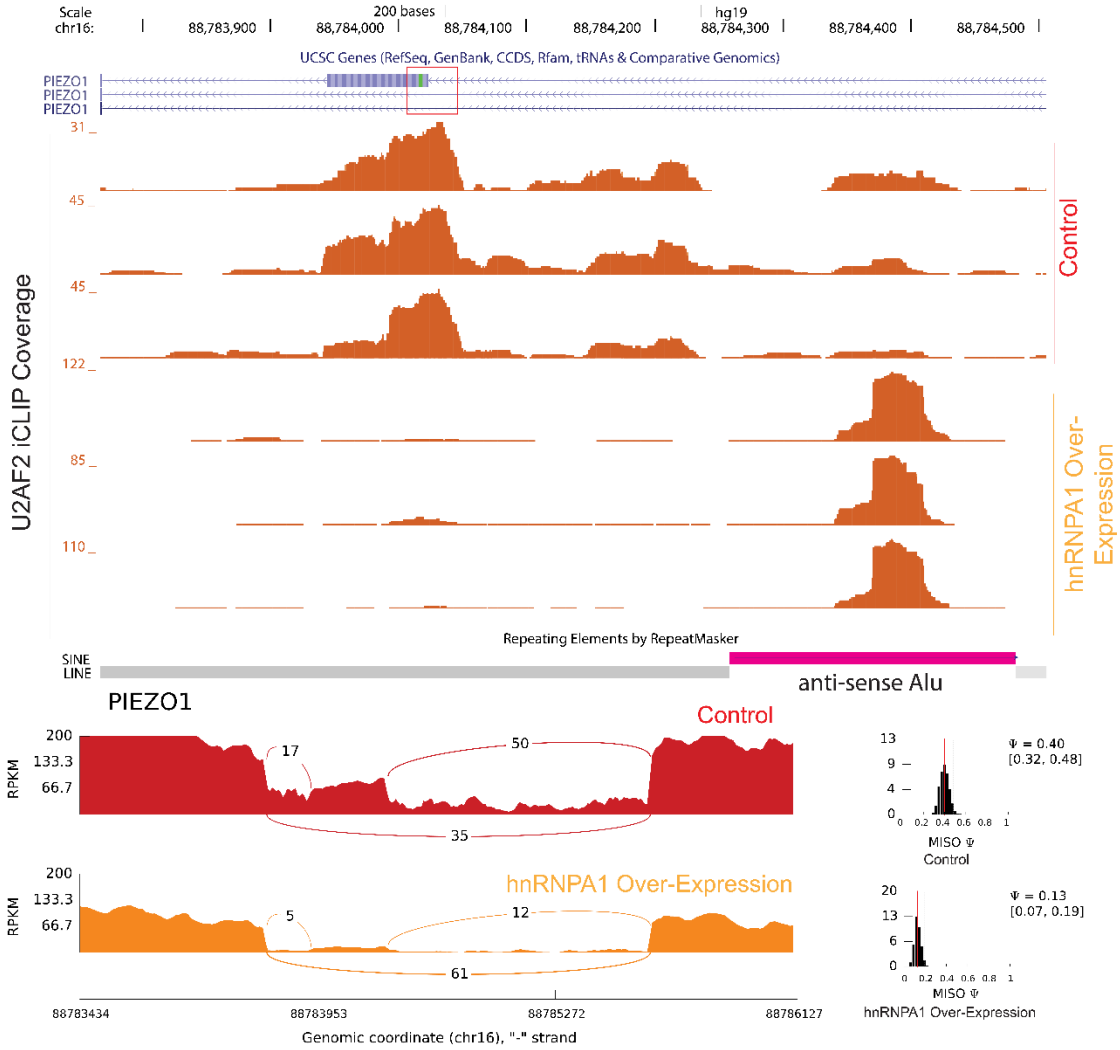

Figure 16 Example of hnRNPA1-dependent modulation of U2AF2-Alu interaction on PIEZO1. UCSC genome browser snapshot depicting U2AF2 redistribution to anti-sense Alu elements in PIEZO1. iCLIP data (read coverage) for U2AF2 in control and hnRNPA1 overexpression cell lines. Sashimi plot representing MISO analysis of RNA sequencing data from samples used for iCLIP. This plot for PIEZO1 represents splicing data corresponding to the exon show in the USCS genome browser snapshots.

3.3.4 Alu Elements May Function as Cis-regulatory Elements that Compete with Authentic Exons for Binding to Splicing Factors

Alu elements influence alternative splicing, although the mechanisms are poorly understood [7, 144, 157-159]. We investigated the positions of Alu-elements with hnRNPA1-dependent changes in U2AF2 crosslinking relative to the 3' splice site of constitutive or skipped exons. As expected, we observed that Alu elements are closer to skipped than constitutive exons ($p < 1.4e-47$, Wilcoxon rank-sum test, Figure 17, compare green and yellow boxes). But yet, those Alu elements with hnRNPA1-dependent increases in U2AF crosslinking are significantly closer to exons than those that are unchanged ($p < 9.5e-93$, Figure 17). Taken together our data suggest the intriguing hypothesis that Alu-elements may function as cis-regulatory elements that compete with authentic exons for binding to splicing factors.

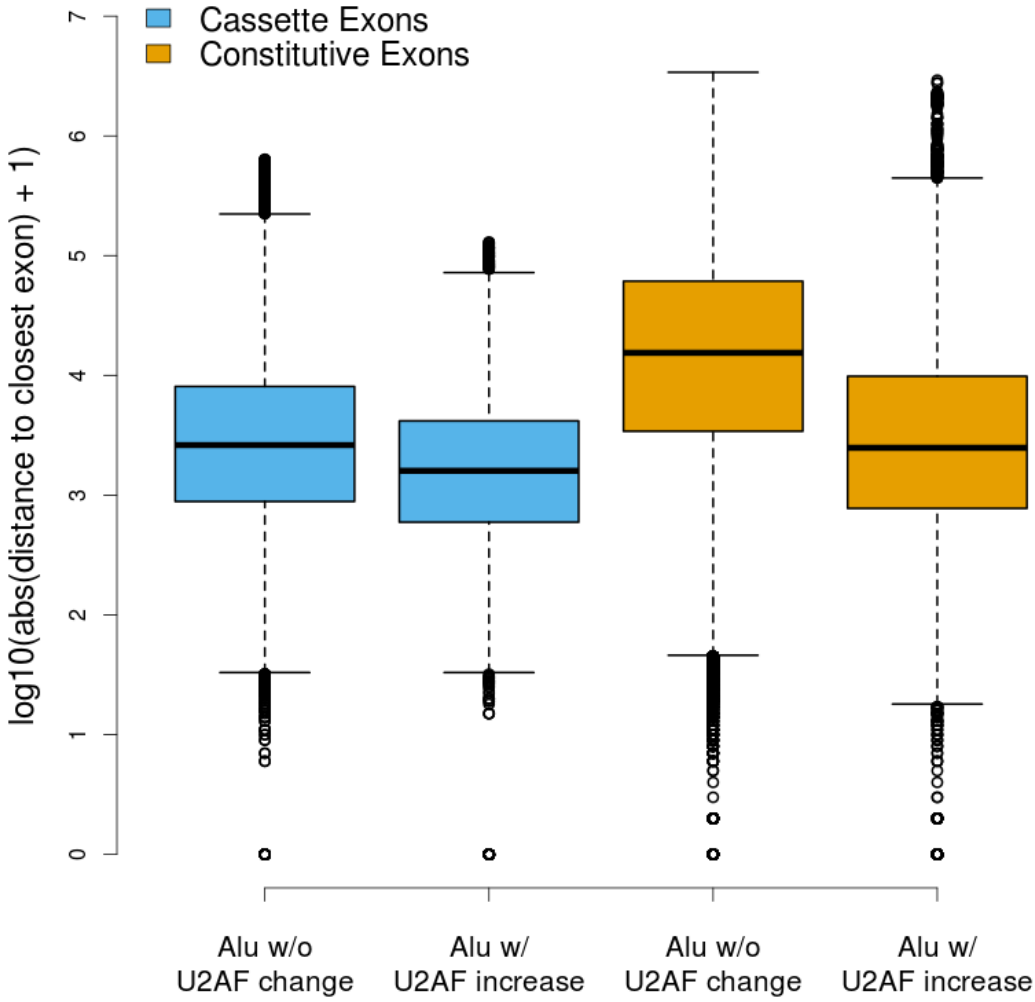

Figure 17 Distance of Alu elements to the closest exons. Box plot representing the distance of Alu elements from cassette exons (blue) and constitutive exons (orange) that show no change in U2AF2 cross-linking versus those that show an increase in U2AF2 crosslinking.

3.4 Discussion

Alu elements influence gene expression in diverse ways [144, 160-164]. The results presented here implicate Alu elements in splicing regulation. The proximity of hnRNPA1-responsive Alu-U2AF2 interaction sites to exons supports this hypothesis. Recently, Zarnack et al. demonstrated that hnRNPC competes with U2AF2 to repress inclusion of Alu-derived exons in mRNA [10]. We find that hnRNPA1 overexpression correlates with increased U2AF2 association with Alu-derived RNA sequences. This occurs with no change in hnRNPC protein expression or localization. We hypothesize that Alu elements function as a sink for U2AF2. In this model, U2AF2 dissociation from Alu-derived sequences maybe prevented by hnRNPA1. Alternatively, hnRNPA1 may alter U2AF2 RNA binding specificity thereby enhancing association with Alu elements. Taken together, our data demonstrate that Alu-derived sequences function as RNA regulatory elements that respond to changes to the intracellular concentration of splicing factors. Our data are also consistent with recent U2AF2 CLIP-seq results which identified a strong correlation between exon skipping and atypical intronic binding sites upstream of exons or within exons [89]. As a primate-specific retrotransponson elements, Alu may speed up human evolution by weakening the splice sites of certain constitutive exons. Such changes may enable them to be excluded from the final transcript, and therefore increased diversity of human proteome. Our results suggest the intriguing hypothesis that retrotransposons contribute to species-specific differences in alternative splicing throughout the primate lineage (Figure 18).

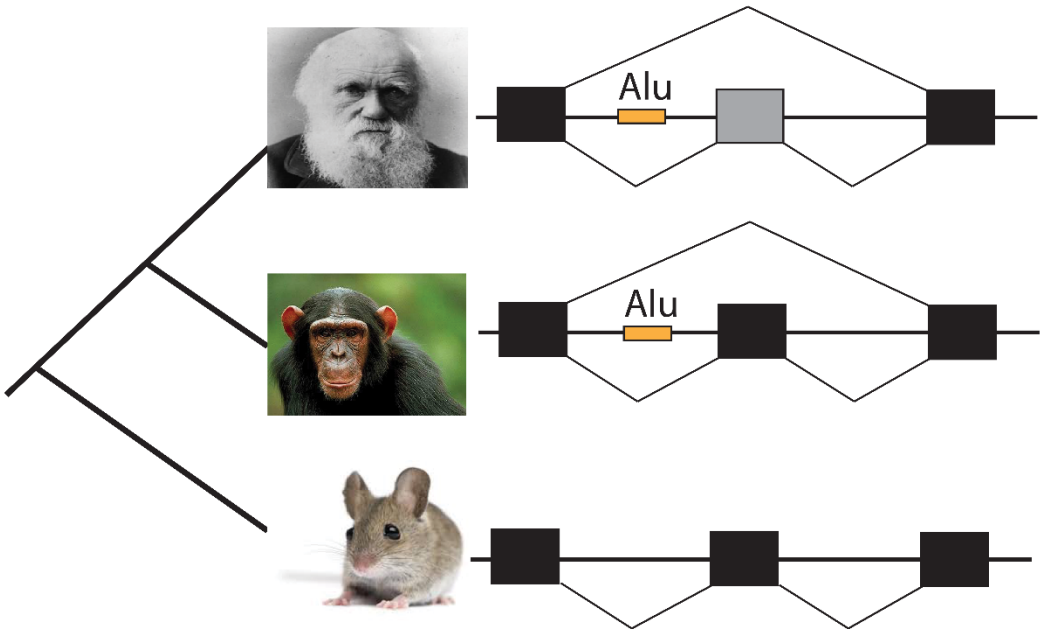

Figure 18 Schematic describing a role for Alu-elements in the evolution of primate-specific alternative splicing.

Chapter 4 regSNP-intron: A Tool for Prioritizing Intronic Single Nucleotide Substitution

4.1 Introduction

Human genome carries a large number of single nucleotide variants (SNVs) and many SNVs are believed to be involved in disease. While most studies focus on SNVs that locate in exonic regions, more and more studies suggest that intronic SNVs (iSNVs) may cause disease by affecting splicing regulation [12, 33, 36, 165]. Intronic SNVs typically affect alternative splicing by disrupting the recognition of splicing sites. For example, studies show that 99.5% cases in familial dysautonomia (FD) have the intronic mutation near 5'-splice site of exon 20 in IKBKAP gene, which can cause the exon skipping of exon 20 and result in the malfunction of IKBKAP [36, 38, 39]. Other evidence suggests that iSNVs can perturb the RBP binding affinity of cis-regulatory elements [165-167]. Since there are a larger number of iSNVs, but only a small fraction of them can potentially cause altered biological functions, it is important to be able to prioritize these SNVs using computational approaches. The selected variants with higher pathogenic potentials can be further subjected to additional experimental validation.

In order to effectively analyze the large number of iSNVs for further functional and clinical studies, efficient bioinformatics algorithms are needed for prioritizing iSNVs. One bioinformatics tool, SPANR (Splicing-based Analysis of Variants) [16], is designed for evaluating how individual SNVs impact splicing regulation, by predicting its maximum changes on the percentage of inclusion (dPSI) of the

nearby exons. It extracts over 1000 genomic features around SNVs and predict the potential splicing outcome by training a neural network with 16 tissues from human BodyMap 2.0 RNA-Seq data. This tool, however, is not designed for assessing whether the target iSNVs cause deleterious phenotypes since it does not further evaluating the level of the resultant splicing changes impacts the protein function. Another widely used tool, CADD (Combined Annotation Dependent Depletion), predicts pathogenic variants with support vector machine (SVM) by combining annotations from multiple sources including conservation scores such as PhyloP, regulatory information such as transcription factor binding, and protein-level scores such as SIFT and PloyPhen [134]. Regarding to iSNVs, however, the available annotations are limited to genomic features such as conservation score and distance to splicing sites. Our earlier studies on small insertions/deletions (INDELs) [17, 18], alternatively spliced exons [168], and synonymous SNVs indicate that simply including or excluding a stretch of amino acid residues does not guarantee an alteration on the protein function. This is also consistent with numerous reports that splicing variations can be “passenger” events, and thus inconsequential to the organismal phenotype [20]. Therefore, in order to fully appreciate the molecular implications of functional iSNVs, it is critical to integrate the protein structure-related features that determine the effects of alternatively spliced exons on protein structure and function.

In this study, we combined the information of both the impact of iSNVs on splicing regulation and the protein structural features of potentially altered exons. We

extracted pathogenic iSNVs from Human Gene Mutation Database (HGMD) and carefully selected a subset of iSNVs from 1000 Genome project as neutral SNVs [169, 170]. A random forest classifier [171] was built to prioritize the disease-causing probability of iSNVs. The classifier was also tested on the independent dataset from ClinVar database. The result suggests that our method can predict the disease-causing probability of iSNVs with high accuracy by combining the genomic features and structural features around iSNVs. As the price of whole-genome sequencing drops below 1000 dollars, there are more and more iSNV data available. Our study can help to prioritize and screen the potential pathogenic iSNVs for further experimental analysis and better understand the role of iSNVs in diseases. The tool is available at http://clark.compbio.iupui.edu/regsnp_intron_web. The pre-computed prediction result for all the possible iSNVs in human genome is also integrated in the ANNOVAR tool [172].

4.2 Materials and Methods

4.2.1 Data Set

To prioritize the pathogenic probability of iSNVs, we construct our training dataset by combining the known pathogenic SNVs from HGMD and neutral SNVs from 1000 genome project (Figure 19, Figure 20). For pathogenic iSNVs in HGMD, we only include the ones that labeled as “DM” (disease causing). Those are the variants that are manually curated to cause disease phenotypes through aberrant mRNA splicing. Neutral iSNVs were selected from the variants documented in the

1000 genomes dataset, which were derived from the genomic sequencing data of 2,500 individuals without apparent clinical phenotype. In order to minimize false negative records in our training dataset, we only select those iSNVs with a minor allele frequency (MAF) greater than 10%. This selection scheme results in 2,438 and 2,104,613 deleterious and neutral SNVs, respectively.

We further divided all selected iSNVs into two categories, the ones that are close to the splicing junction sites, or on splicing sites (on_ss), and the ones that are more distal into the intronic regions (off_ss). These variants are further separated since the variants in these regions may impact splicing outcome through different molecular mechanisms. Specifically, the variants close to the splicing junction sites (on_ss) may directly interfere with the splicesome formation, while the ones further away from the junction sites (off_ss) may impact the splicing regulation by affecting the binding of regulatory RNA binding proteins. Splicing sites are defined as upstream 3bp to downstream 7bp for donor sites and upstream 13bp to downstream 1bp for acceptor sites [173]. In total, 1,865 on_ss and 573 off_ss pathogenic iSNVs were extracted from the HGMD; 3,386 on_ss and 2,104,613 off_ss neutral ones were derived from the 1000 genomes database. As shown in Figure S2, comparing to the neutral variants, pathogenic variants tend to be closer to the splicing junction sites. To avoid the potential biases led by the differences of the distances from junction sits, we randomly selected 852 off_ss iSNVs from the 1000 genomes variants by matching the distance distribution of the HGMD variants.

This process ensured a more balanced training dataset with similar distance distribution between the pathogenic and neutral variants (Figure 21).

For each of the on_ss and off_ss dataset, we randomly select 2/3 of the data as training set to build a random forest classifier (Figure 21). The remaining 1/3 are used as validation set. To further test the performance of our model, we extracted pathogenic and benign iSNVs from ClinVar database as independent testing set [21]. In order to ensure the quality of variant disease association in the ClinVar data set, we only keep the iSNVs with at least 2 submitters, with the exception of pathogenic off_ss iSNVs where we only require single submitter due to the limited number of iSNVs of this category in the ClinVar database. In total, we extracted 121 on_ss and 51 off_ss pathogenic iSNVs, and 167 on_ss and 883 off_ss benign iSNVs from ClinVar.

In order to select the features that best distinguish the differences between pathogenic and neutral iSNVs, we classified all the features into three categories, genomic features characterizing how individual iSNVs affect splicing regulation, structural features evaluating how the inclusion/exclusion of alternatively spliced exons affect protein function, and evolution conservation. Most of these features show significant separation power between pathogenic and neutral iSNVs base on Wilcoxon rank-sum test (Figure 22, Table 7).

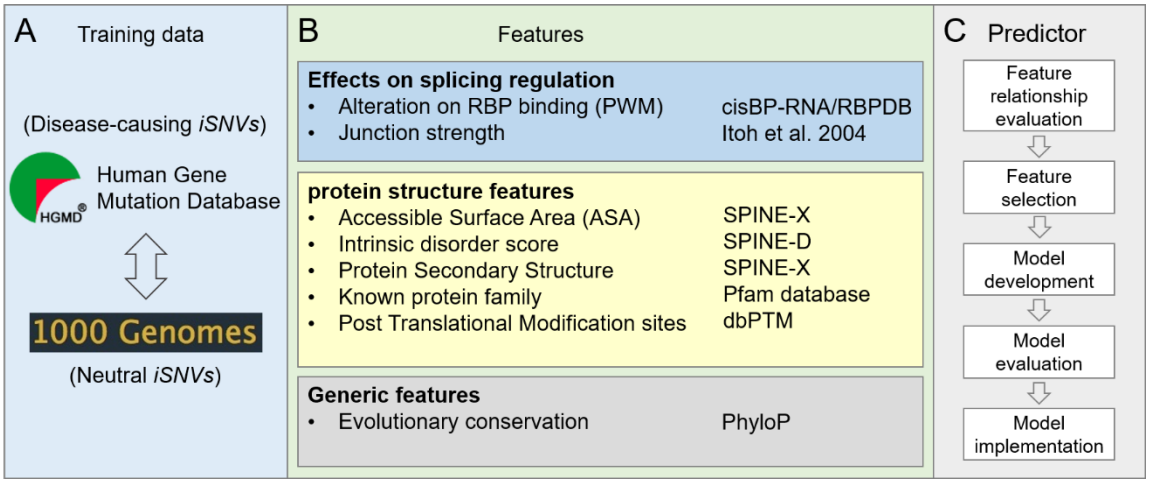

Figure 19 General workflow and feature sources.

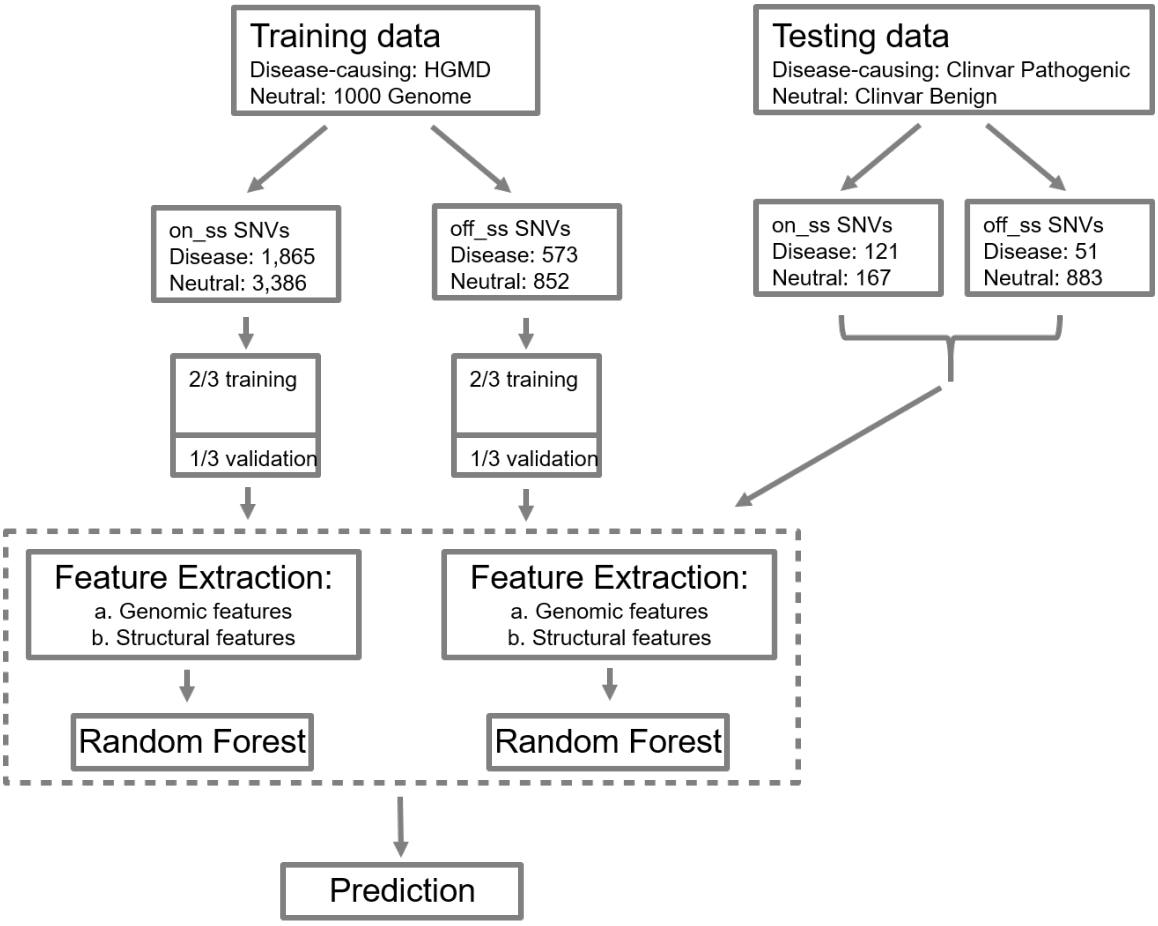

Figure 20 Model training and testing process. SNVs from 1000 Genome project and HGMD are divided into on_ss and off_ss iSNVs. Each category is then split into 2/3 for training and the rest 1/3 for validation. SNVs from Clinvar database are used for independent testing.

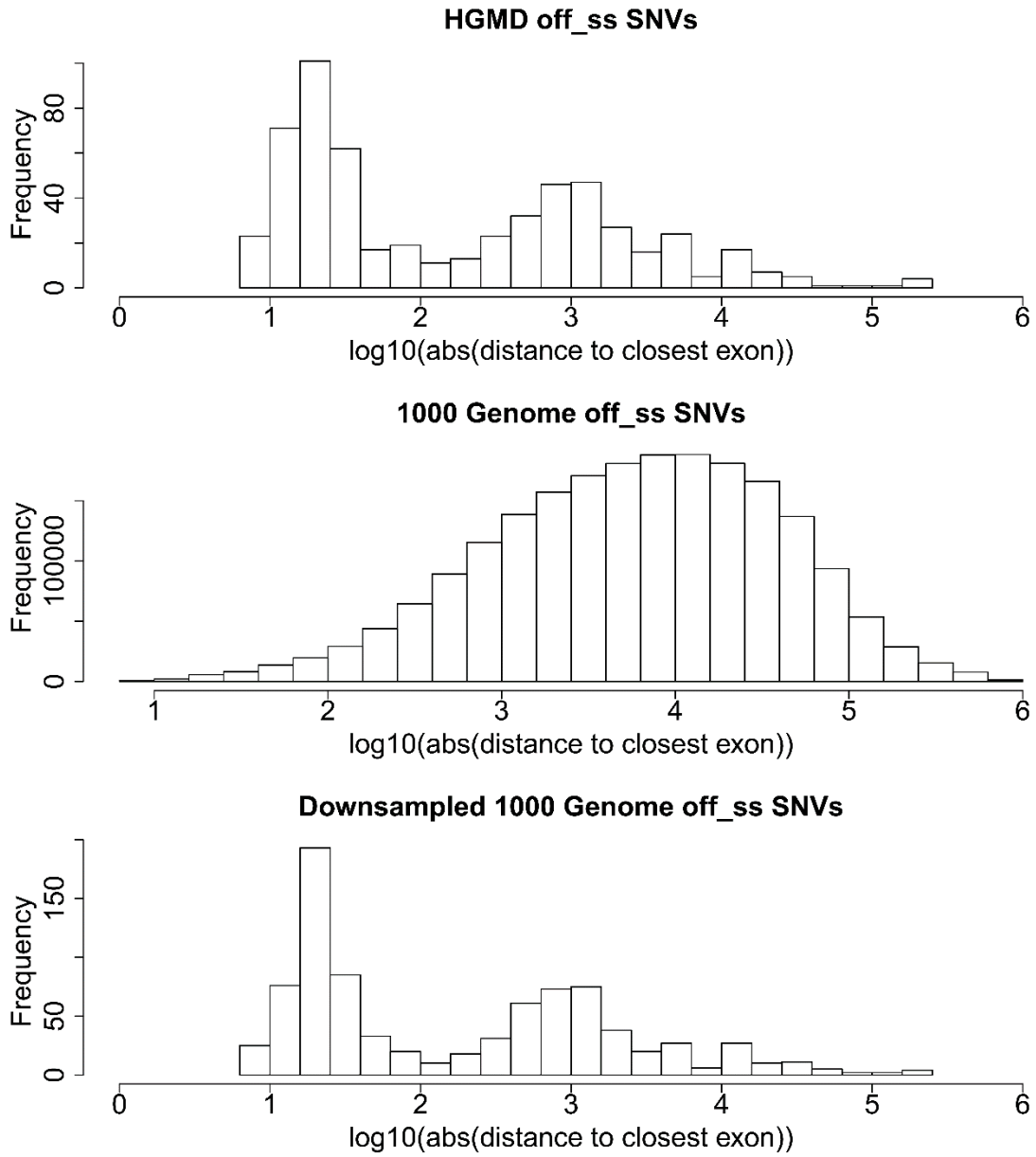

Figure 21 Histogram of distance between off_ss iSNVs and the closest exons. iSNVs from 1000 genome project are farther away from exon boundaries comparing to the iSNVs from HGMD. A down-sampling is needed to correct such bias.

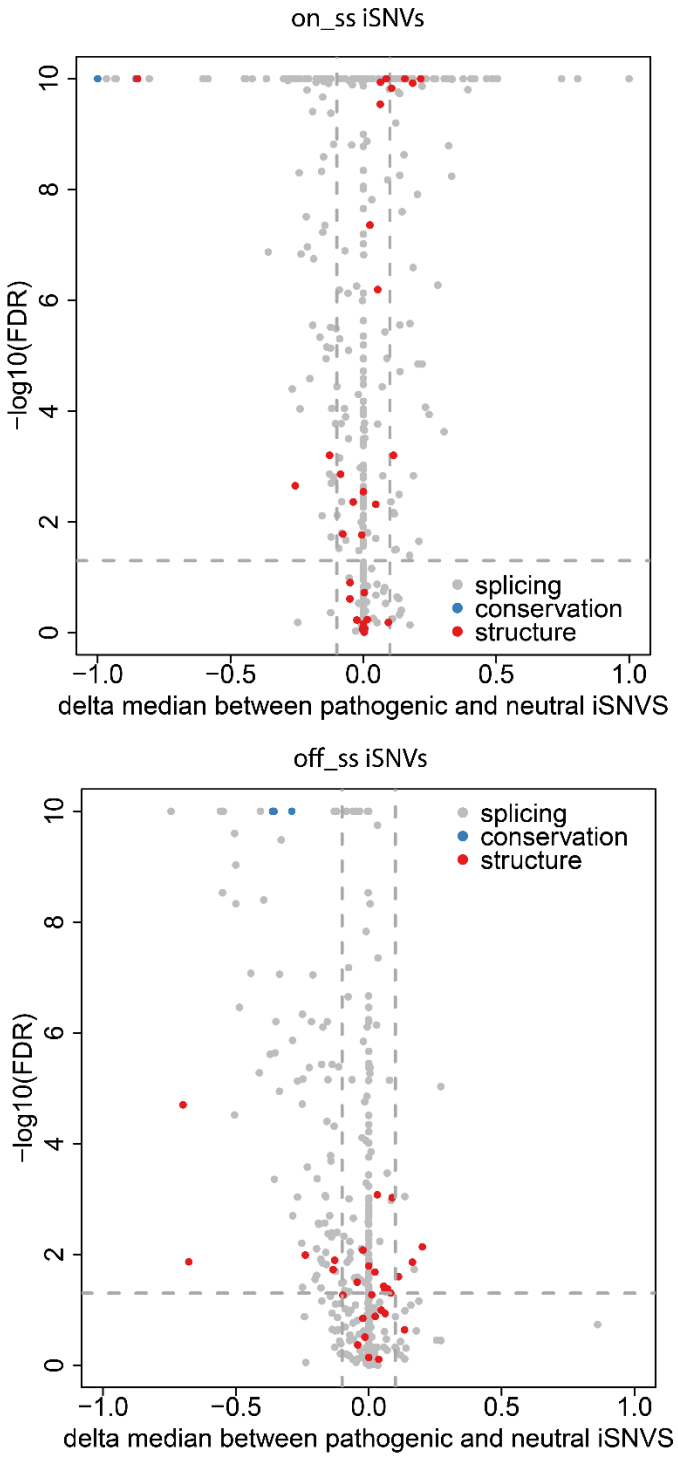

Figure 22 Volcano plot of all features. Each dot represents a feature in each of three categories: splicing (grey), conservation (blue), and structure (red). X-axis shows the difference of median value between pathogenic and neutral iSNVs. Y-axis represents adjusted p-value of Wilcoxon rank-sum test between pathogenic and neutral iSNVs.

4.2.2 Splicing Features

Junction strength of closest exon boundary is calculated for each iSNV based on the position weight matrices (PWMs) derived from canonical splicing sites [173]. The junction strength is measured by adding up information contents of positions -3 to +7 for donor sites and positions -13 to +1 for acceptor sites. Further, for on_ss iSNVs, the change of junction strength caused by allelic substitution is also computed.

The impact of iSNVs on RBP binding affinity is measured based on the PWMs obtained from RBPDB and cisBP-RNA databases [24, 174]. A total of 201 PWMs are collected from those two databases. The magnitude and posterior probability of RBP binding change are measured using methods described in previous study[17]. The matching score is calculated as:

$$S = \sum_{i=1}^{k} s_{i,j} \ , j \in \{A, C, G, T\}$$

where k represents the width of RBP binding site, and $s_{i,j}$ measures the logarithmic ratio between observed frequency and random background frequency:

$$s_{i,j} = log_2 \frac{(n_{i,j} + c_{i,j})/(N + \sum_{j \in \{A,C,G,T\}} c_{i,j})}{d_j}$$

Here, $n_{i,j}$ is the count of base j on position i in the PWM, $c_{i,j}$ is the pseudocount. N is the total number of binding sites used to derive the PWM. And d_j is the prior base frequency of base j (d_j = 0.25 for j = A, C, G, T).

The mean and variance of matching score distributions for binding and non-binding events are estimated as:

$$M_s = \sum_{i=1}^{k} \sum_{j \in \{A,C,G,T\}} f_{i,j} \times s_{i,j}$$

$$V_s = \sum_{i=1}^{k} \sum_{j \in \{A,C,G,T\}} f_{i,j} \times s_{i,j}^2 - (f_{i,j} \times s_{i,j})^2$$

where $f_{i,j}$ is the approximation of the true frequency of base j at position i. For binding events,

$$f_{i,j} = \frac{2^{s_{i,j}}}{4}$$

and for non-binding events, $f_{i,j} = 0.25$.

The magnitude of an iSNV affecting the RBP binding is defined as the log-likelihood ratio between alternative allele and reference allele:

$$M = log_2 \frac{P(S_A|B)/P(S_A|NB)}{P(S_R|B)/P(S_R|NB)} = log_2 \frac{\int_{-\infty}^{S_A} \frac{1}{\sqrt{2\pi V_s}} e^{-\frac{1}{2}\left(x-\frac{M_s}{V_s}\right)^2} d(x)}{\int_{-\infty}^{S_R} \frac{1}{\sqrt{2\pi V_s}} e^{-\frac{1}{2}\left(x-\frac{M_s}{V_s}\right)^2} d(x)}$$

Where S_R and S_A represent the matching score of sequence with reference allele and alternative allele. $P(S_A|B)$ and $P(S_R|B)$ denotes the probability of the given sequence being a binding site ,while $P(S_A|NB)$ and $P(S_R|NB)$ denotes the probability of being non-binding site.

A Bayesian-based posterior probability of RBP binding change is calculated as:

$$P = P(R = B, A = NB|S_R, S_A) + P(R = NB, A = B|S_R, S_A)$$

$$= \int_0^1 \frac{P(B)\big(1 - P(B)\big)\big(P(S_R|R = B)P(S_A|A = NB) + P(S_R|R = NB)P(S_A|A = B)\big)}{P(S_R)P(S_A)} d(B)$$

4.2.3 Evolutionary Features

Basewise conservation scores (PhyloP) of 99 vertebrate genomes with Human were downloaded from UCSC Genome Browser [175]. The scores on iSNVs loci as well as the average scores of +/-3bp and +/-7bp regions around the iSNVs were extracted and used in the machine learning model.

4.2.4 Protein Structure Features

Protein structural features of closest exons of iSNVs were evaluated. The protein disorder score, protein secondary structure and solvent accessible surface area (ASA) were precomputed for all the known protein-coding genes using SPINE-D and SPINE-X [25, 176]. The known protein domains were extracted from Pfam database [177]. The percentage of closest exon region overlaps with Pfam domains were measured. The post-translational modification sites (PTMs) were extracted from dbPTM 3.0 database [178]. The number of PTM sites per 100 amino acids on the closest exon were calculated.

4.2.5 Machine Learning Model

Two different random forest classifiers were built for on_ss and off_ss iSNVs respectively. Since random forest performs an implicit feature selection, only the important features will be selected to build each tree in the forest. The grid search

with 3-fold cross validation was used to fine-tune the hyperparameters such as number of trees and max depth. For on_ss iSNVs, 52 trees with depth equals to 13 were built. For off_ss iSNVs, the random forest contains 59 trees with depth equals to 20.

4.2.6 Allele Frequency of SNVs in GTEx

SNVs and corresponding allele frequencies were obtained from GTEx analysis v6 whole-genome sequencing data. We extracted iSNVs within 300bp of exon boundaries for prediction. 17,194 on_ss iSNVs and 630,557 off_ss iSNVs are left for further analysis.

4.3 Results

4.3.1 Pathogenic iSNVs Tend to Affect Alternative Splicing

Two sets of features were used to evaluate the impacts of individual iSNVs on splicing regulation: a) splicing junction score that quantifies splicing strength and b) how iSNVs to be predicted to affect the binding affinities of RNA-binding proteins (RBPs). Junction score is computed by position weighted matrices (PWMs) that measuring sequence features around canonical splicing sites. (Methods). The higher the score is, the more likely the corresponding exon is included in the final transcribed product. Our result shows that, while junction score change caused by neutral on_ss iSNVs can either increase or decrease (Figure 23), pathogenic on_ss iSNVs tend to decrease the junction strength of both donor and acceptor splicing sites. The median junction score change of pathogenic iSNVs (-2.96 for donor sites; -2.23 for acceptor sites) is significantly larger than the median score change of neutral ones (0.034 for donor sites; -0.059 for acceptor sites) (adjusted Wilcoxon rank-sum test p-value = $1.63*10^{-211}$ and $1.28*10^{-143}$ for donor site and acceptor site separately) (Table 7, Figure 23). This suggests that pathogenic iSNVs have a tendency to disrupt protein function by causing exon skipping.

To evaluate the impacts of iSNVs on RBP binding, we computed both the magnitude and the probability of matching score changes caused by iSNVs for 201 RBPs with known position weight matrices (PWMs, Methods). We show that pathogenic and neutral iSNVs showed significant differences in impacting the binding of many RBPs.

Among those 201 RBPs, a large proportion of them have significant separation power between pathogenic and neutral iSNVs (137 and 43 RBPs with adjusted Wilcoxon rank-sum test p-value < 0.05 for on_ss and off_ss iSNVs separately) (Figure 22).

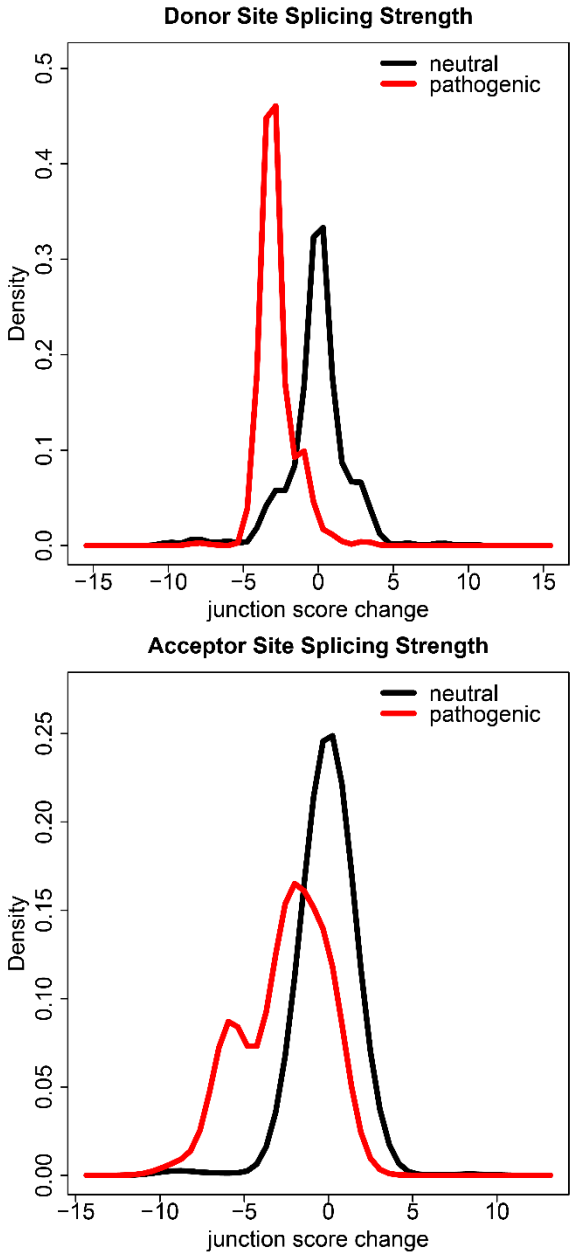

Figure 23 Distribution of junction score change caused by in pathogenic (red) and neutral (black) on_ss iSNVs. Top and bottom panels represent donor sites and acceptor sites separately.

Table 7 Wilcoxon rank-sum test of all features. Delta median is the difference between median values of neutral and pathogenic iSNVs.

	on_ss iSNV		off_ss iSNV	
Feature	Delta median	P-value	Delta median	P-value
Min disorder score	0.025	2.27E-008	0.033	1.77E-004
Max disorder score	0.156	7.30E-017	0.059	1.70E-002
Mean disorder score	0.186	8.89E-012	0.083	2.20E-002
Mean disorder structure region	0.003	9.76E-001	-0.041	3.00E-001
Mean disorder disorder region	0.087	6.60E-016	0.023	7.64E-003
Switch number	0.000	2.07E-003	0.000	6.26E-001
Min disorder length	0.064	8.78E-011	0.024	6.83E-002
Max disorder length	0.064	6.69E-012	0.062	6.00E-002
Mean disorder length	0.106	2.18E-011	0.047	5.04E-002
Min structure length	-0.127	4.31E-004	-0.044	1.27E-002
Max structure length	-0.038	3.27E-003	-0.128	4.25E-003
Mean structure length	-0.086	9.62E-004	-0.134	6.79E-003
Min secondary structure score	0.015	5.46E-001	0.071	1.75E-002
Max secondary structure score	-0.050	1.06E-001	0.135	1.38E-001
Mean secondary structure score	-0.025	5.56E-001	0.202	2.22E-003
Min alpha	0.046	3.63E-003	0.089	2.11E-004
Max alpha	-0.004	8.37E-001	0.012	2.40E-002
Mean alpha	-0.051	2.18E-001	0.112	9.54E-003
Min beta	0.006	8.18E-001	-0.014	2.03E-001
Max beta	-0.257	1.61E-003	-0.677	4.59E-003

Mean beta	-0.077	1.30E-002	-0.239	3.30E-003
Min coil	0.004	1.64E-001	-0.022	7.54E-002
Max coil	-0.006	1.35E-002	-0.022	2.59E-003
Mean coil	0.094	6.18E-001	-0.097	2.47E-002
Min ASA	0.054	3.54E-007	0.057	1.54E-002
Max ASA	0.113	4.31E-004	0.038	7.02E-001
Mean ASA	0.216	1.15E-013	0.165	4.72E-003
PTM	0.000	6.62E-001	0.000	5.76E-003
PFAM	-0.850	6.28E-017	-0.699	3.23E-006
Acceptor strength	0.187	1.40E-007	0.136	1.96E-004
Donor strength	0.209	1.79E-002	0.171	6.68E-003
Acceptor strength change	0.746	4.08E-145	NA	NA
Donor strength change	1.000	1.49E-213	NA	NA
PhyloP 1bp	-0.854	0.00E+000	-0.362	7.35E-041
PhyloP 3bp	-1.000	0.00E+000	-0.290	3.76E-025
PhyloP 7bp	-1.000	1.51E-284	-0.356	9.39E-028

4.3.2 Pathogenic iSNVs Happen at Conserved Regions

Previous studies show that evolutionary conservation can be an important feature in predicting the disease-causing probability of SNVs [131, 179]. Our analysis confirm this observation. We calculate the PhyloP 100-way conservation score on the locus of iSNV itself, 3bp region and 7bp upstream and downstream the iSNV. Our result show that the pathogenic iSNVs locate at the loci with significantly higher conservation scores than the neutral ones do (Figure 24). For on_ss iSNVs, the median conservation score of pathogenic iSNV loci (3.08) is 3.11 higher than the median score of neutral ones (-0.03) with adjusted Wilcoxon rank-sum test p-value $< 1.00 * 10^{-300}$. The large positive PhyloP scores on pathogenic iSNV loci suggest those loci evolve much slower than expected. For off_ss iSNVs, the median conservation score of pathogenic iSNV loci (0.31) is 0.59 higher than the median score of neutral ones (-0.28) with adjusted p-value $= 3.21 * 10^{-38}$. This indicates that the iSNVs happen at more conserved loci are more likely to be disease-causing.

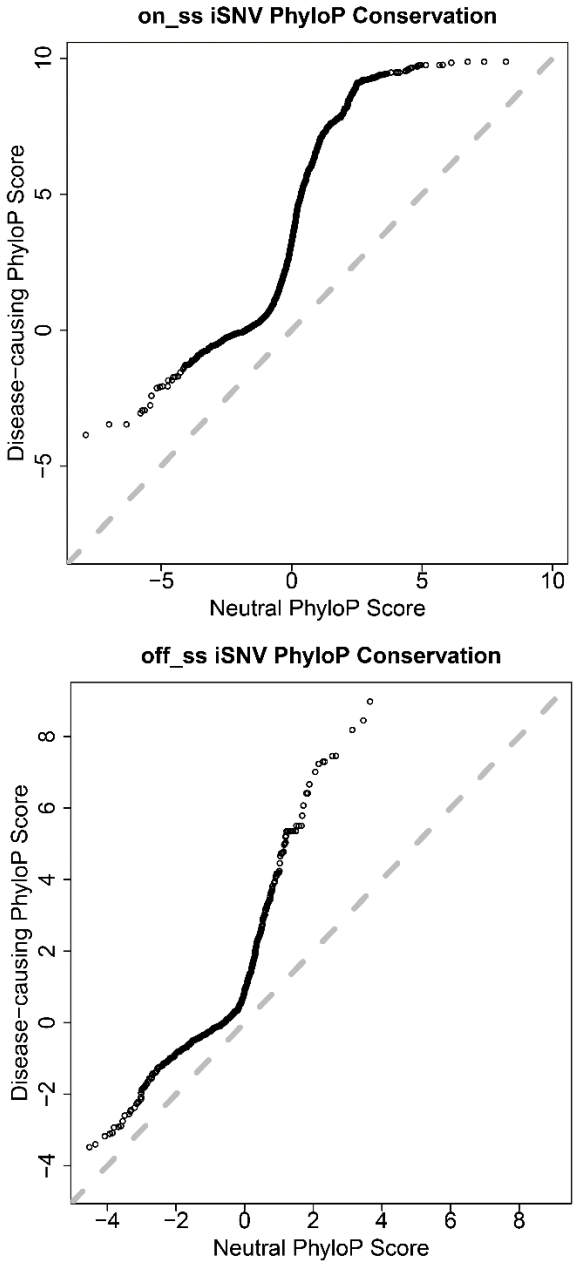

Figure 24 Q-q plot of PhyloP score between pathogenic and neutral iSNVs.

4.3.3 Pathogenic iSNVs Tend to Locate Close to Exons of Functional Important Regions

To further evaluate the impact of iSNVs on protein function, we calculate the structural features that characterize the functions of potentially alternatively spliced exons. We hypothesize that pathogenic iSNVs tend to disrupt the splicing of the exons in key protein structural regions. We capture the protein structural features of the closest exons of iSNVs such as the probability of being intrinsically disordered, secondary structure (probability of being alpha helix, beta sheet or random coil) and soluable accessible surface areas (ASA) (Table 7). We also calculate the percentage of affected exon region that overlaps with known protein domains and contains post-translational modification sites. (Table 7)

Our result suggests that, comparing to neutral iSNVs, exons that affected by pathogenic iSNVs have lower average disorder score and contain longer structured region (adjusted Wilcoxon rank-sum test p-value $2.03 * 10^{-11}$ and $4.97 * 10^{-2}$ for on_ss and off_ss iSNVs respectively) (Table 7, Figure 25). It indicates that pathogenic iSNVs have higher probability of locating close to the exons within structured regions. In addition, exons close to pathogenic iSNVs have significantly smaller average ASA score (adjusted p-value $2.90 * 10^{-13}$ and $1.03 * 10^{-2}$ for on_ss and off_ss iSNVs respectively), which indicates that they are more likely to be located in the core protein regions as opposed to the surface of the protein molecule. Moreover, the closest exons of pathogenic iSNVs tend to have higher percentage of residues overlaping with known protein family domains (adjusted p-

value $1.91 * 10^{-16}$ and $1.98 * 10^{-5}$ for on_ss and off_ss iSNVs respectively). All those results suggest that the exons affected by pathogenic iSNVs have a higher chance to reside in the functionally important protein regions. Therefore, including protein structural features can provide us additional information to predict the disease-causing probability of iSNVs.

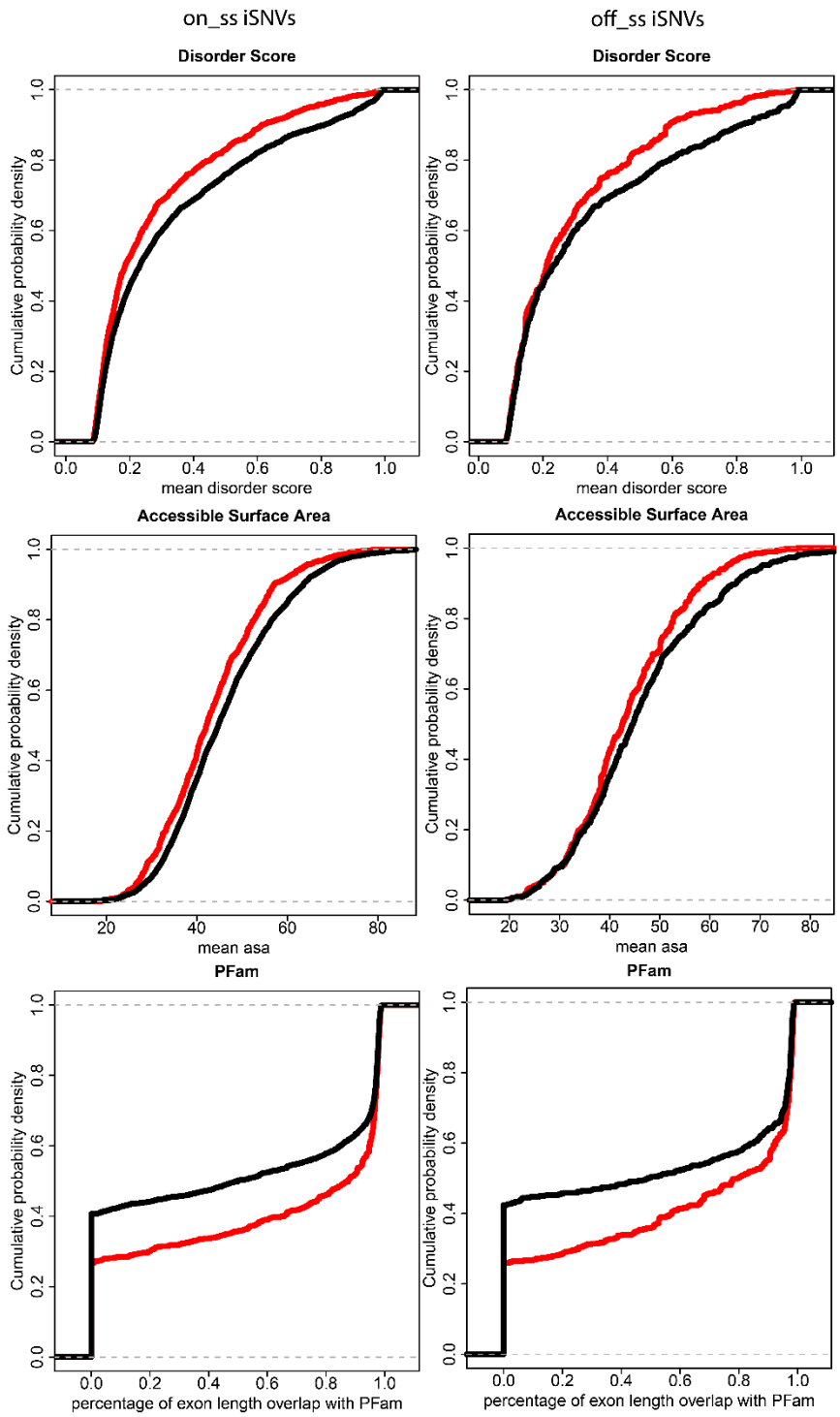

Figure 25 Empirical cumulative distribution of structural features. Red lines represent pathogenic iSNVs. Black lines represent neutral iSNVs. Top panel shows mean disorder score. Middle panel shows mean accessible surface ares (ASA). Bottom shows percentage of exon length overlap with PFam.

4.3.4 Prioritizing iSNVs Based on Their Impact of Splicing Regulation and Protein Structure

As stated above, since on_ss and off_ss iSNVs may affect splicing through different molecular mechanisms. While on_ss iSNVs may inhibit the assembly of spliceosome by decreasing the junction strength directly, off_ss iSNVs may affect the splicing by disrupting the RBPs binding to cis-regulatory elements. Therefore, we built two separate random forest classifiers for on_ss and off_ss iSNVs separately. The models are built on training set (2/3 of original dataset) and validated on validation set (1/3 of original dataset). Hyperparameters such as number of trees and maximum depth are grid searched by 3-fold cross-validation on training set. For on_ss iSNVs, 52 trees with maximum depth 13 are built. For off_ss iSNVs, the random forest contains 59 trees with maximum depth of 20.

Based on the validation dataset (1/3 of the original dataset) that was not used in model training, our algorithm reaches AUROC 0.96 and Matthews correlation coefficient (MCC) 0.79 for on_ss iSNVs. This significantly outperformed SPANR with AUROC 0.77 and CADD with AUROC 0.81. For off_ss iSNVs, the AUROC is 0.84 with MCC 0.52. As a contrast, the AUROC for SPANR and CADD are 0.54 and 0.69 separately (Figure 26). Since SPANR focuses on evaluating the impact of variants on splicing regulation and CADD mainly uses genomic features such as conservation scores and distance to splicing sites for intronic variants, our results suggest that inclusion of the protein structure features of the affected exon significantly increases the model performance.

To further evaluate the prediction power of features related to splicing, conservation and structure features, we built separate models based on each of the three categories of features. For on_ss iSNVs, the AUROCs for splicing, conservation, and structure features are 0.92, 0.92, and 0.72 respectively (Figure 27). For off_ss iSNVs, the AUROCs are 0.75, 0.68, and 0.63 for each category. These results demonstrate that each category of features provide important information in model prediction. The combination of all three types of features can help us reach the highest prediction performance.

To control the false positive rate (FPR) of the prediction results, we report the iSNVs with FPR < 0.05 as “Damaging”, iSNVs with 0.05 <= FPR < 0.1 as “Possibly Damaging” and iSNVs with FPR >= 0.1 as “Benign”. As shown in Figure 3, the reported “Damaging” category (FPR<=0.05) can reach true positive rate (TPR) of 0.85 and 0.45 for on_ss and off_ss iSNVs, respectively (Figure 26), while the sensitivity for the “Possibly Damaging” category is 0.90 and 0.52 for on_ss and off_ss iSNVs.

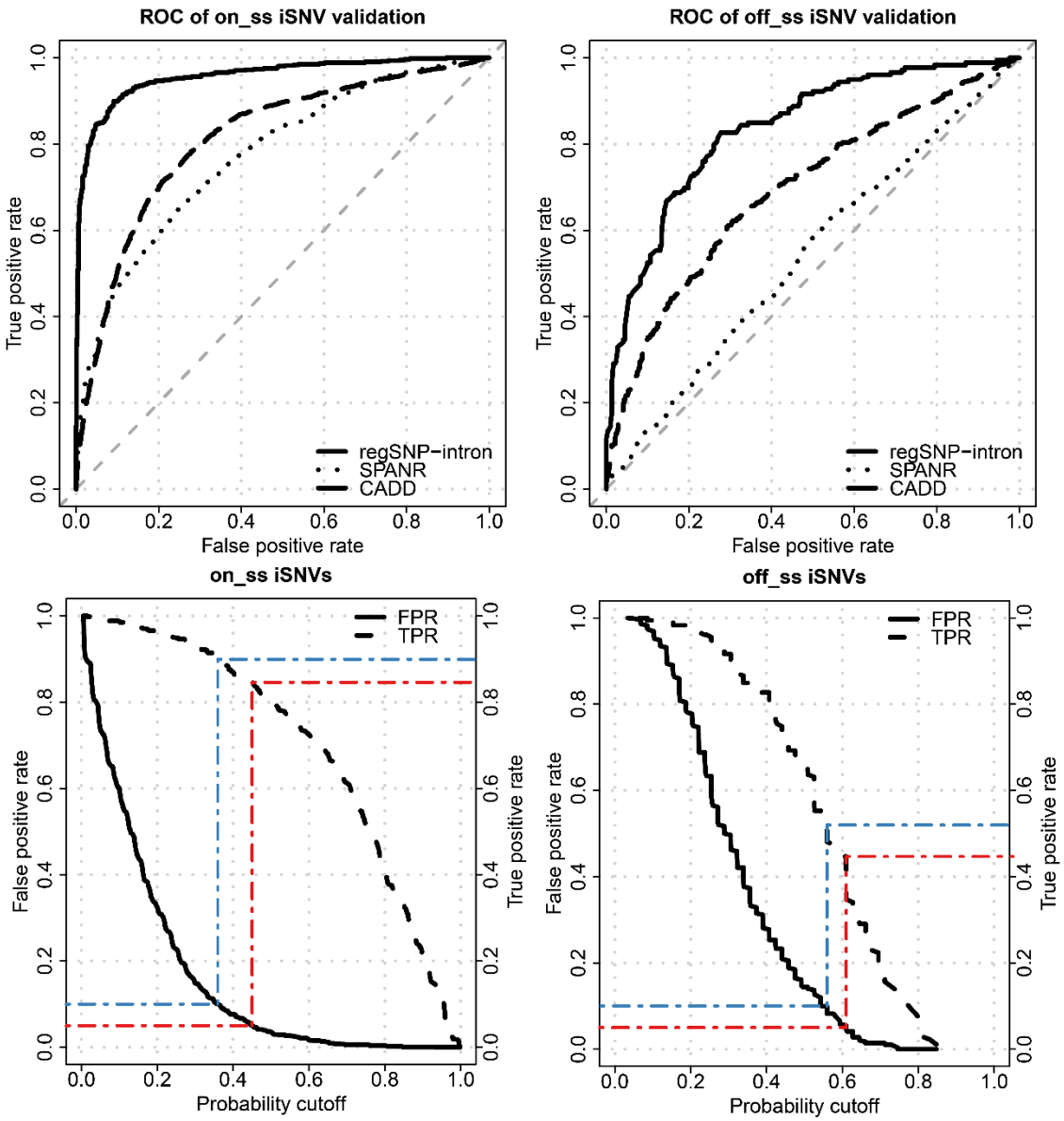

Figure 26 ROC curves on validation set. Top panel shows the comparison of ROC among regSNP-intron, SPANR, and CADD. Bottom panel shows the corresponding probability cutoffs and TPR when FPR = 0.05 (red) and FPR = 0.1 (blue).

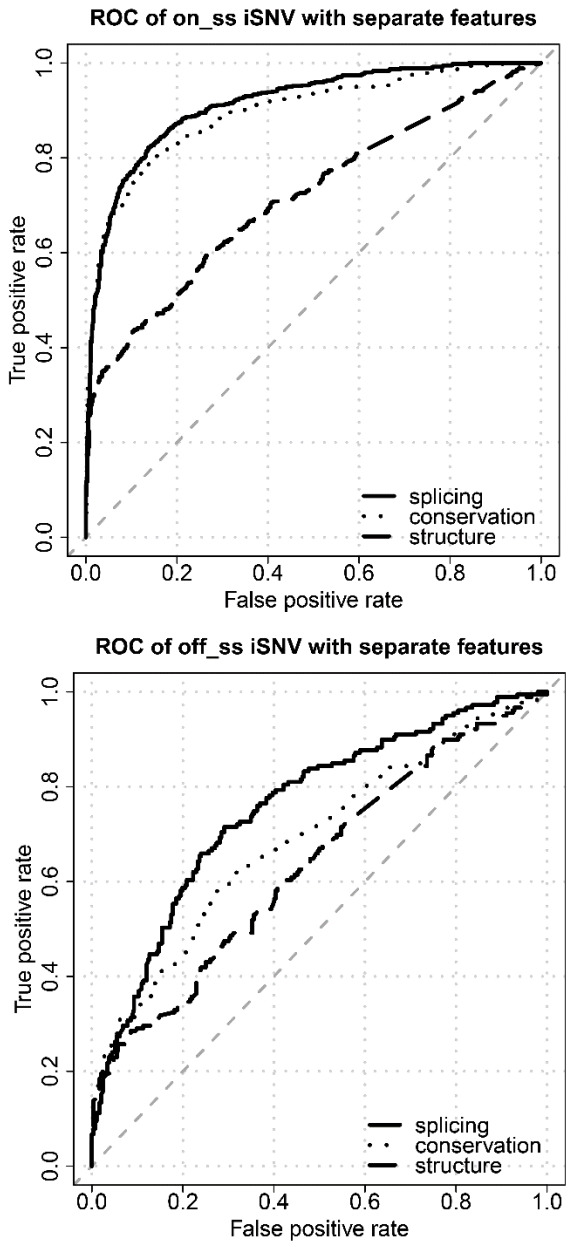

Figure 27 ROC curves of models built on three categories of features separately.

4.3.5 Evaluating Model Performance on Independent Testing Set

We further evaluated our prediction performance with an independent test data set from ClinVar database. We collected 121 on_ss and 51 off_ss pathogenic iSNVs, and 167 on_ss and 883 off_ss benign iSNVs with at least two submitters from ClinVar as our independent testing set. All the ClinVar iSNVs which are also observed in HGMD and 1000 Genome projects were excluded from the training set during the training stage to avoid overfitting. Consistent with the results on validation set, our model also shows a better performance comparing with splicing effect predictor (SPANR) on the test set (Figure 28). The AUROC of regSNPs-intron are 0.96 and 0.95 for on_ss iSNVs and off_ss iSNVs, while the AUROC of SPANR are 0.89 and 0.72 for on_ss iSNVs and off_ss iSNVs separately. These results suggest that our model has stable performance with high prediction accuracy over different data set.

It is interesting to see the performance of both our method and SPANR are improved on the independent test set comparing to validation set. The potential reason might be that there are greater differences on specific features, such as evolution conservation, that are more likely to drive the separation between the pathogenic and benign iSNVs (Figure 29). Such features are likely to be one of the criteria that are used by the ClinVar data contributors. Despite of this potential bias, the ClinVar dataset can still serve as an independent test set to compare the relative performance among prediction algorithms, even though we think the absolute measurement of AUROC on ClinVar data set might be overly optimistic.

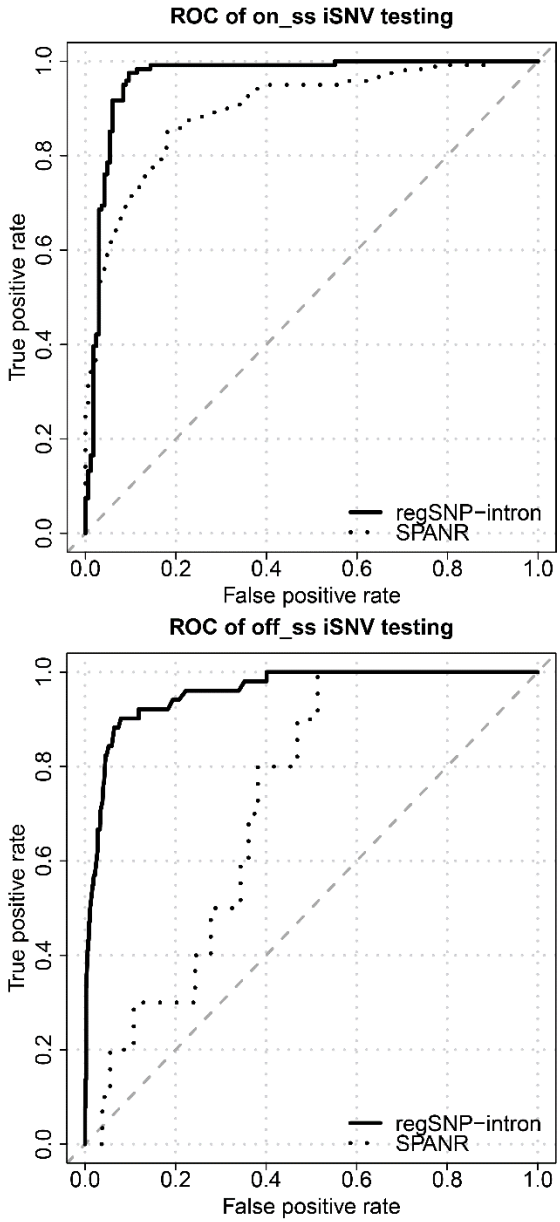

Figure 28 ROC curves on independent Clinvar testing set.

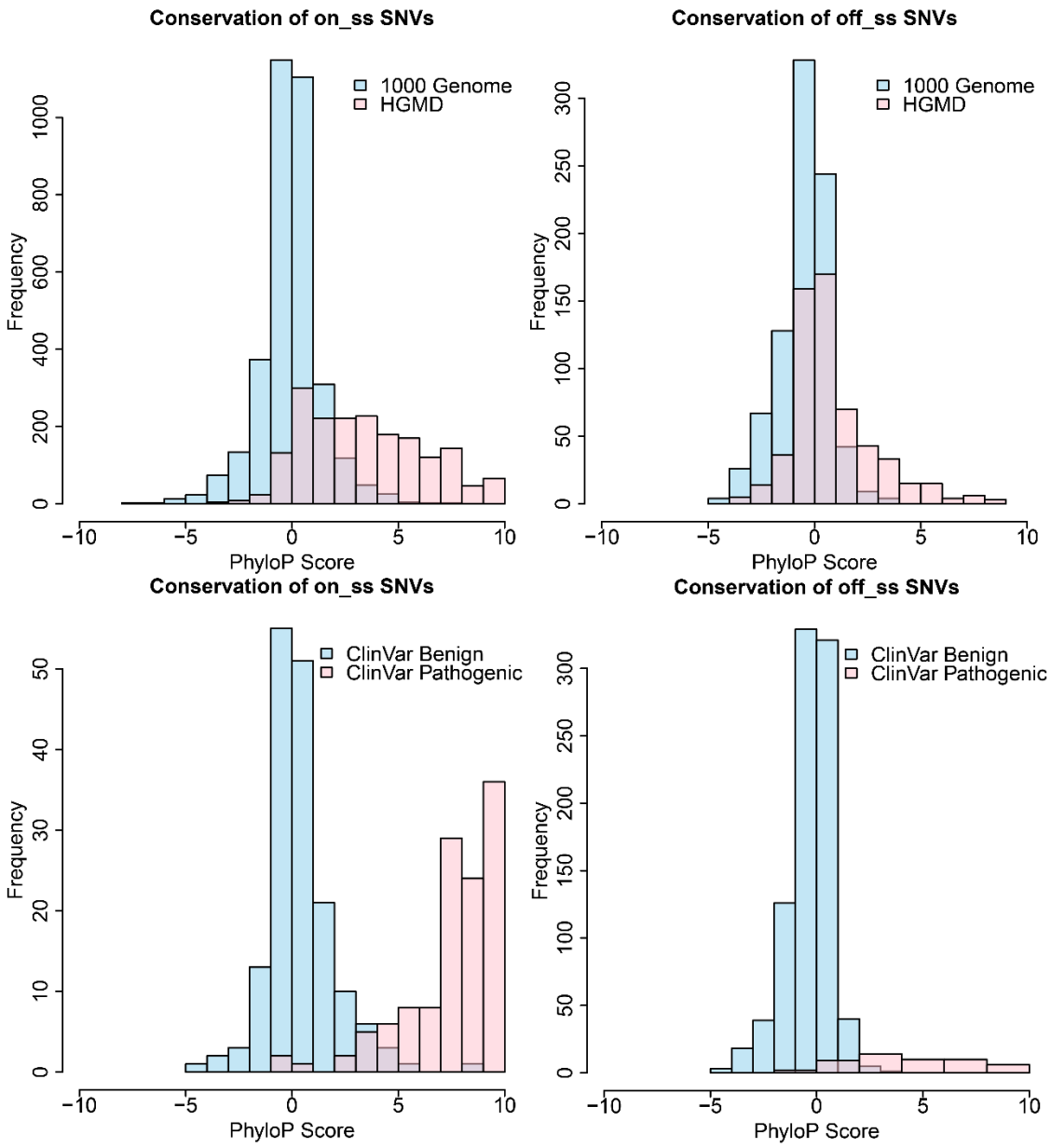

Figure 29 Distribution of PyhloP scores on training and testing set. PhyloP scores are more separable between pathogenic and benign iSNVs in Clinvar database (bottom) comparing with the ones in HGMD and 1000 genome (top).

4.3.6 Allele Frequency is Reversely Correlated with Disease-causing Probability

In general, the allele frequency in the population should reflect the importance of allele's biological function [180-184]. We hypothesize that the disease-causing probability of iSNVs obtained from Genotype-Tissue Expression Project (GTEx) using our model. GTEx project contains high quality genotyping information of 449 healthy individual. Here GTEx data is used independent of 1000 Genome project to avoid overfitting. The iSNVs are divided into 20 bins based on their allele frequency. And the average disease-causing probabilities are calculated for each bin. As expected, we observed a strong negative correlation between allele frequency and disease-causing probability for both on_ss iSNVs (R^2 = 0.47) and off_ss iSNVs (R^2 = 0.90) (Figure 30). This observation suggests that the prediction of our model is consistent with previous studies which concluded that variants with high disease-causing probability are less likely to happen in the population [18].

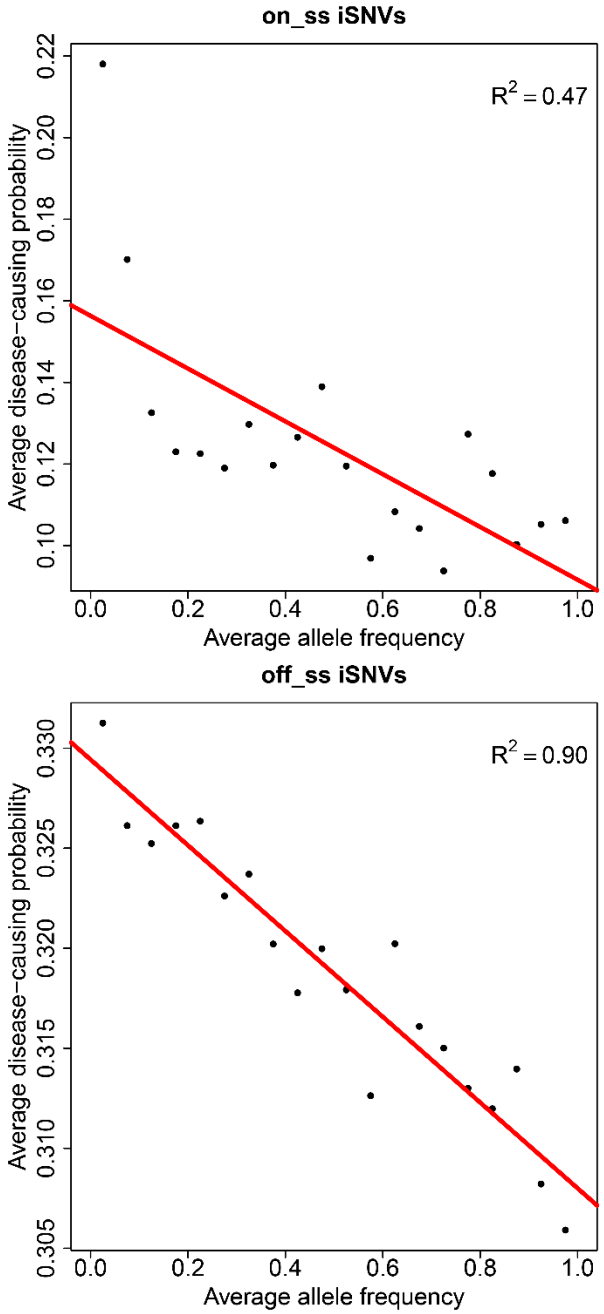

Figure 30 Correlation between average predicted disease-causing probability and average allele frequency in GTEx whole-genome sequencing data.

4.3.7 Pathogenic iSNVs Tends to Happen Near Exons which Can Tolerate Less iSNVs

We further evaluated the prediction results of our model by investing the functional importance of nearby exons which are potentially affected by iSNVs. We hypothesize that functionally important exons can tolerate less iSNVs nearby, and therefore iSNVs are more likely to be predicted as pathogenic if they locate near functionally important exons. To test this hypothesis, we extracted 160,230 exons which have at least one iSNV within ±300bp of intronic regions based on GTEx whole-genome sequencing data. One iSNV is randomly selected per exon and the disease-causing probabilities are predicted using our model. Based on our hypothesis, iSNVs should have higher disease-causing probability if the affected exons have less iSNVs nearby, and vise versa. Our result confirmed this negative correlation between average disease-causing probability and the number of iSNVs around exons (R^2 = 0.61, p-value = 0.007) (Figure 31). This indicates that pathogenic iSNVs tend to happen near functionally important exons.

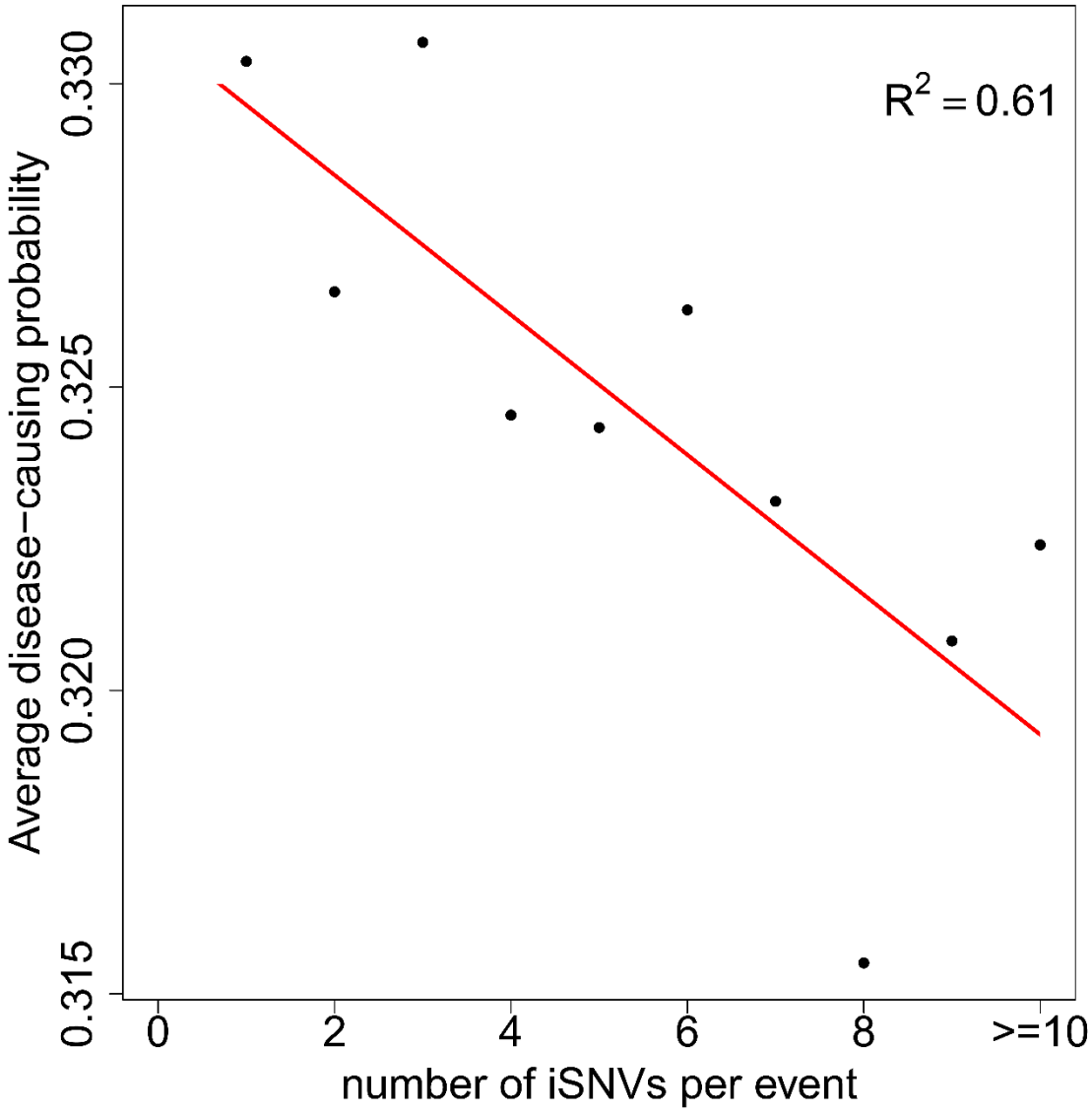

Figure 31 Correlation between average predicted disease-causing probability and the number of iSNVs around exons. Pathogenic iSNVs tend to locate near exons which can tolerate less iSNVs.

4.4 Discussion

Human genetic variants are known to be associated with many diseases. While next-generation sequencing technology enables exploring variants across the whole genome, many studies still focus on exonic regions. The main reason is that the variants locate outside the protein-coding regions are difficult to interpret and validate. However, introns, as the largest source of genetic variants in gene region, contain approximately a hundred times of more variants comparing to exons [185]. Cumulative evidence indicates that variants locating inside the intronic regions may have pathogenic effects through affecting splicing regulation. For instance, over 3,000 iSNVs in the Human Gene Mutation Database are documented to impact splicing, and further causes disease [169]. These iSNVs may affect splicing regulation through different mechanisms depending on the regions they locate in. Intronic SNVs that locate on the junction boundaries may have severe impacts on splicing by disrupting the recognition of splicing sites directly. For example, evidence shows that the donor site variant in intron 4 of adenomatous polyposis coli (APC) gene can lead to the initiation of colon cancer by causing the exon skipping of exon 4 [34, 51]. On the other hand, the iSNVs inside the intronic regions may have moderate effects which weaken the binding affinities of RNA-binding proteins that facilitate the splicesome formation. For instance, studies show that a G -> A substitution within an intronic splicing enhancer (ISE) downstream of exon 3 in growth hormone gene (GH1) can cause familial isolated GH deficiency type II (IGHD II) by suppressing the binding of splicing factors [186-188]. Similarly, an intronic A -> T substitution 32bp downstream of BRCA1 exon 22 can inhibit the

splicing and result in the exon skipping in breast cancer [165, 189]. Due to the large number of intronic variants to be detected in the whole genome/exome sequencing projects and their complex regulatory mechanism in affecting splicing regulation, a computational approach is needed to prioritize the potential functional impacts of intronic variants detected by high-throughput assays.

Although large-scale intronic variants are typically identified from the whole-genome sequencing (WGS) experiment, a proportion of intronic variants, especially for the ones that are close to the splicing junctions, can be captured in the whole-exome sequencing (WES) experiments when they are locating in the flanking regions of the exon. Since the price of WES is relatively cheaper than WGS, many studies utilize WES to allow sequencing more samples. Hence, it is important to survey how many intronic variants can be identified through WES for further computational prediction and experimental validation. To evaluate this, we collected all the genetic variants from the exome aggregation consortium (ExAC). ExAC is an integrated data set which contains high-quality exome sequencing data of 60,706 unrelated individuals from a variety of large-scale sequencing projects such as NHLBI exome sequencing project (ESP) and 1000 genome project [190]. Among 7,908,659 SNVs documented in ExAC, over 50% of them (4,126,724) are locate within intronic regions. Excluding those iSNVs from downstream analysis would not only waste a large proportion of data generate from WES, but also limit the ability to identify functional important non-coding variants. Thus, the algorithm

we developed in this paper can server as a valuable component in the variant analysis of whole-exome sequencing projects.

In this study, we evaluated the impact of genetic features and structural features on prioritizing pathogenic iSNVs. The disruption of variants on splicing regulation and the conservation of variants are already known to be powerful predictors of pathogenic SNVs in exon regions. In fact, such features have been used to evaluate impacts of synonymous variants in alternating splicing outcome [16], and whether they can lead to pathological phenotypes. In addition to these two types of features (impacts on splicing and evolution conservation), our previous study showed that the protein structural features can help us prioritize pathogenic micro-insertions and deletions [18], as well as dys-regulated alternative splicing events [168]. Here, we demonstrated that integrating protein structural features can provide additional information in prioritizing the pathogenic effects of intronic variants, for both on- and off-splice site variants. This is consistent with our observation that pathogenic iSNVs tend to locate near exons within functionally important regions. As a result, by carefully categorizing iSNVs into two subtypes, we successfully build two models to prioritize disease-causing on_ss and off_ss iSNVs with low FPR and relative high TPR. Although tools like SPANR can be used to predict tissue-specific splicing changes induced by individual iSNVs, to our knowledge, there is no bioinformatics algorithm specifically designed for prioritizing pathogenic iSNVs. In conclusion, our method allows effectively prioritizing and screening pathogenic iSNVs generated by genome-wide scale genetic studies.

Chapter 5 Conclusions and Discussions

5.1 Conclusions

5.1.1 Conclusions on RBP Regulatory Network and Alu Emelemts

As one of two key components of splicing regulation, trans-regulatory RBPs play critical roles in the regulatory machinery. However, among hundreds of splicing-related RBPs, the majority of them are under-examined. In this project, we systematically investigated the interaction of three key splicing regulators: hnRNP A1, SRSF1, and U2AF in HEK293 cell line based on the high-throughput sequencing data generated by our collaborators. We observed that the overexpression of hnRNP A1 can trigger the transcriptome-wide redistribution of SRSF1 and U2AF. More specifically, the hnRNP A1 overexpression can inhibit the binding of SRSF1 and U2AF near 3' splicing sites. This indicates that hnRNP A1 can suppress the assembly of spliceosome by interrupting the SRSF1-mediated U2AF recruitment. In addition, we observed that the significant enrichment of U2AF in Alu-derived transcripts after hnRNP A1 overexpression. Our results show that Alu elements may act as cis-acting regulatory elements that compete with authentic exons for binding to U2AF. These results can help us understand the association among key RBPs in splicing regulatory machinery. The discovery on the function of Alu elements in splicing regulation can also provide meaningful information for understanding the role of Alus in primate evolution.

5.1.2 Conclusions on iSNV Prioritization

Intronic variants are known to be related to many human diseases. However, due to the limitation of available data, there are few of algorithms developed to predict the disease-causing probability of intronic variants. As the increasing application of high-throughput sequencing technology, intronic variants start to attract more and more spotlights. In this study, we evaluated the impact of genetic features and structural features on prioritizing disease-causing iSNVs. The disruption of variants on splicing regulation and the conservation of variants are already known to be powerful predictors of pathogenic SNVs in exon regions. However, the alternation on pre-mRNA splicing does not necessarily mean the disruption on protein functions and phenotypes. Whether the potentially altered exon locates in key structural regions also plays a critical role on the functional impacts. Our previous study also shows that the protein structural features of alternatively spliced exons can increase the prediction power on pathogenic micro-insertions and deletions, as well as synonymous SNVs. Here, we demonstrated each category of features can be used to predict the disease-causing probability of iSNVs with a reasonably high accuracy. This suggests protein structural features can contribute an additional layer of information in predicting the disease-causing probability of iSNVs. By carefully categorizing iSNVs into different subtypes, we successfully built separate models to prioritize pathogenic on_ss and off_ss iSNVs with low false positive rate (FPR) and relative high true positive rate (TPR). With all the features combined, our model significantly outperform the widely used algorithm such as SPANR and CADD. As a result, our method allows effective prioritizing

and screening of pathogenic iSNVs generated by high-throughput platform such as whole-genome sequencing and SNP array.

5.2 Future Directions

5.2.1 Future Directions on RBP Regulatory Network

To further understand the relationship between RBPs in splicing regulatory network, a machine learning approach can be used to study the relationship among different RNA binding proteins. The relationship among those proteins will be analyzed systematically based on different genomic context. It will help people to understand the function of those RNA binding proteins in alternative splicing regulation. The effects of each RNA binding proteins are combined together to predict the splicing outcome more precisely.

After the comparison of iCLIP data of each RBP between control and hnRNP A1 overexpressed cell line, a series of features such as peak changes near splicing sites will be collected. A Bayesian network will be built to interpret the relationship among different RNA binding proteins as well as predict the splicing outcome. Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) [191]. Unlike other “black box” machine learning methods, Bayesian network can provide an interpretable model to help understand the splicing regulation. The binding status of hnRNP A1, SF2, U2AF as well as splicing change will be

considered as hidden variables. Those states will be inferred from observed evidence such as RBP motifs, iCLIP data and RNA-Seq. The joint probability of an exon can be calculated based on the network. The binding of hnRNP A1 and SRSF1 will be used to predict the U2AF binding. Their combination effects can be used to predict the splicing outcome. This method can not only predict the change of alternative splicing but also help understand how hnRNP A1 compete with SRSF1 to affect the binding of U2AF and further regulate the alternative splicing.

The method used in this study can serve as a framework for other RNA binding protein analysis. With a similar study design, a computational model can be built on different RNA binding proteins to understand the splicing mechanism and predict the splicing outcome. The data and model generated from this study can also be integrated with the new data set to build more comprehensive models. An RNA binding protein network can be built. It will help people to understand the mechanism of splicing regulation.

5.2.2 Future Directions on the Role of Alu Elements in Alternative Splicing Evolution

Although our results suggest that Alu elements can recruit U2AF binding after hnRNP A1 over expression, the underlying mechanism still needs to be further investigated. The Alu family contains two major subfamilies: AluJ and AluS, as well as other small subfamilies [192]. They may play different roles in splicing

regulation. A machine learning approach can be applied to predict the impact of Alu elements insertion on alternative splicing outcome. Besides the sequence differences among different Alu subfamilies, various genomic features will be collected to help understand the cause of U2AF binding change on Alu elements. Those genomic features include: the distance to the closest exon, the splicing score of closest splice site, whether the closest exon is a skipped exon, etc. Also, the hnRNP A1 and SRSF1 binding changes on Alu elements will also be used to analyze the binding change of U2AF. The result will help us understand the function of Alu elements during splicing regulation, as well as the role of Alu in primate evolution.

5.2.3 Future Directions on Functional Prediction of iSNVs

Our long-term goal is to utilize the developed algorithm to screen candidate intronic variants for downstream functionally test on drug-induced cytotoxicity. Our collaborator has developed high-throughput assay which allows us to evaluate the impact of thousands of intronic variants on pre-mRNA splicing and drug cytotoxicity. With such ability, we can validate our prediction result in large-scale experiments. At the meantime, due to the size limitation of current pathogenic intronic variants database, more data is needed for training. The validation result from high-throughput assay can provide additional information for us to refine the predictive model and achieve higher prediction power.

REFERENCES

1. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.
2. Wang, G.S. and T.A. Cooper, *Splicing in disease: disruption of the splicing code and the decoding machinery.* Nat Rev Genet, 2007. **8**(10): p. 749-61.
3. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.
4. Wang, Z. and C.B. Burge, *Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.* RNA, 2008. **14**(5): p. 802-13.
5. Schmid, C.W. and P.L. Deininger, *Sequence organization of the human genome.* Cell, 1975. **6**(3): p. 345-58.
6. Szmulewicz, M.N., G.E. Novick, and R.J. Herrera, *Effects of Alu insertions on gene function.* Electrophoresis, 1998. **19**(8-9): p. 1260-4.
7. Sorek, R., G. Ast, and D. Graur, *Alu-containing exons are alternatively spliced.* Genome Res, 2002. **12**(7): p. 1060-7.
8. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.
9. Lin, L., et al., *Diverse splicing patterns of exonized Alu elements in human tissues.* PLoS Genet, 2008. **4**(10): p. e1000225.
10. Zarnack, K., et al., *Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.* Cell, 2013. **152**(3): p. 453-66.

11. Agrawal, A., et al., *An intronic ABCA3 mutation that is responsible for respiratory disease.* Pediatr Res, 2012. **71**(6): p. 633-7.

12. Law, A.J., et al., *Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia.* Hum Mol Genet, 2007. **16**(2): p. 129-41.

13. Mele, C., et al., *Characterization of a New DGKE Intronic Mutation in Genetically Unsolved Cases of Familial Atypical Hemolytic Uremic Syndrome.* Clin J Am Soc Nephrol, 2015. **10**(6): p. 1011-9.

14. Cooper, D.N. and M. Krawczak, *Human Gene Mutation Database.* Hum Genet, 1996. **98**(5): p. 629.

15. Stenson, P.D., et al., *The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution.* Curr Protoc Bioinformatics, 2012. **Chapter 1**: p. Unit1 13.

16. Xiong, H.Y., et al., *RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease.* Science, 2015. **347**(6218): p. 1254806.

17. Zhang, X., et al., *Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation.* Hum Mol Genet, 2014. **23**(11): p. 3024-34.

18. Zhao, H., et al., *DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels.* Genome Biol, 2013. **14**(3): p. R23.

19. Folkman, L., et al., *DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence*

and structural properties at nucleotide and protein levels. Bioinformatics, 2015. **31**(10): p. 1599-606.

20. David, C.J. and J.L. Manley, *Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged.* Genes Dev, 2010. **24**(21): p. 2343-64.
21. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
22. Yang, Y.C., et al., *CLIPdb: a CLIP-seq database for protein-RNA interactions.* BMC Genomics, 2015. **16**: p. 51.
23. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
24. Ray, D., et al., *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.
25. Zhang, T., et al., *SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method.* J Biomol Struct Dyn, 2012. **29**(4): p. 799-813.
26. Faraggi, E., et al., *SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles.* J Comput Chem, 2012. **33**(3): p. 259-67.

27. Ezkurdia, I., et al., *Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes.* Hum Mol Genet, 2014. **23**(22): p. 5866-78.

28. Hu, Z., et al., *Revealing Missing Human Protein Isoforms Based on Ab Initio Prediction, RNA-seq and Proteomics.* Sci Rep, 2015. **5**: p. 10940.

29. Keren, H., G. Lev-Maor, and G. Ast, *Alternative splicing and evolution: diversification, exon definition and function.* Nat Rev Genet, 2010. **11**(5): p. 345-55.

30. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

31. Kalsotra, A. and T.A. Cooper, *Functional consequences of developmentally regulated alternative splicing.* Nat Rev Genet, 2011. **12**(10): p. 715-29.

32. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.

33. Scotti, M.M. and M.S. Swanson, *RNA mis-splicing in disease.* Nat Rev Genet, 2016. **17**(1): p. 19-32.

34. Tazi, J., N. Bakkour, and S. Stamm, *Alternative splicing and disease.* Biochim Biophys Acta, 2009. **1792**(1): p. 14-26.

35. Krawczak, M., et al., *Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing.* Hum Mutat, 2007. **28**(2): p. 150-8.

36. Douglas, A.G. and M.J. Wood, *RNA splicing: disease and therapy.* Brief Funct Genomics, 2011. **10**(3): p. 151-64.

37. Lopez-Bigas, N., et al., *Are splicing mutations the most frequent cause of hereditary disease?* FEBS Lett, 2005. **579**(9): p. 1900-3.

38. Slaugenhaupt, S.A., et al., *Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia.* Am J Hum Genet, 2001. **68**(3): p. 598-605.

39. Cheishvili, D., et al., *IKAP/hELP1 deficiency in the cerebrum of familial dysautonomia patients results in down regulation of genes involved in oligodendrocyte differentiation and in myelination.* Hum Mol Genet, 2007. **16**(17): p. 2097-104.

40. Cartegni, L., et al., *Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2.* Am J Hum Genet, 2006. **78**(1): p. 63-77.

41. Kashima, T., et al., *hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing.* Hum Mol Genet, 2007. **16**(24): p. 3149-59.

42. Monani, U.R., et al., *A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2.* Hum Mol Genet, 1999. **8**(7): p. 1177-83.

43. Neumann, M., et al., *Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis.* Science, 2006. **314**(5796): p. 130-3.

44. Zhang, Y.J., et al., *Progranulin mediates caspase-dependent cleavage of TAR DNA binding protein-43.* J Neurosci, 2007. **27**(39): p. 10530-4.

45. Ayala, Y.M., T. Misteli, and F.E. Baralle, *TDP-43 regulates retinoblastoma protein phosphorylation through the repression of cyclin-dependent kinase 6 expression.* Proc Natl Acad Sci U S A, 2008. **105**(10): p. 3785-9.

46. Stickeler, E., et al., *Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis.* Oncogene, 1999. **18**(24): p. 3574-82.

47. Fischer, D.C., et al., *Expression of splicing factors in human ovarian cancer.* Oncol Rep, 2004. **11**(5): p. 1085-90.

48. Narla, G., et al., *Targeted inhibition of the KLF6 splice variant, KLF6 SV1, suppresses prostate cancer cell growth and spread.* Cancer Res, 2005. **65**(13): p. 5761-8.

49. Narla, G., et al., *A germline DNA polymorphism enhances alternative splicing of the KLF6 tumor suppressor gene and is associated with increased prostate cancer risk.* Cancer Res, 2005. **65**(4): p. 1213-22.

50. Narla, G., et al., *KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis.* J Clin Invest, 2008. **118**(8): p. 2711-21.

51. Neklason, D.W., et al., *Intron 4 mutation in APC gene results in splice defect and attenuated FAP phenotype.* Fam Cancer, 2004. **3**(1): p. 35-40.

52. Matera, A.G. and Z. Wang, *A day in the life of the spliceosome.* Nat Rev Mol Cell Biol, 2014. **15**(2): p. 108-21.

53. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.* Nat Rev Mol Cell Biol, 2009. **10**(11): p. 741-54.

54. Du, H. and M. Rosbash, *The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing.* Nature, 2002. **419**(6902): p. 86-90.

55. Guth, S. and J. Valcarcel, *Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF.* J Biol Chem, 2000. **275**(48): p. 38059-66.

56. Graveley, B.R., K.J. Hertel, and T. Maniatis, *The role of U2AF35 and U2AF65 in enhancer-dependent splicing.* RNA, 2001. **7**(6): p. 806-18.

57. Wahl, M.C., C.L. Will, and R. Luhrmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.

58. Sun, J.S. and J.L. Manley, *A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing.* Genes Dev, 1995. **9**(7): p. 843-54.

59. Raghunathan, P.L. and C. Guthrie, *RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2.* Curr Biol, 1998. **8**(15): p. 847-55.

60. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.

61. Cheng, Z. and T.M. Menees, *RNA splicing and debranching viewed through analysis of RNA lariats.* Mol Genet Genomics, 2011. **286**(5-6): p. 395-410.

62. Ng, B., et al., *Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes.* J Allergy Clin Immunol, 2004. **114**(6): p. 1463-70.

63. Patel, A.A. and J.A. Steitz, *Splicing double: insights from the second spliceosome.* Nat Rev Mol Cell Biol, 2003. **4**(12): p. 960-70.

64. Konig, H., et al., *Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation.* Cell, 2007. **131**(4): p. 718-29.

65. Lim, K.H., et al., *Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes.* Proc Natl Acad Sci U S A, 2011. **108**(27): p. 11093-8.

66. Howard, J.M. and J.R. Sanford, *The RNAissance family: SR proteins as multifaceted regulators of gene expression.* Wiley Interdiscip Rev RNA, 2015. **6**(1): p. 93-110.

67. Long, J.C. and J.F. Caceres, *The SR protein family of splicing factors: master regulators of gene expression.* Biochem J, 2009. **417**(1): p. 15-27.

68. Fu, X.D., *The superfamily of arginine/serine-rich splicing factors.* RNA, 1995. **1**(7): p. 663-80.

69. Zhou, Z. and X.D. Fu, *Regulation of splicing by SR proteins and SR protein-specific kinases.* Chromosoma, 2013. **122**(3): p. 191-207.

70. Lavigueur, A., et al., *A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding.* Genes Dev, 1993. **7**(12A): p. 2405-17.

71. Sun, Q., et al., *General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer.* Genes Dev, 1993. **7**(12B): p. 2598-608.

72. Shen, H., J.L. Kan, and M.R. Green, *Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly.* Mol Cell, 2004. **13**(3): p. 367-76.

73. Blencowe, B.J., et al., *SR-related proteins and the processing of messenger RNA precursors.* Biochem Cell Biol, 1999. **77**(4): p. 277-91.

74. Roscigno, R.F. and M.A. Garcia-Blanco, *SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome.* RNA, 1995. **1**(7): p. 692-706.

75. Schneider, M., et al., *Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes.* Mol Cell, 2010. **38**(2): p. 223-35.

76. Erkelenz, S., et al., *Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms.* RNA, 2013. **19**(1): p. 96-102.

77. Chaudhury, A., P. Chander, and P.H. Howe, *Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1's multifunctional regulatory roles.* RNA, 2010. **16**(8): p. 1449-62.

78. Martinez-Contreras, R., et al., *hnRNP proteins and splicing control.* Adv Exp Med Biol, 2007. **623**: p. 123-47.

79. Zhu, J., A. Mayeda, and A.R. Krainer, *Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins.* Mol Cell, 2001. **8**(6): p. 1351-61.

80. Okunola, H.L. and A.R. Krainer, *Cooperative-binding and splicing-repressive properties of hnRNP A1.* Mol Cell Biol, 2009. **29**(20): p. 5620-31.

81. Tavanez, J.P., et al., *hnRNP A1 proofreads 3' splice site recognition by U2AF.* Mol Cell, 2012. **45**(3): p. 314-29.

82. Zamore, P.D., J.G. Patton, and M.R. Green, *Cloning and domain structure of the mammalian splicing factor U2AF.* Nature, 1992. **355**(6361): p. 609-14.

83. Zhang, M., et al., *Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit.* Proc Natl Acad Sci U S A, 1992. **89**(18): p. 8769-73.

84. Agrawal, A.A., et al., *An extended U2AF(65)-RNA-binding domain recognizes the 3' splice site signal.* Nat Commun, 2016. **7**: p. 10950.

85. Valcarcel, J., et al., *Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected].* Science, 1996. **273**(5282): p. 1706-9.

86. Singh, R., J. Valcarcel, and M.R. Green, *Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins.* Science, 1995. **268**(5214): p. 1173-6.

87. Chusainow, J., et al., *FRET analyses of the U2AF complex localize the U2AF35/U2AF65 interaction in vivo and reveal a novel self-interaction of U2AF35.* RNA, 2005. **11**(8): p. 1201-14.

88. Ruskin, B., P.D. Zamore, and M.R. Green, *A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly.* Cell, 1988. **52**(2): p. 207-19.

89. Shao, C., et al., *Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome.* Nat Struct Mol Biol, 2014. **21**(11): p. 997-1005.

90. Soares, L.M., et al., *Intron removal requires proofreading of U2AF/3' splice site recognition by DEK.* Science, 2006. **312**(5782): p. 1961-5.

91. Voith von Voithenberg, L., et al., *Recognition of the 3' splice site RNA by the U2AF heterodimer involves a dynamic population shift.* Proc Natl Acad Sci U S A, 2016. **113**(46): p. E7169-E7175.

92. Ellington, A.D. and J.W. Szostak, *In vitro selection of RNA molecules that bind specific ligands.* Nature, 1990. **346**(6287): p. 818-22.

93. Manley, J.L., *SELEX to identify protein-binding sites on RNA.* Cold Spring Harb Protoc, 2013. **2013**(2): p. 156-63.

94. Jangi, M., et al., *Rbfox2 controls autoregulation in RNA-binding protein networks.* Genes Dev, 2014. **28**(6): p. 637-51.

95. Lambert, N., et al., *RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins.* Mol Cell, 2014. **54**(5): p. 887-900.

96. Zhao, J., et al., *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Mol Cell, 2010. **40**(6): p. 939-53.

97. Zhao, J., et al., *Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.* Science, 2008. **322**(5902): p. 750-6.

98. Jain, R., et al., *RIP-Chip analysis: RNA-Binding Protein Immunoprecipitation-Microarray (Chip) Profiling.* Methods Mol Biol, 2011. **703**: p. 247-63.

99. Keene, J.D., J.M. Komisarow, and M.B. Friedersdorf, *RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts.* Nat Protoc, 2006. **1**(1): p. 302-7.

100. Mili, S. and J.A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses.* RNA, 2004. **10**(11): p. 1692-4.

101. Konig, J., et al., *Protein-RNA interactions: new genomic technologies and perspectives.* Nat Rev Genet, 2012. **13**(2): p. 77-83.

102. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 2003. **302**(5648): p. 1212-5.

103. Ule, J., et al., *CLIP: a method for identifying protein-RNA interaction sites in living cells.* Methods, 2005. **37**(4): p. 376-86.

104. Zhang, C. and R.B. Darnell, *Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data.* Nat Biotechnol, 2011. **29**(7): p. 607-14.

105. Darnell, R., *CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein.* Cold Spring Harb Protoc, 2012. **2012**(11): p. 1146-60.

106. Licatalosi, D.D., et al., *HITS-CLIP yields genome-wide insights into brain alternative RNA processing.* Nature, 2008. **456**(7221): p. 464-9.

107. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.

108. Hafner, M., et al., *PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins.* J Vis Exp, 2010(41).

109. Burger, K., et al., *4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response.* RNA Biol, 2013. **10**(10): p. 1623-30.

110. Yao, C., L. Weng, and Y. Shi, *Global protein-RNA interaction mapping at single nucleotide resolution by iCLIP-seq.* Methods Mol Biol, 2014. **1126**: p. 399-410.

111. Huppertz, I., et al., *iCLIP: protein-RNA interactions at nucleotide resolution.* Methods, 2014. **65**(3): p. 274-87.

112. Graveley, B.R., *Sorting out the complexity of SR protein functions.* RNA, 2000. **6**(9): p. 1197-211.

113. Eldridge, A.G., et al., *The SRm160/300 splicing coactivator is required for exon-enhancer function.* Proc Natl Acad Sci U S A, 1999. **96**(11): p. 6125-30.

114. Blencowe, B.J., et al., *The SRm160/300 splicing coactivator subunits.* RNA, 2000. **6**(1): p. 111-20.

115. Shen, H. and M.R. Green, *A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly.* Mol Cell, 2004. **16**(3): p. 363-73.

116. Hui, J., et al., *Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing.* EMBO J, 2005. **24**(11): p. 1988-98.

117. Pozzoli, U. and M. Sironi, *Silencers regulate both constitutive and alternative splicing events in mammals.* Cell Mol Life Sci, 2005. **62**(14): p. 1579-604.

118. Lin, C.H. and J.G. Patton, *Regulation of alternative 3' splice site selection by constitutive splicing factors.* RNA, 1995. **1**(3): p. 234-45.

119. Nasim, F.U., et al., *High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism.* RNA, 2002. **8**(8): p. 1078-89.

120. Zheng, Z.M., *Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression.* J Biomed Sci, 2004. **11**(3): p. 278-94.

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

122. McCullough, A.J. and S.M. Berget, *G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.* Mol Cell Biol, 1997. **17**(8): p. 4562-71.

123. Chen, C.D., R. Kobayashi, and D.M. Helfman, *Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene.* Genes Dev, 1999. **13**(5): p. 593-606.

124. Ule, J., et al., *An RNA map predicting Nova-dependent splicing regulation.* Nature, 2006. **444**(7119): p. 580-6.

125. Wang, Z., et al., *Systematic identification and analysis of exonic splicing silencers.* Cell, 2004. **119**(6): p. 831-45.

126. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol, 2004. **11**(2-3): p. 377-94.

127. Barash, Y., et al., *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-9.

128. Shannon, C.E., *The mathematical theory of communication. 1963.* MD Comput, 1997. **14**(4): p. 306-17.

129. Leung, M.K., et al., *Deep learning of the tissue-regulated splicing code.* Bioinformatics, 2014. **30**(12): p. i121-9.

130. Bengio, Y., A. Courville, and P. Vincent, *Representation learning: a review and new perspectives.* IEEE Trans Pattern Anal Mach Intell, 2013. **35**(8): p. 1798-828.

131. Liu, X., et al., *dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.* Hum Mutat, 2016. **37**(3): p. 235-41.

132. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.

133. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7 20.

134. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.

135. Loh, T.J., et al., *CD44 alternative splicing and hnRNP A1 expression are associated with the metastasis of breast cancer.* Oncol Rep, 2015. **34**(3): p. 1231-8.

136. Pino, I., et al., *Altered patterns of expression of members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family in lung cancer.* Lung Cancer, 2003. **41**(2): p. 131-43.

137. Ushigome, M., et al., *Up-regulation of hnRNP A1 gene in sporadic human colorectal cancers.* Int J Oncol, 2005. **26**(3): p. 635-40.

138. Yu, C., et al., *Oral squamous cancer cell exploits hnRNP A1 to regulate cell cycle and proliferation.* J Cell Physiol, 2015. **230**(9): p. 2252-61.

139. Ghigna, C., et al., *Altered expression of heterogenous nuclear ribonucleoproteins and SR factors in human colon adenocarcinomas.* Cancer Res, 1998. **58**(24): p. 5818-24.

140. Eperon, I.C., et al., *Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1.* Mol Cell Biol, 2000. **20**(22): p. 8303-18.

141. Zahler, A.M., et al., *SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing.* J Biol Chem, 2004. **279**(11): p. 10077-84.

142. Blanchette, M. and B. Chabot, *Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization.* EMBO J, 1999. **18**(7): p. 1939-52.

143. Chiou, N.T., G. Shankarling, and K.W. Lynch, *hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly.* Mol Cell, 2013. **49**(5): p. 972-82.

144. Pastor, T. and F. Pagani, *Interaction of hnRNPA1/A2 and DAZAP1 with an Alu-derived intronic splicing enhancer regulates ATM aberrant splicing.* PLoS One, 2011. **6**(8): p. e23349.

145. Sanford, J.R., et al., *Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts.* Genome Res, 2009. **19**(3): p. 381-94.

146. Ji, X., et al., *SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase.* Cell, 2013. **153**(4): p. 855-68.

147. Huelga, S.C., et al., *Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins.* Cell Rep, 2012. **1**(2): p. 167-178.

148. Konig, J., et al., *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.* Nat Struct Mol Biol, 2010. **17**(7): p. 909-15.

149. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

150. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biology, 2013. **14**: p. R36.

151. Flynn, R.A., et al., *Dissecting noncoding and pathogen RNA-protein interactomes.* RNA, 2015. **21**(1): p. 135-43.

152. Friedersdorf, M.B. and J.D. Keene, *Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs.* Genome Biol, 2014. **15**(1): p. R2.

153. Lovci, M.T., et al., *Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges.* Nat Struct Mol Biol, 2013. **20**(12): p. 1434-42.

154. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.

155. Wu, J.Y. and T. Maniatis, *Specific interactions between proteins implicated in splice site selection and regulated alternative splicing.* Cell, 1993. **75**(6): p. 1061-70.

156. Buvoli, M., F. Cobianchi, and S. Riva, *Interaction of hnRNP A1 with snRNPs and pre-mRNAs: evidence for a possible role of A1 RNA annealing activity in the first steps of spliceosome assembly.* Nucleic Acids Res, 1992. **20**(19): p. 5017-25.

157. Schwartz, S., et al., *Alu exonization events reveal features required for precise recognition of exons by the splicing machinery.* PLoS Comput Biol, 2009. **5**(3): p. e1000300.

158. Gal-Mark, N., et al., *The pivotal roles of TIA proteins in 5' splice-site selection of alu exons and across evolution.* PLoS Genet, 2009. **5**(11): p. e1000717.

159. Lev-Maor, G., et al., *Intronic Alus influence alternative splicing.* PLoS Genet, 2008. **4**(9): p. e1000204.

160. Hasler, J. and K. Strub, *Alu elements as regulators of gene expression.* Nucleic Acids Res, 2006. **34**(19): p. 5491-7.

161. Gong, C. and L.E. Maquat, *lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.* Nature, 2011. **470**(7333): p. 284-8.

162. Chen, L.L. and G.G. Carmichael, *Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA.* Mol Cell, 2009. **35**(4): p. 467-78.

163. Kelley, D.R., et al., *Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions.* Genome Biol, 2014. **15**(12): p. 537.

164. Tajnik, M., et al., *Intergenic Alu exonisation facilitates the evolution of tissue-specific transcript ends.* Nucleic Acids Res, 2015. **43**(21): p. 10492-505.

165. Pagani, F. and F.E. Baralle, *Genomic variants in exons and introns: identifying the splicing spoilers.* Nat Rev Genet, 2004. **5**(5): p. 389-96.

166. Kashima, T., N. Rao, and J.L. Manley, *An intronic element contributes to splicing repression in spinal muscular atrophy.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3426-31.

167. Santoro, A., et al., *Mutations affecting mRNA splicing are the most common molecular defect in patients with familial hemophagocytic lymphohistiocytosis type 3.* Haematologica, 2008. **93**(7): p. 1086-90.

168. Li, M., et al., *ExonImpact: Prioritizing Pathogenic Alternative Splicing Events.* Hum Mutat, 2017. **38**(1): p. 16-24.

169. Stenson, P.D., et al., *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.* Hum Genet, 2014. **133**(1): p. 1-9.

170. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

171. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

172. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.

173. Itoh, H., T. Washio, and M. Tomita, *Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes.* RNA, 2004. **10**(7): p. 1005-18.

174. Cook, K.B., et al., *RBPDB: a database of RNA-binding specificities.* Nucleic Acids Res, 2011. **39**(Database issue): p. D301-8.

175. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.

176. Faraggi, E., et al., *Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction.* Structure, 2009. **17**(11): p. 1515-27.

177. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.

178. Lu, C.T., et al., *DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications.* Nucleic Acids Res, 2013. **41**(Database issue): p. D295-305.

179. Veltman, J.A. and H.G. Brunner, *De novo mutations in human genetic disease.* Nat Rev Genet, 2012. **13**(8): p. 565-75.

180. Lupski, J.R., et al., *Clan genomics and the complex architecture of human disease.* Cell, 2011. **147**(1): p. 32-43.

181. Tennessen, J.A., et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes.* Science, 2012. **337**(6090): p. 64-9.

182. Gorlov, I.P., et al., *Evolutionary evidence of the effect of rare variants on disease etiology.* Clin Genet, 2011. **79**(3): p. 199-206.

183. Marth, G.T., et al., *The functional spectrum of low-frequency coding variation.* Genome Biol, 2011. **12**(9): p. R84.

184. Subramanian, S., *Quantifying harmful mutations in human populations.* Eur J Hum Genet, 2012. **20**(12): p. 1320-2.

185. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

186. Faustino, N.A. and T.A. Cooper, *Pre-mRNA splicing and human disease.* Genes Dev, 2003. **17**(4): p. 419-37.

187. Cogan, J.D., et al., *Familial growth hormone deficiency: a model of dominant and recessive mutations affecting a monomeric protein.* J Clin Endocrinol Metab, 1994. **79**(5): p. 1261-5.

188. Cogan, J.D., et al., *A novel mechanism of aberrant pre-mRNA splicing in humans.* Hum Mol Genet, 1997. **6**(6): p. 909-12.

189. Matsushima, M., et al., *Mutation analysis of the BRCA1 gene in 76 Japanese ovarian cancer patients: four germline mutations, but no evidence of somatic mutation.* Hum Mol Genet, 1995. **4**(10): p. 1953-6.

190. *ExAC project pins down rare gene variants.* Nature, 2016. **536**(7616): p. 249.

191. Zhang, C., et al., *Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls.* Science, 2010. **329**(5990): p. 439-43.

192. Jurka, J. and T. Smith, *A fundamental division in the Alu family of repeated sequences.* Proc Natl Acad Sci U S A, 1988. **85**(13): p. 4775-8.

CURRICULUM VITAE

Hai Lin

Education

2011 - 2017 Indiana University–Purdue University Indianapolis, Indianapolis, IN
Major: Informatics with Bioinformatics Specification
Minor: Computer Science

2005 - 2009 University of Science and Technology of China, Hefei, Anhui, China
Major: Biological Science

Research, Teaching and Work Experience

2011 - 2017 Indiana University, Indianapolis, IN
Department of Medical & Molecular Genetics
Research Assistant in Dr. Yunlong Liu's Lab

2013 - 2013 Indiana University, Indianapolis, IN
Department of Medical & Molecular Genetics
Teaching Assistant for Introduction to Next Generation Sequencing

2008 - 2009 University of Science and Technology of China, Hefei, Anhui, China
Department of Bioscience
Research Assistant in the Lab of Molecular & Cell Immunology

Publications

[in preparation] Jonathan M. Howard*, Hai Lin*, Sol Katzman, Masoud Toulue, Yunlong Liu, Jeremy R. Sanford: *hnRNPA1 modulates the association of U2AF2 with Alu-derived RNA.*

[in preparation] Hai Lin, Meng Li, Xinjun Zhang, Yunlong Liu: *regSNP-intron: prioritizing intronic single nucleotide substitution.*

Sara K. Custer, Timra D. Gilson, Hongxia Li, A. Gary Todd, Jacob W. Astroski, Hai Lin, Yunlong Liu, Elliot J. Androphy: *Altered mRNA splicing in SMN-depleted motor neuron-like cells*. PLoS ONE 10/2016; 11(10). DOI:10.1371/journal.pone.0163954

Eric A. Benson, Michael T. Eadon, Zeruesenay Desta, Yunlong Liu, Hai Lin, Kimberly S. Burgess, Matthew W. Segar, Andrea Gaedigk, Todd C. Skaar:

Rifampin Regulation of Drug Transporters Gene Expression and the Association of MicroRNAs in Human Hepatocytes. Frontiers in Pharmacology 04/2016; 7. DOI:10.3389/fphar.2016.00111

Patrícia B S Celestino-Soper, Anisiia Doytchinova, Hillel A Steiner, Andrea Uradu, Ty C Lynnes, William J Groh, John M Miller, Hai Lin, Hongyu Gao, Zhiping Wang, Yunlong Liu, Peng-Sheng Chen, Matteo Vatta: *Evaluation of the Genetic Basis of Familial Aggregation of Pacemaker Implantation by a Large Next Generation Sequencing Panel*. PLoS ONE 12/2015; 10(12). DOI:10.1371/journal.pone.0143588

Janice L. Farlow, Laurie A. Robak, Kurt Hetrick, Kevin Bowling, Eric Boerwinkle, Zeynep H. Coban-Akdemir, Tomasz Gambin, Richard A. Gibbs, Shen Gu, Preti Jain, Joseph Jankovic, Shalini Jhangiani, Kaveeta Kaw, Dongbing Lai, Hai Lin, Hua Ling, Yunlong Liu, James R. Lupski, Donna Muzny, Paula Porter, Elizabeth Pugh, Janson White, Kimberly Doheny, Richard M. Myers, Joshua M. Shulman, Tatiana Foroud: *Whole-Exome Sequencing in Familial Parkinson Disease*. 11/2015; 73(1). DOI:10.1001/jamaneurol.2015.3266

Janice L Farlow, Hai Lin, Laura Sauerbeck, Dongbing Lai, Daniel L Koller, Elizabeth Pugh, Kurt Hetrick, Hua Ling, Rachel Kleinloog, Pieter van der Vlies, Patrick Deelen, Morris A Swertz, Bon H Verweij, Luca Regli, Gabriel J E Rinkel, Ynte M Ruigrok, Kimberly Doheny, Yunlong Liu, Joseph Broderick, Tatiana Foroud: *Lessons Learned from Whole Exome Sequencing in Multiplex Families Affected by a Complex Genetic Disorder, Intracranial Aneurysm*. PLoS ONE 03/2015; 10(3). DOI:10.1371/journal.pone.0121104

R. Mourad, P-Y. Hsu, L. Juan, C. Shen, P. Koneru, H. Lin: *Correction: Estrogen induces global reorganization of chromatin structure in human breast cancer cells*. PLoS ONE 03/2015; 10(3). DOI:10.1371/journal.pone.0118237

Kwangsik Nho, Sungeun Kim, Shannon L Risacher, Li Shen, Jason J Corneveaux, Shanker Swaminathan, Hai Lin, Vijay K Ramanan, Yunlong Liu, Tatiana M Foroud, Mark H Inlow, Ashley L Siniard, Rebecca A Reiman, Paul S Aisen, Ronald C Petersen, Robert C Green, Clifford R Jack, Michael W Weiner, Clinton T Baldwin, Kathryn L Lunetta, Lindsay A Farrer, Simon J Furney, Simon Lovestone, Andrew Simmons, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kloszewska, Hilkka Soininen, Brenna C McDonald, Martin R Farlow, Bernardino Ghetti, Matthew J Huentelman, Andrew J Saykin: *Protective variant for hippocampal atrophy identified by whole exome sequencing*. Annals of Neurology 01/2015; 77(3). DOI:10.1002/ana.24349

Raphaël Mourad, Pei-Yin Hsu, Liran Juan, Changyu Shen, Prasad Koneru, Hai Lin, Yunlong Liu, Kenneth Nephew, Tim H. Huang, Lang Li: *Estrogen Induces Global Reorganization of Chromatin Structure in Human Breast Cancer Cells*. PLoS ONE 12/2014; 9(12-12). DOI:10.1371/journal.pone.0113354

Xinjun Zhang, Hai Lin, Huiying Zhao, Yangyang Hao, Matthew Mort, David N Cooper, Yaoqi Zhou, Yunlong Liu: *Impact of Human Pathogenic Micro-Insertions and Micro-Deletions on Post-Transcriptional Regulation.*. Human Molecular Genetics 01/2014; 23(11). DOI:10.1093/hmg/ddu019

Anuradha Ramamoorthy, Yunlong Liu, Santosh Philips, Zeruesenay Desta, Hai Lin, Chirayu Goswami, Andrea Gaedigk, Lang Li, David A Flockhart, Todd C

Skaar: *Regulation of MicroRNA Expression by Rifampin in Human Hepatocytes*. Drug metabolism and disposition: the biological fate of chemicals 08/2013; 41(10). DOI:10.1124/dmd.113.052886

Kwangsik Nhoemail, Jason Corneveaux, Sungeun Kim, Hai Lin, Shannon Risacher, Li Shen, Shanker Swaminathan, Vijay Ramanan, Yunlong Liu, Tatiana Foroud, Mark Inlow, Rebecca Reiman, Paul Aisen, Ronald Petersen, Robert Green, Clifford Jack, Michael Weiner, Clinton Baldwin, Lindsay Farrer, Simon Lovestone, Andrew Simmons, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kloszewska, Hilkka Soininen, Brenna McDonald, Martin Farlow, Matthew Huentelman, Andrew Saykin: *Protective variant for rate of hippocampal volume loss identified by whole exome sequencing in APOE-ε3ε3 males with MCI*. Alzheimer's & dementia: the journal of the Alzheimer's Association 07/2013; 9(4). DOI:10.1016/j.jalz.2013.04.238

K Nho, J J Corneveaux, S Kim, H Lin, S L Risacher, L Shen, S Swaminathan, V K Ramanan, Y Liu, T Foroud, M H Inlow, A L Siniard, R A Reiman, P S Aisen, R C Petersen, R C Green, C R Jack, M W Weiner, C T Baldwin, K Lunetta, L A Farrer, S J Furney, S Lovestone, A Simmons, P Mecocci, B Vellas, M Tsolaki, I Kloszewska, H Soininen, B C McDonald, M R Farlow, B Ghetti, M J Huentelman, A J Saykin: *Identification of functional variants from whole-exome sequencing, combined with neuroimaging genetics*. Molecular Psychiatry 07/2013; 18(7). DOI:10.1038/mp.2013.81

Katherine J Kelly, Yunlong Liu, Jizhong Zhang, Chirayu Goswami, Hai Lin, Jesus H Dominguez: *Comprehensive Genomic Profiling in Diabetic Nephropathy Reveals the Predominance of Pro-inflammatory Pathways.*. Physiological Genomics 06/2013; 45(16). DOI:10.1152/physiolgenomics.00028.2013

K Nho, J J Corneveaux, S Kim, H Lin, S L Risacher, L Shen, S Swaminathan, V K Ramanan, Y Liu, T Foroud, M H Inlow, A L Siniard, R A Reiman, P S Aisen, R C Petersen, R C Green, C R Jack, M W Weiner, C T Baldwin, K Lunetta, L A Farrer, S J Furney, S Lovestone, A Simmons, P Mecocci, B Vellas, M Tsolaki, I Kloszewska, H Soininen, B C McDonald, M R Farlow, B Ghetti, M J Huentelman, A J Saykin: *Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment*. Molecular Psychiatry 04/2013; 18(7). DOI:10.1038/mp.2013.24

David F.B. Miller, Pearlly S Yan, Aaron Buechlein, Benjamin A Rodriguez, Ayse S Yilmaz, Shokhi Goel, Hai Lin, Bridgette Collins-Burow, Lyndsay V Rhodes, Chris Braun, Sunila Pradeep, Rajesha Rupaimoole, Mehmet Dalkilic, Anil K Sood, Matthew E Burow, Haixu Tang, Tim H Huang, Yunlong Liu, Douglas B Rusch, Kenneth P Nephew: *A New Method for Stranded Whole Transcriptome RNA-seq.*. Methods 04/2013; 63(2). DOI:10.1016/j.ymeth.2013.03.023

Huiying Zhao, Yuedong Yang, Hai Lin, Xinjun Zhang, Matthew Mort, David N Copper, Yunlong Liu, Yaoqi Zhou: *DDIG-in: Discriminating between disease-associated and neutral non-frameshifting micro-indels*. Genome biology 03/2013; 14(3). DOI:10.1186/gb-2013-14-3-r23

Adrian G Todd, Hai Lin, Allison D Ebert, Yunlong Liu, Elliot J Androphy: *COPI transport complexes bind to specific RNAs in neuronal cells*. Human Molecular Genetics 11/2012; 22(4). DOI:10.1093/hmg/dds480

X Rao, J Evans, H Chae, J Pilrose, S Kim, P Yan, R-L Huang, H-C Lai, H Lin, Y Liu, D Miller, J-K Rhee, Y-W Huang, F Gu, J W Gray, Th-M Huang, K P Nephew: *CpG island shore methylation regulates caveolin-1 expression in breast cancer*. Oncogene 11/2012; 32(38). DOI:10.1038/onc.2012.474

Kwangsik Nho, Jason Corneveaux, Sungeun Kim, Hai Lin, Shannon Risacher, Li Shen, Yunlong Liu, Tatiana Foroud, Mark Inlow, Asheley Siniard, Rebecca Reiman, Robert Green, Clifford Jack, Michael Weiner, Matthew Huentelman, Andrew Saykin, Alzheimer's Disease Neuroimaging Initiative (ADNI: *Integrating whole-exome sequencing and imaging genetics to identify single-nucleotide variants associated with rate of hippocampal neurodegeneration in APOE-ε3/ε3 males with mild cognitive impairment*. Alzheimer's & dementia: the journal of the Alzheimer's Association 07/2012; 8(4). DOI:10.1016/j.jalz.2013.08.116

David F. Miller, Yunlong Lui, Pearlly Yan, Benjamin Rodriguez, Hai Lin, Shokhi Goel, Edra Jani, Anurag Bhattrai, Kevin Koekesh, Cong Guo, T. H.-M. Huang, Kenneth P. Nephew: *Abstract 4188: Whole transcriptome analyses of platinum-resistant ovarian cancer cells*. Cancer Research 06/2012; 72(8 Supplement). DOI:10.1158/1538-7445.AM2012-4188

J. Farlow, H. Lin, K. Hetrick, H. Ling, D. Lai, L. Sauerbeck, D. Woo, C. Langefeld, R. Brown, E. Pugh, K. Doheny, Y. Liu, T. Foroud, J. Broderick: *The Use of Linkage Data To Prioritize Results from Whole Exome Sequencing in Familial Intracranial Aneurysm (S53.001)*. Neurology 04/2012; 78(Meeting Abstracts 1). DOI:10.1212/WNL.78.1_MeetingAbstracts.S53.001