Using positional frequency distribution to identify functional splicing elements and predict pre-mRNA processing defects in human genes

By

Kian Huat (Eric), Lim

B.S., University of North Texas, 2005
M.S., Brown University, 2007

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE DEPARTMENT OF MOLECULAR BIOLOGY, CELL BIOLOGY, AND BIOCHEMISTRY AT BROWN UNIVERSITY

Providence, RI

May 2012

This dissertation by Kian Huat (Eric), Lim is accepted in its present form by the Department of Molecular Biology, Cell Biology, and Biochemistry as satisfying the dissertation requirement for the degree of Doctor of Philosophy

Date: _____ _____

William Fairbrother, Ph.D., Advisor

Recommended to the Graduate Council

Date: _____ _____

Richard Freiman, Ph.D., Reader

Date: _____ _____

Alexander Brodsky, Ph.D., Reader

Date: _____ _____

Mark Zervas, Ph.D., Reader

Date: _____ _____

Daniel Aalberts, Ph.D., Reader
Williams College

Approved to the Graduate Council

Date: _____ _____

Sheila Bonde
Dean of the Graduate School

Kian Huat (Eric), Lim

PhD candidate • Computer Science and Computational Molecular Biology • Brown University

+1 (401) 484-1155 • KH.EricLim@Alumni.brown.edu

Qualifications Summary:

- Five years of hands-on research experience in large-scale data analysis from inception to completion.
- Demonstrated in-depth knowledge in designing testable hypotheses and performing computational analysis in conjunction with molecular techniques.
- Proficient in modeling and simulation methodologies, object-oriented programming, statistical analysis and numerical computing.
- Excellent in public speaking; college-level teaching experience of multiple semesters in interdisciplinary courses.
- Certified in technical writing; experience in documenting lab manuals and molecular protocols.
- Languages: Excellent in English and Mandarin; fair in Bahasa Malaysia and other dialects of Chinese.
- Citizenship: Malaysia.

Education:

Ph.D. in Computational Molecular Biology (2011) – Brown University, Providence, RI

Sc.M. in Computer Science (2007) – Brown University, Providence, RI

Sc.B. in Computer Science (2005) – University of North Texas, Denton, TX

Professional Experience Highlights:

Senior Research Associate (August 2006 – current) – *concurrent with education*

Center of Computational Molecular Biology, Brown University, Providence, RI

Job descriptions:

- Analyzed large-scale biological data using a combination of computational biology and high-throughput genomic techniques to identify functional splicing elements in human genome.
- Designed hypotheses and performed computational analysis in conjunction with molecular assays.

Key achievements:

- Completed research objectives with numerous first-author publications in top-tier scientific journals.
- Discovered possible causes of hereditary diseases and received press releases from leading biotechnology news networks.
- Transformed raw experimental data into relational databases; delivered analysis through a fully functional web-based software.

Associate Organizer and Software Developer (January 2004 – September 2004)

International Symposium of Graph Drawing 2004, Harlem, NY

Job descriptions:

- Gathered requirements, designed, and implemented a PHP/MySQL web-based software to facilitate conference registration, paper submission, and committee discussions.
- Provided technical support for the production deployment of the system during the conference.

Key achievements:

- Software received outstanding feedbacks from conference chair, committee members, conference participants; continue to be used for subsequent years.

Publications:

K.H. Lim, L. Ferraris, M. Filloux, B. Raphael, W. Fairbrother. "Using positional distribution to identify splicing elements and predict mRNA processing defects in human genes." *Proceeding of the National Academy of Science, U.S.A.* (2011). In press.

K.H. Lim, W. Fairbrother. "Spliceman – A computational web server that predicts sequence variation in human pre-mRNA splicing." *Bioinformatics as Application Notes*. (2011) In review.

Teaching Experience:

Division of Biology and Medicine August 2006 - August 2011

Brown University, Providence RI

- Teaching Assistant in Algorithmic Foundations of Computational Biology
- Workshop Instructor in Perl programming
- Instructor in Perl programming
- Teaching Assistant in Responsible Conduct in Research
- Teaching Assistant and Lab Instructor in The Foundations of Living System

Professional Presentations

RNA 2010 June 2010

University of Washington, Seattle, Washington

"Understanding sequence determinants of splice site selection."

"A novel class of nuclear intron requires a pan-intronic stem loop for correct processing."

Student Poster Day on Computational Biology May 2009

Brown University, Providence, RI

"Fishing for hairpin: secondary structure in fish introns for correct mRNA processing."

Student Poster Day on Computational Biology May 2008

Brown University, Providence RI

"Assessing the molecular evolution of pre-mRNA secondary structure in RNA splicing."

Center of Computational Biology Student Poster Session June 2007

Brown University, Providence, RI

"Using genomic distribution pattern to identify splicing elements."

SMBE Annual Conference June 2006

Arizona State University, Tempe, AZ

"High performance sequence alignment tool."

Computing Skills

Programming: Excellent in PHP and Perl; fair in C and C++

Statistical analysis: Excellent in R; fair in SAS and Stata

Numerical computing: Excellent in Matlab

Entrepreneur Experience

CyberServers Web Hosting Services, Dallas TX August 2003 - June 2005

- Provided superb quality web hosting services to over 400 active customers.
- Developed automation tool using PHP/MySQL/Perl to facilitate customer signup, fraud detection, and support system.
- Managed eight dedicated and collocated servers in both Unix- and Windows-based environments.

Volunteering Experience

Volunteer August 2006 - January 2008

Providence Animal Rescue League (PARL), Providence, RI

Community Liaison June 2005 - January 2008

WebHostingTalk Forum, Kansas City, KS

Volunteer January 2003 - December 2003

Petmarts, Inc, Denton, TX

Others

Member of the IEEE Computer Society: Charter signer of the North Texas chapter.

Member of the RNA Society.

President and captain of the North Texas Men's Volleyball Club team

President and captain of the Brown Men's Volleyball Club team

Table of Contents

List of Illustrations

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 2

Chapter 1

Introduction to pre-mRNA splicing

1.1 Introduction to pre-mRNA splicing and its importance

Splicing of RNA is driven by the split nature of eukaryotic genes, in which the protein-coding sequences (exons) are interrupted by non-coding sequences (introns) in the DNA. Pre-mRNA splicing is an important regulatory stage in the pathway of gene expression by which introns are removed and exons are ligated to form the final mRNA products – blueprints of protein products found in our body. Alternative splicing, in which different combinations of exons can be joined to each other, results in the generation of different isoforms from a single gene and is the basis for the discrepancy between the estimated 24,000 protein-coding genes in the human genome and more than 100,000 different proteins that are postulated to be synthesized (Modrek 2002). Splicing in general, and alternative splicing in particular, is important for regulation of all levels and tissue specificity of gene expression, and if disrupted, can lead to disease (Tazi 2009; Wang 2007; Cartegni 2002).

Analyses of expressed sequence tags (ESTs) and cDNA datasets conservatively estimated that ~40-60% of human genes are alternatively spliced (Modrek 2002). A more recent study estimated that up to 95% of genes could be alternatively spliced to some degree, indicating that splicing is a common way to generate vast diversity within cells (Pan 2009). Given the vast majority of human genes contain introns and that more pre-mRNAs undergo alternative splicing, it is not surprising that disruption of normal splicing patterns can cause or modify human diseases. In fact, mutations that disrupt splicing play a larger role in hereditary diseases than other non-coding lesions. For example, the Human Gene Mutation Database (HGMD) reports that 6,428 hereditary disease alleles disrupt splicing elements, whereas only about 900 disease alleles are reported to disrupt

transcriptional elements (Stenson 2003). These estimates are likely to underestimate splicing mutations as they only reflect base changes that occur within the annotated splicing elements. In addition to these canonical elements, there are a multitude of other *cis*-regulatory motifs that play crucial roles in either activating or repressing splice site usage. Part of my research suggested that approximately one third of all diseases causing mutations alter pre-mRNA splicing (Lim 2011).

The splicing process is orchestrated by the spliceosome, a macromolecular ribonucleoprotein complex that rivals the ribosome in size and complexity. The core of the spliceosome is composed of five small nuclear ribonucleoproteins (snRNPs) that assemble onto the pre-mRNA in a series of complexes (Burge 1999; Jurica 2003), and similar to the ribosome, transition between different assembled states corresponds to different stages of functions. The complex assembly of the spliceosome is guided by the basic *cis*-elements presented in both ends of the introns (Yeo 2004): branch point, polypyrimidine tract, 3' splice site (3'ss), and 5' splice site (5'ss), a family of subsidiary elements known as intron and exon splicing enhancers and silencers (Black 2003), the intron-exon architecture (Fox-Walsh 2005), the process of transcription by RNA polymerase II (Kornblihtt 2005), and more recently, pre-mRNA secondary structure (Hiller 2007; Raker 2009).

A splicing error that either adds or removes even 1 nucleotide will disrupt the open reading frame of an mRNA; yet exons are correctly recognized from within tens of thousands, and in some cases, hundreds of thousands of intronic nucleotides. Like other levels of gene expression control, this remarkable precision involves both *cis* and *trans* components, which are composed of sequences in the pre-mRNA and interacting cellular factors. The basepaired snRNAs target specific phosphate bonds for cleavage.

The challenge of the spliceosome comes in recognizing the correct splice sites prior to the concomitant joining of exons. The relative functional strength of splice sites influences the frequency with which an exon is selected. However, several computational analyses estimate these basic splicing elements only contain at most half the information necessary for correct splice site recognition (Lim 2001; Sun 2000). In fact, a pair of strong splice sites is not always sufficient to define an exon; many pseudo-exons that are flanked by predicted splice sites are not spliced by the spliceosome (Lim 2001) because *bona fide* splice sites must be distinguished from pseudo splice-site sequences that resemble classical splice sites but are never used. Recent global analyses have also indicated that the relative enrichment in the more variable subsidiary elements, known as intron and exon splicing enhancers (ISEs and ESEs, respectively) and silencers (ISSs and ESSs, respectively), help to distinguish authentic exons and pseudo-exons (Cartegni 2002; Faustino 2003; Pagani 2004). These auxiliary elements are highly variable in sequence, but they are crucial in defining constitutive and alternative exons.

Many ESEs contain binding sites for members of the SR family of proteins. SR proteins have roles in several steps of spliceosome assembly, and function as both essential splicing factors and regulatory factors (Graveley 2005). Although they show redundancy in their essential splicing roles, they can show distinct alternative splicing activities, and the lethal effects of individual SR protein knockouts have established that they have non-overlapping roles *in vivo* (Ring 1994; Xu 2005; Wang 1996; Jumaa 1999; Ding 2004). SR proteins have a modular structure with one or two N-terminal RNA recognition motif (RRM)-type domains that bind RNA, and C-terminal domains that are enriched in arginine and serine residues (RS domains). RS domains are also found in core splicing

factors such as U2AF65 (U2 auxiliary factor 65 kDa) and U2AF35, as well as alternative splicing regulators such as *D. melanogaster* TRA. They mediate both protein-protein and protein-RNA interactions (Shen 2004; Shen 2004). Some of the splicing silencers (ESSs and ISSs), on the other hand, bind members of the extended family of heterogeneous nuclear ribonucleoproteins (hnRNPs). In addition to having RRM-type binding domains, these diverse proteins contain an hnRNP K homology (KH)-type domain that binds single-stranded RNA, as well as auxiliary domains that are often involved in protein-protein interactions, and they have multiple roles in pre-mRNA and mRNA metabolism (Krecic 1999; Dreyfuss 2002). A number of hnRNPs function as splicing repressors including hnRNP A1, polypyrimidine tract binding protein (PTB/hnRNP I), and *D. melanogaster* SXL.

In the midst of all this complexity, it has also been proposed that pre-mRNA secondary structures can potentially influence splicing efficiency. The most recent observation is that the predictive ability in the search for novel RNA binding targets for well-known proteins can be greatly improved if secondary structure is taken into consideration. It has also been noted that the binding of several positive (B52, SRp55, and NOVA-1) and negative (hnRNP A1) regulators of splicing have been shown to depend on RNA secondary structures as well as on the target nucleotides sequences (Buckanovich 1997; Damgaard 2002; Nagel 1998; Shi 1997). Recently, the fact that most major members of the SR protein family have been observed to be potentially affected by the conformation of a target RNA may indicate that structural influences maybe a widespread occurrence (Buratti 2004).

A final contextual point is that alternative splicing decisions are not made on a static, pre-synthesized template. Many older and simplified models suggest that pre-mRNAs

are fully transcribed, then fold, and are spliced. Instead, current research shows as the pre-mRNA emerges from RNA polymerase, it immediately becomes packaged with various processing and general RNA binding factors (Bentley 2005). The C-terminal domain of the RNA polymerase II large subunit helps to recruit RNA processing factors to the emerging transcripts (Kornblihtt 2004; Maniatis 2002). The rate of polymerase elongation can also influence splicing patterns by accelerating or delaying the synthesis of competing splice sites or regulatory elements (Kornblihtt 2004). In addition, slow-folding RNA structures, even if quite stable, might not have time to form in the pre-mRNA before splicing occurs. Therefore the competition between the folding of RNA structures and proteins that bind single-stranded conformations of the RNA can also alter splicing efficiency. It is in this dynamic setting that alternative splicing decisions are made.

1.2 Mechanisms of enhancement and silencing of alternative splicing

1.2.1 Mechanisms of activation. A tremendous amount of word has been carried out on ESEs that activate 3'ss and 5'ss by binding SR proteins. An early and simple model for ESE-dependent 3'ss activation by SR proteins involves the recruitment of U2AF65 to weak polypyrimidine tracts. An example of this activity is the female-specific ESE, which comprises a purine-rich element and 6 repeats of a 13-nucleotide *dsx* repeat element (dsxRE) (Lynch 2995). Each element cooperatively binds a conventional SR protein along with two RS-domain-containing regulators, TRA and TRA2 (Lynch 2996). The separation of ~300 nucleotides between the ESE and the regulated 3'ss is crucial for sex-specific regulation, as the experimental reduction of this distance results in TRA-independent activation by TRA2 and SR proteins (Tian 1994).

TRA functions as the crucial female-specific regulator by promoting the cooperative stabilization of the enhancer complex, which allows it to function from a distance. The enhancer has been shown *in vitro* to function through the interaction of the recruited RS domains with U2AF35, thereby promoting U2AF65 binding to weak polypyrimidine tract (Zuo 1996). Nucleotide substitutions that strengthen the 3'ss bypass the ESE requirement. Despite the evidence in its favor, this model is challenged by the fact that RS domain of U2AF35 is not required for regulated splicing of *dsx* exon 4 *in vivo* (Rudner 1998). In addition, other ESEs have been shown to activate splicing without affecting U2AF65 binding (Kan 1999). An alternative mechanism for enhancer activity involves the bridging interactions between ESEs and spliceosomal components, which are mediated by the SRm160 and SRm300 splicing co-activators (Li 1999; Eldridge 1999; Blencowe 2000). These co-activators contain RS domains but lack RRM domains, and can form multiple interactions with snRNPs and enhancer-bound SR proteins. ESEs can also

activate 5'ss. Female-specific selection of an alternative 5'ss in the *D. melanogaster* fruitless gene is dependent on an upstream ESE with 3 copies of the 13-nucleotide dsxRE motif that cooperatively binds TRA and TRA2 proteins (Heinrichs 1998). Through *in vitro* artificial recruitment experiments, RS-domain interactions that are involved in 5'ss and 3'ss activation by TRA and TRA2 are similar (Lam 2003).

The preceding models for ESE function are based on the ability of RS domains to mediate protein-protein interactions (Wu 1993). Recent evidence indicates a series of sequential RS-domain-RNA contacts at the branch point and the 5'ss during splicing complex assembly (Shen 2004; Shen 2004). In particular, RS domains that are recruited artificially or naturally to an ESE specifically contact the branch point of the RNA in the spliceosomal A complex (Shen 2004; Shen 2004). The strong correlation between the ability to activate splicing and to contact the branch point, coupled with the detection of the interaction in the A complex, has led to the suggestion that the RS-domain-branch-point interaction might be responsible for splicing activation by ESEs. However, contacts of RS domains with proteins and with RNA are not mutually exclusive, and their relative contribution to ESE function remains to be elusive.

1.2.2 Mechanisms of repression. Repression of splicing often involves hnRNPs such as SXL, PTB, and hnRNP A1. In the simplest scenario, silencer-bound proteins directly antagonize splicing factor binding. However, repression often requires cooperative repressor binding to multiple silencer elements of a common type. Two models for cooperative repression involve the creation of a "zone of silencing", either by the propagation of repressor binding from high-affinity silencers or by looping out the RNA between bound repressors. In each case, exons, splice sites, or enhancers are made inaccessible to their cognate splicing factors (Cartegni 2002; Wagner 2001).

Alternative splicing of *tra* pre-mRNA in *D. melanogaster* is an example of simple steric inhibition of early spliceosome assembly. In female flies, SXL protein binds to the polypyrimidine tract, thereby blocking access to U2AF65. A downstream 3'ss, which binds U2AF65 but not SXL, is then selected (Valcarcel 1993). Splicing repression by mammalian PTB can also occur by competing with U2AF65 binding (Singh 1995; Zhang 1999). However, repression usually requires cooperative PTB binding to multiple sites, which are often not associated with the polypyrimidine tract, so repression can often be more complex than simple binding competition with U2AF65 (Wagner 2001; Zhang 1999; Southby 1999; Chou 2000; Wagner 2002). Autoregulation of *Sxl* splicing occurs by the cooperative binding of SXL protein to multiple sites around exon 3, leading to exon skipping (Wang 1994). Remarkably, *in vitro* experiments indicate that rather than inhibiting early spliceosome assembly, repression by SXL occurs after the first catalytic step of splicing (Lallena 2002).

HnRNP A1 functions as a general sequence-independent regulator of alternative splicing by antagonizing SR proteins such as SF2/ASF (Mayeda 1992; Eperon 2000), and variations in the ratios of SF2/ASF and hnRNP A1 can also influence many alternative splicing events (Caceres 1994). Variations in the relative activities can result from a change in the sub-cellular localization of these proteins. Phosphorylation of hnRNP A1 as a result of the stress-mediated $MKK_{3/6}$-p38 signaling cascade results in its export to the cytoplasm, with consequent changes in alternative splicing regulation in the nucleus (van der Houven van Oordt 2002). High-affinity hnRNP A1 binding sites function as specific silencers (Blanchette 1999; Zhu 2001; Del Gatto Konczak 1999; Tange 2001). The artificial recruitment of just the C-terminal glycine-rich domain of hnRNP A1 oligomerization can restore ESE function (Del Gatto Konczak 1999). The mechanism of

repression depends on the relative location of binding sites and can include antagonism of binding of U1 and U2 snRNPs and SR proteins. In the case of hnRNP A1-induced alternative splicing events in the *hnRNP A1* gene, the skipped exon is looped out between high-affinity sites, but the important effect might be to bring the external exons into proximity rather than to directly inhibit the skipped exon (Blanchette 1999; Villemaire 2003). This is reminiscent of the effects caused by the secondary structure of RNA, which can similarly influence splice site selection by bringing widely separated *cis*-acting elements into juxtaposition (Baraniak 2003).

1.2.3 Combinatorial interplay between positive and negative regulators. In contrast to the examples mentioned above, which mostly involve the promotion or inhibition of early spliceosome assembly by individual regulators, many alternative splicing events involve a more complex interplay between positive and negative regulators that function through cognate enhancers and silencers. For example, an ESS in HIV1 *tat* exon 3 binds hnRNP A1 with high affinity and inhibits splicing by propagating further hnRNP A1 binding towards the 3'ss. This propagative binding can be inhibited, and splicing activated, by the binding of SF2/ASF to an upstream ESE. In contrast to other ESEs, in which artificial recruitment of just an RS domain can restore function, at this ESE it is only the RRM domains of the SR protein that are required (Zhu 2001). A purine-rich ESE in the M2 exon of the *IgM* gene functions in a similar way, by antagonizing an adjacent PTB-binding ESS (Shen 2004). The ESS itself promotes the formation of a 'dead-end' spliceosomal A-like complex, rather than blocking complex formation completely (Kan 1999). The formation of dead-end, splicing-related complexes as a mechanism of repression has also been observed in other systems. Retention of intron 3 of the *D. melanogaster* P-element transposase is somatic cell specific. The KH-

domain protein P-element somatic inhibitor (PSI) contacts U1 70K protein and diverts U1 snRNP binding away from the authentic 5'ss to a pair of pseudo-5'ss (Labourier 2001). Non-productive interaction of the 5'ss of caspase-2 exon 9 with a 'decoy' 3'ss leads to exon skipping (Wu 2001). In these cases, sequences that resemble splice sites function as silencer elements.

Antagonism between members of the SR protein family can also mediate alternative splicing decisions. For example, the regulation by SRp20 of its own splicing is antagonized by SF2/ASF (Jumaa 1997). Several SR-related factors have also been identified recently that can modulate the splicing regulatory activity of canonical SR proteins. The *D. melanogaster* RSF1 repressor antagonizes the ability of SF2/ASF to promote the binding of U1 to pre-mRNA. RSF1 comprises an N-terminal RRM, which binds purine-rich ESE-like sequences and competes directly with SR proteins, and a C-terminal Gly/Arg/Ser-rich region that interacts with the RS domain of SF2/ASF (Labourier 1999; Labourier 1999*)*. The mammalian SR-related protein SRrp86 exerts differential control over the splicing regulatory effects of specific canonical SR proteins. It enhances SRp20 activity but represses other family members including SC35, SF2/ASF and SRp55 (Barnard 2000), and is itself antagonized by the SR protein 9G8, hnRNP G and SAFB (Scaffold Attachment Factor B) (Li 2003). SRrp86 shows tissue-restricted expression at the translational level, whereas its targets are present ubiquitously; this results in a unique ratio of factors in each tissue type (Li 2003).

Functional antagonism between different hnRNPs is another emerging theme in the regulation of alternative splicing. The opposing effects of repressive PTB and activating CELF (CUG-BP and ETR3-like factors) proteins (Ladd 2001) modulate the splicing of several target pre-mRNAs (Charlet-B 2002; Zhang 2002; Suzuki 2002; Gromak 2003). In

the splicing of *N*-methyl-D-aspartate (NMDA) receptor subunit NR1 and α-actinin pre-mRNAs, direct competition between binding of PTB and CELFs to adjacent sites is involved (Zhang 2002; Suzuki 2002; Gromak 2003). In the case of cardiac troponin-T, functional antagonism occurs without binding competition (Charlet-B 2002). In addition to their anti-repressor activity, CELF proteins also function as repressors of other splicing events, sometimes elsewhere in the same transcript, by binding to CUG- or UG-repeat elements (Zhang 2002; Suzuki 2002; Gromak 2003).

Neuronal-specific alternative splicing of the mammalian *src* N1 exon is a well-studied model system with many regulatory factors, none of which alone seems to be sufficient to explain the tissue specificity of regulation (Black 2003). Inhibition of N1 splicing in non-neuronal cells depends on four flanking CUCUCU elements. These bind PTB cooperatively, as indicated by reduced PTB binding at the upstream sites after mutation of the downstream elements and vice versa (Chou 2000). U2 snRNP binding to the upstream intron correlates inversely with PTB binding. By contrast, U1 snRNP binding to the N1 exon seems to be unaffected by PTB-mediated repression, which argues for cooperative binding with looping of intervening RNA, rather than propagative PTB binding, which would be expected to displace U1 snRNP. N1 inclusion in neuronal cells depends on a downstream ISE, which overlaps with one of the CUCUCU motifs and interacts with hnRNP F and hnRNP H, and KH-type splicing regulatory protein (KSRP). Assembly of the ISE complex is promoted by nPTB (Markovtsov 2000), a neuronally restricted paralog of PTB, which binds the CUCUCU elements but is less repressive than PTB. nPTB seems to have a dual role in antagonizing a repressor as well as promoting assembly of an ISE complex. Despite its restricted expression, nPTB is not sufficient to activate N1 splicing. An ESE in the N1 exon binds SF2/ASF and hnRNP F and hnRNP

H, whereas hnRNPA1 binding mediates repression by a mechanism that is distinct from that of PTB. The mechanistic details of N1 splicing activation by the ISE and ESE are not yet well understood; however, this system illustrates the complexity of splicing regulation.

In some cases, the presence of absence of an individual mammalian tissue-specific regulator is sufficient to determine splicing patterns, The brain-specific KH-domain-containing Nova proteins are involved in determining a number of tissue-specific alternative splicing events through binding to ISEs, ISSs, and ESSs that contain UCAY motifs (Jensen 2000; Ule 2003). Noval promotes alternative exon inclusion in glycine receptor $\alpha 2$, and functions antagonistically with nPTB (Kanopka 1998). However, most tissue-specific alternative splicing events studied so far seem to be regulated by a more complex network of synergistic and antagonistic influences between groups of regulators. In general, the individual *cis* elements that have been identified are not sufficient to direct tissue-specific splicing pathways.

1.2.4 Location dependencies between splicing regulators. As discussed earlier, RS-domain-containing factors are known primarily for their roles as activators when interacting with enhancer elements found in the exons. However, these activators can also function as repressors when bound at intronic locations. The adenovirus L1 transcription unit provides an example of splicing repression mediated by the binding of SR proteins to an ESE-like sequence that, owing to its position upstream of the branch point, functions as a silencer. Artificial recruitment shows that the second RRM domain of SF2/ASF mediates repressor activity (Dauksaite 2002). Other viral alternative splicing systems have also demonstrated that SR proteins, when bound at intronic locations, tend to function negatively in splice site recognition (Manley 1996; Martinez-Contreras

2006; McCullough 2000). On the other hand, hnRNP proteins which are more commonly viewed as repressors can stimulate the use of alternative splice sites when bound at intronic locations. Activation of a weak 5'ss in the *eNOS* (endothelial nitric oxide synthase) transcript is affected by the binding of hnRNP L to a CA-repeat downstream of the 5'ss. Similarly, the closely related TIA1 and TIAR proteins function at uracil-rich ISEs to promote the interaction of U1 snRNP with weak 5'ss, apparently by direct contacts with the U1 snRNP-specific C protein (Del Gatto-Konczak 2000; Forch 2000; Forch 2002; Le Guiner 2001). Experiments that relocated these exonic silencers into introns converted them into enhancers (McNally 1996) and the reverse experiment of moving natural exonic enhancers into intronic locations resulted in splicing repression (Ibrahim 2005).

Consistent with the observation that the location in which a splicing regulator binds can have opposite consequences, as part of this thesis work, I have developed a computational method and software to interpret the effect of point mutation in splicing. In this work, I exploited the relationship between location and function as a discovery tool. The method assumes splicing elements exhibit non-uniform distributions around annotated splice sites. In other words, splicing elements tend to be abundant where they function positively and rare where they are inhibitory. This unique property – most exons are not under negative regulation but rather defined by predominantly positive signals in regions depleted of negative elements – is helpful in locating functional splicing elements. Methods, results, and experimental evidence of my proposed computational framework are discussed in Chapter 2. In addition, I have included a public web-based software to annotate the effect of sequence variation in splicing based on my proposed method. Description of this software can be found in Chapter 3. This computational

method was used to perform a large cross-species analysis. Computational and experimental methods, results, and summaries are presented in Chapter 4.

1.3 Global identification of regulatory motifs

Global approaches that permit genome-wide identification of regulators and target sequences are beginning to contribute to our understanding of alternative splicing. An initial aim of experimental and computational screens is to provide a catalog of regulatory sequence motifs, and substantial progress has been made on this front.

1.3.1 *in vitro* approaches. *In vitro* binding SELEX (systematic evolution of ligands by exponential enrichment) experiments have been carried out with a number of SR proteins and hnRNPs. Optimal binding sites vary between individual SR proteins, but they often function as splicing enhancers (Tacke 1999) and frequently resemble 'classical' purine-rich ESEs. Similarly, optimal binding sites for repressor hnRNPs, such as hnRNP A1 (Burd 1994), PTB/hnRNP I (Perez 1997) and SXL (Singh 1995), correspond to known splicing silencers. The SELEX-selected sequence for FOX1 protein (GCAUG) corresponds to a regulatory sequence motif that had previously been identified in a number of pre-mRNAs and independently by computational analysis of brain-specific exons (Brudno 2001).

In vitro functional SELEX experiments have revealed various classes of ESE, some of which could be associated with particular SR proteins (Liu 2000; Liu 1998; Schaal 1999; Tian 1995). Splicing that is dependent on the addition of individual SR proteins to cytoplasmic extract allowed the selection of distinct motifs for SF2/ASF, SC35, SRp40 and SRp55 (Liu 2000; Liu 1998). These functionally defined motifs, which were more redundant than *in vitro* binding, SELEX selected sequences for the same proteins (possibly reflecting the fact that ESE function does not require the tightest binding), have led to the highly successful ESEfinder web resource (Cartegni 2003).

1.3.2 *in vivo* approaches. *In vivo* functional selection approaches have used cell-transfection assays. The first example of this approach was based on the selection for ESEs in a mini-gene construct, and involved the replacement of a necessary ESE in an internal exon by a random 13-mer sequence. Following transfection, active ESEs were identified by PCR after reverse transcription of RNA (RT-PCR) and sequencing of the included exons (Coulter 1997). This revealed both the expected purine-rich ESEs and a class of AC-rich ESEs, which were subsequently shown to be binding sites for the RNA-binding protein YB1 (Stickeler 2001). An elaboration of this approach, known as 'fluorescence-activated screen for exonic splicing silencers', was reported by the Burge laboratory (Wang 2004). A three-exon mini-gene was engineered to produce enhanced green fluorescent protein (eGFP) only when exon 2 is skipped. A library of random decamers was cloned into exon 2, stably transformed cells were selected, and fluorescence-activated cell sorting (FACS) was used to select cells that produced eGFP. PCR of genomic DNA then allowed ESS sequences to be derived. Using this approach, 141 decamers with ESS activity were identified. These could be grouped into seven putative ESS hexamer motifs, some of which resembled optimal binding sites for hnRNP A1 and hnRNP H. The cell-based screens for ESEs and ESSs (Coulter 1997; Wang 2004) could be carried out in various cell types, potentially revealing motifs that mediate cell-type-specific regulation. Importantly, the experimental design of Wang and colleagues (Wang 2004) could readily be extended to ISEs and ISSs.

1.3.3 Computational approaches. Purely computational approaches, followed by experimental validation, have also contributed to the global definition of splicing regulatory motifs (Coulter 1997; Wang 2004; Fedorov 2001; Fairbrother 2002; Yeo 2004). The usual approach is to consider a sequence dataset that is expected to be

enriched in binding sites of interest, and to compare it with one or more control sets. For example, a search for ESE motifs in coding exons from multi-exon genes versus coding sequences from intron-less genes as a control (Fedorov 2001). Hexamers (or other N-mers, depending on the size of the datasets) that are significantly enriched are considered as putative binding sites, and generally group into families of highly similar sequences that, together, define a binding motif. Computational approaches are necessarily focused on common binding sites for which hexamer overrepresentation can be observed.

In the RESCUE-ESE (Relative Enhancer and Silencer Classification by Unanimous Enrichment) approach, hexamers were identified in constitutive human exons by enrichment in exons versus introns, and in exons with weak splice sites versus exons with strong splice sites (Fairbrother 2002). Using stringent cut-offs, 238 distinct hexamers were identified as possible ESEs. These clustered into ten motifs, of which two were specific to 5'ss, five were specific to 3'ss, and three associated with both splice sites. Representatives of all ten classes were validated as ESEs in transfection assays. In addition, an average of 5.2 ESE 'clumps' (multiple overlapping ESE hexamers) were found per exon, with most exons having between 3 and 7 ESE clumps. Lower ESE densities were observed in a dataset of 2,000 alternative exons. In a comparative analysis between human, mouse and *Fugu* species, potential ESEs and ISEs were identified (Yeo 2004). In the RESCUE-ISE approach, hexamers were selected by enrichment in introns versus exons and in introns with weak splice sites versus introns with strong splice sites. Two primary ISE motifs, the known ISE triplet GGG motif and a C-rich motif, were found in human and mouse, but not in *Fugu* species.

To avoid the complicating effects of protein-coding biases, Zhang and Chasin (Zhang

2004) compared octamers (allowing a single mis-match) from non-coding 5' UTR constitutive exons with both pseudo-exons and intron-less 5' UTR sequences. They identified 2,069 putative ESE octamers and 974 putative ESS octamers (5% of all octamers), grouping into 80 putative ESEs and 69 putative ESSs. The ESE octamers had substantial overlap with the RESCUE-ESE identified motifs (Fairbrother 2002). Of 20 potential ESEs and ESSs tested, 18 were experimentally validated. Furthermore, the motifs identified in this study and in the work by Wang and colleagues (Wang 2005) could be used to distinguish between authentic exons and pseudo-exons.

In addition to these searches for general motifs that regulate splicing, some computational analyses have searched for motifs that are associated specifically with alternative exons. Analysis of a set of 25 brain-specific exons reported in the literature allowed Brudno *et al*. (Brudno 2001) to show that the hexanucleotide UGCAUG, a binding site for FOX1 protein (Jin 2003), was overrepresented in the downstream intron, and this enrichment was conserved for the orthologous exons in other species Minovitsky 2005). Miriami *et. al*. identified longer C- and G-rich elements in the introns that flank skipped exons, prompting the suggestion that these might be involved in looping out skipped exons (Miriami 2003). Finally, a set of motifs associated with exons (or splice sites) that are skipped in brain or liver was found to be enriched in known silencers (Yeo 2004).

Chapter 2

Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes

Lim KH, Ferraris L, Filloux ME, Raphael BJ, and Fairbrother WG. (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.* **108**:11093-11098.

2.1 Abstract

We present an intuitive strategy for predicting the effect of sequence variation on splicing. In contrast to transcriptional elements, splicing elements appear to be strongly position dependent. We demonstrated that exonic binding of the normally intronic splicing factor, U2AF65, inhibits splicing. Reasoning that the positional distribution of a splicing element is a signature of its function, we developed a method for organizing all possible sequence motifs into clusters based on the genomic profile of their positional distribution around splice sites. Binding sites for serine/arginine rich (SR) proteins tended to be exonic whereas heterogeneous ribonucleoprotein (hnRNP) recognition elements were mostly intronic. In addition to the known elements, novel motifs were returned and validated. This method was also predictive of splicing mutations. A mutation in a motif creates a new motif that sometimes has a similar distribution shape to the original motif and sometimes has a different distribution. We created an intraallelic distance measure to capture this property and found that mutations that created large intraallelic distances disrupted splicing in vivo whereas mutations with small distances did not alter splicing. Analyzing the dataset of human disease alleles revealed known splicing mutants to have high intraallelic distances and suggested that 22% of disease alleles that were originally classified as missense mutations may also affect splicing. This category together with mutations in the canonical splicing signals suggest that approximately one third of all disease-causing mutations alter pre-mRNA splicing.

2.2 Introduction

Splicing is catalyzed by the spliceosome, a riboprotein complex that rivals the ribosome in size and complexity. The ribosome has a large and small subunit whose assembly on the mRNA substrate corresponds to a functional switch from initiation to elongation. The spliceosome is composed of five subunits that appear to exist in at least four different stable configurations and, like the ribosomal subunits, transition between different assembled states corresponding to different stages of function (Jurica 2002; Jurica 2004; Nilsen 2002). Mass spectroscopy has identified at least 300 RNA and protein components in this catalytic complex and studies have demonstrated heterogeneity in spliceosomal complexes isolated from different splicing substrates (Chen 2007; Nilsen 2003; Zhou 2002). The spliceosomal components that recognize the basic *cis*-elements of the splicing process are known. How the spliceosome assembles and reorganizes on these elements is also fairly well understood. However, several computational analyses estimate that these basic splicing elements contain at most half the information necessary for splice site recognition (Lim 2001; Sun 2000). The remaining information lies outside these splice sites presumably as enhancers or silencers.

This information required to specify splicing presents a considerable mutational target—estimates of the fraction of disease mutations that affect splicing range from 15% (Stenson 2003) to 62% (Lopez-Bigas 2005). Transcript analysis of genotyped cell lines has discovered numerous cases of allelic splicing demonstrating that polymorphisms also disrupt splicing (Kwan 2008; Kwan 2007). These types of functional variants likely account for a similarly large fraction of the detected genetic risk for complex disease and could eventually be a target for molecular intervention. As physical methods for the detection of alternative splicing require large panels of genotyped accessible tissue,

these studies will probably continue to be limited to samples harvested from human blood. An alternate approach is the prediction of causative variations from single-nucleotide polymorphism (SNPs) that fall within splicing elements. The key to this approach is being able to identify what the splicing elements are and whether a variation is disruptive.

Recently, a variety of experimental and computational methods have emerged to identify sequence elements capable of functioning as enhancers and silencers (Fairbrother 2002; Zhang 2005). Considerable data has been gathered on the proteins that recognize these elements. The prototypical splicing activator that recognizes exonic splicing enhancers (ESEs) is one of the serine/arginine rich (SR) proteins (Manley 1996). The heterogeneous ribonucleoprotein (hnRNP) family of proteins has generally been regarded as repressors as they inhibit splicing when bound to exons in pre-mRNA. However, hnRNP A, B, C, F, and H stimulate splicing when bound at intronic positions (Martinez-Contreras 2006; Swanson 2988). Conversely, SR proteins do not always promote splicing; SR proteins bound at intronic positions tend to function negatively in splice site recognition, a fact exploited by several viral alternative splicing systems (Cook 1998; McNally 1996; Wang 1998; Kanopka 1996). Experiments that relocated these intronic silencers into exons converted them into enhancers (McNally 1996), and the reverse experiment of moving a natural ESE into an intronic location resulted in splicing repression (Ibrahim 2005). Positional effects on function appear at a finer scale than binning sequence into intron versus exon. Indeed an element's location within an exon can also affect its function (Goren 2006). This notion that an element's activity is a function of its position has led to the routine use of "RNA maps" in cross-linking immunoprecipitation (CLIP) studies. An RNA map separates immunoprecipitated tags

that fall around positively regulated exons from tags that fall around negatively regulated exons and plots the location of each tag set relative to the regulated splice site. In the genome-wide CLIP studies of hnRNP C, nova, and Fox1/2 specificity, the RNA maps illustrate that function differs according to positional distribution (Ule 2006; Yeo 2009; Konig 2010).

In this work, we exploit the relationship between location and function as a discovery tool. We show that splicing elements have signature positional distributions around constitutively spliced exons—they are abundant where they function positively and rare where they are inhibitory. Thus in a dataset of successful splicing events an element's positional distribution is a proxy measure for where it enhances splicing. As different types of elements will have different positional distributions, we hypothesize that different positional distributions will define different splicing elements. Here, we describe the development of this discovery tool. All possible hexamers are mapped around splice sites. We discover 51 types of positional distributions (splicing elements) and demonstrate that these are predictive of function in vivo. We find that mutations that create new hexamers with radically different positional distributions are more likely to cause striking differences in splicing in vivo. We use this tool to analyze disease alleles within the human population.

2.3 Materials and methods

2.3.1 UV cross-Linking. RNA probes were T7 transcribed from DNA oligos (all sequences listed in Table 1). Briefly, radiolabeled RNA was transcribed with incorporating alpha 32 P UTP. Approximately 10,000 cpm of substrate was incubated in splicing buffer for 5 min prior to a 2-min exposure (4 cm) under a 30 W germicidal lamp (Sanko Denki G30T8). Irradiated mixture was placed on ice prior to a 30 min digestion in a mixture of RNAse A1/T1(I unit each). Samples were fractionated by 12% SDS-PAGE, dried, and visualized by phosphoimager. Certain experiments were performed with extract preincubated with anti U2AF65 antibody. Here 2 uL of supernatant from the MC2 hybridoma line was added to the extract, 2 min prior to the addition of RNA.

2.3.2 Motif discovery in systematic evolution of ligands by exponential enrichment (SELEX) output. The binding specificity can be determined by SELEX experiments for several RNA binding proteins (Table 3). The output of a SELEX experiment is typically 10–40 sequences that have been identified via iterative selection protocol as high affinity binding sites for the factor. From these sequences, motifs were derived using a Gibbs sampling strategy (Thompson 2005). To determine the length of these motifs for different RNA binding proteins, we examined the maximum a posteriori probability (MAP) of the alignment given by Gibbs sampling. The MAP value is measured relative to an empty null alignment, by taking the difference between the log of the probability of the alignment and the log of the probability of an empty alignment. We reason that motif is the most informative with length at which the MAP value is the highest (Table 3).

2.3.3 Computing and comparing positional distributions. We used the following

algorithm to compute positional distribution of hexamers.

1. Create alignments of nucleotides surrounding all 3′ splice sites (3′ss) and 5′ splice sites (5′ss). We include up to two hundred nucleotides of intronic sequence flanking either side of the splice site, and up to one hundred nucleotides of exonic sequence flanking either side of the splice site.
    a. For exons less than 200 nucleotides, the exonic sequence is divided in half and assigned to the closest splice site.
    b. For introns less than 400 nucleotides, the intronic sequence is divided in half and assigned to the closest splice site.
2. Count the number of occurrences of all 4,096 hexamers.
3. For each hexamer, construct a feature vector of length 600 by concatenating counts from the 3′ss alignment and the 5′ss alignment. We use the following indexing of positions in the vector.
    a. Positions starting at –200 and ending at 1 correspond to the upstream intronic region.
    b. Positions starting at 0 and ending at 199 correspond to the exonic region.
    c. Positions starting at 200 and ending at 399 correspond to the downstream intronic region.
4. Compute z-score normation in each entry of each feature vector. This normalization is done region-by-region. In other words, we separated our feature vector into 3 distinct regions as outlined in Step 3 of the algorithm: 200 nucleotides in the upstream 3'ss (Step 3a); 200 nucleotides in the exonic region (Step 3b); and 200 nucleotides in the downstream 5'ss (Step 3c).
5. Cluster the 4,096 feature vectors using a *k*-means algorithm (MacQueen 1967) and

the L1 distance as the distance between feature vectors. We determine the number of clusters (*k*) with the Calinski and Harabasz (CH) index (Calinski 1974).

a. **Preparation of exon database**. An exon/intron database was made from the RefSeq Genes annotation of the NCBI36/hg18 assembly stored at the UCSC table browser (Karolchik 2004). Each entry in the database consists of at most 600 nucleotides: two 200 nucleotide intronic flanks and two 100 nucleotide exonic flanks on each side of the splice sites. In the case where intronic or exonic length is less than 400 or 200 nucleotides, respectively, the sequence is divided by half and each half is assigned to its nearest splice site. All duplicated entries were screened and removed from the database.

b. **Selecting word size and counting words**. Word size is typically guided by a power calculation, often with some consideration of the expected size of the motifs that are being detected (Fairbrother 2002). RNA binding proteins typically contain one to four of RNA recognition motif domains so the motifs recovered are expected to be of heterogeneous length. It is therefore unlikely that there is a single word size that is appropriate for all motifs present in the data. Previous implementations of dictionary methods illustrate how a smaller word size choice is generally self-correcting. The final alignment of enriched words allows motifs to grow beyond their word size. Our analysis of prior SELEX studies indicates RNA binding proteins recognize motifs between the length of 6 and 10 nucleotides. For this reason we selected hexamers for the analysis presented here. Because overlapping occurrences of internally repetitive words (i.e., AAAAAA) can occur more frequently than complex words, overlapping occurrences of any words will be counted as a single occurrence.

c. **Vector construction, normalization, and clustering**. These counts are then separated for the three regions (Step 3, a-c) and compiled as a feature vector,

$$z = \frac{x - \mu}{\sigma}$$ Eq.1

where *x* are the counts of feature vector, μ is the average regional count and σ is the regional standard deviation. The vector is comprised of normalized counts and not the raw counts because the ultimate goal is to compare the different shapes of word distributions around splice sites. Performing this comparison with raw counts results in the undesirable outcome of simply clustering according to frequency, common words together into one group and rare words into an- other group. It is therefore desirable to perform this comparison with normalized frequencies so the clustering separates words according to the "shape" of their distribution around critical sites. We performed standard z-score normalization for each word. The normalization was performed separately for exonic and intronic space to decrease the dominant influence of the protein coding function of exons on word distribution. For example, words that contain common codons tend to shift to dramatically higher frequencies on the exonic side of the splice site. This hexamer bias from codon usage is eliminated by subdividing all sequence into coding (exons) and noncoding regions (upstream and down- stream introns). The standard z-score normalization was per- formed separately on each of these subdivisions (Eq. 1). In this way upward and downward trends across exons and introns are captured but bulk frequency differences

between regions are not. These vectors will then be clustered into similarly shaped groups.

The first step of clustering is to compare all possible pairs of vectors, position by position, and to return a distance metric that quantifies the “closeness” of the two vectors. The clustering algorithm uses the distance metric to organize the data into groups of similar vectors. An obvious choice for distance metric is Euclidean distance however, the sharp peaks created by the splice sites themselves dominated the comparison and prevented the detection of more subtle signals This was remedied by using Manhattan distance also referred to as the city block distance or L1 distance. The Manhattan distance is calculated as the sum of the absolute value of the difference in frequency between the two words at each of the 600 positions of the critical site region.

$$d_1(\vec{p},\vec{q}) = \|\vec{p}-\vec{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i| \qquad \text{Eq 2}$$

We use the *k*-means algorithm to cluster similar distributions (MacQueen 1967). The data was clustered with a range of values for *k*, 1–100. The optimal value for *k* was determined empirically by choosing a value of *k* that maximizes the CH index, a metric that scores the compactness of clusters and intercluster distances (Ars 2000). High CH indices are achieved with well clustered data (i.e., dense clusters that are spaced well apart from each other). In our evaluation, each k-means computation is run with 100 different

initial configurations to reduce artifacts created by nonoptimal initial centroid locations.

In addition, L1 distance metric is ranked and compared with the neighborhood inference (NI) set of exonic splicing enhancers (ESE)/exonic splicing silencer (ESS) hexamers from the work of Stadler et al. (Stadler 2006). Using their suggested cut-off scores, the dataset is comprised of 979 ESEs (NI score >= 0.8), 496 ESSs (NI score <= 0.8), and 2,621 neutral elements (NI score anywhere in between 0.8 to –0.8). We separated our pairwise distance comparisons (L1 distance metric; Eq. 2) into 2 distinct sets: one consists of the top 25% of highest L1 distances (hypothesized to be functional change) and the other consists of the bottom 25% (hypothesized to be neutral change). We then asked if the two distinct sets exhibited different functional changes according to the NI set referred above. Given two hexamers, a functional change corresponds to any changes on the NI set (i.e., ESE to ESS or vice versa, ESE to neutral or vice versa, ESS to neutral or vice versa), whereas a neutral change corresponds to no changes on the NI set (i.e., ESE to ESE, ESS to ESS, or neutral to neutral).

2.3.4 Minigene reporter experiments. All splicing reporter experiments used variations of the pZW4 splicing reporter minigene (i.e., Fig. 1 and “wt” vector in Fig. 4) (Wang 2004). Additional constructs with variations characterized as splicing mutations in prior reports (Chen 1993) were designed as sensitized reporters to test the affects of point mutations on inserted sequences. Inserts were selected on the basis of their match to the cluster motif. For each insert we can rank all possible substitutions by their difference in L1 distance. The most extreme difference between wild-type and mutant hexamers

represents M2, the point mutation that the method would predict most likely to disrupt a motif. Conversely, the most similar distributions, M1, would be predicted to function similarly to the wild type sequence. These variations were introduced into the reporter, transfected into 293 cells and assayed by RT-PCR. Both alleles of missense mutations were tested with 15 nt of flank as a 31-mer ligated into the minigene. We selected mutations with high L1 intraallelic distances (top 50) and some evidence of alternative splicing, reasoning that variants in these cases would be more sensitive to alteration in splicing definition and, therefore, have a greater chance of being modeled by the sensitized minigene reporter assays.

2.3.5 Comparing disease-causing mutations with random genetic variation. All disease-causing splicing and missense/nonsense mutations within 200 nucleotides of a splice site in the intron or within 100 nucleotides of a splice site in the exon were down-loaded from the Human Gene Mutation Database (HGMD). For each of these observed point mutations, we extracted a wild type and a mutant 11-mers (the wild-type and mutant allele of the point mutation each flanked by five base pairs of sequence). Each 11-mer sequence was then fragmented into hexamers. In these tiled paths of singly shifted hexamers, a nucleotide substitution alters six hexamers. Each mutant hexamer was compared to its wild-type counterpart. We reasoned that a low L1 distance would reflect a similar distribution and function and would therefore be predicted neutral with regards to splicing whereas a large L1 distance should have an impact on splicing. For a given point mutation, the representative L1 distance was taken to be the largest intraallelic distance of the six distances calculated by comparing each tiled wild-type hexamer with its mutant counterpart. L1 distances were calculated in this way for the 8,027 disease-causing splicing mutations and 42,532 missense/nonsense mutations

downloaded from the HGMD.

To determine if the L1 distances of HGMD mutations were higher than expected by chance, we created a background model of random genetic variations. For each HGMD mutation, we chose a random position within the same region (upstream intron, exon, or downstream intron) and mutated it using equal rates of transversions and transitions. L1 distances were calculated (as above) for these random mutations. In this way, a background dataset for each class of HGMD mutants was created. To gain insight into the distribution of L1 distances for the HGMD and background mutations, averages were calculated with the L1 distances for background vs. HGMD mutations for the four classes of mutants: (i) exonic missense/nonsense mutants, (ii) splicing mutants in the intron preceding the 3′ss, (iii) exonic splicing mutants, and (iv) splicing mutants in the intron following the 5′ss (Fig. 4A).

2.4 Results

2.4.1 The splicing activator, U2AF65, inhibits splicing when bound at an exonic site. To test the relationship between the function of a splicing factor and the location of its predicted binding element, we initially focused on one well-characterized factor-ligand binding event, U2AF65's recognition of the polypyrimidine tract. The binding motif consists of a Poly U-rich tract that typically contains runs of four or five uridines followed by cytosine frequently initiated with a G (Fig. 1A). Mapping U2AF65's binding motif across all exons revealed the largest peak occurring immediately upstream of the 3′ splice site (3′ss). This location was consistent with its role as the principal recognizer of the polypyrimidine tract. The U2AF motif was overrepresented in the regions where it was known to function positively (i.e., in 3′ss recognition) and depleted in the exon (where U2AF binding has not been shown to support the normal spliceosomal complex). This suggested that the positional distribution pattern of an element around the splice sites was indicative of the transacting factor's function in splicing.

To experimentally test the role of the binding location of a particular factor in splicing function, we relocated the normally positive-acting intronic U2AF65 binding site into an exonic location and assayed splicing. For this study we utilized two polypyrimidine tracts. One tract was a synthetic consensus U2AF65 binding site derived from a Systematic Evolution of Ligands by Exponential Enrichment (SELEX) study and another was a natural polypyrimidine tract located upstream of the 3′ss of exon 5 of the KCNN1 gene (Singh 1995). UV cross-linking indicated that numerous cellular proteins contacted both probes after incubation. The 65 kD interaction was blocked by preincubation with anti-U2AF65 antibodies thereby establishing specific U2AF65 contacts with the polypyrimidine tract with both of these inserts (Fig. 1B lanes 2 and 4 compared to no

antibody control lanes 3 and 5) but not in the “no insert” control (Fig. 1B, lane 1).

The sequences used to probe binding were then assayed for function in the test exon of pZW4, an in vivo splicing reporter. The splicing phenotype was assayed by RT-PCR from total RNA following transfection into 293 cells. Whereas the no insert control spliced normally (Fig. 1C, black arrow in lane 6), both reporters containing U2AF65 binding elements exhibited evidence of disrupted splice site recognition by skipping exon 2 in some fraction of the transcripts observed. The polypyrimidine tract from the KCNN1 gene also generated an intron inclusion product and several other aberrant species that were not characterized. This result demonstrated that U2AF65, a factor with a well-characterized role of activating splicing when bound in the intron, disrupts splicing when bound in the exon.

To determine if the relationship observed between U2AF65 binding and its function was general, we expanded our analysis to some members of the SR and hnRNP protein family. As SR proteins are generally regarded as activators that function by binding exonic splicing enhancers, we examined the positional distribution of the in vitro SELEX-derived position weight matrix for three SR proteins: ASF/SF2, SC35, and 9G8 (Fig. 5) (Cavaloc 1999; Tacke 1995). Three hnRNP proteins were also analyzed in this study: hnRNP A1, hnRNP L, and hnRNP C (Fig. 5) (Burd 1994; Gorlach 1994; Hui 2005). This analysis largely supported the role of SR proteins as activators that bind ESEs whereas hnRNP binding sites are located at predominantly intronic locations. Binding motifs for hnRNP C were concentrated around the 3′ss consistent with early reports of the location of hnRNP C dependant functional elements (Swanson 1988). Both hnRNP L and hnRNP A1 also bound intronic elements albeit further away from the splice sites. The analysis of the binding sites of known splicing factors revealed a nonuniform positional distribution

that was indicative of their function.

If the position of a splicing motif relative to a splice site is a signature of that motif's function in splicing, then motifs with similar positional distributions should play similar roles in splicing and motifs with different positional distributions should play different roles in splicing. Therefore, by clustering the motifs according to their positional distribution around splice sites, we expected to organize elements into distinct functional classes.

2.4.2 Clustering words by positional distribution recovers splicing elements. We developed an algorithm to cluster sequence motifs according to their positional distribution around splice sites. We first tabulated the frequency of every possible sequence motif around all the annotated splice sites in the human genome. This was accomplished by mapping 4,096 hexamers to all three hundred nucleotide windows around annotated 3′ss. This mapping associated each hexamer with a vector that contained the genomic occurrence of that hexamer at each position around all the 3′ss. This 300 unit long vector had a first position of -200 and a last position of +99 relative to the 3′ss. Counts were normalized to enable comparisons between hexamer positional distributions based on shape and not frequency. Repeating this procedure for the regions around the 5′ splice sites (5′ss) created a second vector that together with the 3′ss vector were used to summarize the positional distribution of hexamers around exon junctions in the human genome.

The overall goal of this method was to cluster hexamers into subsets that shared a similar positional distribution. This clustering required a method for pairwise comparison of two shapes. The difference in positional distribution shapes between two hexamers

was calculated by determining the L1 distances between all possible pairwise combinations of these 4,096 vectors (Fig. 2A and Eq. 1). In a graph of normalized hexamer counts, L1 distance is simply the area between two positional distributions (shaded blue in Fig. 2A). These L1 distances were used to cluster (*k*-means) the hexamers into 51 distinct groups. The optimal value of k was determined by the CH index (Calinski 1974). The hexamers within each cluster were aligned without gaps and displayed as pictogram motifs (Fig. 2C). The resulting motifs returned by this analysis had distinct positional distributions around the 3′ and 5′ss (Fig. 2C).

An immediately obvious feature of all 51 clusters was the sequence similarity between the hexamers that clustered together. In other words, hexamers that were highly similar in positional distribution were also highly similar in sequence. Hamming distance (i.e., the number of shifts or mismatches in the optimal ungapped alignment of two hexamers) was used to compare the sequence similarity of hexamers within a cluster. Intracluster similarity of hexamer sequence was much higher than expected by chance (all p values < 0.01; 1,000 trials per cluster, 51 clusters). As there is no a priori reason for similar sequences to share similar positional distributions relative to splice sites, we interpreted the strong sequence motifs found in the clusters as binding motifs of splicing factors that function at an optimal distance from a splice site. Consistent with this observation, we found motifs that match the known canonical splicing elements (i.e., branch point, polypyrimidine tract, 3′ss, and 5′ss) at the correct location relative to exon/intron boundaries (Fig. 2C). Cluster 24 peaks at position -26 nt and represents the branchpoint sequence with a core TRAY motif flanked by extended complimentarity to U2snRNA (i.e., 4 nucleotides upstream and 3 nucleotides downstream of the bulged A). It is important to note that the motif returned by this algorithm is a far better fit to the

known mechanism of U2 snRNA mediated branch point recognition than motifs built from alignments of experimentally defined branchpoints. Similarly, the 5′ss motif (cluster 51, Fig. 2C) contains GTAAGT—a perfect stretch of complementarity to the mammalian U1 snRNA. Interestingly, this motif is avoided in the downstream exon proximal to the bona fide 5′ss. The polypyrimidine tracts are U-rich and covered by several clusters. A motif identical to the U2AF65 SELEX result (Fig. 1A) was found. The 3′ss AG and the polypyrimidine tract cluster separately presumably because of the variable spacing often found between these elements in natural splicing substrates and because they are recognized by separate factors.

2.4.3 Point mutations that create mutant hexamers with large L1 distances from wildtype hexamers alter splicing *in vivo*. To validate elements from different clusters in vivo we assayed their effect on exon inclusion in a variety of splicing reporter minigenes. Test cases (exemplars) chosen to represent a cluster were cloned into reporter constructs, transfected into 293 cells and assayed by RT-PCR. To determine if the positional distribution distance measurements used in the clustering were predictive in identifying substitutions that disrupt a splicing element, we selected point mutations based on the degree to which they shifted the intraallelic L1 distance of the insert. There are eighteen different point mutations that can be introduced into a hexamer. Each of these mutations creates a new hexamer with a different positional distribution around splice sites. Substitutions with a large L1 distance were predicted to be most likely to disrupt splicing. Ranking all possible point mutations by L1 distance we found the top 25% to have twice as many ESE or exonic splicing silencer (ESS) changing mutations than the bottom 25% of this ranked list (Stadler 2006). We used L1 distance to design predicted splicing mutants for functional analysis in splicing reporter constructs (Fig. 2C).

This analysis was performed for exemplars drawn from three clusters that represented unique splicing elements. For all three exemplars, the inserts and mutants spliced normally when ligated into the vector that contained wild-type splice sites (Fig. 3B, lanes 2, 3, 8, 9, 14, and 15). However when introduced into the context of mutation NS92 where the test exon was weakened by a mutation in the 5′ss, two of the three wild-type/mutant pairs displayed divergent splicing phenotypes (i.e., the wild-type sequence spliced differently than the predicted point mutant for cluster 30 and cluster 29—Fig. 3B, lanes 5, 6, 11, and 12). Neither the wild type nor the mutant of cluster 35 affected splicing (C35.1 in Fig. 3B). To see if the results observed in the mutant context of NS92 were general, we repeated the assay with different cluster exemplars (C35.2 and C30.2 in Fig. 3C) and different mutant context (NS20—weakened polypyrimidine tract) with identical results. This consistency between exemplars across different conditions suggested that the clusters are effectively characterizing the splicing activity of sequence elements. It is, however, possible that any variation in the sequence would disrupt this splicing activity. To establish the specificity of this prediction we tested variations that would be predicted to be neutral (i.e., variations in the same hexamer that results in low L1 distances). In all cases examined, these negative control (M1) mutants were spliced similarly to wild-type inserts in the splicing assay. The wild-type splicing pattern was similar to the predicted neutral mutant (Fig. 3C, lanes 7 and 8 and lanes 10 and 11). The mutation with high L1 distance was spliced differently than both the wild type and predicted neutral mutations (Fig. 3C, lane 9 versus lanes 7 and 8).

Exemplars were also selected from two additional clusters that represent a variety of intronic splicing enhancers (i.e., positional distributions are enriched in the intronic regions). The predicted neutral mutants (M1) were spliced similarly to wild type (Fig. 3C

comparing lanes 13 and 14, 16 and 17, 19 and 20, and 22 and 23), whereas the change-of-function mutants (M2) were spliced differently (Fig. 3C comparing lanes 13 and 15, 16 and 18, 19 and 21, and 22 and 24). In both cases, mutating an intronic element in the exon exhibited positive splicing phenotypes.

2.4.4 High Intraallelic Distance Is Predictive of Splicing Mutations. To test the predictive power of using intraallelic L1 distance to discover splicing mutations, we computed the intraallelic L1 distances of splicing mutations that were downloaded from the Human Gene Mutation Database (HGMD). Disease-causing alleles specifically associated with splicing exhibited significantly higher L1 distances than simulated mutations (p-value < 0.001 for the upstream intron, exon, and downstream intron) (Fig. 4A). The simulation incorporated mutational bias toward transitions (see Materials and Methods). Interestingly missense disease alleles downloaded from HGMD also displayed a significantly higher intraallelic L1 distance than expected (p-value < 0.001). This data suggests that even human disease alleles located outside of the canonical splice sites are more likely to cause aberrant splicing than natural variations that do not cause disease. We roughly estimated the fraction of splicing mutants by modeling the missense category of HGMD mutations as a mixture of exonic HGMD mutations that are known to cause splicing defects and simulated mutations (which are presumed not to cause splicing defects). In other words a hypothetical set comprised of 78% simulated mutations and 22% splicing mutants had the same average intraallelic L1 distance as the HGMD missense mutants. Accounting for these mutants along with HGMD entries that were formally classified as splicing mutants suggested that about a third of all disease-causing mutations display some sort of aberrant splicing phenotype. To explore the usefulness of L1 distance in predicting splicing mutations, we performed receiver

operating characteristic (ROC) curve analysis, comparing the true to false positive rates at different thresholds of L1 (Fig. 4B). The ROC curve analysis suggests that an L1 prediction threshold that can identify 50% of the exonic splicing mutations in a sample (i.e., $y \approx 0.50$ in Fig. 4B), would also return 20% false positives (i.e., $x \approx 0.2$). This analysis demonstrated that the model was significantly predictive of splicing mutants—especially 5′ss and exonic mutants (Fig. 4B). As the later category of exonic mutants falls outside of the well-defined canonical splice sites, there are few other options to evaluate the effect of mutations. This method could be applied to finding splicing mutations in exons. To investigate this idea that missense mutations disrupt splicing, we tested six missense mutations with high L1 distances in the minigene splicing assay (Fig. 4). RT-PCR analysis of these exemplars uncovered an obvious difference in splicing between wild-type and mutant inserts in four of the six exemplars tested (Fig. 4C). This data confirmed the presence of processing mutations in exonic mutations. A web interface has been written to facilitate the analysis of variations in human pre-mRNA (http://fairbrother.biomed.brown.edu/data/mutations).

2.5 Discussion

In the output of the clustering, the canonical splicing elements segregated into discrete clusters. Strong 5′ss motifs (cluster 51) and 3′ss motifs (cluster 14) emerged as independent clusters. The hexamers in cluster 27 represented the polypyrimidine tract with their well-characterized signal located 4–20 nucleotides upstream of the 3′ss (Fig. 2C). Clusters 23 and 24 both appeared to fit the T(A/G)A(C/T) of the eukaryotic branchpoint sequence. ESEs mostly fell within 5 clusters (clusters 29–33, Fig. 2B). Further sorting the ESE hexamers into five prime specific ESEs, 3′ splice site ESEs and shared ESEs revealed that ESEs specific to the 3′ss fell mostly within cluster 30 and the smaller 5′ss specific ESEs segregated into cluster 29. In addition to ESEs, a variety of intronic splicing enhancers (ISEs) could be recognized within the cluster results. A prominent ISE, the G triplet, was found in cluster 8 (Caputi 2001; McCullough 2000; Reid 2009; Yeo 2004). We found G triplets and C triplets to possess distinct nonoverlapping positional distributions around human splice sites (compare cluster 8 to cluster 35). Whereas both C and G triplets have a predominantly intronic positional distribution, C triplets tend to occur closer to the splice sites than G triplets. C triplets could be a recognition element for a protein like hnRNP C. Like many intronic enhancers, both C and G triplets occur at lower frequency on the exonic side of splice sites suggesting that they are not tolerated in the constitutively spliced exons that comprise the majority of the database used in this study. We did not find that mutations in exonic C triplets alter their effect on splicing (Fig. 3). C triplets may require other splicing elements for their activity and cannot function in isolation in a minigene. One candidate for this auxiliary element is the G triplet as these elements cooccur. C triplets are predominantly located upstream of the 3′ss, roughly around 30 nucleotides downstream of the local G triplet peak. Across

the database, 22% of introns have G triplets between positions -65 and -50 relative to the 3′ss. If the intron contains a C triplet, the likelihood of a G triplet increases from 22% to 34% (p-value ≈ 0, chi-square test). It is possible that this co-occurrence may reflect a function synergy such as their potential to form structure or a larger ribonucleoprotein (RNP) complex through their transacting factors.

The general observation of intronic motifs that increase in frequency with decreasing distance to the splice site and then decrease in frequency when approaching the splice site from the exonic side is not consistent across all motif classes. Certain motifs (cluster 17) appear to increase in frequency with decreasing distance to the splice sites on both the intronic and exonic side of the junction. This type of distinction would not have been discovered by previous computational approaches. One possible explanation for this outlier might be that this motif is not an RNA element but rather a recognition element for a DNA binding protein. Polymerase pausing and chromatin formation with specific histone modifications are two DNA binding phenomena that have been implicated in enhanced splicing (Kadener 2002). A/T rich elements are often found in recognition sites of DNA bending proteins or could form the weak RNA:DNA duplexes that promote the polymerase backtracking associated with some types of transcriptional pauses (Kulish 2001).

Although describing the mechanism of each element is beyond the scope of this study, we demonstrate that mutations that are disruptive to positional distribution are disruptive to splicing. We also find evidence that missense mutations that cause human disease are more likely to disrupt splicing than simulated mutations. Because of the difficulty of assaying splicing in patients, very little is known about the prevalence of splicing defects in human disease. About 15% of the mutations in the HGMD are described as splicing

mutants (Stenson 2003). Some have been validated directly but many of these mutations colocalize with critical regions of splice sites and so are assumed to disrupt splicing. A more problematic class of identification is the set of mutations that fall outside of well-defined sites. It is possible that many of these disease alleles are associated with subtle defects in splicing that could exacerbate the disease phenotype. Using an approach that models the missense mutations as a mixture of exonic splicing mutants and simulated mutations, we estimate that 22% of missense disease alleles alter splicing. A reanalysis of missense mutations supports the notion that many disease alleles originally classified as missense also disrupt splicing (Ars 2000). Furthermore, another recent study finds a similar fraction (i.e., 25%) of > coding mutations alters splicing (Sterne-Weiler 2011). This class of "undiagnosed" splicing mutations along with known splicing mutations predicts that about one third of all mutations alter splicing.

It is important to be able to identify the many human disease alleles that alter splicing and characterize missense mutations for their effect on pre-mRNA processing. In the future, new molecular therapies that correct splicing defects may ameliorate many genetic disorders (Hua 2010). The ability to correctly identify splicing mutations by their elevated L1 distance and the ability to predict mutations in the minigene system demonstrate that this is a useful tool in predicting causal alleles.

2.6 Figures and tables

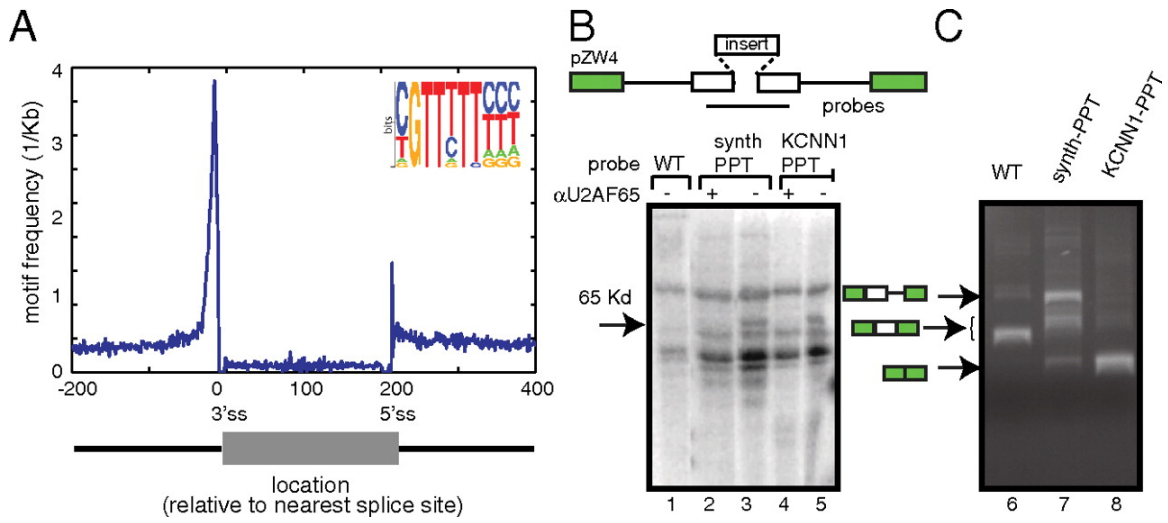


Figure 2.1. Exonic binding of the intronic activator, U2AF65, inhibits splicing. (A) SELEX motifs were mapped to a dataset of 312,275 human splice site regions and plotted on an amalgamated exon. (B) The synthetic polypyrimidine tract returned by the SELEX consensus U2AF65 motifs and a genomic polypyrimidine tract were ligated into an exon and tested for U2AF65 binding by UV cross-linking in extract without antibody (lane 1, 3, and 5) or in extract that was blocked by an anti-U2AF65 antibody (lane 2 and 4). The radiolabel transferred to several products of differing mobility—a 65 kD interaction that was sensitive to preincubation with antiU2AF65 antibody is indicated with an arrow. (C) The sizes of RT-PCR products reflecting varying degrees of splicing are shown by the arrows. The disruptive effects of ligating the synthetic and natural PPT into the test exon of pZW4 is shown by RT-PCR in lane 7 and 8.

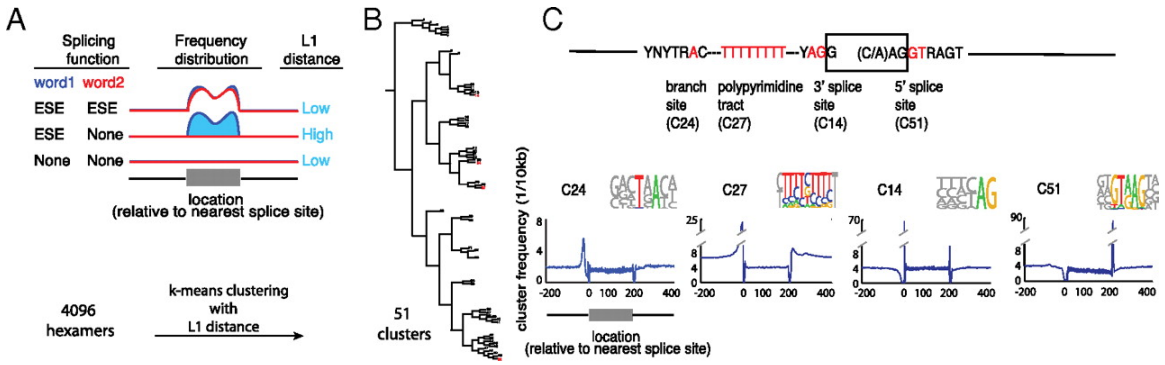


Figure 2.2. Clustering motifs according to their positional distribution around splice sites. The positional distributions of all 4,096 possible hexamers were plotted around a database of human splice sites. (A) Several comparisons of two hypothetical hexamers (word 1 and word 2) are drawn to illustrate three different scenarios. L1 distance (shaded blue area) is used to compare normalized frequency distributions. Low L1 distance indicates there are small differences between two positional distributions and the two hexamers have the same or no difference in splicing function. High L1 distance denotes the two positional distributions are vastly different and likely differ in their role in splicing. (B) L1 distance was used to cluster the hexamers into 51 distinct groups based on the shape of their positional distributions around splice sites. Motifs and positional distributions of all 51 clusters can be found in the supplement. The clusters that correspond to the canonical splicing elements are indicated in red. (C) The arrangement of these elements on a prototypical pre-mRNA is annotated on the exon diagram. Hexamers within these clusters were aligned into motifs. Average occurrence frequencies of all the cluster's hexamer were calculated at each position around the splice site database.

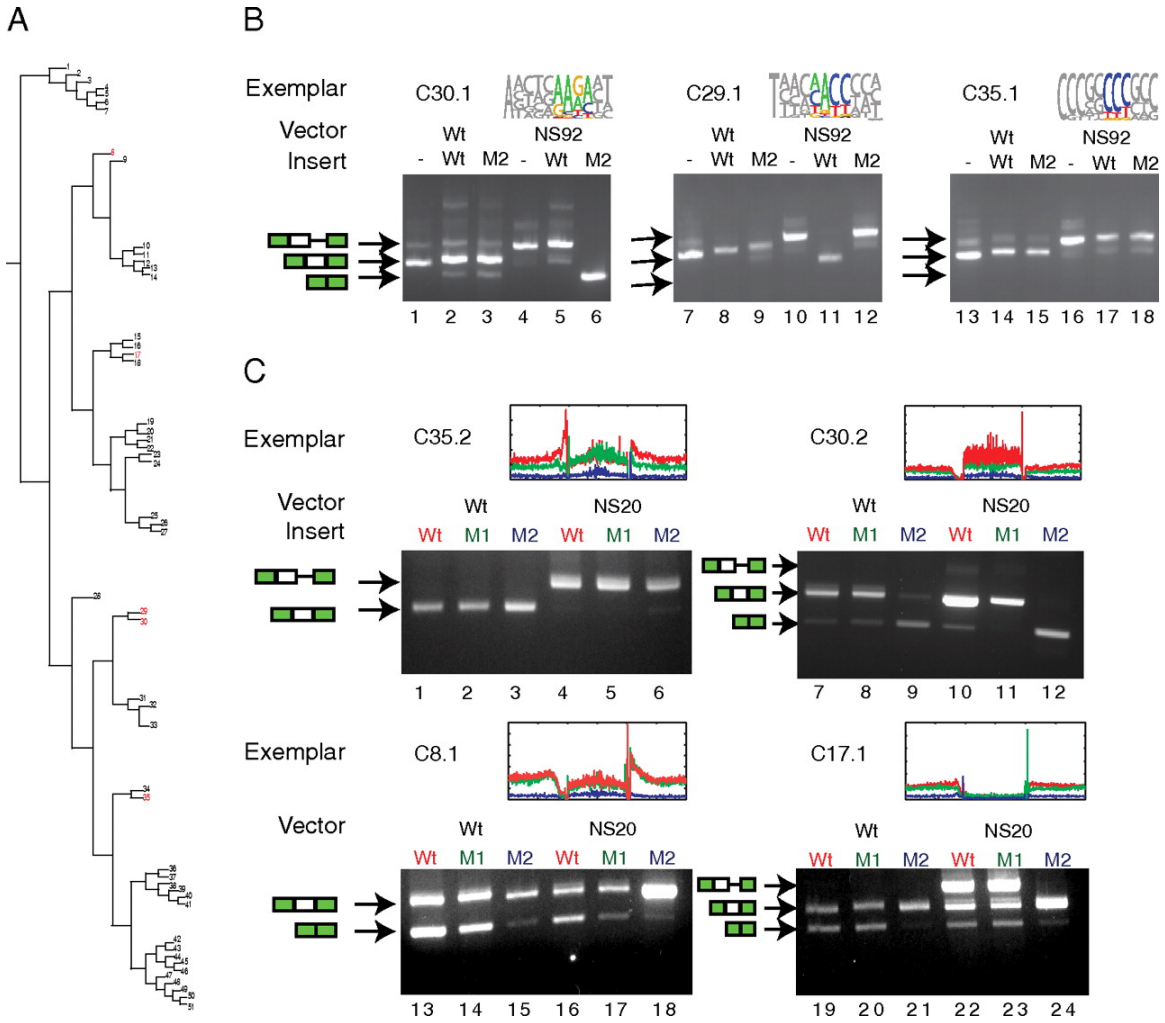


Figure 2.3. Minigene assay of element function confirms splicing differences between wild-type cluster exemplars and predicted mutants. (A) The clusters selected for functional analysis are indicated in red. (B) Exemplars drawn from each cluster are tested with their variants and no insert controls in several splicing reporter constructs. Total RNA from transfection into 293 cells was analyzed by RT-PCR. Arrows indicate the nature of the splicing product. M2 denotes the point mutant with the highest intraallelic L1 distance predicted to be most deleterious to the splicing function of the wild-type insert. (C) Additional exemplars for clusters 30 and 35, along with exemplars for clusters 8 and 17 were used to contrast the effect of predicted neutral mutations (M1) or the effect of predicted change-of-function mutations (M2) with wild-type splicing. As before, the M2 mutation is the variation with the highest intraallelic L1 distance, and the negative control, the M1 mutation, has the lowest intraallelic L1 distance.

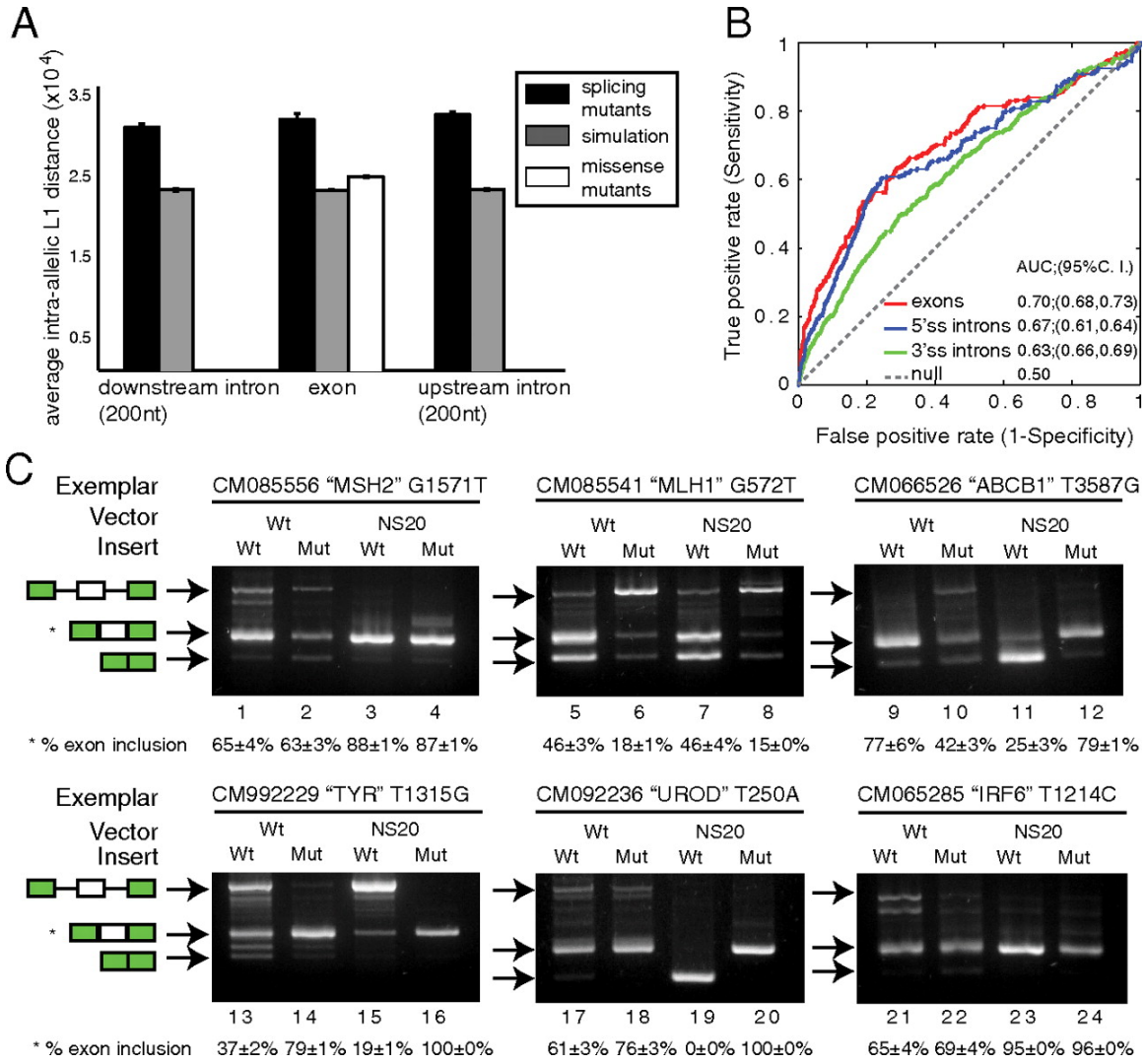


Figure 2.4. Human disease alleles are predicted to disrupt splicing. (A) Average intraallelic L1 distances for each category of mutation (HGMD splicing and HGMD missense/nonsense) and their corresponding background models of simulated mutations divided by location with respect to the splice sites. Error bars denote 95% confidence intervals. (B) Receiver operating characteristics (ROC) curve analysis using HGMD splicing mutants in regions around the 3′ss and 5′ss as "true positives" and simulated mutations as "true negatives." ROC curve analysis classifies these mutations at decreasing thresholds of L1 stringency plotting the false against true positive rates. The exonic region is shown in red; upstream and downstream intronic regions are shown in green and blue, respectively. (C) Exemplars were selected from the HGMD missense mutants with the highest intraallelic L1 distance. Total RNA from transfection into 293 cells was analyzed by RT-PCR. The HGMD ID, gene name, and the mutational position are shown for each experiment. Quantifications on exon inclusion products are also shown. Arrows indicate the identity of the splicing product.

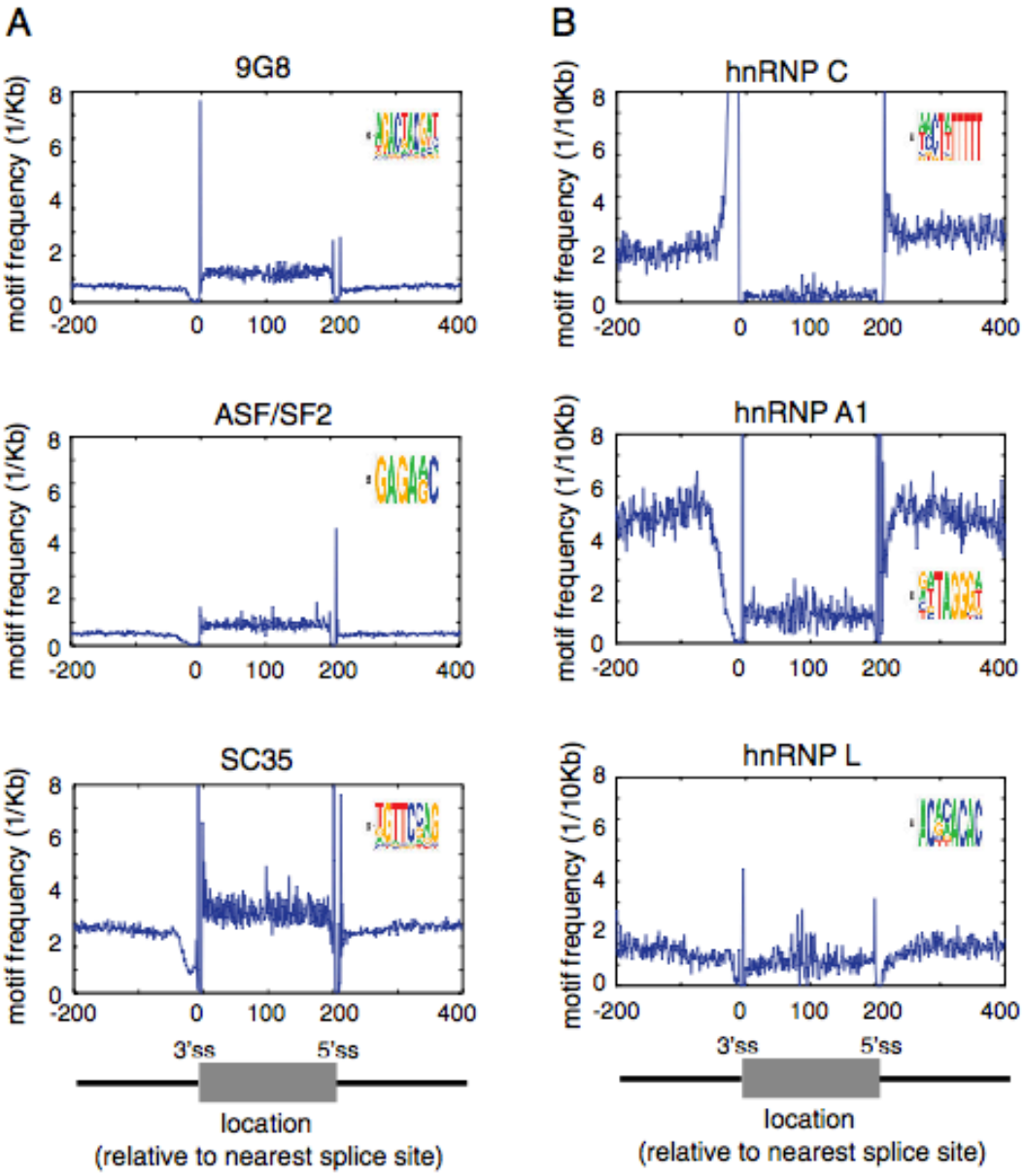


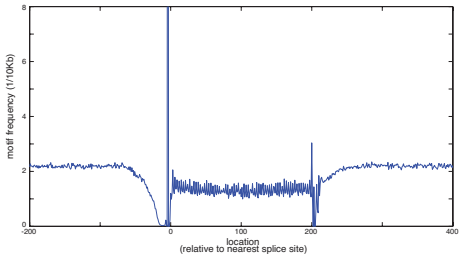
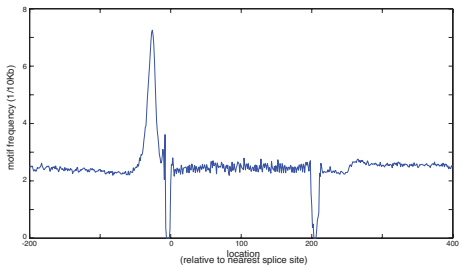
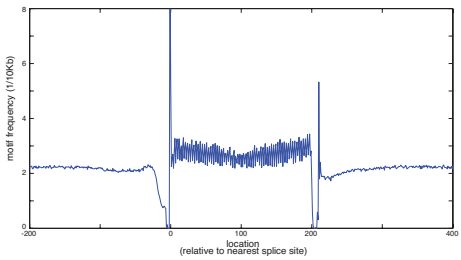
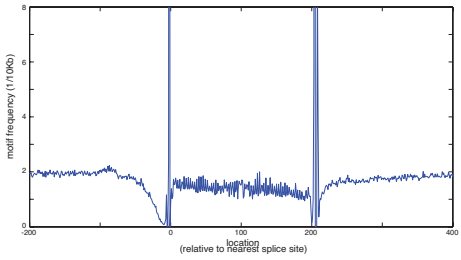
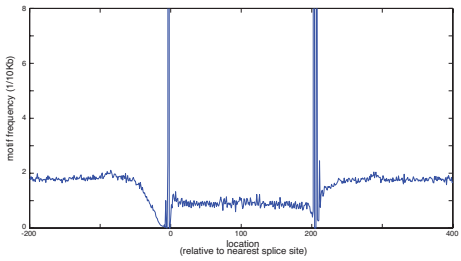
Figure 2.5. Binding motifs for serine/arginine rich (SR) proteins are exonic and heterogeneous ribonucleoprotein (hnRNP) motifs are intronic. SELEX motifs for (A) SR proteins and (B) hnRNP proteins were discovered from prior SELEX studies (see Material and Methods) and mapped to a dataset of regions around 312,275 human splice sites. Motif frequency was plotted on an amalgamated exon graph as a function of distance to the nearest splice site (as in Fig. 1A).

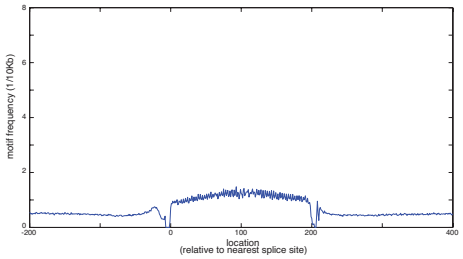
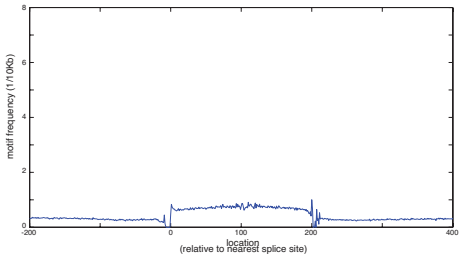
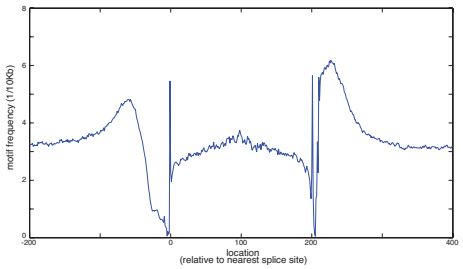
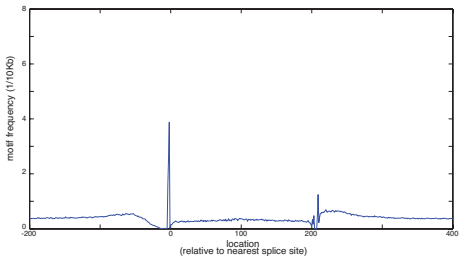
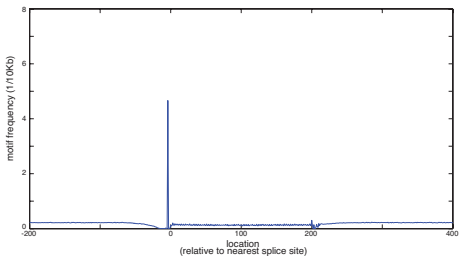
Table 2.1. Oligonucleotides used in these experiment.

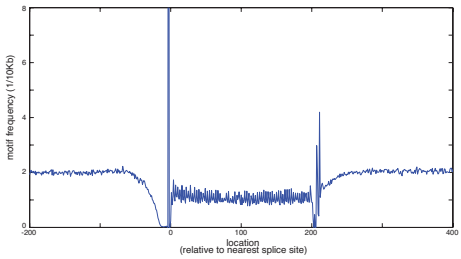
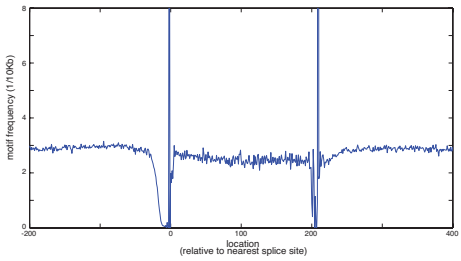
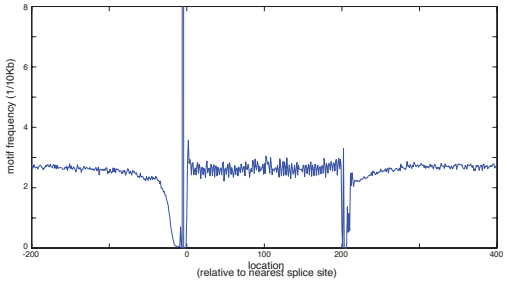
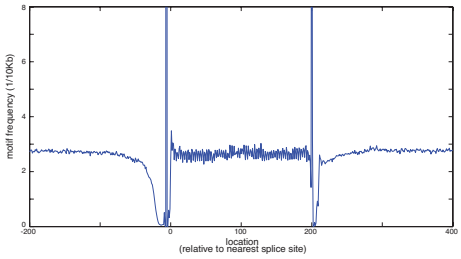
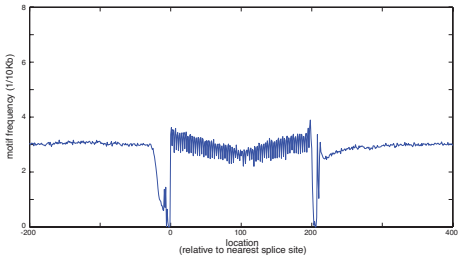
Name	**wt**	**mt_1**	**mt_2**
synth-PPT	CTTTTTTTTTCCTTTTTTTTTC		
KCNN1-PPT	CTTGCCTTGTGCTTTTCTCCTGCTCTGCCCC		
C17.1	CTATTTAAAAAAAA	CTATTAAAAAAAAA	CTATTTAAACAAAA
C29.1	ATCTCTGCCCACCACCCCC		ATCTCTGCCGACCACCCCC
C35.1	GCCCCCGCCCCGCCGTGTG		GCCCCCGCCACGCCGTGTG
C8.1	TGGCCCTGGGGGTC	TGGCGCTGGGGGTC	TGGCCCTCGGGGTC
C35.2	TGTGCCTGCCACGC	TGTGGCTGCCACGC	TGTGCGTGCCACGC
C30.1	TTTGCACTCAAGACCTACT		TTTGCACTCTAGACCTACT
C30.2	GGCACTGGAGTTTG	GGCACTGGAATTT	GGCACTGGCGTTTG

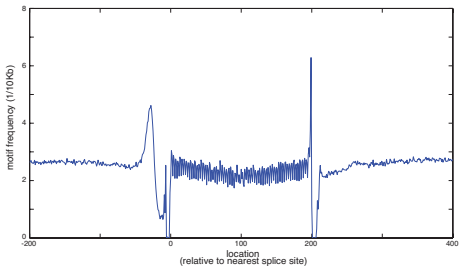
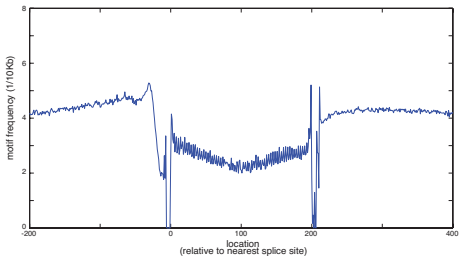
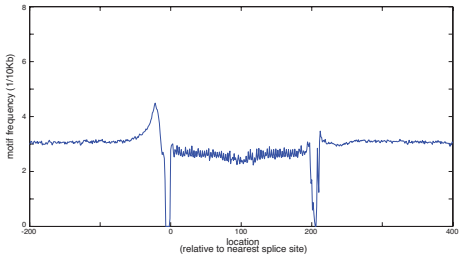
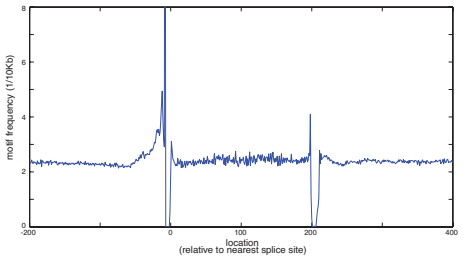
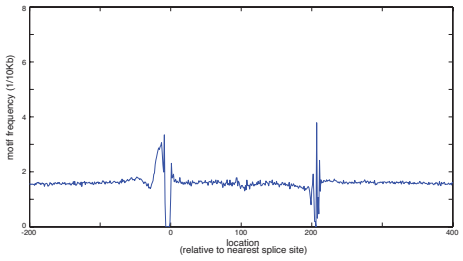
		G	
G1571T	TTGGATATTACTTTCGTGTAACCTGTAAGGA		TTGGATATTACTTTCTTGTAACCTGTAAGGA
G572T	ACAATGCAGGCATTAGTTTCTCAGTTAAAAA		ACAATGCAGGCATTATTTTCTCAGTTAAAAA
T3587G	TTAGACAGCCTCATATTTTGCTTTTGGATGA		TTAGACAGCCTCATAGTTTGCTTTTGGATGA
T1315G	AGAAATGGTGATTTCTTTATTTCATCCAAAG		AGAAATGGTGATTTCGTTATTTCATCCAAAG
T250A	GATGCTGCCATCATTTTCTCCGACATCCTTG		GATGCTGCCATCATTATCTCCGACATCCTTG
T1214C	GGATGATCTACGAGATGTTTTCTGGTGATTT		GGATGATCTACGAGACGTTTTCTGGTGATTT

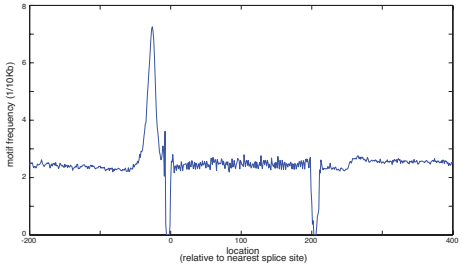
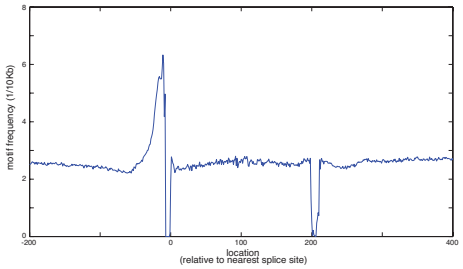
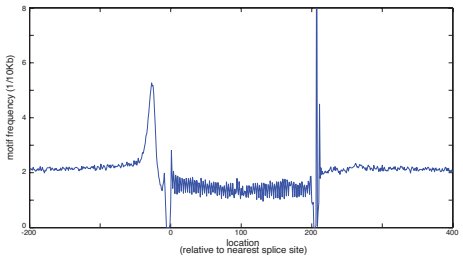
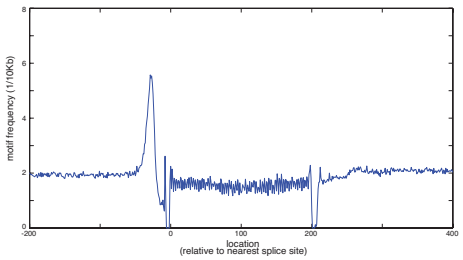
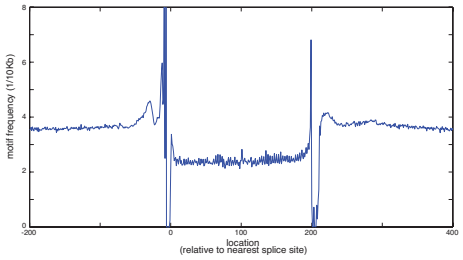
Table 2.2. Clustering results of this study.

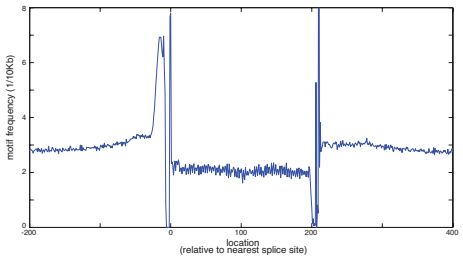
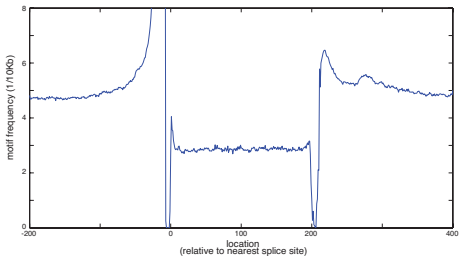
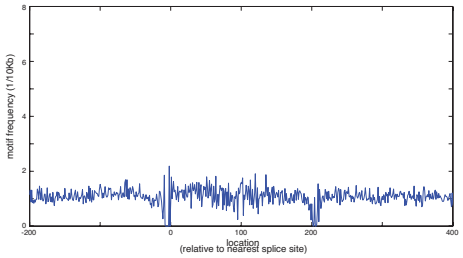
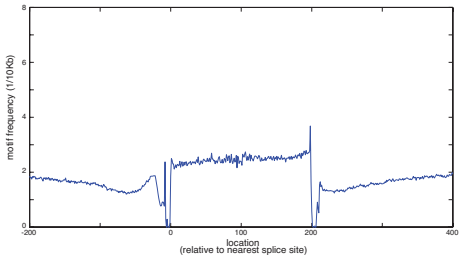
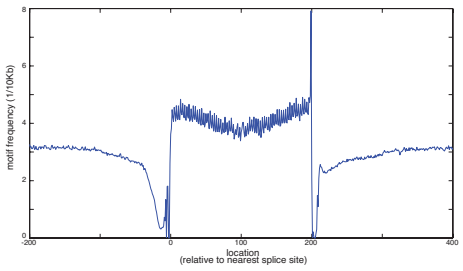
Cluster	Total elements	Motif	Average frequency distribution profile
1	9		
2	57		
3	68		
4	93		
5	98		

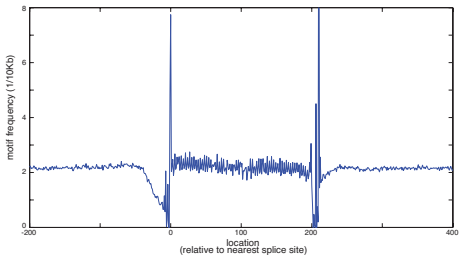
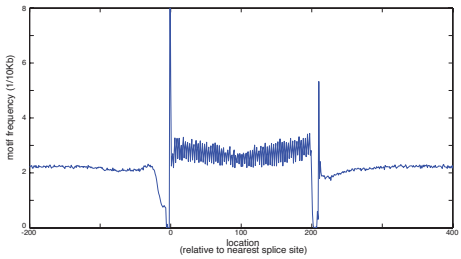
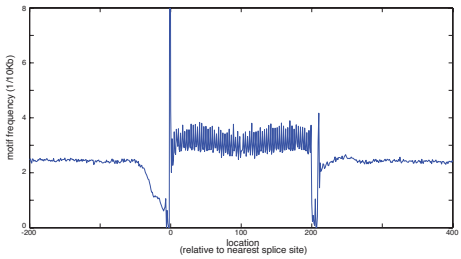
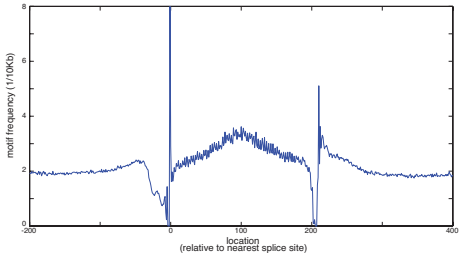
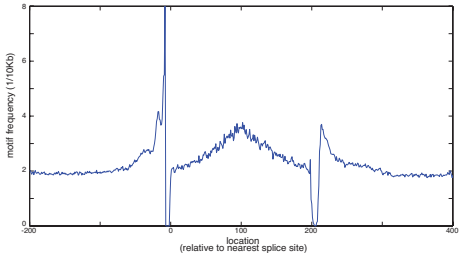
6	89		
7	102		
8	115		
9	20		
10	50		

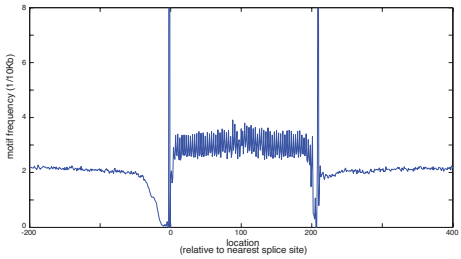
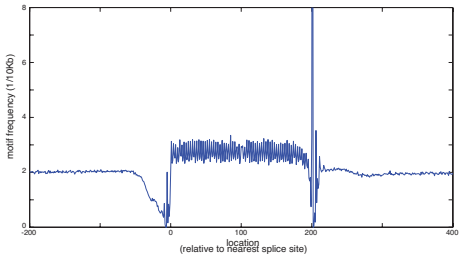
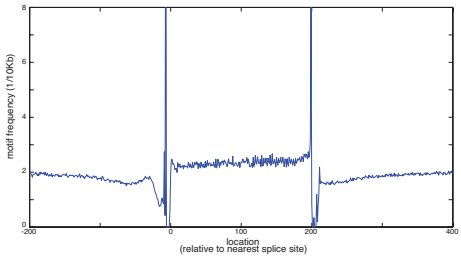
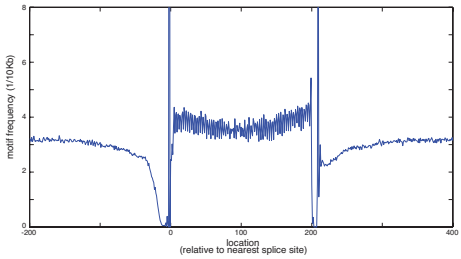
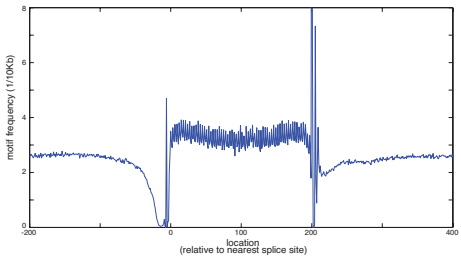
11	43		
12	46		
13	59		
14	107		
15	136		

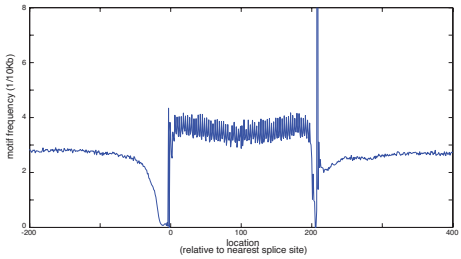
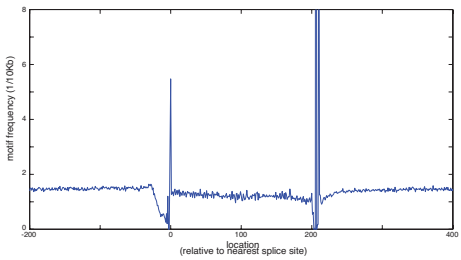
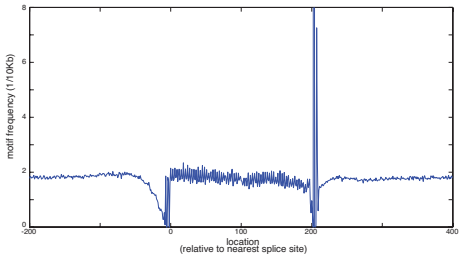
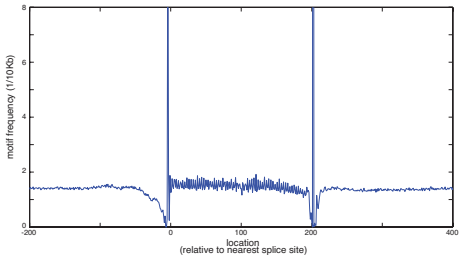
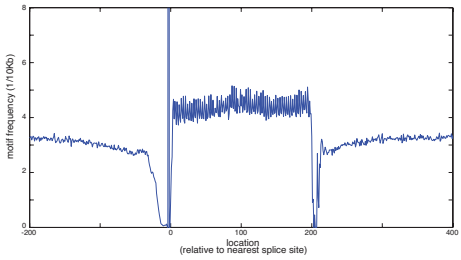
16	74		
17	83		
18	151		
19	80		
20	91		

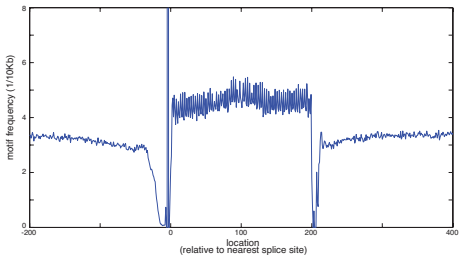
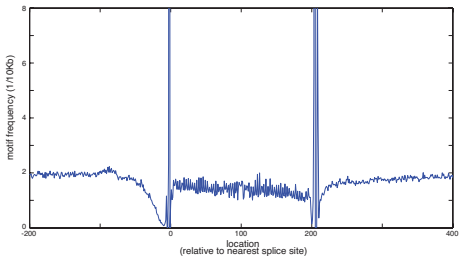
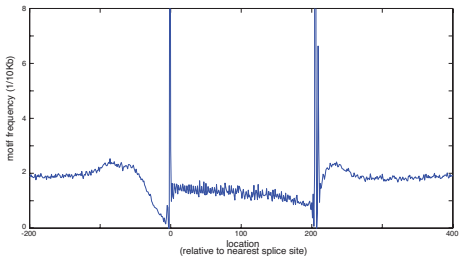
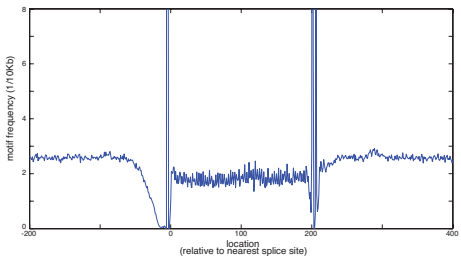
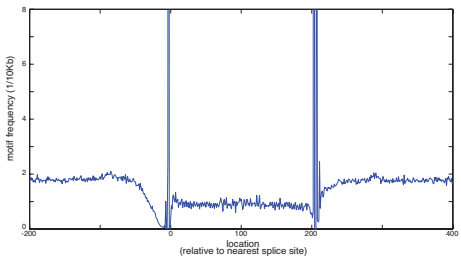
21	71		
22	138		
23	57		
24	40		
25	101		

26	66		
27	229		
28	2		
29	98		
30	140		

31	60		
32	130		
33	109		
34	97		
35	65		

36	82		
37	133		
38	94		
39	80		
40	86		

41	112		
42	60		
43	77		
44	89		
45	49		

46	48		
47	28		
48	39		
49	37		
50	27		

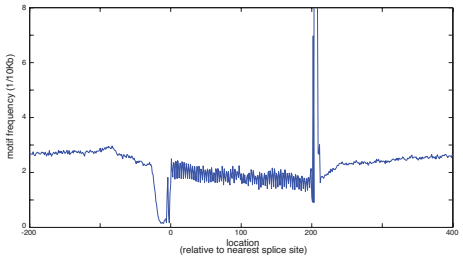
51	81		 motif frequency (1/1.0Kb) location (relative to nearest splice site)

Table 2.3. SELEX experiments used in this study. PPT, Polypyrimidine Tract; ESE, Exonic Splicing Enhancer; ISE; Intronic Splicing Enhancer.

Factors	Function	Length	Reference
U2AF	PPT	10	Singh 1995
9G8	ESE	10	Cavaloc 1999
ASF/SF2	ESE	6	Manley 1996
SC35	ESE	8	Manley 1996
hnRNP C	ISE, PPT	10	Burd 1994
hnRNP L	ISE	8	Hui 2005

Chapter 3

Spliceman – An online prediction tool to screen for sequence variations that disrupt pre-mRNA splicing

Lim KH, Fairbrother WG. (2011) Spliceman – An online prediction tool to screen for sequence variations that disrupt pre-mRNA splicing. *Bioinformatics as Application Note*. In review.

3.1 Abstract

It was previously demonstrated that splicing elements are strongly positional dependent. We exploited this relationship between location and function by comparing and contrasting positional frequency distributions between all possible 4096 hexamers around splice sites. The intra-allelic distance measure used in this study found that point mutations that produced higher distances disrupted splicing, whereas point mutations with smaller distances generally had no affect on splicing. Reasoning the idea that splicing elements have signature positional distributions around constitutively spliced exons, we introduce Spliceman – a computational web server that predicts the effect of sequence variation on pre-mRNA splicing. In addition to a distance measure, this tool utilizes positional conservation scores to assess and to predict how likely a mutation around the human splice sites is to disrupt splicing. ROC curve statistics demonstrate the predictive power of this online tool.

3.2 Introduction

Pre-mRNA splicing is an important regulatory step in gene expression pathway: non-coding sequences in pre-mRNA (introns) are removed and protein-coding sequences (exons) are concatenated to form mRNA messages. The splicing process is performed by the spliceosome, a macromolecular ribonucleoprotein complex that rivals the ribosome in size and complexity. The intricate assembly of the spliceosome is guided by the sequences in the pre-mRNA substrate such as the branch point, polypyrimidine tract, 3' and 5' splice sites, and a family of subsidiary elements known as intron and exon splicing enhancers and silencers. Estimates of the fraction of disease mutations that cause aberrant splicing had been reported to range from 15% (Stenson 2003) to 62% (Lopez-Bigas 2005).

In a previous study (Lim 2011), we demonstrated that splicing elements have signature positional distributions around annotated splice sites. The elements occur frequently where they function positively and rarely when they act to inhibit splicing. We captured these positional properties for hexamers in an index called intra-allelic L1 distance and used it to cluster genomic distributions of all possible 4096 hexamers around human splice sites. In addition to recognizing the consensus splice site sequences, our method successfully identified various classes of intronic and exonic splicing enhancers and silencers. Experimental verifications of the computational results strongly indicated the power of this method to be predictive of variations that disrupt splicing. Specifically, we found point mutations that produced higher intra-allelic L1 distances disrupted splicing in an *in vivo* minigene system, whereas point mutations with small distances generally had no affect on splicing.

To facilitate the analysis of splicing mutations, we present Spliceman – a computational web server to predict if a variation is disruptive to splicing. User-specified target (point mutations) are randomly sampled according to their intra-allelic L1 distances and positional conservation scores in order to derive an average percentile ranking for each target. Each ranking is then compared with randomly drawn targets from within the same region of the splice site to assess the importance of a user-specified target around its surrounding sequences. High L1 distance and conservation percentile rankings indicate the point mutations are more likely to disrupt splicing. To analyze the predictive power of this algorithm, we generated Receiver Operating Characteristic (ROC) curve statistics from a set of 8,027 splicing mutational targets from the Human Gene Mutation Database (HGMD) (Stenson 2003). The Area Under Curve (AUC) measurements suggest our proposed method is extremely predictive and is significantly improved by adding conservation (Figure 1). This method does not rely upon specific binding models, thus making it ideal for detecting classes of mutations that may be occurring in uncharacterized elements. For the convenience of the user, the web site includes links to other online splicing element prediction tools.

3.3 Overview of online splicing annotation tools

3.3.1 Programs to evaluate canonical splice sites.

1. **MES (MaxEntScan)**. MaxEntScan is based on the approach for modeling the sequences of short sequence motifs such as those involved in RNA splicing which simultenously accounts for non-adjacent as well as adjacent dependencies between positions. This method is based on *Maximum Entropy Principle* and generalizes most previous probabilistic models of sequence motifs such as weight matrix models and inhomogeneous Markov models.

   URL: http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html

2. **SplicePort**. SplicePort is a splice site analysis tool that makes splice site predictions for submitted sequences, and allows browsing of predictive signals and motif exploration. This collection of signals is capable of achieving high classification accuracy on human splice sites.

   URL: http://spliceport.cs.umd.edu/

3. **ASSA (Annotated Splice Site Analysis)**. ASSA provides a tool to predict the effects of sequence changes that alter mRNA splicing in human diseases. The system was designed to evaluate changes in splice site strength based on information theory-based models of donor and acceptor splice sites.

   URL: https://splice.uwo.ca/

4. **NetGene2**: The NetGene2 server is a service producing neural network predictions of splice sites in human, *C. elegans*, and *A. thaliana* DNA.

URL: http://www.cbs.dtu.dk/services/NetGene2/

5. **HSF (Human Splicing Finder)**: HSF is an online bioinformatics tool to predict splicing signals. Specifically, it predicts the effects of mutations on splicing signals or to identify splicing motifs in human sequence.

   URL: http://www.umd.be/HSF/

6. **CRYP-SKIP**: SRYP-SKIP employs multiple logistic regression to predict the two aberrant transcripts from the primary sequence. The server takes up nucleotide sequence of the mutated exon together with flanking intronic sequences and provides the overall probability of cryptic splice-site activation as opposed to exon skipping. It also shows the values of the most important predictor variables for each sequence.

   URL: http://cryp-skip.img.cas.cz/

7. **NNSplice**: NNSplice analyzes the structure of donor and acceptor sites using a separate neural network recognizer for each site.

   URL: http://www.fruitfly.org/seq_tools/splice.html

3.3.2 Programs to evaluate splicing regulatory elements (SREs).

1. **ESEfinder 3.0**: ESEfinder is a web-based resource that facilitates rapid analysis of exon sequences to identify putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55, and to predict whether exonic mutations disrupt such elements.

   URL: http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home

2. **RESCUE-ESE**: RESCUE-ESE was developed for identifying sequences with ESE activity. In this approach, specific hexanucleotide sequences are identified as candidate ESEs on the basis that they have both significantly higher frequency of occurrence in exons than in introns and also significantly higher frequency in exons with weak (non-consensus) splice sites than in exons with strong (consensus) splice sites.

   URL: http://genes.mit.edu/burgelab/rescue-ese/

3. **ExonScan**: ExonScan expects a primary transcript sequence, preferably excluding the first and last exon. It also currently needs at least about 20 bases upstream of the first exon it can locate, and at least about 60 bases downstream of the last. Splice sites are scored using a maximum entropy model using datasets from various other sources such as RESCUE-ESE.

   URL: http://genes.mit.edu/exonscan/

4. **Splicing Rainbow**: The tool identifies binding sites for splice factors in RNA sequences.

   URL: http://www.ebi.ac.uk/asd-srv/wb.cgi?method8

3.3.3 General splicing utility.

1. **Sroogle**: This server provides a graphic display of splicing related data on DNA segments. These include:
   - Splice site scores, based on different metrics

- Mapping of putative exonic and intronic splicing regulatory sequences (SRSs).
- Data regarding SRSs that would occur as a result of point mutations.
- Percentile scores, comparing your target exon to precompiled datasets of constitutive and alternative exons.

3.4 Materials and methods

3.4.1 Selecting word size, counting and normalizing words. Word size is typically guided by a power calculation (Fairbrother 2002). Similarly to our previous study (Lim 2011) we chose a word size of six (hexamer) for the web server analysis presented here. Counting of hexamers is done by simply moving one position at a time along the sequences in the exon database until hexamers are assigned to each position throughout the exon database. This counting procedure is done in three distinct regions: upstream intron (up to 200 nucleotides upstream of the 3'ss), exon (up to 100 nucleotides from each side of the splice sites), and downstream intron (up to 200 nucleotides downstream of the 5'ss). In the case where intronic or exonic sequence length is less than 400 or 200 nucleotides, respectively, the sequences are equally divided and each half is assigned to its nearest splice site. The positional frequencies for each hexamer will result in a 600-position vector. We performed standard *z*-score normalization for each vector separately for exonic and intronic space to decrease the influence of protein coding function.

3.4.2 Calculating intra-allelic L1 distance and assigning positional conservation scores. The mutation analysis fragments sequence into a set of overlapping singly shifted hexamers. A point mutation alters six hexamers to new hexamers with new L1 distances. In order to compare and contrast between normalized vectors, we employed L1 distance metric. The L1 distance is calculated as the sum of the absolute value of the difference in frequencies between two vectors. In addition to L1 distance, each genomic position in the exon database is assigned a normalized conservation score. These scores were obtained from alignments of 16 vertebrate genomes with human (phastCons17way) (Siepel 2004).

3.4.3 Sampling. We used a sampling with replacement strategy to convert both L1 distances and conservation scores into percentile rankings. Given a mutational target, the algorithm randomly selects a position within the same region (upstream intron, exon, downstream intron) and simulate a mutation by incorporating biases observed *in vivo* (Lim 2011). This sampling step repeats itself for 1000 trials, and the mutational target is compared with each sampled mutation using the L1 distance metric to derive a percentile ranking (i.e. how many times the sampled mutation has smaller L1 distance). The algorithm repeats the same sampling strategy for positional conservation scores. The sampling step results in two percentile rankings, each and the average of both are reported.

3.5 Results

3.5.1 Design and implementation. The computational engine and web interface were developed in Perl and with the use of the Bioperl toolkit (Stajicn 2002). The web server is designed to accept DNA sequences with mutational targets in FASTA format. For convenience, we have added an additional search function from which our users can load sequences by searching either the gene name, RefSeq accession ID, or the description of the gene of interests. This search engine pulls sequences directly from a locally maintained hg18 exon database. The web server will only process records with at least one point mutational target and the target must be flanked by at least 5 nucleotides on each side of the mutation (i.e. acgtc(a/c)gtagc). Users also have the option to reverse complement the input sequences.

3.5.2 Output. The results are displayed in table format. Each mutational target will be returned with several information: intra-allelic L1 distances, conservation scores, sampled percentile rankings derived from either the L1 distance or the positional conservation scores, and the average percentile ranking. Higher L1 distance indicates the mutation is likely to affect mRNA processing; and higher conservation score suggests the genomic location of the mutation is highly conserved. Average percentile ranking derived from our sampling strategy can be used to estimate how likely the mutation is to disrupt splicing. A detailed explanation of each returned information can also be found on the web server.

3.6 Figure and tables

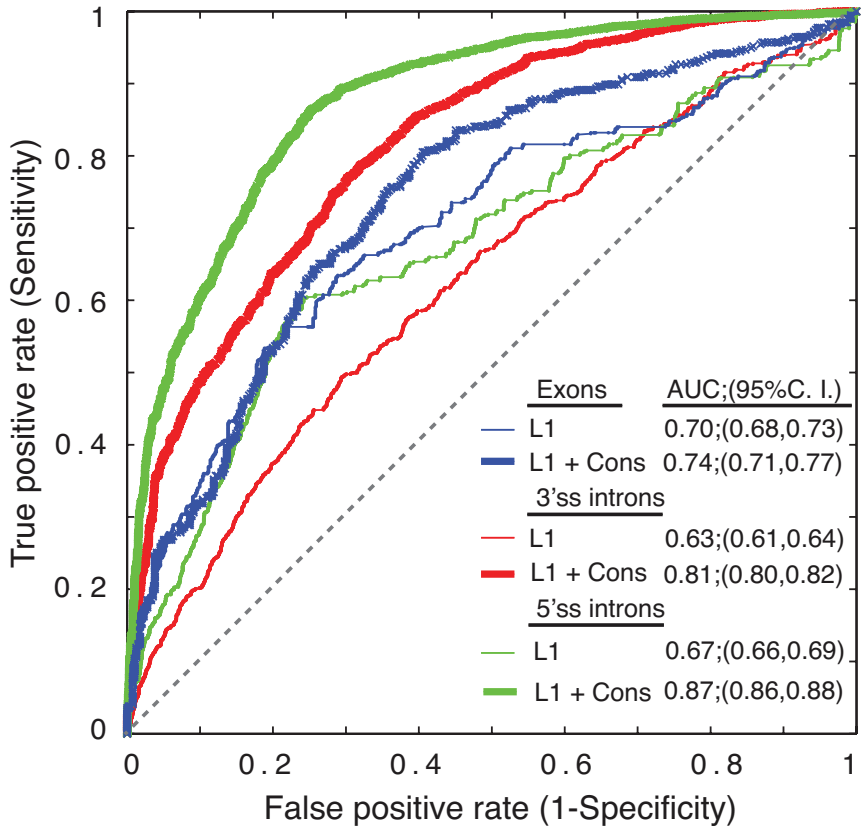


Figure 1. Incorporating conservation scores into classifier improves predictive power of L1 distance. ROC curve analysis using HGMD splicing mutants to compare L1 distance (L1) with L1 distance averaged with conservation (L1 + Cons) in regions around the annotated human splice sites. True positive samples are derived from 8,027 HGMD splicing mutants and false positive samples are constructed from simulated mutations. The exonic region is shown in blue; upstream and downstream introns are shown in red and green, respectively. AUC, Area Under Curve; C.I., Confidence Interval.

Chapter 4

A novel class of nuclear intron requires a pan-intronic stem loop for correct processing

4.1 Abstract

RNA secondary structure plays an integral role in the function of many catalytic, ribosomal, small nuclear and transfer RNAs. A global role for secondary structure within pre-mRNA that modulates splicing has been suspected and is currently an area of growing interest. By utilizing a combination of genome-wide computational screens, chemical mapping of RNA secondary structure and functional splicing assays; we present evidence for a class of nuclear intron that relies upon cis-structural elements for correct splicing. Phylogenetic analysis suggests that this ancient mechanism was present in the ancestral vertebrate lineage prior to the divergence of tetrapods. While largely lost from land dwelling vertebrates, this class of introns comprises as much as 1.8% of all introns in extant ray-finned fish. This novel class of intron is defined by simple repeat expansions of complimentary AC and GT dimers at opposite boundaries of an intron that, via a mechanism of intron bridging, enforce correct splice site pairing.

4.2 Introduction – regulation of splicing by RNA structures

Pre-mRNA splicing is driven by the split nature of eukaryotic genes, in which exons that will be used to produce the final mRNA transcripts are interrupted by non-coding introns. The removal of introns and the joining of exons are processed by the spliceosome, which catalyzes the joining of the 5' and 3 ends of an intron to create a new phosphodiester bond at the splice junction. The core of the spliceosome consists of five small nuclear ribonucleoproteins (snRNPs) that assemble onto the pre-mRNA through a series of complexes (Burge 1999). This assembly can be characterized as early steps involved in the correct identification and pairing of the splice sites followed by later steps that catalyze the reaction.

Even though much is understood about the biochemical pathway of RNA splicing, the mechanism of accurate splice site choice remains elusive. The majority of introns have highly conserved splicing signals at the 5' and 3' splice sites, a putative branchpoint near the 3' splice site, and a pyrimidine rich region between the branchpoint and the 3' splice site. Splice site selection can also be influenced by the interplay between positive and negative regulators that function through a combinatorial control of enhancers and silencers. These elements, however, seem to be insufficient to unambiguously assign splice sites pairing. Recently an increasing amount of evidence suggesting the importance of RNA secondary structure in modulating splice site selection has been documented (Hiller 2007; Raker 2009). Our focus is to understand a mechanism by how pre-mRNA secondary structure enforces pairing of exons. We previously introduced a novel computational framework for the identification of functional cis-elements in splicing (Lim 2011), and here we present preliminary data to support how a sequence-specific

pre-mRNA structure regulates splice sites pairing in a vast variety of nuclear introns in zebrafish.

The RNA portion of the snRNPs involved in the initial interaction with the pre-mRNA is known to basepair to sequences within the pre-mRNA. It is therefore reasonable to hypothesize that structures within pre-mRNA can inhibit splicing by preventing binding access to these regulatory sequences. However these internal structures might not always be extremely stable, and a competition would potentially exist between the internal structure and the basepairing of the snRNP to pre-mRNA. The equilibrium between intramolecular and intermolecular interactions can vary drastically, and therefore the subsequent splicing efficiency will be directly correlated to the stability of the internal structure. Structures that block snRNP binding appear to be a common mechanism through which RNA structures can inhibit splicing. Many examples of this type of inhibition have been documented.

4.2.1 Structures that inhibit splicing. In several cases, native RNA structures have been identified to sequester canonical splicing signals (i.e. 5' or 3' splice sites, branchpoint, and the polypyrimidine tract) in basepaired structures (Abbink 2008; Singh 2007; Loeb 2002). In other examples, artificially inserted pre-mRNA structures have been used to sequester canonical splicing signals to assess their roles in the splicing process. When the 5' splice site or branch point is structured, recruitment of U1 and U2 snRNP is inhibited (Blanchette 1997; Jiang 2000). In other cases, structures might also inhibit proteins from binding the pre-mRNA. The U2AF heterodimer consists two auxiliary factors UAF65 and U2AF35, which bind to the polypyrimidine tract and the 3' splice site, respectively, and help recruit the U2 snRNP to the branchpoint (Stenson 2003). Biochemical and structural data indicate that U2AF65 recognizes the polypyrimidine tract

as a single stranded site (Sickmier 2006), and not surprisingly, when the polypyrimidine tract is structured U2AF65 binding is inhibited, thereby splicing is blocked.

A recent bioinformatics study found that introns that are predicted to contain stable RNA structures are less likely to be spliced efficiently compared with introns of less predicted structures[12]. This finding suggests that RNA structure might be a common method used to regulate alternative splicing by repressing exon usage.

4.2.2 Structures that promote splicing. Surprisingly, there are just as many examples of RNA structure aiding splicing. Unlike structures that inhibit splicing, one general theme is that pre-mRNA structures can bring important splicing signals into closer proximity. For instance, 3'ss and branchpoint are, in general, in close proximity (< 20 nucleotides). However, in some introns they can be hundreds of nucleotides apart, yet the introns still splice efficiently (Chebli 1989; Deshler 1991). In some of these introns, structured regions between 3'ss and branchpoint have been found that make effective distance shorter, allowing distal branch points to be used. In other examples, structures have been found in introns that bring 5'ss and 3'ss into closer proximity, and are hypothesized to increase splicing due to the effects on distance between splice sites (Libri 1995; Goguel 1993; Charpentier 1996).

Some introns have been identified to contain structures with very long-range interactions, which are also proposed to bring the 5' and 3'ss closer together (Del Gatto 1997; Baraniak 2003; Preker 2006). In many examples, proposed stems have loops that are greater than 1kb in size, and disruption of the stems reduces splicing efficiency. Another interesting example of a long-range interaction is observed in *D. melanogaster Dscam* gene (Down syndrome cell adhesion molecule) (Graveley 2005). Dscam is a

complex gene with three regions of mutually exclusive splicing. Each region contains a tandem array of mutually exclusive splicing events. The 5' half of the hairpin is located upstream of the alternate exons and pairs with one of the Y alternate 3' spliced exon. Each of these 3' half hairpin is located upstream of the alternatively spliced exon. For instance, exon 6 has 48 possible exons, and only one is chosen in any given mRNA transcript. Long-range RNA structures are proposed to determine the selection of exon 6, where a selector sequence near the 5'ss of the intron basepairs with a docking sequence at the 5' end of the intron.

Due to substantially large introns (multiple kb in size) are commonly found in higher eukaryotes, model such as long-range interaction is an appealing model for bringing distant splice sites into closer proximity. A recent study provided evidence that this might be a common regulatory mechanism. Raker and colleagues identified 202 phylogenetically conserved potential structures in 12 *Drosophila* genomes, with most proposed stems having extensive loop spanning over 1kb (Raker 2009). The proposed stems were enriched within introns surrounding alternatively spliced exons, and modifications to these structures alter splicing patterns.

4.2.3 Dependency between two simple dinucleotide expansions in zebrafish. Although there have been reports of secondary structure sequestering particular elements, at the time of this writing we are not aware of obvious cases of secondary structure bridging nuclear introns in vertebrates. Secondary structure is an integral part of RNA biology and probably the principal determinants of catalytic recognition processes in splicing. It is surprising that a larger role for structure has not been described in the mRNA processing literature. Here we report a computational method to study evolution of non-coding elements that we have applied to vertebrates. Using a

method that exploits the positional distribution bias of splicing elements around annotated splice sites, we have created a discovery tool for rapidly evolving splicing elements across the animal kingdom. Applying this novel tool across the vertebrate clade reveals elements that are functioning as enhancers in one species but not in humans (or vice versa). The most rapidly evolving class of elements are AC and GT repeats that were discovered in fish. We demonstrate that these elements are enhancers that cooccur across introns and use chimeric splicing constructs and chemical mapping to demonstrate the dependency between AC and GT repeats within an intron.

4.3 Materials and Methods

4.3.1 Preparing a set of exon annotations and with orthologs in multiple species. Critical to the identification of splicing motifs and determination of conservation is the possession of a reliable set of exon annotations and with orthologs in multiple species. This type of annotation is often performed by some combination of *ab initio* gene prediction and transcript support (Hsu 2006). Gene prediction programs use splice site and sequence composition to determine potential exon-intron junctions. This approach selectively filters splicing substrates with poor splice sites or unusual composition from the database, thereby creating undesirable bias.

Relative to the normal length distribution of human-curated *knowngenes*, EST defined genes are enriched for short aligned blocks (Fig. 7A). As these short blocks are greatly reduced with increased transcript evidence, they are likely alignment artifacts. To test this, 235,493 human EST genomic alignment blocks bounded by an AG at the 5' end (the most conserved part of the 3' splice site) were scored for agreement to the 3' splice site motif using a first order Markov model. Two distinct populations were shown in the slate blue histogram (Fig. 7B): a high scoring peak (~90% of all counts) and a broad low scoring tail (~10% of all counts). Sampling random AG positions within these alignment regions and scoring them for splice site strength results in a similar broad distribution, suggesting that 10% of these EST genomic alignments describe exon models so weak as to be indistinguishable from background. To counter this, we required at least three ESTs to define an exon and also included alignment coordinates of known gene transcripts in the annotations.

These genomic coordinates were translated across different species using the UCSC *liftover* program, which parses the output of the multiple alignment tool, *blastz*, and returns orthologous coordinates. Preliminary analysis was performed on exons from chromosome 11 using a range of identity thresholds to the genome of eleven vertebrates (Fig. 7C). This analysis demonstrate that the probability of identifying a vertebrate orthologs is roughly proportional to its phylogenetic distance from human. The *liftover* program was used to generate the 22 sets discussed in this study.

4.3.2 Motif discovery in SELEX output. The binding specificity can be determined by SELEX experiments for several RNA binding proteins (Chapter 2; Table 3). The output of a SELEX experiment is typically 10–40 sequences that have been identified via iterative selection protocol as high affinity binding sites for the factor. From these sequences, motifs were derived using a Gibbs sampling strategy (Thompson 2005). To determine the length of these motifs for different RNA binding proteins, we examined the maximum a posteriori probability (MAP) of the alignment given by Gibbs sampling. The MAP value is measured relative to an empty null alignment, by taking the difference between the log of the probability of the alignment and the log of the probability of an empty alignment. We reason that motif is the most informative with length at which the MAP value is the highest (Chapter 2; Table 3).

4.3.3 Comparing positional distributions across species. We used the following algorithm to compare positional distribution across species. Given a set of exon coordinates in human,

1. Used *liftover* to parse and converted human exon coordinates into 22 other species and vice versa using a reciprocal best hit strategy, where the highest scoring match

to a human exon in a second species is deemed the orthologs if the best match in the reverse direction is the original human exon. This resulted in 22 pairs of exon database consisted of orthologous coordinates.

2. For each pair of orthologous coordinates (i.e. human vs. rat), we included up to 200 nucleotides of intronic sequence flanking either side of the splice site, and up to 100 nucleotides of exonic sequence flanking either side of the splice site.
    a. For exons less than 200 nucleotides, the exonic sequence is divided in half and assigned to the closest splice site.
    b. For introns less than 400 nucleotides, the intronic sequence is divided in half and assigned to the closest splice site.
3. Counted the number of occurrences of all 4,096 hexamers in each pair.
4. For each hexamer, we constructed a feature vector of length 600. We use the following indexing of positions in the vector.
    a. Positions starting at –200 and ending at 1 correspond to the upstream intronic region.
    b. Positions starting at 0 and ending at 199 correspond to the exonic region.
    c. Positions starting at 200 and ending at 399 correspond to the downstream intronic region.
5. Normalized the counts (i.e., *z*-score; Eq. 1) in each entry of each feature vector.

$$\forall x \in \vec{x},\ x' = \frac{x-\mu}{\sigma},$$ Eq.1

6. Computed L1 distance (Eq. 2) for each hexamer across species. For instance, a feature vector of hexamer acgtag in human is compared with the same hexamer acgtag in rat.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$ Eq. 2

7. Summarized the comparison of 4096 hexamers in histogram and screened for hexamers that produced with high L1 distances.
8. Repeated the procedures with another pair of orthologous coordinates.

4.3.4 Minigene reporter experiments. All splicing reporter experiments used the pZW4 splicing reporter minigene (Wang 2004). The chimeric constructs tested all possible pairings of 5' and 3' splice sites derived from two endogenous introns that contained splice sites of identical strength. The endogenous control intron contains no AC and GT repeats and the other endogenous intron contains both long stretches of AC and GT repeats at both side of the splice sites. The chimeras were constructed with the 5' end of the intron from the control and 3' side of the intron from the AC-GT intron, and vice versa. These variations were introduced into the reporter, transfected into 293 cells and assayed by RT-PCR.

4.4 Results

4.4.1 Splicing elements are positional dependent. To study the evolution of splicing elements across vertebrates, we employed a computational method to define conserved and rapidly evolving splicing elements and mechanisms. Unlike transcriptional enhancers, splicing elements are often strictly positional dependent. It was previously demonstrated that exonic binding of the normally intronic splicing factor, U2AF65, repressed splicing (Chapter 2; Lim 2011). This was also evident by deriving binding model of U2AF65 from SELEX data and plotting the significant matches to this binding model on a database of exons and introns (Fig. 1). The positional frequency distribution of U2AF65 motif revealed a significant enrichment at around the sites of polypyrimidine tract, and consistent with previous experimentations, the positional frequency of U2AF65 was heavily depleted in the exonic region of an exon. This observation suggested that the positional distribution pattern of an element around splice sites was indicative of the transacting factor's function in splicing. To determine if the relationship observed between U2AF65 binding and its function was general, we expanded our analysis to some members of the SR and hnRNP proteins (Chapter 2). This analysis largely supported the role of SR proteins as activators that interacted with splicing elements within the exon, whereas hnRNP binding sites were predominantly located within the intronic space. Since the distribution of a motif relative to splice sites appeared to be a signature of its function, we hypothesized that changes in this signature in a particular lineage would indicate rapid evolution of a *cis*-elements in that lineage.

4.4.2 Positional distribution of a motif around annotated splice sites is indicative of its function. We implemented this strategy by considering the distribution of all possible hexamers relative to the 5' and 3' splice sites. Each of the 4,096 hexamers

would have a distribution pattern around the annotated splice sites that we represented as a feature vector of word counts at positions relative to splice sites. Each hexamer (word) was counted in a database of orthologous genes with 200 intronic flanks and 100 exonic flanks on each side of the splice sites. Each feature vector was consisted of 600 positions summarizing the counts of occurrences for a word in positions relative to splice sites. These vectors were normalized by their *z*-scores and compared across species by their L1 distances (i.e. the absolute sum of the differences between two distribution patterns). The L1 distance effectively becomes a measure in the difference of distribution patterns of a hexamer around splice sites in two species. As splicing elements tend to accumulate at optimal locations relative to splice sites, an element that is functional in one species and non-functional in another would likely result in different distributions, and therefore be associated with high L1 distance. Non-functional sequence and elements that perform similar function in both species will likely have similar distribution, thus such comparison will result in low L1 distance. This distance metric is useful for discovering functional splicing elements or mechanisms that have either evolved or become lost following a particular speciation event.

We performed this distance comparison between human and 22 other genomes ranging from plants to other vertebrate species (Fig. 6). In each of the pairwise comparisons the analysis was restricted to exon regions that form orthologous pairs. Any differences observed capture lineage specific processing mechanisms that evolved within the set of identifiable ancestral genes, and not due to the quality of annotation in a given species. As expected, obligatory splicing signals such as splice sites and branchpoints appear to share similar positional distributions and therefore are associated with low L1 distances (Fig. 2A). However the most extreme L1 distances in all cross species comparisons

arose from simple dinucleotide expansions from AC and GT repeats in zebrafish (Fig. 2B).

4.4.3 Enrichments of AC and GT repeats are specific to splice sites. To understand the differences in positional distributions between AC and GT repeats in fish versus other vertebrates, we plotted the average hexamer frequencies as a function of positions relative to splice sites (Fig. 2C). The most obvious feature in the comparisons were the locations of the two prominent peaks: AC repeats are predominantly enriched downstream of the 5' splice site, whereas GT repeats were found immediately upstream of the 3' splice site. Adding Tetraodon and Fugu to the species analyzed confirmed that the two dinucleotide repeats were a feature of the teleosts lineage (i.e. bony fish). These repeat pairs were also detected in lampreys, suggesting an ancient chordate origin where the repeats were lost as a putative splicing mechanisms in tetrapods. Unlike many intronic splicing enhancers that were dramatically reduced in frequency across splice sites, the AC repeats were also slightly enriched on the exonic side of the 5' splice site as were the GT repeats on the exonic side of the 3' splice site. It is interesting to note that AC repeats and GT repeats were slightly enriched in the rodent lineage but on the opposite sides of the intron. As runs of AC and GT dinucleotides were complimentary, these elements could potentially form basepair with each other via mechanisms through exon looping or intron bridging. Both mechanisms would predict AC and GT repeats to cooccur across either exon or intron.

4.4.4 AC and GT repeats co-occur across introns but not exons. To test the co-occurrence we compared the observed positional frequency of the two repeats to a null model of uniform AC and GT distributions. As it was not clear what length of dinucleotide repeat was necessary for robust hairpin formation, we analyzed this dataset at several

repeat lengths ($2 < n < 30$, where n was the number of repeats). In each case we calculated the enrichment value, which represented the fold increase that a pair was observed above expectation. Plotting this data as a heatmap demonstrated a higher enrichment of cooccurrence across introns than exons (Fig. 3). The populations across exon were distributed across subpopulations of high AC/low GT and low AC/high GT strongly indicating an avoidance of long repeats of AC and GT across exons. In contrast there was a population of high AC/high GT observed across intron implying a selection for pairings of longer of tracts of these repeats across introns. As this analysis favored a model of secondary structure across introns, we design an *in vivo* minigene system to assess the dependency of the two repeats across an intron during pre-mRNA splicing.

4.4.5 AC and GT repeats effectively bring the distance of adjacent splice sites closer. We utilized a number of computational prediction algorithms to predict structures within the intronic space of a pre-mRNA. We designed a panel of chimeric introns that we systematically used Sfold to predict their secondary structures. Introns that contain long stretches of AC and GT repeats often exhibited strong secondary structures across introns. In the case of the endogenous intron that contained both long stretches of AC and GT repeats, Sfold predicted a structure that effectively brought the distance of its adjacent splice sites drastically closer (Fig 5A). In contrast, the chimera, an intron that only had GT repeats on one side of the intron, did not exhibit structures that would suggest to bring two splice sites in closer proximity (Fig 5B).

4.4.6 Dependency between AC/GT repeats enforces correct 3' splice site utilization *in vivo*. While detecting secondary structure forming across a transient RNA species is challenging, one of the features secondary structure is the dependency required on both strands of the duplex. In order to test this dependency in a functional

assay, we designed a panel of chimeric introns that we ligated into a splicing reporter construct. The chimeric introns tested all possible pairing of 3' and 5' splice sites derived from two endogenous introns that contain splice sites of similar strength. We found that the control intron (i.e. contains no AC and GT repeats) spliced correctly in the reporter (Fig 5; Lane 1), as did the chimera that combined the 5' AC repeats with the 3' end of the control intron (AC-control) (Fig 5; Lane 2). Chimeric constructs typically replicates splicing patterns of the endogenous gene and are widely utilized in the splicing field. The chimera that combines the 5' end of the control intron with the 3' GT repeats (control-AC) did not splice accurately (Fig 5; Lane 3). In addition to the correct splicing product, it was spliced to create two additional isoforms. cDNA cloning and sequencing revealed these isoforms to be spliced via cryptic 3' splice site in the downstream exon. This poor specificity of 3' splice site use was rescued by the reintroduction of AC repeats which reconstitutes the endogenous intron (AC-GT) (Fig. 5; Lane 4). These results demonstrated that while control and AC-GT splice sites were capable of splicing in all contexts, the 3' splice site of the CONT-GU intron required pairing with the AC repeats to enforce correct 3' splice site utilization.

4.5 Discussion

We present an example of a pair of simple dinucleotide repeats being utilized as paired elements of secondary structure across the chordate phylogeny. We discovered this signals in the introns of teleosts but not in tetrapods. The fact that we observed this signal in the Agnathans (lamprey) suggests an ancient origin of this mechanism and that repeat based secondary structure splicing mechanisms did not arise in the fish but were lost from the tetrapod and other lineages (Fig. 6).

The likelihood of coordinated evolution of elements across introns is lower than elements that function independently. However, dinucleotide repeats are known to be hypermutable. The expansions and contractions of simple repeats occur at a frequency several orders of magnitude higher than other indels or mutations. It is therefore more likely that a mechanism that employs repeat expansions across an intron would arise than a mechanism that employs RNA stems of complementary but heterologous sequences. It is still puzzling why a particular nucleotide dimer pair rose to such a high frequency within the family of nuclear introns. This may relate to the role other simple repeats play as target of RNA binding proteins (poly A tracts, polypyrimidine tracts and purine repeats are all RNA elements utilized as ligands in gene expression). This restriction does not explain the preference for AC and GT at particular ends of the intron. However a system of secondary structure that utilizes a single ordered element pair describes a modular system of generic across-intron structure. Modularity is a defining feature of many regulatory *cis*-elements sequences and this feature has evolutionary implications (i.e. most elements originated through mechanisms of a duplication and recombination from preexisting elements). We speculate that, if a particular dimer pair like AC and GT came to occupy a significantly large fraction of a genomes' introns, it

could further expand through intragenic recombination or exon shuffling mechanisms with other AC and GT introns. Conversely, if the hairpin mechanisms involved complements of heterogeneous sequence, It is unlikely that a new intron could be made through the combination of two different introns as their 5' and 3' stems would be unlikely to pair with each other.

The existence of group II intron which rely heavily on secondary structure for precise splicing offers some parallels to the AC and GT introns. It appears that these introns were lost in tetrapod speciation although elevated levels of both AC and GT repeats can be found in the rodents. It is possible that GT repeats which appear enriched in 3' splice site have been taken over by RNA binding proteins. hnRNP K has been implicated in binding these sequences for example. In addition, hnRNP L has been documented to interact with AC repeats. These elements do not always seem to function as enhancers. In fact, the cystic fibrosis receptor gene has a polymorphic run of GT repeats upstream of the 3' splice site that appears to reduce its recognition

4.6 Future directions

4.6.1 Chemical mapping of AC- and GT-containing intron reveals a juxtaposition of 3' and 5' splice sites.

We employed computational prediction algorithms to predict structures within the intronic space of a pre-mRNA, and showed that intron with long stretches of AC and GT repeats exhibited more stable structures relative to control introns with no AC and GT repeats. Our computational analysis also suggested a strong dependency between the co-occurrence of AC and GT repeats at both ends of the intron, an evidence observed by structures predicted computationally when one side of hairpin was removed. In addition, distance between splice sites were dramatically reduced for introns with AC and GT repeats.

As RNA undergoes conformational changes in response to folding and ribonucleoprotein assembly reactions, changes in the solution conditions, and binding by small molecular and protein ligands, we employed SHAPE (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension) to chemically map the secondary structure of an RNA sequence in vitro (Wilkinson 2006). Computational prediction algorithms have been demonstrated to accurately identify RNA structures, however most RNAs only function correctly after they fold back upon themselves to form 3-dimensional structures, which are extremely challenging to predict computationally. Experimentally determining the local flexibility of RNA as a function of nucleotide position therefore provides important information that enables both the sequence and the structure of RNA to be related to its biological function.

Local nucleotide structure can be detected by treating RNA with chemical and enzymatic reagents (Ehresmann 1987). SHAPE chemistry is based on the discovery that the nucleophilicity of the ribose 2'-position is exquisitely sensitive to the electronic influence of the adjacent 3'phosphodiester bond (Chemberlin 2002). Unconstrained nucleotides sample more conformations that enhance the nucleophilicity of the 2'hydroxy group than do basepaired or otherwise constraint nucleotides. Therefore hydroxyl-selective electrophiles such as NMIA form stable 2'-O-adducts (Raker 2009) with flexible RNA nucleotides. The NMIA reagent is also consumed by competing hydrolysis reaction, which advantageously causes the reaction to be self-limiting (Merino 2005). The sites of 2'-O-adduct formation are identified as stops to primer extension by reverse transcriptase. The extended products are separated by standard high-resolution gel electrophoresis. Absolute NMIA reactivity at each nucleotide is then determined by comparing band intensities from the modification reaction to those of no NMIA control. One or more dideoxy sequencing lanes are used to assigned lanes within the NMIA reaction and control lanes.

Folded RNA is generated by in vitro transcription. In order to eliminate multimeric forms, RNA is first heated and snap cooled in a low ionic strength buffer. A folding solution is added to allow the RNA to achieve an appropriate conformation and to prepare it for NMIA modification. The RNA is folded in a single reaction and is later separated into a (+) and (-) NMIA reactions. After NMIA reaction, the modified RNA is precipitated with ethanol in order to purify the RNA from reaction products and buffer components that may be detrimental to the primer extension reaction. Sequence reactions are performed concurrently with the primer extension to localize the structurally open regions in the RNA transcript.

4.6.2 Demonstrating the regulation of mutually exclusive splicing by RNA structures. An interesting aspect of AC and GT repeats is their distribution within transcripts. The two repeats do not exist in the one-to-one stoichiometry as predicted by secondary structure. Our preliminary data shows that GT repeats are approximately 2-fold more abundant than AC repeats, and the majority of GT repeats are clustered within the same transcript. Could the architecture of this class of introns be similar to hose observed in the Dscam gene (Graveley 2005), where mutually exclusive splicing is regulated by a driver element (AC repeats in this case) located upstream of an array of target elements (GT repeats)? In order to experimentally test this observation, total RNA can be extracted from zebrafish and RT-PCR is performed using primers specific to each transcripts that exhibits the above-described distributions of AC and GT repeats. The PCR products of tubulin can be used as a control, and all PCR products are separated by gel electrophoresis. Further analysis including sequencing of the PCR products can be performed to accurately demonstrate the different splicing isoforms of each transcript.

4.6.3 Determining the roles of trans-acting factors in AC and GT repeats. Due to sequence specific of the dinucleotide repeats, we seek to demonstrate if trans-acting factors play roles in binding to the AC and GT repeats in order to aid in precise recognition and pairing of splice sites. To identify polypeptides that interact with the repeats, we can prepare pre-mRNA containing a single radioactive phosphate (^{32}P) within the AC and GT repeats. We may use a set of AC and GT containing introns that have been shown to exhibit secondary structure both computationally and experimentally. The splicing assay may be done in ZF4 cell nuclear extract with typical kinetics. At various times aliquots of the splicing reaction mixture are collected, then are exposed with UV light and treated with RNase A (for single-stranded RNA) and RNase

H1 (for double-stranded RNA). ^{32}P polypeptides can be detected by SDS-PAGE and autoradiography.

4.6.4 Predicting the rates of conservations among AC and GT repeats. Another future direction of this work involves the rate of conservation on genes that contain AC and GT dinucleotide repeats. This can be done in several ways.

a. *Detecting molecular adaptation by Ka/Ks ratio.* Measuring the Ka/Ks ratio (number of non-synonymous substitutions over number of synonymous substitutions) provides a direct way to measure the selective pressure acting on codons. Evaluating whether Ka/Ks ratio is significantly greater than 1 constitutes a test for the action of positive selection. The ratio is often measured by the so-called counting methods, for which all proposed measures (Miyata 1980; Nei 1986) rely on similar procedures: (1) count the number of synonymous and non-synonymous changes by taking into account of all possible evolutionary pathways between aligned nucleotides or codons; (2) count the number of synonymous and non-synonymous sites; (3) correct for multiple substitutions at the same site by using an explicit evolutionary model (i.e. Jukes Cantor for protein sequences and Jerzy Neyman for DNA/RNA sequences). Averaging such counting method drastically reduces the power of these approaches because the selection signal is diluted (Gillespire 1991). In order to accurately predict selection on functional site, however, we must also take codon models within a probabilistic framework when estimating different ratios at different sites. This is often achieved by modeling at the codon level, instead of at the nucleotide or amino acid levels (Goldman 2004), as this approach provides a more realistic

model by incorporating the differences in evolutionary rates at different codon positions.

b. *Determining positive selection by analyzing polymorphism data*. In addition to looking at changes between species and comparing synonymous vs. non-synonymous substitutions, signals of selections can be sought by comparing allelic frequencies within species. Our goal here is to quantify the relative contribution of positive Darwinian selection and random genetic drift in determining levels of variability within exons among the same species. Specifically, we are interested in if pre-mRNA splicing can be completely driven by intronic features due to exon variability. We first perform tests by rejecting neutrality and allow detection of advantageous mutations that are in the process of becoming, or have recently become, fixed in the population. Positive selection can generate genetic differentiation between populations at a loci under selection in a heterogeneous environment. Therefore extreme differentiation between population at a given locus relative to other loci may be evidence for positive selection.

4.6.5 Demonstrating how RNA splicing can be driven solely by features found within introns. An exon adjacent to the computationally predicted structures found in the set of zebrafish introns can be tested for their intronic roles in splicing by transient transfection. The function in promoting splice sites pairing by forming stem loop structure across intronic space should be diminished if such structure is disrupted. Here we are interested in demonstrated if RNA splicing can be solely driven by features found within introns (i.e. secondary structure). While keeping the intronic structure that we hypothesize to bring two splice sites in closer proximity, we may design reporter constructs by intentionally inserting exonic splicing silencers (ESS) into both adjacent

exons and measure if the presence of ESSs disrupt efficient splicing of the two exons. ESSs have been implicated to bind hnRNP proteins in the exons, which effectively silence the usage of those exons (Krecic 1999). Wildtype and its modified constructs are transfected into a zebrafish fibroblast cell line to screen for splicing patterns. Minigene mRNA is collected, analyzed by RT-PCR, and separated on an agarose gel to screen for the differences in splicing isoforms.

4.7 Figures and Tables

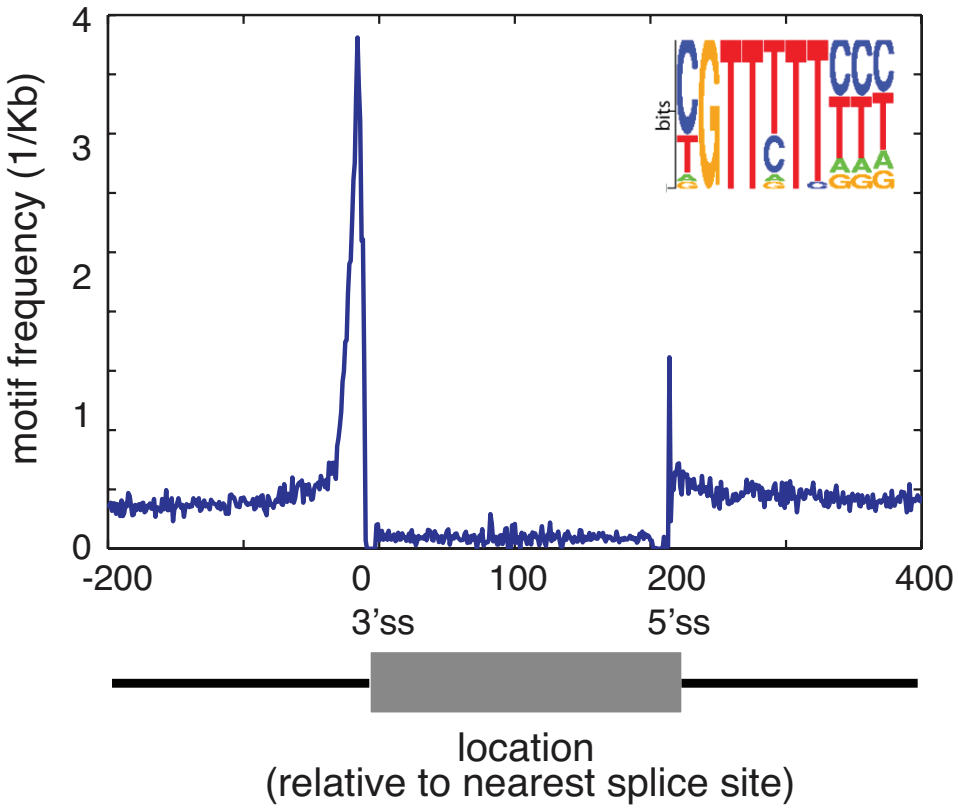


Figure 1. Positional frequency distribution of U2AF65 revealed significant enrichment at the sites of polypyrimidine tract and depleted in exonic region. U2AF SELEX motif was mapped to a dataset of 312,275 human splice sites and plotted on an amalgamated exon.

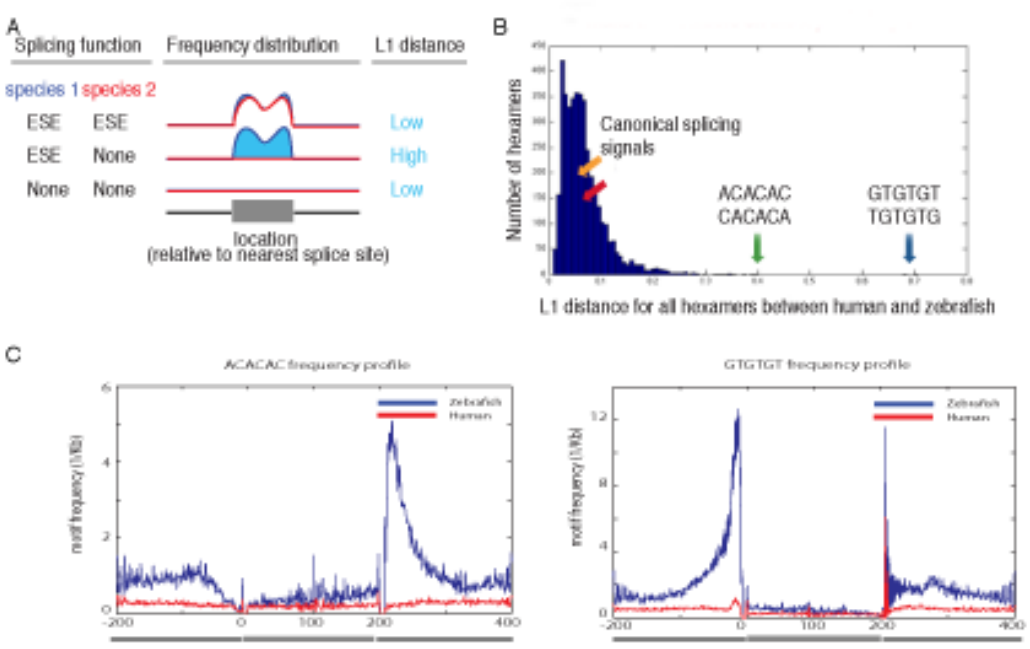


Figure 2. Cross species analysis revealed significant enrichments of two simple dinucleotide repeats in zebrafish that cooccur on specific side of the splice sites. The positional frequency distributions of all 4096 hexamers were plotted around a database of orthologous exons between human and a second species. (A) Several comparisons of the same hexamer between two hypothetical species were drawn to illustrate three possible scenarios. L1 distance metric (shared blue area) was used to compare normalized frequency distributions. Low L1 distance indicated there were small differences and the hexamer had similar or no difference in splicing function across the two species. High L1 distance denoted the two positional distributions were vastly different among species and likely differed in their role in splicing. (B) L1 distance was used to compare the human genome against other 22 species. In the analysis between human and zebrafish, we discovered two outliers: AC and GT repeats, whereas the canonical splicing signals (branchpoint, polypyrimidine tract, 3' and 5' splice sites) resulted in low L1 distances. (C) Positional frequency distributions of AC and GT repeats

revealed that these elements were rare in human, and when enriched in zebrafish, the repeats were specified to individual splice site. AC repeats were found to concentrate at around the 3' splice site whereas GT repeats were predominantly found at around the 5' splice site.

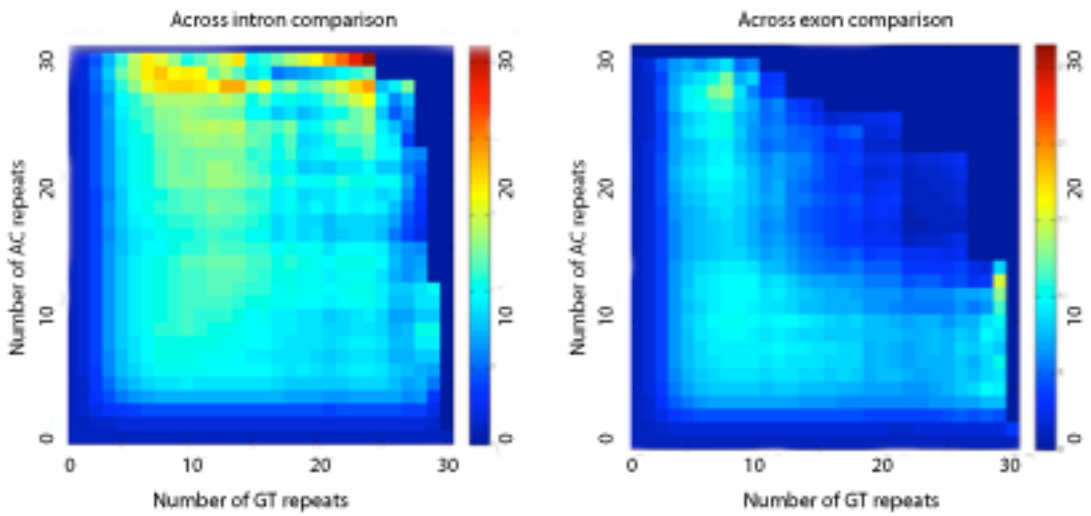


Figure 3. AC and GT repeats cooccur across introns but not exons. Observations of AC and GT pairing in zebrafish across either exons or introns were compared to values expected from a null model of uniform AC and GT distributions. Each square in the heatmap represented the fold enrichment above expectation assuming the occurrences of AC and GT were independent from each other.

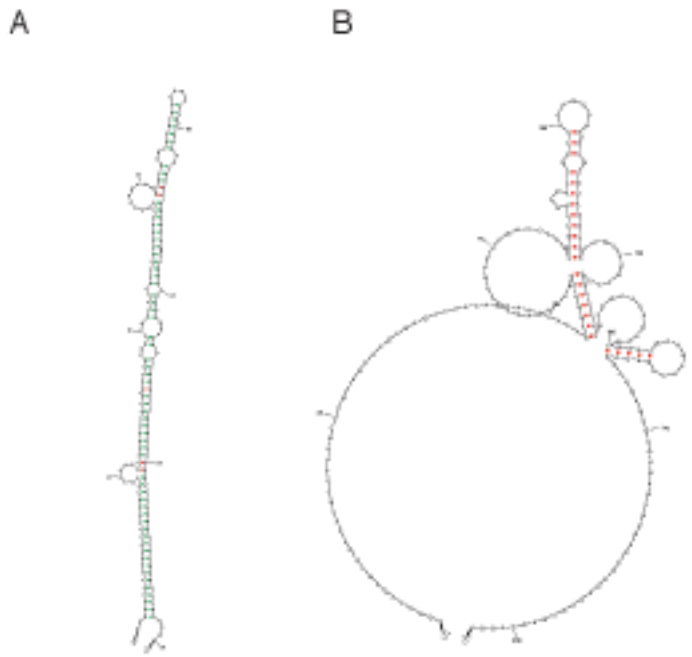


Figure 4. Dependency between AC and GT repeats minimizes the distance between splice sites. Sfold predicted centroids for (A) intron that contains both AC and GT repeats and (B) intron that contains only GT repeats. The structures were selected from the ensemble centroids during the Sfold sampling process. These structures were also in agreement with the lowest energy structures.

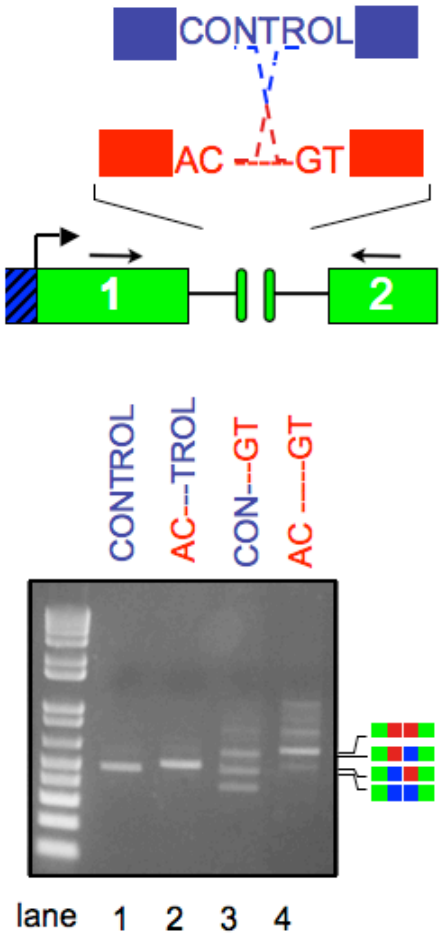


Figure 5. In vivo experimentation suggest strong dependency between AC and GT repeats. We designed a chimeric construct that we ligated into a splicing reporter construct. The chimeric introns tested all possible combinations of splice sites from two endogenous 200nt long introns (CONTROL and AC-GT). AC-TROL includes the 100nt 5' AC repeats with the 100nt 3' end of the control intron. CON-GT combines the 5' end of the control intron with the 3' GT repeats. Total RNA from transfection into 293 cells was analyzed by RT-PCR. Arrows indicate the correct splicing isoform.

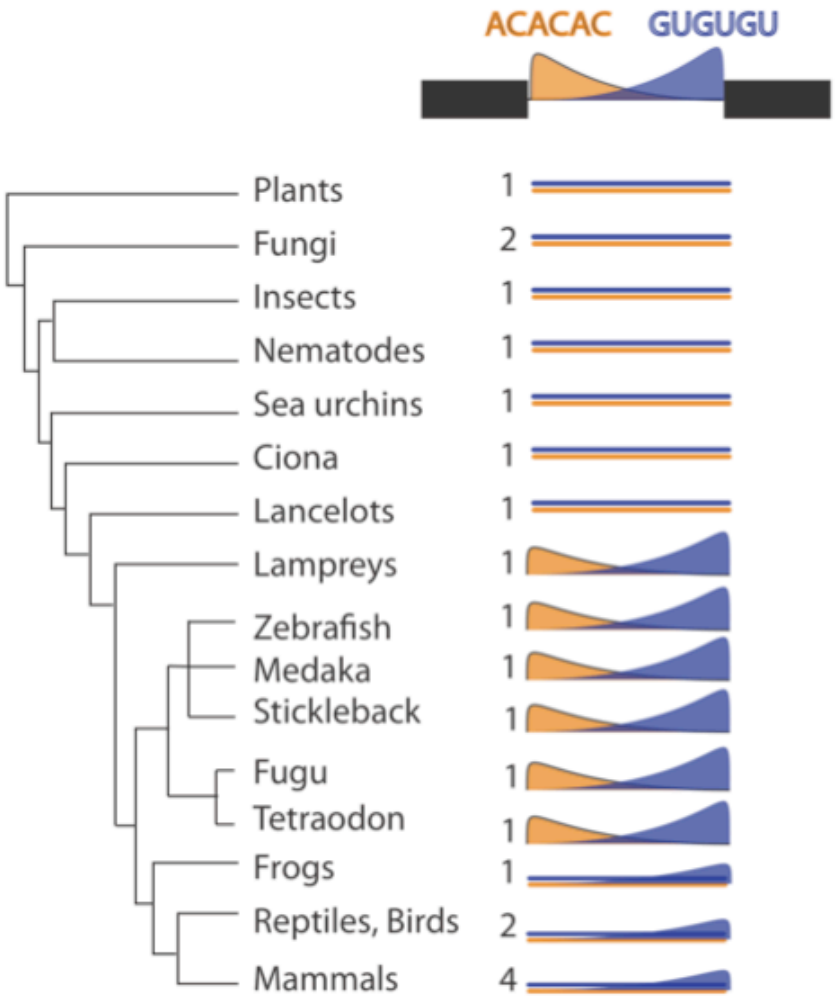


Figure 6. AC and GT repeats were found in bony fish and lampreys, suggesting an ancient origin. The genome wide conservation of AC and GT repeats across eukaryotic nuclear introns. A total of 22 genomes were analyzed for AC and GT repeats occurring across introns. For each clade in the unscaled dendogram, the genomes from the indicated numbers of species were downloaded and analyzed by the word counting procedure described above. The results were summarized as cartoon histograms of the total repeat counts as a function of position from the 3' and 5' splice sites.

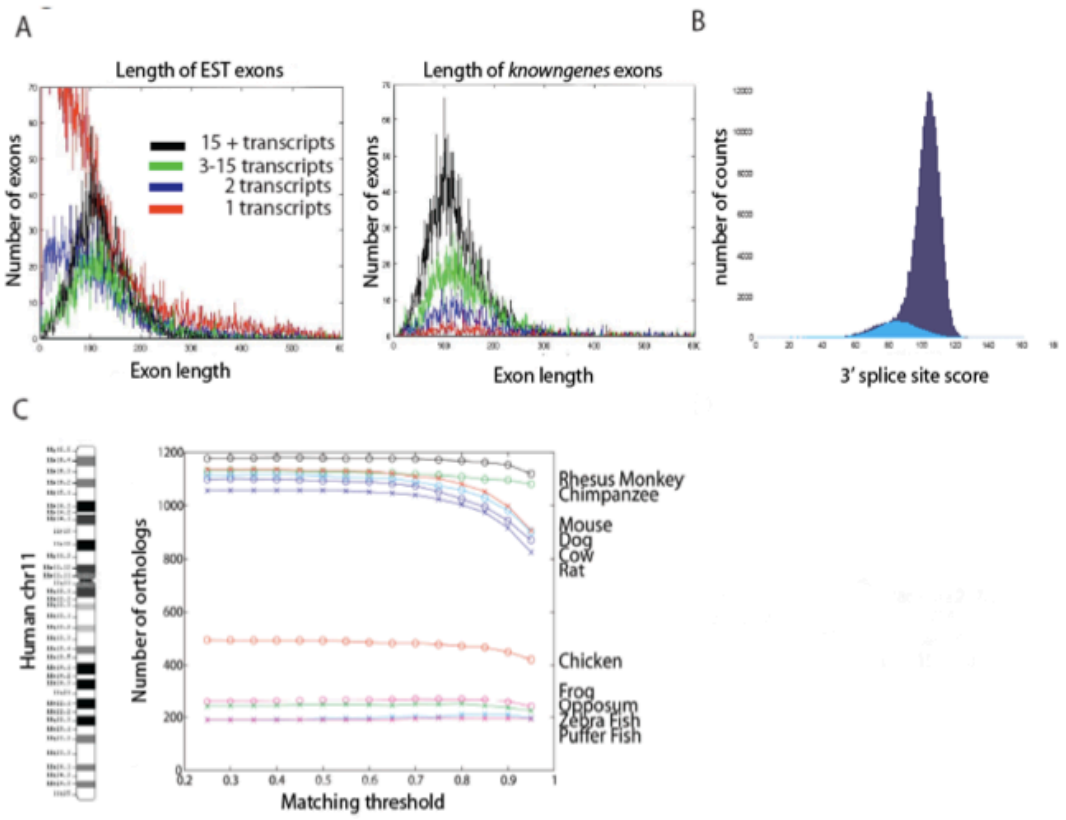


Figure 7. Multiple EST records were required to represent an exon. (A) Compared human-curated *knowngenes* exons, EST defined genes were enriched for short aligned blocks. (B) EST genomic alignments were scored for agreement to the 3' splice site motif and revealed roughly 10% of EST alignments (shaded light blue in the histogram) \were of background level. (C) Using *liftover* to convert genomic coordinates from one species to another was roughly proportional to its phylogenetic distance from human.

Chapter 5

Discussions and future directions

In this thesis work, I introduced a novel computational method for which it quantified the positional frequency distributions among splicing elements and calculated the mathematical distance from the norm. This proposed method successfully captured and classified 4,096 hexamers into 51 distinct groups, which contained classes of canonical splicing motifs, splicing enhancers and silencers, transcriptional elements that play roles in splicing, and other elements that weren't characterized prior to this study. Many of these elements were experimentally tested to demonstrate their influence on RNA splicing. We found that point mutations that altered their positional frequency distributions were more likely to change their splicing patterns. The computational method, in conjunction with experimental assays, was the foundation of this thesis.

In Chapter 2, I applied this method to assess the importance of sequence variation that affected mRNA processing defects in human genes (Lim 2011). Specifically, this method was used to analyze disease alleles from the Human Gene Mutation Database. The work found more than one-third of hereditary disease-causing mutations may be caused, at least in part, by errors made in pre-mRNA splicing. This discovery potentially unlocked a brighter outlook on patients with genetic diseases using a new therapeutic application called antisense oligonucleotides (ASO). This technology can serve as an easier and less risky way to interfere with errant splicing operations, as compared to gene replacement therapy, a technology that attempts to replace a whole gene caused by malfunctioned three-dimensional protein products.

An ASO is a single-stranded deoxyribonucleotide that is typically 20 nucleotides (nt) long and is complementary to the target mRNA of interest. Hybridization of ASO to the target mRNA can be achieved by Watson-Crick base pairing, which results in specific inhibition

of gene expression by multiple mechanisms (Crooke 2004). This technology is not only a useful tool for studies of loss-of-gene function and target validation, but also highly valuable as a novel therapeutic strategy to treat diseases that are linked to dysregulated gene expression. Depending on the chemical makeup of ASO and the location at which the ASO interacts with the target mRNA, the hybridization often reduces levels of translation of the target transcript by triggering a non-specific endonuclease, *RNase H*, which catalyzes the cleavage of mRNA via a hydrolytic mechanism. This leads to the degradation of the target mRNA while leaving the ASO intact (Wu 2004). Other ASO-driven mechanisms include translational arrest by steric hindrance of ribosomal activity (Kurreck 2002), destabilization of pre-mRNA in the nucleus (Kurreck 2003), and perhaps more relevant to this thesis, interference with mRNA maturation by inhibiting binding access of splicing elements and changes of local secondary structures within the mRNA during RNA processing (Chan 2006).

One example of induced alternative splicing by ASO technology is the *Apolipoprotein B* (*APOB*) mRNA (Khoo 2007). *APOB* is the primary apolipoprotein of low-density lipoproteins (*LDL*), which is responsible for carrying cholesterol to tissues, and research has shown that excess level of *APOB*-containing particles may lead to atherogenesis. The *APOB* pre-mRNA comprises of 29 exons, all constitutively spliced into the final mRNA product. Khoo and colleagues have investigated the ability of ASO to induce the skipping of exon 27 in endogenous *APOB* mRNA and demonstrated the ability to induce a specific truncated version of *APOB* was a potential therapeutic utility in lowering *LDL* and cholesterol levels *in vivo* (Khoo 2007). It is also important to mention that the skip 27 APOB mRNA was able to bypass the RNA quality check in the nucleus and transported into the cytoplasm for ribosomal processing.

The strength and stability of interactions between ASO and target mRNA highly depend on the interplay of multiple cellular parameters, such as thermodynamic stability (binding energy prediction), secondary structure of the target mRNA, and proximity of the hybridization site to functional motifs on the designated transcripts (5' cap region and translational start sites). There are screening approaches to obtain potent ASO for therapeutic applications, such as mRNA walking (Sohail 2000), oligonucoetide array (Cho 2001), and RNase H mapping (Ho 1998), but as these techniques are generally expensive and labor intensive, computational algorithms have been developed to assist the design of ASO. These computational algorithms – most of which are freely available on the web – are certainly the most economical way to ASO design and often generate potent ASO design from a handful of candidates.

Despite the advancement in algorithmic development, currently there is no standalone program in predicting potent ASO (Chan 2007). One, however, can utilize multiple computational programs to improve the rate of hybridization. As ASO needs to interact with a single-stranded RNA molecule, it is generally accepted that the most effective ASO design strongly depends on the accurate prediction of RNA secondary structures (Vickers 2000; Andronescu 2005). Although predicting structure has been a strong interest for RNA biologists for decades, a truly reliable algorithm to predict RNA secondary structure has remained to be a challenging task. In addition, the strategies employed within the core prediction algorithms differed from programs to programs. For instance, the widely used *mfold* program predicts overall minimum free energy of different possible structures (Zuker 2003), whereas another well-known program, *sfold*, statistically samples an ensemble of structures and predicts only the best secondary structure of the target transcript (Ding 2004). Using a combination of different prediction

algorithms often serves as the best strategy to determine the most frequently occurring secondary structure of the target mRNA. A list of prediction algorithms was outlined in Chapter 4.

After confirmation of RNA secondary structure within the target mRNA, one must settle on some well-defined activity-enhancing motifs and discard those activity-decreasing motifs in the ASO. Prior study demonstrated a strong positive correlation between ASO-mediated knockdown and the presence of the following motifs: CCAC, TCCC, ACTC, GCCA, and CTCT motifs in the ASO (Matveeva 2000). On the other hand, motifs such as GGGG, ACTG, AAA, and TAA motifs in ASO weakened ASO activity. Although it is commonly believed that the formation of ASO-mRNA heteroduplex stimulates *RNase H* activity, Ho and colleagues found that *RNase H* activity is sequence independent (Ho 1996). Interestingly, GC content is strongly correlated with the thermodynamic stability of the ASO-mRNA duplexes and RNase H activity. Study has found that strong ASO effects with a minimum of 11 G or C residues in a 20nt ASO, whereas poor inhibition was observed by ASO having 9 or fewer G or C residues.

Another consideration of ASO design comes from the accurate prediction of binding energy between ASO and target mRNA. The program suite *RNAStructure* offers a program called *OligoWalk*, which can be used to calculate binding energy of ASO/ASO and ASO/mRNA interactions. By utilizing two large databases from ISIS Pharmaceuticals and published literatures, Matveeve and colleagues showed that the hit rate of developing a potent ASO was roughly 6-fold higher if the binding energy between ASO and mRNA is greater than -8 kcal/mol (Matveeva 2003). Simply following the above criteria, another research group were able to design ASO for the downregulation of

vascular endothelial growth factor protein expression with a success rate higher than 85% (Fei 2005).

ASO design continues to be a challenging task, and the delivery of ASO presents yet another layer of difficulty to this new technology. An unmodified ASO will normally be modified by various nucleases in biological environments, thus its overall property will be changed and be prevented from penetrating through cell membrane. In addition, unmodified ASO has a net negative charge and can barely penetrate the plasma membrane. The amount of well-designed ASO that successfully penetrates through cell membrane is low enough to prompt the development of a variety of delivery strategies. For *in vivo* cell cultures studies, mechanical techniques such as electroporation and microinjection were used to increase the internalization of ASO. Cellular uptake of ASO is primarily driven by an adsorptive endocytosis process (Lysik 2003), and upon entry into the intracellular milieu in the form of an endosome, the ASO needs to escape from the endosomal vesicles to avoid lysosomal degradation. In several studies Dioleoylphosphatidylethanolamine (DOPE) was added into the lipid carrier in order to promote endosomal membrane destabilization, thus leading to release of ASO into the cytosol (Lysik 2003; Fattal 2004). Other methods such as covalent conjunction of ASO to a macromolecular-like dendrimer and to cell-penetrating peptides have been demonstrated as an efficiency strategy to promote cellular uptake of ASO (Yoo 2000; Jarver 2004). Dendrimers are spherical and highly branched polymers; their ability to form covalent complexes with the ASO enhances delivery into both the cytosol and the nucleus, and increases retention time of ASO once it is in the cell.

The computational method I developed assists scientists and clinicians to predict target mRNA that may be responsible for their diseases of interests. In addition, I transformed the method into a functional web server to facilitate the accessibility of my proposed method. The functionality and usability of the software were discussed in Chapter 3. In short, users are required to provide sequences in FASTA format with mutational data. The software gathers the sequence and mutational information and predicts, through a sampling without replacement strategy, which mutation is likely to affect RNA processing. In addition, the software provides a search function for which users can search thousands of sequences by gene name, accession ID, and gene descriptions.

To further explore the possibilities of my proposed method, I extended the scope of my research by performing 22 cross-species analyses. The specifics of this novel work can be found in Chapter 4. Specifically, I discovered two sequence-specific dinucleotides expansion in fish that potentially enforced the correct pairing of splice sites through RNA secondary structures. More importantly, extensive phylogenetic analysis strongly suggested an ancient mechanism was present in the ancestral lineage prior to the divergence of tetrapods. This novel mechanism comprised as much as 1.8% of all introns in extant ray-finned fish, and had the ability to rescue splicing defects when pairing of the two dinucleotides expansion were placed in optimal locations relative to individual splice sites. This serves as another potential yet powerful application of my proposed method.

In summary, I first proposed a novel computational method that utilized the positional dependencies between splicing elements. Through extensive amounts of experimentation, I showed this method was extremely useful in identifying existing and novel splicing elements in the human genome. Furthermore, I illustrated how this method

could be used to capture point mutations that were likely to alter splicing phenotypes. For instance, analyzing the Human Genome Mutation Database revealed nearly one-third of all disease-causing mutations could be linked to splicing errors. I then implemented a computational web server to facilitate an easier access of this intuitive and powerful method. This software is freely available on the World Wide Web for all non-commercial use. Lastly, I applied this method to a large cross-species analysis and revealed a novel splicing mechanism by which two sequence-specific dinucleotides expansions that together form secondary structure to enforce correct splice sites selection.

References

Abbink, T.E.M., and Berkhout, B. (2008). RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. J Virol *82*, 3090-3098.

Aitchison, J.D., Blobel, G., and Rout, M.P. (1996). Kap104p: a karyopherin involved in the nuclear transport of messenger RNA binding proteins. Science *274*, 624-627.

Allawi, H.T., Dong, F., Ip, H.S., Neri, B.P., and Lyamichev, V.I. (2001). Mapping of RNA accessible sites by extension of random oligonucleotide libraries with reverse transcriptase. RNA *7*, 314-327.

Andronescu, M., Zhang, Z.C., and Condon, A. (2005). Secondary structure prediction of interacting RNA molecules. J Mol Biol *345*, 987-1001.

Arima, H., Takahashi, M., Aramaki, Y., Sakamoto, T., Yuda, K., Akiyama, K., Goto, T., and Tsuchiya, S. (1999). Design of potent phosphorothioate antisense oligonucleotides directed to human interleukin 10 gene product and their evaluation of antisense activity in U937 cells. Pharm Res *16*, 1163-1171.

Babitzke, P., Yealy, J., and Campanelli, D. (1996). Interaction of the trp RNA-Binding attenuation protein (TRAP) of Bacillus subtilis with RNA: effects of the number of GAG repeats, the nucleotides separating adjacent repeats, and RNA secondary structure. J Bacteriol *178*, 5159-5163.

Baraniak, A.P., Lasda, E.L., Wagner, E.J., and Garcia-Blanco, M.A. (2003). A stem structure in fibroblast growth factor receptor 2 transcripts mediates cell-type-specific splicing by approximating intronic control elements. Mol Cell Biol *23*, 9327-9337.

Barnard, D.C., and Patton, J.G. (2000). Identification and characterization of a novel serine-arginine-rich splicing regulatory protein. Mol Cell Biol *20*, 3049-3057.

Bartel, D.P., and Unrau, P.J. (1999). Constructing an RNA world. Trends Cell Biol *9*, M9-M13.

Bentley, D.L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol *17*, 251-256.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem

72, 291-336.

Blanchette, M., and Chabot, B. (1997). A highly stable duplex structure sequesters the 5' splice site region of hnRNP A1 alternative exon 7B. RNA *3*, 405-419.

Blanchette, M., and Chabot, B. (1999). Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. EMBO J *18*, 1939-1952.

Blaustein, M., Pelisch, F., Coso, O.A., Bissell, M.J., Kornblihtt, A.R., and Srebrow, A. (2004). Mammary epithelial-mesenchymal interaction regulates fibronectin alternative splicing via phosphatidylinositol 3-kinase. J Biol Chem *279*, 21029-21037.

Blencowe, B.J., Bauren, G., Eldridge, A.G., Issner, R., Nickerson, J.A., Rosonina, E., and Sharp, P.A. (2000). The SRm160/300 splicing coactivator subunits. RNA *6*, 111-120.

Brudno, M., Gelfand, M.S., Spengler, S., Zorn, M., Dubchak, I., and Conboy, J.G. (2001). Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res *29*, 2338-2348.

Buckanovich, R.J., and Darnell, R.B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Mol Cell Biol *17*, 3194-3201.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol *24*, 10505-10514.

Buratti, E., Muro, A.F., Giombi, M., Gherbassi, D., Iaconcig, A., and Baralle, F.E. (2004). RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. Mol Cell Biol *24*, 1387-1400.

Burd, C.G., and Dreyfuss, G. (1994). RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. EMBO J *13*, 1197-1204.

Burge, C.B. (1999). The modern RNA world. Nat Struct Biol *6*, 595-596.

Burge, P.S., Pantin, C.F., Newton, D.T., Gannon, P.F., Bright, P., Belcher, J., McCoach, J., Baldwin, D.R., and Burge, C.B. (1999). Development of an expert system for the interpretation of serial peak expiratory flow measurements in the diagnosis of occupational asthma. Midlands Thoracic Society Research Group. Occup Environ Med *56*, 758-764.

Caceres, J.F., and Kornblihtt, A.R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet *18*, 186-193.

Caceres, J.F., Stamm, S., Helfman, D.M., and Krainer, A.R. (1994). Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. Science *265*, 1706-1709.

Calapez, A., Pereira, H.M., Calado, A., Braga, J., Rino, J., Carvalho, C., Tavanez, J.P., Wahle, E., Rosa, A.C., and Carmo-Fonseca, M. (2002). The intranuclear mobility of messenger RNA binding proteins is ATP dependent and temperature sensitive. J Cell Biol *159*, 795-805.

Caputi, M., and Zahler, A.M. (2001). Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. J Biol Chem *276*, 43850-43859.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet *3*, 285-298.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res *31*, 3568-3571.

Carter, P.J., Breiner, K.M., and Thorp, H.H. (1998). Effects of secondary structure on DNA and RNA cleavage by diplatinum(II). Biochemistry *37*, 13736-13743.

Cavaloc, Y., Bourgeois, C.F., Kister, L., and Stevenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA *5*, 468-483.

Chamberlin, S.I., Merino, E.J., and Weeks, K.M. (2002). Catalysis of amide synthesis by RNA phosphodiester and hydroxyl groups. Proc Natl Acad Sci U S A *99*, 14688-14693.

Chang, C., and Craven, G.R. (1977). Identification of several proteins involved in the messenger RNA binding site of the 30 S ribosome by inactivation with 2-methoxy-5-nitrotropone. J Mol Biol *117*, 401-418.

Chang, M.L., Chen, J.C., Alonso, C.R., Kornblihtt, A.R., and Bissell, D.M. (2004). Regulation of fibronectin splicing in sinusoidal endothelial cells from normal or injured liver. Proc Natl Acad Sci U S A *101*, 18093-18098.

Charpentier, B., and Rosbash, M. (1996). Intramolecular structure in yeast introns aids the early

steps of in vitro spliceosome assembly. RNA *2*, 509-522.

Chen, I.T., and Chasin, L.A. (1993). Direct selection for mutations affecting specific splice sites in a hamster dihydrofolate reductase minigene. Mol Cell Biol *13*, 289-300.

Chen, Y.-I.G., Moore, R.E., Ge, H.Y., Young, M.K., Lee, T.D., and Stevens, S.W. (2007). Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. Nucleic Acids Res *35*, 3928-3944.

Chou, M.Y., Underwood, J.G., Nikolic, J., Luu, M.H., and Black, D.L. (2000). Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. Mol Cell *5*, 949-957.

Chung, N.H., and Tabata, M. (2002). Selective extraction of gold(III) in the presence of Pd(II) and Pt(IV) by salting-out of the mixture of 2-propanol and water. Talanta *58*, 927-933.

Cook, C.R., and McNally, M.T. (1998). SR protein and snRNP requirements for assembly of the Rous sarcoma virus negative regulator of splicing complex in vitro. Virology *242*, 211-220.

Coulter, L.R., Landree, M.A., and Cooper, T.A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. Mol Cell Biol *17*, 2143-2150.

Crooke, S.T. (1994). Progress in evaluation of the potential of antisense technology. Antisense Res Dev *4*, 145-146.

Crooke, S.T. (2000). Progress in antisense technology: the end of the beginning. Methods Enzymol *313*, 3-45.

Crooke, S.T. (2004). Progress in antisense technology. Annu Rev Med *55*, 61-95.

Damgaard, C.K., Tange, T.O., and Kjems, J. (2002). hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. RNA *8*, 1401-1415.

Dauksaite, V., and Akusjarvi, G. (2002). Human splicing factor ASF/SF2 encodes for a repressor domain required for its inhibitory activity on pre-mRNA splicing. J Biol Chem *277*, 12579-12586.

Del Gatto-Konczak, F., Olive, M., Gesnel, M.C., and Breathnach, R. (1999). hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. Mol Cell Biol *19*, 251-260.

Deschamps, S., Viel, A., Denis, H., and le Maire, M. (1991). Purification of two thermostable components of messenger ribonucleoprotein particles (mRNPs) from Xenopus laevis oocytes, belonging to a novel class of RNA-binding proteins. FEBS Lett *282*, 110-114.

Deshler, J.O., and Rossi, J.J. (1991). Unexpected point mutations activate cryptic 3' splice sites by perturbing a natural secondary structure within a yeast intron. Genes Dev *5*, 1252-1263.

Ding, J.-H., Xu, X., Yang, D., Chu, P.-H., Dalton, N.D., Ye, Z., Yeakley, J.M., Cheng, H., Xiao, R.-P., Ross, J.*, et al.* (2004). Dilated cardiomyopathy caused by tissue-specific ablation of SC35 in the heart. EMBO J *23*, 885-896.

Ding, J.H., Xu, X., Yang, D., Chu, P.H., Dalton, N.D., Ye, Z., Yeakley, J.M., Cheng, H., Xiao, R.P., Ross, J.*, et al.* (2004). Dilated cardiomyopathy caused by tissue-specific ablation of SC35 in the heart. EMBO J *23*, 885-896.

Ding, Y., Chan, C.Y., and Lawrence, C.E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res *32*, 135-141.

Dirksen, W.P., Mohamed, S.A., and Fisher, S.A. (2003). Splicing of a myosin phosphatase targeting subunit 1 alternative exon is regulated by intronic cis-elements and a novel bipartite exonic enhancer/silencer element. J Biol Chem *278*, 9722-9732.

Dove, H., and Mayes, R.W. (2006). Protocol for the analysis of n-alkanes and other plant-wax compounds and for their use as markers for quantifying the nutrient supply of large mammalian herbivores. Nat Protoc *1*, 1680-1697.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol *3*, 195-205.

Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.P., and Ehresmann, B. (1987). Probing the structure of RNAs in solution. Nucleic Acids Res *15*, 9109-9128.

Eldridge, A.G., Li, Y., Sharp, P.A., and Blencowe, B.J. (1999). The SRm160/300 splicing coactivator is required for exon-enhancer function. Proc Natl Acad Sci U S A *96*, 6125-6130.

Eperon, L.P., Graham, I.R., Griffiths, A.D., and Eperon, I.C. (1988). Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? Cell *54*, 393-401.

Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007-1013.

Fattal, E., Couvreur, P., and Dubernet, C. (2004). "Smart" delivery of antisense oligonucleotides by anionic pH-sensitive liposomes. Adv Drug Deliv Rev *56*, 931-946.

Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Genes Dev *17*, 419-437.

Fedorov, A., Saxonov, S., Fedorova, L., and Daizadeh, I. (2001). Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. Nucleic Acids Res *29*, 1464-1469.

Fei, J., and Zhang, Y. (2005). Prediction of VEGF mRNA antisense oligodeoxynucleotides by RNA structure software and their effects on HL60 and K562 cells. Cell Biol Int *29*, 737-741.

Forch, P., Puig, O., Kedersha, N., Martinez, C., Granneman, S., Seraphin, B., Anderson, P., and Valcarcel, J. (2000). The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. Mol Cell *6*, 1089-1098.

Forch, P., Puig, O., Martinez, C., Seraphin, B., and Valcarcel, J. (2002). The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites. EMBO J *21*, 6882-6892.

Geng, H., Liu, Y.-M., Chan, W.-S., Lo, A.W.-I., Au, D.M.-Y., Waye, M.M.-Y., and Ho, Y.-Y. (2005). The putative protein 6 of the severe acute respiratory syndrome-associated coronavirus: expression and functional characterization. FEBS Lett *579*, 6763-6768.

Giegerich, R., Voss, B., and Rehmsmeier, M. (2004). Abstract shapes of RNA. Nucleic Acids Res *32*, 4843-4851.

Gimautdinova, O.I., Karpova, G.G., Knorre, D.G., and Kobetz, N.D. (1981). The proteins of the messenger RNA binding site of Escherichia coli ribosomes. Nucleic Acids Res *9*, 3465-3481.

Goguel, V., and Rosbash, M. (1993). Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. Cell *72*, 893-901.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. Mol Cell *22*, 769-781.

Gorlach, M., Burd, C.G., and Dreyfuss, G. (1994). The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. J Biol Chem *269*, 23074-23078.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197-1211.

Graveley, B.R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell *123*, 65-73.

Gromak, N., Matlin, A.J., Cooper, T.A., and Smith, C.W.J. (2003). Antagonistic regulation of alpha-actinin alternative splicing by CELF proteins and polypyrimidine tract binding protein. RNA *9*, 443-456.

Grunweller, A., Gillen, C., Erdmann, V.A., and Kurreck, J. (2003). Cellular uptake and localization of a Cy3-labeled siRNA specific for the serine/threonine kinase Pim-1. Oligonucleotides *13*, 345-352.

Grunweller, A., Wyszko, E., Bieber, B., Jahnel, R., Erdmann, V.A., and Kurreck, J. (2003). Comparison of different antisense strategies in mammalian cells using locked nucleic acids, 2'-O-methyl RNA, phosphorothioates and small interfering RNA. Nucleic Acids Res *31*, 3185-3193.

Heinrichs, V., Ryner, L.C., and Baker, B.S. (1998). Regulation of sex-specific selection of fruitless 5' splice sites by transformer and transformer-2. Mol Cell Biol *18*, 450-458.

Hiller, M., Zhang, Z., Backofen, R., and Stamm, S. (2007). Pre-mRNA secondary structures influence exon recognition. PLoS Genet *3*.

Ho, S.P., Bao, Y., Lesher, T., Malhotra, R., Ma, L.Y., Fluharty, S.J., and Sakai, R.R. (1998). Mapping of RNA accessible sites for antisense experiments with oligonucleotide libraries. Nat Biotechnol *16*, 59-63.

Ho, S.P., Britton, D.H., Stone, B.A., Behrens, D.L., Leffet, L.M., Hobbs, F.W., Miller, J.A., and

Trainor, G.L. (1996). Potent antisense oligonucleotides to the human multidrug resistance-1 mRNA are rationally selected by mapping RNA-accessible sites with oligonucleotide libraries. Nucleic Acids Res *24*, 1901-1907.

Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J *24*, 1988-1998.

Hussain, M., Shchepinov, M., Sohail, M., Benter, I.F., Hollins, A.J., Southern, E.M., and Akhtar, S. (2004). A novel anionic dendrimer for improved cellular delivery of antisense oligonucleotides. J Control Release *99*, 139-155.

Huttley, G.A., Easteal, S., Southey, M.C., Tesoriero, A., Giles, G.G., McCredie, M.R., Hopper, J.L., and Venter, D.J. (2000). Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. Nat Genet *25*, 410-413.

Ibrahim, E.C., Schaal, T.D., Hertel, K.J., Reed, R., and Maniatis, T. (2005). Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. Proc Natl Acad Sci U S A *102*, 5002-5007.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., and Darnell, R.B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25*, 359-371.

Jiang, Z., Cote, J., Kwon, J.M., Goate, A.M., and Wu, J.Y. (2000). Aberrant splicing of tau pre-mRNA caused by intronic mutations associated with the inherited dementia frontotemporal dementia with parkinsonism linked to chromosome 17. Mol Cell Biol *20*, 4036-4048.

Jin, Y., Lee, H., Zeng, S.X., Dai, M.-S., and Lu, H. (2003). MDM2 promotes p21waf1/cip1 proteasomal turnover independently of ubiquitylation. EMBO J *22*, 6365-6377.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Jumaa, H., and Nielsen, P.J. (1997). The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. EMBO J *16*, 5077-5085.

Jumaa, H., Wei, G., and Nielsen, P.J. (1999). Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. Curr Biol *9*, 899-902.

Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N., and Moore, M.J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. RNA *8*, 426-439.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell *12*, 5-14.

Jurica, M.S., Sousa, D., Moore, M.J., and Grigorieff, N. (2004). Three-dimensional structure of C complex spliceosomes by electron microscopy. Nat Struct Mol Biol *11*, 265-269.

Kadener, S., Fededa, J.P., Rosbash, M., and Kornblihtt, A.R. (2002). Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. Proc Natl Acad Sci U S A *99*, 8185-8190.

Kan, J.L., and Green, M.R. (1999). Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. Genes Dev *13*, 462-471.

Kanopka, A., Muhlemann, O., Petersen-Mahrt, S., Estmer, C., Ohrmalm, C., and Akusjarvi, G. (1998). Regulation of adenovirus alternative RNA splicing by dephosphorylation of SR proteins. Nature *393*, 185-187.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res *32*, 493-496.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol *17*, 909-915.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., and Nogues, G. (2004). Multiple links between transcription and splicing. RNA *10*, 1489-1498.

Krecic, A.M., and Swanson, M.S. (1999). hnRNP complexes: composition, structure, and function. Curr Opin Cell Biol *11*, 363-371.

Kulish, D., and Struhl, K. (2001). TFIIS enhances transcriptional elongation through an artificial

arrest site in vivo. Mol Cell Biol *21*, 4162-4168.

Kurreck, J. (2003). Nucleic acids chemistry and biology. Angew Chem Int Ed Engl *42*, 5384-5385.

Kurreck, J. (2003). Antisense technologies. Improvement through novel chemical modifications. Eur J Biochem *270*, 1628-1644.

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T.J., Sladek, R., and Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. Nat Genet *40*, 225-231.

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Serre, D., Zuzan, H., Clark, T.A., Schweitzer, A., Staples, M.K., Wang, H.*, et al.* (2007). Heritability of alternative splicing in the human genome. Genome Res *17*, 1210-1218.

Labourier, E., Adams, M.D., and Rio, D.C. (2001). Modulation of P-element pre-mRNA splicing by a direct interaction between PSI and U1 snRNP 70K protein. Mol Cell *8*, 363-373.

Labourier, E., Allemand, E., Brand, S., Fostier, M., Tazi, J., and Bourbon, H.M. (1999). Recognition of exonic splicing enhancer sequences by the Drosophila splicing repressor RSF1. Nucleic Acids Res *27*, 2377-2386.

Labourier, E., Bourbon, H.M., Gallouzi, I.E., Fostier, M., Allemand, E., and Tazi, J. (1999). Antagonism between RSF1 and SR proteins for both splice-site recognition in vitro and Drosophila development. Genes Dev *13*, 740-753.

Ladd, A.N., Charlet, N., and Cooper, T.A. (2001). The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol Cell Biol *21*, 1285-1296.

Lallena, M.J., Chalmers, K.J., Llamazares, S., Lamond, A.I., and Valcarcel, J. (2002). Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45. Cell *109*, 285-296.

Lam, B.J., Bakshi, A., Ekinci, F.Y., Webb, J., Graveley, B.R., and Hertel, K.J. (2003). Enhancer-dependent 5'-splice site control of fruitless pre-mRNA splicing. J Biol Chem *278*, 22740-22747.

Lambert, D., and Draper, D.E. (2007). Effects of osmolytes on RNA secondary and tertiary structure stabilities and RNA-Mg2+ interactions. J Mol Biol *370*, 993-1005.

Le Cao, K.-A., Gonzalez, I., and Dejean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics *25*, 2855-2856.

Levy, M., and Miller, S.L. (1999). The prebiotic synthesis of modified purines and their potential role in the RNA world. J Mol Evol *48*, 631-637.

Li, J., Hawkins, I.C., Harvey, C.D., Jennings, J.L., Link, A.J., and Patton, J.G. (2003). Regulation of alternative splicing by SRrp86 and its interacting proteins. Mol Cell Biol *23*, 7437-7447.

Li, J.Q., Xu, Z.J., Liu, F., and Du, S.C. (2007). [The effects of variations of HBV S gene on its RNA secondary structure in serious hepatitis B patients]. Zhonghua Gan Zang Bing Za Zhi *15*, 387-388.

Li, Y., and Blencowe, B.J. (1999). Distinct factor requirements for exonic splicing enhancer function and binding of U2AF to the polypyrimidine tract. J Biol Chem *274*, 35074-35079.

Liao, H.X., and Spremulli, L.L. (1990). Effects of length and mRNA secondary structure on the interaction of bovine mitochondrial ribosomes with messenger RNA. J Biol Chem *265*, 11761-11765.

Libri, D., Stutz, F., McCarthy, T., and Rosbash, M. (1995). RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. RNA *1*, 425-436.

Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J., and Fairbrother, W.G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A *108*, 11093-11098.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A *98*, 11193-11198.

Liu, H.X., Chew, S.L., Cartegni, L., Zhang, M.Q., and Krainer, A.R. (2000). Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol Cell Biol *20*, 1063-1071.

Liu, H.X., Zhang, M., and Krainer, A.R. (1998). Identification of functional exonic splicing

enhancer motifs recognized by individual SR proteins. Genes Dev *12*, 1998-2012.

Liu, X., Zhang, Z., Yang, L., Chen, D., and Wang, Y. (2000). [Inhibition of the activation and collagen production of cultured rat hepatic stellate cells by antisense oligonucleotides against transforming growth factor-beta 1 is enhanced by cationic liposome delivery]. Hua Xi Yi Ke Da Xue Xue Bao *31*, 133-135, 142.

Loeb, D.D., Mack, A.A., and Tian, R. (2002). A secondary structure that contains the 5' and 3' splice sites suppresses splicing of duck hepatitis B virus pregenomic RNA. J Virol *76*, 10195-10202.

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

Lynch, K.W., and Maniatis, T. (1995). Synergistic interactions between two distinct elements of a regulated splicing enhancer. Genes Dev *9*, 284-293.

Lynch, K.W., and Maniatis, T. (1996). Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer. Genes Dev *10*, 2089-2101.

Lysik, M.A., and Wu-Pong, S. (2003). Innovations in oligonucleotide drug delivery. J Pharm Sci *92*, 1559-1573.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature *416*, 499-506.

Manley, J.L., and Tacke, R. (1996). SR proteins and splicing control. Genes Dev *10*, 1569-1579.

Markovtsov, V., Nikolic, J.M., Goldman, J.A., Turck, C.W., Chou, M.Y., and Black, D.L. (2000). Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. Mol Cell Biol *20*, 7463-7479.

Martinez-Contreras, R., Fisette, J.-F., Nasim, F.-u.H., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol *4*.

Matteson, R.J., Biswas, S.J., and Steege, D.A. (1991). Distinctive patterns of translational reinitiation in the lac repressor mRNA: bridging of long distances by out-of-frame translation and

RNA secondary structure, effects of primary sequence. Nucleic Acids Res *19*, 3499-3506.

Matveeva, O.V., Mathews, D.H., Tsodikov, A.D., Shabalina, S.A., Gesteland, R.F., Atkins, J.F., and Freier, S.M. (2003). Thermodynamic criteria for high hit rate antisense oligonucleotide design. Nucleic Acids Res *31*, 4989-4994.

Matveeva, O.V., Tsodikov, A.D., Giddings, M., Freier, S.M., Wyatt, J.R., Spiridonov, A.N., Shabalina, S.A., Gesteland, R.F., and Atkins, J.F. (2000). Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity. Nucleic Acids Res *28*, 2862-2865.

Mayeda, A., and Krainer, A.R. (1992). Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. Cell *68*, 365-375.

McCullough, A.J., and Berget, S.M. (2000). An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. Mol Cell Biol *20*, 9225-9235.

McNally, L.M., and McNally, M.T. (1996). SR protein splicing factors interact with the Rous sarcoma virus negative regulator of splicing element. J Virol *70*, 1163-1172.

Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J Am Chem Soc *127*, 4223-4231.

Minovitsky, S., Gee, S.L., Schokrpur, S., Dubchak, I., and Conboy, J.G. (2005). The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucleic Acids Res *33*, 714-724.

Miriami, E., Margalit, H., and Sperling, R. (2003). Conserved sequence elements associated with exon skipping. Nucleic Acids Res *31*, 1974-1983.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet *30*, 13-19.

Morgan, D., Rutebemberwa, A., Malamba, S., Ross, A., Whitworth, J., Kaleebu, P., and Gotch, F. (1999). HIV-1 RNA levels in an African population-based cohort and their relation to CD4 lymphocyte counts and World Health Organization clinical staging. J Acquir Immune Defic Syndr *22*, 167-173.

Morishita, R., Gibbons, G.H., Horiuchi, M., Nakajima, M., Ellison, K.E., Lee, W., Kaneda, Y., Ogihara, T., and Dzau, V.J. (1997). Molecular Delivery System for Antisense Oligonucleotides: Enhanced Effectiveness of Antisense Oligonucleotides by HVJ-liposome Mediated Transfer. J Cardiovasc Pharmacol Ther *2*, 213-222.

Nagel, R.J., Lancaster, A.M., and Zahler, A.M. (1998). Specific binding of an exonic splicing enhancer by the pre-mRNA splicing factor SRp55. RNA *4*, 11-23.

Nilsen, T.W. (2002). The spliceosome: no assembly required? Mol Cell *9*, 8-9.

Nilsen, T.W. (2003). The spliceosome: the most complex macromolecular machine in the cell? Bioessays *25*, 1147-1149.

Ovchinnikov, L.P., Avanesov, A.T., Seriakova, T.A., Alzhanova, A.T., and Radzhabov, H.M. (1978). A comparison of the RNA-binding proteins with the proteins of polyribosomal messenger ribonucleoproteins in rabbit reticulocytes. Eur J Biochem *90*, 527-535.

Owen, N., Zhou, H., Malygin, A.A., Sangha, J., Smith, L.D., Muntoni, F., and Eperon, I.C. (2011). Design principles for bifunctional targeted oligonucleotide enhancers of splicing. Nucleic Acids Res.

Pagani, F., and Baralle, F.E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet *5*, 389-396.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Perez, I., Lin, C.H., McAfee, J.G., and Patton, J.G. (1997). Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. RNA *3*, 764-778.

Polydorides, A.D., Okano, H.J., Yang, Y.Y., Stefani, G., and Darnell, R.B. (2000). A brain-enriched polypyrimidine tract-binding protein antagonizes the ability of Nova to regulate neuron-specific alternative splicing. Proc Natl Acad Sci U S A *97*, 6350-6355.

Preker, P.J., and Guthrie, C. (2006). Autoregulation of the mRNA export factor Yra1p requires inefficient splicing of its pre-mRNA. RNA *12*, 994-991006.

Pret, A.M., Balvay, L., and Fiszman, M.Y. (1999). Regulated splicing of an alternative exon of beta-tropomyosin pre-mRNAs in myogenic cells depends on the strength of pyrimidine-rich intronic enhancer elements. DNA Cell Biol *18*, 671-683.

Raghavan, A., Robison, R.L., McNabb, J., Miller, C.R., Williams, D.A., and Bohjanen, P.R. (2001). HuA and tristetraprolin are induced following T cell activation and display distinct but overlapping RNA binding specificities. J Biol Chem *276*, 47958-47965.

Raker, V.A., Mironov, A.A., Gelfand, M.S., and Pervouchine, D.D. (2009). Modulation of alternative splicing by long-range RNA structures in Drosophila. Nucleic Acids Res *37*, 4533-4544.

Reid, D.C., Chang, B.L., Gunderson, S.I., Alpert, L., Thompson, W.A., and Fairbrother, W.G. (2009). Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. RNA *15*, 2385-2397.

Ring, H.Z., and Lis, J.T. (1994). The SR protein B52/SRp55 is essential for Drosophila development. Mol Cell Biol *14*, 7499-7506.

Rooke, N., Markovtsov, V., Cagavi, E., and Black, D.L. (2003). Roles for SR proteins and hnRNP A1 in the regulation of c-src exon N1. Mol Cell Biol *23*, 1874-1884.

Rudner, D.Z., Breger, K.S., and Rio, D.C. (1998). Molecular genetic analysis of the heterodimeric splicing factor U2AF: the RS domain on either the large or small Drosophila subunit is dispensable in vivo. Genes Dev *12*, 1010-1021.

Schaal, T.D., and Maniatis, T. (1999). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol Cell Biol *19*, 1705-1719.

Scholz, G., Hartson, S.D., Cartledge, K., Hall, N., Shao, J., Dunn, A.R., and Matts, R.L. (2000). p50(Cdc37) can buffer the temperature-sensitive properties of a mutant of Hck. Mol Cell Biol *20*, 6984-6995.

Schubert, S., Gul, D.C., Grunert, H.P., Zeichhardt, H., Erdmann, V.A., and Kurreck, J. (2003). RNA cleaving '10-23' DNAzymes with enhanced stability and activity. Nucleic Acids Res *31*, 5982-5992.

Shen, H., and Green, M.R. (2004). A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. Mol Cell *16*, 363-373.

Shen, H., Kan, J.L.C., Ghigna, C., Biamonti, G., and Green, M.R. (2004). A single polypyrimidine tract binding protein (PTB) binding site mediates splicing inhibition at mouse IgM exons M1 and M2. RNA *10*, 787-794.

Shen, H., Kan, J.L.C., and Green, M.R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Mol Cell *13*, 367-376.

Shepard, P.J., and Hertel, K.J. (2008). Conserved RNA secondary structures promote alternative splicing. RNA *14*, 1463-1469.

Shi, H., Hoffman, B.E., and Lis, J.T. (1997). A specific RNA hairpin loop structure binds the RNA recognition motifs of the Drosophila SR protein B52. Mol Cell Biol *17*, 2649-2657.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., and Kielkopf, C.L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. Mol Cell *23*, 49-59.

Siegel, R.W., Bellon, L., Beigelman, L., and Kao, C.C. (1999). Use of DNA, RNA, and chimeric templates by a viral RNA-dependent RNA polymerase: evolutionary implications for the transition from the RNA to the DNA world. J Virol *73*, 6424-6429.

Siepel, A., and Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol *11*, 413-428.

Singh, N.N., Singh, R.N., and Androphy, E.J. (2007). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Res *35*, 371-389.

Singh, R., Valcarcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science *268*, 1173-1176.

Sohail, M., and Southern, E.M. (2000). Selecting optimal antisense reagents. Adv Drug Deliv Rev *44*, 23-34.

Southby, J., Gooding, C., and Smith, C.W. (1999). Polypyrimidine tract binding protein functions

as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons. Mol Cell Biol *19*, 2699-2711.

Spirin, A.S. (1979). Messenger ribonucleoproteins (informosomes) and RNA-binding proteins. Mol Biol Rep *5*, 53-57.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat *21*, 577-581.

Stickeler, E., Fraser, S.D., Honig, A., Chen, A.L., Berget, S.M., and Cooper, T.A. (2001). The RNA binding protein YB-1 binds A/C-rich exon enhancers and stimulates splicing of the CD44 alternative exon v4. EMBO J *20*, 3821-3830.

Sun, H., and Chasin, L.A. (2000). Multiple splicing defects in an intronic false exon. Mol Cell Biol *20*, 6414-6425.

Suzuki, H., Jin, Y., Otani, H., Yasuda, K., and Inoue, K. (2002). Regulation of alternative splicing of alpha-actinin transcript by Bruno-like proteins. Genes Cells *7*, 133-141.

Swanson, M.S., and Dreyfuss, G. (1988). RNA binding specificity of hnRNP proteins: a subset bind to the 3' end of introns. EMBO J *7*, 3519-3529.

Szathmary, E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. Trends Genet *15*, 223-229.

Tacke, R., and Manley, J.L. (1995). The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. EMBO J *14*, 3540-3551.

Tacke, R., and Manley, J.L. (1999). Determinants of SR protein specificity. Curr Opin Cell Biol *11*, 358-362.

Tange, T.O., Damgaard, C.K., Guth, S., Valcarcel, J., and Kjems, J. (2001). The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. EMBO J *20*, 5748-5758.

Taylor, N.R., and Rossi, J.J. (1991). Ribozyme-mediated cleavage of an HIV-1 gag RNA: the effects of nontargeted sequences and secondary structure on ribozyme cleavage activity. Antisense Res Dev *1*, 173-186.

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

Tian, H., and Kole, R. (1995). Selection of novel exon recognition elements from a pool of random sequences. Mol Cell Biol *15*, 6291-6298.

Tian, M., and Maniatis, T. (1994). A splicing enhancer exhibits both constitutive and regulated activities. Genes Dev *8*, 1703-1712.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212-1215.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580-586.

Valcarcel, J., Singh, R., Zamore, P.D., and Green, M.R. (1993). The protein Sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA. Nature *362*, 171-175.

Vaux, D.L., and Silke, J. (2005). IAPs, RINGs and ubiquitylation. Nat Rev Mol Cell Biol *6*, 287-297.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Vickers, T.A., Wyatt, J.R., and Freier, S.M. (2000). Effects of RNA secondary structure on cellular antisense activity. Nucleic Acids Res *28*, 1340-1347.

Villemaire, J., Dion, I., Elela, S.A., and Chabot, B. (2003). Reprogramming alternative pre-messenger RNA splicing through the use of protein-binding antisense oligonucleotides. J Biol Chem *278*, 50031-50039.

Wagner, E.J., and Garcia-Blanco, M.A. (2001). Polypyrimidine tract binding protein antagonizes exon definition. Mol Cell Biol *21*, 3281-3288.

Wagner, E.J., and Garcia-Blanco, M.A. (2002). RNAi-mediated PTB depletion leads to enhanced exon definition. Mol Cell *10*, 943-949.

Wahl, M.C., Will, C.L., and Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

Walstrom, K.M., Dozono, J.M., and von Hippel, P.H. (1998). Effects of reaction conditions on RNA secondary structure and on the helicase activity of Escherichia coli transcription termination factor Rho. J Mol Biol *279*, 713-726.

Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet *8*, 749-761.

Wang, G.S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet *8*, 749-761.

Wang, J., and Bell, L.R. (1994). The Sex-lethal amino terminus mediates cooperative interactions in RNA binding and is essential for splicing regulation. Genes Dev *8*, 2072-2085.

Wang, J., Takagaki, Y., and Manley, J.L. (1996). Targeted disruption of an essential vertebrate gene: ASF/SF2 is required for cell viability. Genes Dev *10*, 2588-2599.

Wang, J., Xiao, S.H., and Manley, J.L. (1998). Genetic analysis of the SR protein ASF/SF2: interchangeability of RS domains and negative control of splicing. Genes Dev *12*, 2222-2233.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.

Wells, K.E., Fletcher, S., Mann, C.J., Wilton, S.D., and Wells, D.J. (2003). Enhanced in vivo delivery of antisense oligonucleotides to restore dystrophin expression in adult mdx mouse muscle. FEBS Lett *552*, 145-149.

Wu, H., Lima, W.F., Zhang, H., Fan, A., Sun, H., and Crooke, S.T. (2004). Determination of the role of the human RNase H1 in the pharmacology of DNA-like antisense drugs. J Biol Chem *279*, 17181-17189.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061-1070.

Xu, X., Yang, D., Ding, J.-H., Wang, W., Chu, P.-H., Dalton, N.D., Wang, H.-Y., Bermingham, J.R., Ye, Z., Liu, F.*, et al.* (2005). ASF/SF2-regulated CaMKIIdelta alternative splicing temporally

reprograms excitation-contraction coupling in cardiac muscle. Cell *120*, 59-72.

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

Yarus, M. (1999). Boundaries for an RNA world. Curr Opin Chem Biol *3*, 260-267.

Yasuda, S., Wada, S., Arao, Y., Kogawa, M., Kayama, F., and Katayama, S. (2004). Interaction between 3' untranslated region of calcitonin receptor messenger ribonucleic acid (RNA) and adenylate/uridylate (AU)-rich element binding proteins (AU-rich RNA-binding factor 1 and Hu antigen R). Endocrinology *145*, 1730-1738.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

Yeo, G., Holste, D., Kreiman, G., and Burge, C.B. (2004). Variation in alternative splicing across human tissues. Genome Biol *5*.

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

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.-D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol *16*, 130-137.

Yoo, H., and Juliano, R.L. (2000). Enhanced delivery of antisense oligonucleotides with fluorophore-conjugated PAMAM dendrimers. Nucleic Acids Res *28*, 4225-4231.

Zhang, L., Liu, W., and Grabowski, P.J. (1999). Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB. RNA *5*, 117-130.

Zhang, W., Liu, H., Han, K., and Grabowski, P.J. (2002). Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. RNA *8*, 671-685.

Zhang, X.H.F., and Chasin, L.A. (2004). Computational definition of sequence motifs governing

constitutive exon splicing. Genes Dev *18*, 1241-1250.

Zhang, X.H.F., Kangsamaksin, T., Chao, M.S.P., Banerjee, J.K., and Chasin, L.A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. Mol Cell Biol *25*, 7323-7332.

Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. Nature *419*, 182-185.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol Cell *8*, 1351-1361.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res *31*, 3406-3415.

Zuo, P., and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes Dev *10*, 1356-1368.