A trio of unique alternative splicing patterns: The splicing of tandem NAGNAG acceptors, transcription-start-site-dependent and mutually dependent cassette exons

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Tzu-Ming Chern

aus
South Africa

Basel, 2008

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Professor Mihaela Zavolan und Professor Torsten Schwede.

Professor Mihaela Zavolan

Basel, den 20. Mai 2008

Abstract

With the rapid increase in the volume of genomic and transcript data in mouse and human, a diverse set of alternative splicing patterns can be discovered. We have set out to explore in more depth some of these unique splicing patterns which include: i.) tandem acceptor splice sites termed 'NAGNAGs', which cause an exon length variation of three nucleotides, ii.) the splicing of internal cassette exons, whose inclusion and exclusion are strongly associated with their transcriptional start sites, iii.) and groups of internal cassette exons that are always observed to be either included or excluded together within a transcript. We did not find much evidence of functional potential for the majority of variant acceptor splice sites carrying the NAGNAG motif and thus conclude that their abundance is due mostly to the stochastic behaviour of the spliceosome. We inferred that a large fraction (15-30%) of internal cassette exons in transcription units with multiple start sites are included and skipped in a transcription-start-site-dependent manner. We did not find that this relationship is conserved in orthologous human and mouse exons. Our first analysis has revealed several interesting trends in mutually dependent exons when compared to mutually exclusive and constitutive exons: these exons have a stronger pressure to maintain the reading frame as a group of exons rather than individually, and they generally have shorter intron lengths. Ours are the first analyses of transcription-start-site-dependent and mutually-dependent splicing. Their mechanisms remain to be further elucidated and our results provide a good starting point for future computational and experimental studies.

Declaration

I, Tzu-Ming Chern, declare that I have written the thesis entitled "A trio of unique alternative splicing patterns: The splicing of tandem NAGNAG acceptors, transcription-start-site-dependent and mutually dependent cassette exons". The thesis was conducted under the supervision of Professor Mihaela Zavolan and it is submitted for the degree of Doctor of Philosophy in the Faculty of Science at the University of Basel, Basel. No part of this research has been submitted to any other University.

Acknowledgements

Special thanks goes to the following people:

1. Professor Mihaela Zavolan for supervising my project and always offering constructive advice.

2. South African National Research Foundation for my doctoral scholarship - special thanks to Rose Robertson for facilitating the processing of funds and documents.

3. To my wonderful sister, Tzu-Mei, for her loving support in times of need.

4. Professor Gregory Blatch for believing in me.

5. Warm thanks to my friends (Sandra,Monika,Barbara,Suchi,Edward,Kylie,Jeannine) for their understanding and support.

6. Much thanks to my colleagues for their patience and understanding.

7. Thanks to Torsten for letting me use his empty office during my thesis writeup.

8. Dr. Gonzalos Lopez for his help on FireDB.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What is alternative splicing?

Splicing is an mRNA processing step whereby introns are removed from the pre-mRNA and exons are ligated to form the mature mRNA. The molecular tool responsible for removal of introns is a ribonucleoprotein complex called the spliceosome. The active spliceosome consists of five small nuclear ribonucleoprotein particles (snRNPs) that are associated with a large number of splicing factors. The spliceosome generally recognizes oligonucleotide sequences as well as splice sites (donor and acceptor) which are found at the ends of the intron and are highly conserved throughout eukaryotes. How the spliceosome specifically selects a particular splice site depends on several factors which will be discussed in section 1.2. The usage of a particular donor splice site may be used in conjunction with different acceptor splice sites and vice versa, which may result in different variants of the gene.

The initial steps of splicing are depicted in Fig. 1.1. The process begins with the positioning and rearrangement of splicing factors on the pre-mRNA, which eventually leads to the assembly of an active spliceosome that carries out intron removal.

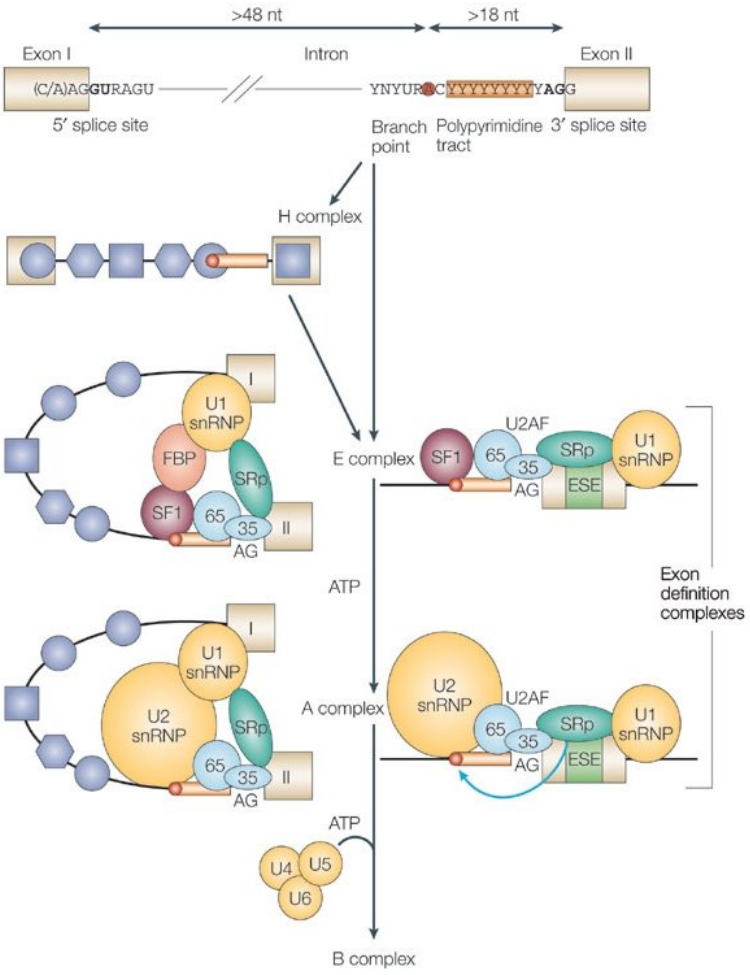


Figure 1.1: **Mammalian spliceosomal assembly.** The assembly of the spliceosome begins with the assembly of five small nuclear ribonucleoprotein particles (snRNPs) - U1, U2, U4, U5, and U6, in which these snRNPs associate with other proteins. The H-complex consists of pre-mRNAs associating with heterogeneous nuclear ribonucleoproteins (hnRNPs). In the E-complex, all the consensus splice-site signals are recognized - U1 snRNP binds to the 5' splice site, splicing factor 1 (SF1) binds to the branch point, the large and small subunits of U2 auxiliary factor(U2AF) bind to the polypyrimidine tract and acceptor splice site respectively, and lastly serine-arginine proteins (SRp) bind to exonic splicing enhancers, U2AF, U1 snRNP, and the branch point. Spliceosomal assemblies across the intron and exon are shown for E and A complexes, however, assembly across the intron must occur before the assembly of the active spliceosome. The formation of the A-complex results from ATP hydrolysis, during which the U2 snRNP binds to the branch point and the 3' splice-site sequence. The B-complex and the final catalytic complex are formed through the incorporation and rearrangements of the tri-snRNPs (U4,U5, and U6). Figure extracted from Matlin et al. [1].

The splicing reaction is essentially a chemical reaction. The chemistry of intron removal and exon ligation occurs via two transesterification reactions as illustrated in Fig.1.2.

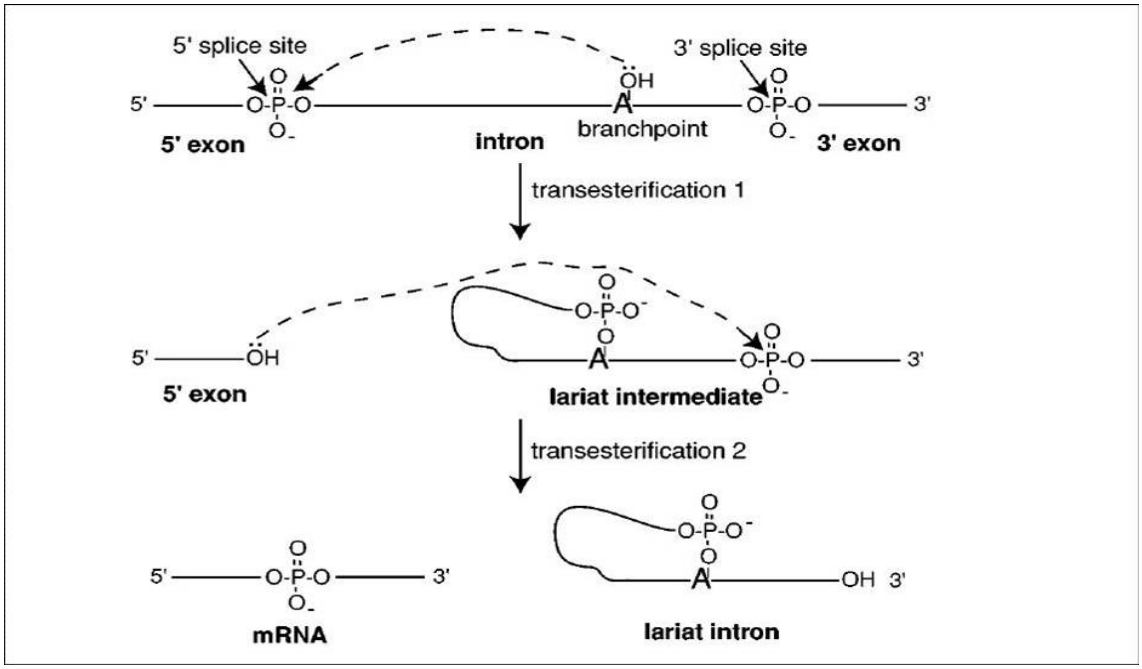


Figure 1.2: **The chemistry of the splicing reaction.** During the first transesterification reaction, the 2' oxygen of the adenosine residue (at the branch point) forms a bond with the phosphorus atom at the donor splice site, resulting in a lasso structure between the 5' end of the intron and 3' exon. The oxygen atom at the 3' hydroxyl group of the 5' exon then forms a bond with the phosphorus atom at the acceptor splice site, thus resulting in the ligation of exons and excision of the lariat intron. Figure extracted from Brow, D. [2].

1.2 Factors affecting alternative splice-site choice

The selection of different splice sites can generate different alternative splicing patterns as illustrated in Fig. 1.3. The frequency of occurrence of these splicing patterns varies across species, with exon skipping being most frequent in mammalians and intron-retention occurring most frequently in plants [1].

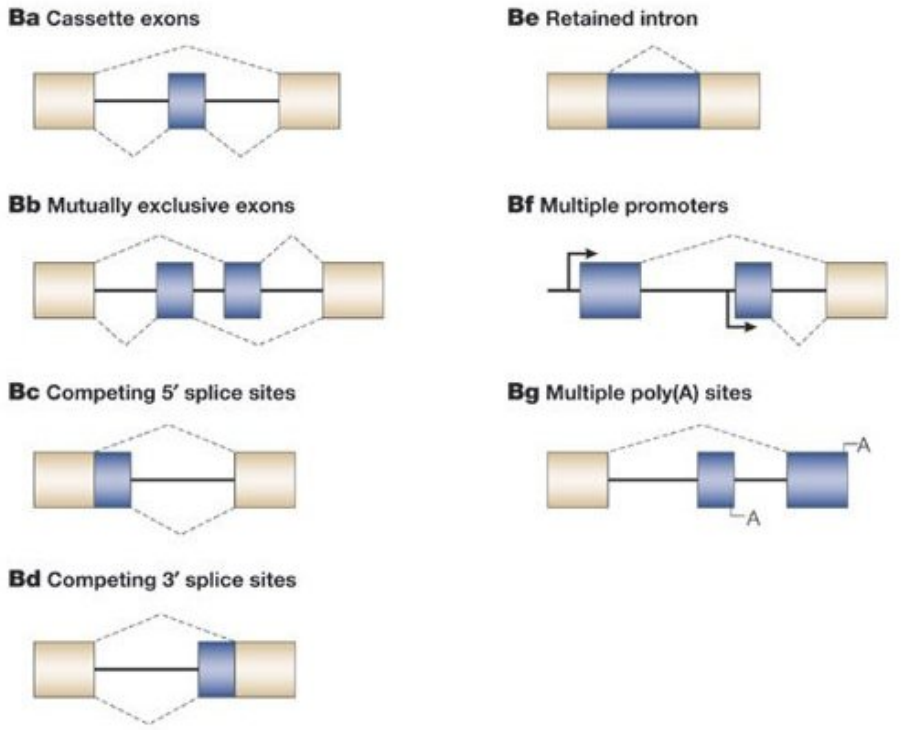


Figure 1.3: **The different types of alternative splicing.** The term 'cassette exons' describes exons that are skipped in some transcripts and included in others (Ba). Mutually exclusive exons are alternative exons that are never included or excluded together in the same transcript (Bb). Competing donor (Bc) and acceptor splice sites (Bd) result in exon-length variations from the donor and acceptor splice sites respectively. Intron retention (Be) is when the intron is not removed and is present within the mature transcript. The usage of alternative promoters (Bf) and polyadenylation sites (Bg) can result in different transcript variants. Figure extracted from Matlin et al. [1].

The actual mechanisms for splice-site selection may vary from gene to gene, but they typically involve repressing or enhancing the use of specific alternative splice-sites through recruitment of competing splicing factors and usage of inhibitory or enhancing regulatory elements. In essence, as said by Franciso Baralle [3] "it is the fine balance of power between a myriad of controlling factors" that determines the final splicing decision. These splice-site enhancing or repressing factors and sequences that they bind are described in the following subsections.

1.2.1 Splicing regulatory elements

Splicing regulatory elements are classified into two types: enhancer and silencer elements, both of which can be found in introns and exons. As their name implies, splicing enhancer and silencer elements are bound by factors that enhance or repress respectively, the usage of a particular splice site. Some examples of different enhancer and silencer elements affecting the usage of splice sites are tabulated in Table 1.1.

Table 1.1: **Enhancer and silencer elements affecting alternatively spliced events**

Sequence element	Splicing factor	Mechanism	Effect	References
Intronic enhancer	Rodent TIA1	Enhances the binding of U1 snRNP to a weak 5' splice site	Leads to the inclusion of the K-SAM alternative exon	Del Gatto-Konczak et al. 2000 [4]
Exonic enhancer	Drosophila RBP1 protein, TRA, and TRA2	Enhances recruitment of U2AF65 to weaker polypyrimidine tracts	Inclusion of female-specific exon	Lynch and Maniatis 1996 [5]
Intronic silencer	Sex-lethal	Blocks U2AF binding	Allows usage of downstream female-specific acceptor site	Valcarcel J. 1993 [6]
Exonic silencer	hnRNP A1	Blocks U2AF binding	Leads to inclusion of HIV1 tat exon 3	Tange et al. 2001 [7]

1.2.2 Concentration of splicing factors

The splicing regulatory elements mentioned previously can be bound by both splicing factors such as SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). It is known that the activity of SF2/ASF can be antagonized by hnRNPA1 [8, 9] and it has been shown that the relative concentrations of these antagonistic splicing factors can influence the outcome of splice events. Caceres et al. [10] have shown that over-expression of SF2/ASF activates the usage of proximal donor splice sites, and promotes inclusion of a neuron-specific exon, whereas increased expression of hnRNP A1 promotes the usage of distal donor splice sites. Hanamura et al. [11] have suggested that the alternative splicing of specific transcripts may be subjected to different levels of hnRNP A/B and SR proteins, but at present - one cannot determine which and how many transcripts are differentially spliced as a consequence of variable concentrations of these splicing factors.

1.2.3 Dual functions of splicing factors

Some splicing factors can function as both activators and repressors. Recently, the hnRNP M protein has been shown to promote the exclusion of the FGFR2-IIIb exon presumably by binding to the ISS flanking this exon. On the other hand, hnRNP M promotes the inclusion of a troponin-derived exon flanked by muscle-specific ISEs [12]. Other splicing regulators such as Nova-1 [13], Fox-1 [14], the CELF protein family [15], and hnRNPL [16] have also been shown to promote exon inclusion and exclusion.

1.2.4 Splice-site strengths

Constitutive exons usually have stronger splice-site signals when compared to their alternative exons, which suggest that alternative exons can be more easily regulated than their constitutive exons [17]. Evidence from literature have shown that strengthening of a weak donor splice site results in exon 4 inclusion of the rat preprotachykinin gene (encodes a neuropeptide) [18]. The polypyrimidine tract is important for efficient branch point usage and acceptor splice-site recognition and

its mutation leads to either increased or decreased splice-site strength. For instance, Dominski and Kole [19] have demonstrated that exon skipping is reversed when the strength of the polypyrimidine tract have been increased. The strength of the splice sites flanking cassette exons also influences its responsiveness to transcriptional elongation, whereby Nogues et al. [20] have observed that weaker acceptor splice sites flanking the fibronectin extra domain I exon, increases its responsiveness to the rate of elongation of the RNA polymerase.

1.2.5 Exon- and intron-length

Although, the exon- and intron-lengths do not directly affect alternative splicing, size constraints imposed by interactions between splicing factors do exist. Sterner et al. [21] investigated the relationship between intron and exon size in pre-mRNA processing and have found that splicing became problematic when intron and exon sizes exceeded 500 nucleotides. Furthermore, his findings has revealed an inverse relationship between exon and intron size. Small exons are efficiently included when flanked by large introns. As the exon size increases, they are less efficiently included when flanked by large introns, but are included when the intron size decreases.

1.2.6 Impact of RNA secondary structure on alternative splicing

Despite the unavailability of experimental pre-mRNA secondary structures, there has been an increase in the observations of highly conserved complementary sequences that can form putative RNA secondary structures surrounding alternative exons, thereby suggesting that RNA secondary structure formation is a likely mechanism for influencing alternative splicing events.

Models of how RNA secondary structures can affect splicing have been put forth by Buratti and Baralle [22]. They proposed that if RNA secondary structures did form in splicing factor binding sites, such as exon-intron junctions and splicing regulatory elements - this would affect the proper binding of splicing factors and thus influence the downstream splicing events.

An example of putative RNA secondary structures regulating the splicing of the *Drosophila* homothorax gene (*hth*) is illustrated in Figure 1.4.

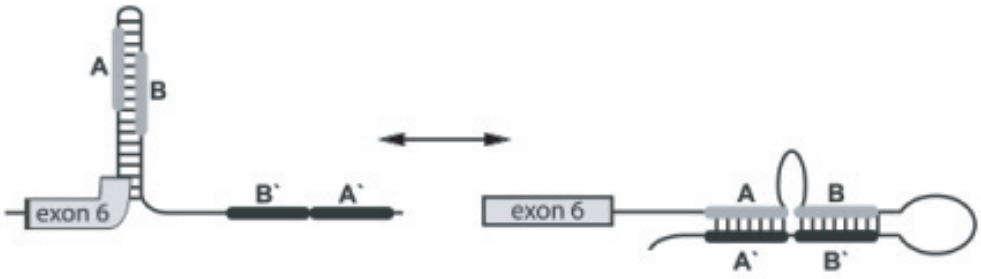


Figure 1.4: **Alternative RNA secondary structures formed by the conserved elements of the *hth* gene.** Alternative splicing of the *hth* gene involves partial intron retention of the intron between exon 6 and 7, which occurs when the donor splice site flanking exon 6 is not used. The figure shows proposed RNA secondary structures that can either block U1 snRNP binding to the donor splice site (left RNA secondary structure) or allow accessibility at the donor splice site of exon 6. The ultraconserved elements, A and B, can base-pair with each other to form an RNA hairpin structure or base-pair with another downstream conserved pair of elements (A′ and B′) to form another RNA hairpin. Figure extracted from Glazov et al. [23].

1.3 Consequences of alternative splicing

1.3.1 A diversity of molecular properties of proteins and mRNAs arising from alternative splicing

Alternative splicing can lead to many changes in the molecular properties of proteins and mRNAs - the majority of the alterations being changes in binding specificities and functionalities. Many apoptotic factors involved in programmed cell death are also alternatively spliced. Some important functional consequences of alternative splicing of known proteins are summarized in Table 1.2.

Table 1.2: **Proteins with altered molecular properties as a consequence of alternative splicing**

Gene name	Cellular function	Effect of splicing	References
Dopamine D(2) receptor	Binds to the neurotransmitter, dopamine, which has a variety of functions in the brain	Affects the degree of retention of dopamine D(2) receptor isoforms within the endoplasmic reticulum	Prou et al. 2001 [24]
Staufen RNA-binding protein	Required for the localization of multiple mRNA species during oogenesis and zygotic development	Rat staufen isoforms bind with different affinities to RNA	Monshausen 2001 [25]
Agrin (a proteoglycan)	Induces aggregation of acetylcholine receptors	Agrin isoforms bind with different binding affinities to heparin, which inhibits aggregation of acetylcholine receptors	Gesemann 1996 [26]
Cytochrome P450	Involved in drug metabolism and involved in the synthesis of steroid hormones	Cytochrome P450 isoforms differ in substrate specificity	Christmas et al. 2001 [27]
Thyroperoxidase	Involved in thyroid hormone synthesis	Reduces the protein half-life	Niccoli-Sire et al.2001 [28]
Beta-site amyloid precursor protein cleaving enzyme (BACE)	Cleaves Alzheimer's beta amyloid precursor protein to form the pathogenic amyloid beta peptides found in senile plaques of Alzheimer patients	Loss of glycosylation sites in BACE isoforms leads to weaker enzyme activity	Tanahashi and Tabira 2001 [29]
Caspases (caspase-2,-9,-10)	Initiator caspases, which activate other caspases during apoptosis	Results in antagonistic functions during apoptosis, dominant-negative phenotype, and changed activity	Schwerk and Schulze-Osthoff 2005 [30]

1.3.2 *Drosophila* sex determination is regulated by alternative splicing

Perhaps the most well-known example of alternative splicing regulation is the sex determination in *Drosophila*. The *Drosophila* sex-lethal (*Sxl*) protein is the key RNA-binding protein that determines the sex of a fly. Active *Sxl* proteins are produced in female flies primarily because transcription of the *Sxl* gene utilizes the early upstream promoter. In the male flies, the downstream promoter is utilized instead, leading to inactive *Sxl* products. Active *Sxl* products, promote the splicing and production of downstream active RNA-binding proteins such as transformer (*tra*), which in turn interacts with Transformer2 protein to produce female-specific doublesex (*dsx*) isoforms. In contrast, inactive *Sxl* products result in the downstream production of male-specific *dsx* isoforms. The selection of the early upstream promoter in females is due to four transcription factors present on the X-chromosome, that bind directly to the upstream promoter and are expressed prior to dosage compensation. The expression of these transcription factors are twice as high in females compared to males, thereby enabling a stable expression from the early, upstream promoter [31].

1.3.3 The role of nonsense-mediated decay in alternative splicing

Nonsense-mediated decay (NMD) is a mechanism by which cells remove deleterious transcripts. In mammals, NMD occurs when a termination codon resides more than 50-55 nucleotides upstream of an exon-exon junction [32]. The link between NMD and alternative splicing has been supported by several studies [33, 34], that suggest that NMD might play an important role in the removal of PTC-containing (premature termination codon containing) isoforms. However, not all PTC-containing splice variants are regulated by NMD, in fact only a small proportion of these transcripts are regulated by NMD, since the authors have observed uniformly low levels of PTC-containing transcripts in normal mammalian tissues which are independent of NMD [35].

1.3.4 Tissue-specific alternative splicing

An important function of alternative splicing would be to generate tissue-specific isoforms. One observed mechanism that controls the tissue-specific splicing patterns of variant transcripts is their regulation by tissue-specific splicing factors. Of particular note, the splicing factor, NOVA-1, is a neuron-specific splicing factor which is essential for neuronal viability [36]. Other essential splicing factors such as SR proteins and hnRNPs have been observed to have varying concentrations in different tissues [11], which suggest that the selective concentrations of these antagonistic factors might play a role in altering splicing patterns in a tissue-dependent manner.

1.3.5 Altered splicing in cancer and diseases arising from mutations or factors affecting enhancer regulatory elements

Mutations in splicing regulatory elements, overexpression of splicing factors, and overexpression of abberant alternatively spliced transcripts can give rise to diseases whose fundamental pathogenic mechanism is aberrant expression of splice forms. Missplicing has been observed in a number of tumour-associated genes. The breast cancer susceptibility gene, BRCA1, is an example in which mutations in regulatory elements lead to exon skipping [37], disruption of exonic splicing enhancers [38], and use of cryptic splice sites [39]. Other splicing defects implicated in cancer may be due to alterations in the relative concentrations of splicing factors, leading to different ratios of alternative splice forms between tumours and their corresponding non-tumourous tissues. It is still unclear whether the change in splicing is a trigger for tumorigenesis or a symptom resulting from disease progression [40]. However, recent work by Kim et al. [41] have shown that a number of cancer-specific genes are involved in mRNA processing, suggesting that aberrant splicing patterns are a result of alterations in genes involved in the splicing machinery rather than being the result of disease progression.

A recent review by Solis et al. [42] focuses on human genetic diseases arising from mutations in splicing enhancer elements. These include spinal muscular at-

rophy, frontal-temporal dementia with parkinsonism linked to chromosome 17, and muscular dystrophy (See Fig. 1.4). In essence, mutations within enhancer elements for these genetic diseases result in exon skipping.

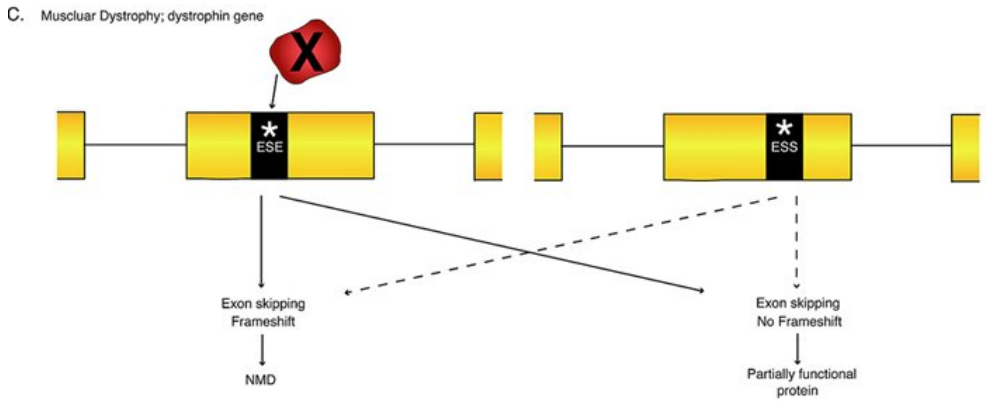


Figure 1.5: **Mutations in the dystrophin gene results in exon skipping.** Mutations in the dystrophin gene that causes muscular dystrophy (MD), which is an X-linked recessive disease. The disease is characterized by the degeneration of skeletal muscles due to the fact that dystrophin is a structural cytoskeletal protein which maintains membrane stability and communication between the extracellular matrix and the cytoskeleton. Two types of MD are known: Duchenne and Becker's with the former of the two being more severe. In Duchenne MD, a mutation affects the exonic enhancer element resulting in a frameshift, nonsense-mediated decay, and complete loss of dystrophin whereas in Becker's MD, a nonsense mutation creates an exonic silencer element that does not result in a frameshift and some dystrophin can be produced. Figure extracted from Solis et al. [42].

1.4 Computational approaches to study alternative splicing

To address alternative splicing using computational approaches requires one to understand the limitations of the computational tools and data sources available. This section aims to give an introduction to the relevant tools available and to explain why certain tools and data sources were utilized in the current study.

1.4.1 Tools used to detect alternatively spliced events

Two essential components that are pertinent for alternative splicing analysis are a robust algorithm for accurate splice-junction inferences and an automated splice-analysis pipeline to annotate and store the most up-to-date splicing variations when new genome or transcript data becomes available.

The purpose of splice alignment algorithms is to align transcripts to their respective genomes, which allows us to obtain locations of intron-exon boundaries necessary to detect alternative splice variation. The challenge for splice alignment algorithms is that transcripts such as ESTs, cDNAs, and mRNAs are never completely identical to the genome due to mutations and sequencing errors, thus complicating the detection of splice sites.

To date, several heuristic splice alignment programs such as sim4 [43], spidey [44], BLAT (BLAST-like alignment tool) [45], GMAP [46], SPA [47], and most recently PALMA [48] have been developed. Most of the splice alignment programs are similar in the sense that they start by using a local alignment search tool such as BLAST or BLAT to quickly find approximate matches between a query mRNA and target genomic sequence, and then filtering these matches for the best mRNA-genomic alignment. The programs vary in their approaches to identifying splice sites, parameters used during processing, speed and memory usage, sensitivity and specificity, as well as other specific functions such as microexon identification which has been incorporated in programs such as GMAP and more recently PALMA.

Despite the availability of other alignment programs, SPA (Spliced Alignment Algorithm) has been developed by our group for various reasons. One reason is that

SPA is capable of incorporating prior information to infer the best gene structure for a particular organism or dataset under study. SPA's scoring function consists of contributions from two models: i.) sequencing error model - in which we assume that the resulting sequenced cDNA is different from the original transcript due to sequencing errors. Thus, the aim is to compute how likely one observes the cDNA given the likelihoods of misincorporations, insertions, and deletions of bases of variable lengths that can arise during sequencing, and ii.) gene structure prior model - in which we compute how likely a given mapping is observed based on the prior probabilities of intron and exon lengths and the occurrences of canonical and noncanonical splice-site sequences. Thus, the unique feature of the scoring function of SPA is that it does not assume the same adhoc scoring parameters for all species and different sequencing technologies, but rather it adapts itself to the given situation by initially performing parameter estimations either directly on the input or on a subset of the dataset (See Fig. 1.6 for a schematic summary of SPA). These estimated parameters are then used to find the best mapping.

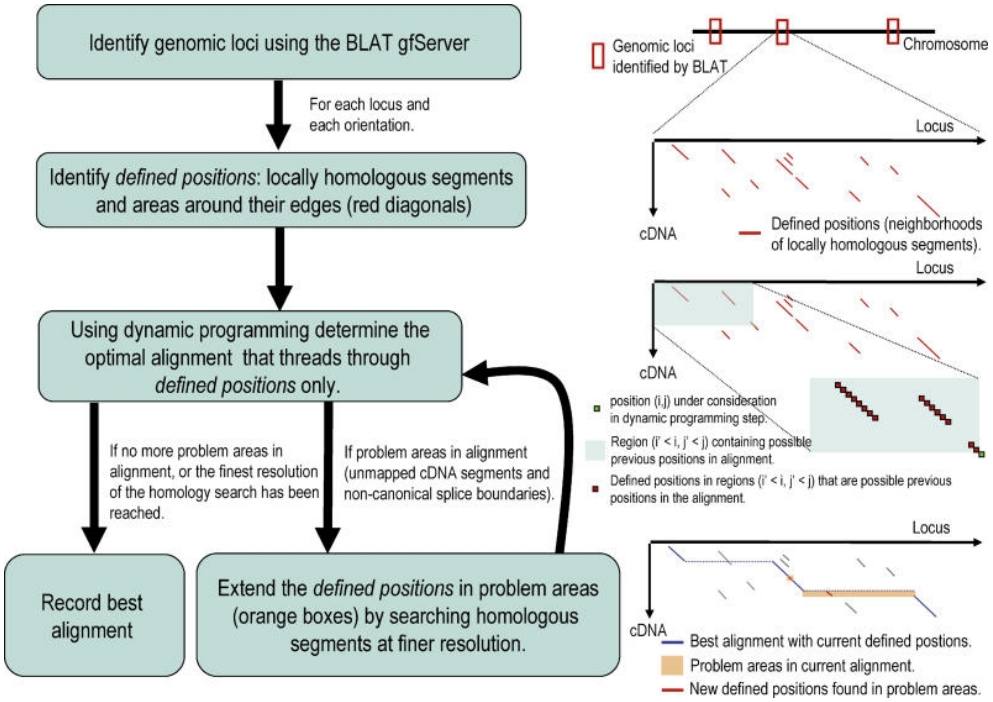


Figure 1.6: **A flow-chart of the SPA algorithm.** Figure extracted from van Nimwegen et al. [47].

Although, other alternative splicing and gene structure databases exist, it is still crucial to have one's own automated splicing pipeline - the primary reason being that one needs to have better control on how the alternative splicing data is generated as new transcript and genome data becomes available. The studies described here utilize our in-house pipeline, namely SPAED (SPlice Analyis of EST Data), which consists of the following steps: i.) Alignment of transcripts to the genome, ii.) Clustering of transcripts according to their orientation and genomic location, and iii.) Annotation of splice variation for all transcripts in each cluster. The splicing pipeline has been developed and utilized previously by Zavolan et al. [49] to estimate the frequencies of different patterns of alternative splicing in mouse full-length cDNA transcripts and to further examine the functional significance of alternative transcriptional start/stop sites giving rise to alternative initiation and termination exons. The current version of SPAED uses SPA to perform the alignments. A public website is also available at http://www.spaed.unibas.ch, where one can query our databases or view results of our analyses.

1.4.2 Sequence data sources used in computational studies of alternative splicing

Sequence data sources that are utilized in computational biology contain genome and transcript data. Genome assemblies are publicly available from several data sources such as the National Center of Biotechnology (NCBI) and the University of California Santa Cruz (UCSC). Genome data can be generated from three different types of sequencing techniques: 1.) clone-by-clone shotgun sequencing, 2.) whole-genome shotgun sequencing, and 3.) hybrid approach - a combination of the previous two techniques. The clone-by-clone shotgun sequencing technique involves the construction of a clone-based physical map and then selection of a minimal path of clones that covers a particular genomic region. A physical map is an ordered set of DNA fragments, which may be linked to several genetic markers in order to allow one to identify more clearly the gene's location. Then, the individual clones are subcloned into smaller insert libraries whereby thousands of sequence reads for each random subclone are derived. In whole-genome shotgun sequencing, instead

of using a physical map, clone libraries are prepared directly from the genome and millions of sequence reads are generated from these clone libraries. These are then assembled computationally into the genomic sequence.

The hybrid sequencing approach consist of generating reads from clone-by-clone and whole-genome shotgun sequencing. The advantage of this approach is that reads generated from the whole genome approach rapidly provide the entire genomic landscape whereas the clone-by-clone approach provides enhanced sequence coverage and thus can correct for any sequence misassemblies [50]

It is difficult to say which genome assembly is the better one since the methods by which they were generated are so different. However, Rouchka et al. [51] have previously compared the UCSC Goldenpath assembly against NCBI genome assembly by looking at 50 markers within a small region of genome and have observed that the NCBI assembly may be more accurate in determining the order of contigs whereas UCSC Goldenpath is more accurate in determining orientation of contigs. Nonetheless, this was a small scale study and the question of which genome assembly is more accurate remains to be investigated further.

The transcript data utilized in computational biology can be grouped according to their size such as full-length complementary DNAs (cDNAs), which are usually thousands of base pairs long and short sequence reads (200-800bp) such as expressed sequence tags (ESTs), which are used quite commonly. ESTs are generated from reverse-transcribed mRNAs and are commonly utilized to locate the promoters, poly-A tails and 3' UTRs, and to profile tissue expression. Since ESTs are generated from random single-pass sequencing, they can be highly error-prone and need to be subjected to EST pre-processing. Despite these errors, ESTs still remain a rich resource for gene analyses in organisms whose large genomes are too expensive to sequence [52]. Full-length cDNAs are also generated from reverse-transcribed mRNAs and they can be used to identify complete open reading frames and untranslated regions, which allow for research into mRNA stability and translational efficiency. One can obtain mouse and human cDNAs from the Fantom3 group [53] and the H-Invitational project [54]. The cDNA datasets are not always representative of transcript abundance in the cell, as there are various sources of errors

incurred during cDNA library generation, albeit there have been methods to circumvent many of these errors. Some of these errors include intron contamination as a result of inefficient RNA fractionation between cytoplasmic and nuclear mRNA species, RNA transcripts lacking or having shortened poly(A) tails which are selected against, competition between mRNA species of different sizes during clone transformation, selection against transcripts with strong RNA secondary structures. Certain cDNA libraries such as the bacteriophage cDNA libraries contain clones with different growth rates which can lead to the loss of clone representation. In any event, full-length cDNAs provide a better quality dataset than ESTs due to their wider coverage of the pre-mRNA, but it is still important to keep in mind the putative sources of error during data interpretation [55].

1.4.3 Genomic approaches to study alternative splicing

The genomic approach involves the use of genome and transcript data to examine alternative splicing. Several studies to date have utilized these data sources to uncover more information with regards to alternative splicing and they will be discussed in the following subsections.

Estimates of the frequency of alternative splicing

The frequency of alternative splicing in individual organisms such as human and mouse has been shown to be high and percentage estimations are as high as 74% when using microarrays [56]. Other studies focus on the estimation of the frequency of conserved alternative splicing between mouse and human. For example, a study from Sugnet et al. [57] estimated that 10% (1964/19156) of mouse and human alternative splicing events are conserved. Of these conserved events, exon skipping is most frequent (38.4%), followed by alternative acceptor (18.4%) and donor (7.9%) variations, with intron-retention and other splicing patterns coming last. Thanaraj et al. [58] have also shown that the frequency of human alternative splice donor and acceptor splice-site usage is quite high in mouse (61%).

Sequence features of alternatively spliced exons

Based on genomic studies, several features of alternatively spliced exons have been uncovered: 1.) higher intronic sequence conservation flanking alternatively spliced exons compared to constitutively spliced exons [57, 59, 60], 2.) shorter exon lengths relative to constitutively spliced exons [60], 3.) higher frequency of reading frame preservation amongst those alternatively spliced exons that are included in a minority of transcripts [Resch et al. 2004], 4.) weaker donor and acceptor splice-site strengths relative to constitutive exons [61, 60], 5.) higher density of regulatory sequences (purine-rich motifs that resemble enhancer elements) relative to constitutive exons [61], and 6.) higher frequency of premature termination codons relative to their constitutive transcripts [33]. These features have mostly been studied in human and mouse datasets.

Microarray approaches to detection of tissue-specific alternative splicing events

We define tissue-specific alternative splicing events as splicing events that occur specifically in a restricted number of tissues. The typical approach to detecting alternatively spliced events using microarrays is to design probes that span the splice junctions confirming alternative splicing events. The detection of exon-skipping events appears to be more common as the probes are more easier to design. A sketch of one approach to finding tissue-specific cassette exons using a 2-color cDNA microarray approach is illustrated in Fig. 1.7.

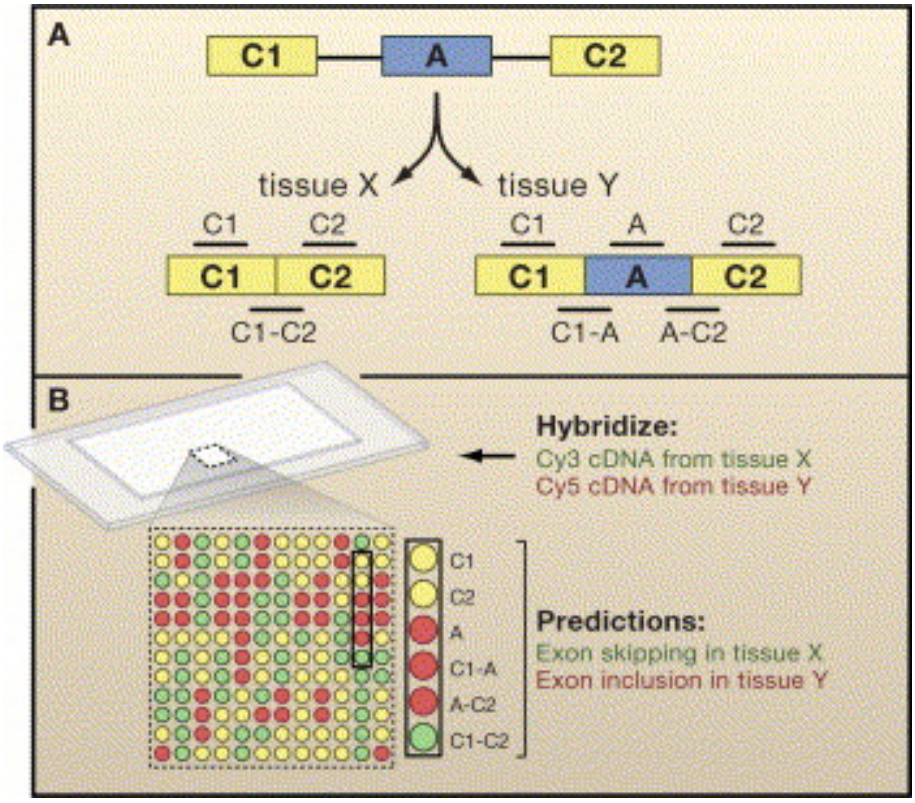


Figure 1.7: **Microarray approach towards detection of alternatively spliced tissue-specific events.** In this example, cassette exons are being detected in two different tissues, X and Y. The cDNAs from these tissues are labeled with green and red fluorescent dyes. Constitutive exons are labeled C1 and C2, whereas the cassette exon is annotated as exon A. In panel A, probes covering the splice junctions and every exon are shown. If exon A was missing in transcripts derived from tissue X, then we would not get signals from the probes: C1-A, A, A-C2. In panel B, a fictional hybridization pattern is obtained if we used Cy3-labeled tissue X cDNA (green) and Cy5-labeled tissue Y cDNA (red). The different colored spots indicate detection of signals from probes hybridized to the labeled cDNA from tissues X and Y. Data needs to be processed using an appropriate algorithm to predict the alternative splicing levels of alternative exons. Figure extracted from Blencowe,B. [62].

Microarray approaches towards detecting tissue-specific alternative splicing events have been used by Johnson et al. [56] to identify tissue-specific alternative splicing events in human as well as by Sugnet et al. who identified cassette exons that are differentially expressed between brain/muscle and non-brain/non-muscle tissues [63]. In addition, Clark et al. [64] have used human exon arrays to profile with probes covering only the individual exons in the human genome in 16 adult normal tissues. By comparing gene-level-normalized and exon-specific expression values, they were able to detect tissue-specific alternatively spliced exons.

It seems that many alternative splice forms are tissue specific. The frequency of alternative splicing events appears to be highest in brain, testis, and the immune system [65, 66], thus suggesting a more complicated and diverse regulation within these tissues.

1.4.4 Structural insights from proteomic approaches

The proteomic approach to addressing questions about alternative splicing essentially boils down to how the splicing event affects the resulting properties of proteins. Properties such as protein folding, protein stability, and disruption of protein domains can be affected by alternative splicing events. We need to consider several key issues when performing such analyses: i.) finding enough protein 3D structures to perform the analysis, ii.) correctly mapping the alternative splicing event to the correct region within the 3D structure, and iii.) accurate assessment of the effect of the alternatively spliced event on the protein function. Although the number of protein 3D structures in Protein Data Bank (PDB) is still limiting when compared to the number of transcript sequences in public databases, several studies have begun to investigate the functionality of alternative splicing using 3D protein structures and we will discuss their results shortly.

Earlier studies that evaluated the effect of alternative splicing on protein 3D structures reported mixed results: one study [67] observed that small splice variations (<50 nucleotides) are likely to influence the tertiary structure of the proteins. Another study found that the majority of human alternatively spliced junctions occur in protein domain boundaries with an insignificant tendency to avoid being

located inside protein domains [68]. Several other studies [69, 54, 70, 71] with larger datasets found that alternative splicing events affect protein domains that are involved in signal subcellular localization and protein-interaction binding sites. Finally some studies suggested that alternatively spliced regions may be located to avoid structural implications: i.) Romero et al. have reported that alternatively spliced events tend to occur in protein regions that are intrinsically disordered [72], ii.) Alternatively spliced junctions tend to avoid annotated protein domains and have strong preferences for loop and exposed regions in protein secondary and tertiary structures respectively [73, 74].

Recently, Tress et al. [69] raised an interesting, yet debatable viewpoint that although many alternative splicing events, that may be deleterious, can greatly affect the structure of a protein, there is not much evidence in literature to prove that alternative splicing gives rise to much protein functional diversity. The author has suggested that alternative splicing may have another role other than to expand the protein functional potential and that perhaps the organism is capable of tolerating a high number of abberant protein isoforms. The mechanisms of such tolerance have not been elucidated. Given this observation, the picture is likely to be clearer as more protein structures become available.

FireDB and *Firestar* - useful tools for assessing functionally important protein residues

FireDB [75] is a database containing functionally important residues from proteins of known structure. The database consists of all PDB sequences in which only those sequences with >97% identity are grouped into one cluster and consensus sequences are generated for each cluster through multiple protein sequence alignments. In addition, each consensus sequence also contains residue reliability information, which is a measure on how conserved and reliably aligned a particular amino acid is within a multiple sequence alignment. The reliability score calculations are done using SQUARE [76], which is a web-based version of the method developed by Tress et al. [77]. Their method assesses the quality of alignments between unknown query and template sequences of known structure and the aligned positions are

scored using profiles generated from template sequences. The alignment scores are then smoothed using a triangular five-residue window and from these smoothed scores, reliably aligned regions are then determined. The consensus sequence for each cluster is annotated with functional binding-site information obtained from Catalytic Site Atlas (CSA), the literature, databases containing protein-ligand interaction annotations, and conserved binding sites from other proteins.

To exploit the potential of FireDB, the same authors have also developed another tool namely *Firestar*, which predicts ligand-binding residues in protein query sequences. *Firestar* [78] takes in a protein query sequence as input and performs a pairwise PSI-BLAST against the PDB, then a final search is made to the FireDB consensus database. As mentioned before, binding-site annotations are associated with each consensus sequence and residue reliability scores are also calculated between the query and consensus sequence using SQUARE. However, these reliability scores are calculated from PSI-BLAST profiles, which are pre-generated for each consensus sequence (profiles are generated from the nonredundant database (nrdb90) from the European Bioinformatics Institute). Thus, depending on how good the reliability scores are, we can transfer the functional binding-site annotations from the consensus sequences in FireDB to our query sequence.

In chapter 2, we have utilized FireDB and *Firestar*, to assess the functional impact of single amino acid insertions/deletions arising from the usage of short tandem acceptor splice sites.

1.5 Evolutionary insights into alternative splicing from computational analyses

This section addresses how mammalian splice sites and other splicing regulatory features have evolved and provides some insights into the mechanism of alternative splicing from an evolutionary standpoint. Most of the points addressed here have already been covered by the review written by Xing and Lee [79], but here we offer a brief summary.

1.5.1 Orthologous exons and their inclusion levels

Conserved alternatively spliced cassette exons between mouse and human represent approximately 10-20% of all cassette exons, thus suggesting that there is much species-specific alternative splicing to be explored [80, 81, 82]. Modrek and Lee [80] observed a linear correlation between the conservation of mouse and human orthologous exons and their inclusion levels: highly conserved orthologous exons have high inclusion levels and vice versa.

1.5.2 Strength of splice-site signals

Recently, Schwarz et al. [83] have performed comparative analyses (from fungi to mammalians) to study the evolution of splicing signals. The authors have found that the strength of the polypyrimidine tract has slowly increased from fungi to metazoans. In addition, the 5'ss and branch sites were most degenerate in the majority of the organisms examined with the exception of fungi. Lastly, their analyses revealed that the strength of the polypyrimdine tract is correlated with changes in residues in their corresponding splicing factors, indicating that the splicing regulatory signals are evolving simultaneously with their corresponding splicing factors.

1.5.3 Is organismal complexity correlated with increasing alternative splicing?

A few studies have attempted to examine rates of alternative splicing between vertebrates and invertebrates and have yielded mixed results. According to the work performed by Brett et.al [84], alternative splicing occurs at comparable frequencies amongst vertebrates and invertebrates whereas other studies have reported higher frequencies in vertebrates compared to invertebrates [79, 85].

1.5.4 Correlations between alternative splicing and tandem exon duplications

Letunic et al. [86] have estimated that about 10% of all genes in human, worm and fly contain conserved duplicated exons. They took a subset of those annotated exons with duplications and estimated that between $59-62\%$ of these exons are subjected to mutually exclusive splicing, a type of alternative splicing, in which alternative exons are not observed in the same transcript. Interestingly, the authors have identified a duplicated exon within the human glycine receptor alpha-2 gene, which was also known to be a candidate for alternative splicing. With cDNA evidence, the authors confirmed that the duplicated exons were mutually exclusively spliced into resulting gene products. By mapping the duplicated exon to their relevant protein structure, the authors found that the substitution between the two duplicated exons might affect the function of the receptor.

1.5.5 Correlations between alternative splicing and Alu transposable elements

The majority of Alu transposable elements are found in approximately 4% of human protein-coding regions. The Alu elements contain potential splice donor and acceptor splice sites and new exons are created upon their insertion into the intron. According to Sorek and colleagues, most Alu-containing exons are alternatively spliced [87]. One possible exon-creation mechanism proposed by Lev-Maor et al. [88], is that it only takes one mutation for intronic Alu elements to convert the intronic sequence to an exon. How transposable elements may play a role in the alternative splicing of protein-coding genes is not completely clear, but a few examples such as the Bim-beta3 [89], suggest that introduction of an Alu element into the coding region of genes can result in the creation of alternative translational start sites or the generation of new Alu-encoded exons.

1.5.6 Major and minor exons

Modrek and Lee [80] classified alternative exons into 'major' and 'minor' form exons, whereby the major form is defined as the alternative transcript confirmed by more transcript copies than the minor form. Major form exons are more conserved between mouse and human than minor form exons and they are also more similar to constitutive exons. Minor form exons, being less conserved, appear to have been created recently [90, 91]. In addition, the flanking introns of minor-form exons are more conserved between mouse and human than constitutive exons suggesting their functional importance [92]. Xing and Lee [79] suggested that the purpose of introducing minor-form exons through alternative splicing is to allow the newly created alternative exon to experience little negative selection pressure and thus, permit evolutionary changes to occur in genes.

Chapter 2

Investigation of short tandem acceptor splice sites

2.1 Introduction/Rationale

In the literature review (Chapter 1), we discussed several functional, biological consequences of alternative splicing. However, alternative splicing may be due simply to the noise in the splicing process. The idea of spurious splice variation as a result of stochastic choice of splice sites was first suggested by Kan et al. [93] and Zavolan et al. [49]. Kan and others have observed that many alternatively spliced events detected in ESTs could not be confirmed in other sequence data or through conservation and that theoretically these spurious events might exist in the cell, but occur at lower frequencies. In support of this view, Zavolan and others have also observed frequent small variations (less than 10 nucleotides) at donor and acceptor splice sites, thus suggesting that the spliceosome may slip near acceptor splice-site boundaries.

In line with the interest in small exon-length variations occurring at splice-site boundaries, another observation by Hiller et al. [94] reported the frequent occurrence (up to 30% in human genes) of short tandem, acceptor splice sites, termed 'NAGNAGs'. These acceptor splice sites are adjacent to each other and selection of one of the acceptor splice sites will lead to either the inclusion of one amino acid (if the first acceptor splice site is chosen) or deletion of a single amino acid (if the sec-

ond site is selected). The same authors have proposed that these NAGNAGs have the potential to increase the protein functional diversity through the inclusion and exclusion of one amino acid. Although there are only a few experimentally-validated functional NAGNAG acceptor splice sites in literature, the presumably first, functional NAGNAG case did make its appearance 14 years ago, when Condorelli and others [95] have found distinct biological differences of NAGNAG isoforms of the human insulin-like growth factor I receptor. The interest in NAGNAG acceptor sites increased slightly after Hiller et al.'s study highlighted its prevalence in the human genome [94]. Thereafter, other research efforts have begun to focus on single nucleotide polymorphisms (SNPs) in NAGNAGs [96] and on finding tissue-specific differences between the included and excluded NAG forms [97]. NAGNAG events have also been detected in plants. Recently, Iida et al. [98] have detected between 300-400 NAGNAG events in Arabidopsis and rice. The precise mechanism by which these tandem alternative acceptor splice sites are chosen by the splicing machinery is not well understood. However, recent mutation experiments by Tsai and others [99] suggest that the region from the branchpoint sequence to the NAGNAG splice sites is important for splice-site selection.

We have explored, in more depth, the frequencies of small exon-length variations, the evolutionary and functional characterization of NAGNAG tandem acceptor splice-site motifs, and the spliceosomal binding affinities of NAGNAG tandem acceptor splice sites computed from our weight matrix (WM) models. The work has revealed that the binding of the spliceosome at NAGNAG tandem acceptor splice sites can be competitive, which would also allow possible stochastic binding at these sites confirming earlier studies on the possibility of noise in the splicing process. In support of this view, we did not have enough evidence from our evolutionary and functional characterization of NAGNAG motifs at tandem acceptor splice sites to suggest that the majority of these single amino acid insertions and deletions can increase the functional protein repertoire.

2.2 Methods and Results

The work performed in sections 2.2.1 - 2.2.9 has already been published in the PLoS Genetics article entitled 'A simple physical model predicts small exon-length variations' and a copy is included in Chapter 5. Please refer to the Materials and Methods section in the paper for a detail description of sections 2.2.1 - 2.2.9. Other methods not present in the paper will be described in the relevant sections.

2.2.1 Abundance of small exon-length variations at acceptor splice sites

We started our analyses in mouse by first obtaining datasets that are annotated with alternative acceptor and donor splice-site variation. In brief, we did this by mapping mouse full-length cDNA transcripts to their genome using SPA (See literature review) and using our in-house splicing pipeline (SPAED) to annotate the splice variation. From our initial dataset of acceptor and donor sites, we have subdivided them into three categories: 1.) those sequences that have coding sequence annotation (CDS), but have their site-of-variation fall within the CDS, 2.) those sequences that have coding sequence annotation, but have their site-of-variation fall outside of the CDS, which we term as the 'UTR' group, and 3.) those sequences that do not have any coding sequence annotations and come from transcription units in which none of the transcripts had a CDS annotation, which we call the 'noncoding exons'. We totalled the number of unique exon-length variations at donor and acceptor splice sites and have observed that the total number of acceptor variations (2295) was greater than the donor variations (1689). Please see Fig. 2.1 on the frequency of exon-length variations for each exon-type category.

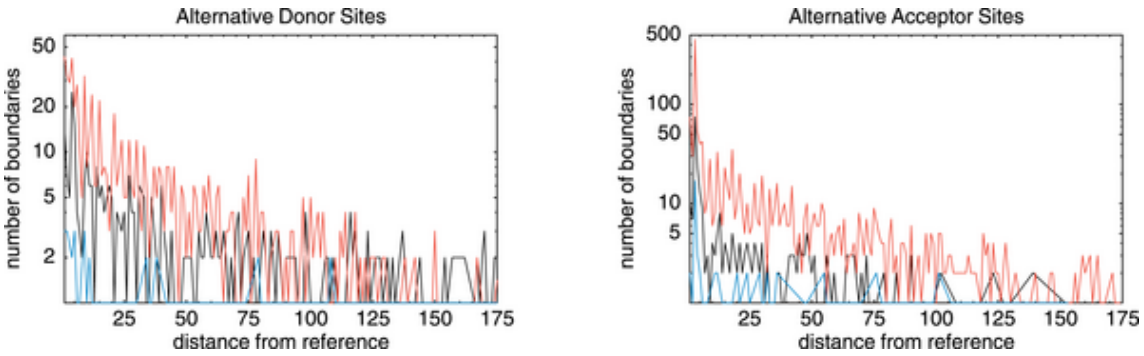


Figure 2.1: **Frequency of length variations of different sizes at donor(left) and acceptor(right) splice sites.** The figures show the observed counts of splice events occurring at different distances from the most commonly used splice site in transcripts originating from the same gene, which we termed as reference boundary. The red, black, and blue distributions correspond to alternative exons belonging to the CDS-, UTR-, and noncoding exons respectively. The figures highlight that variations, which are multiples of three are common in CDS exons at both donor and acceptor splice sites, whereas this is not the case for splice sites of noncoding exons. Figure taken from Chern et al. [100].

In addition, we have also tallied those exon-length variations, in which the exon-length difference was less than 10 nucleotides, for donor and acceptor splice sites, and have observed once again that small acceptor variations (955) were larger than their donor variations (348). Interestingly, small acceptor splice-site variations constitute 42% of all acceptor splice-site variations versus only 21% was observed for small donor splice-site variations.

We have decided to examine which small exon-length variation contributes most in small acceptor and donor splice-site variations, and have found that for small acceptor splice-site variations, those that cause an exon length difference of 3 nucleotides, constitute 58% (554/955) of all small acceptor splice-site variations. In contrast, among small donor splice-site variations, those of one and four nucleotides contributed the most (18.4% and 20.1% respectively) to all small donor variations.

2.2.2 Abundance of in-frame exon-length variations at acceptor splice-site boundaries

Exon-length variations that are not a multiple of three most likely cause frame-shifts that lead to transcript degradation by NMD. Thus, we computed the frequencies of all multiple of three (in-frame) variations in both donor and acceptor splice sites and in all exon-types (CDS, UTR, and noncoding exons). In addition, we compared the frequencies to those expected by chance, which is 1/3. We have observed that the frequency of in-frame variations for acceptor splice-site boundaries in all exon types exceeded 1/3, whereas for donor splice-site boundaries, only the frequency in CDS exons was greater than expected by chance (See Fig.2.2).

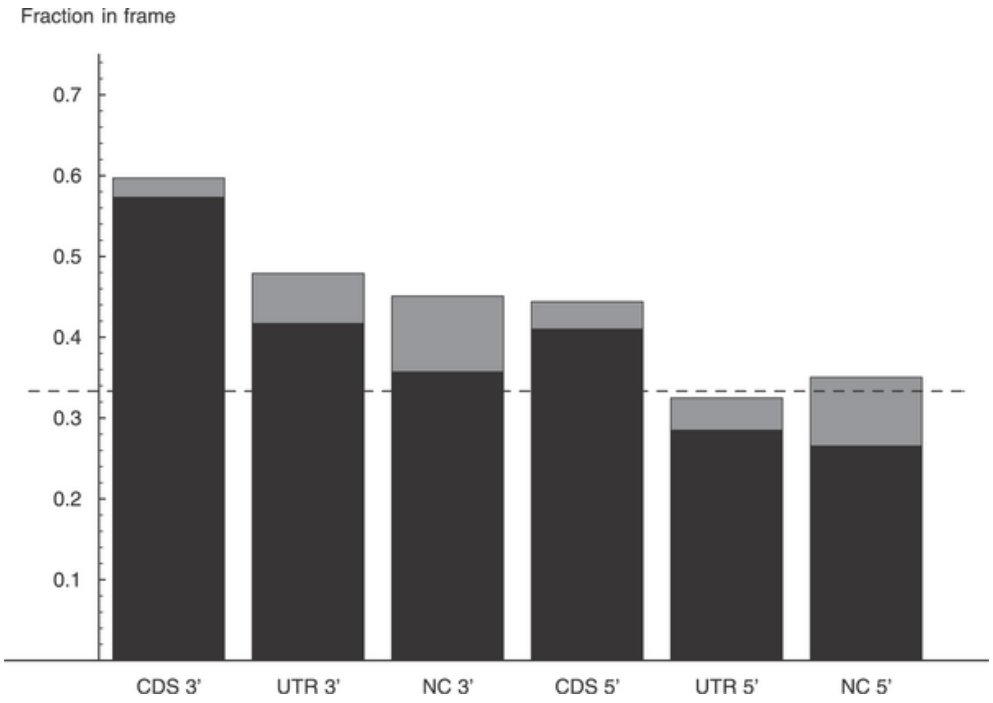


Figure 2.2: **Proportion of in-frame alternative acceptor and donor splice-site events.** The histogram shows alternative acceptor (3') and donor (5') splice-site events of CDS, UTR, and noncoding (NC) exons, in which the splicing event leads to in-frame shifts with respect to their reference boundaries. The estimated fraction is in the middle of the gray bar, with the width of the gray bar indicating two standard errors. The horizontal dashed line represents the fraction 1/3, that would be expected by chance. Figure taken from Chern et al. [100].

We will explain shortly that the differences in the proportion of in-frame variations between donor and acceptor splice sites is due to the differences in their sequence composition within the first few bases of donor and acceptor splice sites of small exon-length variations that causes nucleotide shifts of 1-4 bases.

2.2.3 Overrepresentation of in-frame exon-length variations that causes nucleotide shifts of more than 4 bases in coding sequence exons

At a closer examination, we observed that for exon-length variations of more than 4 nucleotides, the patterns of variation are similar between donor and acceptor sites: only variations in CDS exons are enriched in multiples of 3 (See Fig. 2.3). This suggests that a common mechanism that is sensitive to the reading frame such as NMD is responsible for the observed enrichment.

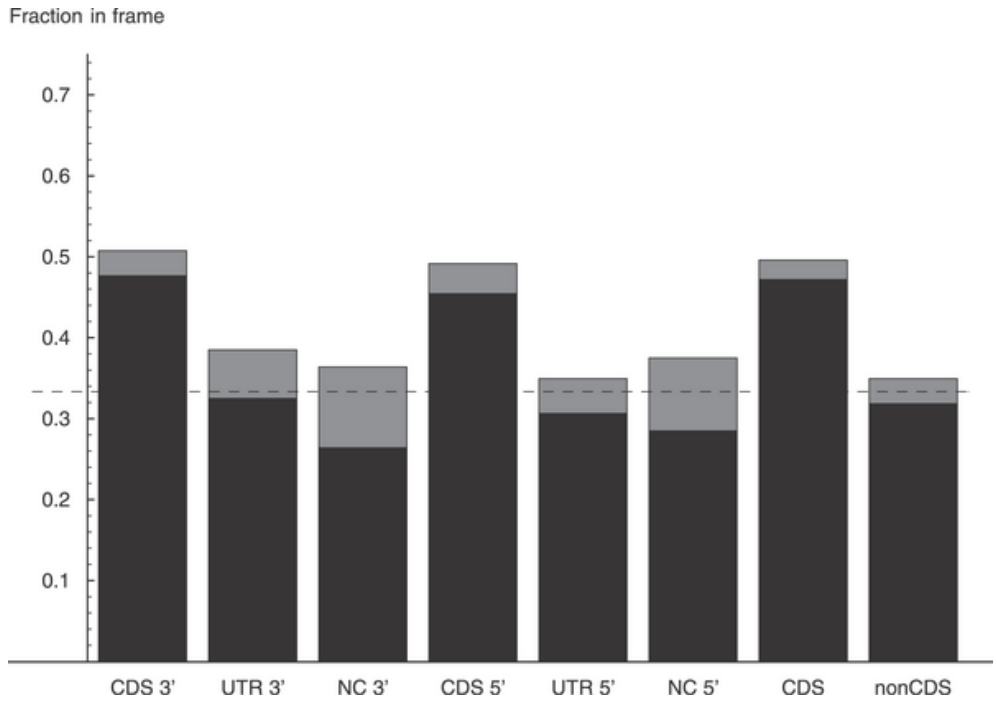


Figure 2.3: **Proportion of in-frame alternative acceptor and donor splice-site events that causes exon-length differences greater than 4 nucleotides.** The histogram shows alternative acceptor (3') and donor (5') splice site events of CDS, UTR, and noncoding (NC) exons, in which the splicing event leads to in-frame shifts with respect to their reference boundaries. The estimated fraction is in the middle of the gray bar, with the width of the gray bar indicating two standard errors. The horizontal dashed line represents the fraction (1/3), that would be expected by chance. The last two bars on the right represent the data from all CDS and non-CDS exons pooled. Figure taken from Chern et al. [100].

2.2.4 The proportion of in-frame putative donor and acceptor splice sites is close to the proportion expected by chance

We tried to explain the excess of in-frame variations at CDS regions in donor and acceptor groups in terms of the sequence bias at different relative positions with respect to the reference splice boundary. For this, we extracted the intronic sequences of 100 nucleotides flanking all alternative acceptor and donor splice sites and counted the number of times an intronic AG or GT occurs upstream or downstream respectively of their corresponding reference splice-site boundary. The occurrences of intronic, upstream AGs or intronic, downstream GTs within the four nucleotides of the reference boundary were not counted. We have found that the fraction of in-frame occurrence of the putative splice-site dinucleotides is close to 1/3 for all exon types, thus suggesting that there is no bias in the sequence composition flanking CDS exons of donor and acceptor groups (See Fig. 2.4) and we therefore set to investigate whether NMD might be responsible for the excess of in-frame variations at CDS exons.

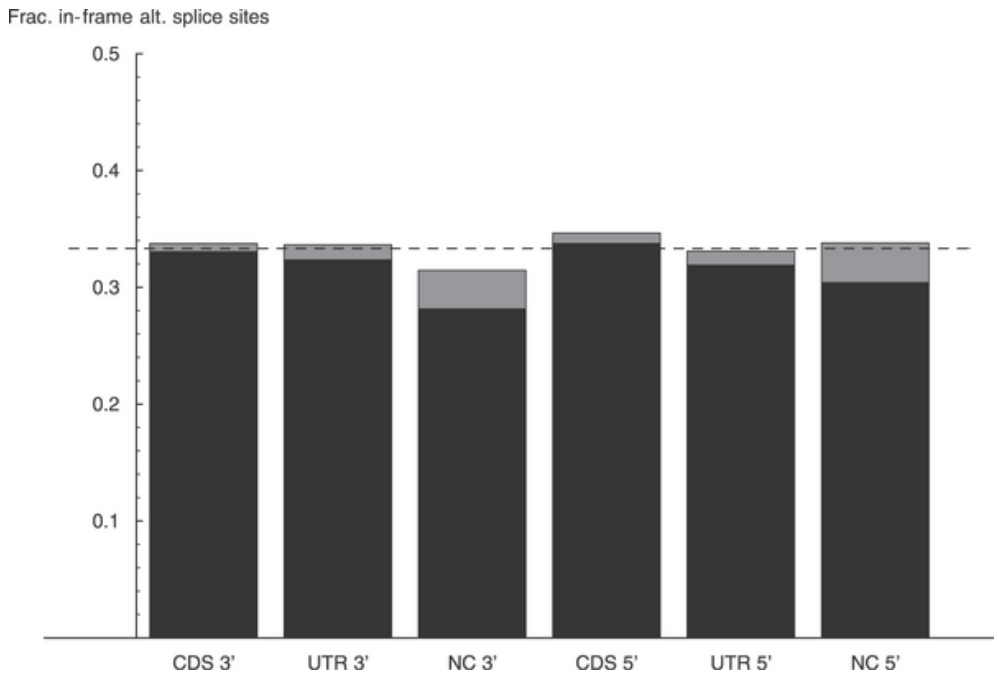


Figure 2.4: **Proportion of putative donor and acceptor splice sites that are situated in-frame relative to the splice sites in CDS, UTR, and noncoding regions.** The histogram shows the counts of AG (3',acceptor splice site) and GT (5',donor splice site) dinucleotides appearing within 100 intronic bases upstream and downstream respectively from their splice sites. AG and GT occurrences within the first 4 nucleotides were excluded from the counts. The estimated fraction is in the middle of the gray bar, with the width of the gray bar indicating two standard errors. The dashed line indicates the fraction (1/3) that would be expected by chance. Figure taken from Chern et al. [100].

2.2.5 Estimation of the fraction of frameshifting exon-length variations that survive NMD

It is well known that transcripts containing premature stop codons are subjected to NMD [32]. We reasoned that it is likely that the observed enrichment of in-frame variations in CDS exons is due to NMD removing some of the out-of-frame variants, and we set to estimate the fraction of out-of-frame variations that survive NMD. If we assume our fraction of in-frame variations at CDS exons before NMD is 1/3, which we term as p_i, and that a proportion of out-of-frame variants survive NMD, then the observed proportion of in-frame variants that we observe after NMD is given by:

$$p_o = \frac{p_i}{p_i + f(1 - p_i)} \tag{2.1}$$

Solving this equation for f with $p_i = 1/3$ and $p_o = 0.48$, we obtain f=0.53. This suggest that only 47% of out-of-frame variations are removed by NMD. We then used this fraction (f=0.53) to estimate the proportion of exon-length variations of 1-4 nucleotides before NMD.

So far we have shown that in-frame variations were more frequent at acceptor splice sites than at donor splice sites (See Fig. 2.2), yet when we focus our attention to only those exon-length variations of more than 4 nucleotides, we found that the proportion of in-frame variations for donor and acceptor splice sites were quite similar (See Fig. 2.3). The scenario is different for very small exon-length variations of one to four nucleotides. As shown in Fig. 2.5 (A and C), we observe more in-frame variations of precisely three nucleotides at acceptor splice sites than at donor splice sites, where frameshifting variations of 1 and 4 abound. In previous sections of this chapter, we argued that the excess of in-frame variations at CDS exons is due to NMD and now we set to demonstrate that very small exon-length variations (1-4nt) can be explained by a combination of stochastic binding of the spliceosome at competing splice sites and NMD.

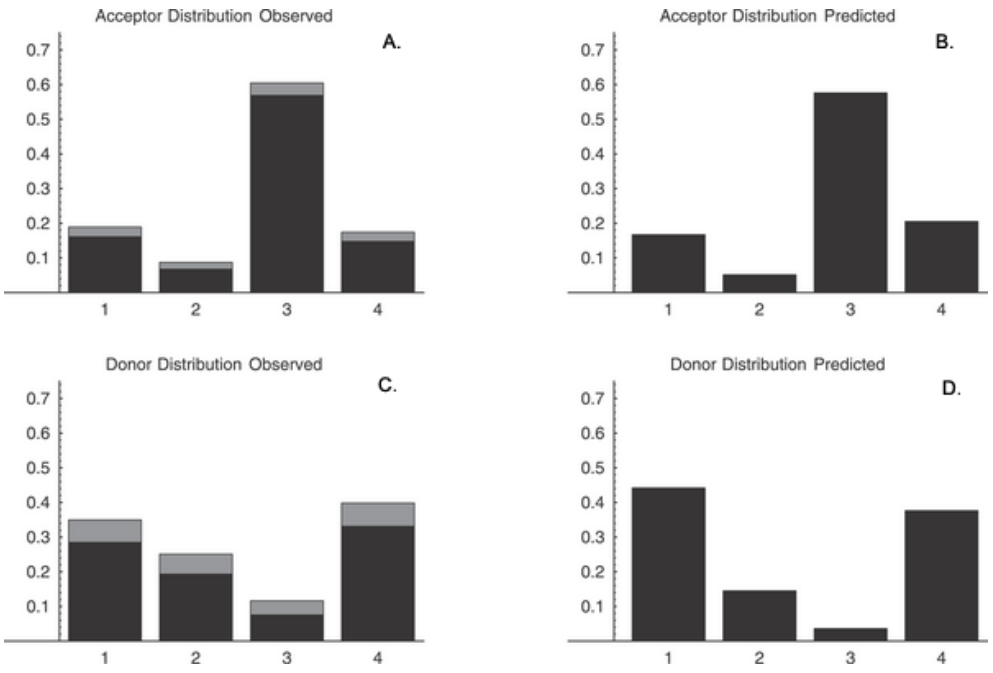


Figure 2.5: **Distributions of very small exon-length variations occurring at acceptor and donor splice sites that cause exon-length differences of one to four nucleotides.** Histograms on the left column indicate observed acceptor (top) and donor (bottom) distributions of small exon-length variations before NMD. The fractions of all exon types (CDS, UTR, and noncoding) were pooled since their frequencies were similar. The frequency of variations of length 3 was rescaled to take into account the effects of NMD. On the right column, predicted acceptor and donor distributions are shown from our weight matrices, which we will explain shortly. The estimated relative frequency is in the middle of the gray bar, with the width of the gray bar corresponding to two standard errors. Figure taken from Chern et al. [100].

2.2.6 A high frequency of NAGNAG motifs was found in alternative acceptor splice sites that causes an exon length difference of 3 nucleotides

Earlier, we have mentioned that small acceptor variations that cause an exon length difference of 3 nucleotides, contribute towards 58% of all small acceptor variations (<10 nucleotides). Now, we focus our attention to these three-nucleotide acceptor variations and we divide this dataset into 3 categories: those transcripts that contain our site-of-variation in the coding sequence and have coding sequence annotation (CDS), those transcripts that contain our site-of-variation in the untranslated region, but also have coding sequence annotation (UTR), and those transcripts that do not have coding sequence annotations (NC) and come from transcription units in which none of the transcripts have coding sequence annotations. Within each of these groups, we have computed the frequency of NAGNAG motif occurrence, computed the most frequent NAGNAG motif(s) within that group, and calculated the NAGNAG motif percentage conservation to its human ortholog (See Table 2.1 below). We have defined sequence conservation as perfect conservation of the nucleotides within the NAGNAG motif. We find that 95% of all three-nucleotide acceptor variations in CDS group contain NAGNAG motifs, followed by over 50% of NAGNAG-containing splice boundaries in other groups. The most frequent motif observed in each group was CAGCAG and the percentage conservation was highest in the CDS group.

Table 2.1: **Frequency of NAGNAG motifs in alternative acceptor splice-site boundaries causing an exon length difference of 3 nucleotides**

Categories	Frequency of NAGNAG motifs	Most frequent NAGNAG motif	Conservation of NAGNAG motif
CDS	499/526=94.9%	CAGCAG=223/499 (44.7%) and TAGCAG=97/499(19.4%)	213/499=42.7%
UTR	7/11=63.6%	CAGCAG=5/7 (71.4%)	0
noncoding(NC)	15/17=88.2%	CAGCAG=5/15 (33.3%)	3/15=20%

It is difficult to imagine how the spliceosome can choose in a regulated manner a splice site among variants that are 1-4 nucleotides apart with such high precision, we reasoned that a likely factor determining splice-site selection is the strength of the alternative splice sites. If one of the alternative acceptor splice sites is stronger than the other, then one expects that the stronger acceptor would be selected by the spliceosome. To test this, we have measured the relative strength of alternative splice sites (NAGNAG tandem acceptor splice sites) as follows: We extracted the intronic and exonic sequences spanning the splice boundaries of variant and invariant exons. We constructed a weight matrix (WM) based on invariant splice sites and used it to estimate the strength of NAGNAG-containing alternative acceptor and invariant acceptor splice sites.

The reason for using a WM constructed from invariant splice sites to score tandem acceptor sites is that for invariant splice-site sequences, the spliceosome will efficiently and reliably recognize the stronger splice site (See Fig. 2.6). The likelihood of a candidate splice site is given by the product of the base probabilities at each position in the splice site as given by the WM. We obtained log-likelihood scores for all variant and invariant acceptor splice sites containing NAGNAG motifs. We have also constructed a donor site WM and have used this WM to predict the relative frequencies of small exon-length variations (See section 2.2.9).

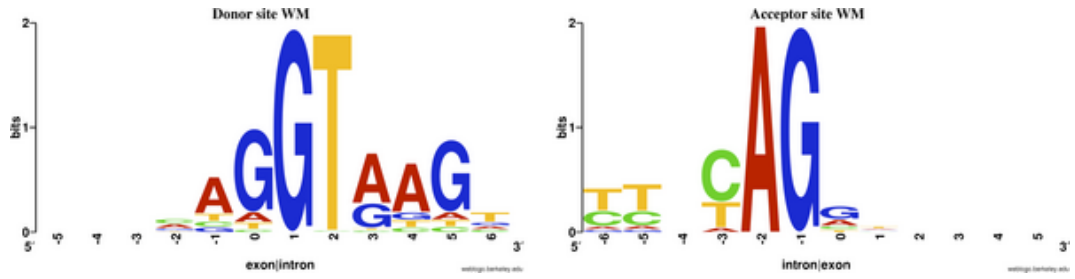


Figure 2.6: **Sequence composition of invariant donor and acceptor splice sites.** The weblogos illustrate the weight matrices of donor (left) and acceptor (right) splice sites. The weight matrices are constructed from six intronic and six exonic nucleotides spanning the splice boundary. The relative size of the letters reflect the relative frequency of each base at each position. However, the total height of each column represents the information content in that column, expressed in bits. Figure taken from Chern et al. [100].

We then sought to determine whether the relative likelihoods of the two putative sites in NAGNAG acceptors can explain the relative frequency with which the two sites are used in splicing reactions. The results are shown in Fig. 2.7. In blue, we show the distribution of log-likelihood differences of variant NAGNAG sites (both sites have been used in the splicing reaction), in red we show the log-likelihood differences of invariant NAGNAG sites (in which the first site is used), and the green distribution shows that log-likelihood differences of invariant NAGNAG sites (in which the second NAG site is used). The figure indicates that, as expected, when the likelihoods of the competing acceptor sites are similar, they are both used in splicing reactions, whereas when one or the other of the competing tandem acceptor site has a much higher likelihood than the other, then only the site with the higher likelihood is used in the splicing reaction. Thus, based on their log-likelihood differences, we were able to distinguish between variant and invariant NAGNAG splice sites. In addition, the fact that the likelihoods of variant competing acceptor sites are similar suggests that the spliceosome may choose one of the NAG sites in a stochastic manner.

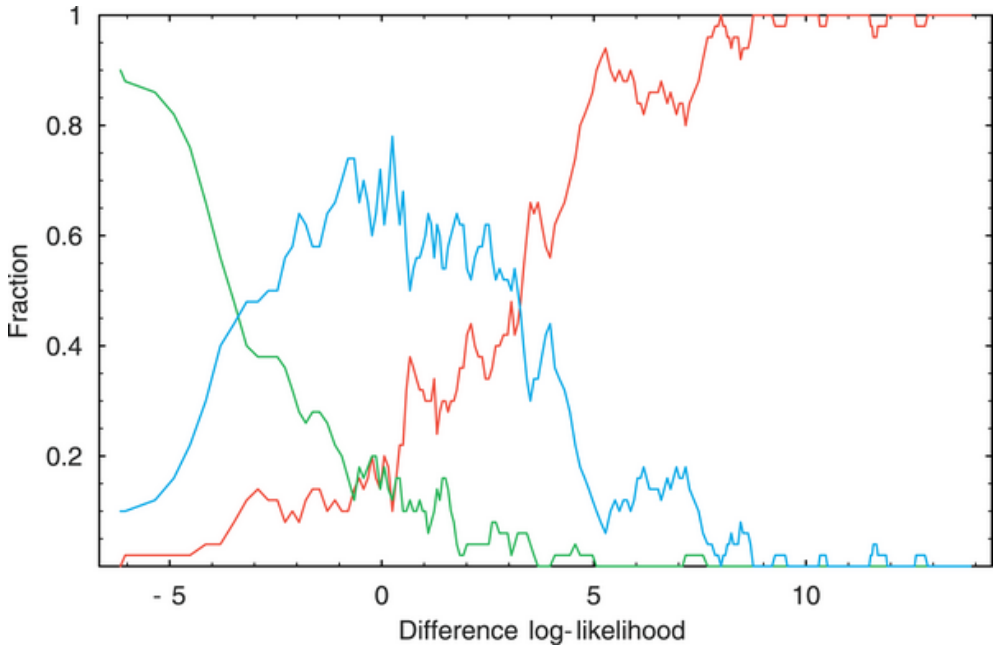


Figure 2.7: **Dependency of the frequency of alternative splicing at NAGNAG sites on the relative likelihood of the two putative acceptor sites.** The relative likelihood is calculated by taking the difference between the log likelihood of the first and second NAG using our WM constructed from invariant splice sites. The red, green, and blue distributions correspond to tandem acceptor splice sites containing NAGNAG motifs in which the spliceosome splices only at the first NAG (red), only at the second NAG (green), and splices at both NAGs (blue). Figure taken from Chern et al. [100].

2.2.7 NAGNAG motifs are also frequent and even more conserved at invariant acceptor sites

In order to infer the functional importance of NAGNAG motifs, we have also checked whether the high NAGNAG motif frequency is only observed at alternative acceptor splice sites. To address this question, we have obtained 130689 invariant, constitutive exons that do not show any alternative splicing at their acceptor splice sites. We have searched for NAGNAG motifs in two distinct regions surrounding the constitutive splice boundary for each exon - (i) six intronic nucleotides upstream of the splice boundary, and (ii) six nucleotides spanning the splice boundary (3 intronic nucleotides upstream and 3 exonic nucleotides downstream from the splice boundary) were extracted. As a control, we have also extracted all possible hexameric motifs spanning the invariant constitutive splice boundary.

We find that the majority of the NAGNAG motifs occurs across the splice boundary in invariant exons (6052/130689=4.6%), although these motifs were also found in the intronic portion upstream of the splice boundary (663/130689=0.37%). The most common NAGNAG motifs observed for invariant exons were CAGGAG (45.6% occurring across the splice boundary) and GAGCAG (34.8% six intronic nucleotides occurring upstream of the splice boundary).

We then compared the conservation levels of the NAGNAG motif observed in all groups of invariant exons and have found that NAGNAG motifs occurring across the splice boundaries of invariant exons were more conserved (2897/6052=47.9%) than those motifs extracted only from the intronic portion (183/663=27.6%). Interestingly, if we compare the proportions of those NAGNAGs occurring in variant CDS exons against the invariant NAGNAGs occurring across the splice boundary, the conservation is higher for NAGNAGs in invariant exons (p-value =0.03) than the NAGNAGs occurring in variant CDS exons. In addition, if we compare the conservation levels of any hexameric motif across the splice boundary of invariant exons (63001/126817=49.7%) to the NAGNAG conservation levels of variant CDS exons, we find that the conservation levels of any hexameric motif across invariant splice boundaries is significantly higher than the NAGNAG conservation levels for

variant exons (p-value=0.002). These results suggest that NAGNAG motifs at variant splice boundaries are not under stronger selection pressure than any other motif occurring at the splice boundary.

2.2.8 Frequencies of NAGNAG motifs at the splice boundaries of noncoding exons are significantly higher than at the splice boundaries of coding exons

To check whether the presence of NAGNAG motifs is preferred in CDS exons, we have computed the frequencies of NAGNAG motifs occurring at all coding exons and noncoding exons. For noncoding exons, we pooled the frequencies of NAGNAG motifs occurring in both UTR and noncoding groups. We have found the NAGNAG frequency in noncoding exons to be 6.2% (958/15379), which is significantly higher than the 5.4% (6239/115086) observed in coding exons (p-value=$4.1e^{-05}$). These results indicate that NAGNAG motifs are not more frequent in regions in which alternative splicing would result in higher proteome diversity.

2.2.9 Prediction of relative frequencies of observed small exon length variations of $1-4$ nucleotides from weight matrices are accurate

In addition to using our WMs to predict the binding affinities of the spliceosome on alternative acceptor splice sites, we used our donor and acceptor WMs to estimate the relative frequencies of small (1-4nt) exon-length variations. To do this, we summed over all splice boundaries the probability that a site is located at 1,2,3,4 nucleotides from the reference site that is used in the splicing reaction (according to its likelihood given by the WM). In order to compare our predicted relative frequencies to observed frequencies of splicing at small exon-length variations, we have computed the observed frequencies of small exon-length variations "before NMD" because we only want to see the effects of splicing and not the effects of splicing and NMD. Fig. 2.5 (B and D) shows that the predicted acceptor and donor distribu-

tions are quite similar to those distributions observed before NMD (Fig. 2.5 A and C), except for the donor distributions in which shifts of 1 and 4 nucleotides were slightly different between the observed and predicted distributions. Thus our results show that our predicted relative frequencies are quite accurate when compared to the observed frequencies.

2.2.10 Extending the polypyrimidine tract does not change the spliceosomal binding affinity distributions of variant and invariant acceptor splice sites

Earlier, we had constructed weight matrices of length 12 nucleotides, consisting of 6 upstream intronic nucleotides and 6 downstream exonic nucleotides from the splice boundary, which were used to score variant and invariant acceptor splice sites containing NAGNAG motifs. Our next objective is to check whether increasing the length of the polypyrimidine tract would affect distributions shown previously in Fig. 2.7.

We have also mentioned earlier in our literature review that the strength of the splice site, which includes the strength of the polypyrimidine tract, also affects splice-site selection. That said, we have constructed weight matrices of length 30, consisting of 24 upstream intronic nucleotides and 6 downstream exonic nucleotides, for both mouse and human. For the human data, we used the hg17 University of California Santa Cruz genome assembly. Using the same procedure to calculate the likelihood scores for variant and invariant NAGNAG acceptor splice sites, we show the mouse and human distributions in Figures 2.8 and 2.9 respectively. The results are qualitatively similar to those obtained with WM of length 12: alternative splicing occurs at NAGNAG acceptors in which the competing sites have comparable likelihoods, otherwise the splicing occurs at the site with the highest likelihood.

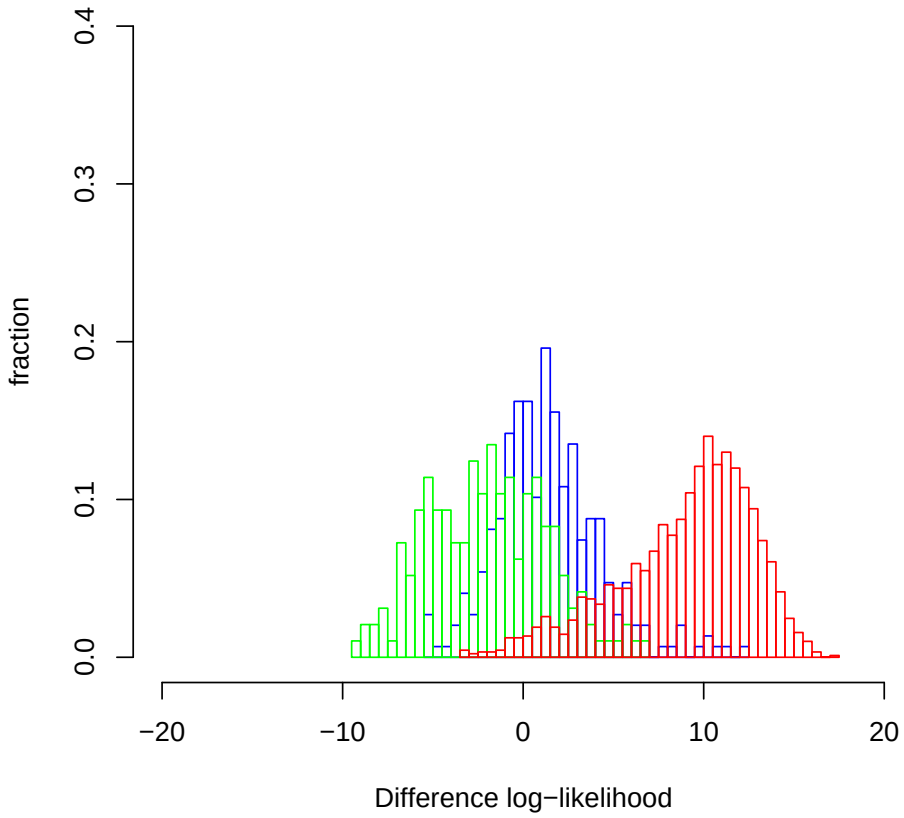


Figure 2.8: **Mouse log-likelihood distributions using our extended polypyrimidine tract WM model.** The figure shows the mouse difference log-likelihood distributions calculated using our extended polypyrimidine tract WM model. The red, green, and blue distributions refer to the fraction of all NAGNAG boundaries splicing at the first NAG acceptor, second NAG acceptor, and at both NAG acceptors respectively.

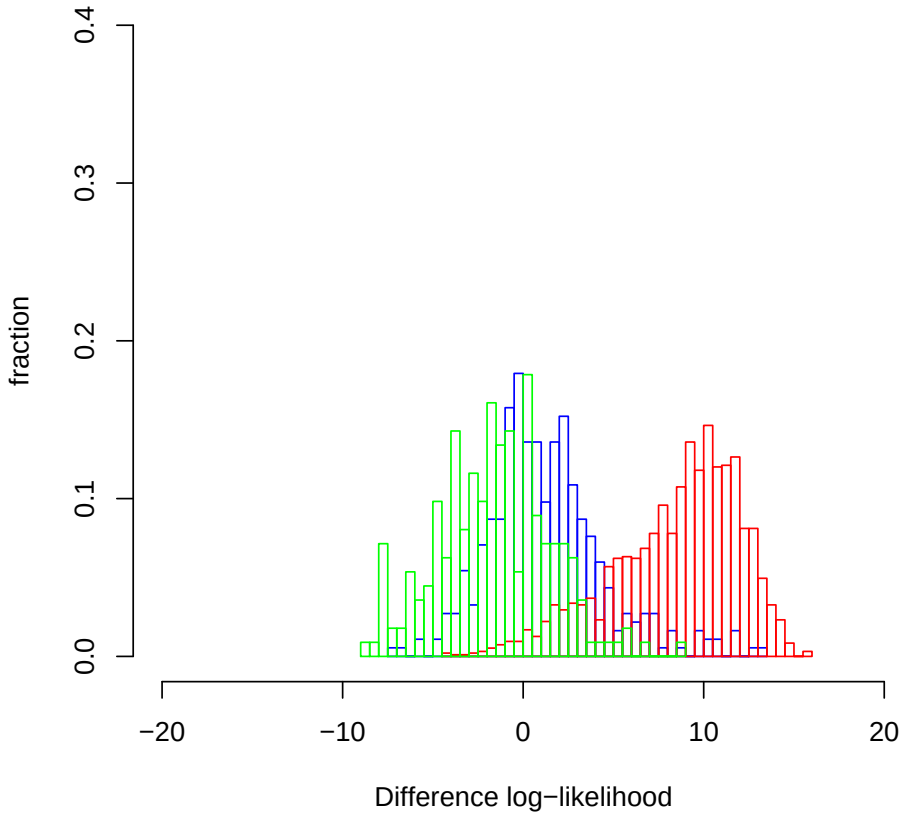


Figure 2.9: **Human log-likelihood distributions using our extended polypyrimidine tract WM model.** The figure shows the human difference log-likelihood distributions calculated using our extended polypyrimidine tract WM model. The red, green, and blue distributions refer to the fraction of all NAGNAG boundaries splicing at the first NAG acceptor, second NAG acceptor, and at both NAG acceptors respectively.

2.2.11 Similar splicing at NAGNAG acceptors in fly and human

We have checked whether the splicing at NAGNAG acceptors in mouse is similar to other species. We have calculated the likelihood scores for the fly and human using the same procedure as we did for mouse, and aside from the differences in the frequencies of distributions and extent of overlaps between the distributions, we observed similar behaviour in both fly and human when compared to mouse: splicing occurs at the acceptor site with the highest likelihood, and when the tandem acceptor sites have comparable likelihoods, one observes alternative splicing (Figs. 2.10 and 2.11 for fly and human respectively).

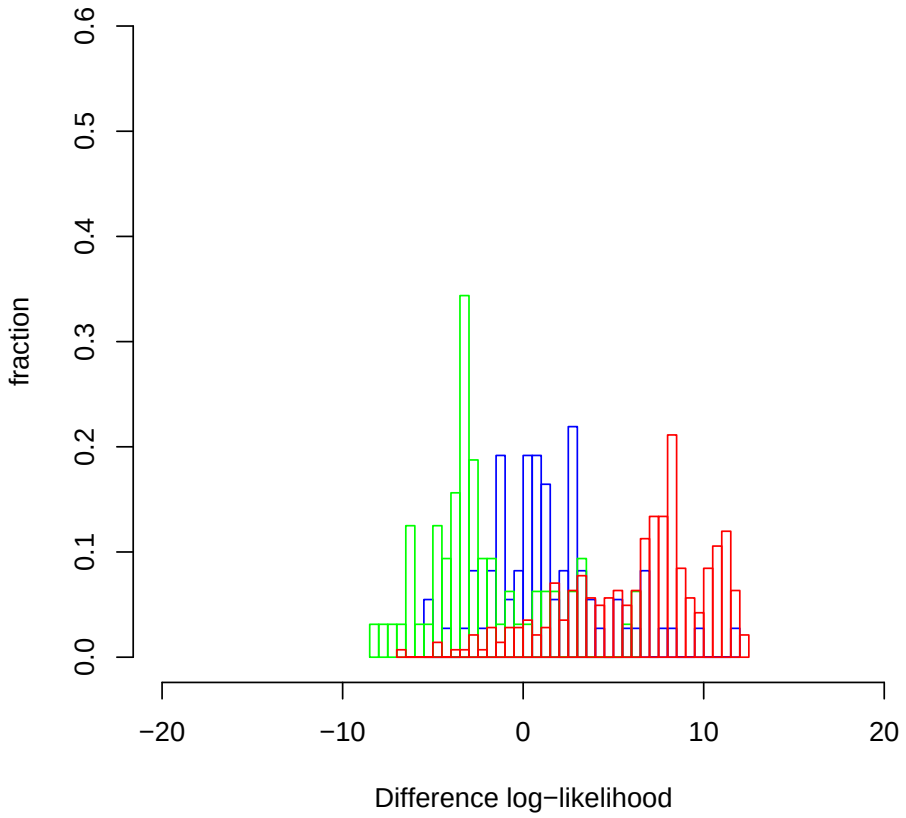


Figure 2.10: **NAGNAG acceptor splicing in the fly.** The figure shows the fly difference log-likelihood distributions calculated using the fly WM model. The red, green, and blue distributions refer to the fraction of all NAGNAG boundaries splicing at the first NAG acceptor, second NAG acceptor, and at both NAG acceptors respectively.

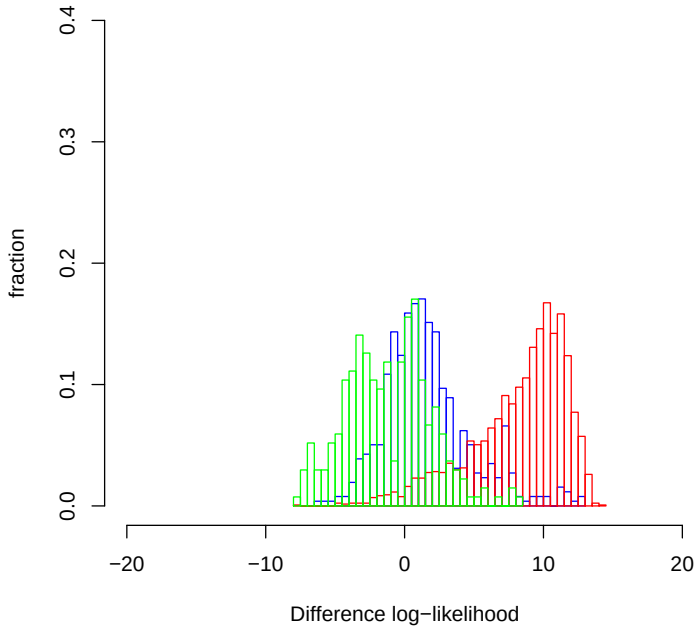


Figure 2.11: **NAGNAG acceptor splicing in human.** The figure shows the human difference log-likelihood distributions calculated using our human WM model. The red, green, and blue distributions refer to the fraction of all NAGNAG boundaries splicing at the first NAG acceptor, second NAG acceptor, and at both NAG acceptors respectively.

2.2.12 Single amino acid insertions/deletions resulting from the usage of NAGNAG acceptor splice sites do not correspond to annotated protein functional residues in FireDB

In our final analysis of NAGNAG tandem acceptor splice sites, we checked whether these single amino acid insertions/deletions, which are introduced during alternative tandem acceptor splice-site usage, map to annotated functional residues in proteins. From our total NAGNAG mouse and human datasets, we extracted those cases in which the spliceosome would be able to bind to either alternative acceptor splice site with similar affinity. We took only those cases with log-likelihood differences between -5 and 5. Then, we took only those sequences with coding sequence annotations and subjected them to the following criteria: translational start codons at the beginning of the extracted coding sequence, followed by no internal stop codons, and ending with a translational stop codon. The filtering resulted in 273 human and 211 mouse protein sequences containing a single amino acid insertion. We took these protein sequences and used the *Firestar* tool to search FireDB (See literature review on introduction to *Firestar* and FireDB), which is a database consisting of annotated functionally important residues from proteins of known structure.

We have obtained very few hits for both mouse (2 matches) and human (3 matches) from FireDB and from these matches - after carefully assessing the scores - we did not observe any biologically functional protein residue matches. Our deciding factors for assessing good quality matches lies in two crucial factors: i.) functional residue binding-site annotations which can come from two sources: a.) the number of evolutionarily related sites, which gives an indication of the number of homologous templates that bind to a particular ligand in equivalent positions as our template and this value gives us an indication on the functional relevance of the amino acids that are conserved between our template and other proteins and b.) binding-site information from the Catalytic Site Atlas or literature, and ii.) the reliability scores, which indicates how well we can transfer the binding-site annotations from our template to our query sequence. The reliability scores are computed by *Firestar*, which

uses SQUARE [76], which assigns reliability scores by extracting values for each aligned residue from PSI-BLAST profiles pre-generated from the FireDB consensus sequences. The SQUARE reliability scores provide information on which residues are aligned reliably and can be used to infer sequence conservation. The lack of biologically functional protein residue matches is primarily due to the low reliability scores and the low number of homologous proteins having conserved amino acid matches at our interested positions (See Table 2.2). Other factors that were taken into account included the PSI-BLAST E-value and the fraction coverage, which is the length of our query sequence matching to the template divided by the total length of our query sequence.

Table 2.2: **Mouse and human single amino acid matches to consensus sequences in FireDB**

Query ID	PSI-BLAST E-value	Fraction coverage of query sequence to the template	Amino acid in query, amino acid position in query	Amino acid in template	SQUARE reliability score*	Number of evolutionarily related sites (other evidence)
			Human matches			
HIT000010959.5	0.21	0.27	A,458	A	1	6
HIT000068284.2	0.005	0.04	1729,S	S	2	0
HIT000278529.2	0.71	0.35	33,K	K	1	18
			Mouse matches			
AF274321	$1e^{-69}$	0.81	39,S	S	1	0(annotated catalytic site)
BC024798	$7e^{-14}$	0.18	393,E	E	C	1

*The SQUARE reliability scores range from 1-5;C, where C is reliable with 99% reliability, a score of 1 indicates 45% reliability, 2 indicates 60% reliability, 3 indicates 75% reliability, 4 indicates 85% reliability, and 5 has 90% reliability.

2.2.13 The frequencies of incontiguous NAG(X)$_n$NAG motifs are not significantly higher in variant acceptor splice sites compared to invariant splice sites

We next focused our attention to finding incontiguous NAG(X)$_n$NAGs (where n > 0) that have significantly higher frequencies in variant acceptor splice sites compared to invariant splice sites. The NAG(X)$_n$NAG motif is observed in acceptor splice sites when the first NAG acceptor is separated from the second NAG acceptor by a certain number (n) of nucleotides. We first extracted all mouse exons that show exon-length variation at their acceptor splice sites in coding sequence regions (CDS) and then for each alternative exon, we extracted 45-nucleotide sequences, consisting of 15 intronic and 30 exonic nucleotides from the splice boundary, and then searched for NAG(X)$_n$NAG patterns within these sequences. We searched for the NAG(X)$_n$NAG satisfying the following criteria: i.) the first NAG position in an intronic position is kept while the second NAG acceptor is located within the exonic portion, separated by a distance of n bases (up to a maximum of 12) ii.) there are no other NAG motifs within the sequence between the first two NAGs. For invariant splice sites, we applied the same procedure as described for variant splice sites. The data obtained in this section utilizes the same number of mouse full-length cDNAs (FANTOM3) as previously described, but they were mapped to the mm7 genome assembly obtained from the University of California, at Santa Cruz (http://hgdownload.cse.ucsc.edu/goldenPath/mm7/bigZips/).

The frequencies of NAG(X)$_n$NAG motifs are presented in Table 2.3. We did not find any NAG(X)$_n$NAG motifs in the various distance categories (where n>0) that have significantly higher NAG(X)$_n$NAG frequencies in variant acceptor splice sites when compared to invariant acceptor splice sites. In addition, NAG(X)$_n$NAG motifs occur at much lower frequencies when compared to the tandem NAGNAG motifs in variant acceptor splice sites. This suggests that the tandem NAGNAG motifs are specifically abundant to variant acceptor splice sites when we compare them to NAG(X)$_n$NAG-containing splice sites.

Table 2.3: **Counts of variant and invariant incontiguous NAG(X)$_n$NAGs**

n category	Variant CDS exons(% freq.)	Invariant CDS exons(% freq.)	p-values
0(NAGNAG)	561(25%)	4637(4.1%)	$<2.2e^{-16}$
1	83(3%)	5559(4.2%)	0.005
2	48(2%)	4850(4.4%)	$3.2e^{-07}$
3	39(2%)	6484(5.9%)	$<2.2e^{-16}$
4	7(0.3%)	6660(6.02%)	$<2.2e^{-16}$
5	8(0.4%)	5732(5.2%)	$<2.2e^{-16}$
6	20(0.9%)	6393(5.8%)	$<2.2e^{-16}$
7	10(0.4%)	5253(4.8%)	$<2.2e^{-16}$
8	9(0.4%)	4381(4%)	$<2.2e^{-16}$
9	27(1.2%)	4547(4.1%)	$<7.18e^{-12}$
10	13(0.6%)	4102(3.71%)	$<2.2e^{-16}$
11	6(0.3%)	3468(3.14%)	$<2.2e^{-16}$
12	27(1.2%)	3666(3.32%)	$<3.8e^{-08}$
Total exon boundaries	2241	110581	

2.3 Discussion

Previously, the abundance of small (<10 nucleotides) exon-length variations at acceptor and donor splice sites has been reported by Zavolan et al. [49]. Interestingly, we found that at acceptor splice sites, 58% of these small variations constitute in-frame, variations that causes the length of an exon to increase or decrease by 3 nucleotides (3nt-variations). Even more intriguing is the fact that the abundance of these 3nt-variations was not completely due to the effects of NMD, since we computed the fraction of these variations before NMD and still found an abundance of these 3nt-variations compared to other small exon-length variations of 1-4 nucleotides. We thus conjecture that the abundance of these 3nt-variations is also influenced in part by sequence biases that cause the spliceosome to "slip" at neighboring, competing splice sites.

Many of these 3nt-variations contain tandem NAGNAG motifs, but we did not find evidence to support their functionality. For instance, we did not find evidence of highly conserved NAGNAG motifs in mouse variant exons, quite on the contrary - NAGNAG motifs were significantly more conserved in invariant exons. Moreover, although we have observed a high frequency of NAGNAG motifs in variant CDS exons, we found that the frequencies of NAGNAG motifs at the splice boundaries of all noncoding exons were in fact significantly higher when compared to all coding exons. Therefore, we conclude that the occurrence of NAGNAG motifs is not indicative of their function in proteome diversification. We further suggested that the observed variations at NAGNAG motifs are due to the stochasticity in spliceosome binding.

We therefore wanted to compare the binding affinities of the spliceosome acting on the NAGNAG acceptor splice sites. For this purpose, we used our WM model of invariant acceptor splice sites to estimate the binding affinities of the spliceosome splicing at the first and second NAG. We found that the likelihood difference between the two NAGNAG sites is compatible with a model in which the spliceosome binds stochastically at one or the other of the sites.

It is, however, possible that selection of a particular NAG can be driven by specific factors affecting splice-site selection. But here in our study we show that

based on sequence alone one can determine which NAG the spliceosome prefers and that we can distinguish between variant and invariant NAGNAG acceptor splice sites even when their polypyrimidine tracts have been extended. Our WM makes a number of simplifications and it is possible that a more accurate model would be able to quantitatively predict the relative usage of competing splice sites. For instance, Tsai et al. showed that the sequence between the branchpoint and the NAGNAG splice site as well as the nucleotide preceding the AG dinucleotide is important for splice-site selection [99]. Refining the spliceosome-premRNA interaction model is an interesting problem which can be addressed in future work.

Without deviating too much from the discussion of NAGNAG tandem acceptor splice sites, we have found another use for our WM model - that is we were able to use the probabilities computed from our donor and acceptor WMs to predict, with reasonable accuracy, the relative frequencies of very small exon-length variations of 1-4 nucleotides. The general order of predicted frequencies were similar when compared to those observed distributions, except for the donor distributions in which the predicted distributions of shifts of 1 and 4 nucleotides were slightly higher than the observed fractions. Most importantly, predicted frequencies of 3nt-variations were highly similar to their observed abundant frequencies before NMD. Thus, these results further support our hypothesis that the abundance of small exon-length variations must be due to the stochasticity in spliceosomal binding to closely spaced splice sites.

The distributions of log-likelihood differences in fly and human are similar to that in mouse suggesting that NAGNAG splice-site acceptors are quite common and the spliceosomal mode of action - splicing only at the first NAG with higher affinity, splicing only at the second NAG with higher affinity, and splicing at both NAGs with roughly equal affinities - is also similar amongst the species examined.

The functional consequence of using NAGNAG acceptors is either the inclusion or exclusion of a single amino acid. Hiller and others [94] proposed that a single amino acid can increase the protein functional potential and we have tested this hypothesis by taking those NAGNAG acceptors that have competitive binding affinities and checked whether the single amino acid residues have good matches to functionally

annotated protein residues in FireDB. We did not find any reasonable matches and this indicates that there is no evidence so far to say that the inclusion or deletion of the amino acid generally affects the function of proteins. Literature searches also fail to reveal abundant evidence for functional NAGNAG single amino acid insertions or deletions. Thus we conclude from our results that the majority of the NAGNAGs do not contribute significantly to the protein functional diversity as proposed by Hiller and others.

Our preliminary findings on incontiguous NAG$(X)_n$NAG motifs, in which the first NAG site is separated by a short distance to the second NAG site, has revealed that NAG$(X)_n$NAG motifs are present at lower frequencies in variant acceptor splice sites when compared to the frequencies of tandem NAGNAG motifs in variant acceptor splice sites. Furthermore, they have significantly lower frequencies in variant acceptor splice sites when compared to invariant acceptor splice sites suggesting that the abundance of tandem NAGNAG motifs are specific to variant acceptor splice sites.

Chapter 3

Investigation into the dependency of the inclusion/exclusion of cassette exons on the transcription start sites of their transcripts

3.1 Introduction/Rationale

The link between transcription and splicing is mediated by the C-terminal domain of RNA polymerase II (RNA pol II). The C-terminal domain (CTD) is composed of heptapeptide repeats containing the sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser, with yeasts having 26 repeats and mammals 52. The serine residues in the repeats can either have a low level of phosphorylation during or prior to transcription initiation or a high level of phosphorylation after 25-30 nucleotides have been polymerized and during transcription elongation. The function of CTD phosphorylation is to enhance productive RNA pol II elongation (reviewed in Ref [101]).

Previously, two models of how transcription can affect alternative splicing have been proposed: i.) The CTD of RNA pol II serves to recruit splicing factors, which will influence splicing decisions downstream [102]. An example is the SRp20 protein, which is recruited by RNA pol II and promotes exon-skipping of the fibronectin cassette exon [103]. ii.) The rate of transcription elongation determines whether

exons with weak splice sites are included (at low elongation rates) and skipped (at high elongation rates) from the mature mRNA [104, 105]. There are some overlaps between the two models proposed in which the components of the transcription machinery modulates the RNA pol II elongation rate as well as interacting and recruiting known splicing factors to the site of alternatively spliced exons. This can be seen with the illustrated example of the alternative splicing of CD44 variant exons (See Fig. 3.1), which is regulated by the BRM(Brahma) subunit of the SWI/SNF chromatin remodeling complex [106].

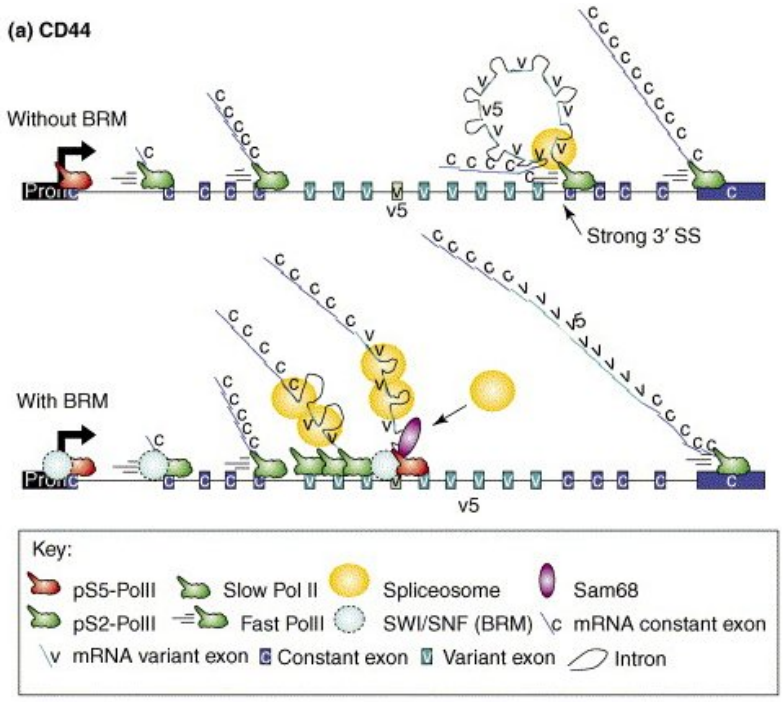


Figure 3.1: **The BRM-regulated splicing of CD44 variant exons.** The CD44 gene contains ten clustered, variable exons with weak acceptor splice sites that are included or excluded during splicing. The SWI/SNF complex is a chromatin remodeling complex, such that it regulates the binding of transcription factors to DNA embedded in chromatin and it requires ATP in order to alter the nucleosome structure. BRM, a subunit of the SWI/SNF complex, interacts with Sam68, a factor that enhances the inclusion of exon v5 by binding to this exon. In addition, BRM is also known to associate with other splicing factors. As shown in the figure, the combined complex of BRM and SAM68 at the variant exonic region, may provide an entry point for incoming spliceosomes or create an obstacle for RNA Pol II (bottom figure), and thus slowing the elongation rate relative to the situation when BRM is lacking (top of the figure). The suggested role of BRM is to reduce the processivity of RNA pol II by altering the CTD phosphorylation pattern. Figure extracted from Auboeuf et al. [107].

Here we have studied human and mouse internal cassette exons and have found that their inclusion/exclusion from the mature mRNA frequently depends on the transcription start site (TSS) used to generate their corresponding pre-mRNA. We further show that those cassette exons that are strongly associated with their TSSs may have been created recently.

3.2 Methods and Results

The work performed in sections 3.2.1 - 3.2.6 has already been published in the DNA Research article entitled 'Computational analysis of full-length cDNAs reveals frequent coupling between transcription and splicing programs' and a copy is included in Chapter 5. Please refer to the Materials and Methods section in the paper for a detail description of sections 3.2.1 - 3.2.6.

3.2.1 Construction of TSS-associated,TSS-independent, and constitutive exon datasets

We begin our analyses by first mapping mouse and human full-length transcripts to their respective genomes and using our in-house splicing pipeline, SPAED, to obtain 4964 mouse and 11664 human internal cassette exons with no other splice variation and these cassette exons are part of transcription units with multiple TSSs. We defined our TSS quite conservatively as the initial exon of the transcript since the position of the TSS can be variable [108]. Hence, different initial exons within a transcription unit correspond to different TSSs. For each cassette exon and each individual TSS we counted the number of times the cassette exon was included and the number of times the exon was skipped when that particular TSS was used. We then needed a method to quantify the coupling between the inclusion or exclusion of the cassette exon and the usage of individual TSSs. The method we have used to quantify such an association was developed by Professor Erik van Nimwegen and can be briefly described as follows. We estimate the relative likelihood of two models, a "dependent" model, which assumes that there is a TSS-associated probablity for the inclusion of the cassette exon, and an "independent" model which assumes that

the frequency of exon inclusion is the same for all TSSs. Using this model (given in detail in Chern et al. [109]), we estimated that the fraction of mouse and human internal cassette exons whose inclusion is dependent on the TSS is 24% and 30% respectively. We further compute for each cassette exon the posterior probability that its inclusion is dependent on the TSS. We then obtained cassette exons whose posterior probabilities were either within the top or bottom 10%. We will refer to those cassette exons that are within the top 10% as 'TSS-associated' exons and the bottom 10% as the 'TSS-independent' exons. For mouse, we have obtained 496 TSS-associated and TSS-independent exons and for human, we obtained 1166 cassette exons in each of the two categories.

We further obtained 5136 mouse and 6377 human internal exons that were not annotated with any splice variation and were included in more than 10 transcripts as constitutive exons whose properties we wanted to compare to those of TSS-associated and TSS-independent exons.

3.2.2 Comparisons of sequence features between TSS exons and constitutive exons

In Table 3.1 (see below), we have summarized the sequence features that were compared between TSS exon types (TSS-associated and TSS-independent) and constitutive exons. For sequence conservation, we have used the UCSC pairwise genome alignments to obtain orthologous regions for both mouse and human TSS exon types and constitutive exons. We have also checked the fraction of exons conserved over some percentage thresholds (>=50%,>=80%). We observed that in general (for both human and mouse) TSS-independent and constitutive exons have significantly higher conservation levels than TSS-associated exons (mouse p-values=0.0007,$<2.2e^{-16}$;human p-values=$1.1e^{-06}$,$<2.2e^{-16}$), with constitutive exons having the highest conservation level.

We have also examined the inclusion rate, which is the total number of transcripts that include the exon-of-interest divided by the total number of transcripts that span the exon-of-interest. We have found that the TSS-associated exons have lower average inclusion rates than the other exon types. We observed similar average

inclusion trends between human and mouse.

We further examined the proportion of symmetrical exons (i.e. exons whose length is a mutiple of 3) and found that this proportion is lower for TSS-associated exons compared to TSS-independent exons (for mouse p-value=0.097, for human p-value=0.072). However, the symmetrical exon proportions for constitutive exons were even lower when compared to TSS-associated exons, but this difference was only significant for human exons (p-value=0.02). Once again, we have observed similar trends of symmetry in our mouse and human data.

Table 3.1: **Comparison of TSS-associated, TSS-independent, and constitutive exons**

Data set	Number exons	With orthologs	Fraction conserved(>= 50% cutoff)	Fraction conserved(>= 80% cutoff)	Average inclusion rate	Proportion symmetrical
Human exons						
TSS-associated	1166	1006	0.77	0.54	0.75	0.41
TSS-independent	1166	1078	0.87	0.70	0.86	0.45
constitutive	6377	6343	0.95	0.81	1	0.38
Mouse exons						
TSS-associated	496	438	0.81	0.58	0.75	0.42
TSS-independent	496	469	0.90	0.73	0.88	0.48
constitutive	5136	5117	0.93	0.78	1	0.4

3.2.3 Weak correlation between the posterior probabilities of orthologous, TSS-associated human and mouse exons

To obtain orthologous mouse and human exons, we used genome sequence alignments. We started with our query mouse TSS-associated exons and obtained the corresponding human exons using our genome alignments obtained from the University of California (http://hgdownload.cse.ucsc.edu/goldenPath/mm7/vsHg18/axtNet/). We also obtained the genomic coordinates of these human exons and matched them to our dataset of human TSS-associated exons (described in the previous section) to identify orthologous TSS-associated exons. This procedure was performed in the same way starting from human TSS-associated exons and using genome alignments from University of California to obtain our corresponding mouse exons (http://hgdownload.cse.ucsc.edu/goldenpath/hg18/vsmm7/axtNet/). An orthologous pair of TSS-associated exons is defined as a pair of orthologous human and mouse exons which were both considered TSS-associated. We plotted the posterior probabilities of 668 mouse-human orthologous pairs and have observed a weak, significant correlation (p-value=0.002) between mouse and human TSS association (See Fig. 3.2).

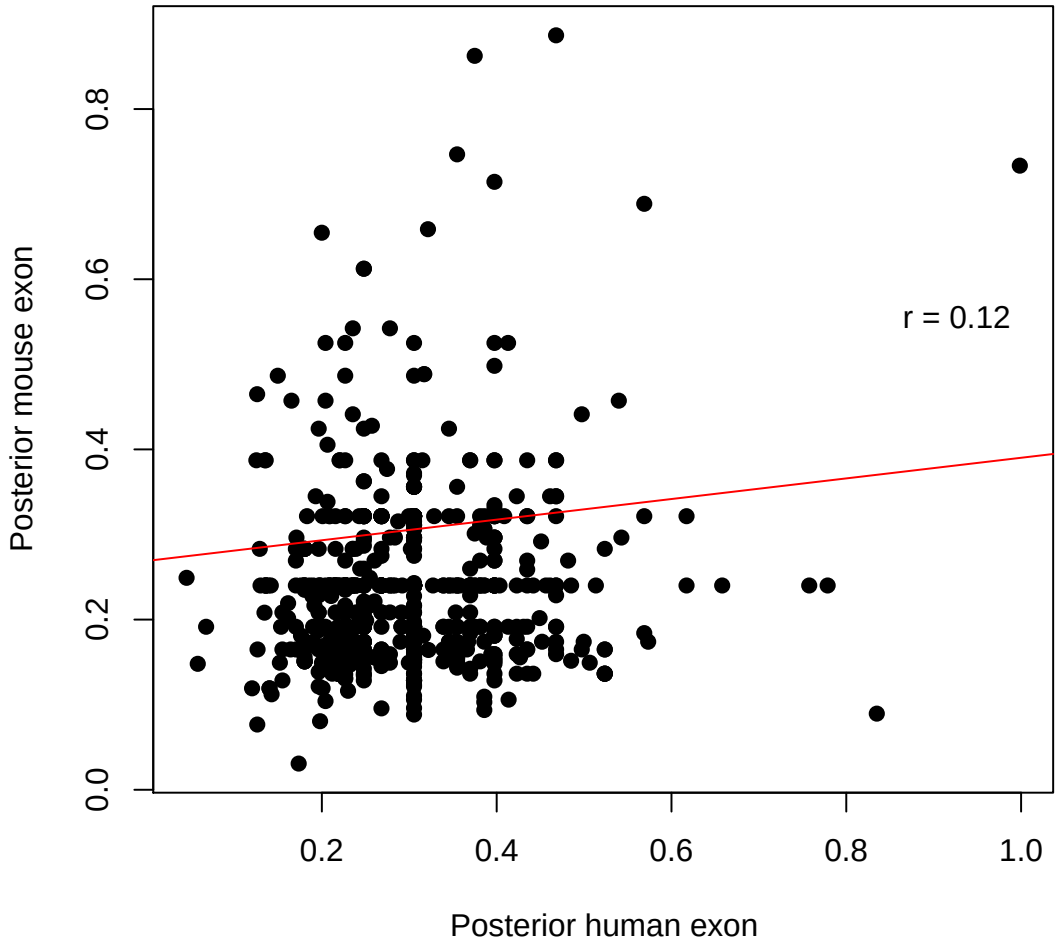


Figure 3.2: **Correlation between the posterior probabilities of TSS association for orthologous human-mouse cassette exons.** Each dot refers to a human-mouse orthologous pair, with the posterior probability of TSS-association of the mouse exon on the y-axis and of the human exon on the x-axis. We have computed the correlation coefficient to be 0.12, p-value = 0.002. Figure taken from Chern et al. [109].

3.2.4 Distance to the TSS is not predictive for the inclusion or for the exclusion of TSS-associated exons

We questioned whether TSSs that are closer to the TSS-associated cassette exons tend to promote inclusion and those that are far, promote exclusion, or vice versa. We performed this experiment by first calculating the distance between cassette exon and their TSSs. We define those TSS that promote the inclusion of the cassette exon as 'inclusion-promoting TSS' and those TSSs that promote the skipping of the cassette exon as 'skipping-promoting TSS'. We then computed the average distance to an inclusion-promoting TSS by summing the distances for all transcripts that included the cassette exon for each TSS and dividing this number by the total number of transcripts that include the cassette exon. We perform the same procedure for computing the average distances to skipping-promoting TSSs. We then took the logarithm of the difference between the inclusion-promoting and skipping-promoting average distances for each TSS-associated exon, and constructed a histogram from these distances for mouse and human (See Fig. 3.3).

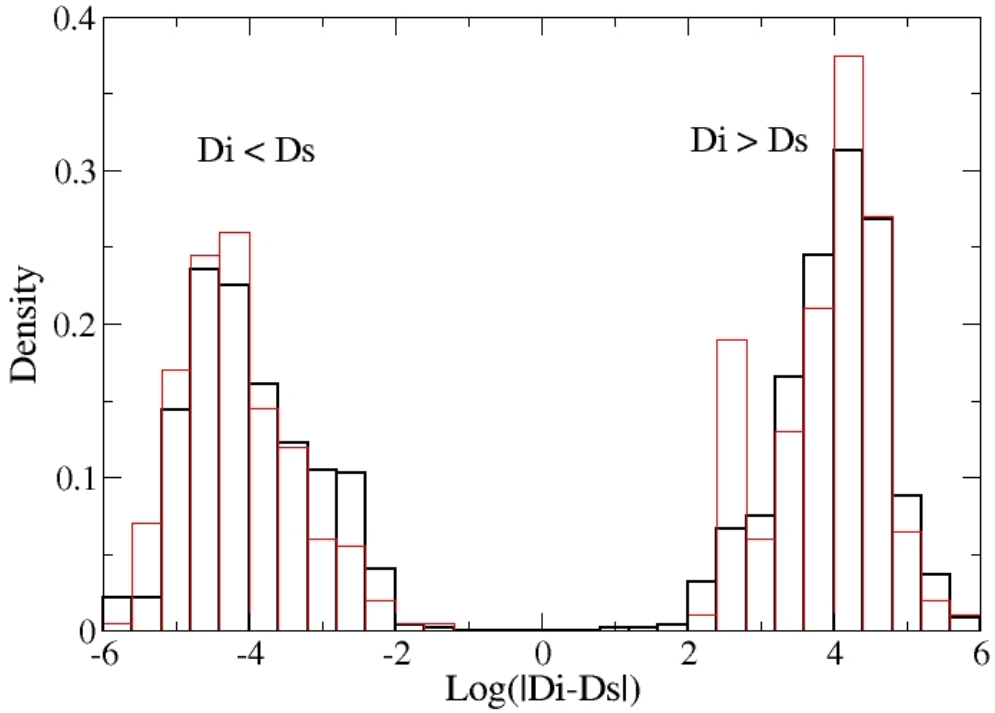


Figure 3.3: **Histogram of the signed difference between the logarithm (base 10) of the inclusion-promoting and skipping-promoting average distances.** The histograms shown in black and red represent the human and mouse TSS-associated exons respectively. The distribution on the right indicates when the average distance to the inclusion-promoting TSS is larger than the distance to the exclusion-promoting TSS, and the left distribution indicates that the average distance to the inclusion-promoting is smaller than the distance to the exclusion-promoting TSS. Figure taken from Chern et al. [109].

3.2.5 The frequency of the inclusion and exclusion of TSS-associated exons occurring within the same tissue is between 14-21% in mouse and human

One mechanism in which the inclusion and exclusion of TSS-associated exons can occur is via the regulation by tissue-specific factors. Our goal for this experiment is to estimate the frequency of TSS-association cases in which the transcripts that includes or excludes the TSS-associated exons are expressed within the same tissue, thus ruling out the possibility of tissue-specific regulation. We collected the transcripts for each TSS-associated exon and obtain their tissue annotations from GenBank. We have obtained 14% (58/419) and 21% (223/1042) of mouse and human TSS-association cases respectively.

3.2.6 The strength of the splice sites flanking TSS cassette exon types does not agree with the kinetic model of transcription-coupled splicing

According to the second proposed model of transcription-coupled splicing (the kinetic model), when cassette exons are flanked by weaker splice sites compared to their flanking exons, then they may be skipped unless the polymerase elongation rate is sufficiently slow to allow exon recognition by the spliceosome. We have tested this hypothesis by examining the splice-site strengths of all exons in our datasets. We have extracted the acceptor and donor splice sites flanking all TSS cassette exon types (referred to as "central exons" in Table 3.2) and their flanking exons (referred to as "upstream and downstream exons" in Table 3.2). We have also extracted a reference dataset composed of constitutive exons for comparison. We used a webserver (http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm), which implements the Shapiro and Senapathy model [110], to calculate the strengths of the splice sites in our datasets. Table 3.2 shows the computed mean splice-site scores for each category of exons.

Table 3.2: **Mean splice-site scores of TSS exon types and constitutive exons**

Exon types	Human	Mouse
Acceptor splice site of central exon		
TSS-associated	81.2	80.0
TSS-independent	81.6	81.8
constitutive	79.7	82.7
Acceptor splice site of upstream exon		
TSS-associated	73.7	71.7
TSS-independent	78.1	79.4
constitutive	77.5	77.2
Acceptor splice site of downstream exon		
TSS-associated	81.1	80.5
TSS-independent	80.6	81.7
constitutive	81.3	83.2
Donor splice site of central exon		
TSS-associated	80.5	78.5
TSS-independent	80.3	77.5
constitutive	81.5	79.2
Donor splice site of upstream exon		
TSS-associated	79.9	75.3
TSS-independent	80.9	79.3
constitutive	81.3	83
Donor splice site of downstream exon		
TSS-associated	76	73.5
TSS-independent	73.8	73
constitutive	77.3	83.2

From Table 3.2, we summarize our findings as follows:

1.) For TSS-associated exons in human and mouse, we did not observe that their donor and acceptor splice-site strengths of their flanking exons were stronger than the splice-site strengths of their central cassette exons. In fact, the donor and acceptor splice-site strengths of their flanking exons were similar or significantly weaker (human upstream acceptor p-value=$6e^{-07}$, mouse upstream acceptor p-value=$4e^{-04}$, human downstream donor p-value=0.0 2) than their central exons. Thus, our results do not meet the requirements of the kinetic model.

2. We did not find any significant differences in the acceptor and donor splice-site strengths between TSS-associated and TSS-independent groups of their central exons.

3. In most cases, the acceptor and donor splice-site strengths of constitutive exons were stronger than their TSS exon types, which was expected.

3.3 Discussion

In this study, we have characterized the sequence properties of TSS-associated exons compared to other exon types. The lesson we have learned from comparing various sequence features between TSS exons and constitutive exons is that TSS-associated exons may have been created recently given their lower conservation levels and average inclusion rates compared to other exon types. Furthermore, the weak correlation between the posterior probabilities of mouse and human TSS-associated, orthologous pairs suggest that the inclusion of cassette exons may depend on factors that change on fast evolutionary timescales. On the other hand, it is clear that more data is necessary in order to make definite statements about the TSS-dependence of orthologous exons.

One proposed mechanism (also known as the kinetic model) in how transcription affects alternative splicing involves the elongation rate of RNA polymerase and the strength of the splice sites. Kornblihtt [104] proposed that if an alternative exon has a weak acceptor splice site, then a fast-elongating polymerase would lead to its skipping and a slower one would lead to its inclusion. Based on our observations that

the splice sites of our TSS-associated exons have comparable or stronger splice-site strengths compared to their upstream and downstream exons, we conclude that the TSS-dependency of our exons cannot be explained by the kinetic model.

The fact that RNA polII appears to pause at the $5'$ region of the transcription units of genes [111, 112, 113, 114] during early transcription elongation led us to question whether the TSS-associated exons could be explained by this process. That is, we asked whether the majority of TSS-associated exons are included when the most proximal promoters are used and skipped when the most distal promoters are used. We found that the proportion of those TSS-associated exons that were preferentially included when the most upstream TSSs were used (where $D_i < D_s$) was similar to the proportion of TSS-associated exons that were preferentially included when the most downstream TSSs (where $D_i > D_s$) were utilized. Thus we cannot explain their inclusion by the RNA polymerase pausing at $5'$ ends of transcripts.

A second mechanism that could explain the TSS-association could involve tissue-specific factors that play a role in tissue-specific transcription and tissue-specific splicing. By filtering for only those TSS-associated exons for which the transcripts that validate the exon inclusions and exclusions were expressed in the same tissue, we can rule out the possibility that only tissue-specific factors regulate transcription and splicing. We have found that approximately 14-21% of mouse and human TSS-cassette exons have their inclusions/exclusions expressed in the same tissue and thus conclude that at least for some of these exons, alternative splicing is likely to be regulated through direct coupling of transcription and splicing events. In addition, if one were to further study experimentally the mechanism of TSS-dependency of cassette exons, we would recommend using this particular subset of TSS-associated exons.

Chapter 4

A first look at mutually dependent and mutually exclusive splicing events *in silico*

4.1 Introduction/Rationale

Mutually exclusive (ME) splicing describes the situation when alternatively spliced exons are never found together in the same transcript. This alternative splicing pattern appears to be quite common in receptor and channel subunits (fibroblast growth factor receptor 2 [115], glutamate receptor subunits 1-4 [116], sodium channel alpha subunit [117], calcium$_v$L-type channels [118]), in structural proteins (*C. elegans let2* gene [119], alpha and beta tropomyosins [120]), and they are also found in the muscle protein subunits such as the Troponin T subunit [121]. The mutually exclusive splicing found within these genes typically results in developmentally- or tissue-specific isoforms. In mammalian genes, we often observe two ME exons involved in the ME splicing, whereas in *Drosophila*, the process appears to be far more complex. The well-known *Dscam* gene in *Drosophila* can give rise to over 38000 different alternatively spliced transcripts. The splicing of *Dscam* is achieved through the alternative splicing of one alternative exon from each of the four available exon clusters, with each cluster containing ME exons (See Fig 4.1).

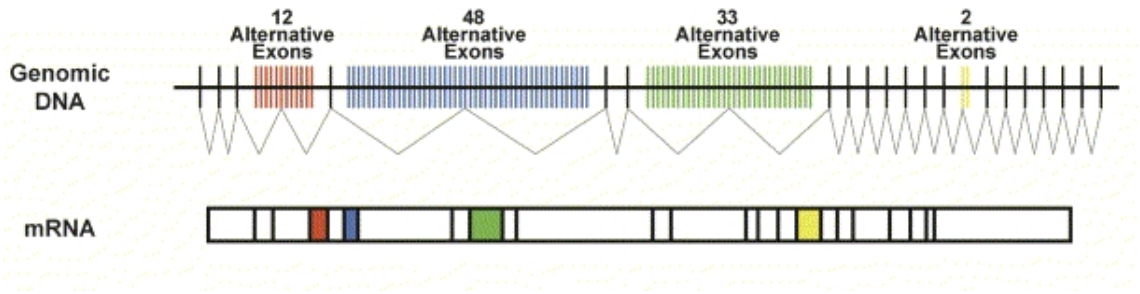


Figure 4.1: **Mutually exclusive splicing of the *Dscam* gene.** Only one exon from each of the alternative exon clusters, which are shown in the different colors, are included in the final mRNA transcript. Figure extracted from Wojtowicz et al. [122].

The RNA-binding protein asd-2 is responsible for regulating the ME splicing of the *let-2* gene in *C. elegans* [123], whereas in the *Drosophila Dscam* gene, it is the combination of a splicing repressor, hrp36, and alternative RNA structures that are regulating the alternative splicing[124].

The current study is focused on groups of consecutive cassette exons (more than one) that are always either skipped or included together within transcripts, which we define as "mutually dependent" exons (MD). This kind of splicing have not been previously described in literature as a separate form of alternative splicing, thus this is the first known study to characterize the properties of MD exons relative to ME and constitutive exons. In addition, no large-scale computational study has so far examined mutually exclusive splicing in both mouse and human genomes, although this type of splicing does exist as we have seen from our literature examples described above. Our preliminary analyses of MD and ME splicing in mouse and human, revealed subtle differences in the features of the two types of exons, which may prove to be useful for future MD and ME studies.

4.2 Methods and Results

4.2.1 Transcript mapping and alternative splicing annotations

We have used full-length mouse and human cDNAs and have mapped these transcripts to their respective genomes using our in-house splicing pipeline, SPAED. The alternative splice annotations were retrieved from our mouse and human full-length database generated for the studies in Chapters 2 and 3. The procedure for mapping and obtaining our splice variations are also the same as described in the previous chapters.

4.2.2 Comparison of symmetry, exon- and intron-lengths, and sequence conservation between the different cassette exon types and constitutive exons

We define an ME exon cluster as consecutive cassette exons that are observed to be skipped, but are never found together on the same transcript. Similarly, an MD exon cluster is defined as a group of consecutive cassette exons that are always included or skipped together within the same transcript. We have obtained 60 ME and 221 MD examples for mouse, and in human, we have obtained 91 ME and 382 MD examples. The total number of ME and MD exons in mouse was 120 and 520 respectively, and in human - we obtained 183 ME and 896 MD exons. We have also obtained groups of consecutive cassette exons that do not correspond to either of the ME and MD type, which we have defined as a 'mixed' exon cluster. We have obtained 19 (38 exons in total) and 51 (103 exons) such cases in mouse and human respectively. For constitutive exons, we have retrieved internal exons with no splice variation and have obtained 102083 mouse and 123467 human exons. For all exon types (ME, MD, constitutive, and mixed), we have computed the proportion of exons and introns whose length is a multiple of 3. We determined the sum of all exon lengths within an ME, MD, constitutive, and mixed cassette exon cluster and the length of intron between the exons present within each ME, MD, mixed, and constitutive exon cluster. The human-mouse and mouse-human orthologous alignments were extracted for all exons within the ME, MD, and mixed exon clusters, using human and mouse whole-genome alignments obtained from the University of California, at Santa Cruz (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsMm8/ and http://hgdownload.cse.ucsc.edu/goldenPath/mm7/vsHg17/). We then filtered the alignments that have a percentage identity of greater or equal to 85% to obtain our fraction of highly conserved exons. The results are summarized in Table 4.1 below.

Table 4.1: **Comparison of ME, MD, mixed, and constitutive exons**

Sequence Feature	ME exons	MD exons	Constitutive exons	Mixed cases
Human exons				
Proportion of symmetrical exon clusters	0.29	0.49	0.39	0.47
Proportion of individual symmetrical exons	0.38	0.37	0.33	0.56
Avg. intron length	7924.4	3609.1	4831.5	9864.5
Avg. intron length (less than 40kb)	4474	3204.5	-	-
Avg. individual exon length	181.5	138.5	144.1	119.6
Fraction of highly conserved exons	0.46	0.46	-	0.60
Mouse exons				
Proportion of symmetrical exon clusters	0.4	0.54	0.39	0.53
Proportion of individual symmetrical exons	0.43	0.36	0.4	0.74
Avg. intron length	5938.3	2879.5	3694.3	5319.8
Avg. intron length (less than 40kb)	3383.1	2475.8	-	-
Avg. individual exon length	223.5	129.1	156.8	90.5
Fraction of highly conserved exons	0.68	0.50	-	0.74

Although the proportions of symmetrical exon clusters were not significantly different between mouse ME and MD exons (p-value=0.08), the expected trend that this proportion was higher in MD than in ME exons was observed. In human, the proportion of symmetrical exon clusters was significantly higher in MD compared to ME exons as expected (p-value=0.0006) and also to constitutive exons (human p-value=0.0001). The latter comparison also gave a significant difference in mouse (p-value=$1.53e^{-06}$). However, comparisons of MD exons to mixed cases did not yield any significant differences in either of the species which suggest that the pressure of maintaining the exon reading frame in mixed cases may be similar to MD exons.

Although the proportions of individual symmetrical ME exons were not significantly higher than MD exons, the higher proportion observed for ME exons was expected and we can see this trend more clearly in mouse. Both human and mouse proportions of individual symmetrical exons were significantly lower than the proportions of mixed exons (Human p-value=0.0003,mouse p-value=$9e^{-06}$). However, comparisons to constitutive exons did not reveal any consistent trends.

We have found that human MD exon clusters contain significantly shorter intron lengths compared to the intron lengths in other exon cluster groups (ME exon clusters p-value=0.01; constitutive exon clusters p-value=0.003; mixed exon clusters p-value=0.006). The trend is similar in mouse, but only significantly shorter when compared to constitutive exon clusters (p-value=0.03). When we focused only on intron lengths less than 40 kilobases(kb) between ME and MD groups, we still observe the same trend in both species, but with no significance in either of the species. This suggests that extremely large intron lengths were present in the human ME dataset, however the trend that MD exon clusters have shorter intron lengths compared to other datasets still remains.

Since we have observed significantly shorter intron-lengths in human MD exon clusters compared to other exon groups, we have examined their intron-length distributions more closely (See Fig. 4.2).

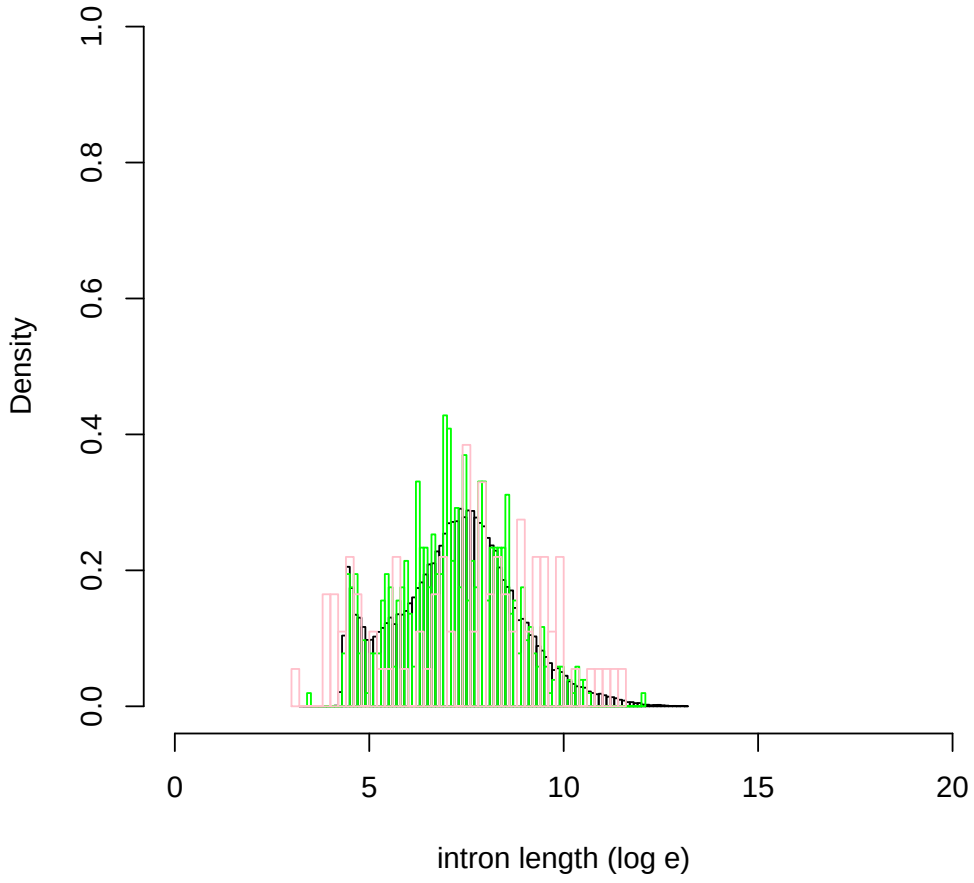


Figure 4.2: **Distribution of intron length for MD (green), ME (pink), and constitutive (black) exon clusters**

Comparison of the human MD intron-length distribution to the ME intron-length distribution was found to be significant according to the kolmogorov-smirnov test (p-value= 0.0008) and comparisons to other distributions did not yield any significance.

We have observed the same trends in mouse and human that ME exons have the largest exon lengths, followed by constitutive, MD, and mixed exons. However, ME exon lengths were not significantly larger than MD exon lengths. Comparisons of MD exon lengths to other exon types yielded significance in mouse (constitutive p-value=$3.65e^{-11}$, mixed p-value=0.002). Lastly, we did not observe a consistent trend in the conservation pattern of the different categories of exons except that the proportion of highly conserved exons was highest in the mixed exon category.

4.3 Discussion

Our analysis of MD and ME exons revealed that the properties of these exons matched our expectations. The proportion of symmetrical exon clusters was higher in MD exons than in ME exons since MD exons are always included or excluded together and not independently of each other. Conversely, we did observe higher proportion of symmetrical ME individual exons since they are individually included and excluded and never occur together.

Interestingly, we observed that the average intron length between the individual exons in MD clusters was smaller than in ME clusters. We envision several possible explanations: i.) the inclusion and exclusion of the individual exons within the MD cassette are dependent upon each other and the shorter intron length may facilitate the loading of the spliceosome at consecutive downstream introns, ii.) long introns may result in high rates of RNA polymerization and poor exon recognition during co-transcriptional splicing.

The relatively small number of these various exon categories made it difficult to infer statistically significant differences in their properties. Nonetheless, we were able to draw some conclusions about their symmetry and about the lengths of introns in these exon clusters. These conclusions can be interpreted in light of the mechanism of splicing of these exons and we therefore believe that our study provides a good

basis for follow-up studies of MD exon clusters.

Chapter 5

Publications

The following publications for chapters 2 and 3 are included in this section.

5.1 Publication for Chapter 2

OPEN ACCESS Freely available online PLoS GENETICS

A Simple Physical Model Predicts Small Exon Length Variations

Tzu-Ming Chern[1☯], Erik van Nimwegen[1☯], Chikatoshi Kai[2], Jun Kawai[2,3], Piero Carninci[2,3], Yoshihide Hayashizaki[2,3], Mihaela Zavolan[1*]

1 Division of Bioinformatics, Biozentrum, University of Basel, Basel, Switzerland, **2** Genome Exploration Research Group (Genome Network Project Core Group) RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Japan, **3** Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, Wako, Japan

One of the most common splice variations are small exon length variations caused by the use of alternative donor or acceptor splice sites that are in very close proximity on the pre-mRNA. Among these, three-nucleotide variations at so-called NAGNAG tandem acceptor sites have recently attracted considerable attention, and it has been suggested that these variations are regulated and serve to fine-tune protein forms by the addition or removal of a single amino acid. In this paper we first show that in-frame exon length variations are generally overrepresented and that this overrepresentation can be quantitatively explained by the effect of nonsense-mediated decay. Our analysis allows us to estimate that about 50% of frame-shifted coding transcripts are targeted by nonsense-mediated decay. Second, we show that a simple physical model that assumes that the splicing machinery stochastically binds to nearby splice sites in proportion to the affinities of the sites correctly predicts the relative abundances of different small length variations at both boundaries. Finally, using the same simple physical model, we show that for NAGNAG sites, the difference in affinities of the neighboring sites for the splicing machinery accurately predicts whether splicing will occur only at the first site, splicing will occur only at the second site, or three-nucleotide splice variants are likely to occur. Our analysis thus suggests that small exon length variations are the result of stochastic binding of the spliceosome at neighboring splice sites. Small exon length variations occur when there are nearby alternative splice sites that have similar affinity for the splicing machinery.

Citation: Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, et al. (2006) A simple physical model predicts small exon length variations. PLoS Genet 2(4): e45. DOI: 10.1371/journal.pgen.0020045

The Genome Network Project – FANTOM3 article collection

Introduction

Anticipating that the sequencing and initial annotation of the human [1,2] and mouse [3] genomes will not be able to uncover all the complexities of mammalian gene structures, several groups have focused on producing high-quality, annotated transcript data such as the Riken Clone Collection [4,5], the Mammalian Gene Collection [6], the NCBI Reference Sequence [7], and the NCBI unfinished high-throughput cDNA sequences. In conjunction with genome sequences, these data have revealed that alternative splicing [1,8–11], alternative transcriptional initiation, and alternative polyadenylation [5,12,13] are extremely common, affecting over 70% of mammalian genes [14].

Many factors at the molecular level appear to play a role in the regulation of splicing, from the recognition of the primary splice signals by the components of the spliceosome, to modulation of splicing via interactions of signaling and RNA processing pathways with the spliceosome. More specifically, the choice of splice site appears to be determined by a combination of (1) the strength of the splice signal, i.e., the affinity for the splicing machinery of the sequence around the splice site, (2) structural constraints set on the interactions of spliceosomal components by the lengths and sequences of introns and exons and possibly by the secondary structure of the mRNA, (3) the presence of enhancer or repressor elements that may serve, respectively, to activate a weak splice site or repress a strong one, and (4) the effective concentrations of splicing factors such as SR proteins and heterogeneous nuclear ribonucleoproteins that can be regulated through post-translational modifications such as phosphorylation [15]. The extensive literature elaborating on the mechanisms of repression and activation of specific splice sites has been reviewed by Smith and Valcarcel [16] and more recently by Matlin et al. [17].

The high estimates of the frequency of alternative splicing in human [8,14,18] and mouse [13] genes raise the question to

Editors: Judith Blake (The Jackson Laboratory, US), John Hancock (MRC-Harwell, UK), Bill Pavan (NHGRI-NIH, US), and Lisa Stubbs (Lawrence Livermore National Laboratory, US), together with *PLoS Genetics* EIC Wayne Frankel (The Jackson Laboratory, US)

Received August 16, 2005; **Accepted** February 10, 2006; **Published** April 28, 2006

DOI: 10.1371/journal.pgen.0020045

Abbreviations: CDS, coding region; NMD, nonsense-mediated decay; UTR, untranslated region; WM, weight matrix

* To whom correspondence should be addressed. E-mail: mihaela.zavolan@unibas.ch

☯ These authors contributed equally to this work.

Synopsis

It has recently become clear that splice variation affects most mammalian genes. It is, however, less clear to what extent these splice variations are functional and regulated by the cell as opposed to simply a result of noise in the splicing process.

One of the most frequently observed forms of splice variation are small variations in exon length in which the boundary of an exon is shifted by small amounts between different transcripts. In this work the authors study the statistics of these splice variations in detail, and the results suggest that these variations are mostly the result of noise in the splicing process. In particular, they propose a simple physical model in which the last step of splicing involves the sequence-specific binding of the splicing machinery to the splice site. In this model, small length variations can occur when there are nearby splice sites with comparable affinity for the splicing machinery. The authors show that this model not only accurately predicts the relative abundances of different splice variations but also predicts which splice sites are likely to undergo small exon length variations.

what extent these splice variations are functional, with their production controlled and regulated by the cell, versus being the result of inherent noise in the molecular process of splicing. The molecular mechanisms mentioned in the previous paragraph are all susceptible to noise, e.g., thermodynamic noise, fluctuations in the concentrations of splicing factors, fluctuations in elongation rates, and any other fluctuations that are not under the control of the cell. It is thus clear that some of the splice variation observed in the sequence data might simply be a result of noise [19]. At the same time, one can easily imagine that many of the molecular mechanisms just mentioned could be exploited by the cell to regulate the expression of different splice variants under different conditions.

In a previous study [13], we found that the second most common form of splice variation (after "cassette" or "alternative" exons that are included in some but not all transcripts) is a small change in exon length due to the use of closely spaced alternative donor or acceptor sites. Intuitively, the simplest explanation for these abundant small exon length variations is that they are a result of noise in the splicing process that causes the spliceosome to "slip" by a small number of nucleotides, perhaps to a competing neighboring splice site. However, as we previously reported [13], three-nucleotide variations at tandem acceptor sites are by far the most common among these small exon length variations, and are much more common than any other exon length variation. This seems to suggest that processes other than simple noise must be causing these small in-frame shifts. Indeed, in this context Hiller et al. [20] have proposed an intriguing hypothesis, namely, that splice variations involving only three nucleotides at so-called NAGNAG tandem acceptor sites are introduced in a regulated manner to "fine-tune" the protein sequence. More recently, these three-nucleotide splice variations at NAGNAG sites have attracted considerable attention [21], including two papers [22,23] that appeared after our submission of the current work.

Here we extensively study the statistics of small exon length variations. We revisit our original hypothesis that these small exon length variations are a result of noise in the splicing process. In particular, we show that a combination of the effects of nonsense-mediated decay (NMD) and a simple physical model of the splicing machinery binding in a stochastic manner to nearby splice sites can efficiently explain all the observed statistics. In addition, we show that our physical model can predict which NAGNAG tandem acceptor sites are likely to undergo alternative splicing, which will splice exclusively at the first NAG, and which will splice only at the second NAG.

Results

Splice Variations at Acceptor and Donor Sites

The use of alternative splice donor and acceptor sites leads to exons whose length varies between transcripts. By far the most common variation of this kind is a difference of precisely three nucleotides at acceptor sites. To investigate the origin of such variations we selected all exons that showed variation at only one of their two splice sites, i.e., only at their acceptor site or only at their donor site. For each exon with an alternative acceptor (or donor) site we chose the most common splice site as a reference site and counted the total number of alternative splice events at different distances from the reference site. Figure 1 shows the distribution of distances for both acceptor and donor sites, calculated separately for coding, untranslated region (UTR), and noncoding exons.

The first thing to note is that the total number of alternative acceptor sites is larger than the overall number of alternative donor sites. This observation is consistent with our previous reports on the FANTOM2 dataset [13] as well as

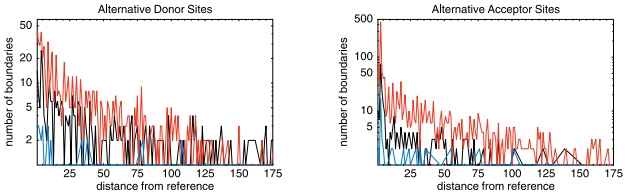


Figure 1. The Number of Splice Events Involving Alternative Donor and Acceptor Sites at a Specified Distance Relative to the Reference (Most Commonly Used) Splice Site

The horizontal axis shows the distance from the reference splice site corresponding to each genomic exon for both donor sites (left) and acceptor sites (right). The red lines correspond to coding exons, the black lines to UTR exons, and the blue lines to exons from non-protein-coding transcription units. The vertical axis is shown on a logarithmic scale.

DOI: 10.1371/journal.pgen.0020045.g001

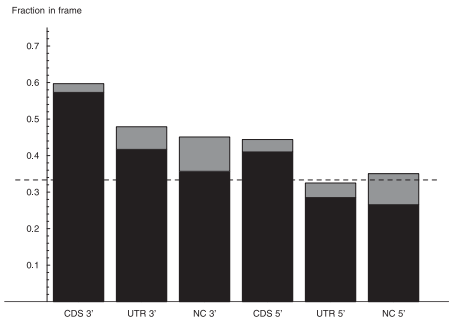


Figure 2. Proportion of In-Frame Variations at Donor and Acceptor Splice Sites That Are Located within CDS, UTR, and Noncoding Regions

This figure shows the fractions of alternative splice events that lead to an in-frame shift with respect to the reference boundary at acceptor (3′) and donor (5′) splice sites of CDS, UTR, and noncoding (NC) exons. The estimated fraction is in the middle of the gray bar, with the gray bar indicating two standard errors. The dashed line shows the fraction 1/3 that would be expected by chance.

DOI: 10.1371/journal.pgen.0020045.g002

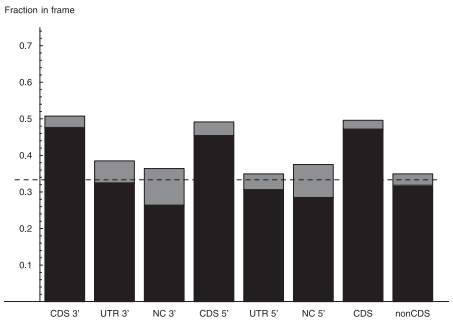


Figure 3. Proportion of In-Frame Variations of More Than Four Nucleotides at Donor and Acceptor Sites Located within CDS, UTR, and Noncoding Regions

This figure shows the fractions of alternative splice events that lead to an in-frame shift with respect to the reference boundary at acceptor (3′) and donor (5′) splice sites of CDS, UTR, and noncoding (NC) exons, when only splice events that are more than four nucleotides shifted with respect to the reference boundary are considered. The two rightmost columns show the fractions when the data from all CDS exons and all non-CDS exons are pooled.

The estimated fraction is in the middle of the gray bar, with the gray bar indicating two standard errors. The dashed line shows the fraction 1/3 that would be expected by chance.

DOI: 10.1371/journal.pgen.0020045.g003

with the observation made by Sugnet et al. [24] that conserved mouse and human alternative acceptor splice sites are twice as common as conserved alternative donor splice sites.

The second thing to note is that small length variations are very common: 23.7% of all donor site variations and 43.7% of all acceptor site variations involve ten or fewer nucleotides. This is suggestive of a "noise" process in which the spliceosome "slides" a few nucleotides from its initially chosen position.

Preference for Reading Frame Preservation and NMD

Most of the exon length variations shown in Figure 1 are not a multiple of three in length, and would therefore have dramatic effects on translation whenever the splice boundary overlaps the coding region (CDS). We will refer to exon length variations as "in-frame" and "frame-shifting" depending on whether the change in exon length is or is not a multiple of three. Figure 2 shows the fraction of in-frame and frame-shifting variations for each category of splice sites. We see that in-frame variations are overrepresented at the acceptor boundaries of all exon types. In contrast, in-frame variations are overrepresented at donor sites only of CDS exons, and they amount to roughly 1/3 of the variations at the donor sites of UTR and noncoding exons.

The different behavior of donor and acceptor sites is the result of the very different distribution of very small exon length variations of 1–4 nucleotides. As we show in detail below, the frequencies of these very small exon length variations at acceptor and donor sites are the result of the different sequence composition of the first few intronic bases at donor and acceptor sites. If we focus on exon length variations of more than four nucleotides, we find that, strikingly, both donor and acceptor splice sites show the same pattern of in-frame variation across exon types (Figure 3). Namely, the frequency of in-frame variations at both donor and acceptor splice sites of noncoding and UTR exons is statistically indistinguishable from 1/3, which is what one would expect by chance. Moreover, CDS exons show the same overrepresentation (approximately 48%) of in-frame variations at both donor and acceptor splice sites.

One possible explanation for the overrepresentation of in-frame variations could be that the sequences flanking CDS exon boundaries are biased such that alternative splice sites occur more often in frame than out of frame. To test this hypothesis we extracted the 100 nucleotides of the intronic sequence flanking the acceptor splice site of each exon that shows length variation at the acceptor boundary and counted the number of times an AG dinucleotide occurs at different distances from the boundary. Similarly, we counted the number of times the dinucleotide GT occurs at different distances from donor sites of exons that show variation at their donor site. We then determined the fraction of times AG and GT occur in frame relative to the acceptor and donor splice sites, respectively. The results are shown in Figure 4. We see that, for both donor and acceptor sites, and for all exon types, the frequency of in-frame occurrence of dinucleotides that could form alternative splice sites is very close to 1/3. It thus appears that biases in the sequence composition flanking CDS exons cannot explain the overrepresentation of in-frame exon length variations at either donor or acceptor sites.

The most plausible explanation for the statistics of the in-frame variations is that NMD removes a fraction of transcripts that have frame-shifting exon length variations in CDS exons. The details of the NMD process are not completely understood, but it is generally thought to function as follows [25,26]. After the splicing process, the exon junction complexes remain attached and are carried along with the transcript. During a preliminary round of translation these complexes are removed by the translation machinery. If any

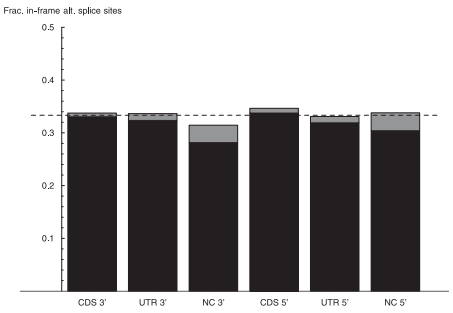


Figure 4. Proportion of Putative Donor (GT) and Acceptor (AG) Splice Sites That Are Located In-Frame Relative to the Splice Sites in CDS, UTR, and Noncoding Regions

This figure shows the fraction of AG dinucleotides that occur at distance that is a multiple of three in the first 100 intronic bases upstream of acceptor (3′) splice sites of exons that show splice variation at their acceptor sites, and the fraction of GT dinucleotides that occur at a distance that is a multiple of three in the first 100 intronic bases downstream of donor (5′) splice sites of exons that show splice variations at their donor sites. Occurrences of AG or GT within the first four bases flanking the splice sites were not counted. The estimated fraction is in the middle of the gray bar, with the gray bar indicating two standard errors. The dashed line shows the fraction 1/3 that would be expected by chance.
DOI: 10.1371/journal.pgen.0020045.g004

such complex remains further than some critical distance away from the stop codon, the transcript is targeted by NMD. Thus, frame-shifting length variation in CDS exons leads to premature stop codons, which in turn increase the chance of the transcript being targeted by NMD. In contrast, the boundaries of noncoding and UTR exons already do not overlap the CDS; therefore, the shifts that occur here do not alter the probability of the transcript being targeted by NMD.

In summary, it is reasonable to assume that some fraction of all frame-shifting exon length variations at CDS exons are targeted by NMD, and that only a fraction f survive NMD. Let us assume that before NMD a fraction ρ_i of all exon length variations at CDS exons are in frame. Since in-frame exon length variations are not affected by NMD, and a fraction f of frame-shifting variations make it past NMD, it follows that the observed frequency ρ_o of in-frame exon length variations in CDS exons is given by

$$\rho_o = \frac{\rho_i}{\rho_i + f(1-\rho_i)}. \quad (1)$$

There is no reason to believe that the fraction ρ_i of in-frame variations before NMD is different from 1/3. This assumption is supported by the fact that the fraction of potential acceptor and donor dinucleotides within 100 nucleotides of exon boundaries is 1/3 for both CDS and non-CDS exons (see Figure 4). We thus assume that $\rho_i = 1/3$ for CDS exons as well. Since $\rho_o = 0.484$ (Figure 3) for CDS exons, it then follows that $f = 0.53$. That is, the data suggest that slightly more than 50% of the transcripts that contain a frame-shifting exon length variation in a CDS exon survive the NMD process.

Exon Length Variations of 1–4 Nucleotides and NAGNAG Acceptor Boundaries

We now turn to the exon length variations of 1–4 nucleotides, whose relative frequencies are shown in Figure 5. To take into account the effects of NMD we have rescaled the numbers of variations of length three by a factor $f = 0.53$, as calculated in the previous section. Moreover, since the frequencies of such variations at CDS, UTR, and noncoding exons are very similar, we have pooled the data for each type

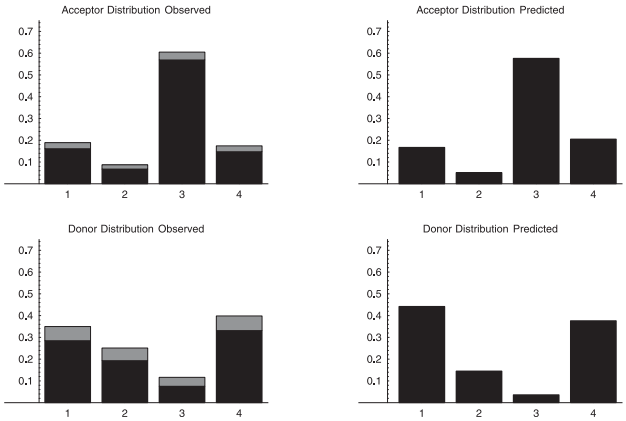


Figure 5. The Distribution of Alternative Splice Events That Are Shifted by One, Two, Three, or Four Nucleotides with Respect to the Reference Splice Site

The two left panels show the observed distributions at acceptor sites (above) and donor sites (below). The estimated relative frequency is in the middle of the gray bar, with the width of the gray bar corresponding to two standard errors. The panels on the right show the predicted relative frequency of alternative splice events of lengths 1–4 based on the splice site WMs and the sequences around exon boundaries that show splice variation.
DOI: 10.1371/journal.pgen.0020045.g005

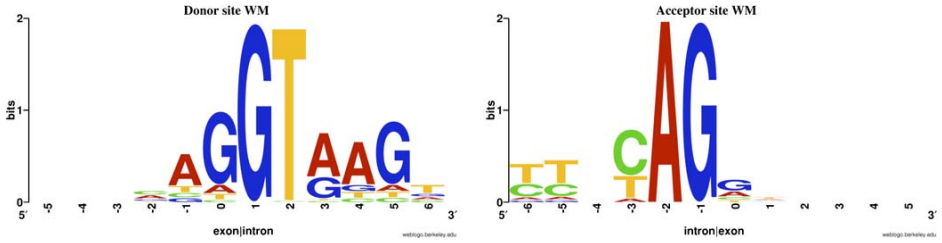


Figure 6. WMs Representing the Sequence Specificity of the Spliceosome at Invariant Donor and Acceptor Splice Sites

WMs have been constructed from six exonic and six intronic nucleotides flanking each type of splice site. The relative sizes of the letters are proportional to the frequency w_α^i of each nucleotide α at position i. The total height in each column is given by the information score $I = \Sigma_\alpha w_\alpha^i \log(4w_\alpha^i)$.

DOI: 10.1371/journal.pgen.0020045.g006

of splice site (Table S1 lists these distributions separately for each exon type). Figure 5 thus shows estimates of the relative frequencies of splice variations of lengths 1–4 nucleotides before NMD, which appear to be very different for donor compared to acceptor splice sites. At donor sites, the most common variations are shifts of one or four nucleotides while in-frame variations of length three are rare. In contrast, at acceptor sites the in-frame variations of length three are highly overrepresented, as we have already reported in our analysis of the FANTOM2 dataset [13]. The very large majority (94%) of these three-nucleotide variations involve tandem acceptor sites whose sequence is of the form NAGNAG. These have been the topic of a recent paper by Hiller et al. [20], who proposed that the role of such variations is to "fine-tune" protein forms by addition/deletion of a single amino acid. That is, Hiller et al. [20] suggested that these NAGNAG sites have been specifically selected to provide the cell with alternative protein forms, and to express these different protein forms in a regulated manner.

If the three-nucleotide variations at NAGNAG acceptor sites were indeed important for fine-tuning protein forms, then one would expect these variations to be more abundant at CDS than at non-CDS exons. However, once we correct for the overrepresentation of *all* in-frame variations due to NMD, this is not what we observe. As shown in Table S1, the relative frequency of three-nucleotide variations is not significantly different at CDS, UTR, and noncoding exons. In addition, the NAGNAG sequence motif is not overrepresented at acceptor sites of CDS exons. On the contrary, the frequency of NAGNAG sites at the splice boundaries of UTR and noncoding exons is 6.7% compared to only 5.9% at the boundaries of coding exons. This is statistically significantly lower in a χ^2 test at a p-value of 0.00014. Thus, NAGNAG sequences are in fact a little less frequent at the acceptor sites of CDS exons than at those of non-CDS exons.

We next investigated the evolutionary conservation of NAGNAG acceptor sites. If the NAGNAG sites that show splice variation were explicitly selected to do so, one would expect them to be better conserved evolutionarily than NAGNAG sites in exons that show no splice variation. To test this hypothesis, we extracted the human sequences that correspond to NAGNAG sites in mouse from the pairwise mm5–hg17 alignments provided by the University of California Santa Cruz Genome Bioinformatics group (http://hgdownload.cse.ucsc.edu/goldenPath/mm5/vsHg17/axtNet).

We found that the proportion of mouse NAGNAG acceptor sites that have corresponding human NAGNAG sites (allowing the nucleotides at the N positions to vary) is in fact slightly higher in invariant exons (59.5%) than in variant exons (54.7%) (3,998/6,715 versus 273/499, χ^2 test, $p = 0.04$). The frequency of perfectly conserved NAGNAG sites (where the nucleotides at the N positions are conserved as well) is 45.7% in invariant exons and 42.7% in variant exons, which statistically is not significantly different. Thus, NAGNAG sites of variant exons are not more conserved in evolution than NAGNAG sites of invariant exons. If anything they are less conserved than NAGNAG sites of invariant exons. We thus cannot find any evidence that the abundance of three-nucleotide variations at acceptor sites is due to specific selection, or that it is related to the coding potential of the transcript.

We instead consider the much simpler hypothesis that the small exon length variations are a result of inherent noise in the splicing process. We hypothesize that the final step in the process of splice site selection involves the sequence-specific binding of the splicing machinery to the mRNA. This binding process is subject to thermal noise just like any physicochemical process. Whenever multiple "binding sites" with comparable affinity occur near each other, the splicing machinery may bind to these alternative sites in a stochastic fashion, and this will lead to small shifts of the splice site.

To test this hypothesis we first constructed computational models of the sequence specificity of the splicing machinery at acceptor and donor sites. Specifically, we gathered acceptor and donor boundaries of *invariant* exons and reconstructed weight matrices (WMs) of length 12 from the six intronic bases and the six exonic bases flanking each boundary. These WMs are shown in Figure 6. The WMs demonstrate the well conserved GT dinucleotide immediately following the donor boundary and the AG dinucleotide immediately preceding the acceptor boundary. The WMs also show the known preference for a second GT dinucleotide four nucleotides downstream of the donor site, the polypyrimidine tract 5–6 nucleotides upstream of the acceptor site, and the preference for a cytosine immediately preceding the AG of the acceptor site.

We now assume that the probability of the splicing machinery binding to a particular sequence is proportional to the probability of observing that sequence when sampling from the WM representing that boundary. That is, we assume that the probability $P(i)$ that the splicing machinery will bind

and splice at a particular location i is proportional to the probability of the local sequence $s_{i-6}s_{i-5}...s_{i+4}s_{i+5}$ under the WM

$$P(i) = \prod_{k=-6}^{k=5} w^{k}_{s_{i+k}}, \quad (2)$$

where w^{k}_{α} is the frequency of base α at position k of the WM, and i is the location of the putative splice site.

For the donor site WM we collected all exons that show variation at the donor site and calculated the probabilities $P(i)$ at the positions that are shifted by between one and four nucleotides (either to the left or right) with respect to the observed splice site. By summing all $P(i)$ that correspond to shifts of the same length, we calculated the relative frequencies of length variations of lengths 1–4 that our model predicts. These predictions are shown in the lower right panel of Figure 5. In the same way, using the acceptor site WM and the sequences flanking acceptor sites in exons that show variation at their acceptor site, we calculated the relative frequencies of variations of lengths 1–4 that our model predicts at acceptor sites. These predictions are shown in the upper right panel of Figure 5. Given the simplicity of our model, its predictions of the relative abundances of different length variations match the data surprisingly well. At the donor sites, it correctly predicts that shifts of length one and four are the most abundant and that shifts of length three are the least abundant. At the acceptor sites, the model predicts relative abundances that are quantitatively very close to the observed abundances. In particular, the predicted abundance of three-nucleotide variations is within 1% of the observed abundance. We thus see that a very simple model of splice site selection based on the sequence specificity of the splicing machinery at each splice site can correctly predict the relative abundances of small exon length variations. This further supports the hypothesis that these small exon length variations are mostly the result of inherent noise in the splicing process.

Local Sequence Distinguishes Variant Tandem Acceptor Sites from Nonvariant Tandem Acceptor Sites

If our model for the small exon length variations is correct, it should also be possible to predict, from the sequence, which acceptor sites are most likely to be prone to three-nucleotide length variations. To this end we focused on all acceptor sites that show the sequence pattern NAGNAG at their splice site. We collected all acceptor sites with sequence NAGNAG (irrespective of which of the two NAG sequences is used as splice site) and then selected only those sites for which there are at least two transcripts in the data. We then counted, for each NAGNAG site, how many times we observed splicing at the first and how many times at the second NAG site. Based on these counts we separated the NAGNAG sites into three categories: those that splice only at the first NAG, those that splice only at the second NAG, and those that splice at both NAGs.

We then investigated to what extent we could predict the category of each NAGNAG site by using the WM constructed from the invariant acceptor sites. We again assumed that the binding affinity of the splicing machinery to a putative acceptor site sequence is proportional to the log-likelihood of the putative acceptor site sequence given the acceptor site WM. If we additionally assumed that the probability of splicing occurring at the different NAG sites is proportional to the equilibrium frequencies with which the splicing machinery binds at these sites, then the category of a NAGNAG site is only a function of the difference in log-likelihood of the neighboring acceptor sites.

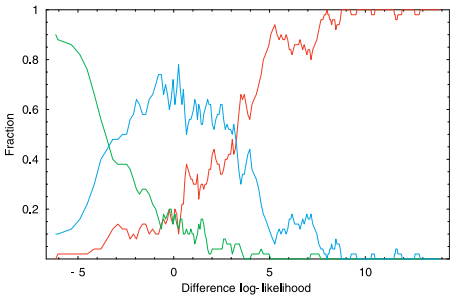


Figure 7. Dependency of the Frequency of Alternative Splicing at NAGNAG Sites on the Relative Likelihood of the Two Putative Acceptor Sites

The figure shows the fraction of all NAGNAG boundaries that splice only at the first NAG (red), only at the second NAG (green), or at both NAGs (blue) as a function of the log-likelihood difference of the first and second putative splice sites for the acceptor site WM.

DOI: 10.1371/journal.pgen.0020045.g007

Figure 7 shows the fractions of NAGNAG sites that splice only at the first NAG, only at the second NAG, or at both NAGs as a function of the log-likelihood difference of the two sites. The results are quite striking. We see that, using the simple WM model, one can reasonably accurately predict which NAGNAG sites splice only at the first NAG, which splice only at the second, and which show three-nucleotide length variations. Whenever the log-likelihood of the first site is larger by five or more than the log-likelihood of the second site, then splicing virtually always occurs at the first site only. Similarly, if the log-likelihood of the second site is larger by five or more than the log-likelihood of the first site, then splicing virtually always occurs at the second site. When there is almost no difference in the log-likelihood of the two sites, then in the large majority of examples one observes splicing at both sites. These results further support our hypothesis that splice site selection is simply based on the affinity of the splice site for the splicing machinery, and that small exon length variations occur whenever there are neighboring splice sites with comparable affinity.

Discussion

Recent large-scale sequencing efforts have made clear that the great majority of mammalian genes are subject to splice variation, and that some genes show a very large number of different transcript forms. One of the most basic questions to ask about these splice variations is to what extent they are regulated by the cell rather than being the result of noise in the molecular process or of other fluctuations that are not controlled by the cell. We investigated this issue for small exon length variations caused by the alternative usage of nearby acceptor and donor sites.

Small exon length variations are the second most common form of splice variation. Among these, three-nucleotide variations at tandem acceptor sites containing a NAGNAG sequence pattern are by far the most common. It has been

suggested that cells can "fine-tune" the expression of protein forms through these subtle splice variations [20]. However, we were unable to find any evidence that NAGNAG sites are either under specific evolutionary selection or are more abundant at coding exons than at noncoding exons. In contrast, we collected and compared a number of statistics on observed small exon length variations, and showed that all these statistics can be accurately explained by the combined effect of NMD and the stochastic binding of the splicing machinery to competing nearby splice sites.

Our analysis of the relative abundances of exon length variations of five and more nucleotides at donor and acceptor sites of CDS, UTR, and noncoding exons strongly suggests that NMD targets about 50% of frame-shifting exon length variations at CDS exons. This effect partly explains the overrepresentation of three-nucleotide variations in the data. When the effects of NMD are taken into account, donor and acceptor sites show very distinct relative abundances of small exon length variations, with an overrepresentation of shifts of length one and four at donor sites, and an overrepresentation of shifts of length three at acceptor sites. However, shifts of length three at acceptor sites are not more abundant at CDS exons than they are at UTR and noncoding exons. This observation again supports our conjecture that the abundance of three-nucleotide shifts is special to the specifics of splicing at acceptor sites and is not related to the coding potential of the exon.

To explain the relative abundances of these small exon length variations we introduced a model that assumes that the last step in splice site selection involves the sequence-specific binding of the splicing machinery to a splice site and that the probability of splicing occurring at a particular site is proportional to the affinity of the splice site sequence for the splicing machinery. In this model, small exon length variations can occur when there are multiple binding sites with comparable affinity near each other. Our model is similar to the scanning model proposed by Smith et al. [27] for the selection of 3′ splice sites: the spliceosome recognizes and binds the region of the branch point and polypyrimidine sequence, but once there, it can still "see" a limited stretch of sequence in which competing splice sites may be present. The relative frequencies with which these competing splice sites are used in the splicing reaction are determined by the affinity of the spliceosome for the sequences of the different sites.

From the splice sites of invariant exons we constructed WMs representing the sequence specificity of the splicing machinery at donor and acceptor sites. We used these WMs to predict the relative frequencies of small exon length variations from the sequences flanking the observed splice sites of variant exons and found that the predictions accurately reproduce the observed relative abundances. For example, the relatively high abundance of shifts of four nucleotides at donor sites is explained by the common occurrence of a GT pattern at position +5 of the intron [13,28]. The relatively high abundance of three-nucleotide variations at acceptor splice sites is explained by the fact that the WM of the acceptor sites strongly disfavors a guanine at position −3 (directly upstream of the AG) and that positions −5, −6, and further upstream are part of the polypyrimidine tract that disfavors the occurrence of purines in general. Therefore, whenever an alternative AG dinucleotide does occur, it is almost always at position −3 and not at −1, −2, or −4.

Apart from explaining the relative abundances of small exon length variations at both donor and acceptor sites, our simple model also predicts, with reasonable accuracy, which NAGNAG acceptor sites show splice variation and which do not. We showed that when one of the two neighboring sites has much higher affinity than the other, one observes splicing almost exclusively at the site with the higher affinity. When the neighboring sites have similar affinity, three-nucleotide variations are observed in the large majority of cases.

Materials and Methods

Transcript mapping and splice analysis. The transcripts used in the study consist of the 102,797 FANTOM3 mouse full-length cDNAs [29] and 52,070 GenBank mouse mRNAs. We mapped all these transcripts to the mm5 assembly of the mouse genome available from the University of California Santa Cruz.

For the identification of splice variants we used a new implementation of the automated splicing analysis pipeline that we developed for the FANTOM2 project [13]. Briefly, we first mapped all cDNAs to the mouse genome using our novel spliced alignment algorithm, SPA [30], which produces better quality alignments than other commonly used cDNA-to-genome alignment programs. In particular, it has fewer alignment errors around splice boundaries, and has a better coverage of the 5′ and 3′ ends of the cDNAs. The details of the comparisons of SPA's mappings to those of other methods are described in van Nimwegen et al. [30].

To avoid biases from transcripts that are badly mapped we selected only those transcripts that had at least 75% of their nucleotides mapped to the genome, with at least 95% identity or fewer than ten mismatches in each exon. This procedure yielded 129,655 mapped transcripts, which we clustered such that the mapping of each transcript in a cluster shared at least one exonic nucleotide on the same strand with at least one other transcript in the cluster [11,13]. We obtained 42,023 clusters (transcription units) that we analyzed for splice variation. We refer the reader to Zavolan et al. [13] for the details of the annotation procedure. Briefly, all exons whose genomic mappings overlap were clustered into "genomic exons." For each genomic exon we then compared the set of exons corresponding to it to identify splice variation.

For our analysis of exon length variations we extracted all genomic exons that show only variation at their donor splice site and all exons that show only variation at their acceptor splice site. For each such genomic exon we then extracted the set of all "clean" exons corresponding to it. These "clean" exons were selected based on the following criteria: (1) the first and last ten nucleotides of every clean exon must be perfectly aligned (no mismatches or gaps) to the genome, and (2) the first ten nucleotides of the flanking exon(s) must be perfectly aligned. We used the FANTOM3 and GenBank annotation of CDSs to separate the exon boundaries of the clean exons into boundaries that overlap with the CDS, boundaries that are located in the UTRs of transcripts that have a CDS annotated, and boundaries of exons from noncoding transcription units. A noncoding transcription unit has no CDS annotated for any of its corresponding transcripts.

For each genomic exon with variation only at the acceptor site or only at the donor site we then determined the number of times each alternative boundary was observed and took the most abundant boundary as the "reference boundary." We then counted the total number of times other boundaries were observed for each of these exons, and recorded the distances of these alternative boundaries from the reference boundary. Finally, we constructed from these counts the histograms of the number of observed exon length variations as a function of the distance to the reference boundary for each boundary type and each class of exon. The total numbers of observed exon length variations for donor sites were 871 in CDS exons, 524 in UTR exons, and 117 in noncoding exons, and for acceptor sites were 1,620 in CDS exons, 366 in UTR exons, and 109 in noncoding exons.

The relatively low number of variations in noncoding exons is a result of the fact that transcripts from noncoding transcription units in general have far fewer exons than do coding transcripts. In addition, there are many transcripts for which no CDS is annotated but that occur in a transcription unit that does contain at least one transcript with an annotated CDS. We exclude these transcripts because it is unclear whether they are indeed noncoding or whether their CDS has simply not been annotated.

For every exon that has variation only at its acceptor boundary we collected all transcripts that contain this exon and extracted the first 100 intronic nucleotides upstream of the acceptor site for this exon. Similarly, for every exon with variation only at its donor boundary we collected all transcripts that contain this exon and extracted the first 100 nucleotides downstream of the donor site in each of these transcripts.

Selection of the sequences used for constructing the WMs for donor and acceptor splice sites. We used the full set of clean invariant exons to extract sequences at the acceptor and donor splice sites. We removed all sequences that contained ambiguous characters and obtained a set of 130,827 sequences in each set. We wanted to use the sequences of these invariant exons to construct WMs for the acceptor and donor site sequences. However, since we also wanted to score NAGNAG sites for these WMs, we split the set of all invariant exons into two halves and used only the first half to construct the WMs of the six intronic and six exonic nucleotides flanking the splice site. The other half we used for extracting NAGNAG sites of invariant exons.

Extraction of NAGNAG acceptor splice sites and log-likelihood histogram. From the second half of acceptor site sequences of invariant exons just described we collected all boundaries that have a NAGNAG sequence. There were 2,444 cases of NAGNAG invariant exons that splice at the first NAG and 228 cases of NAGNAG invariant exons that splice at the second NAG. The set of variant exons with NAGNAG sites consisted of all clean exons that contain a NAGNAG motif at their acceptor site and whose only splice variation involves the use of alternative acceptor sites precisely at the NAGNAG boundary. We obtained 404 exons with such variant NAGNAG acceptor sites.

For each of these NAGNAG sites we calculated the difference in log-likelihood of the tandem putative acceptor sites for the acceptor site WM. We then ordered all 3,076 NAGNAG sites by the log-likelihood difference (from small to large) and calculated the average log-likelihood difference and fraction of sites variant, invariant splicing at the first NAG, and invariant splicing at the second NAG in consecutive groups of 50 sites. That is, the horizontal value of the leftmost data points of the red, green, and blue curves in Figure 7 were obtained by averaging the log-likelihood differences of the first 50 NAGNAG sites, and the vertical values were obtained by calculating the fraction of variant NAGNAG sites (blue), invariant NAGNAG sites splicing at the first boundary (red), and invariant NAGNAG sites splicing at the second boundary (green) among those first 50 NAGNAG sites. Similarly, the second leftmost set of data points was obtained by calculating the same averages and fractions over NAGNAG sites 11 through 60, the third set of points over NAGNAG sites 21 through 70, etc.

Supporting Information

Table S1. The Relative Frequencies and Two Standard Errors of Exon Length Variations of Length 1–4 at Donor and Acceptor Sites of Different Exon Types

Note that, in order to correct for NMD, the number of variations of length three has been multiplied by 0.53 and rounded to the nearest integer.

Found at DOI: 10.1371/journal.pgen.0020045.st001 (11 KB PDF).

Acknowledgments

Author contributions. EvN and MZ conceived and designed the experiments. TMC, EvN, and MZ performed the experiments and analyzed the data. TMC, EvN, CK, JK, PC, YH, and MZ contributed reagents/materials/analysis tools. EvN and MZ wrote the paper.

Funding. TMC thanks the South African National Research Foundation for their doctoral fellowship support. YH acknowledges the support through the research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan, a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan, a grant for the Strategic Programs for R&D of RIKEN, and research grants for Preventure Program C of Japan Science and Technology Agency (JST).

Competing interests. The authors have declared that no competing interests exist. ∎

References

1. Lander E, Linton L, Birren B, Nusbaum C, Zody M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
2. Venter J, Adams M, Myers E, Li P, Mural R, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
3. Waterston R, Lindblad-Toh K, Birney E, Rogers J, Abril J, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
4. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, et al. (2001) Functional annotation of a full–length mouse cDNA collection. Nature 409: 685–690.
5. Suzuki Y, Taira H, Tsunoda T, Mizushima–Sugano J, Sese J, et al. (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep 2: 388–393.
6. Gerhard D, Wagner L, Feingold E, Shenmen C, Grouse L, et al. (2004) The status, quality, and expansion of the NIH full–length cDNA project: The Mammalian Gene Collection (MGC). Genome Res 14 (10B): 2121–2127.
7. Pruitt K, Tatusova T, Maglott D (2005) NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–D504.
8. Mironov A, Fickett J, Gelfand M (1999) Frequent alternative splicing of human genes. Genome Res 9: 1288–1293.
9. Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29: 2850–2859.
10. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. Nat Genet 30: 29–30.
11. Zavolan M, van Nimwegen E, Gaasterland T (2002) Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. Genome Res 12: 1377–1385.
12. Beaudoing E, Gautheret D (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res 11: 1520–1526.
13. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume D, et al. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13: 1290–1300.
14. Johnson J, Castle J, Garrett-Engele P, Kan Z, Loerch P, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302: 2141–2144.
15. Stoilov P, Daoud R, Nayler O, Stamm S (2004) Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. Hum Mol Genet 13: 509–524.
16. Smith C, Valcarcel J (2000) Alternative pre-mRNA splicing: The logic of combinatorial control. Trends Biochem Sci 25: 381–388.
17. Matlin A, Clark F, Smith C (2005) Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Biol 6: 386–398.
18. Croft L, Schandorff S, Clark F, Burrage K, Arctander P, et al. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat Genet 24: 340–341.
19. Kan Z, States D, Gish W (2002) Selecting for functional alternative splices. Genome Res 12: 1837–1845.
20. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, et al. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet 36: 1255–1257.
21. Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, et al. (2005) Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: The case of Gln in DRPLA affects subcellular localization of the products. J Hum Genet 50: 382–394.
22. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, et al. (2006) Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. Am J Hum Genet 78: 291–302.
23. Martin A, Yael M (2006) Alternative splicing regulation at tandem 3′ splice sites. Nucleic Acids Res 34: 23–31.
24. Sugnet C, Kent W, Ares M, Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput 2004: 66–77.
25. Le Hir H, Gatfield D, Izaurralde E, Moore M (2001) The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J 20: 4987–4997.
26. Maquat L (2004) Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5: 89–99.
27. Smith C, Chu T, Nadal-Ginard B (1993) Scanning and competition between AGs are involved in 3′ splice site selection in mammalian introns. Mol Cell Biol 13: 4939–4952.
28. Iida Y (1985) Splice-site signals of mRNA precursors as revealed by computer search. Site-specific mutagenesis and thalassemia. J Biochem 97: 1173–1179.
29. Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
30. van Nimwegen E, Paul N, Sheridan R, Zavolan M (2006) SPA: A probabilistic algorithm for spliced alignment. PLoS Genet 2: e24. DOI: 10.1371/journal.pgen.0020024

5.2 Publication for Chapter 3

DNA Research Advance Access published February 14, 2008

DNA Research pp. 1–10, (2008) doi:10.1093/dnares/dsm036

Computational Analysis of Full-length cDNAs Reveals Frequent Coupling Between Transcriptional and Splicing Programs

Tzu-Ming Chern, Nicodeme Paul, Erik van Nimwegen, and Mihaela Zavolan*

Division of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50-70, Basel CH-4056, Switzerland[4]

(Received 31 August 2007; accepted 25 December 2007)

Abstract

High-throughput sequencing studies revealed that the majority of human and mouse multi-exon genes have multiple splice forms. High-density oligonucleotide array-based measurements have further established that many exons are expressed in a tissue-specific manner. The mechanisms underlying the tissue-dependent expression of most alternative exons remain, however, to be understood. In this study, we focus on one possible mechanism, namely the coupling of (tissue specific) transcription regulation with alternative splicing. We analyzed the FANTOM3 and H-Invitational datasets of full-length mouse and human cDNAs, respectively, and found that in transcription units with multiple start sites, the inclusion of at least 15% and possibly up to 30% of the 'cassette' exons correlates with the use of specific transcription start sites (TSS). The vast majority of TSS-associated exons are conserved between human and mouse, yet the conservation is weaker when compared with TSS-independent exons. Additionally, the currently available data only support a weak correlation between the probabilities of TSS association of orthologous exons. Our analysis thus suggests frequent coupling of transcriptional and splicing programs, and provides a large dataset of exons on which the molecular basis of this coupling can be further studied.

Key words: alternative splicing; transcription initiation

1. Introduction

The most common form of splice variation is the inclusion of an exon in some, but not all, of the transcripts of a gene.[1,2] Numerous studies have been dedicated to specific instances of such exons, which are known by various names such as 'cassette', 'alternative', 'skipped', and 'cryptic' exons. The regulatory signals leading to the inclusion or exclusion of a cassette exon also form a vast topic of research. Computational studies are converging toward the view that cassette exons are generally less recognizable to the splicing machinery than constitutive exons due to their shorter length,[3] lower strength of splice sites,[4] and poor representation of general splice enhancers,[1,5] The tissue-specific inclusion of these exons appears to be dependent upon specific regulatory elements, at least some of which are located in the strongly conserved intronic regions that flank the cassette exons.[2,6–8]

One attractive hypothesis concerning the mechanism of tissue-specific inclusion of cassette exons involves the direct coupling between tissue-dependent transcription and splicing. It has been shown, for instance, that the promoter from which transcription is initiated can affect the inclusion of downstream exons through the recruitment of transcription factors and co-activators that modulate the elongation rate of RNA polymerase II[9,10] (kinetic model). In turn, a low polymerase elongation rate can promote the

Edited by Osamu Ohara

* To whom correspondence should be addressed. Tel. +41 61-267-1576. Fax. +41 61-267-1584. E-mail: mihaela.zavolan@unibas.ch

inclusion of some exons that are skipped when the elongation rate is high, as has been shown for the fibronectin EDI exon.[11] Alternatively, protein factors that are recruited at the stage of transcription initiation may interact with splicing factors (recruitment model). For instance, the inclusion of the fibronectin EDII exon, normally promoted by the SRp40 protein,[12] is inhibited when transcription is initiated at a promoter containing a binding site for the PPARγ transcription factor. This is because the PPARγ transcription factor recruits an SRp40 inhibitor, the PGC-1 co-activator.[13] The role of the promoter architecture on internal splicing has also been demonstrated for the cystic fibrosis transmembrane regulator[14] and for the steroid-sensitive genes.[15] Such studies have been limited, however, to very few genes, prompting us to evaluate the extent to which transcriptional and splicing programs appear to be coordinated at the level of the whole transcriptome.

2. Materials and methods

2.1. Transcript mapping and splice analysis

The sequences used in the study consist of the 102 797 Fantom3 mouse cDNAs,[16] 52 070 mouse mRNAs from Genbank, and 167 992 human cDNAs from version 3.0 (http://www.h-invitational.jp/) of the H-Invitational project.[17] We have mapped the mouse cDNAs to the mm7 assembly of the mouse genome and the human cDNAs to the hg18 assembly of the human genome, both available from the University of California at Santa Cruz.

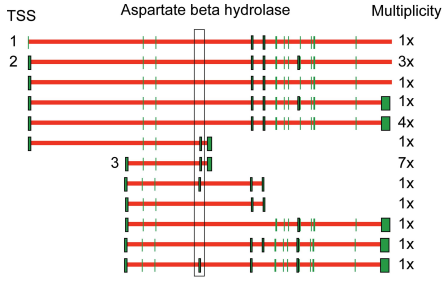
For the identification of splice variants, we used the automated splicing analysis pipeline that we have previously developed.[1,18] Briefly, we first mapped all cDNAs to their respective genome using our spliced alignment algorithm (SPA).[19] To avoid biases from transcripts that are badly mapped due to a high rate of sequencing errors or erroneous assembly, we select only those transcripts that have at least 75% of their nucleotides mapped to the genome, with at least 95% identity or less than ten mismatches in each exon. This procedure yielded 132 681 mouse and 110 978 human mapped transcripts, which we clustered such that the mapping of each transcript in a cluster shares at least one exonic nucleotide with at least one other transcript in the cluster.[1,18,20] We obtained 42 407 mouse and 22 116 human clusters (transcription units) that we analyzed for splice variation. We were interested only in cassette exons with no other form of splice variation. We identified these as internal exons that were completely contained in an intron implied by the mapping of another transcript in the cluster, having the same splice boundaries in all transcripts in which they were included. Our final mouse dataset consisted of 29 416 transcripts and 4 964 internal cassette exons, and the human dataset of 79 030 transcripts and 11 664 internal cassette exons.

2.2. Quantifying the evidence for coupling between the choice of transcription start sites and the inclusion/exclusion of internal exons

For each transcription unit, we first identified (1) the set of internal cassette exons and (2) the set of transcription start sites (TSSs). The cassette exon annotation was determined as outlined above. To identify different TSSs used within a transcription unit, we had to define precisely what we mean by a unique TSS. The analysis of mammalian TSSs by Carninci et al.[21] has shown that most TSSs show some amount of variability. Especially at TSSs located in CpG islands, one finds transcripts starting from many different nearby sites covering tens and sometimes hundreds of nucleotides. We, therefore, needed to group transcripts whose apparent start sites were 'near' each other and then identify different TSSs with the different clusters of apparent start sites. We decided to take a conservative approach to this clustering of apparent start sites in a transcription unit by considering all transcripts that started within the same exon to derive from the same TSS. That is, in our analysis different TSSs correspond to *different initial exons* in the transcripts.

In addition, we tested the validity of our results, on a separate dataset in which we use only transcripts whose initial exons were confirmed by CAGE tag data.[22] In the latter case, the initial exon was considered confirmed as a TSS if one or more CAGE tags were found within 100 bp of the start of the exon in the genome. Since the results did not change, and the requirement of CAGE validation of TSSs reduced the size of our data-set significantly, we did not use the CAGE validated TSSs further.

For each internal cassette exon, we collected all transcripts that could have included the exon as an internal exon, i.e. those transcripts that contained exons both upstream and downstream of the genomic location of the exon in question, and determined the TSS that was used for each of these transcripts. We thus obtained a list of TSSs that were used in the set of transcripts in which the cassette exon could have been included. For further analyses, we kept only cassette exons for which multiple TSSs were identified. We then counted, for each TSS in the list, how many transcripts starting from this TSS included the exon, and how many transcripts excluded the exon. For each internal exon, we thus obtained counts of the number of times each TSS was used in a transcript whose locus covered the



TSS	Exon inclusion	Exon skipping
1	0	1
2	1	9
3	9	3

Figure 1. Example of an internal cassette exon whose inclusion is strongly correlated with specific TSSs. Exons are indicated by (green) boxes and introns by red lines. Cassette exons are shown with a black frame. The internal cassette exon that has a high probability of TSS association is indicated by the thin black rectangle. The number of times a specific TSS-splicing pattern was observed in the data is indicated by the multiplicity on the right. For simplicity, we show only those transcripts whose genomic locus contains the cassette exon, and we truncate the first three transcripts past the TSS-associated cassette exon. Table 1 summarizes the data in the figure, indicating how many times the exon was included and how many times skipped when a particular TSS was used.

exon, and the number of times the exon was included and excluded with each of the TSSs.

To identify exons whose inclusion depends on which TSS was used, we used a Bayesian model selection procedure that compared the probabilities of the observed counts under a TSS-independent model and a TSS-dependent model. Considering a particular cassette exon, let t_i denote the total number of times TSS i was used, n_i the number of times the exon was included when TSS i was used, t the total number of transcripts, and n the total number of times that the exon was included. For the independent model, we assumed that the n inclusions are distributed at random among the t transcripts. Under this model, the probability of the observed counts $\{n_i\}$, given the counts $\{t_i\}$, and n is

$$P_{\text{indep}}(\{n_i\}|n,\{t_i\}) = \frac{n!(t-n)!}{t!}\prod_i \frac{t_i!}{n_i!(t_i-n_i)!}. \quad (1)$$

For the dependent model, we assumed that the rates of inclusion and exclusion for the different TSSs are set by some unknown mechanism. Given our general ignorance about the mechanism or mechanisms determining these rates, there is no reason to assume that any set of counts $\{n_i\}$ is more or less likely than any other set of counts. We, therefore, assumed that all possible counts $\{n_i\}$ that are consistent with the totals $\{t_i\}$ and the total number of inclusions n are all equally likely, meaning that

$$P_{\text{dep}}(\{n_i\}|n,\{t_i\}) = \frac{1}{C(n,\{t_i\})}, \quad (2)$$

where $C(n, \{t_i\})$ is the total number of different sets of counts $\{n_i\}$ that are possible, given the totals n and $\{t_i\}$. The total count numbers $C(n, \{t_i\})$ can be determined recursively. Let $C_i(r)$ be the number of different inclusion counts n_1 through n_i that can be assigned to TSSs 1 through i, such that r of the n total inclusions remain. We have the following recursion relation for $C_i(r)$:

$$C_i(r) = \sum_{n_i=0}^{t_i} C_{i-1}(r+n_i). \quad (3)$$

We initialize the recursion by setting

$$C_0(r) = \delta_{rn}, \quad (4)$$

that is, before we assign counts to any TSS there have to be precisely n inclusions left. Once we arrive at the last (pth) TSS, our count $C(n, \{t_i\})$ is given by $C_p(0)$, i.e. there should be no inclusions left.

To estimate the total fraction f of cassette internal exons whose inclusion is dependent on TSSs, we calculated the probability $P(D|f)$ of the data of all cassette exons assuming that a fraction f was dependent. Let $P_{\text{indep}}(k)$ and $P_{\text{dep}}(k)$ denote the probabilities of the counts for the kth cassette exon given the independent and dependent model, respectively. For each exon k, these quantities are computed according to Equations (1) and (2). We then have

$$P(D|f) = \prod_k \left[P_{\text{indep}}(k)(1-f) + P_{\text{dep}}(k)f\right]. \quad (5)$$

Using a uniform prior over f the posterior probability $P(f|D)$ for the fraction of dependent exons given the data simply becomes

$$P(f|D) = \frac{P(D|f)}{\int P(D|\tilde{f})\mathrm{d}\tilde{f}}. \quad (6)$$

This distribution is shown as the solid line in Fig. 2. We calculated the expectation value $\langle f\rangle = \int f\, P(f \mid D)$

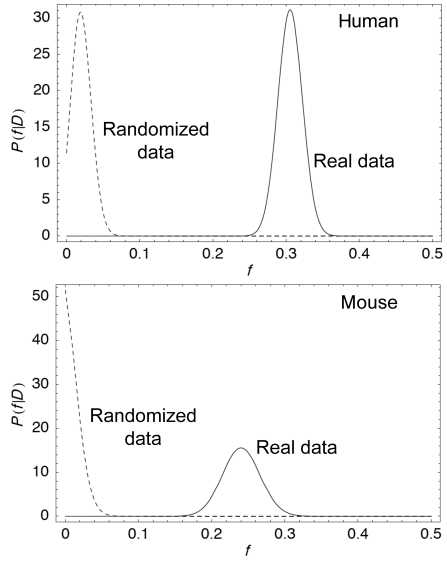


Figure 2. The posterior probability $P(f|D)$ that the inclusion of cassette exons is dependent on the TSS in a fraction f of all cassette exons (solid line). The dashed line shows the same distribution $P(f|D_{rand})$ for a randomized dataset containing the same marginal counts of inclusion and TSS usage for each exon. The upper panel shows the human data, and the lower panel the mouse data.

df and using $\langle f \rangle$ as a prior probability of the independent model we computed, for each individual exon, the posterior probability that the inclusion of the exon is TSS associated.

We also generated a randomized dataset D_{rand} by, for each exon, randomly distributing the n inclusions of the exon among the different TSSs, in such a way that the total number of transcripts t_i for each TSS i stays the same. That is, the data D_{rand} were generated in accordance with the independent model, keeping the total inclusion counts n, and total TSS counts $\{t_i\}$ of the real data. The distribution $P(f|D_{rand})$ is shown as the dotted line in Fig. 2.

2.3. *Extraction of constitutive exons*

We used our database to identify internal exons that were included in more than ten transcripts and did not have any splice variation. We obtained 5136 and 6377 such exons for mouse and human, respectively, and we used these as internal constitutive exons.

2.4. *Computation of the exon conservation statistics*

We used the whole-genome alignments provided by the University of California, at Santa Cruz (http://hgdownload.cse.ucsc.edu/goldenPath/mm7/vsHg18/axtNet/ and http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsmm7/axtNet/) to extract alignments of mouse exons in our dataset with the corresponding orthologous regions from the human genome, and of human exons with the corresponding orthologous regions from the mouse genome. We found orthologous human regions for 438 of the 496 TSS-associated, 469 of the 496 TSS-independent, and 5122 of the 5136 constitutive internal mouse exons. Similarly, we found orthologous mouse regions for 1003 of the 1166 TSS-associated, 1078 of the 1166 TSS-independent, and 6344 of the 6377 constitutive internal human exons. On the basis of the extracted alignments, we determined the fraction of all mouse exon nucleotides that are perfectly conserved in human, and the fraction of human exon nucleotides that are conserved in mouse.

To compute the correlation between posterior probabilities of TSS association in human and mouse, we used the following procedure. We started with cassette exons from transcription units with multiple TSSs from human. On the basis of the whole-enome alignments, we found the orthologous mouse exons as described above, and then we intersected the coordinates of these orthologs with the coordinates of mouse cassette exons that were part of transcription units with multiple TSSs. This procedure gave us the list of orthologous cassette exons that were part of transcription units with multiple TSSs in both human and mouse. We checked that we obtain the same list of exons if we start from the mouse cassette exons and compute their human orthologs. We did not set a threshold of minimal conservation between orthologous exons, as the dataset is already relatively small, and our previous results were not sensitive to the precise threshold of conservation that we used.

2.5. *Computation of the average distance to an inclusion-promoting TSS and to a skipping-promoting TSS*

We identified all transcripts in which a particular cassette exon was included, and for each of them, we determined the distance in the genome between the TSS and the start of the cassette exon. We then averaged these distances to obtain the average distance of the exon to an inclusion-promoting TSS. Similarly, we identified all transcripts in which a particular cassette exon was skipped, and we computed the average distance of the exon to a skipping-promoting TSS.

2.6. *Analysis of 5′ EST data*

To address the issue of biases in the coverage of gene structures that may have been introduced by the selection of clones for full-length cDNA sequencing, we have performed the same analysis using only 5′ EST sequences. On the basis of the October 2007 UniGene database, we extracted 3 738 929 human 5′ EST sequences. We removed from this set those ESTs with low sequence quality (over 3% ambiguous nucleotides), we then trimmed the polyA tails and discarded those ESTs whose length was <95 after polyA tail removal (using the trimpoly program of the SeqClean package from http://compbio.dfci.harvard.edu/tgi/software/). We thus obtained 3 700 028 ESTs, which we mapped to the hg18 assembly of the human genome using our SPA program.[19] After mapping, we retained for further analysis only ESTs, which contained at least two exons, were mapped with overall 95% identity, and whose every exon was mapped with at least 98% identity to the human genome. This selection left 1 276 669 good quality ESTs, which we clustered based on exon overlap and annotated for splice variation as described in Section 2.1.

3. Results and discussion

3.1. *The inclusion/skipping of internal exons is correlated with the usage of specific TSSs*

We used a Bayesian approach to estimate the fraction f of internal cassette exons whose inclusion depends on the choice of TSS. We thus considered two models: the first assumes that the probability of exon inclusion is independent of the TSS used to transcribe the pre-mRNA, i.e. the probability of exon inclusion is the same for all TSSs, and the second assumes that for each TSS there is an independent probability of exon inclusion, which can be different between different TSSs. For a given f, we can write the probability of the data as

$$P(D|f) = \prod_{k\in\text{exons}} \left[P_{\text{indep}}(k)(1-f) + P_{\text{dep}}(k)f\right], \qquad (7)$$

with $P_{\text{dep}}(k)$ and $P_{\text{indep}}(k)$ being the probabilities of the data for cassette exon k under the dependent and independent models, respectively. In order to obtain these probabilities we collected, for each internal exon, the set of transcripts in the dataset that could have contained the exon, i.e. those transcripts that contain exons both upstream and downstream of the genomic location of the cassette exon in question. We divided this set of transcripts into groups that use the same TSSs, and counted the number of transcripts in which the exon was included, and the number of transcripts in which the exon was excluded in each TSS group. A specific example of this computation is shown in Fig. 1.

As described in Section 2.2, we can use Equation (7) and Bayes' theorem to calculate the posterior probability $P(f|D)$ that a fraction f of all cassette exons is dependent on the TSS. The distributions $P(f|D)$ obtained for both the human and mouse data are shown as solid lines in Fig. 2. They suggest that the inclusion of about 24% (99% posterior probability interval 17.4−30.6%) of all cassette exons in our mouse dataset and 30% (99% posterior probability interval 26.2−34.9%) of all cassette exons in our human dataset is dependent on the TSSs of the corresponding transcripts. To additionally test the statistical significance of this result, we created randomized datasets D_{rand} by permuting, for each cassette exon, the inclusions and exclusions among the TSSs in such a way that the total number of times each TSS was used, and the total number of times the exon was included remained unchanged. The posterior distributions $P(f|D_{\text{rand}})$ obtained by applying the Bayesian procedure to the randomized data are shown as dashed lines in Fig. 2. These distributions show that the Bayesian procedure correctly infers that the inclusion of $<5\%$ (and likely none) of the cassette exons in D_{rand} depends on TSS. Moreover, Fig. 2 shows the striking difference between the real and randomized data, which is due to the enrichment of cassette exons with high posterior probability of TSS dependency in the real data.

Using the estimated fraction $\langle f \rangle$ (the mean of the posterior distribution $P(f|D)$) as a prior that the inclusion of a cassette exon depends on TSS, we computed the posterior probability that the inclusion of each exon in our dataset is TSS dependent (see Section 2.2). These data are given in the Supplementary table, and can also be further explored using the server that we established at http://www.spaed.unibas.ch/Promoter_data/TSS_cassette_exons_spaed_human.html and http://www.spaed.unibas.ch/Promoter_data/TSS_cassette_exons_spaed_mouse.html.

Fig. 1 shows the cassette exon with the highest posterior probability of TSS dependence in our mouse data. The exon belongs to the gene aspartate beta hydrolase has a posterior probability of TSS dependency of 0.89, and is indicated in the figure by the thin black rectangle. For clarity, we showed only the transcripts whose genomic loci contain the cassette exon, and we also truncated some of the transcripts after the exon in question. There are three different TSSs upstream of the exon, and a total of 23 transcripts that could have included this exon. Of the 11 transcripts originating in the two upstream TSSs,

only one includes the cassette exon. In contrast, nine of the 12 transcripts originating from the third TSS include the cassette exon.

One may wonder to what extent our results are affected by imperfect efficiency of full length cDNA capture. That is, if a significant fraction of the cDNAs is not full length, the apparent TSSs for these transcripts would be incorrect, and one may wonder how they would influence our results. We have addressed this question by performing the same analysis using only transcripts whose start site was confirmed by CAGE tag data,[22] and obtained essentially the same results. However, since requiring additional confirmation of TSSs substantially reduces the sizes of our datasets, we did not use this dataset further.

It is important to note that we do not need to find the precise locations of the TSSs which may in fact be much less precise than initially thought,[21] but we only need to separate our transcripts into sets that arose from the same transcription initiation regions, controlled by specific sets of regulatory signals. We decided to simply assume that transcripts with the same initial exon arose from the same TSS and that transcripts with different initial exons arose from different TSSs. Two types of errors may occur in this classification. First, transcripts that arose from two different, but nearby TSSs may be assigned to a single common TSS. Second, if a transcript is severely truncated due to cloning or sequencing errors, it may appear to start in an exon which is downstream from its real initial exon. Since the first type of error reduces our ability to distinguish different TSSs, and the second error per definition must be uncorrelated with splicing, the effect of both types of errors will be to *reduce* the correlations between TSS usage and splicing. Therefore, the clear correlations that we observe in spite of these potential errors should be considered to provide a lower bound on the correlations that do exist.

Another concern may be that the gene structures inferred from full-length cDNA are not representative, because the full-length cDNA sequencing projects generally included a prioritization step, that may have caused an apparent enrichment in rare splice variants. To address this issue, we have constructed a database of splice variants using solely human 5′-end ESTs, which we obtained based on the UniGene annotation. We analyzed these data using the same model as we used for full-length cDNAs. The 99% probability interval computed using this dataset was 0.74–0.78, compared with 0–0.009 obtained using the corresponding randomized dataset. This indicates that TSS-associated splice events are in fact even more frequent than initially estimated from full-length cDNA data. The likely reason why the estimate of the fraction of TSS-associated exons is larger when using 5′ EST data compared with full-length cDNA data is illustrated in Fig. 3. The exon indicated by the box belongs to the cAMP-dependent protein kinase catalytic beta subunit (PRKACB), is always skipped when the two upstream promoters are used (136 ESTs), and is generally included with the most downstream promoter (46 of 56 ESTs). These counts are very unlikely under a model in which the promoter usage and exon inclusion are uncoupled. Generally, many exons that in the cDNA data did not have sufficient coverage to allow us to detect their TSS association do have sufficient coverage in the 5′ EST data to allow this inference to be made, and consequently, the fraction of exons inferred to be TSS associated is larger.

3.2. *Evolutionary conservation of TSS association*

If the correlation that we inferred between transcriptional and splicing events is functionally relevant, one would expect that the TSS dependence tends to be conserved between orthologous cassette exons. This could, for instance, manifest itself in a correlation of the posterior probabilities of TSS dependence of orthologous cassette exons. To check this, we started with the human and mouse cassette exons that are part of transcription units with multiple start sites,

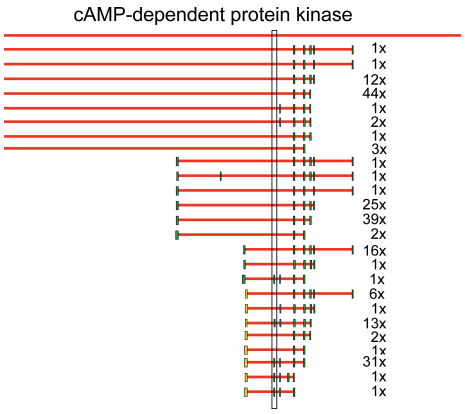


Figure 3. Example of an internal cassette exon whose inclusion is strongly correlated with specific TSSs as inferred from 5′ EST data. The conventions used in this representation are the same as for Fig. 1. Exons shown in green have invariant splice boundaries, those in yellow have variable splice donor sites, and those in cyan intron inclusion. The inclusion of the exon indicated by the black box occurs only when the most downstream TSSs are used.

and we used the UCSC human-to-mouse and mouse-to-human whole-genome alignments to identify orthologous exons (see Section 2.4). This procedure yielded 668 pairs of orthologs. We then computed the correlation coefficient between the posterior probabilities of orthologous exons. As shown in Fig. 4 we obtained a weak, but significant ($P = 0.002$) correlation between the probabilities of TSS association of orthologous exons, providing some evidence that TSS dependence of exon inclusion is evolutionarily conserved.

One of the best examples of an evolutionarily conserved relationship is shown in Fig. 5. The exon with a high probability of TSS association ($P = 1$ in human and $P = 0.73$ in mouse) is indicated by an arrow. It is included in the skeletal form of tropomyosin, which uses the most upstream TSS, and is excluded in other forms of tropomyosin, which also tend to use downstream TSSs.

The level of evolutionary conservation of cassette exons has previously been related to the rate of inclusion of the exons in mature mRNAs: the so-called 'major form' exons, which are predominantly included, are as conserved as constitutive exons, whereas 'minor form' exons, which are predominantly skipped, appear to be of a more recent evolutionary origin,[23] rarely having orthologs between human and rodents. To understand where TSS-associated exons fit in this evolutionary scenario, we analyzed the degree of human–mouse conservation of the following categories of exons: (1) TSS-associated exons—those with the top 10% values of posterior probability of TSS association, (2) TSS-independent exons—those with the bottom 10% values of posterior probability of TSS association, and (3) constitutive exons—exons included with no variation in more than ten transcripts. For each of these exons, we extracted the mouse—human and human–mouse alignments from the whole genome alignments provided by the UCSC (see Section 2.4). We then computed the fraction of exons that have orthologs in the other species and the fraction of nucleotides in each exon that are conserved. As shown in Table 1, we found that the large majority of TSS-associated exons are conserved between mouse and human ($438/496 = 88.3\%$ of mouse and $1003/1166 = 86.02\%$ of human TSS-associated exons). Particularly, TSS-associated exons are much more strongly conserved than 'minor form' exons, only 27–31% of which having been reported to be conserved between human and rodents,[23] However, TSS-associated exons are significantly less conserved than TSS-independent exons ($469/496 = 94.6\%$ in mouse and $1077/1166 = 92.4\%$ in human, P-value of the χ^2 test $= 6.7 \times 10^{-4}$ for mouse and 1.12×10^{-6} for human). These results are not sensitive to the precise threshold beyond which we consider an exon 'conserved'. Among those TSS-associated exons that do have orthologs, the proportion of conserved nucleotides is lower compared with TSS-independent exons (P-value: 2.9×10^{-12} for human and 2.2×10^{-3} for mouse), as well as compared with constitutive exons (P-value of the Wilcoxon test $< 2.2 \times 10^{-16}$ for human and 4.7×10^{-2} for mouse). Consistent with previous results relating the degree of evolutionary conservation to the inclusion rate of the exons,[23] we found that the overall inclusion rate of TSS-associated

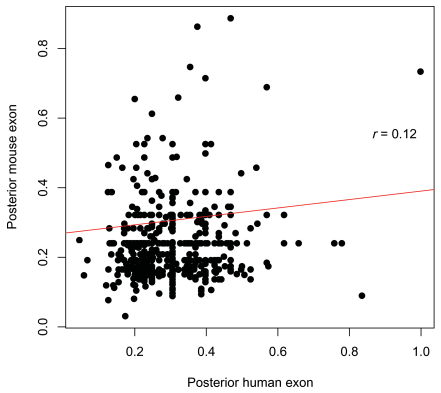


Figure 4. Correlation between the posterior probabilities of TSS association for orthologous human–mouse cassette exons. Each dot represents one exon, with the *x*-coordinate being the posterior probability of TSS association of the human exon and the *y*-coordinate being the the posterior probability of TSS association of the orthologous mouse exon. The correlation coefficient is $r = 0.12$, *P*-value = 0.002.

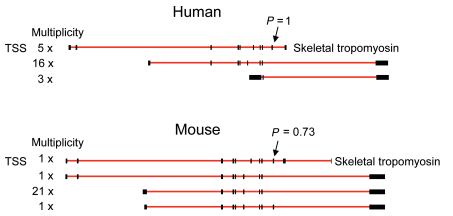


Figure 5. Conserved TSS association for a tropomyosin exon. The intron/exon structure of tropomyosin transcripts initiated from various TSSs is shown, with exons represented as black boxes and introns as red lines connecting the exons. The number of times each transcript form was observed in our dataset is indicated by the 'multiplicity' column on the right-hand side of the transcripts. The orthologous cassette exons are indicated by arrows, and their posterior probabilities of TSS association are indicated.

Table 1. Comparison of TSS associated, TSS independent, and constitutive exons

Data set	Number exons	With orthologs	Median proportion conserved nucleotides	Inclusion rate	Proportion symmetrical
Human exons					
TSS associated	1166	1003	0.84	0.75	0.41
TSS independent	1166	1077	0.87	0.86	0.45
Constitutive	6377	6343	0.88	1	0.38
Mouse exons					
TSS associated	496	438	0.86	0.75	0.42
TSS independent	496	469	0.88	0.88	0.48
Constitutive	5136	5117	0.87	1	0.4

exons is lower than the inclusion rate of TSS-independent exons (Table 1). Thus, the results suggest that some of the TSS-associated exons are of relatively recent evolutionary origin, and it will be interesting to establish whether the TSSs that promote their inclusion have also undergone recent evolutionary changes. On the other hand, the relatively weak correlation between the probabilities of TSS association of orthologous exons is likely to be in part due to the fact that the human and mouse sequencing projects did not cover sufficiently similar sets of tissues. This would be reflected in different relative usage of the alternative TSSs and exons between human and mouse and these, in turn, would be reflected in disparate probabilities of TSS association of orthologous exons in the two species.

Finally, we did not find a consistent trend in the proportion of 'symmetrical' exons, i.e. the proportion of exons whose length is a multiple of three, among our different categories of exons. We did not specifically select the exons for being part of coding regions, but rather we only considered internal exons in our analysis. As shown in Table 1, the proportion of symmetrical exons is significantly higher than expected by chance, i.e. 1/3, for all exon types (TSS-associated, TSS-independent, constitutive). In human, TSS-associated exons have a significantly higher tendency for symmetry compared with constitutive internal exons (*P*-value of the χ^2 test = 0.02), and lower, but not significantly, compared with TSS-independent exons (*P*-value of the χ^2 test = 0.07). The tendencies are similar in mouse, but the differences are not statistically significant (*P*-value of the χ^2 test = 0.36 in the comparison with constitutive exons and 0.097 in the comparison with TSS-independent exons).

3.3 *Insights into the mechanism of coupling between transcriptional and splicing events*

We can envision two mechanisms that could give rise to the correlation that we observe between transcriptional and splicing events. One is that tissue-specific transcriptional and splicing events are induced by independent, but tissue-specific transcription and splicing factors. The second mechanism involves a direct influence of tissue-specific transcription factors on internal splicing. We reasoned that if an exon is always included when one TSS is used and always skipped when another TSS is used, yet the two TSSs are both used in the same tissue, then the TSS association may be due to direct coupling of transcription with exon selection. To identify such cases, we used the library annotation of the transcripts in our datasets. We found that for 21% of the human and 14% of the mouse TSS-associated exons the inclusion- and skipping-promoting TSS have been both used in the same tissue.

One model that has been proposed for the coupling between transcriptional and splicing events is known as the 'kinetic model',[10] which postulates that cassette exons with weaker splice signals compared with constitutive exons tend to be skipped when the transcript is produced by a fast-elongating polymerase. In contrast, when the polymerase elongation rate is low, the spliceosome has sufficient time to assemble on these cassette exons, which are then spliced into the mature mRNA. We used a web server implementing the Shapiro and Senapathy model[24] (http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm) to evaluate the strength of the splice sites of the cassette exons as well as of the exons flanking them. We found that the strength of the splice sites is comparable between the cassette exons and the exons flanking them (data not shown), suggesting that one of the pre-requisites of the kinetic model (weaker splice sites) is not met by our sets of TSS-associated exons.

Given that the elongation rate is not homogeneous along a gene, and that promoter-proximal pausing of RNA polymerase II is common,[25] we reasoned that the elongation rate of the polymerase may be generally lower at the start of the transcripts, allowing better recognition of cassette exons that are close to the TSS. Therefore, we asked whether the TSS-associated exons tend to be included when the TSS closest to

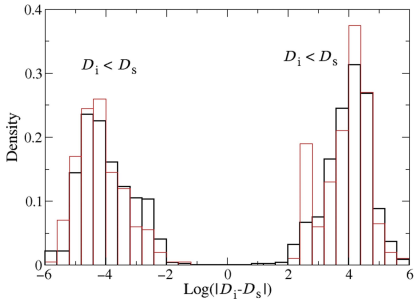


Figure 6. Histogram of the signed difference between the logarithm (base 10) of the average distance to TSS when the exon is included and the average distance to TSS when the exon is skipped. Human data are shown in black and mouse data are shown in red.

them is used, as this may allow sufficient time for spliceosome assembly. We found this not to be the case: for all TSS-associated exons, we computed the average distance D_i between the exon and the TSSs of transcripts that included the exon, and the average distance D_s between the exon and the TSSs of transcripts that excluded the exon. We then constructed a histogram of the difference $D_i - D_s$ across cassette exons (Fig. 6). To improve visibility of the histogram, we split it into a part where $D_i - D_s > 0$ and a part where $D_i - D_s < 0$, and plotted the absolute distance $|D_i - D_s|$ on a logarithmic scale. We found that approximately the same proportion of TSS-associated exons are preferentially included with the most upstream TSSs as are included with the most downstream TSSs (Fig. 6). This suggests that proximity to the TSS, which may be indicative of lower polymerase elongation rate, cannot explain the inclusion pattern of TSS-associated cassette exons.

To conclude, we estimated that exons whose inclusion in the mature mRNA is correlated with specific TSSs are rather common, i.e. they represent at least 15% of all internal cassette exons. The correlation may be due to direct coupling between transcription and splicing or indirect coupling, due, for instance, to a tissue-specific signaling pathway that activates both the transcription factors responsible for determining the TSS as well as the splicing factors responsible for the inclusion or exclusion of the cassette exon. Nonetheless, for at least 14–21% of the TSS-associated exons, TSSs that are associated with exclusion and TSSs that are associated with inclusion occur *both* in the same tissue, suggesting a more direct connection between the TSS and exon inclusion for at least these exons. The details of the molecular mechanism underlying this dependency remain to be uncovered, and may involve a gene-specific component, as suggested by our observation that some exons are predominantly included when the proximal TSSs are used, while other exons are predominantly skipped. One way of implementing a gene-(and tissue-) specific component could be through dynamic changes in the local chromatin structure that induce in turn local variations in the polymerase elongation rate.[26] This is a topic for future study and the cassette exons for which we estimated a high probability of TSS association provide good starting points for experimental investigations.

Supplementary Data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

This work has been supported in part by the European Network for Alternative Splicing (EURASNET). TM Chern is also supported by a fellowship from the South African National Research Foundation. We gratefully acknowledge the VitalIT team of the Swiss Institute of Bioinformatics, in particular Ioannis Xenarios, for computational support.

References

1. Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D., Hayashizaki, Y., et al., 2003, Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome, *Genome Res.*, **13**, 1290–1300.
2. Sugnet, C., Kent, W., Ares, M. and Haussler, D. 2004, Transcriptome and genome conservation of alternative splicing events in humans and mice, *Pac. Symp. Biocomput.*, 66–77.
3. Berget, S. 1995, Exon recognition in vertebrate splicing, *J. Biol. Chem.*, **270**, 2411–2414.
4. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M. 2000, An alternative-exon database and its statistical analysis, *DNA Cell Biol.*, **19**, 739–756.
5. Wang, J., Smith, P., Krainer, A. and Zhang, M. 2005, Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes, *Nucleic Acids Res.*, **33**, 5053–5062.
6. Sorek, R. and Ast, G. 2003, Intronic sequences flanking alternatively spliced exons are conserved between human and mouse, *Genome Res.*, **13**, 1631–1637.
7. Baek, D. and Green, P. 2005, Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing, *Proc. Natl Acad. Sci. USA*, **102**, 12813–12818.
8. Plass, M. and Eyras, E. 2006, Differentiated evolutionary rates in alternative exons and the implications for splicing regulation, *BMC Evol. Biol.*, **6**, 50.

9. Kornblihtt, A. 2005, Promoter usage and alternative splicing, *Curr. Opin. Cell. Biol.*, **17**, 262–268.
10. Kornblihtt, A. 2006, Chromatin, transcript elongation and alternative splicing, *Nat. Struct. Mol. Biol.*, **13**, 5–7.
11. Nogues, G., Kadener, S., Cramer, P., Bentley, D. and Kornblihtt, A. 2002, Transcriptional activators differ in their abilities to control alternative splicing, *J. Biol. Chem.*, **277**, 43110–43114.
12. Lim, L. and Sharp, P. 1998, Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats, *Mol. Cell. Biol.*, **18**, 3900–3906.
13. Monsalve, M., Wu, Z., Adelmant, G., Puigserver, P., Fan, M. and Spiegelman, B. 2000, Direct coupling of transcription and mRNA processing through the thermogenic coactivator PGC − 1, *Mol. Cell*, **6**, 307–316.
14. Pagani, F., Stuani, C., Zuccato, E., Kornblihtt, A. and Baralle, F. 2003, Promoter architecture modulates CFTR exon 9 skipping, *J. Biol. Chem.*, **278**, 1511–1517.
15. Auboeuf, D., Honig, A., Berget, S. and O'Malley, B. 2002, Coordinate regulation of transcription and splicing by steroid receptor coregulators, *Science*, **298**, 416–419.
16. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N. et al. 2005, The transcriptional landscape of the mammalian genome, *Science*, **309**, 1559–1563.
17. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. et al. 2004, Integrative annotation of 21,037 human genes validated by full-length cDNA clones, *PLoS Biol.*, **2**, 856–875.
18. Chern, T., van Nimwegen, E., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. et al. 2006, A simple physical model predicts small exon length variations, *PLoS Genet.*, **2**, e45.
19. van Nimwegen, E., Paul, N., Sheridan, R. and Zavolan, M. 2006, SPA: a probabilistic algorithm for spliced alignment, *PLoS Genet.*, **2**, e24.
20. Zavolan, M., van Nimwegen, E. and Gaasterland, T. 2002, Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome, *Genome Res.*, **12**, 1377–1385.
21. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J. et al. 2006, Genome-wide analysis of mammalian promoter architecture and evolution, *Nat. Genet.*, **38**, 626–635.
22. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M. et al. 2006, CAGE: cap analysis of gene expression, *Nat. Methods.*, **3**, 211–222.
23. Modrek, B. and Lee, C. 2003, Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss, *Nat. Genet.*, **34**, 177–180.
24. Shapiro, M. and Senapathy, P. 1987, RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression, *Nucletic. Acids Res.*, **15**, 7155–7174.
25. Krumm, A., Hickey, L. and Groudine, M. 1995, Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation, *Genes Dev.*, **9**, 559–572.
26. Batsche, E., Yavin, M. and Muchardt, C. 2006, The human SWI/SNF subunit Brm is a regulator of alternative splicing, *Nat. Struct. Mol. Biol.*, **13**, 22–29.

Chapter 6

Bibliography

Bibliography

[1] A.J. Matlin, F. Clark, and C.W. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell. Biol.*, 6:386–398, 2005.

[2] D.A. Brow. Allosteric cascade of spliceosome activation. *Annu.Rev.Genet.*, 36:333–360, 2002.

[3] E. Buratti, M. Baralle, and F.E. Baralle. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucleic Acids Res.*, 34:3494–3510, 2006.

[4] F. Del Gatto-Konczak, C.F. Bourgeois, C. Le Guiner, L. Kister, M.C. Gesnel, Stvenin J., and R. Breathnach. The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5' splice site. *Mol. Cell. Biol.*, 20:6287–6299, 2000.

[5] Lynch K.W. and Maniatis T. Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer. *Genes Dev.*, 10:2089–2101, 1996.

[6] J. Valcarcel, R. Singh, P.D. Zamore, and M.R. Green. The protein sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA. *Nature*, 362:171–175, 1993.

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

[8] A. Mayeda and A.R. Krainer. Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicng factor SF2. *Cell*, 68:365–375, 1992.

[9] I.C. Eperon, O.V. Makarova, A. Mayeda, S.H. Munroe, J.F. Ccere, D.G. Hayward, and A.R. Krainer. Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol. Cell. Biol.*, 20:8303–8318, 2000.

[10] J.F. Caceres, S. Stamm, D.M. Helfman, and A.R. Krainer. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science*, 265:1706–1709, 1994.

[11] A. Hanamura, J.F. Caceres, A. Mayeda, B.R. Franza Jr., and A.R. Krainer. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA*, 4:430–444, 1998.

[12] R.H. Hovhannisyan and R.P. Carstens. Heterogeneous ribonucleoprotein M is a splicing regulatory protein that can enhance or silence splicing of alternatively spliced exons. *J. Biol. Chem.*, 282:36265–74, 2007.

[13] B.K. Dredge, G. Stefani, C.C. Engelhard, and R.B. Darnell. Nova autoregulation reveals dual functions in neuronal splicing. *EMBO J.*, 24:1608–1620, 2005.

[14] Y. Jin, H. Suzuki, S. Maegawa, H. Endo, S. Sugano, K. Hashimoto, and et al. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J.*, 22:905–912, 2003.

[15] J. Han and T.A. Cooper. Identification of CELF splicing activation and repression domains in vivo. *Nucleic Acids Res.*, 33:2769–2780, 2005.

[16] L.H. Hung, M. Heiner, J. Hui, S. Schreiner, V. Benes, and A. Bindereif. Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. *RNA*, 14:284–196, 2008.

[17] S.M. Berget. Exon recognition in vertebrate splicing. *J. Biol. Chem.*, 270: 2411–2414, 1995.

[18] HC. Kuo, FH. Nasim, and PJ. Grabowski. Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. *Science*, 251:1045–1050, 1991.

[19] Z. Dominski and R. Kole. Selection of splice sites in pre-mRNAs with short internal exons. *Mol.Cell.Biol.*, 11:6075–83, 1991.

[20] G. Nogues, M.J. Munoz, and A.R. Kornblihtt. Influence of polymerase II processivity on alternative splicing depends on splice site strength. *J. Biol. Chem.*, 278:52166–52171, 2003.

[21] D.A. Sterner, T. Carlo, and S.M. Berget. Architectural limits on split genes. *Proc. Natl. Acad. Sci. USA*, 93:15081–15085, 1996.

[22] E. Buratti and F.E Baralle. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol.Cell.Biol.*, 24:10505–10514, 2004.

[23] E.A. Glazov, M. Pheasant, S. Nahkuri, and J.S. Mattick. Evidence for control of splicing by alternative RNA secondary structures in Dipteran homothorax pre-mRNA. *RNA Biology*, 3:36–39, 2006.

[24] D. Prou, W.J. Gu, S. Le Crom, J.D. Vincent, J. Salamero, and P. Vernier. Intracellular retention of the two isoforms of the D(2) dopamine receptor promotes endoplasmic reticulum disruption. *J.Cell Sci.*, 114(Pt 19):3517–3527, 2001.

[25] M. Monshausen, U. Putz, M. Rehbein, M. Schweizer, L. DesGroseillers, D. Kuhl, and et al. Two rat brain staufen isoforms differentially bind RNA. *J.Neurochem.*, 76:155–165, 2001.

[26] M. Gesemann, V. Cavalli, A.J. Denzer, A. Brancaccio, B. Schumacher, and M.A. Ruegg. Alternative splicing of agrin alters its binding to heparin,dystroglycan, and the putative agrin receptor. *Neuron.*, 16:755–767, 1996.

[27] P. Christmas, J.P. Jones, C.J. Patten, D.A. Rock, Y. Zheng, and S.M. et al. Cheng. Alternative splicing determines the function of CYP4F3 by switching substrate specificity. *J.Biol.Chem.*, 276:38166–38172, 2001.

[28] P. Niccoli-Sire, L. Fayadat, S. Siffroi-Fernandez, Y. Malthierry, and J.L. Franc. Alternatively spliced form of human thyroperoxidase,TPOzanelli: activity,intracellular trafficking, and role in hormonogenesis. *Biochemistry*, 40: 2572–2579, 2001.

[29] H. Tanahashi and T. Tabira. Three novel alternatively spliced isoforms of the human beta-site amyloid precursor protein cleaving enzyme (BACE) and their effect on amyloid beta-peptide production. *Neurosci. Lett.*, 307:9–12, 2001.

[30] C. Schwerk and K. Schulze-Osthoff. Regulation of apoptosis by alternative pre-mRNA splicing. *Mol.Cell*, 19:1–13, 2005.

[31] D.A. Harrison. Sex determination: controlling the master. *Curr. Biol.*, 17: R328–R330, 2007.

[32] E. Nagy and L.E. Maquat. A rule for temination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.*, 23:198–199, 1998.

[33] Y. Xing and C.J. Lee. Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy. *Trends Genet.*, 20: 472–475, 2004.

[34] F. Lejeune and L.E. Maquat. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr. Opin. Cell. Biol.*, 17:309–315, 2005.

[35] Q. Pan, A.L. Saltzman, Y.K. Kim, C. Misquitta, O. Shai, L.E. Maquat, B.J. Frey, and B.J. Blencowe. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.*, 20:153–158, 2006.

[36] K.B. Jensen, B.K. Dredge, G. Stefani, R. Zhong, R.J. Buckanovich, H.J. Okano, and et al. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, 25:359–371, 2000.

[37] A.A. Tesoriero, E.M. Wong, M.A. Jenkins, J.L. Hopper, M.A. Brown, G. Chenevix-Trench, and et al. Molecular characterization and cancer risk associated with BRCA1 and BRCA2 splice site variants identified in multiple-case breast cancer families. *Hum. Mutat.*, 26:495, 2005.

[38] H.X. Liu, L. Cartegni, M.Q. Zhang, and A.R. Krainer. A mechanism for exon skipping caused by nonsense or missense mutations in brca1 and other genes. *Nat.Genet.*, 27:55–58, 2001.

[39] Y. Yang, S. Swaminathan, B.K. Martin, and S.K. Sharan. Aberrant splicing induced by missense mutations in BRCA1: clues from a humanized mouse model. *Hum. Mol. Genet.*, 12:2121–2131, 2003.

[40] C.A. Pettigrew and M.A. Brown. Pre-mRNA splicing aberrations and cancer. *Front Biosci.*, Jan1;13:1090–1105, 2008.

[41] E. Kim, A. Goren, and G. Ast. Insights into the connection between cancer and alternative splicing. *Trends Genet.*, Jan 24(1):7–10, 2008.

[42] A.S. Solis, N. Shariat, and J.G. Patton. Splicing fidelity, enhancers, and disease. *Front Biosci.*, Jan1,13:1926–1942, 2008.

[43] L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, 8:967–974, 1998.

[44] S.J. Wheelan, D.M. Church, and J.M. Ostell. Spidey: A tool for mRNA-to-Genomic alignments. *Genome Res.*, 11:1952–1957, 2001.

[45] W.J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res.*, 12:656–664, 2002.

[46] T.D. Wu and C.K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21:1859–1875, 2005.

[47] E. van Nimwegen, N. Paul, R. Sheridan, and M. Zavolan. SPA: A probabilistic algorithm for spliced alignment. *PLoS Genet.*, 2:e24, 2006.

[48] U. Schulze, B. Hepp, C.S. Ong, and G. Ratsch. PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics*, 23:1892–1900, 2007.

[49] M. Zavolan, S. Kondo, C. Schonbach, J. Adachi, D.A. Hume, Y. Hayashizaki, T. Gaasterland, RIKEN GER Group, and GSL Members. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, 13:1290–1300, 2003.

[50] E.D. Green. Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.*, 2:573–583, 2001.

[51] E.C. Rouchka, W. Gish, and D.J. States. Comparison of whole genome assemblies of the human genome. *Nucleic Acids Res.*, 30:5004–5014, 2002.

[52] S.H. Nagaraj, R.B. Gasser, and S. Ranganathan. A hitchhiker's guide to expressed sequence tag (est) analysis. *Brief. Bioinform.*, 8:6–21, 2006.

[53] P Carninci, T Kasukawa, S. Katayama, J. Gough, M.C. Frith, N. Maeda, and et al. The transcriptional landscape of the mammalian genome. *Science*, 309: 1559–1563, 2005.

[54] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K. Koyanagi, and et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biology*, 2:856–875, 2004.

[55] M. Das, I. Harvey, L.L. Chu, M. Sinha, and J. Pelletier. Full-length cDNAs: more than just reaching the ends. *Physiol. Genomics*, 6:57–80, 2001.

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

[57] C.W. Sugnet, W.J. Kent, M. Jr Ares, and D. Haussler. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput.*, 2004:66–77, 2004.

[58] T.A. Thanaraj, F. Clark, and J. Muilu. Conservation of human alternative splice events in mouse. *Nucl. Acids Res.*, 31:2544–2552, 2003.

[59] R. Sorek and G. Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, 13:1631–1637, 2003.

[60] C.L. Zheng, XD. Fu, and M. Gribskov. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, 11:1777–1787, 2005.

[61] H. Itoh, T. Washio, and M. Tomita. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA*, 10:1005–1018, 2004.

[62] B.J. Blencowe. Alternative Splicing: New insights from global analyses. *Cell*, 126:37–47, 2006.

[63] C.W. Sugnet, K. Srinivasan, T.A. Clark, G. O'Brien, M.S. Cline, H. Wang, A. Williams, D. Kulp, J.E. Blume, D. Haussler, and M. Ares. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput. Biol.*, 2:e4, 2006.

[64] T.A. Clark, A.C. Schweitzer, T.X. Chen, M.K. Staples, G. Lu, H. Wang, A. Williams, and J.E. Blume. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, 8:R64, 2007.

[65] G. Yeo, D. Holste, G. Kreiman, and C.B Burge. Variation in alternative splicing across human tissues. *Genome Biol.*, 5:R74, 2004.

[66] F.L. Watson, R. Puttmann-Holgado, F. Thomas, D.L. Lamar, M. Hughes, M. Kondo, V.I. Rebel, and D. Schmucker. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science*, 309:1874–1878, 2005.

[67] F. Wen, F. Li, H. Xia, X. Lu, X. Zhang, and Y. Li. The impact of very short alternative splicing on protein structures and functions in the human genome. *Trends Genet.*, 20:232–236, 2004.

[68] K. Homma, R.F. Kikuno, T. Nagase, O. Ohara, and K. Nishikawa. Alternative splice variants encoding unstable protein domains exist in the human brain. *J. Mol.Biol.*, 343:1207–1220, 2004.

[69] M.L. Tress, P.L. Martelli, A. Frankish, G.A. Reeves, J.J. Wesselink, and C. et al. Yeats. The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci. USA*, 104:5495–5500, 2007.

[70] K. Yura, M. Shionyu, K. Hagino, A. Hijikata, Y. Hirashima, T. Nakahara, T. Eguchi, K. Shinoda, A. Yamaguchi, K. TAkahashi, T. Itoh, T. Imanishi, T. Gojobori, and M. Go. Alternative splicing in human transcriptome: Functional and structural influence on proteins. *Gene*, 380:63–71, 2006.

[71] M. Nakao, R.A. Barrero, Y. Mukai, C. Motono, M. Suwa, and K. Nakai. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res.*, 33:2355–2363, 2005.

[72] P.R. Romero, S. Zaidi, Y.Y. Fang, V.N. Uversky, P. Radivojac, C.J. Oldfield, M.S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, and A.K. Dunker. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. USA*, 103:8390–8395, 2006.

[73] P. Wang, B. Yan, J.t. Guo, C. Hicks, and Y. Xu. Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc. Natl. Acad. Sci. USA*, 102:18920–18925, 2005.

[74] E.V. Kriventseva, I. Koch, R. Apweiler, M. Vingron, P. Bork, M.S. Gelfland, and S. Sunyaev. Increase of functional diversity by alternative splicing. *Trends Genet.*, 19:124–128, 2003.

[75] G. Lopez, A. Valencia, and M. Tress. Firedb - a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.*, 35: D219–D223, 2006.

[76] M.L. Tress, O. Grana, and A. Valencia. SQUARE - determining reliable regions in sequence alignments. *Bioinformatics*, 20:974–975, 2004.

[77] M.L. Tress, D. Jones, and A. Valencia. Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. biol.*, 330:705–718, 2003.

[78] G. Lopez, A. Valencia, and M.L. Tress. firestar - prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, 35:W573–577, 2007.

[79] Y. Xing and C.J. Lee. Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, 7:499–509, 2006.

[80] B. Modrek and C.J. Lee. Alternative splicing in the human,mouse,and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, 34:177–180, 2003.

[81] R. Sorek, R. Shamir, and G. Ast. How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, 20:68–71, 2004.

[82] Q. Pan, M.A. Bakowski, Q. Morris, W. Zhang, B.J. Frey, and T.R. Huges. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.*, 21:73–77, 2005.

[83] S. Schwartz, J. Silva, D. Burstein, T. Pupko, E. Eyras, and G. Ast. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.*, 18:88–103, 2007.

[84] D. Brett, H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nat. Genet.*, 30:29–30, 2002.

[85] E. Kim, A. Magen, and G. Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, 35:125–131, 2007.

[86] I. Letunic, R.R. Copley, and P. Bork. Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.*, 11:1561–1571, 2002.

[87] R. Sorek, G. Ast, and D. Graur. Alu-containing exons are alternatively spliced. *Genome Res.*, 12:1060–1067, 2002.

[88] G. Lev-Maor, R. Sorek, N. Shomron, and G. Ast. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science*, 300:1246–1247, 2003.

[89] M. Wu, L. Li, and Z. Sun. Transposable element fragments in protein-coding regions and their contributions to human functional proteins. *Gene*, 401:165–171, 2007.

[90] B.P. Cusack and K.H. Wolfe. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol.Biol.*, 22:2198–2208, 2005.

[91] W. Wang, H. Zheng, S. Yang, H. Yu, J. Li, and H. et al. Jiang. Origin and evolution of new exons in rodents. *Genome Res.*, 15:1258–1264, 2005.

[92] D. Kaufmann, O. Kenner, P. Nurnberg, W. Vogel, and B. Bartelt. In NF1, CFTR, PER3, CARS, and SYT7 alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons. *Eur. J. Hum. Genet.*, 12:139–149, 2004.

[93] Z. Kan, D. States, and W. Gish. Selecting for functional alternative splices. *Genome Res.*, 12:1837–1845, 2002.

[94] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, and M. Platzer. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, 36:1255–1257, 2004.

[95] G. Condorelli, R. Bueno, and R.J. Smith. Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics. *J.Biol.Chem.*, 269:8510–8516, 1994.

[96] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, and M. Platzer. Single-nucleotide polymorphisms in NAGNAG acceptors

are highly predictive for variations of alternative splicing. *Am.J. Hum.Genet.*, 78:291–302, 2006.

[97] K. Tadokoro, M. Yamazaki-Inoue, M. Tachibana, M. Fujishiro, K. Nagao, M. Toyoda, and et al. Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. *J. Hum. Genet.*, 50: 382–394, 2005.

[98] K. Iida, M. Shionyu, and Y. Suso. Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals. *Mol.Biol.Evol.*, Epub ahead of print, 2008.

[99] K.W. Tsai, W.Y. Tarn, and W.C. Lin. Wobble splicing reveals the role of the branch point sequence-to-NAGNAG region in 3' tandem splice site selection. *Mol.Cell Biol.*, 27:5835–5848, 2007.

[100] Chern T.M., E. van Nimwegen, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and M. Zavolan. A simple physical model predicts small exon length variations. *PLoS Genet.*, 2:e45, 2006.

[101] K.J. Howe. RNA polymerase II conducts a symphony of pre-mRNA processing activities. *Biochimica et Biophysica Acta*, 1577:308–324, 2002.

[102] A.R Kornblihtt. Promoter usage and alternative splicing. *Curr. Opin. Cell. Biol.*, 17:262–268, 2005.

[103] M. de la Mata and A.R. Kornblihtt. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat. Struct. Mol. Biol.*, 13:973–980, 2006.

[104] A.R Kornblihtt. Chromatin, transcript elongation and alternative splicing. *Nature Structural AND Molecular Biology*, 13:5–7, 2006.

[105] M. de la Mata, C.R. Alonso, S. Kadener, J.P. Fededa, M. Blaustein, F. Pellsch, P. Cramer, D. Bentley, and A.R. Kornblihtt. A slow RNA polymerase II affects alternative splcing in vivo. *Mol. Cell.*, 12:525–532, 2003.

[106] E. Batsche, M. Yaniv, and C. Muchardt. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat. Struct. Mol. Biol.*, 13:22–29, 2006.

[107] D. Auboeuf, E. Batsche, M. Dutertre, and C. Muchardt. Coregulators: transducing signal from transcription to alternative splicing. *Trends Endocrinol. Metab.*, 18:122–129, 2007.

[108] P Carninci, T.A. Sandelin, B Lenhard, S Katayama, K Shimokawa, J. Ponjavic, and et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, pages 626–635, 2006.

[109] Chern T.M., N. Paul, E. van Nimwegen, and M. Zavolan. Computational analysis of full-length cDNAs reveals frequent coupling between transcriptional and splicing programs. *DNA Res.*, 2:63–72, 2008.

[110] M. Shapiro and P. Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, 15:7155–7174, 1987.

[111] I. Montanuy, R. Torremocha, C. Hernandex-Munain, and C. Sune. Promoter influences transcription elongation: TATA-box element mediates the assembly of processive transcription complexes responsive to cyclin-dependent kinase 9. *J. Biol. Chem.*, 283:7368–7378, 2008.

[112] L.J. Core and J.T. Lis. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, 319:1791–1792, 2008.

[113] C. Lee, X. Li, A. Hechmer, M. Eisen, M.D. Biggin, B.J. Venters, and et al. NELF and GAGA factor are linked to promoter proximal pausing at many genes in drosophila. *Mol.Cell.Biol.*, Mar 10, published ahead of print, 2008.

[114] A. Krumm, L. Hickey, and M. Groudine. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev.*, 9:559–572, 1995.

[115] R.P. Carstens, J.V. Eaton, H.R. Krigman, P.J. Walther, and M.A. Garcia-Blanco. Alternative splicing of fibroblast growth factor receptor 2(FGF-R2) in human prostate cancer. *Oncogene*, 15:3059–3065, 1997.

[116] B. Sommer, K. Keinanen, T.A. Verdoorn, W. Wisden, N. Burnashev, A. Herb, and et al. Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science*, 249:1580–1585, 1990.

[117] T.A. Gustafon, E.C. Clevinger, T.J. O′Neill, P.J. Yarowsky, and B.K. Krueger. Mutually exclusive exon splicing of type III brain sodium channel alpha subunit RNA generates developmentally regulated isoforms in rat brain. *J.Biol.Chem.*, 268:18648–18653, 1993.

[118] Z.Z. Tang, P. Liao, G. Li, F.L. Jiang, D. Yu, X. Hong, and et al. Differential splicing patterns of L-type calcium channel Cav1.2 subunit in hearts of spontaneously hypertensive rats and wistar Kyoto rats. *Biochim. Biophys. Acta.*, 1783:118–130, 2008.

[119] P.L. Graham, J.J. Johnson, S. Wang, M.H. Sibley, M.C. Gupta, and J.M. Kramer. Type IV collagen is detectable in most, but not all, basement membranes of caenorhabditis elegans and assembles on tissues that do not express it. *J.Cell Biol.*, 137:1171–1183, 1997.

[120] D.M. Helfman. The generation of protein isoform diversity by alternative RNA splicing. *Soc.Gen.Physiol.Ser.*, 49:105–115, 1994.

[121] JP. Jin, J. Wang, and O. Ogut. Developmentally regulated mouse type-specific alternative splicing of the COOH-terminal variable region of fast skeletal muscle troponin T and an aberrant splicing pathway to encode a mutant COOH-terminus. *Biochem. Biophys. Res. Commun.*, 242:540–544, 1998.

[122] W.M. Wojtowicz, J.J. Flanagan, S.S. Millard, S.L. Zipursky, and J.C. Clemens. Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118:619–633, 2004.

[123] G. Ohno, M. Hagiwara, and H. Kuroyanagi. STAR family RNA-binding protein ASD-2 regulates developmental switching of mutually exclusive alternative splicing in vivo. *Genes Dev.*, 22:360–374, 2008.

[124] S. Olson, M. Blanchette, J. Park, Y. Savva, G.W. Yeo, J.M. Yeakley, and et al. A regulator of dscam mutually exclusive splicing fidelity. *Nat. Struct. Mol. Biol.*, 14:1134–1140, 2007.

Chapter 7

Curriculum Vitae

Ms. Tzu-Ming Chern
Nationality: South African
Date of birth: Nov. 19th 1976
Address: Ahornstrasse 51, Basel, Switzerland 4055
Work Tel #: +41 (61) 267-1573
Mobile #:+41 0786546195
Email: t.chern@unibas.ch

Recent Education

- **PhD bioinformatics student at University of Basel, Switzerland (2004-present)**
- MSc. Bioinformatics cum laude, South African National Bioinformatics Institute, South Africa (2000-2002)
- BSc. Honours (Protein biochemistry), University of Witwatersrand, South Africa (1999)
- BSc (Biochemistry and Genetics) cum laude, University of Witwatersrand, South Africa (1996-1998)

Recent Awards & Publications

Notable awards

- Overseas prestigious PhD scholarship from the South African National Research Foundation (2004-2008)
- Best Biological Science Student for BSc degree (1998) awarded SA breweries gold medal

Recent Publications

- Chern et al. (2006) A simple physical model predicts small exon length variations. Plos Genet. Apr:2(4)e45
- Chern et al. (2008) Computational Analysis of Full-length cDNAs Reveals Frequent Coupling Between Transcriptional and Splicing Programs.

References

PhD thesis advisor:
Professor Mihaela Zavolan
Biozentrum, University of Basel, Switzerland
Email: mihaela.zavolan@unibas.ch
Tel: +41 61 267 1577

MSc thesis advisor:
Professor Winston Hide
SANBI
University of Western Cape
South Afrcia
Email: winhide@sanbi.ac.za
Tel: +27 21 959 3645

May 30, 2008