Genome Signals and Evolution for Fidelity and Regulation of Pre-mRNA Splicing

A Dissertation Presented

by

Chaolin Zhang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Biomedical Engineering

Stony Brook University

May 2008

Stony Brook University

The Graduate School

Chaolin Zhang

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Michael Q. Zhang – Dissertation Advisor
Professor, Biomedical Engineering & Cold Spring Harbor Laboratory

Michael Hadjiargyrou – Chairperson of Defense
Associate Professor, Biomedical Engineering

Scott Powers, Associate Professor
Biomedical Engineering & Cold Spring Harbor Laboratory

John Castle, Merck Research Fellow
Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck & Co., Inc

Adrian R. Krainer, Professor
Cold Spring Harbor Laboratory

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of Graduate School

Abstract of the dissertation

Genome Signals and Evolution for Fidelity and Regulation of Pre-mRNA Splicing

by

Chaolin Zhang

Doctor of Philosophy

in

Biomedical Engineering

Stony Brook University

2008

A majority of eukaryotic genes have alternating exons and introns. Introns in pre-mRNAs are removed and exons are joined to generate mature transcripts in a process called splicing. By different combinations of exons and splice sites, i.e., alternative splicing, one gene can produce multiple transcript and protein isoforms, providing a major source of proteomic diversity, novel mechanisms of gene expression regulation, and new paths of gene evolution. Splicing and alternative splicing are dictated by interactions of many *cis*-regulatory elements and *trans*-acting splicing factors in a cellular machinery called spliceosome. However, the splicing code that elucidates how these interactions determine the splicing outcome, sometimes specific for particular tissues, developmental stages, and different species or populations, is still poorly understood. This study aims to advance the mechanistic understanding of the fidelity and regulation of both constitutive and alternative splicing. To approach the aim, I mainly use statistical and computational analysis of genomewide, high-throughput data to generate experimentally testable hypotheses, combined with experimental validations by collaborative bench biologists. In this dissertation, I first demonstrate the limited fidelity of splicing and describe a new and unusual type of alternative splice site—dual-specificity splice site, which implies the evolutionary selective pressure to reduce splicing fluctuations and eliminate evolutionary intermediates. Then, I show how such selective pressure results in non-random, distinct distributions of splicing-regulatory elements of different classes in exons and introns, including deep intronic sequences, for optimal exon and intron discrimination. The distribution of splicing-regulatory elements is gauged by a neutral evolution model developed from DNA strand-asymmetry patterns, and the principle is also very effective to predict new regulatory elements. To achieve a better understanding of the organization and functional impacts of splicing-regulatory networks, I use the tissue-specific splicing factors Fox-1/2 as a model. By comparative analysis of 28 vertebrate species, this study predicts thousands of conserved Fox-1/2 targets with high specificity and sensitivity; at least 50-60% of predicted targets can be experimentally verified in HeLa cells. This analysis reveals a surprising extensiveness and complex patterns of tissue-specific splicing regulation. The regulatory network is highly organized and modular, with many predicted targets important for neuromuscular functions and diseases. Lastly, in addition to splicing fidelity and regulation in normal conditions, I also describe one of the first surveys of splicing dysregulation in prostate cancer, which demonstrates the importance of splicing in tumorigenesis, and the unique advantage of splicing profiling in cancer sample classification and biomarker identification.

Dedicated to my family

Contents

List of Tables

Chapter 2: Data sources, compilation and visualization

Chapter 3: Evolutionary impact of limited splicing fidelity

Chapter 4: Dual-specificity splice sites function alternatively as 5' and 3' splice sites

Chapter 5: RNA landscape of evolution for optimal exon and intron discrimination

Chapter 6: Defining the splicing regulatory network of tissue-specific splicing factors Fox-1 and Fox-2

Chapter 7: Profiling alternatively spliced mRNA isoforms for prostate cancer classification

List of Figures

Chapter2: Data sources, compilation and visualization

Chapter 3: Evolutionary impact of limited splicing fidelity

Chapter 4: Dual-specificity splice sites function alternatively as 5’ and 3’ splice sites

Chapter 5: RNA landscape of evolution for optimal exon and intron discrimination

Chapter 6: Defining the splicing regulatory network of tissue-specific splicing factors Fox-1 and Fox-2

Chapter 7: Profiling alternatively spliced mRNA isoforms for prostate cancer classification

Acknowledgements

I most gratefully acknowledge my advisor, Michael Q. Zhang, for leading me into the field, and the guidance and support during my doctoral study. I also gratefully acknowledge Adrian R. Krainer for allowing me to attend their group seminars, and many collaborations, suggestions and helps in various projects. I would like to thank my past and current colleagues in the Zhang lab (Zhenyu Xuan, Gengxin Chen, Andrew Smith, Pavel Sumazin, Fang Zhao, Xiaoyue Zhao, Debopriya Das, Dustin Schones, and Jeffrey Rosenfeld), and in the Krainer lab (Michelle Hastings, Zuo Zhang, Philip Smith, Shuying Sun, Rotem Karni, Yimin Hua, Xavier Roca, Rahul Sinha, and Hazeem Okunola), from whom I benefit a lot through many discussions. Among them, I would like to particularly acknowledge Zhenyu, Michelle, Zuo, Shuying, and Phil for collaborations in different projects. I also acknowledge other inside and outside collaborators, including Wen-Hsiung Li (U. Chicago), Xiang-Dong Fu (UCSD), John Castle and Jason Johnson (Rosetta/Merck), John Conboy (LBNL), Ken Zaret (Fox Chase Cancer Center) and Scott Lowe (CSHL) for providing the opportunities of productive collaborations. Finally, I would like to thank my mother, wife Shu, and other family members for their love and support all these years.

Chapter 1

Introduction

The completion of genome sequencing projects for human and many other species shifted the focus of genomic research to understanding mechanisms of gene expression regulation. This task includes two main aspects: first, how the genetic information coded in the genome confers the orchestration of gene expression to determine the complex phenotypes of an organism; second, how perturbations of the genetic information, as well as environmental factors, results in phenotypic differences among different species, different human populations, and various genetic diseases. In the past two decades or so, computational biologists from different fields, such as biophysics, statistics, and computer sciences, have started to play a vital role in addressing these questions, because of the overwhelming amount of data generated by large-scale sequencing, annotations, and comparisons of multiple genomes, as well as by the recently developed high-throughput technologies probing the genomes from different angles.

While the process of gene expression, or the transformation of genetic information, was summarized very elegantly as the central dogma—"DNA makes RNA makes protein"—50 years ago, the mechanisms of gene expression regulation in mammalian systems are extremely complex, including regulations at each individual step, i.e., transcription, pre-mRNA splicing, mRNA export and editing, translation and post-translational modifications. These steps are dictated by different cellular machineries, which are, however, coupled with each other, forming complex, multilayered gene-expression regulatory networks. This dissertation focuses on the first step of post-transcriptional regulation, i.e., pre-mRNA splicing and alternative splicing, which were first discovered in 1970's (1, 2). Pre-mRNA splicing is almost universal among all eukaryotes, and even more essential for vertebrates and mammals. Splicing and alternative splicing are key to expanding proteomic diversity, regulating gene expression, and creating new paths of gene evolution. Mutations resulting in aberrant splicing are implicated in numerous genetic diseases, ranging from neurological disorders to cancer, and in some cases they account for almost half of all disease-causative mutations (3, 4). However, the mechanisms of splicing fidelity, regulation and dysregulation are still poorly understood. This dissertation addresses multiple aspects of these questions. To provide a context of my research, this chapter gives a brief introduction of background information related to splicing and recent advances in this field.

1.1 Eukaryotic gene splicing and alternative splicing

A majority of eukaryotic genes are split, with coding segments (exons) separated by noncoding segments (introns). After transcription, introns of premature messenger RNAs (pre-mRNAs) are removed and exons are joined, to produce mature mRNA transcripts, a process called splicing. Splicing was first discovered in adenovirus (1, 2), and later found to occur in almost all eukaryotes, such as yeast, plants, and animals. The fidelity of splicing is critical for expressing correct protein products, as aberrant exon insertions (deletions), or lengthenings (shortenings) result in alterations of amino acid sequences in particular parts, or in dramatic changes of the whole proteins by disrupting the reading-frame. These arguments raised the question of why eukaryotic genes are in pieces.

Alternative splicing expands proteomic diversity

Immediately after the discovery of splicing, it was proposed that one of its most important implications is the possibility of producing multiple transcript and protein isoforms by different combinations of exons, i.e., alternative splicing (AS), which provides a mechanism of expanding proteomic diversity and therefore organismal complexity (5). This prediction was soon validated and alternative splicing was found to be common in metazoans. Typical types of alternative splicing are inclusion or skipping of one or more exons (cassette exons), shortening or lengthening of an exon by alternative 5' and 3' splice site usages, mutual exclusion of two or more exons, and retained introns. Alternative splicing can also be coupled with alternative promoter or polyA usage. More complex alternative splicing patterns can be formed by combinations of different basic types.

A striking example of alternative splicing is the *Drosophila Dscam* gene, which is a member of the immunoglobulin superfamily required for axon guidance (6). This gene has a very complex architecture, including 95 alternative exons organized in four groups. The mutually exclusive splicing in these exon groups potentially generates 38,016 distinct axon guidance receptors. Although it is still unclear whether every isoform is required, the repertoire of axon guidance receptors must be large enough to generate the extraordinary complexity of synaptic connections (7-9) [reviewed by (10)]. In the fly mushroom body (a higher brain center that processes olfactory information), axons initially project as part of a single fascicle (bundle) in the central peduncle region and then bifurcate. Dscam appears to mediate the repulsive interactions between newly formed axon branches, because mutant flies in absence of *Dscam* often fail to separate the branches. Further studies suggested that the repulsion depends on the remarkably specific homophilic interactions among Dscam isoforms. Isoforms that differ in only a few amino acids exhibit no or very weak protein-protein associations. These results suggest that neighboring neurons express different subsets of isoforms for self versus nonself discrimination. Similar observations were also made for mammalian genes. In the case of neurexin genes for example, thousands of neurexin isoforms can be generated by three genes, through alternative promoter usage and alternative splicing (11).

Besides these extreme cases in which one gene generates numerous isoforms, alternative splicing is prevalent, especially in mammalian genes, although fewer isoforms were produced for each gene. With the completion of the human genome project and other related projects, it is are now clear that human has only 20,000-25,000 protein-coding genes (12); the number is in a similar range as in fish (13), slightly more than the worm (~19,000) (14), and only three times of unicellular yeast (~6,000) (15), although human is phenotypically much more complex than the other named species. To resolve this paradox, alternative splicing is regarded as a major source of proteomic diversity, generating biochemically distinct, sometimes even antagonistic, protein products from a limited set of genes. A typical mammalian gene comprises of 8-9 exons. The estimates of alternatively spliced genes keep rising, from 5% as initially estimated (16) to more than two thirds as currently estimated (17), largely due to the advance of sequencing and microarray technologies. Moreover, this is still not the final estimate, given the observation that almost all genes have alternative isoforms if the transcriptome is sampled with a sufficient depth (18).

Alternative splicing regulates gene expression

A second important implication of alternative splicing is to express functionally distinct isoforms in specific tissues or developmental stages. It is also possible to switch on/off gene expression in particular conditions. By inserting or deleting an exon or part of an exon whose length is not a multiple of three into the coding region, the reading frame is shifted, creating premature stop codons (PTCs) before the end of the transcript. PTCs can also be introduced by inserting a “poisonous” exon, which carries an in-frame stop codon. A resulting nonsense transcript carrying a PTC >50 nt upstream of a splice junction is generally degraded in a process called nonsense-mediated mRNA decay (NMD) (19, 20). It was reported that alternative splicing and NMD are widely coupled with each other; one-third of alternative transcripts observed from expressed sequence tags (ESTs) are potentially subject to NMD (21). Therefore, such a mechanism may serve as a post-transcriptional on/off switch of gene expression, which is important to specify tissue-identities or developmental stages. One best studied example is the splicing of the *Drosophila* Transformer (*Tra*) gene, which is critical to initiate a cascade of sex-specific alternative splicing events to determine the sex phenotype. The *Tra* gene is not expressed in males due to the use of the proximal splice site from a pair of alternative 3’ splice sites, which introduces a PTC. However, the recognition of the proximal 3’ splice site is blocked by a female-specific splicing factor sex lethal (*Sxl*), which forces the use of the distal 3'splice site, generating a female-specific transcript without PTC.

Another interesting example of NMD induced by alternative splicing is related to the homeostatic expression of two classes of important splicing factors, i.e., SR proteins and hnRNPs (see below) (22, 23). More specifically, a majority of these splicing factors have frame-shifting or poisonous alternative exons, whose alternative splicing patterns are highly conserved between human and rodents, suggesting their important functional roles. It was proposed that when the expression of these splicing factors is too high, the unproductive splicing pattern is triggered to generate PTC-containing transcripts subject to NMD.

Alternative splicing increases the rate of exon creation and loss

Despite a number of reported examples of regulated unproductive splicing and translation (RUST) (23), it is still controversial how prevalent alternative splicing and NMD are coupled for gene expression regulation. Alternatively, NMD induced by alternative splicing might serve as a quality-control mechanism to eliminate aberrant splicing products, which might be toxic or dominant negative. The latter argument was supported by the observation that a vast majority of PTC-containing transcript isoforms are of very low abundance independent of the action of NMD (24).

The question raised above is related to the fact that splicing is stochastic in nature, similar to other biological processes, such as DNA replication and transcription. Therefore, there is an inherent limit of splicing fidelity, which is good or bad. As aforementioned, almost all genes are alternatively spliced when the transcriptome is sampled with sufficient depth; this is becoming clearer, as the ultra-high-throughput sequencing technologies are available very recently. In this sense, the splicing machinery must be accurate enough, which will otherwise introduce a burden of energy cost to eliminate the overwhelming splicing noises in the quality control pathways. On the other hand, alternative splicing may increase the rate of gene evolution, by exon creation and loss in new transcripts (25). Because the original isoform is maintained and the new isoforms are usually of low abundance, the toxicity and other deleterious effects, if any, are minimized. New transcripts with adaptive benefits can be positively selected and become more abundant during the course of evolution.

Evaluating the global functional and evolutionary impacts of alternative splicing events can shed light onto the understanding of general characteristics of important alternative splicing events, trends of evolution, and the splicing-regulatory mechanisms. My efforts in this direction are described in Chapters 3 and 4.

1.2 Biochemical reactions and the splicing machinery

So far I have not got into mechanisms of splicing and alternative splicing. This section summarizes the basic splicing reactions, and the splicing machinery that catalyzes such reactions. Regulation of exon/intron recognition and splicing is described in the next section.

To dictate splicing, each intron is almost invariantly marked by a GU dinucleotide at the 5' end (5' splice site, or 5'ss), and an AG dinucleotide at the 3'end (3' splice site or 3'ss), although exceptions exist. Other important splicing signals include a branch point sequence (BPS) upstream of the 3'ss and a polypyrimidine tract between the BPS and the 3'ss. The splice sites and BPS also include longer, less conserved consensus sequences, or motifs. For example, the motif sequences of a typical 5'ss and 3'ss are MAG|GURAGU and CAG|G, respectively, where M represents A or C, and R represents A or G. The consensus of BPS is UCCUR$\underline{\text{A}}$Y, where Y represents C or U, and the branch point is underlined. These primary splicing signals are universal and required for the recognition of every exon and intron, although the level of conservation varies in different exons/introns and organisms.

The basic splicing reactions comprise of two trans-esterification steps. In the first step, the 2'-hydroxyl group of the A residue at the branch point attacks the phosphate group of the upstream 5'ss, which results in a detached 5' exon and an intron lariat ligated by the intron 5' end with the branch point. In the second step, the 3'-hydroxil group of the detached exon attacks the phosphate group of the downstream 3'ss, followed by the ligation of the two exons and the release of the intron lariat.

These reactions, requiring ATP, are catalyzed by a large and highly dynamic complex called spliceosome. The splicesome core is composed of four small ribonucleoprotein particles (snRNP U1, U2, U4/U6 and U5) and numerous auxiliary proteins, which are assembled into the spliceosome in a series of steps. Initially, the U1 snRNP binds to the 5'ss by the base pairing between the 5'ss and the U1 snRNA. At the same time, a branch-point protein SF1 binds to the BPS; the two subunits of the dimeric U2 auxiliary factor (U2AF), U2AF65 and U2AF35, bind to the polypyrimidine tract and the 3' splice site, respectively. The spliceosome at this stage is called the E (early) complex, which commits the substrate transcript to splicing. In the next few steps, the spliceosome is dynamically rearranged, with some factors replaced by others. The U2 snRNP, recruited by SF2, joins the E complex and binds the branch point, forming the A complex, which is then followed by joining of the U4/U6-U5 tri-snRNPs, forming the B complex. In the B complex, dramatic conformational changes occur to replace the U1 snRNP by U6 and to displace the U4 snRNP. The resulting C complex actually catalyzes the splicing reactions.

The process described above is the major pathway of splicing or U2-type splicing, which is the focus of this dissertation. However, there exist a very small group (<0.1%) of introns, which are flanked by an AU dinucleotide at 5'ss and an AC dinucleotide at 3'ss (26). These splice sites, together with a BPS motif distinct from the U2-type introns, are recognized by another set of snRNPs (U11, U12, U4atac, U6atac). This minor splicing machinery and the U12-type splicing have an early origin, which can be traced back to eukaryotic microbes and in a fungus (27). The functional significance of maintaining such rare introns and a second splicing machinery is unclear. A recent study reported that the minor U12-type splicing predominantly takes place in cytoplasm, in contrast to the major U2-type splicing, which occurs in the nucleus. Interestingly, the same study also found that minor splicing plays very specific roles in cell cycle progression (28).

Besides the core spliceosomal components, many more proteins are involved in various steps of splicing. Based on mass spectrometric analysis, more than 300 proteins are directly assembled into the spliceosome in mammals, making it one of the largest protein complexes (29-31). In addition, many other splicing factors, such as those involved in splicing before the assembly of the spliceosome, and those expressed in very limited types of tissues or developmental stages, are difficult to identify in the purified spliceosome. Cataloguing all splicing factors and annotating their roles in splicing regulation is a fundamental work for splicing-regulation studies. I have generated a comprehensive list of known and putative splicing factors or spliceosomal components by a semi-automatic pipeline. Although more efforts are required to be made for more systematic annotations of these proteins, these data will be valuable in the near future and are described in Chapter 2.

1.3 The splicing code

Although the basic biochemical reactions of splicing have been worked out, how the splicing machinery accurately recognizes exons and introns, or the splicing code, remains poorly understood, especially in mammals. In unicellular yeast, the genome is very compact and introns are generally small. In addition, the splice-site and BPS motifs are rather conserved across different introns. Therefore, the primary splicing signals, including the splice sites, BPS and polypyrimidine tract, are sufficient for exon and intron recognition (30). However, mammalian introns expand dramatically to a median size of ~1500 nt, compared to a median size of ~120 nt for exons. Meanwhile, the motifs of the splice sites and BPS become much more degenerate. In short, the information content provided by the primary splicing signals is not sufficient for mammalian exon and intron discrimination.

The general splicing code

For metazoans, many splicing-regulatory elements (SREs), in addition to the primary splicing signals, reside outside the splice sites, either in exons or in introns. This is especially true for mammalian genes, to recognize small exons embedded in much longer introns. These sequence elements can either enhance or repress splicing, depending on their sequence identities and context. Accordingly, splicing-regulatory elements are conventionally divided into four categories: exonic splicing enhancers and silencers (ESEs and ESSs), and intronic splicing enhancers and silencers (ISEs and ISSs) (32). In parallel to the invention of *cis*-regulatory elements, several families of splicing factors, including SR proteins, hnRNP proteins, and kinases, are greatly expanded in metazoans or vertebrates (29). Therefore, evolution of gene structure, *cis*-regulatory elements and the splicing machinery supports the notion that splicing regulation is more complex and essential for mammals and other vertebrates. Although the quantitative comparison in the rate of alternative splicing across different species remains difficult, several groups agreed that mammals appear to have a higher rate than other species (33, 34).

A focus of genomics research in pre-mRNA splicing is to elucidate the "splicing code", i.e., how the interactions between *cis*-regulatory elements and *trans*-acting splicing factors determine the splicing outcome. So far, a well characterized class of splicing-regulatory elements are ESEs interacting with a specific subset of SR proteins, including SF2/ASF, SC35, SRp35, SRp40 and SRp55. SR proteins are an important family of splicing factors with one or two RNA recognition motifs (RRMs) and an RS domain containing repeated arginine/serine dipeptides. They are highly conserved in a wide spectrum of species from yeast to mammals, although they undergo significant expansions during the evolution of metazoans. It has been proposed that the function of SR proteins is to stimulate the recognition of weak upstream 3'ss by recruiting U2AF65/35, to facilitate U1 snRNP binding to 5'ss (35, 36), or to counteract the effects of nearby silencers (37-39). In contrast to SR-protein-binding ESEs, the best characterized ESSs are those bound by heterogeneous ribonucleoproteins (hnRNPs), which coat nascent transcripts. HnRNP A1, one of the most studied hnRNPs, has two RRMs and a glycine-rich auxiliary domain. A1 has been shown to functionally bind to ESSs within *FGFR2* , *HIV-1* and other genes (40, 41). Both SR proteins and hnRNP proteins are ubiquitous and highly expressed in almost all cell types.

The ESEs and ESSs recognized by SR proteins and hnRNPs were initially identified in the context of individual alternative exons. To get more systematic understanding of splicing regulation, later studies attempted to determine the binding preferences of particular splicing factors (e.g. (42)), or to identify more ESEs or ESSs by SELEX (for systematic evolution of ligands by exponential enrichment) (43, 44) or cell-based screen (45). In addition, the accumulation of sequenced transcripts, including full-length cDNAs and ESTs, has made it possible to predict ESEs and ESSs using computational approaches (46-48) (see Section 1.5 for more details). These studies appear to reach a consensus that SR proteins and hnRNPs recognize very degenerate sequence motifs, whose occurrences are abundant in most, if not all, exons. This idea is plausible given the fact that ESEs and ESSs are superimposed onto the more restrictive protein code, and thus required to be flexible and robust enough, so that they can be recognized in various sequence contexts. This degeneracy also implies the importance of combinatorial regulation by SR proteins and hnRNPs, sometimes depending on the relative abundance of these proteins and the elements they recognize. For example, SF2/ASF bound to an ESE element antagonizes the function of hnRNP A1 bound to a juxtaposed ESS by impairing the propagation of cooperative binding of hnRNP A1 along the HIV *tat* exon 3 (49). In the *c-src* gene, there is an ESE within the 5' half of the N1 exon which, when bound by SF2/ASF, activates N1 exon inclusion in the neuronal tissue. However, this activation can be repressed by hnRNP A1 binding to the same elements *in vitro* (50). The antagonistic effects between SR proteins and hnRNPs can be determined precisely by the relative concentrations in *in vitro* splicing assays (51). In many cases of alternative 5' splice-site selection, an excess of hnRNP A1 tends to favor distal 5' splice sites, whereas an excess of SF2/ASF results in utilization of proximal 5' splice sites.

So far, relatively few studies focused on intronic regulatory elements, i.e., ISEs and ISSs (52). However, the effect of ESEs and ESSs on splicing is often context-dependent. For example, an SR-protein-dependent ESE element, when present in an intron, can act as an ISS to repress splicing (53), whereas a number of ESSs, such as the GGG motif, are also potent ISEs (54). These observations suggested the overlapping role of ESEs and ISSs, and ESSs and ISEs, at least to some extent. In addition, a recent study suggested even more complicated, often unpredictable, behaviors of exonic splicing regulatory (ESR) elements depending on their locations in exons (48). However, at least in some cases, insertion of one such element may create or disrupt other elements and/or RNA secondary structures, which may complicate the interpretation of the results.

These extensive efforts to elucidate the global splicing code have resulted in accumulating evidence which appears to suggest no qualitative difference between many constitutive and alternative exons. Constitutive exons tend to have stronger splice sites, longer polypyrimidine tracts, more abundant ESEs, and fewer ESSs. They can be readily converted into alternative exons by weakening any of the splicing signals, artificially in the lab or by nature. Naturally occurring random mutations that change exon/intron strength are eliminated by purifying selective pressure, if they are deleterious, or are positively selected, if they create beneficial new isoforms. However, the pattern and extent how evolutionary selective pressure constrains mammalian exons and introns for the purpose of accurate splicing are unclear in a genomewide scale. An answer to this question can provide important insights into the mechanisms of splicing fidelity and regulation. My efforts towards this direction are described in Chapter 5.

In addition to the enhancers and silencers, RNA secondary structures may also have important influences on splicing (55). However, testing such a hypothesis is very challenging at the current stage, due to the difficulty to reliably determine the secondary structures by computational methods. This difficulty is in two folds. First, current methods of RNA-secondary-structure prediction are mostly based on energy minimization. The search space of such an optimization procedure is large and has many local minima. Adding further complexity, the real RNA structure may not use the "optimal" solution. This is especially true given the fact that splicing is a co-transcriptional process, and that transcribed 5' part is folded before the 3' part is available. However, several interesting examples of RNA secondary structures important for splicing have been found (55), including the regulation of the mutually exclusive exons in *Drosophila Dscam* (56). By comparative analysis of 16 insects, two types of conserved sequence elements were identified in the exon group 6. The first element, called docking site, is located in the intronic region upstream of the first exon 6 variant. The elements of the second type, called selector sequences, are located upstream of each exon 6 variant. Importantly, the docking site and the select sequences are complementary to each other, which may bring upstream exon 5 and one exon 6 variant together. Combined with the evidence that hrp36, an hnRNP whose mammalian homolog is hnRNP A1, is important for the repression of these exon variants (57), it was proposed that the exon 6 splicing is initially repressed and the interaction between the docking site and one selector sequence somehow derepress the nearby exon variant.

The More specific splicing code

Besides SR proteins and hnRNPs, many other splicing factors and the pre-mRNA targets they regulate are poorly characterized. Of particular interest are those splicing factors expressed at specific tissue types or developmental stages, which are able to switch the splicing patterns of their targets. These splicing factors usually recognize specific regulatory elements, although frequently degenerate as well, in a relatively small subset of genes. Very interestingly, these regulatory elements, their context sequences, and the regulated alternative splicing patterns, are usually very conserved across different vertebrate species (58), suggesting important functional roles of such switchable alternative splicing events.

As aforementioned, a well studied model of developmental-specific splicing is Sxl, Tra, Tra-2, and several other splicing factors, which regulate a cascade of alternative splicing events during *Drosophila* sex determination (59, 60). In mammals, splicing factors known to be important for tissue-specific splicing include Nova-1/2 (61), PTB/nPTB (62, 63), Fox-1/2 (64), Muscleblind like (MBNL) (65) and CELF family proteins (66), Hu proteins (67), and TIA1/TIAR (68, 69). Two well characterized splicing factors regulating brain-specific splicing are PTB, Nova, and their paralogs.

PTB has four RRMs and recognizes pyrimidine-rich elements (70, 71). As a widely expressed protein, PTB represses the inclusion of many tissue-specific exons, including the neuron-specific exon N1 of *c-src*, a smooth muscle-specific exon of *α*-actinin and others. The repressive effect of PTB in specific tissues can be counteracted by an activator, or by a neuronal homolog of PTB (nPTB), which is less repressive.

A series of studies by the Darnell lab provided another example of tissue-specific splicing regulation in mammals (72-75). Nova-1/2, the first tissue-specific splicing factors discovered in vertebrates, are neuron-specific antigens targeted in paraneoplastic opsoclonus myoclonus ataxia (POMA). Nova proteins bind to an YCAY motif as determined by SELEX experiments. Clusters of the motif (≥3 copies spaced in ~30 nt) are necessary and sufficient to confer specific target recognition and regulation. Recently, cross-linking and immunoprecipitation (CLIP) and splicing microarrays have been used to identify *in vivo* targets of Nova in high throughput. Analyses of the identified targets revealed important functional implications of Nova-dependent splicing regulation. First of all, in analogy to transcriptional regulation, Nova-1/2 regulated targets have coherent functional roles in neuronal synapse or in axon guidance (73). This modular structure is consistent with the current understanding of gene expression regulation, and therefore strongly suggests that many alternative splicing events in vertebrates can be tightly regulated. Secondly, Nova-1/2 can activate or repress exon inclusion in a predictable way, depending on the location of the YCAY clusters (74). Nova-1/2 binding sites in the downstream intronic region usually enhance splicing of the alternative exon, whereas those inside the alternative exon or immediately upstream repress splicing. This position-dependent effect is likely due to different spliceosomal components Nova-1/2 interacts with. Thirdly, the effect of Nova-1/2 binding sites is highly local; a YCAY cluster in the intron downstream of an alternative exon will not affect splicing of the upstream intron, and vice versa.

In this dissertation, I use splicing factors Fox-1/2, which are specifically expressed in brain, heart and skeletal muscle, as a model to study tissue-specific splicing-regulatory networks.

1.4 Aberrant splicing in genetic diseases

Splicing patterns can be altered by mutations in splice sites, *cis*-regulatory elements, or *trans*-acting factors. Many of such alterations might have only moderate effects either because the magnitude of splicing change is small, or because the disrupted isoforms do not have important physiological functions. These mutations are largely tolerated during evolution. However, other mutations have dramatic effects on splicing and gene expression, and are implicated in various genetic diseases, ranging from neurological disorders to cancer.

An early review reported that 15% of point mutations implicated in genetic diseases affect splicing (76). Since only mutations disrupting splice sites were considered in this survey, the actual fraction of mutations resulting in aberrant splicing, including those in splicing-regulatory elements, is likely much higher. For example, exonic mutations that appear to be silent or affect protein coding might actually disrupt splicing. In the case of the *CFTR* gene, it was shown that a quarter of synonymous substitutions affect splicing. In the *ATM* and *NF1* genes, which are implicated in ataxia telangiectasia and neurofibromatosis type I, respectively, about half of mutations alter patterns or level of splicing (3, 4). Another interesting example is the survival of motor neuron 1 (*SMN1*) gene, the loss of which causes spinal muscular astrophy (SMA). SMA patients usually have an intact *SMN2* gene, which is almost identical to *SMN1* in genomic sequence and potentially codes for the same protein. However, *SMN2* is only partially functional because it generates a low level of full-length transcripts, and predominantly expresses an isoform lacking exon 7. The truncated protein is unstable and nonfunctional. The difference in exon 7 splicing between *SMN1* and *SMN2* is primarily due to a translationally silent, single nucleotide C/T difference at position 6 (77). This mutation disrupts an ESE element recognized by SF2/ASF (78, 79), and probably also creates an ESS element recognized by hnRNP A1/A2 (80, 81).

Disruptions of important *trans*-acting splicing factors usually have more deleterious effects, because they potentially affect splicing of many regulated targets. One interesting example is myotonic dystrophy (DM), one of the several known trinucleotide repeat disorders. This disease has several subtypes, including DM1 and DM2, with CUG expansion in the 3' UTR of the *DMPK* gene (82) and with CCUG expansion the intron1 of *ZNF9* (83), respectively. Although the expansions do not directly affect proteins encoded by the two genes, several models have been invoked to explain the molecular mechanisms underlying the diseases. In one model, the CUG or CCUG repeats are recognized by CUG-binding proteins, which are muscle-specific splicing factors. Therefore, the expansion of these repeats sequesters these proteins and alters their activities, which in turn changes splicing patterns of their targets.

Disruptions of general splicing factors, such as SR proteins, may play important roles in cancer. It was recently reported that SF2/ASF shows amplification and overexpression in various human tumors (84). Overexpression of the protein is sufficient to transform mouse immortal fibroblasts. The detailed mechanisms how abnormal expression of these splicing factors leads to cancer remain elusive, although several targets downstream of the pathway have been identified as tumor suppressors or oncogenes.

In some cases, altered expression of particular isoforms is found to be important in diseases, but the regulators of the altered splicing patterns are unknown. One very recent example is the pyruvate kinase (*PKM*) gene, which normally expresses the M1 isoform in adult tissues. Interestingly, another isoform M2, which is different from the M1 isoform by a mutually-exclusive exon, is expressed in embryos and different tumors. The expression of the embryonic M2 isoform is necessary for the high lactate production in the presence of oxygen in tumors, known as the Warburg effect. However, how this splicing switch is controlled is still unknown. As high-throughput technologies detecting individual splice isoforms in specific conditions are becoming available, it will be much easier to identify differential alternative splicing by comparing disease and normal samples. The gap between splicing factors, alternative splicing events, and their functional implications stand out more than ever (see Section 1.5 below).

1.5 A brief review of bioinformatics and genomics studies

As one can see from the review above, bioinformatics and genomics studies have been one of the driving forces to advance the mechanistic understanding of splicing fidelity and regulation. In addition, the development of new computational methodologies and ideas is also interwoven with technological advances, especially those high throughput genomic technologies.

Bioinformatics studies of splicing at the early stage mainly focused on the collection of splice sites (85) and alternative exons (86) from literature, which provided invaluable insights into the basic characteristics of splicing signals. The revolution of sequencing technologies soon led to an explosive expansion of cDNAs, ESTs, and protein sequences (87). Algorithms and software tools, such as BLAST (88), were developed for the manipulation and comparison of these sequences. This made it possible to detect alternative splicing events in a genomewide scale. Initially, pairwise comparison of transcripts was made to detect alternatively spliced regions. With the completion of genome-sequencing projects for human and other species, it is now rather standard to compare transcripts with genomic sequences. This is facilitated by special tools that consider the presence of introns and splice site consensuses in sequence alignments (89). To be more specific,,an alignment block represents an exon and an alignment gap represents an intron; the alignment at boundaries of each block is optimized locally according to the splice site consensuses. After alignment, splicing graph is the most natural representation of alternatively spliced transcripts and used to detect alternative splicing events and to model full-length isoforms (90). For a long time period, the focus of splicing studies has been to catalogue exons, introns and alternatively spliced isoforms; several databases have been built to host and visualize such data (91-94). The most important conclusion derived from these efforts is probably the extensiveness of alternative splicing. The estimates of genes with alternative splicing keep rising, from 5% as initially estimated (16), to more than two thirds as currently estimated (17). This is not the final estimate, given the observation that almost all genes have alternative isoforms if the transcriptome is sampled with a sufficient depth (18).

The compilation and catalogue of alternative splicing events provided important insights into the global features of splicing. To begin with, is the extensiveness of alternative splicing correlated with organismal complexity? In other words, do human and other mammals have a higher level of alternative splicing? There are still some debates on this question due to the complication of the very different EST coverage in different organisms (33, 34, 95). For example, human and mouse have eight and five million ESTs in dbEST, respectively, far more than other organisms, such as rat (0.9 million). In addition, the samples used to prepare EST libraries have different biases and heterogeneities in different organisms. For human, cancer samples account for about two thirds in all EST libraries, and the proportion is substantially more than that in other species (96). Although different filtering and sampling strategies were applied to reduce or eliminate such biases, the effectiveness of these approaches is of a serious concern. As a bottom line, it appears to be safe to claim that mammals have a higher rate of alternative splicing than other vertebrates and invertebrates, given the large difference in magnitude observed in the EST data, as well as other indirect support, such as gene structure and the number of splicing factors. It is worth noting that introns are much longer in mammals than in lower organisms; this is correlated with the observation that cassette exons are the most prevalent pattern of alternative splicing in mammalian genes, accounting for about half of all alternative splicing events, but are less abundant in invertebrates and plant, whose introns are much shorter. This observation invoked the "exon definition" and the "intron definition" models (97). According to these models, exons, rather than introns, are the basic units of recognition at the early stage of splicing, if introns are long (more than 250 to 500 nt). Therefore, the failure of exon recognition often results in the skipping of the exon, rather than the retention of the intron.

A related question to the extensiveness of alternative splicing is which alternative splicing events are likely functional, whereas others represent splicing byproducts (errors or evolutionary intermediates). A direct evaluation of an event in terms of functional significance is difficult, or at least laborious. Two types of evidence are usually suggestive regarding whether an event is potentially functional or not: tissue-specificity and cross-species conservation. Tissue- or development-specific splicing patterns suggest that there is a switch "controlled" by certain splicing factors; alternative splicing events conserved in different species implies that the event survived against purifying selective pressure after the divergence of two or more species. In contrast, an alternative splicing event, for which one or more isoforms are always in low abundance, is often regarded as a byproduct of leaky splicing reactions. With these assumptions, several studies have found that tissue-specific and conserved cassette exons share several general features, but differ from overall cassette exons and constitutive exons. The former have a much higher level of sequence conservation in exons and flanking intronic sequences (98, 99); in some extreme cases, these regions are "ultra-conserved" (100). This observation is intriguing for splicing regulation. More specifically, regulated alternative splicing events have splicing-regulatory signals in subtle balance, which cannot be easily changed during evolution. For example, for a brain-specific exon, it must contain not only the regulatory information to reliably skip the exon in non-brain tissues, but also signals strong enough to activate the exon in brain. Besides enhancers and silencers that are recognized by specific splicing factors, RNA secondary structures might also play a role in the subtle balance by modulating the accessibility of splicing-regulatory elements. As a second feature, tissue-specific and conserved cassette exons are also shorter than overall cassette exons in general. In addition, their exon size is more frequently multiple of three, which means that the skipping of the exon changes only a local portion of protein coding sequences, instead of shifting the reading-frame.

However, it should also be noted that these are the general features of regulated alternative splicing events. For individual cases, none of the arguments is held absolutely. For example, fluctuations of expression levels in multiple ubiquitous splicing factors may also introduce splicing variations to the pre-mRNAs under their combinatorial regulation. Although conserved alternative splicing events are generally more likely to be functional, phenotypic differences among different species must have resulted from species-specific alternative splicing and other levels of gene expression regulation.

With the compilation of exons, introns and alternative splicing events, another focus of bioinformatics studies of splicing has been the predictive identification of splicing-regulatory elements, especially those important for constitutive splicing, or the “general splicing code” (46-48, 101). The general assumption is that different classes of splicing-regulatory elements have different densities in exonic and intronic sequences. For example, RESCUE-ESEs were a subset of hexanucleotides predicted by two criteria: (i) more enriched in constitutive exons than in constitutive introns; (ii) more enriched in constitutive exons with weaker splice sites than those with stronger splice sites. This approach identified 238 ESE hexanucleotides clustered into several motifs, which, in some cases, match the binding preference of several SR proteins. However, the distribution of ESEs in exons might be complicated by amino acid or codon biases. Two alternatives were proposed to address this potential caveat of the RESCUE-ESE approach. One study compared 5’ UTR noncoding exons with introns and 5’ UTR portions of intronless genes (47). Octanucleotides that are enriched in noncoding exons, relative to introns and 5’ UTR of intronless genes, were predicted as putative ESEs, and putative ESSs vice versa. The other study considered the frequency and human-mouse conservation of each individual codon to estimate the expected frequency and conservation rate of each hexamer, assuming the two consecutive codons in the hexamer are independent with each other (48). A higher level of enrichment and conservation than that expected under the independence assumption was interpreted as constraints at the higher order, including exonic splicing-regulatory (ESR) elements. This approach did not explicitly distinguish ESEs and ESSs. All of these approaches have been successful in practice, because a number of the predicted elements strongly enhance or repress exon inclusion when inserted into the alternative exon of a splicing reporter. Due to the stringency of the thresholds used in these studies, there are probably more regulatory elements that fall below the thresholds. Indeed, a recent study used an idea of "neighbor inference" to predict new elements based on the similarity to known ESEs or ESSs (102). A major drawback of these approaches is that in most cases, it is unclear which elements are recognized by which splicing factors.

Previously, studies of global alternative splicing regulation at specific conditions were largely limited by technical difficulties to monitor splicing isoforms and protein-RNA interactions in high throughput. Significant progress has been made in the past decade, although these technologies still await maturing and wide applications. One of the most important high-throughput technologies is splicing microarrays. The feasibility of using microarrays to study splicing regulation was first demonstrated in yeast (103). It was shown that the loss of key mRNA processing factors led to dramatic splicing defects, which could be measured by microarrays. Compared to conventional microarrays, splicing microarrays are designed to be capable of distinguishing splicing variants using probes interrogating exon bodies and/or exon junctions. Later, splicing microarrays were applied in mammalian species (17, 58, 73, 104-106). More recently, the commercial Affymetrix Human Exon 1.0 ST Array has been released, providing the most comprehensive coverage of the genome (107). The Exon Array contains approximately 5.4 million probes or "features" grouped in 1.4 million probe sets, interrogating over one million known or predicted exons. Therefore, each exon is covered by four probes in average. Among various platforms, a major difference in array design is whether to put exon junction probes according to EST/mRNA evidence of alternative splicing, depending on whether one would like to discover novel alternative splicing events or only to measure the abundance of known events.

A common challenge in splicing microarray data processing is to separate the effect of transcription and splicing. Currently, the most popular approach is to estimate overall gene expression level, using probes targeting common regions of all splice isoforms; the intensities of probes targeting each individual splice isoform was then divided by the estimated gene expression level to obtain normalized intensities, or "splicing indices" (103). This simple approach was successfully applied in several previous studies (103, 107). However, the accuracy is limited when signal intensities are saturated or other noises are dominant, which deviates the underlying linear assumption. In addition, different probes have different affinities or behaviors so that direct comparisons of different probesets are difficult. When probe sets are designed to measure all isoforms of the same alternative splicing events, a higher accuracy can be achieved by considering the reciprocal change of different isoforms (73, 108). Other model-based algorithms were also proposed, and some are specific for particular array designs (109-111). As the accuracy keeps improving, splicing microarrays, combined with genetic perturbations by gene knock-out or RNAi, will be a powerful tool for splicing regulation studies (73, 112).

The combination of splicing microarrays with RNP immunoprecipitation (RIP-chip) provides another powerful tool to identify interactions of splicing factors and their substrates. A related technology is CLIP, which cross-links RNP complex using UV exposure *in vivo* (75). By cross-linking, CLIP allows more stringent immunoprecipitation and the identification of more stable protein-RNA interactions, which likely represent directly-bound targets. Applications of CLIP and RIP-chip have provided important insights into mechanisms of neuron-specific splicing regulated by Nova-1/2 (74, 75), and other aspects of splicing regulation (113). These methods will be further powered by the next-generation sequencing technologies, which have emerged very recently.

1.6 Organization

This dissertation focuses on fidelity and regulation of pre-mRNA splicing, to elucidate both the general splicing code, and the more tissue- and development-specific splicing code. My approach is based on the integration of multiple types of genomic data, such as genomic sequences, cDNA/ESTs, cross-species comparisons and splicing microarrays. Statistical analysis (e.g., different hypothesis-testing methods), machine learning approaches (such as hierarchical clustering and support vector machines), different bioinformatics tools (such as Clustal w and sim4 for sequence alignment), and databases and genome browsers, were heavily used to reveal the hidden information underlying the high throughput data, and to generate experimentally testable hypotheses. Chapter 2 describes the compilation, organization and visualization of the data used this study. This mainly includes a database of classified alternative splicing events (dbCASE) and a database of splicing factors (SpliceFac), which can be combined into a splicing knowledgebase in the future. These data, together with various other data that currently have not fit into the databases, form the basis of my different projects described in the later chapters, a majority of which have been published in the past few years. In Chapters 3 and 4, I explore the limited splicing fidelity for mammalian genes, which has important implications in gene evolution. Chapter 3 describes the general impact of limited splicing fidelity due to purifying selective pressure (114). Chapter 4 focuses on a new type of alternative splice site, named dual-specificity splice site (115). These splice sites can function as both 5' and 3' splice sites in different transcripts. Although rare, they are found in different mammalian species and tissues. Dual-specificity splice site provides a good model to explore how they are recognized by the splicing machinery and how they are originated during evolution. Chapter 5 focuses on the general splicing code, to understand the extent and pattern how mammalian genes are constrained for accurate splicing during the course of evolution (116). This study employed the idea of DNA strand asymmetry to provide a model of neutral evolution, which is key to rigorously evaluating the distribution of known splicing-regulatory elements and to predicting new elements. Chapter 6 focuses on the more tissue-specific splicing code and splicing-regulatory networks, using Fox-1 and Fox-2 as a model. Fox-1/2 are specifically expressed in brain, heart and skeletal muscle, and they specifically recognize a UGCAUG RNA element. Using comparative analysis of 28 vertebrate species, this study predicted thousands of conserved Fox-1/2 targets, many of which are important for neuromuscular functions. Combined with evidence from splicing microarray data analysis and RT-PCR validation, this study suggested that Fox-1/2 can activate or repress splicing depending on the locations of their binding sites, and contribute to more complex splicing patterns. The manuscript describing this work is to be submitted soon. Chapter 7 describes the large-scale profiling of splicing changes in prostate cancer, which is a collaborative project with Xiang-Dong Fu lab at UCSD (117). This study is one of the first surveys that demonstrated the unique advantage of splicing microarrays for the classification of cancer and normal samples and the identification of biomarkers. While a general introduction is provided in the previous sections of this chapter, each of the following chapters is self-contained, including introductions of more specific background related to each projects.

1.7 Author contributions

While I initiated, designed, performed the studies described in this dissertation, a number of collaborators made significant contributions to my research. Specifically, Michelle L. Hasting performed the RT-PCR validation of dual-specificity splice sites described in Chapter 4. In the Fox-1/2 project described in Chapter 5, Zuo Zhang performed the experimental validation using overexpression and knock-down of Fox-1 or Fox-2 in HeLa cells, with the help from Shuying Sun at the early stage of this project. The splicing microarray data used in this study were generated by Rosetta/Merck. In Chapter 7, the splicing microarray data were generated by Xiang-Dong Fu lab as a collaboration with our lab.

1.8 References

1. Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA *Cell* **12,** 1-8.

2. Berget, S. M., Moore, C., & Sharp, P. A. (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA *Proc. Natl. Acad. Sci. USA* **74,** 3171-3175.

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

4. Ars, E., Kruyer, H., Morell, M., Pros, E., Serra, E., Ravella, A., Estivill, X., & Lazaro, C. (2003) Recurrent mutations in the NF1 gene are common among neurofibromatosis type 1 patients *J. Med. Genet.* **40,** e82.

5. Gilbert, W. (1978) Why genes in pieces? *Nature* **271,** 501-501.

6. Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., & Zipursky, S. L. (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity *Cell* **101,** 671-684.

7. Chen, B. E., Kondo, M., Garnier, A., Watson, F. L., Puettmann-Holgado, R., Lamar, D. R., & Schmucker, D. (2006) The molecular diversity of Dscam is functionally required for neuronal wiring specificity in *Drosophila Cell* **125,** 607-620.

8. Wang, J., Ma, X., Yang, J. S., Zheng, X., Zugates, C. T., Lee, C.-H. J., & Lee, T. (2004) Transmembrane/juxtamembrane domain-dependent dscam distribution and function during mushroom body neuronal morphogenesis *Neuron* **43,** 663-672.

9. Zhan, X.-L., Clemens, J. C., Neves, G., Hattori, D., Flanagan, J. J., Hummel, T., Vasconcelos, M. L., Chess, A., & Zipursky, S. L. (2004) Analysis of Dscam diversity in regulating axon guidance in *Drosophila* mushroom bodies *Neuron* **43,** 673-686.

10. Bharadwaj, R. & Kolodkin, A. L. (2006) Descrambling DSCAM diversity *Cell* **125,** 421-424.

11. Missler, M. & Sudhof, T. C. (1998) Neurexins: Three genes and 1001 products *Trends Genet.* **14,** 20-26.

12. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome *Nature* **431,** 931-945.

13. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-m., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A*., et al.* (2002) Whole-genome

shotgun assembly and analysis of the genome of Fugu rubripes *Science* **297,** 1301-1310.

14. The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: A platform for investigating biology *Science* **282,** 2012-2018.

15. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M*., et al.* (1996) Life with 6000 genes *Science* **274,** 546-567.

16. Sharp, P. A. (1994) Split genes and RNA splicing *Cell* **77,** 805-815.

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

18. Kan, Z., States, D., & Gish, W. (2002) Selecting for functional alternative splices in ESTs *Genome Res.* **12,** 1837-1845.

19. Maquat, L. E. & Carmichael, G. G. (2001) Quality control of mRNA function *Cell* **104,** 173-176.

20. Maquat, L. E. (2004) Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics *Nat. Rev. Mol. Cell Biol.* **5,** 89-99.

21. Lewis, B. P., Green, R. E., & Brenner, S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans *Proc. Natl. Acad. Sci. USA* **100,** 189-192.

22. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., & Brenner, S. E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements *Nature* **446,** 926-929.

23. Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. A., Blume, J. E., & Ares, M., Jr. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay *Genes Dev.* **21,** 708-718.

24. Pan, Q., Saltzman, A. L., Kim, Y. K., Misquitta, C., Shai, O., Maquat, L. E., Frey, B. J., & Blencowe, B. J. (2006) 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.

25. Modrek, B. & Lee, C. J. (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.

26. Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes *Nucleic Acids Res.* **28,** 4364-4375.

27. Russell, A. G., Charette, J. M., Spencer, D. F., & Gray, M. W. (2006) An early evolutionary origin for the minor spliceosome *Nature* **443,** 863-866.

28. Konig, H., Matter, N., Bader, R., Thiele, W., & Muller, F. (2007) Splicing segregation: The minor spliceosome acts outside the nucleus and controls cell proliferation *Cell* **131,** 718-729.

29. Barbosa-Morais, N. L., Carmo-Fonseca, M., & Aparicio, S. (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion *Genome Res.* **16,** 66-77.

30. Burge, C. B., Tuschl, T., & Sharp, P. A. (1999) Splicing of precursors to mRNAs by the spliceosomes in *The RNA World* eds. Gesteland, R. F., Cech, T. R., & Atkins, J. F. (Cold Spring Harbor Laborotory Press, Cold Spring Harbor, NY), pp. 525-560.

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

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

33. Kim, E., Magen, A., & Ast, G. (2007) Different levels of alternative splicing among eukaryotes *Nucleic Acids Res.* **35,** 125-131.

34. Kim, H., Klein, R., Majewski, J., & Ott, J. (2004) Estimating rates of alternative splicing in mammals and invertebrates *Nat. Genet.* **36,** 915-916.

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

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

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

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

39. Zhu, J. & Krainer, A. R. (2000) Pre-mRNA splicing in the absence of an SR protein RS domain *Genes Dev.* **14,** 3166-3178.

40. Caputi, M., Mayeda, A., Krainer, A. R., & Zahler, A. M. (1999) hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing *EMBO J.* **18,** 4060-4067.

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

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

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

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

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

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

47. Zhang, X. H.-F. & Chasin, L. A. (2004) Computational definition of sequence motifs governing constitutive exon splicing *Genes Dev.* **18,** 1241-1250.

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

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

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

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

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

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

54. McCullough, A. J. & Berget, S. M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection *Mol. Cell Biol.* **17,** 4562-4571.

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

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

57. Olson, S., Blanchette, M., Park, J., Savva, Y., Yeo, G. W., Yeakley, J. M., Rio, D. C., & Graveley, B. R. (2007) A regulator of Dscam mutually exclusive splicing fidelity *Nat Struct Mol Biol* **14,** 1134-1140.

58. Sugnet, C. W., Srinivasan, K., Clark, T. A., Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D*., et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays *PLoS Computat. Biol.* **2,** e4.

59. Lopez, A. J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation *Annu. Rev. Genet.* **32,** 279-305.

60. Meyer, B. J. (2000) Sex in the worm: counting and compensating X-chromosome dose *Trends Genet.* **16,** 247-253.

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

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

63. Patton, J. G., Mayer, S. A., Tempst, P., & Nadal-Ginard, B. (1991) Characterization and molecular cloning of polypyrimidine tract-binding protein: a component of a complex necessary for pre-mRNA splicing *Genes Dev.* **5,** 1237-1251.

64. Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., & Inoue, K. (2003) A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG *EMBO J.* **22,** 905-912.

65. Ho, T., Charlet-B, N., Poulos, M., Singh, G., Swanson, M., & TA, C. (2004) Muscleblind proteins regulate alternative splicing *EMBO J.* **23,** 3103-3112.

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

67. Zhu, H., Hinman, M. N., Hasman, R. A., Mehta, P., & Lou, H. (2008) Regulation of neuron-specific alternative splicing of neurofibromatosis type 1 pre-mRNA *Mol. Cell Biol.* **28,** 1240-1251.

68. Del Gatto-Konczak, F., Bourgeois, C. F., Le Guiner, C., Kister, L., Gesnel, M.-C., Stevenin, J., & Breathnach, R. (2000) 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.

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

70. Oberstrass, F. C., Auweter, S. D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D. L*., et al.* (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation *Science* **309,** 2054-2057.

71. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing *Annu. Rev. Biochem.* **72,** 291-336.

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

73. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M*., et al.* (2005) Nova regulates brain-specific splicing to shape the synapse *Nat. Genet.* **37,** 844-852.

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

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

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

77. Lorson, C. L., Hahnen, E., Androphy, E. J., & Wirth, B. (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy *Proc. Natl. Acad. Sci. USA* **96,** 6307-6311.

78. Cartegni, L. & Krainer, A. R. (2002) Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1 *Nat Genet* **30,** 377-384.

79. Cartegni, L., Hastings, M. L., Calarco, J. A., Stanchina, E. d., & Krainer, A. R. (2006) Determinants of Exon 7 Splicing in the Spinal Muscular Atrophy Genes, SMN1 and SMN2 *Am J Hum Genet* **78,** 63-77.

80. Kashima, T. & Manley, J. L. (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy *Nat Genet* **34,** 460-463.

81. Kashima, T., Rao, N., David, C. J., & Manley, J. L. (2007) hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing *Hum Mol Genet* **16,** 3149-3159.

82. Philips, A. V., Timchenko, L. T., & Cooper, T. A. (1998) Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy *Science* **280,** 737-741.

83. Liquori, C. L., Ricker, K., Moseley, M. L., Jacobsen, J. F., Kress, W., Naylor, S. L., Day, J. W., & Ranum, L. P. W. (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9 *Science* **293,** 864-867.

84. Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., & Krainer, A. R. (2007) The gene encoding the splicing factor SF2/ASF is a proto-oncogene *Nat Struct Mol Biol* **14,** 185-193.

85. Mount, S. M. (1982) A catalogue of splice junction sequences *Nucleic Acids Res.* **10,** 459-472.

86. Stamm, S., Zhang, M. Q., Marr, T. G., & Helfman, D. M. (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons *Nucl. Acids Res.* **22,** 1515-1526.

87. Boguski, M. S., Lowe, T. M. J., & Tolstoshev, C. M. (1993) dbEST -- database for "expressed sequence tags" *Nat. Genet.* **4,** 332-333.

88. Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990) Basic local alignment search tool *J. Mol. Biol.* **215,** 403-410.

89. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence *Genome Res.* **8,** 967-974.

90. Heber, S., Alekseyev, M., Sze, S.-H., Tang, H., & Pevzner, P. A. (2002) Splicing graphs and EST assembly problem *Bioinformatics* **18,** S181-188.

91. Saxonov, S., Daizadeh, I., Fedorov, A., & Gilbert, W. (2000) EID: the Exon-Intron Database--an exhaustive database of protein-coding intron-containing genes *Nucleic Acids Res.* **28,** 185-190.

92. Kim, N., Alekseyenko, A. V., Roy, M., & Lee, C. (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species *Nucleic Acids Res.* **35,** D93-98.

93. Stamm, S., Riethoven, J.-J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N. L., & Thanaraj, T. A. (2006) ASD: a bioinformatics resource on alternative splicing *Nucleic Acids Res.* **34,** D46-55.

94. Kim, N., Shin, S., & Lee, S. (2005) ECgene: Genome-based EST clustering and gene modeling for alternative splicing *Genome Res.* **15,** 566-576.

95. Harrington, E. D., Boue, S., Valcarcel, J., Reich, J. G., & Bork, P. (2004) Estimating rates of alternative splicing in mammals and invertebrates *Nature Genet.* **36,** 916-917.

96. Xu, Q. & Lee, C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences *Nucleic Acids Res.* **31,** 5635-5643.

97. Berget, S. M. (1995) Exon recognition in vertebrate splicing *J. Biol. Chem.* **270,** 2411-2414.

98. Sorek, R. & Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse *Genome Res.* **13,** 1631-1637.

99. Xing, Y. & Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences *Proc. Natl. Acad. Sci. USA* **102,** 13526-13531.

100. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004) Ultraconserved elements in the human genome *Science* **304,** 1321-1325.

101. Yeo, G. W., Nostrand, E. L. V., & Liang, T. Y. (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements *PLoS Genet.* **3,** e85.

102. Stadler, M., Shomron, N., Yeo, G. W.-M., Schneider, A., Xiao, X., & Burge, C. B. (2006) Inference of splicing regulatory activities by sequence neighborhood analysis *PLoS Genetics* **2,** e191.

103. Clark, T. A., Sugnet, C. W., & Ares, M., Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays *Science* **296,** 907-910.

104. Yeakley, J. M., Fan, J. B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S., & Fu, X. D. (2002) Profiling alternative splicing on fiber-optic arrays *Nat. Biotechnol.* **20,** 353-358.

105. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G*., et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22 *Genome Res.* **14,** 331-342.

106. Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., Morris, Q. D*., et al.* (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform *Molecular Cell* **16,** 929-941.

107. Clark, T., Schweitzer, A., Chen, T., Staples, M., Lu, G., Wang, H., Williams, A., & Blume, J. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays *Genome Biol.* **8,** R64.

108. Fehlbaum, P., Guihal, C., Bracco, L., & Cochet, O. (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants *Nucleic Acids Res.* **33,** e47.

109. Wang, H., Hubbell, E., Hu, J.-s., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M. A., Ares, M., Kulp, D. C*., et al.* (2003) Gene structure-based splice variant deconvolution using a microarry platform *Bioinformatics* **19,** i315-322.

110. Cline, M. S., Blume, J., Cawley, S., Clark, T. A., Hu, J.-S., Lu, G., Salomonis, N., Wang, H., & Williams, A. (2005) ANOSVA: a statistical method for detecting splice variation from expression data *Bioinformatics* **21,** i107-115.

111. Shai, O., Morris, Q. D., Blencowe, B. J., & Frey, B. J. (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data *Bioinformatics* **22,** 606-613.

112. Blanchette, M., Green, R. E., Brenner, S. E., & Rio, D. C. (2005) Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila *Genes Dev.* **19,** 1306-1314.

113. Gama-Carvalho, M., Barbosa-Morais, N., Brodsky, A., Silver, P., & Carmo-Fonseca, M. (2006) Genome-wide identification of functionally distinct subsets of

cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors *Genome Biol.* **7,** R113.

114. Zhang, C., Krainer, A. R., & Zhang, M. Q. (2007) Evolutionary impact of limited splicing fidelity in mammalian genes *Trends Genet* **23,** 484-488.

115. Zhang, C., Hastings, M. L., Krainer, A. R., & Zhang, M. Q. (2007) Dual-specificity splice sites function alternatively as 5' and 3' splice sites *Proc. Natl. Acad. Sci. USA* **104,** 15028-15033.

116. Zhang, C., Li, W.-H., Krainer, A. R., & Zhang, M. Q. (2007) RNA landscape of evolution for optimal exon and intron discrimination *Proc. Natl. Acad. Sci. USA* **10.1073/pnas.0801692105**.

117. Zhang, C., Li, H.-R., Fan, J.-B., Wang-Rodriguez, J., Downs, T., Fu, X.-D., & Zhang, M. (2006) Profiling alternatively spliced mRNA isoforms for prostate cancer classification *BMC Bioinformatics* **7,** 202.

Chapter 2

Data sources, compilation, and visualization

2.1 dbCASE—a database of classified alternative splicing events

Previously, extensive efforts have been made to identify alternative splicing events from literature manually, or from sequenced transcripts deposited into GenBank computationally (1, 2). For those computational approaches, the basic idea is to align transcripts to genomic sequences. Introns are identified by alignment gaps flanked by splice site consensuses and exons by alignment blocks. Drawbacks of using these existing databases include

- The lack of documentation and therefore difficult to parse and use the data.
- Many of them are not up-to-date whereas transcripts in GenBank accumulate very quickly.
- Accurate alignment of transcripts and genomic sequences is critical. More accurate and efficient algorithms designed particularly for this type of alignment are being developed.

- It is usually difficult to track supporting transcripts for each alternative splicing event. This is important when one needs to study alternative splicing in particular tissues and conditions.
- Some of them focused on only one or a few species.

Due to these considerations, I have implemented a computational pipeline to detect alternative splicing events from transcript-genome alignment. The pipeline takes transcripts from UniGene (3), which includes both clustered mRNAs and ESTs, and from Refseq (4), as well as genomic sequences downloaded from the UCSC genome browser. The main procedures of the pipeline are described below.

Extracting gene contigs.

Pre-aligned RefSeq transcripts were downloaded from the UCSC genome browser and assigned to Entrez genes according to the transcript-to-gene mapping data downloaded from Genbank (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2accession.gz). The transcripts with ambiguities (e.g. aligned to multiple loci) were removed. For each Entrez gene, the extreme boundaries determined by the transcripts were extended for 3kb on each side to define the boundaries of the gene contig. The gene contig sequences were then extracted.

Transcript-contig alignment.

UniGene transcripts were aligned to the corresponding gene contigs by sim4, which optimizes alignment at gap termini with splice site consensuses (5). Terminal blocks with less than 25 nt were removed. Only high-quality alignment with idenity > 95% and coverage > 85% and no internal insertions in transcripts were kept. Unspliced transcripts were removed to avoid intron contamination. RefSeq transcripts were aligned using the same criteria.

Building splicing graph.

For each gene, the alignments were converted into a graphic representation using directed acyclic graph (DAG), called splicing graph. Splicing graph has been used previously in several different forms (6-8). I used the most flexible representation in which each node in the graph represents a 5’ or 3’ splice site and an edge represents an exon or intron (7)

(**Fig. 1**). The evidence supporting each exon/intron is also recorded. In addition, the uniqueness of my representation is that I allow the same position to be both a 5' splice site and a 3' splice site, which is observed to be necessary during the implementation. The supporting transcripts where recorded separately for each type in this case.

Detecting typical alternative splicing events.

Alternative splicing events of typical types were detected by examining sub network topologies as shown in **Fig. 1**. The supporting evidence was also dumped.

Detecting constitutive exons and introns.

A collection of constitutive exons and introns are very useful as controls in the study of alternative splicing. Using the graphic theory, for the first time, strictly constitutive exons and introns (in the sense of no violation in existing transcripts) can be identified elegantly and efficiently. To do so, I introduced a measure called cumulative degree of exon elicitation (*CDE*), as shown below, to record the number of exon variants spanning a position.

Denote the coordinates of all nodes of a DAG as N_1, N_2, …, N_n, from left (5'end) to right (3'end) and N_0=0 for convenience. The $CDE[x]$, where x is the coordinate on the contig, changes only at nodes along the contig, and is calculated as follows:

```
Initialize $CDE[N_0]=0$
FOR $i$=1 to $n$
$CDE[N_i] = CDE[N_{i-1}]$ + out_degree of $N_i$, when $N_i$ is used as a 3'SS
$CDE[N_i] = CDE[N_{i-1}]$ - in_degree of $N_i$, when $N_i$ is used as a 5'SS
END
```

Similarly, cumulative degree of intron elicitation (*CDI*) was introduced to measure the number of intron variants spanning a position.

```
Initialize $CDI[N_0]=0$
FOR $i$=1 to $n$
$CDI[N_i] = CDI[N_{i-1}]$ + out_degree of $N_i$, when $N_i$ is used as a 5'SS
$CDI[N_i] = CDI[N_{i-1}]$ - in_degree of $N_i$, when $N_i$ is used as a 3'SS
END
```

An *internal* constitutive exon is defined by two *nonterminal* nodes N_iN_j satisfying:

- N_i is a 3' splice site and N_j is a 5' splice site
- $CDI\ [N_i]=0$
- Every node between N_i and N_j are terminal nodes, or equivalently, $CDE[\mathrm{N}_i]$ - #terminal exon starting from N_i =1

A constitutive intron is defined by two *neighboring* nonterminal nodes N_iN_j satisfying:

- N_i is a 3' splice site and N_j is a 5' splice site
- $CDI\ [N_i]=1$
- $CDE[Ni]=0$

More generally, these measures were used to derive the exon inclusion level and transcript coverage at each individual position. To do this, two additional quantities, *CDE_ts* and *CDI_ts* were introduced to measure the number of transcripts that use the position as exons and introns, respectively.

CDE_ts:

```
    Initialize $CDE\_ts[N_0]=0$
    FOR $i$=1 to $n$
    $CDE\_ts[N_i] = CDE\_ts[N_{i-1}]$ + number of supporting transcripts for all edges out of $N_i$, when $N_i$ is used as a 3'SS
    $CDE\_ts[N_i] = CDE\_ts[N_{i-1}]$ – number of supporting transcripts for all edges out of $N_i$, when $N_i$ is used as a 5'SS
    END
```

CDI_ts:

```
    Initialize $CDI\_ts[N_0]=0$
    FOR $i$=1 to $n$
    $CDI\_ts[N_i] = CDI\_ts[N_{i-1}]$ + number of supporting transcripts for all edges out of $N_i$, when $N_i$ is used as a 5'SS
    $CDI\_ts[N_i] = CDI\_ts[N_{i-1}]$ – number of supporting transcripts for all edges out of $N_i$, when $N_i$ is used as a 3'SS
    END
```

Trivially then, transcript coverage is the total number of transcripts that span the position, i.e. (*CDE_ts* + *CDI_ts*). Exon inclusion level is (*CDE_ts* / (*CDE_ts*+*CDI_ts*)).

Up to now, I have applied this pipeline to human, mouse, rat, zebrafish, fly and worm. In the case of human, which has the most number of transcripts, the whole pipeline ran for a few hours in our linux cluster. Various tricks were included to improve the efficiency.

Some of the results are summarized in **Table 1** and **2**. The large variation in the number of alternative splicing events (**Table 1**) is likely due to the higher EST coverage in human and mouse, although it was argued that higher eukaryotes may have more frequent alternative splicing events (9-11). From the perspective of splicing regulation, a more interesting observation to me is that the percentage of cassette exons increased from ~30% in zebrafish to ~50% in the mammalian lineage. Intuitively, this is consistent with the fact that intron size in mammals is much larger than that in lower vertebrates. It was reported that exon skipping is more prevalent for those exons flanked by long introns (12, 13).

Finally, the data generated using the pipeline were imported into a local UCSC genome browser and the web-based database dbCASE (**Fig. 2**). The web interface of dbCASE has two important features. First, it is closely integrated with the genome browser, so that a number of other tracks and powerful features in the browser can be easily utilized. Secondly, since the supporting transcripts and the EST library information for each splicing pattern were recorded, the abundance of each isoform in each tissue type, developmental stage and health state is also visualized in a color-coded format.

In the process of building the database, a new type of splice site, which I call dual-specificity splice site, has been discovered (**Table 2**). These splice sites can function alternatively as 5' splice sites in some transcripts and 3' splice sites in some other transcripts. The detailed information about dual-specificity splice sites is described in Chapter 4.

2.2 SpliceFac—a database of splicing factors

SpliceFac is a splicing-factor-centric database that is analogous to TransFac, a database devoted to transcriptional regulation studies. The final aim of the database is to host information such as expression, protein domains, binding motifs and target RNAs of splicing factors, all of each are essential for splicing regulation.

At the current stage, I have collected ~480 splicing factors and spliceosomal proteins, among which 409 are from human (**Fig. 3A**). This is the most comprehensive collection of known or potential splicing regulators till now. Among them, 223 human proteins are annotated as “RNA binding” or “nucleotide binding” in gene ontology. Since it was reported that there are ~380 RNA binding proteins in mouse (14) and a similar number can be expected in human, the collection of human splicing factors seems to be relatively complete. The splicing factors in other mammalian species can be obtained by mapping orthologous genes, if not available directly.

The main data sources for the collection are mass-spectrum studies for spliceosomal components and published literature for validated splicing factors. Currently, two mass-spectrum studies were included (15, 16). Results from a comprehensive study of splicing factor evolution was also used (17). Literature was analyzed by a semi-automated pipeline. First, 800 PubMed abstracts containing the keyword “splicing factor” were downloaded (as of Dec. 2005). Then the sentences with the keywords were extracted for manual inspection.

A web interface has been created for curate, search and display of the data (**Fig. 3B**).

2.3 A splicing knowledge base

The final goal of this project is to create a knowledge base of splicing regulation, which will integrate dbCASE and SpliceFac together, to provide information of splicing factors and their expression profiles, alternative splicing events and their splicing profiles, the regulatory relationship. This resource will provide invaluable insight into splicing regulatory networks.

2.4 References

1. Kim, N., Alekseyenko, A. V., Roy, M., & Lee, C. (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species *Nucleic Acids Res.* **35,** D93-98.

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

3. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006) GenBank *Nucleic Acids Res.* **34,** D16-20.

4. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins *Nucleic Acids Res.* **33,** D501-504.

5. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence *Genome Res.* **8,** 967-974.

6. Kim, N., Shin, S., & Lee, S. (2005) ECgene: Genome-based EST clustering and gene modeling for alternative splicing *Genome Res.* **15,** 566-576.

7. Sugnet, C., Kent, W., Ares, M. J., & Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice in *Pac. Symp. Biocomput.*, pp. 66-77.

8. Eyras, E., Caccamo, M., Curwen, V., & Clamp, M. (2004) ESTGenes: alternative splicing from ESTs in ensembl *Genome Res.* **14,** 976-987.

9. Harrington, E. D., Boue, S., Valcarcel, J., Reich, J. G., & Bork, P. (2004) Estimating rates of alternative splicing in mammals and invertebrates *Nature Genet.* **36,** 916-917.

10. Kim, E., Magen, A., & Ast, G. (2007) Different levels of alternative splicing among eukaryotes *Nucleic Acids Res.* **35,** 125-131.

11. Kim, H., Klein, R., Majewski, J., & Ott, J. (2004) Estimating rates of alternative splicing in mammals and invertebrates *Nat Genet* **36,** 915-916.

12. Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S.-p., Baldi, P. F., & Hertel, K. J. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing *Proc. Natl. Acad. Sci. USA* **102,** 16176-16181.

13. Sterner, D., Carlo, T., & Berget, S. (1996) Architectural limits on split genes *PNAS* **93,** 15081-15085.

14. McKee, A., Minet, E., Stern, C., Riahi, S., Stiles, C., & Silver, P. (2005) A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain *BMC Dev. Biol.* **5,** 14.

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

16. Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome *Genome Res.* **12,** 1231-1245.

17. Barbosa-Morais, N. L., Carmo-Fonseca, M., & Aparicio, S. (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion *Genome Res.* **16,** 66-77.

2.5 Tables and Figures

Table 1 Statistics of genes and transcripts in dbCASE.

Species (version)	**Gene contigs**	**UniGene** Version transcripts/clusters	**RefSeq**	**Total No. transcripts with contigs**	**No. transcripts aligned** transcripts/genes
human (Mar. 2006)	17,991	Build 196 6,988,853/83,896	39,188	6,274,364	2,884,883/17,072
mouse (Feb. 2006)	18,172	Build 158 4,277,970/64,632	47,932	3,788,678	1,455,256/16,423
rat (Nov. 2004)	9,513	Build 157 771,182/52,204	40,326	404,272	164,239/7,781
zebrafish (May 2006)	8,280	Build 98 1,067,802/40,426	31,025	633,339	327,945/7,885
fly (Apr. 2004)	13,812	Build 46 504,549/ 15,139	20,507	491,107	256,139/11,562
worm (Mar. 2004)	17,102	Build 32 342,498/20,464	23,470	288,516	176,669/14527

Table 2 Statistics of alternative splicing events in dbCASE.

Species (version)		**Cassette exon**	**Alt' 3SS**	**Alt' 5SS**	**Intron retention**	**Mutual Exclusion**	**Dual splice site**
human (Mar. 2006)	all	29,131 (46.18%)	14,281 (22.64%)	9,278 (14.71%)	7,356 (11.66%)	2,444 (3.87%)	594 (0.94%)
	≥ 2ts	12,048 (55.70%)	4,571 (21.13%)	2,859 (13.22%)	1,297 (6.00%)	771 (3.56%)	85 (0.39%)
mouse (Feb. 2006)	all	11,953 (42.05%)	7,405 (26.05%)	4,780 (16.81%)	3,494 (12.29%)	649 (2.28%)	148 (0.52%)
	≥ 2ts	5,245 (52.66%)	2,501 (25.11%)	1,418 (14.24%)	534 (5.36%)	237 (2.38%)	26 (0.26%)
rat (Nov. 2004)	all	1,704 (43.89%)	995 (25.63%)	572 (14.73%)	498 (12.83%)	100 (2.58%)	13 (0.33%)
	≥ 2ts	571 (53.31%)	265 (24.74%)	130 (12.14%)	69 (6.44%)	34 (3.17%)	2 (0.19%)
zebrafish (May 2006)	all	1,047 (26.01%)	1,243 30.87%	1,018 (25.29%)	616 (15.30%)	61 (1.52%)	41 (1.02%)
	≥ 2ts	339 (32.13%)	363 (34.41%)	239 (22.65%)	81 (7.68%)	31 (2.94%)	2 (0.19%)
fly (Apr. 2004)	all	1,072 (32.11%)	1,070 (32.05%)	429 (12.85%)	632 (18.93%)	120 (3.59%)	16 (0.48%)
	≥ 2ts	535 (41.96%)	388 (30.43%)	152 (11.92%)	121 (9.49%)	72 (5.65%)	7 (0.55%)
worm (Mar. 2004)	all	313 (23.87%)	385 (29.37%)	281 (21.43%)	306 (23.34%)	17 (1.30%)	9 (0.69%)
	≥ 2ts	163 (36.22%)	145 (32.22%)	86 (19.11%)	47 (10.44%)	9 (2.00%)	0 (0.00%)

Both of the total number of events and the number of events with ≥ 2 supporting transcripts for each isoform are shown. The percentage of each type in each species is also provided.

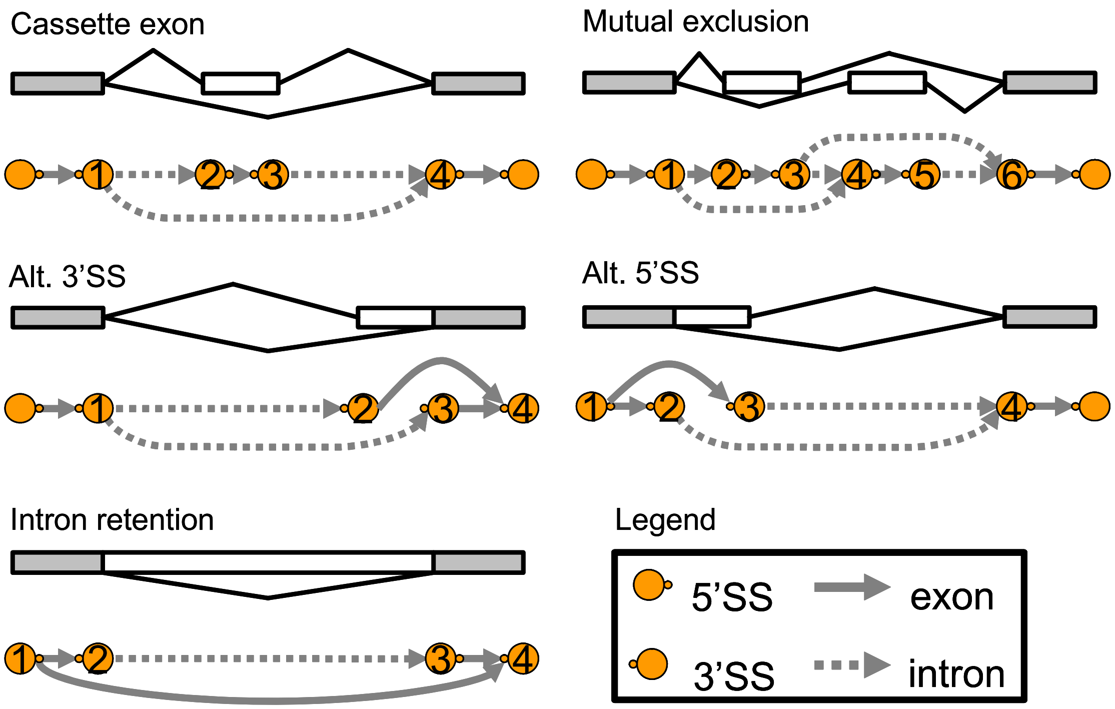


Figure 1 Graph representation of typical AS patterns.

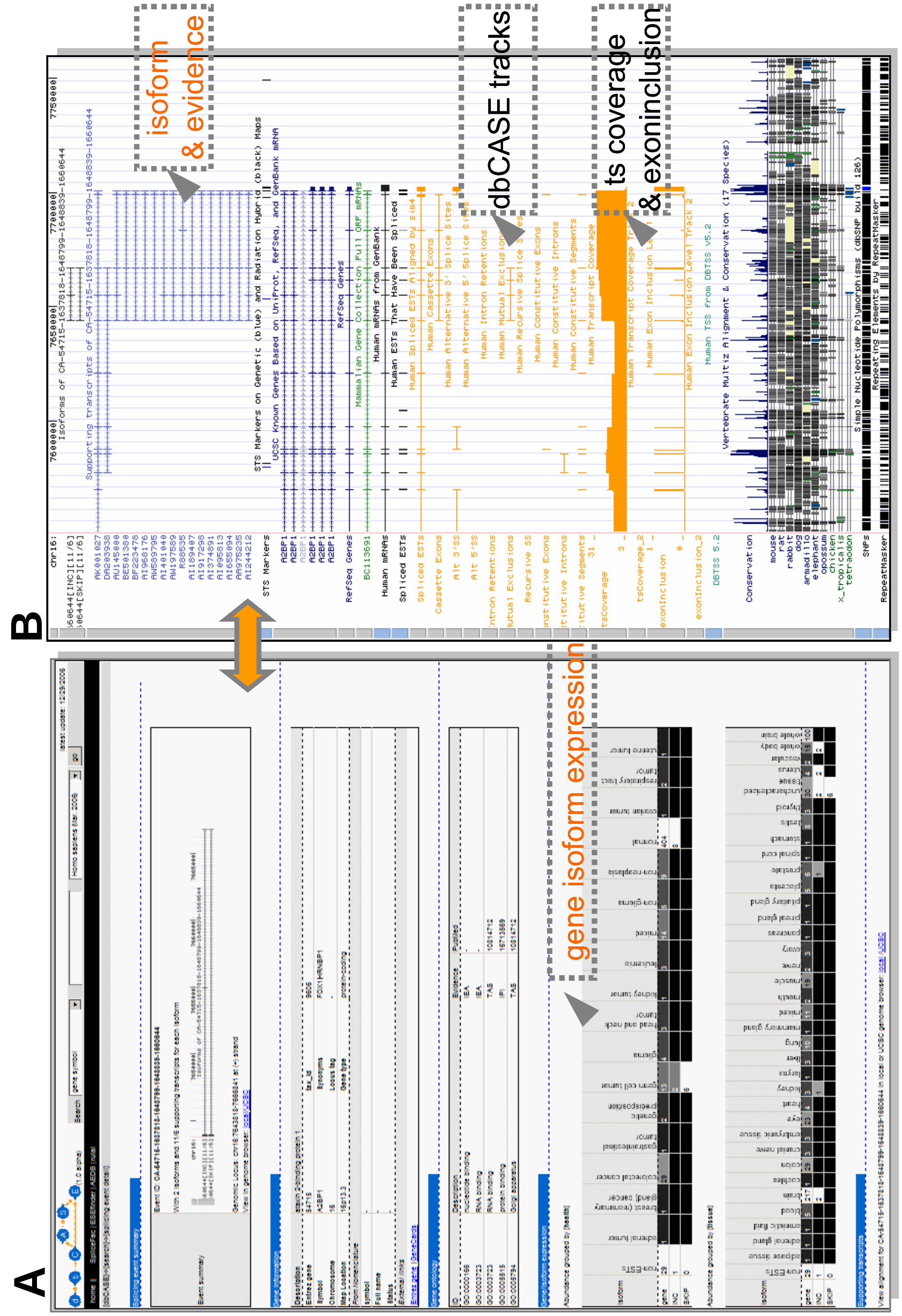
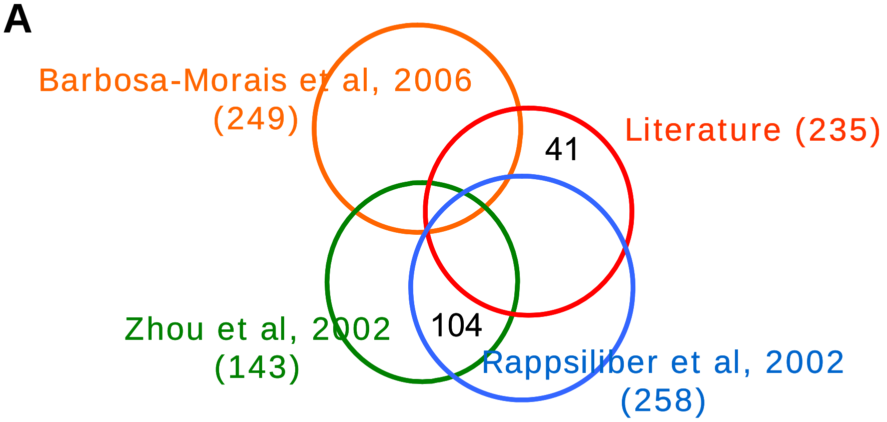
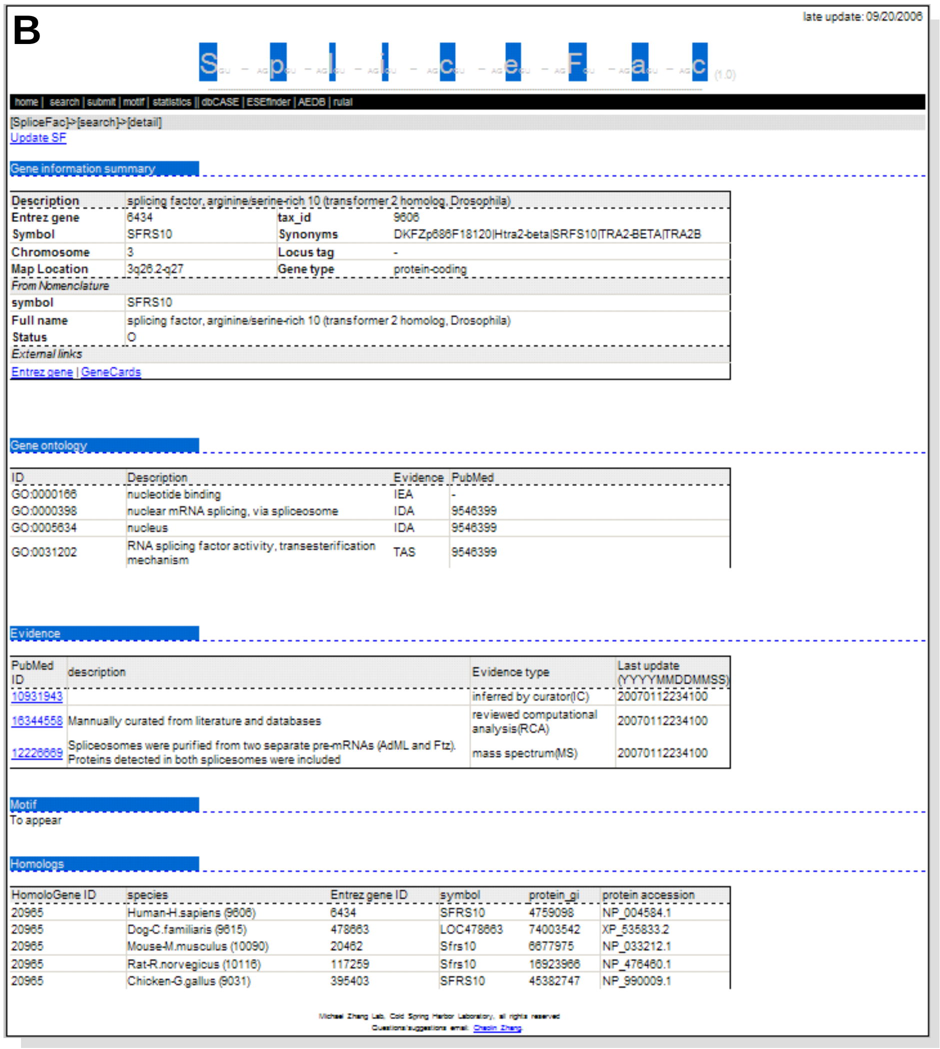


Figure 2 User interface of dbCASE (A) and the local UCSC genome browser (B). The two databases are closely integrated and cross-referenced. Major features are also highlighted.




B
S GU – AG p GU – AG l GU – AG i GU – AG c GU – AG e GU – AG F GU – AG a GU – AG c (1.0)
home | search | submit | motif | statistics || dbCASE | ESEfinder | AEDB | rulal

late update: 09/20/2006

[SpliceFac]->[search]->[detail]

Update SF

Gene information summary

Description	splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila)		
Entrez gene	6434	**tax_id**	9606
Symbol	SFRS10	**Synonyms**	DKFZp686F18120|Htra2-beta|SRFS10|TRA2-BETA|TRA2B
Chromosome	3	**Locus tag**	-
Map Location	3q26.2-q27	**Gene type**	protein-coding
From Nomenclature			
symbol	SFRS10		
Full name	splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila)		
Status	O		
External links			
Entrez gene | GeneCards			

Gene ontology

ID	Description	Evidence	PubMed
GO:0000166	nucleotide binding	IEA	-
GO:0000398	nuclear mRNA splicing, via spliceosome	IDA	9546399
GO:0005634	nucleus	IDA	9546399
GO:0031202	RNA splicing factor activity, transesterification mechanism	TAS	9546399

Evidence

PubMed ID	description	Evidence type	Last update (YYYYMMDDMMSS)
10931943		inferred by curator(IC)	20070112234100
16344558	Mannually curated from literature and databases	reviewed computational analysis(RCA)	20070112234100
12226669	Spliceosomes were purified from two separate pre-mRNAs (AdML and Ftz). Proteins detected in both splicesomes were included	mass spectrum(MS)	20070112234100

Motif

To appear

Homologs

HomoloGene ID	species	Entrez gene ID	symbol	protein_gi	protein accession
20965	Human-H.sapiens (9606)	6434	SFRS10	4759098	NP_004584.1
20965	Dog-C.familiaris (9615)	478663	LOC478663	74003542	XP_535833.2
20965	Mouse-M.musculus (10090)	20462	Sfrs10	6677975	NP_033212.1
20965	Rat-R.norvegicus (10116)	117259	Sfrs10	16923966	NP_476460.1
20965	Chicken-G.gallus (9031)	395403	SFRS10	45382747	NP_990009.1

Michael Zhang Lab, Cold Spring Harbor Laboratory, all rights reserved
Questions/suggestions email: Chaolin Zhang

Figure 3 Collection of splicing factors and spliceosomal proteins.
(**A**) Data sources of SpliceFac include two mass-spectrum studies (15, 16), a previous collection (17) and literature.
(**B**) The web interface of SpliceFac.

Chapter 3

Evolutionary impact of limited splicing fidelity

3.1 Abstract

The functional significance of most alternative splicing (AS) events, especially frame-shifting ones, has been controversial. Using human-mouse comparison, we demonstrate that frame-preserving AS events adapt and get fixed more rapidly than frame-shifting AS events; selection for smaller exon size is stronger in frame-preserving exons than in frame-shifting ones. These results suggest AS events introducing mild changes are generally favored during evolution and explain the excess of shorter, frame-preserving cassette exons in present mammalian genomes.

3.2 Introduction

Alternative splicing (AS), the process of removing introns and joining exons in different combinations, is critical for expanding proteomic diversity and regulating gene expression (1). In humans and rodents, a majority of genes (>60%) express multiple isoforms (2-4). Significant progress has recently been made in understanding AS evolution. It has been proposed that AS accelerates exon creation and loss by relaxing the negative selection pressure against the new minor isoforms, while maintaining the original major isoforms (5-7). In support of this idea, alternative exons in humans and rodents frequently arose after the divergence of these species. Exonization of Alu elements is one evolutionary mechanism and contributes to more than 5% of human AS exons (8-10).

Several reports demonstrated that AS events of functional importance, such as ancestral AS events (e.g., cassette exons with conserved AS pattern across different mammals) and tissue-specific AS events, are associated with a significant increase of frame-preserving preference (FPP) (11-14) and sequence conservation level (15, 16). In contrast, only 40-45% of overall AS events, as shown in cassette exons, preserve the reading frame, whereas the majority introduce premature stop codons (PTCs) and potentially induce nonsense-mediated mRNA decay (NMD). Due to the fact that ancestral and tissue-specific AS events represent a biased and limited subset of the whole population of AS events, it remains controversial in the field whether the widespread frame-shifting AS events are coupled with NMD as a gene expression regulatory mechanism (17) or represent byproducts generated by limited splicing fidelity as evolutionary precursors (18). In this study we sought to address this question in a more general context and evaluate the evolutionary trends of AS from a novel perspective by identifying differential selection pressure for these two categories of AS events.

3.3 Results

Many AS events are probably nonessential

Because isoform abundance of an AS event is positively correlated with its evolutionary age and fitness (5, 7), we examined the distribution of cassette-type AS events extracted from ASD (19) in terms of skipping-to-inclusion ratios (see **Methods** in Online Supplementary Material). The distribution overall is unimodal with a ratio of one at the mode (**Fig. 1 a** and **c**, last row of the heat maps), which is perhaps not surprising. However, it should be noted that transcript counts are discrete in nature. Thus, the distribution is largely truncated, as most AS events have relatively low EST coverage, whereas the minimal transcript count is one. To reduce this effect, we analyzed cassette exons with more supporting transcripts. Intriguing patterns of the distributions arose as we varied the filtering thresholds from 10 to 200 for human, and to 80 for mouse. More precisely, when 10 or more supporting transcripts were required, human and mouse showed a very similar bimodal distribution of the ratios, with two modes at 0.08 and 15 (-1.1 and 1.2 in the log10 scale), respectively (**Fig. 1 a** and **c**, second row from the bottom in each heat map); 60% of these AS events had a ratio of minor to major isoform (RMM) smaller than 0.1. Consistent observations were reported in a recent study focused on genes expressed in stem cells (20). These findings suggest that as the sensitivity of detecting AS events increases by sampling the transcriptome more deeply, it becomes easier to find rare splicing isoforms that probably represent recent evolutionary precursors generated due to limited splicing fidelity. To further support this idea, we examined FPP as a function of RMM. The FPP of AS events with RMM <0.1 had a very similar value (~40%) to that of constitutive exons. Removing these AS events significantly increased the FPP value to the level of known functional AS events (~60%), i.e., ancestral or tissue-specific AS events (**Fig. 1 b** and **d**, **Fig. S1**, **Table S1** and **S2** in Online Supplementary Material). We note that these observations are unlikely to represent an artifact due to degradation of PTC-containing transcripts, because we observed a very similar bimodal distribution of skipping-to-inclusion ratios for frame-preserving cassette exons alone (**Fig. S2** in Online Supplementary Material).

Human-mouse comparison suggests a more rapid fixation of shorter, frame-preserving exons

We reasoned that if certain AS events are nonessential but represent evolutionary precursors, a large fraction of them should be eliminated by negative (purifying) selection pressure during evolution. Frame-shifting and frame-preserving AS events as evolutionary precursors might be under differential selective pressure and thus have different outcomes, because they have distinct effects on protein products. Conversely, differential selective pressure discernible in current species might further support the common incidence of evolutionary precursors. In contrast to the increased evolutionary rate of overall AS events, functional AS events are under much greater purifying selection pressure in both exonic regions and intronic flanks, and evolve more slowly than constitutive exons (13, 15, 16, 21). It is likely that the excess of frame-preserving exons in functional AS events is largely due to differential selective pressure.

To test this hypothesis, we divided cassette exons into frame-preserving and frame-shifting ones, and analyzed the mutation rate of exonic regions and immediate intronic flanks separately for each group. Rarely included cassette exons are more likely to have been recently exonized, and are difficult to match between human and mouse (5). In other words, those that are conserved between human and mouse might represent a very biased sample. Therefore, we further limited our analysis to frequently included cassette exons (skipping/inclusion≤1), for which the skipped-exon isoforms probably evolved later.

Fig. 2a-d shows the results for human cassette exons matched with orthologous mouse exons. Besides the general trend of decreasing mutation rate (which suggests the increase of purifying selection pressure) as the RMM increases, frame-preserving and frame-shifting exons indeed show very different mutation patterns. The extent of conservation increases rapidly for frame-preserving exons as the RMM value increases, whereas the rate of increase is much smaller for frame-shifting exons. More specifically, with the constraint of RMM≥0.6, the synonymous mutation rate, Ks, of frame-preserving exons is 0.32, significantly lower than that of frame-shifting exons, which has a value of 0.54 ($p<10^{-10}$, Fisher's exact test). The opposite trend, namely smaller Ks of frame-shifting AS events than of frame-preserving ones (0.61 vs. 0.54 $p=1.5 \times 10^{-8}$, Fisher's exact test), is observed for AS events with small RMMs (<0.1). Consistent patterns were observed in flanking intronic regions (**Fig. 2c** and **d**). Also, similar results were obtained for mouse cassette exons matched with orthologous human exons (**Fig. 2 e-h**).

Although alternative interpretations might exist, the following scenario can give a parsimonious explanation of the differential mutation pattern. Cassette exons with recently evolved skipping events are essentially similar to constitutive exons. For these AS events, purifying selection pressure tends to eliminate the skipping isoform; because the frame-preserving events affect only local protein sequences, they have a smaller negative selection pressure and a higher tolerance for mutations than frame-shifting AS events. In contrast, for older AS events, negative selection tends to prevent the disruption of either isoform. Frame-preserving events are more enriched in AS events of functional importance and therefore have a stronger purifying selection pressure, at a level similar to what we observed for ancestral cassette exons. The rapid increase of purifying selection pressure for frame-preserving cassette exons suggests that they have a more rapid fixation rate and confer an advantage to the organism, compared to frame-shifting ones. This explains the excess of frame-preserving AS events. Simply put, we propose that AS events that introduce mild changes are generally favored during evolution.

Interestingly, the differential selective pressure also contributes to shaping exon length (**Table 1**). The median length of frame-preserving cassette exons is significantly shorter than that of frame-shifting ones (108 vs. 116, $p=10^{-9}$ for human; 99 vs. 112, $p=8\times10^{-15}$ for mouse, Wilcoxon rank sum test). Similarly, we observed a difference for ancestral cassette exons, which are even shorter (87 vs. 103, $p=2\times10^{-5}$, Wilcoxon rank sum test), but not for constitutive exons. This is entirely consistent with our model. For frame-preserving AS events, inclusion or skipping of short exons as evolutionary precursors introduces even smaller perturbations, whereas for frame-shifting events, exon length plays a more moderate role because most or all of the protein would be affected due to codon changes or to NMD. A long-standing observation from earlier AS studies is that AS exons are generally shorter than constitutive exons (22), although this was interpreted as a result of suboptimal exon definition by the spliceosome (23). However, the difference of exon size between frame-preserving and frame-shifting cassette exons cannot be explained by suboptimal exon definition (see also **Fig. S3** in Online Supplementary Material).

3.4 Concluding remarks

Consistent with the accelerated gene evolution associated with AS, the present study suggests the common occurrence of evolutionary precursors that might be nonessential and negatively selected. We identified differential selective pressure between frame-shifting and frame-preserving AS events, which suggests that AS events introducing mild changes are generally favored. This extrinsic selective force gives a plausible explanation for the excess of shorter, frame-preserving cassette exons in present mammalian genomes, among other possible mechanisms. Our observations are consistent with the work of Wen et al., who suggested that AS events that introduce a short variable region might have a larger functional impact than expected (24). Finally, our results support the notion that NMD is generally more a mechanism for quality control (17, 25) rather than one for geneexpression regulation (18).

3.5 Acknowledgements

This work was supported by NIH grants GM74688 (A.R.K.) and HG01696 (M.Q.Z.). The authors would like to thank Philip J. Smith and Andrew D. Smith for helpful comments on the manuscript.

3.6 References

1. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing *Annu. Rev. Biochem.* **72,** 291-336.

2. Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., & Devon, K. (2001) Initial sequencing and analysis of the human genome *Nature* **409,** 860-921.

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

4. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22 *Genome Res.* **14,** 331-342.

5. Modrek, B. & Lee, C. J. (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.

6. Xing, Y. & Lee, C. (2006) Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes *Nature Rev. Genet.* **7,** 499-509.

7. Zhang, X. H. F. & Chasin, L. A. (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons *Proc. Natl. Acad. Sci. USA* **103,** 13427-13432.

8. Sorek, R., Ast, G., & Graur, D. (2002) Alu-containing exons are alternatively spliced *Genome Res.* **12,** 1060-1067.

9. Lev-Maor, G., Sorek, R., Shomron, N., & Ast, G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons *Science* **300,** 1288-1291.

10. Dagan, T., Sorek, R., Sharon, E., Ast, G., & Graur, D. (2004) AluGene: a database of Alu elements incorporated within protein-coding genes *Nucleic Acids Res.* **32,** D489-492.

11. Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., & Lee, C. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation *Nucleic Acids Res.* **32,** 1261-1269.

12. Sorek, R., Shamir, R., & Ast, G. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20,** 68-71.

13. Xing, Y. & Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences *Proc. Natl. Acad. Sci. USA* **102,** 13526-13531.

14. Xing, Y. & Lee, C. J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing *PLoS Genet.* **1,** e34.

15. Sorek, R. & Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse *Genome Res.* **13,** 1631-1637.

16. Sugnet, C. W., Srinivasan, K., Clark, T. A., Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays *PLoS Computat. Biol.* **2,** e4.

17. Pan, Q., Saltzman, A. L., Kim, Y. K., Misquitta, C., Shai, O., Maquat, L. E., Frey, B. J., & Blencowe, B. J. (2006) 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.

18. Lewis, B. P., Green, R. E., & Brenner, S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans *Proc. Natl. Acad. Sci. USA* **100,** 189-192.

19. Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V., & Muilu, J. (2004) ASD: the alternative splicing database *Nucleic Acids Res.* **32,** D64-69.

20. Pritsker, M., Doniger, T. T., Kramer, L. C., Westcot, S. E., & Lemischka, I. R. (2005) Diversification of stem cell molecular repertoire by alternative splicing *Proc. Natl. Acad. Sci. USA* **102,** 14290-14295.

21. Sugnet, C., Kent, W., Ares, M. J., & Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice in *Pac. Symp. Biocomput.*, pp. 66-77.

22. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., & Zhang, M. Q. (2000) An alternative-exon database and its statistical analysis *DNA Cell Biol.* **19,** 739-756.

23. Berget, S. M. (1995) Exon Recognition in Vertebrate Splicing *J. Biol. Chem.* **270,** 2411-2414.

24. Wen, F., Li, F., Xia, H., Lu, X., Zhang, X., & Li, Y. (2004) The impact of very short alternative splicing on protein structures and functions in the human genome *Trends Genet.* **20,** 232-236.

25. Maquat, L. E. & Carmichael, G. G. (2001) Quality control of mRNA function *Cell* **104,** 173-176.

3.7 Tables and Figures

Table 1: Comparison of exons length between frame-preserving and frame-shifting exons.

p-values were calculated by a Wilcoxon rank sum test.

		Median size		
Exon types		**frame-preserving**	**frame-shifting**	***p***
Human	constitutive	126	127	0.5
	cassette	108	116	10^{-9}
Mouse	constitutive	123	127	0.02
	cassette	99	112	8x10^{-15}
Ancestral cassette		87	103	2x10^{-5}

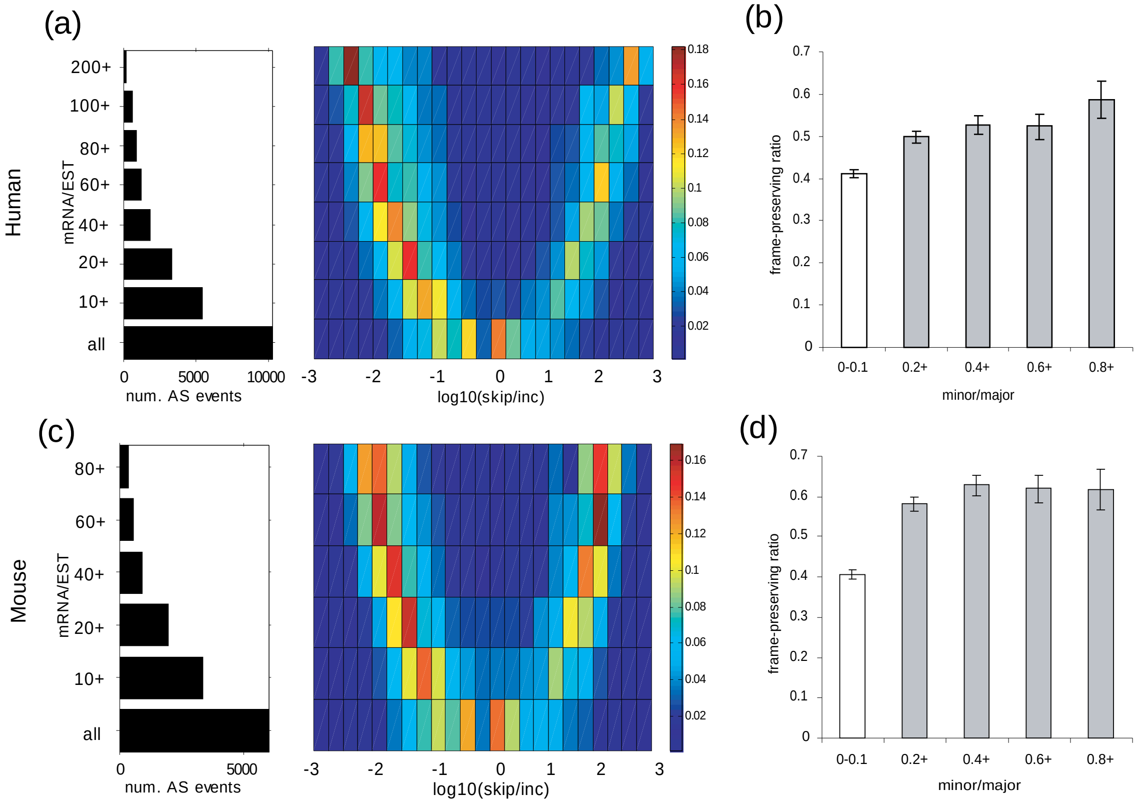
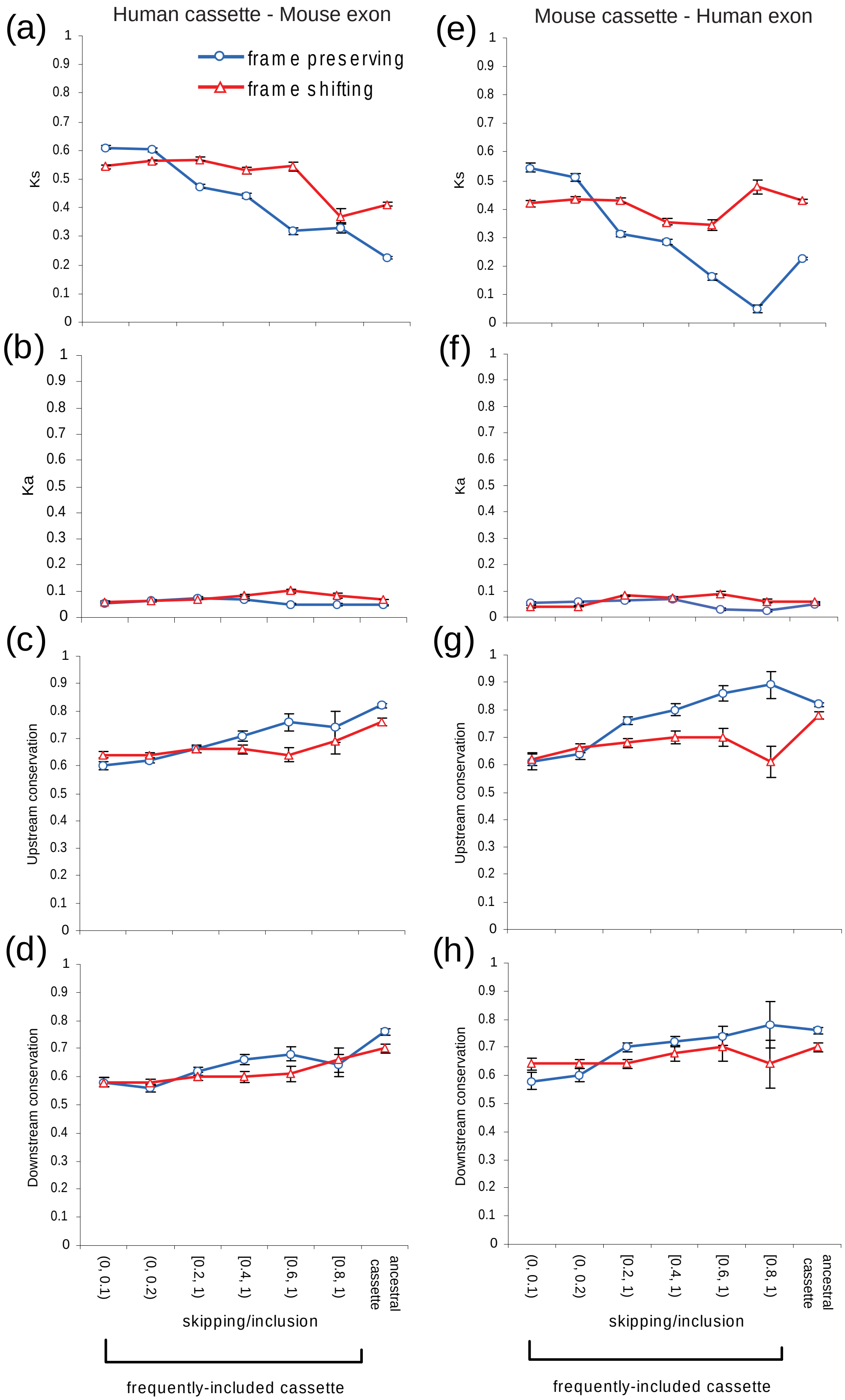


Figure 1: Distribution and frame-preserving preference (FPP) in terms of the relative isoform abundance for human (a and b) and mouse (c and d) cassette-type AS events.

(**a** and **c**) The distribution of skipping-to-inclusion ratios (skip/inc) of AS events is represented as a heatmap in the right panel. Different filtering thresholds (10 to 200 mRNAs or ESTs for human, and to 80 for mouse) were applied to remove AS events with a low EST coverage. With each threshold, the number of remaining events is shown in the bar plot in the left panel and the distribution of the ratio is shown in the corresponding row of the heatmap. The color in each cell represents the probability of AS events with a ratio in the corresponding interval. The color scale is shown at the right of each heatmap.

(**b** and **d**) The frame-preserving preference (FPP) is calculated as a function of the ratio of the minor isoform to the major isoform (RMM). Error bars represent the standard deviation estimated from binomial distribution.



(See legend at the next page)

Figure 2: Exonic and intronic conservation/mutation pattern of orthologous exons in human and mouse.

Frame-preserving exons are shown in green lines with circles and frame-shifting exons are shown in red lines with triangles, respectively.

(**a-d**) show the results of human cassette exons matched with mouse exons while (**e-h**) show the results of mouse cassette exons matched with human exons, respectively.

(**a**) and (**e**), synonymous mutation rate, Ks, in exons;

(**b**) and (**f**), non-synonymous mutation rate, Ka, in exons;

(**c**) and (**g**) the nucleotide conservation level of the intronic region 50 nt upstream of the exon;

(**d**) and (**h**) the nucleotide conservation level of the intronic region 50nt downstream of the exon. Error bars represent the standard error.

3.8 Supplementary materials

Methods

Human and mouse transcript-confirmed exons

Human and mouse transcript (mRNA/EST)-confirmed exons were extracted from the Alternative Splicing Database (ASD) (Release 2, April 2005) (1). The AltSplice database in ASD is a computationally derived collection of alternative splicing (AS) events of human and mouse based on alignment of EST/mRNA sequences to the corresponding genomic sequences with high quality and minimal redundancy. In the ASD database, all transcript/genome alignments with ambiguities were removed. A confirmed intron is defined by an alignment gap of genomic sequence flanked by two splice sites of known types. A confirmed exon is defined by an alignment match flanked by two confirmed introns; therefore, only internal exons are considered as being confirmed. Confirmed introns/exons that overlap with each other indicate alternative splicing events. In human, AltSplice has 16,293 genes, including 9,945 (61%) with one or more alternative splicing events. In mouse, AltSplice has 16,352 genes, including 8,211 (50%) alternatively spliced ones. The higher percentage of alternatively spliced genes in human is probably due to the higher EST coverage.

In this study, we considered only splicing events involving GT-AG intron boundaries. In total, 133,926 and 121,202 exons, plus 50 nt of flanking intronic sequences, were extracted for human and mouse, respectively. Cassette exons are those included in some transcripts but skipped in others, without affecting the two neighboring exons (denoted as SCE, for simple cassette exons, in ASD). We extracted 10,196 and 5,992 cassette exons for human and mouse, respectively. We also compiled a set of 30,892 and 37,313 exons likely to be constitutively spliced in human and mouse, respectively. A brief summary of frame-preserving preference (see below) and human-mouse conservation is given in supplementary Table 1 and supplementary Figure 1. We also compared other features, such as intron phase bias (data not shown). All these general statistics are similar to and consistent with those reported previously (e.g. (2-4)).

Exon inclusion/skipping level

For each cassette exon, the number of supporting transcripts for the inclusion and the skipping isoforms was also extracted from the ASD database (1). The number of supporting transcripts was used as an approximate measure of the abundance of the exon inclusion/skipping isoform, as done previously (5, 6). The ratio of the skipping to inclusion isoform or the ratio of minor isoform to the major isoform (RMM) was used to estimate the relative abundance of the two isoforms.

Frame-preserving preference

An exon is defined as frame-preserving if its length is a multiple of three nt, and as frame-shifting otherwise (e.g. (4)). The inclusion or skipping of a frame-preserving exon will not change the reading frame, thus affecting only the local protein sequence, unless the cassette exon has one or more PTCs, which is relatively infrequent. For a set of exons, the frame-preserving preference (FPP) is defined as the fraction of frame-preserving exons out of the total. The standard deviation of the FPP is estimated by Binomial distribution, std(FPP)=sqrt[FPP×(1-FPP)/n]. The statistical significance of the difference in the FPP between two exon groups is tested using a two-way contingency table, (group1 frame-preserving, group1 frame-shifting; group2 frame-preserving, group2 frame-shifting) by a Fisher's exact test (4).

To generate results given in Fig. 1 (b and d), we used cassette exons with ≥ 10 supporting transcripts and ≥3 transcripts for the minor isoform. The filtering permits a more precise estimate of the relative abundance of the two isoforms. FPPs were calculated for cassette exons with different ranges of relative abundance in the two isoforms. In particular, we regard an isoform as being rare if the relative abundance is less than 0.1. The thresholds of filtering and intervals were somewhat arbitrary and determined empirically, but the results seem to be robust with different thresholds.

Identification of orthologous exons for human mouse comparison

Orthologous exon pairs were identified between human and mouse as previously described, with minor adaptations (4). In brief, othologous gene pairs were downloaded using the Ensembl BioMart tool (formerly known as EnsMart) (7) (November, 2005).

Then, each exon in the human gene was aligned to each exon in the orthologous mouse gene at both the nucleotide and protein levels using Clustalw (8). For the protein-level alignment, nucleotide sequences were translated in all frames. Only those frames without a stop codon were retained for alignment. The reading frame that gave the best amino acid identity in each orthologous comparison was identified. Orthologous exon pairs were defined as those with reciprocal best alignment with nucleotide identity ≥ 60% and amino acid identity ≥ 50%. Generally, a real exon pair has a much higher conservation level than the thresholds. We identified mouse orthologous exons for human cassette exons, and vice-versa. We also identified ancestral cassette exons (orthologous cassette exons that can be included and skipped in both species).

To generate the results presented in Fig. 2, the same filtering criteria as described above (≥ 10 supporting transcripts and ≥3 transcripts for the minor isoform) were used. In addition, cassette exons with skip/inc >1 were not analyzed here due to reasons described in the main text. Subsets of exons with different relative abundance of the two isoforms were defined similarly as in Fig. 1. For each subset, synonymous (Ks) and non-synonymous (Ka) mutation rates in the exons, and sequence conservation level in the flanking intronic regions were estimated as described below.

Calculation of synonymous (Ks) and non-synonymous (Ka) mutation rates

Following a previously described approach (4, 9), the protein alignment generated to identify orthologous exons was used to realign the two nucleotide sequences of each orthologous exon pair. Gaps were removed. Synonymous and non-synonymous substitutions/sites were estimated by the Yang-Nielsen maximum-likelihood method, using the program yn00 in the PAML package (10). For each subset of exons, the number of substitutions/sites was added up to calculate the overall synonymous (Ks) and non-synonymous (Ka) mutation rates by the ratio of the two sums. The standard deviation of the ratio (Ka or Ks) was estimated by Binomial distribution, the same as the estimation of standard deviation of FPP, as described above. The difference in Ks (Ka) for two exon groups was tested using the total number of substitutions/sites and Fisher's exact test, as described above and in previous studies (4).

Intronic sequence conservation

For each orthologous exon pair, we aligned both the upstream and downstream intronic flanking sequences (200 nt in each region) using Clustalw. The 50 positions immediately upstream or downstream of the cassette exons were used to estimate the intronic conservation level. For each subset of exons, the average conservation level and standard error was calculated. We used robust estimates, i.e. median and scaled MAD (median absolute deviation), which impose no assumption of normality. More precisely, the standard error is estimated by MAD/sqrt(n), where n is the number of sequences. Note that in the software package R, MAD is scaled to be equivalent with the standard deviation for normal distributions.

Comparison of exon length for frame-shifting and frame-preserving exons

To generate the results in Table 1, we used all constitutive and cassette exons, as well as ancestral cassette exons. The difference in median of exon size for frame-preserving exons and frame-shifting exons was tested by a Wilcoxon rank sum test. To generate the results presented in Supplementary Fig. 3, we filtered cassette exons by requiring ≥50 supporting transcripts. Exons were then broken down into three subsets according to the relative abundance of the two isoforms. For each subset, we calculated the average and the standard error by robust estimates, i.e. Median and MAD/sqrt(n).

Statistical analyses

All statistical analyses and tests were performed in the open source software R.

URLs

ASD: http://www.ebi.ac.uk/asd/

Biomart: http://www.ensembl.org/Multi/martview.

Supplementary References

1. Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V., & Muilu, J. (2004) ASD: the alternative splicing database *Nucleic Acids Res.* **32,** D64-69.

2. Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., & Lee, C. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation *Nucleic Acids Res.* **32,** 1261-1269.

3. Sorek, R. & Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse *Genome Res.* **13,** 1631-1637.

4. Xing, Y. & Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences *Proc. Natl. Acad. Sci. USA* **102,** 13526-13531.

5. Modrek, B. & Lee, C. J. (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.

6. Kan, Z., States, D., & Gish, W. (2002) Selecting for functional alternative splices in ESTs *Genome Res.* **12,** 1837-1845.

7. Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., & Birney, E. (2004) EnsMart: a generic system for fast and flexible access to biological data *Genome Res.* **14,** 160-169.

8. Thompson, J., Higgins, D., & Gibson, T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice *Nucleic Acids Res.* **22,** 4673-4680.

9. Nekrutenko, A., Chung, W.-Y., & Li, W.-H. (2003) ETOPE: evolutionary test of predicted exons *Nucleic Acids Res.* **31,** 3564-3567.

10. Yang, Z. & Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models *Mol. Biol. Evol.* **17,** 32-43.

11. Sugnet, C. W., Srinivasan, K., Clark, T. A., Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays *PLoS Computat. Biol.* **2,** e4.

12. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse *Nat. Genet.* **37,** 844-852.

Supplementary Tables and Figures

Table S1: Frame-preserving preference (FPP) and exonic/intronic conservation of orthologous exons in human and mouse.

The median and the standard error are shown. Cassette exons overall are more similar to constitutive exons than ancestral cassette exons.

Exon type (hs17 vs mm5)	Exon no.	FPP (%)	Exon nucl. id. (%)	Exon a.a. id. (%)	UIF id.(%)	DIF id. (%)
exon vs exon	86063	39.6(0.2)	87.8(0.0)	93.8(0.0)	60.0(0.0)	56.0(0.1)
const. vs const.	9645	40(0.5)	87(0.1)	93(0.0)	60(0.1)	56(0.1)
exon vs cass.	4700	48(0.7)	88(0.1)	92(0.2)	64(0.2)	58(0.3)
cass. vs exon	2956	49(0.9)	89(0.1)	93(0.2)	66(0.3)	62(0.3)
cass. vs cass.	809	62(1.7)	93.0(0.3)	94.6(0.4)	78.0(0.6)	70.0(0.7)

Table S2: Frame-preserving preference of tissue-specific cassette exons from the literature.

	Frame-preserving	all	Fraction(%)	Data source
Brain-specific	106	171	62	(11)
Muscle-specific	17	28	61	(11)
Validated nova targets†	29	35	83	(12)

†a few exons not matched in ASD were excluded

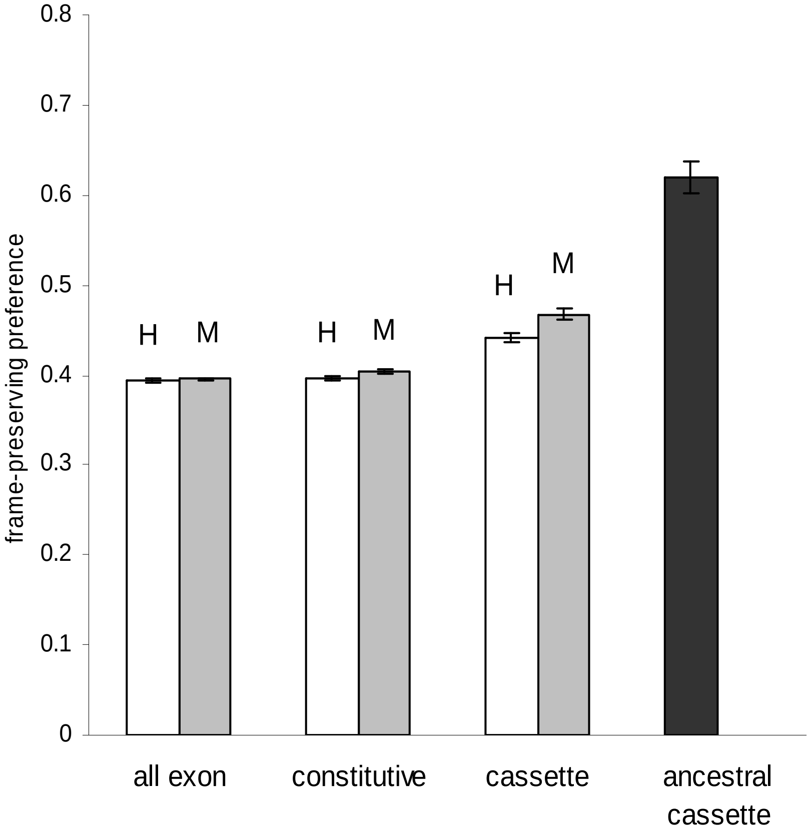


Figure S1: Frame-preserving preference (FPP) of all exons, constitutive exons, cassette exons and ancestral cassette exons in human (H) and mouse (M).

The error bars show the standard deviation estimated from a binomial distribution. Cassette exons overall are more similar to constitutive exons than to ancestral cassette exons.

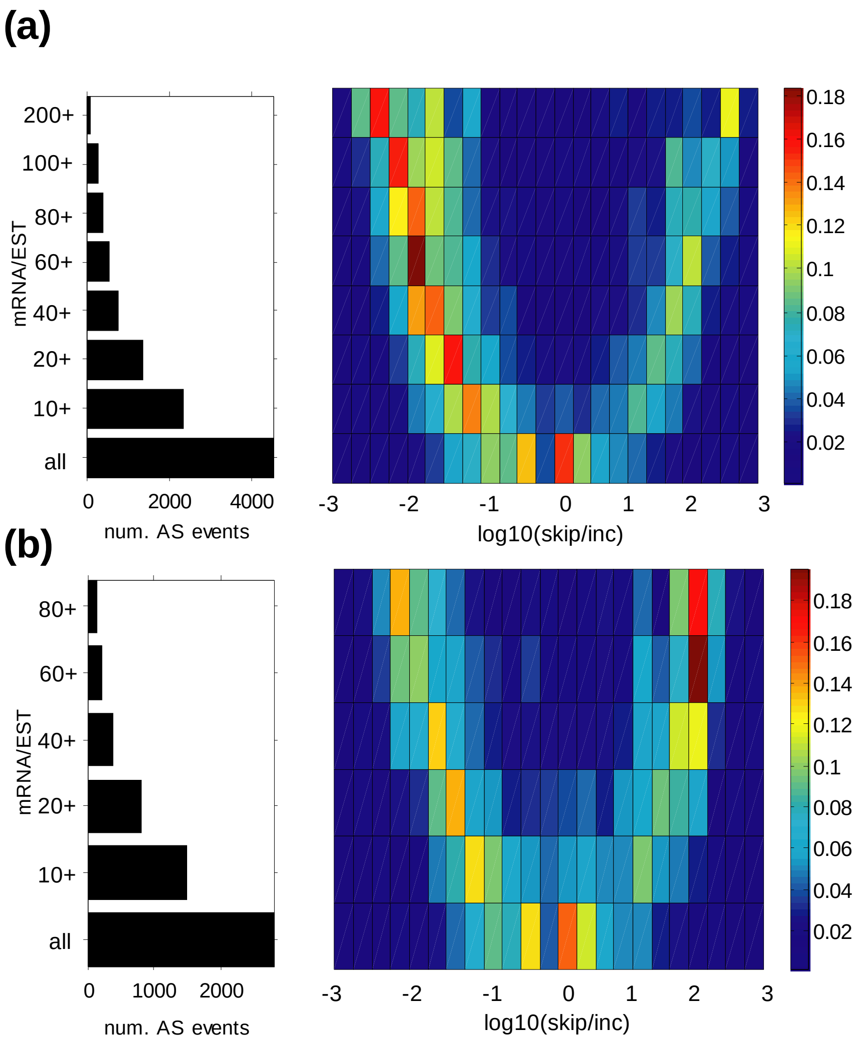


Figure S2: Distribution of frame-preserving cassette exons in terms of relative isoform abundance.

(a) human, (b) mouse. See the legend of Fig. 1 in the main text for details. The bimodal distribution (when a threshold of supporting transcripts was applied) is very similar to that observed for all cassette exons. This suggests that as the transcriptome is sampled more deeply, it is easier to find low-abundance splicing isoforms. This low abundance is largely independent of NMD. Also note that in both Fig. 1 and Fig. S2, the peak of cassette exons with rare-skipping is almost twice that of cassette exons with rare-inclusion, which implies that leaky or aberrant exon skipping is more prevalent than inclusion. As an alternative interpretation, it is easier for random mutations to attenuate splicing signals than to create them in intronic sequences. These observations cannot be explained by NMD either.

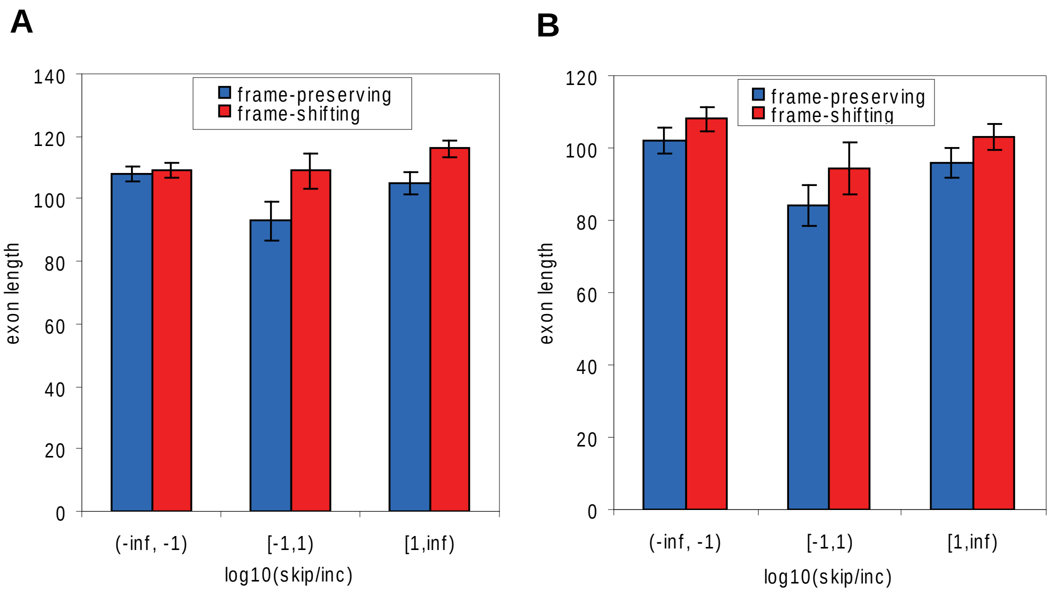


Figure S3: Comparison of exon size for cassette exons with different inclusion levels.

Cassette exons with ≥50 supporting transcripts were divided into three bins according to skipping-to-inclusion ratio (<0.1, between 0.1 and 10 and ≥10). Frame-preserving and frame-shifting exons were compared separately (shown in blue and red respectively). The bars show median exon sizes. Error bars show standard errors. (**a**) human, (**b**) mouse. Note that exons included at intermediate levels are shorter than those predominantly skipped or included. The difference seems to be larger for frame-preserving exons than frame-shifting ones. If suboptimal exon definition by the spliceosome is the primary reason for the shorter size of AS exons, rarely included exons should be even shorter, which contradicts our observation.

Chapter 4

Dual-specificity splice sites function alternatively as 5′ and 3′ splice sites

4.1 Abstract

As a result of large-scale sequencing projects and recent splicing-microarray studies, estimates of mammalian genes expressing multiple transcripts continue to increase. This expansion of transcript information makes it possible to better characterize alternative splicing events and gain insights into splicing mechanisms and regulation. Here we describe a novel class of splice sites we call dual-specificity splice sites, which we identified through genome-wide, high-quality alignment of mRNA/EST and genome sequences, and experimentally verified by RT-PCR. These splice sites can be alternatively recognized as either 5' or 3' splice sites, and the dual splicing is conceptually similar to a pair of mutually exclusive exons separated by a zero-length intron. The dual splice site sequences are essentially a composite of canonical 5' and 3' splice-site consensus sequences, with a CAG|GURAG core. The relative use of a dual site as a 5' or

3' splice site can be accurately predicted by assuming competition for specific binding between spliceosomal components involved in recognition of 5' and 3' splice sites, respectively. Dual-specificity splice sites exist in human and mouse, and possibly in other vertebrate species, although most sites are not conserved, suggesting that their origin is recent. We discuss the implications of this unusual splicing pattern for the diverse mechanisms of exon recognition, and for gene evolution.

4.2 Introduction

Eukaryotic genes are split into exons and introns, which in the vast majority of cases are marked by a GU-dinucleotide (5' splice site) at the exon/intron boundary and an AG-dinucleotide (3' splice site) at the intron/exon boundary. To produce a mature transcript from a pre-mRNA, the introns are spliced out and the exons are ligated by a large protein/snRNA complex, the spliceosome (1, 2). The accuracy and efficiency of exon and intron recognition and splicing are dictated by: (i) primary splicing signals, including the splice sites, a polypyrimidine tract, and a branch site (2); (ii) nearby exonic or intronic regulatory sequences acting as splicing enhancers or silencers (3-5); (iii) spatial and structural constraints, such as exon and intron size (6, 7) and RNA secondary structure (8); and (iv) interactions of these *cis*-acting elements with splicing factors (9). Any compromise or disruption of these splicing elements, or changes in the levels or properties of the factors, may result in regulated alternative splicing (AS) or aberrant splicing events (10).

With the availability of genome sequences and a large amount of mRNA/EST data, especially in human and mouse, genome-wide bioinformatic analysis has revealed that a majority (> 60%) of mammalian genes are alternatively spliced in various patterns (11, 12). Typical types of AS events include exon skipping/inclusion (cassette exons), alternative 5' or 3' splice sites, mutually exclusive exon use, intron retention, and various combinations thereof (10). In spite of the complexity of splicing patterns and regulation, in all of these cases, 5' and 3' splice sites are defined unambiguously and recognized by distinct sets of spliceosomal components, usually at the earliest stages of spliceosome assembly (Fig. 1A) (1). The splice sites have degenerate consensus sequences, although GU and AG are nearly invariant at the 5' and 3' intronic borders, respectively.

Interestingly, CAG|GU defines the consensus sequence of both 5' and 3' splice sites, although with a different extent of degeneracy (Fig. 1C). This observation raises interesting questions concerning how the splicing machinery distinguishes 5' and 3' splice sites, and whether the same site can be used as both a 3' and a 5' splice site.

In this study, we investigate unusual alternative splicing events associated with splice sites that can be used as either a 5' or a 3' splice site. We refer to these sites as dual-specificity splice sites (or dual splice sites). We detected these dual-specificity sites using high-quality mRNA/EST and genome sequence alignment evidence. In these cases, a particular splice site is used as a 3' splice site in some transcripts, and in other transcripts the same site is used as a 5' splice site. When the dual splice site is recognized as a 3' splice site, the sequences upstream of the site are removed as an intron, whereas the sequences downstream are retained as an exon. However, this situation is reversed in altenative isoforms, in which the dual site is used as a 5' splice site and the sequences downstream of the site are removed as an intron (Fig. 1 B). Thus, the resulting exon/intron flip-over in different isoforms affects the nature of the protein products. We experimentally validated the occurrence of dual-specificity splicing *in vivo* by RT-PCR and direct sequencing, and found that the use of the site as a 5' or 3' splice site can vary in a tissue-specific manner. Bioinformatic analysis revealed unique features that are consistent with the dual-specificity character, and predictive of the splicing outcome. The implications for protein coding and gene-structure evolution are also discussed. We conclude that the use of dual-specificity splice sites as either a 5' or 3' splice site represents a novel class of alternative splicing.

4.3 Results

Identification and classification of dual-specificity splice sites

We built a database of classified alternative splicing events (dbCASE) using high-quality transcripts (mRNA/EST) and genome alignment for multiple species. A data structure called splicing graph (13) was applied and extended to efficiently detect various alternative and constitutive splicing events and to track supporting transcripts (see Materials and Methods). During this process, we found that previous data structure could not represent the transcript data in some cases, because of the presence of dual splice sites. In total, we found 594 human (and 195 mouse) putative dual splice sites with supporting transcript (mRNA/EST) evidence. We also extracted strictly constitutive exons and introns (in the sense that no violating transcripts were detected) as a comparative dataset to further analyze the nature of these dual splice sites.

Because most canonical introns have GU and AG dinucleotides at their 5' and 3' termini, respectively (14), we first examined whether the dual splice sites conform to this AG|GU rule. Overall, 155 dual splice sites (26%) conform to the AG|GU rule. This percentage is lower than that expected compared with constitutive splice sites (Table 1). There are several explanations that may account for this difference. First, sites with few supporting transcripts may be unreliable, as they could reflect aberrant splicing or RT-PCR errors. Second, repetitive elements, sequencing errors in the transcripts (especially ESTs) or in the genome, polymorphisms, and transcripts from paralogous or pseudo genes may result in spurious alignments. The third point, which is not mutually exclusive with the two preceding explanations, is that we observed 64 human (and nine mouse) genes with clusters of dual splice sites. These genes seem to be highly conserved across vertebrate species, but are enriched in exonic SNPs (data not shown). They account for about half of the total number of putative dual splice sites. Most of these sites (~85%) do not match the AG|GU pattern, and it is not clear whether they are authentic examples of dual splice sites, or rather represent artifacts.

To increase the level of confidence in dual-splice-site prediction, we explored ways to increase the stringency of our criteria for dual-splice-site classification. The percentage of AG|GU sites increased greatly when two or more supporting transcripts were required for each isoform (Table 1). We also considered gene transcripts with only one dual splice site (singletons) by removing all genes with two or more sites, in order to eliminate potential noise from other classes of transcripts, as described above. This filtering step further increased the proportion of AG|GU sites. For example, 23 of 26 (88.5%) singleton sites with three or more supporting transcripts for each isoform conformed to the AG|GU pattern; this percentage is significantly higher compared to constitutive splice sites (*p*=0.0006 for 5' splice sites, *p*=10^{-14} for 3' splice sites, Fisher's exact test). Thus, we surmise that most authentic dual splice sites follow the AG|GU rule,

which is likely an important feature to specify dual splicing, probably by the U2-type spliceosome (2).
To characterize the features of dual splice sites, we derived a high-confidence dataset by limiting dual splice sites to AG|GU sites with two or more supporting transcripts for each isoform. We further removed nine sites from the *UBC* gene—because this gene contains multiple repetitive coding units (15), which are prone to alignment uncertainties—and also three other sites that lacked perfectly matching alignments in sequences flanking the sites. The final high-confidence dataset has 36 dual-specificity splice sites (supporting information (SI) Table 2), which were used for the analyses below. Among these splice sites, 11 (31%) have RefSeq or mRNA supporting evidence for both isoforms, whereas the remaining 25 (69%) have only ESTs as supporting evidence for one or both isoforms.

The dual splicing pattern can be classified according to the nature of the resulting alternative transcripts. The most prevalent class of dual splice sites is associated with the first exon (12 of 36 cases) (class I, SI Fig. 5). This is unlikely to be due to sequence-alignment artifacts, because all spurious terminal exons shorter than 25 nucleotides were removed, so that each intron is flanked by two reliable exons. Instead, this first-exon preference suggests a possible link between alternative promoters and dual-splice-site choice (SI Fig. 5). Other dual sites create an upstream or downstream alternative exon (class II, SI Fig 6 and class III, SI Fig. 7), or result in intron retention (class IV, SI Fig. 8), or exon truncation (class V, SI Fig. 9).

Dual splice sites resemble the 5' and 3' splice site consensus sequences

To study the specificity of recognition as 5' splice sites and 3' splice sites more quantitatively, we derived the position weight matrices (PWMs) of dual splice sites, and canonical 5' and 3' splice sites from constitutive exons (16) (Fig. 1C). Compared to the constitutive splice sites, it is readily discernible that the PWM of dual splice sites (Fig. 1D) is roughly the juxtaposition of the intronic portions of the constitutive 5' and 3' splice site matrices, with CAG|GURAG (R represents A or G) as a core in the consensus. The GC content around dual splice sites is higher than that of the corresponding portions of constitutive splice sites (SI Table 3). This could reflect either the fact that exonic sequences generally have a higher GC content than intronic sequences (11), or perhaps unknown mechanistic reasons related to the recognition and splicing of dual splice sites. One could argue that the resemblance of the dual splice site matrix to both canonical 5' and 3' splice site matrices of constitutive splice sites may be an artifact of contamination with both types of splice sites, which are erroneously classified as dual splice sites. To exclude this possibility, we scored each individual dual splice site with both canonical 5' and 3' splice site matrices using previous methods (16) (see Materials and Methods for details).

As shown in Fig. 2, the canonical 5' and 3' splice sites of constitutive exons fall into two distinct, yet overlapping, populations in the space of 5' and 3' splice site matrix scores. Most 5' splice sites have low scores using the PWM for 3' splice sites, and vice-versa. In contrast, dual splice sites have relatively high scores using both matrices (Fig. 2A and C). For example, only 2-4% of constitutive splice sites have both scores for a single site greater than the first quantile (0.25 in the abscissa in Fig. 2C), whereas ~50% (19 of 36) of dual splice sites have both matrix scores greater than the same threshold.

To ensure that this difference between constitutive and dual splice sites is not an artifact reflecting our choice of dual splice sites with the AG|GU pattern, which conforms to the consensus of both 3' and 5' splice sites, we performed a stringent comparison of dual splice sites to the subset of constitutive splice sites with the AG|GU pattern (Fig. 2B and C). This increased the percentage of constitutive splice sites with high scores by both PWMs, which nevertheless was still significantly lower than that of dual splice sites. For example, only 8-13% of AG|GU constitutive splice sites have both scores greater than the first quantile, compared with ~50% for dual splice sites ($p<10^{-7}$ in both comparisons with 5' and 3' splice sites, Fisher’s exact test).

Thus, the resemblance of dual splice sites to both 5' and 3' splice site consensus motifs strongly suggests that they are authentic splice sites with dual specificity as both 5' and 3' splice sites. It is also worth noting that relatively few dual splice sites have top scores (e.g., greater than the third quantile, 0.75 in the abscissa) for both matrices (Fig. 2C), most likely reflecting the difficulty to simultaneously satisfy the constraints of both matrices in a perfect manner.

Competitive recognition predicts splicing outcome

We further reasoned that if the dual splice sites are authentic, the sequence should dictate the competition between 5'-splice-site-associated and 3'-splice-site-associated spliceosomal components, which would be reflected in the relative use of each site, and hence the number-of-transcripts evidence for each isoform. The difference between 5' and 3' splice site matrix scores of each dual splice site, Δb, reflects the log-likelihood ratio of the site being recognized as a 5' splice site to it being recognized as a 3' splice site (Equation 2 in Materials and Methods). We assumed a linear relationship between the binding-likelihood ratios to splicing ratios, and examined their Pearson correlation. Indeed, we observed a significant correlation (R^2=0.2, p=0.006), which means that 20% of the variation in splicing outcome can be explained by the binding affinity to the splice sites (Fig. 3A). A simple classifier according to Δb at the threshold of zero gives 26 of 36 (72%) correct predictions. This correlation and accuracy of prediction is surprising, given that the number of sequenced transcripts pooled from all sources is only an approximation of the real splicing outcome (17), and that other sequences around the splice sites are also likely to be important determinants of splice-site selection.

To test the latter hypothesis, we evaluated the importance of upstream and downstream splice sites in determining splicing outcome. We reasoned that the strength of the splice site that pairs with the dual 5' or 3' splice site across the exon [as per exon definition (6)] may influence the splicing outcome. For simplicity, we limited our analysis to dual sites that give rise to alternative exons, i.e., class II and class III, as defined above (see also SI Figs. 6 and 7). In the high-confidence dataset, six of 36 cases belong to this category. To expand the sample size, we examined 109 AG|GU dual sites with a single supporting transcript for either isoform, and with perfect local alignment, and included nine additional cases in the category. For each of the 15 cases in total, we calculated the scores of the upstream 3' splice site and downstream 5' splice site, together with the scores of the dual site, and derived a measure of competition Δb_2 (see Equation 3 in Materials and Methods for details). This measure can have two alternative interpretations: the difference in the strength of exon definition of the two isoforms, or the difference in the strength of alternative 5' and 3' splice-site competition. As shown in Fig. 3B, the competitive recognition of dual splice sites alone measured by Δb explains

19% of the variation, consistent with the results in Fig. 3A, although the significance level drops due to the limited sample size (p=0.1). A classifier according to Δb at a threshold of zero gives 10 of 15 (67%) correct predictions. Importantly, including upstream and downstream splice sites into the competition model explains 32% of the variation (p=0.03) (Fig. 3C). A classifier according to Δb_2 at a threshold of zero gives 12 of 15 (80%) correct predictions. Therefore, we conclude that the strength of the upstream and downstream splice sites, and probably other regulatory sequences, also contribute to the dual splicing pattern.

Splice sites are used as both 5' and 3' splice sites in cells

To confirm that dual splice sites are used as both 5' splice sites and 3' splice sites, we analyzed splicing of the endogenous Smac/Diablo (*DIABLO*), *UBE2C*, *POLR2G*, and *UROD* transcripts in two human cell lines. In each case, RT-PCR analysis verified the presence of the two isoforms with the expected sizes (Fig. 4 and SI Fig. 10; see primer sequences in SI Table 4). Each pair of isoforms was identified in HeLa cells and the neuronally-derived Weri-Rb1 cell line. The use of the predicted 5' and 3' splice sites of the Smac/Diablo and *UBE2C* dual splice sites was further confirmed by sequence analysis of the RT-PCR products (data not shown).

Dual splice sites are essentially alternative splice sites, and are potentially subject to regulation. We tested the relative use of the Smac/Diablo and *UBE2C* dual 5' and 3' splice sites in a number of tissues and cell lines. We observed variations in the use of the splice sites (Fig. 4), suggesting that use of the dual splice sites is regulated. If a specific *trans*-acting factor determines whether the site is used as a 5' or 3' splice site, then a specific cell type or tissue might be expected to show a general preference for the 5' or 3' splice site of all dual splice sites. However, there did not appear to be a consistent bias for the 5' or 3' splice site in any of the tissues or cell-types we tested, for the two pre-mRNAs we examined.

To explore the functional implications of dual splicing, we examined the splicing patterns that can potentially generate functional protein products (SI Table 2). In six cases (17%), we found protein products for both isoforms, and the alternatively spliced region of each isoform was at least partially coding. In 21 cases, protein products for one or both isoforms were not found, most likely due to the incompleteness of the protein sequence database in GenBank (18) and/or to the presence of premature termination codons (PTCs) in some isoforms that are presumably subject to nonsense-mediated mRNA decay (NMD) (19). In the remaining cases, the dual splice sites were in the untranslated regions, making it difficult to link transcripts to protein sequences directly, although protein sequences that are compatible with the transcripts were found.

Dual splice sites in mouse

Dual splice sites are not limited to human: we also found evidence for 195 putative dual splice sites in mouse. Using the same filtering criteria as applied in human, the mouse high-confidence dataset contains 18 sites (SI Table 5). The difference in number is likely due to the fact that EST coverage in human is significantly higher than that in mouse (seven million compared to four million). Dual splice sites were also detected in rat, zebrafish and fly, although infrequently, and with less supporting evidence (data not shown). We performed a detailed analysis of mouse dual splice sites in the same way as we did for human. The properties of mouse dual splice sites, such as the motif itself, were generally very similar to those of human sites, as described above. We performed a human-mouse comparison by examining the conservation of dual-splice-site sequences and the splicing patterns. Although the splice sites are often conserved at the sequence level, it appears that the conservation of flanking exonic and/or intronic sequences is low (SI Figs. 5-9, and data not shown). Most of the sites lack supporting evidence for conservation of the dual splicing pattern, except in two cases: *MYL6* and *PHC* (data not shown). Both of these sites follow the AG|GU rule in both species. However, neither of these two sites was included in our high-confidence set, due to an insufficient number of supporting transcripts. Therefore, the conservation rate in dual splicing appears to be very low, with an upper bound of 5% [2 of 38 (36+2)], which is much lower than that of cassette-type splicing events (10-20%) (20-22). Indeed, we could only detect a single isoform of Smac/Diablo and *UBE2C* in the mouse neuronal cell line NSC-34 and in mouse NIH-3T3 fibroblasts (Fig. 4).

4.4 Discussion

Large-scale sequencing projects in the past decade, and recent applications of splicing microarrays have made clear the extent and complexity of alternative splicing in mammalian genes (23). In this study, we identify a novel class of splice site and associated alternative splicing pattern. A dual splice site is a composite of canonical 5' and 3' splice sites, which makes it possible for a single site to be recognized as either a 5' splice site or a 3' splice site, and results in an exon becoming an intron and vice-versa (exon/intron flipping). There was only one previously documented example of a dual-specificity splice site in the *IRF3* gene (24). In this case, dual splicing generates isoforms that can potentially code for proteins with different functions. We now show that this form of alternative splicing is more prevalent than previously appreciated. We identified hundreds of potential dual splice sites in human and mouse, among which at least 36 in human and 18 in mouse were identified with high confidence. The greatly expanded list of dual sites allowed us to uncover unique features of these sites.

Several lines of evidence, including multiple supporting transcripts, the resemblance of the sites to both 5' and 3' splice site consensus motifs, the correlation between binding specificity and splicing outcome, and the presence in different species, strongly suggest that alternative splicing via dual sites is an authentic pattern. For several cases, the presence of these sites and the splicing pattern was further validated by RT-PCR and sequencing. It is possible that many of the dual splice sites not included in our high-confidence dataset are also authentic, but currently have a limited number of supporting ESTs or mRNA transcripts, for reasons implicit in the splicing event. For example, the alternative splicing events associated with the use of the 5' or 3' splice site may be rare in certain tissues, or one of the splicing events may generate a premature termination codon, resulting in a transcript that is subject to NMD.

The capacity of a dual splice site to switch its identity between a 5' splice site and a 3' splice site has implications for many aspects of pre-mRNA processing, and raises important questions regarding the mechanisms of splice-site recognition, regulation, and competition. First, to our knowledge, dual splice sites are the first type of splice site to lack unambiguous identity as either a 5' or a 3' splice site. The use of a dual splice site likely involves competition between the 5' and 3' splice sites through a stochastic process, as the two isoforms can coexist in the same tissue type. It is of interest to know at which step of spliceosome assembly the choice of 5' or 3' splice site is made. Second, what are the determinants of dual-splice-site use? Except for two cases, the dual alternative splicing pattern does not appear to be conserved between human and mouse. However, in many cases, including two that we tested (Fig. 4), the sequence of the dual splice site is conserved. Despite this sequence conservation, the sites do not appear to be used as dual splice sites in mouse. In fact, there are thousands of splice sites that have dual character, comparable to the observed dual splice sites, but they do not appear to have dual splicing. Our preliminary analysis suggests that the flanking splice sites also contribute to the splicing outcome, together with the dual sites. Other splicing signals, such as the strength of the polyrimidine tract and the distribution of splicing enhancers and silencers, are also likely important determinants of dual-splice-site use.

At the present time, the functional significance of this unusual AS pattern is not clear. Our results suggest that most dual splice sites have a recent evolutionary history, appearing independently in each species. Recently, introns with significant sequence similarities at their 5' and 3' splice sites were described (25). It was proposed that sequences bearing cryptic splice sites can be duplicated to serve as the termini of a new intron. Such a mechanism could be one possible evolutionary origin of dual splice sites, before the duplicated cryptic splice sites had a chance to evolve into unambiguous 5' or 3' splice sites.

As with other alternative splicing patterns, many of the new isoforms might have arisen as splicing errors or may represent evolutionary precursors (26). However, by inserting an exon and simultaneously deleting another exon, dual splicing may in some cases generate a novel transcript with adaptive value, and thus serve as a mechanism for genomic diversification and expansion of coding capacity. In some cases, both isoforms appear to be abundant. For instance, there are 344 transcripts aligned to the *WDR73* locus, among which 43 (13%) and 99 (29%) directly support one of the two isoforms resulting from dual splicing, respectively. In addition to cases that are predicted to yield unproductive transcripts by inducing NMD, we found several cases, including

Smac/Diablo and *UBE2C*, in which both isoforms code for distinct protein products with potentially altered biochemical properties (SI Table 2). Our RT-PCR analysis reveals that both isoforms of Smac/Diablo and *UBE2C* are abundant at the mRNA level. Furthermore, the levels of each isoform vary among tissues and cell types, suggesting regulation of the use of the dual splice site as a functional 5' or 3' splice site. This regulation may have important functional consequences for protein activity. In the case of Smac/Diablo, which codes for a pro-apoptotic protein, the alternative isoforms have different abilities to bind to effector molecules, as well as differential cellular localization, although the dual splicing was not previously noted (27).

There are interesting similarities and differences between dual splice sites and the previously reported recursive splice sites, which are thought to be used as intermediate steps in the splicing of long introns (28, 29). The consensus motifs for both types of splice sites look like a composite of the canonical 5' and 3' splice-site consensus motifs. However, in reported examples of recursive splicing, a splice site first functions as a 3' splice site and then, following ligation to the upstream exon, a 5' splice site is regenerated. This new 5' splice site is subsequently spliced to a downstream 3' splice site. Thus, recursive splice sites are generated, in part, as a result of the splicing reaction, in contrast to dual splice sites, for which both the 5' and 3' splice sites are present in the pre-mRNA and are functional.

Another difference between these two classes of splice sites is that recursive splicing at intronic sites does not directly affect the final mRNA product, whereas alternative splicing of dual splice sites does. In addition, although in principle the two sequential steps of recursive splicing might be reversed, with a 5' splice site used first, and regenerating a functional 3' splice site, a recent study argued against this reversibility (29). Therefore, competition is not involved in recognition of the recursive splice site as a 5' or 3' splice site in the first splicing reaction, because only one functional splice site is initially present and used. In contrast, for dual splice sites, both the 5' splice site and the 3' splice site are present simultaneously and probably compete for binding of their respective splicing factors. Steric hindrance presumably forces the use of a dual site in a given pre-mRNA molecule as either a 5' or 3' splice site, because once 3'-splice-site factors bind to the site, 5'-splice-site factors are effectively excluded, and vice-versa—a consideration that also applies to microexons (30).

Despite the above differences, it is possible that some dual splice sites could function as sites of recursive splicing as well. For 10 of 36 high-confidence dual splice sites, there is transcript evidence that the exon in which the dual splice site resides can be skipped (e.g., Fig. 4A and C). Recursive splicing at a dual splice site would result in an mRNA isoform lacking an exon, which would be indistinguishable from a mature mRNA arising from a conventional exon-skipping event. Examples of recursive splicing resulting in exon skipping have been described (28, 29). More direct experimental evidence will be required to determine whether exon skipping is actually generated by recursive splicing at the dual splice sites we found.

In summary, by using transcripts and genome alignment in human and mouse, as well as experimental validation, we have identified and characterized a novel class of splice sites with dual specificity as 5' and 3' splice sites. The functional significance of these sites and of the AS events they specify is underscored by their direct effects on the corresponding protein products, in some cases in a tissue-specific manner. Importantly, this novel class of alternative splicing via dual splice sites suggests even greater versatility of the splicing machinery than was previously recognized.

4.5 Materials and methods

Detection of splicing patterns with splicing graphs

We built a database of classified alternative splicing events (dbCASE, http://rulai.cshl.edu/dbCASE) using high-quality transcripts (mRNA/EST) and genome alignment for human and mouse (coverage >85%, identity >95%). Briefly, transcripts from UniGene (ftp://ftp.ncbi.nih.gov/repository/UniGene/, build 196 for human, build 158 for mouse) and RefSeq (ftp://ftp.ncbi.nih.gov/refseq/release/, release 20) (31) were aligned to genomic sequences (hg18 and mm8) using sim4 (32). The alignment of all transcripts to the same gene locus was then converted into a splicing graph, in which each splice site is represented by a node and each exon/intron is represented by an edge (13). In contrast to Sugnet et al., we allowed the same position to be both 5' and 3' splice site, and the transcript evidence was recorded for each form separately, which was critical for this study. AS patterns (in particular dual-specificity splice sites) were detected by analyzing subnetwork topologies. In addition, strictly constitutive exon and introns (in the sense of no violation of transcript evidence) can be detected efficiently by graphic analysis.

Construction of position weight matrices for canonical and dual splice sites

To measure the presumptive binding specificity of the spliceosome, we first constructed position weight matrices (PWMs) for canonical 5' and 3' splice sites from constitutively spliced exons (27,556 in human and 36,262 in mouse, with four or more supporting transcripts). Thirty nucleotides surrounding the splice junction (15 nucleotides on each side) were extracted, and PWMs were built from these sequences (16). Each dual splice site, as well as constitutive splice site, was scored by both matrices as follows:

$$s^{5ss} = \sum_i \log_2(f_{i,b_i}^{5ss} / f_{b_i}^{0}) \qquad (1A)$$

and

$$s^{3ss} = \sum_i \log_2(f_{i,b_i}^{3ss} / f_{b_i}^{0}), \qquad (1B)$$

where s^{5ss} (s^{3ss}) is the score of the 5' (or 3') splice site matrix, i is the position in the matrix, b_i is the base of the site at position i. f_{i,b_i}^{5ss}, f_{i,b_i}^{3ss} and $f_{b_i}^{0}$ represent the frequency of base b_i in 5' splice sites, 3' splice sites, and background sequences, respectively.

A matrix of dual splice sites was built in a similar manner.

Competition at dual splice sites or by exon definition

We considered two models of competition to determine the splicing outcome at a dual splice site. In the first model, the recognition of the dual splice site as a 5' or 3' splice site results from the competition of 5'-splice-site-associated and 3'-splice-site-associated spliceosomal components at the dual splice site. Therefore, the difference between 5' and 3' splice site matrix scores of each dual splice site reflects the log-likelihood ratio of the site being recognized as a 5' splice site to it being recognized as a 3' splice site.

$$\Delta b = s^{5ss} - s^{3ss} = \sum_i \log_2(f_{i,b_i}^{5ss} / f_{i,b_i}^{3ss}) = \log_2[P(5ss)/P(3ss)] \qquad (2)$$

In the second model, the splicing outcome results from competition between exon definition by pairing the dual splice site with the upstream 3' splice site, versus with the downstream 5' splice site (Fig. 1B). The competition is represented by

$$\Delta b_2 = \left(s_{up}^{3ss} + s_{dual}^{5ss}\right) - \left(s_{dual}^{3ss} + s_{down}^{5ss}\right) = \left(s_{dual}^{5ss} - s_{down}^{5ss}\right) - \left(s_{dual}^{3ss} - s_{up}^{3ss}\right) \qquad (3)$$

where the scores of the dual sites are shown by the subscripts, s_{up}^{3ss} is the matrix score of the upstream 3' splice site and s_{down}^{5ss} is the matrix score of the downstream 5' splice site. An alternative interpretation of this model is the difference in the strength of alternative 5' (3') splice-site competition, as shown on the right of Equation (3).

Identification of protein products

For each dual splice site, representative supporting transcripts were retrieved from dbCASE and were searched against the non-redundant protein database of GenBank (18). All protein sequences with significant matches (>10 amino acids) were retrieved and BLATed against the genomic sequence in the UCSC genome browser (http://genome.ucsc.edu) (33). The protein sequences that aligned properly with the desired pattern were subsequently identified.

Statistical analysis

Fisher’s exact test in R was used to evaluate the significance of two-by-two contingency tables (34).

RT-PCR

RNA collected from cells using Trizol Reagent (Invitrogen, Carlsbad, CA) or RNA from tissue samples (Clontech, Mountain View, CA) was reverse transcribed using Superscript II reverse transcriptase (Invitrogen) with oligo dT primers. PCR with AmpliTaq Gold (Roche) was carried out for 40 amplification cycles (95 ^{o}C for 30 s, 60 ^{o}C for 60 s, and 72 ^{o}C for 60 s) in reactions containing [-^{32}P]dCTP. Primer sequences are provided in SI Table 4. Products were separated on 6% native polyacrylamide gels. Quantitation was based on phosphorimage analysis (Fujifilm FLA-5100).

4.6 Acknowledgements

This study was supported by NIH grants GM42699 (A.R.K.) and GM74688 (M.Q.Z., A.R.K.), and by the Dart Neurogenomics Alliance Funds (M.Q.Z).

4.7 References

1. Moore, J. M., Query, C. C., & Sharp, P. A. (1993) Splicing of precursors to mRNA by the spliceosome in *The RNA World*, eds. Gesteland, R. F. & Atkins, J. F. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 303-357.

2. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing *Annu. Rev. Biochem.* **72,** 291-336.

3. Ladd, A. & Cooper, T. (2002) Finding signals that regulate alternative splicing in the post-genomic era *Genome Biol.* **3,** reviews0008.

4. Zheng, Z.-M. (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression *J. Biomed. Sci.* **11,** 278-294.

5. Hastings, M. L. & Krainer, A. R. (2001) Pre-mRNA splicing in the new millennium *Curr. Opin. Cell Biol.* **13,** 302-309.

6. Berget, S. M. (1995) Exon recognition in vertebrate splicing *J. Biol. Chem.* **270,** 2411-2414.

7. Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S.-p., Baldi, P. F., & Hertel, K. J. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing *Proc. Natl. Acad. Sci. USA* **102,** 16176-16181.

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

9. Smith, C. W. J. & Valcárcel, J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control *Trends Biochem. Sci.* **25,** 381-388.

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

11. Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., & Devon, K. (2001) Initial sequencing and analysis of the human genome *Nature* **409,** 860-921.

12. Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D. A., RIKEN GER Group, GSL Members, Hayashizaki, Y., & Gaasterland, T. (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.

13. Sugnet, C., Kent, W., Ares, M. J., & Haussler, D. (2004) in *Pac. Symp. Biocomput.*, pp. 66-77.

14. Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes *Nucleic Acids Res.* **28,** 4364-4375.

15. Baker, R. & Board, P. (1989) Unequal crossover generates variation in ubiquitin coding unit number at the human UbC polyubiquitin locus *Am. J. Hum. Genet.* **44,** 534-542.

16. Stormo, G. D. (2000) DNA binding sites: representation and discovery *Bioinformatics* **16,** 16-23.

17. Modrek, B. & Lee, C. J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss *Nature Genet.* **34,** 177-180.

18. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2007) GenBank *Nucleic Acids Res.* **35,** D21-25.

19. Chang, Y.-F., Imam, J. S., & Wilkinson, M. F. (2007) The Nonsense-mediated decay RNA surveillance pathway *Annu. Rev. Biochem.* **76,** 51-74.

20. Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., & Blencowe, B. J. (2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse *Trends Genet.* **21,** 73-77.

21. Sorek, R., Shamir, R., & Ast, G. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20,** 68-71.

22. Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T., & Burge, C. B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse *Proc. Natl. Acad. Sci. USA* **102,** 2850-2855.

23. Blencowe, B. J. (2006) Alternative splicing: new insights from global analyses *Cell* **126,** 37-47.

24. Karpova, A. Y., Howley, P. M., & Ronco, L. V. (2000) Dual utilization of an acceptor/donor splice site governs the alternative splicing of the IRF-3 gene *Genes Dev.* **14,** 2813-2818.

25. Zhuo, D., Madden, R., Elela, S. A., & Chabot, B. (2007) Modern origin of numerous alternatively spliced human introns from tandem arrays *Proc. Natl. Acad. Sci. USA* **104,** 882-886.

26. Zhang, C., Krainer, A. R., & Zhang, M. Q. (2007) Evolutionary impact of limited splicing fidelity in mammalian genes *Trends Genet* **23,** 484-488.

27. Roberts, D. L., Merrison, W., MacFarlane, M., & Cohen, G. M. (2001) The Inhibitor of Apoptosis Protein-binding Domain of Smac Is Not Essential for its Proapoptotic Activity *J. Cell Biol.* **153,** 221-228.

28. Hatton, A. R., Subramaniam, V., & Lopez, A. J. (1998) Generation of Alternative Ultrabithorax Isoforms and Stepwise Removal of a Large Intron by Resplicing at Exon-Exon Junctions *Mol. Cell* **2,** 787-796.

29. Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J., & Lopez, A. J. (2005) Subdivision of Large Introns in Drosophila by Recursive Splicing at Nonexonic Elements *Genetics* **170,** 661-674.

30. Carlo, T., Sierra, R., & Berget, S. M. (2000) A 5' Splice Site-Proximal Enhancer Binds SF1 and Activates Exon Bridging of a Microexon *Mol Cell Biol.* **20,** 3988-3995.

31. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins *Nucleic Acids Res.* **33,** D501-504.

32. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence *Genome Res.* **8,** 967-974.

33. Kent, W. J. (2002) BLAT---The BLAST-Like Alignment Tool *Genome Res.* **12,** 656-664.

34. Ihaka, R. & Gentleman, R. (1996) R: A language for data analysis and graphics *J. Comput. Graph. Statist.* **5,** 299-314.

4.8 Tables and Figures

Table 1: Percentage of dual splice sites conforming to the AG|GU rule.

Splice sites	All	AG|GU sites
5' splice site	27,556	15,455 (56.1%)
3' splice site	27,556	5,022 (18.2%)
dual site	594	155 (26.1%)
dual site.2	85	46 (54.1%)
dual site.3	40	28 (70.0%)
dual site.singleton	319	119 (37.3%)
dual site.singleton.2	39	31 (79.5%)
dual site.singleton.3	26	23 (88.5%)

dual site.2 (dual site.3), dual splice sites with two (three) or more supporting transcripts for each isoform;

dual site.singleton: singleton dual splice sites (no other dual splice sites from the same gene).

dual site.singleton.2 and dual site.singleton.3 are similarly defined.

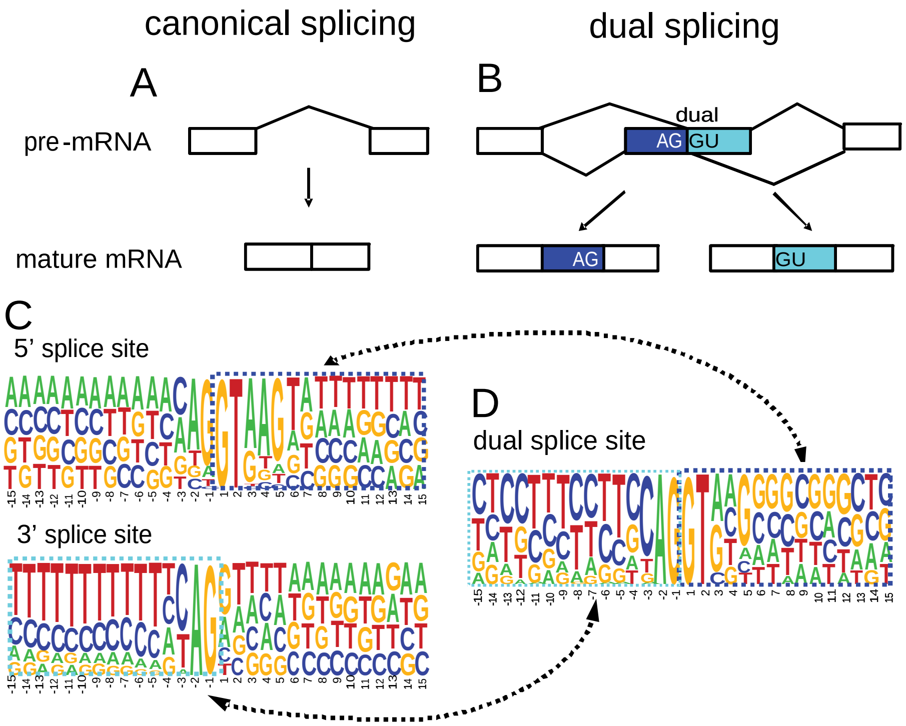


Figure 1: Illustrative representation of dual splicing.

(**A** and **B**) Schematic diagram of canonical splicing (**A**) and dual splicing (**B**). Boxes represent exons and lines are introns. The dual splice site is labeled in (**B**).

(**C** and **D**) The motifs of canonical (constitutive) 5' and 3' splice sites (**C**) and of dual splice sites (**D**).

Dotted arrows and boxes indicate the similarity of dual sites with constitutive splice sites. Uridine is shown as thymine in the logos.

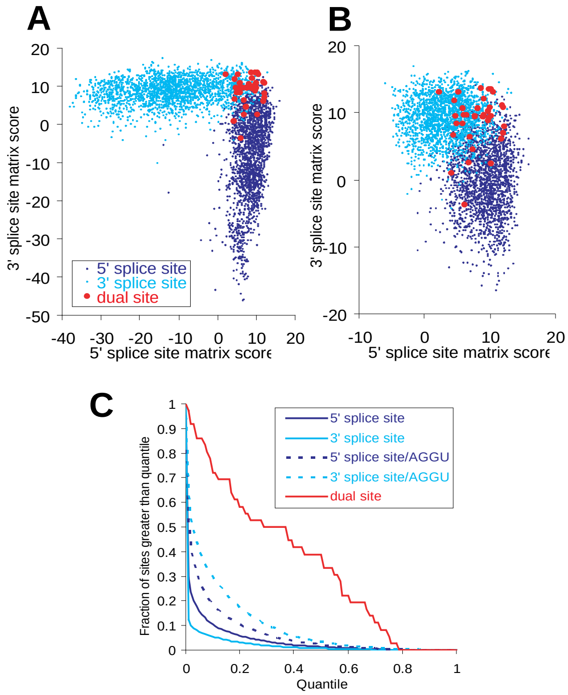


Figure 2: Motif scores of dual and constitutive splice sites.

(**A**) Graphical representation of motif scores of constitutive 5' (blue), 3' (cyan), and dual splice sites (red).

(**B**) Similar to analysis in (**A**) except that only constitutive splice sites with AG|GU pattern are shown.

(**C**) Resemblance of dual splice sites to the canonical 5' and 3' splice-site consensus motifs. Matrix scores were ranked and converted into quantiles according to constitutive splice sites, and different thresholds (quantile 0 to 1, in steps of 0.01) were applied to count the number of sites whose scores exceed both thresholds.

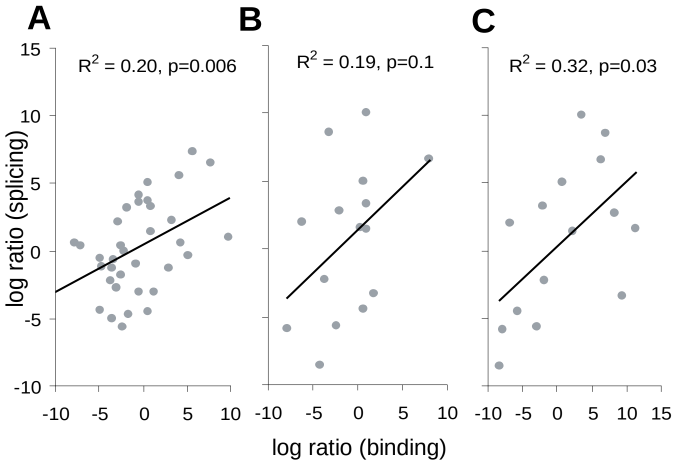


Figure 3: Predicting splicing outcome using binding specificity.

Each point represents a dual splice site. The x-axis shows the log-likelihood ratio of the binding probabilities, and the y-axis shows the log ratio of supporting-transcript number for the 5'-splice-site isoform relative to that of the 3'-splice-site isoform.

(**A**) Competition at the dual splice site. The log-likelihood ratio of binding is calculated as Δb using the high-confidence dataset.

(**B**) Similar to (**A**) except using the extended set.

(**C**) Competition by exon definition or alternative 5'/3' splice sites. The log-likelihood ratio of binding is calculated as Δb_2 using the extended set.

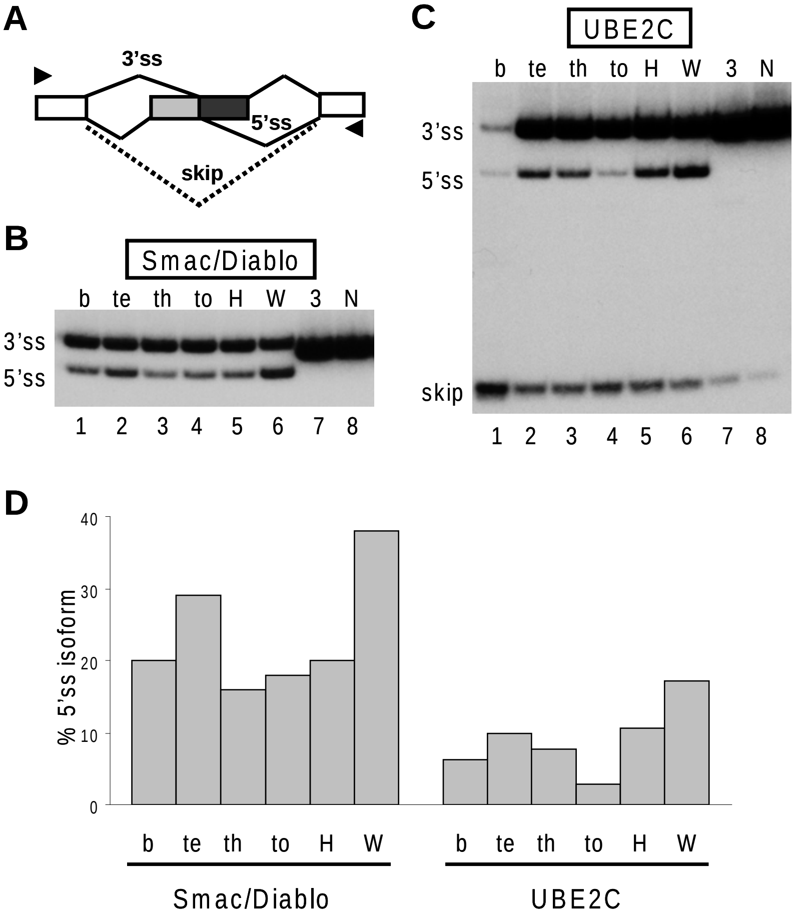


Figure 4: Validation of dual-splice-site recognition in human and mouse.

RT-PCR analysis of RNA from human brain (b), testis (te), tonsil (to), and thymus (th) tissues (Clontech Laboratories, Inc.), HeLa (H), Weri-Rb1(W), NIH-3T3 (3), and NSC-34 (N) cells.

(**A**) Diagram of the general splicing pattern and location of primers (arrows) used to detect isoforms that result from the use of the 5' splice site or 3' splice site of (**B**) Smac/Diablo (3'ss, 234 nt; 5'ss, 205 nt), and (**C**) *UBE2C* (3'ss, 287 nt; 5'ss, 233 nt). In the case of *UBE2C*, the exon containing the dual splice site is also skipped to give an additional isoform (skip, 82 nt).

(**D**) Quantitative analysis of RT-PCR results. The histogram represents the percentage of the products that are generated by use of the 5' splice site.

4.9 Supporting information

SI Table 2: List of human dual splice sites in the high-confidence set

Symbol	Evidence (# of ts)		Evidence type[1]	Protein product[2]		Motif score		Seq. conserv.[3]
	5'ss	3'ss	(5'ss/3'ss)	5'ss	3'ss	5'ss	3'ss	
WDR73	43	99	mRNA/RefSeq	EAX01949?	Q96JZ1	4.70	9.45	mrd
DIABLO	23	108	RefSeq/RefSeq	**Q502X2**	**NP_063940**	5.94	9.61	mrd
IRF3	128	13	RefSeq/mRNA	Q14653	N/A	9.76	8.74	mr
ILVBL	11	24	EST/RefSeq	AAI26914?	O43341	2.27	12.98	mrd
SPIN1	11	17	*mRNA/mRNA*	**BAD92639**	**EAW52016**	10.23	13.51	mrd
ASBABP1	22	8	RefSeq/mRNA	Q4ZIN3	ENSP00000234393?	11.99	11.04	mrd
C16orf24	7	24	*mRNA/RefSeq*	Q6ZWF3?	Q9BQD7	10.48	12.98	mrd
UBE2C	7	154	RefsSeq/RefSeq	**Q9BQP0**	**Q9BQP1**	6.88	6.28	mrd
NSUN5C	5	34	mRNA/RefSeq	N/A	Q6PHS1	6.51	9.56	N/A
LCN2	160	5	mRNA/EST	P80188	N/A	9.87	9.22	mrd
RHOT2	64	5	*RefSeq/mRNA*	Q96C13	N/A	10.16	9.52	mrd
ACADVL	358	4	RefSeq/EST	P49748	N/A	10.21	2.40	mrd
C10orf82	17	4	*mRNA/RefSeq*	NP_653262?	Q8WW14	9.11	12.00	mrd
TMEM141	70	4	RefSeq/EST	Q96I45	EAW88278?	9.79	10.21	mrd
AMHR2	4	5	*mRNA/RefSeq*	N/A	Q16671	12.14	6.95	mrd
EGR2	4	6	*mRNA/mRNA*	EAW54238?	A40492?	5.84	10.66	mrd
POLR2G	4	196	*mRNA/RefSeq*	**EAW74091**	**P62487**	8.18	10.55	d
CUBN	4	3	EST/RefSeq	N/A	O60494	4.58	11.73	mr
RABGGTA	3	81	EST/RefSeq	N/A	Q92696	7.68	9.38	mrd
PIGG	14	3	mRNA/EST	Q5H8A4	N/A	4.18	0.87	mrd
HDLBP	4	3	EST/EST	N/A	N/A	5.84	8.34	N/A
LOC339123	3	63	EST/RefSeq	AAH94850?	Q4VBY1	8.66	13.57	d
GLT28D1	3	3	EST/EST	N/A	N/A	4.47	6.59	mrd
RPS3	4	2	EST/EST	N/A	N/A	6.15	-3.70	mrd
TLE2	2	4	EST/EST	N/A	N/A	10.00	10.70	mrd
DMAP1	2	5	EST/EST	NP_061973?	BAA92663?	4.94	8.37	
UROD	324	2	RefSeq/EST	P06132	N/A	11.84	6.05	mrd
CBS	95	2	mRNA/EST	P35520	N/A	12.15	7.91	mr
TNFAIP3	24	2	RefSeq/EST	P21580	N/A	9.17	9.65	md
RPL8	2	5	EST/EST	N/A	N/A	7.38	4.46	d
DCI	3	2	mRNA/EST	AAH02746	N/A	6.83	2.52	N/A
ASAH1	3	2	EST/EST	N/A	BAD96500?	5.26	12.98	N/A
RECQL4	2	17	EST/RefSeq	N/A	O94761	8.92	9.34	rd
FGFR4	18	2	RefSeq/RefSeq	**P22455**	**Q71TW8**	7.69	9.48	mrd
COCH	2	66	mRNA/RefSeq	N/A	O43405	9.91	13.50	mrd
SLC35B4	2	17	EST/RefSeq	N/A	Q969S0	12.07	10.77	d

1. In some cases (italics), transcripts with higher-quality supporting evidence (RefSeq>mRNA>EST) exist with BLAT alignment but failed to pass filtering criteria used for dbCASE. In these cases, the higher-quality transcript (rather than the type of transcript of lower quality) is provided.

2. In five cases (bold), both isoforms produce potentially functional protein products, and the alternative region is at least partially coding. For dual splice sites in untranslated regions, the protein products cannot be inferred directly, and are shown with question marks. In the *IRF3* gene, both isoforms generate protein products (Karpova et al. 2000, Genes Dev. 14: 2813-2818), but one of them is not present in the GenBank protein sequence database.
3. Abbreviations: m, mouse; r, rat; d, dog; e, elephant.

SI Table 3: Nucleotide composition of flanking sequences around human dual splice sites and constitutive splice sites

Splice site	C,%	GC,%
3' splice site.left (-15 to -3)	30.4%	40.3%
dual site.left (-15 to -3)	39.5%	54.7%
5' splice site.right (3 to 15)	17.8%	44.0%
dual site.right (3 to 15)	26.3%	59.8%

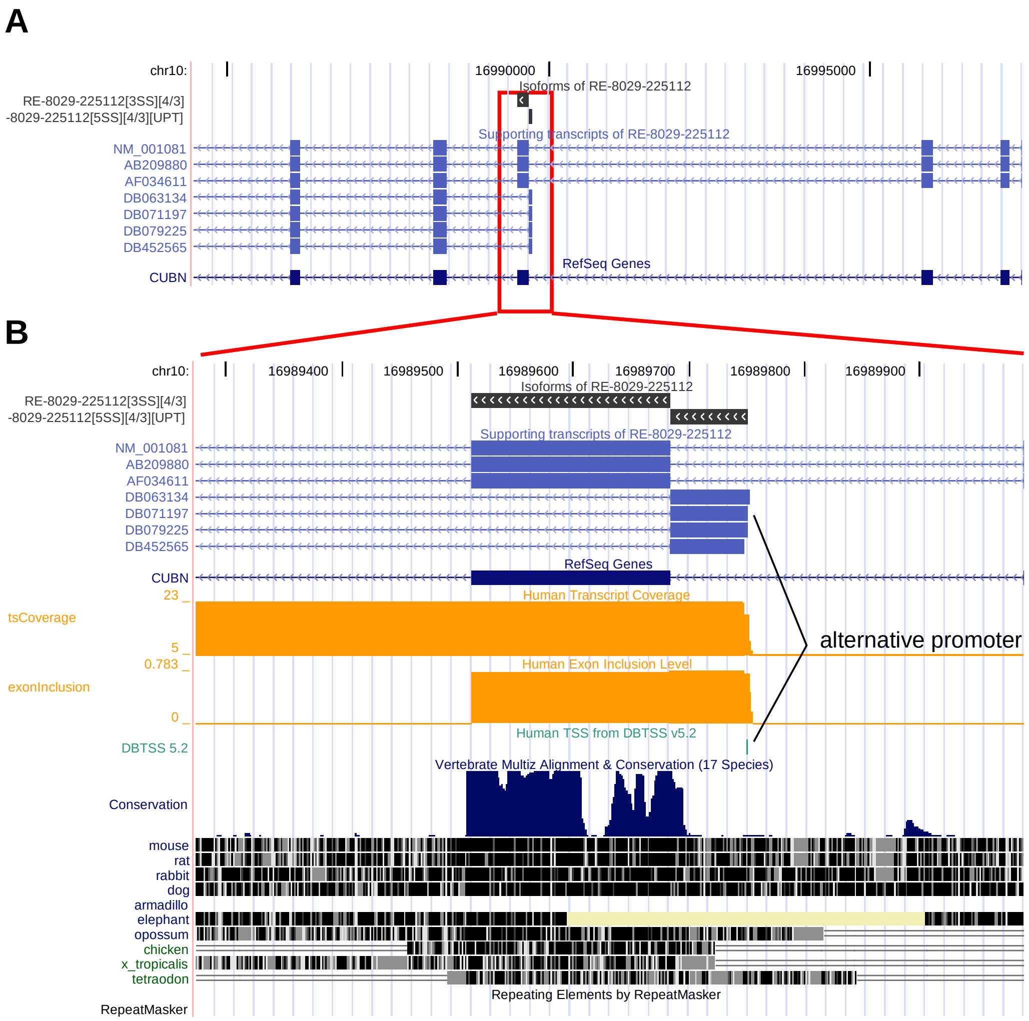
SI Table 4: Sequences of primers used in PCR reactions.

Primer	Sequence (5'→3')
Mouse-DIABLO-F	GGCTCTGAGAAGTTGGGTG
Mouse-DIABLO-R	AGACACAGCCCTCCTCATC
Human-DIABLO-F	CTGCACAATGGCGGCTCTG
Human-DIABLO-R	CTCCTCATCAATGCTTCAC
UBE2C-F	GGACCATCCATGGAGCAGC
UBE2C-R	GGGGTTTTTCCAGAGCTCGGC
UROD-F	ACCTCAGGGTTTTCCGGAGC
UROD3'SS-R	TTGCTAGGTGGCAGACTGAAGG
POLR2G-F	GGAGGGGACCTGCACAGGG
POLR2G-R	CCCAATTTCTGTGAAGAGTCCAAC

F: forward, R: reverse. For *UBE2C*, *UROD*, and *POLR2G*, the same primers were used for human and mouse, as the target sequences are conserved.

SI Table 5 List of mouse dual splice sites in the high-confidence set

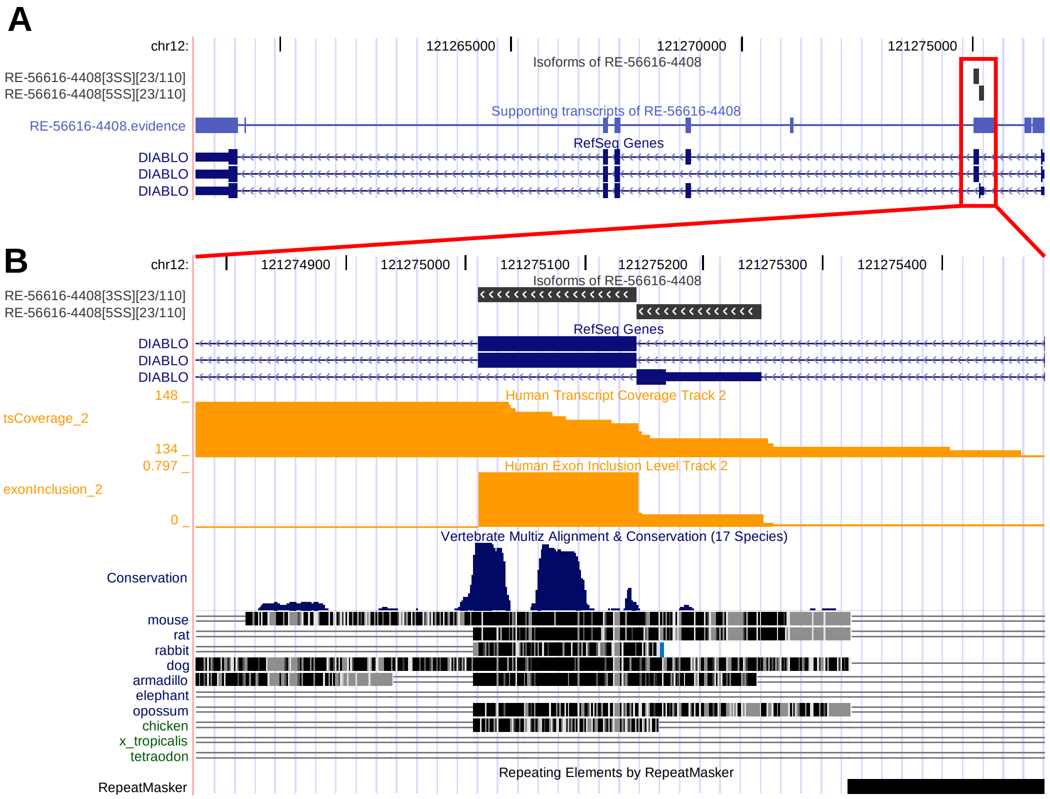
Symbol	**Evidence (# of ts)**		**Evidence type**	**Motif score**		**Seq. Conserv.**
	5'ss	3'ss		5'ss	3'ss	
Gpr137	16	10	RefSeq/RefSeq	7.98	10.46	hrd
Vill	13	9	RefSeq/mRNA	9.77	7.87	rd
Nt5c3l	8	20	RefSeq/mRNA	12.39	5.62	hrd
D3Ertd300e	6	32	mRNA/EST	5.35	7.39	r
Ldha	16	6	EST/RefSeq	9.92	9.20	hrd
Med25	6	101	RefSeq/mRNA	6.56	13.47	hrd
Csf3r	4	22	RefSeq/EST	11.90	10.38	hrd
Psmb10	3	61	RefSeq/EST	6.65	12.68	hrd
Irak1	29	3	EST/RefSeq	9.04	14.32	hrd
Rnf170	3	11	RefSeq/mRNA	9.23	5.44	hrd
1110006G06Rik	3	16	mRNA/RefSeq	11.42	8.72	hrd
Tmem112b	2	33	RefSeq/mRNA	8.29	14.69	hrd
2810453I06Rik	3	2	EST/RefSeq	1.16	5.60	N/A
Gpr114	4	2	EST/RefSeq	8.76	5.59	hd
Dvl1	23	2	EST/RefSeq	9.57	6.05	hrd
2400006H24Rik	2	104	RefSeq/EST	4.15	9.19	hrd
Tmem19	2	49	RefSeq/EST	5.12	11.87	r
Bmf	4	2	EST/mRNA	5.21	11.53	hrd



SI Figure 5: An example of a dual splice site (*CUBN*) associated with alternative promoters. The gene structure and splicing patterns are visualized with the UCSC genome browser (http://genome.ucsc.edu) (Kent et al. 2002, Genome Res. 12, 996-1006). Exons are represented by boxes and introns are represented by lines. The arrows through a transcript indicate its orientation.

(**A**) Zoom-out view. From top to bottom, the first two tracks are custom tracks based on our data in dbCASE. The first track shows the two isoforms resulting from dual splicing; the second track shows transcripts (RefSeqs, mRNAs, and ESTs) supporting the dual splicing pattern; the third track shows RefSeq transcripts.

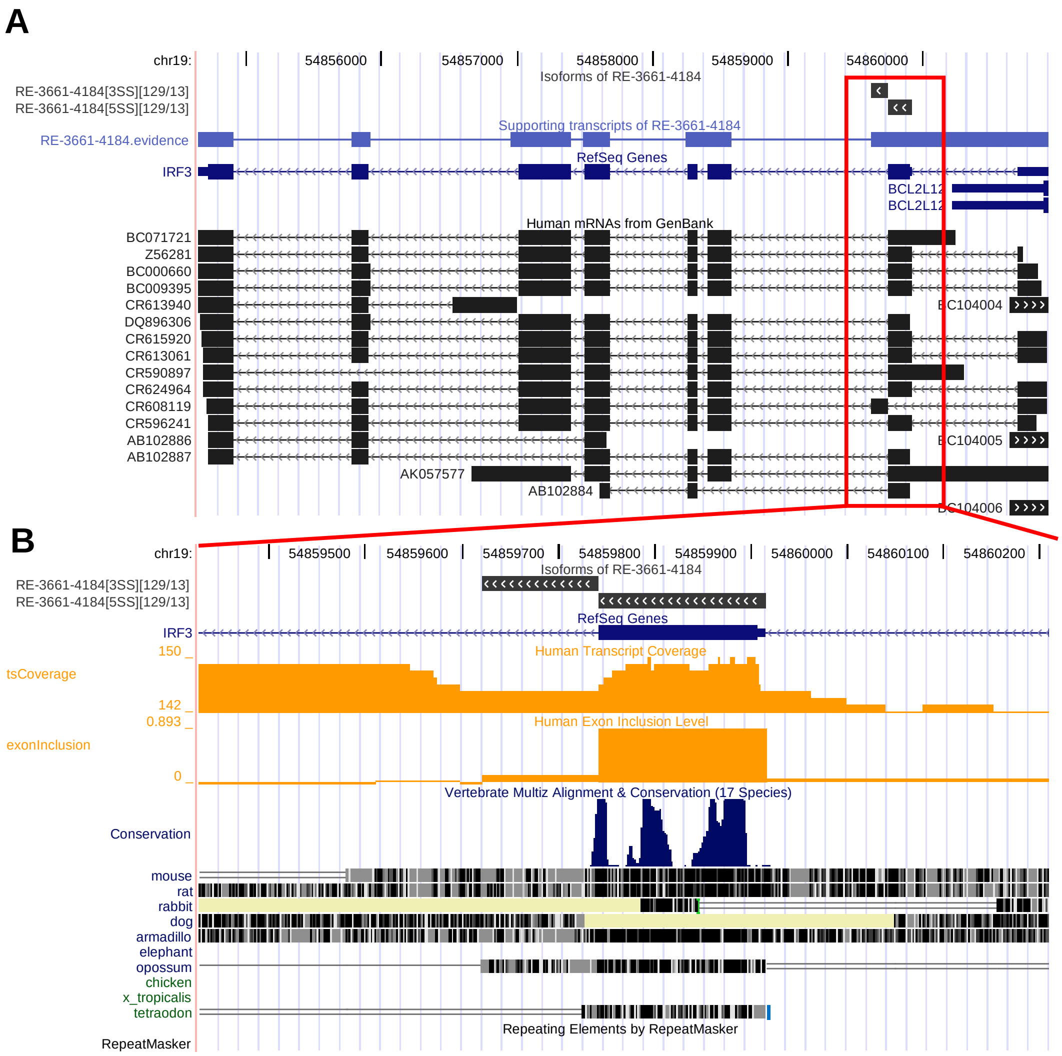
(**B**) Zoom-in view. The transcript coverage and exon inclusion level at each position, based on data from dbCASE, are displayed in the two orange tracks. Cross-species conservation is displayed near the bottom. Note that the alternative promoters are supported by DBTSS (the cyan track) (Yamashita et al. 2006, Nucleic Acids Res. 34: D86-89). The conservation track shows phastCons scores (Siepel et al 2005, Genome Res. 15, 1034-1050), reflecting the level of conservation, at the top, and the quality of pairwise alignment of each species with human in gray scale (the darker, the more reliable) below. The RepeatMasker track is also shown at the very bottom. A repetitive sequence, if any, is represented by a filled box. In this particular case, however, no repetitive sequences were detected in the region.


A
chr12:
121265000
121270000
121275000
Isoforms of RE-56616-4408
RE-56616-4408[3SS][23/110]
RE-56616-4408[5SS][23/110]
Supporting transcripts of RE-56616-4408
RE-56616-4408.evidence
RefSeq Genes
DIABLO
DIABLO
DIABLO
B
chr12:
121274900
121275000
121275100
121275200
121275300
121275400
Isoforms of RE-56616-4408
RE-56616-4408[3SS][23/110]
RE-56616-4408[5SS][23/110]
RefSeq Genes
DIABLO
DIABLO
DIABLO
148
Human Transcript Coverage Track 2
tsCoverage_2
134
0.797
Human Exon Inclusion Level Track 2
exonInclusion_2
0
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
Repeating Elements by RepeatMasker
RepeatMasker

SI Figure 6: An example of a dual splice site (*DIABLO*) associated with the activation of an upstream alternative exon.

(**A**) Zoom-out view. The evidence track is condensed due to the large number of supporting transcripts. Representative supporting transcripts can be seen in the RefSeq track.

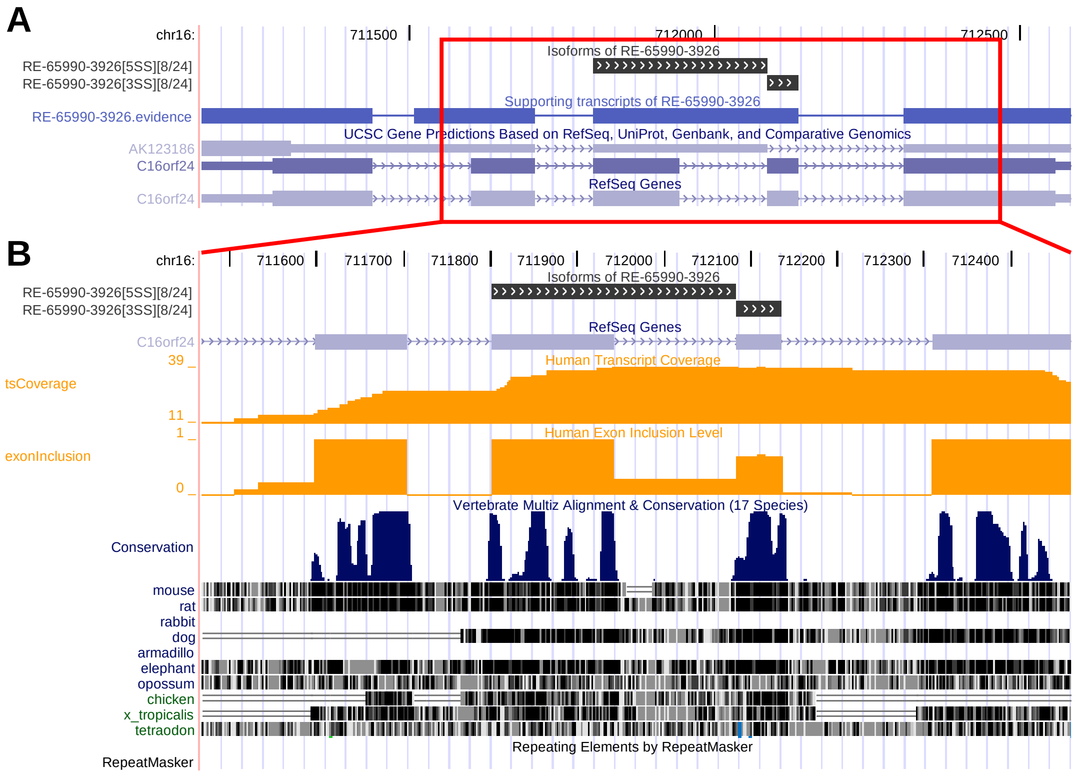
(**B**) Zoom-in view. See the legend of SI Fig. 5 for more details.



SI Figure 7: An example of a dual splice site (*IRF3*) associated with the activation of a downstream alternative exon.

(**A**) Zoom-out view. The evidence track is condensed due to the large number of supporting transcripts. Representative supporting transcripts can be seen in the mRNA track.

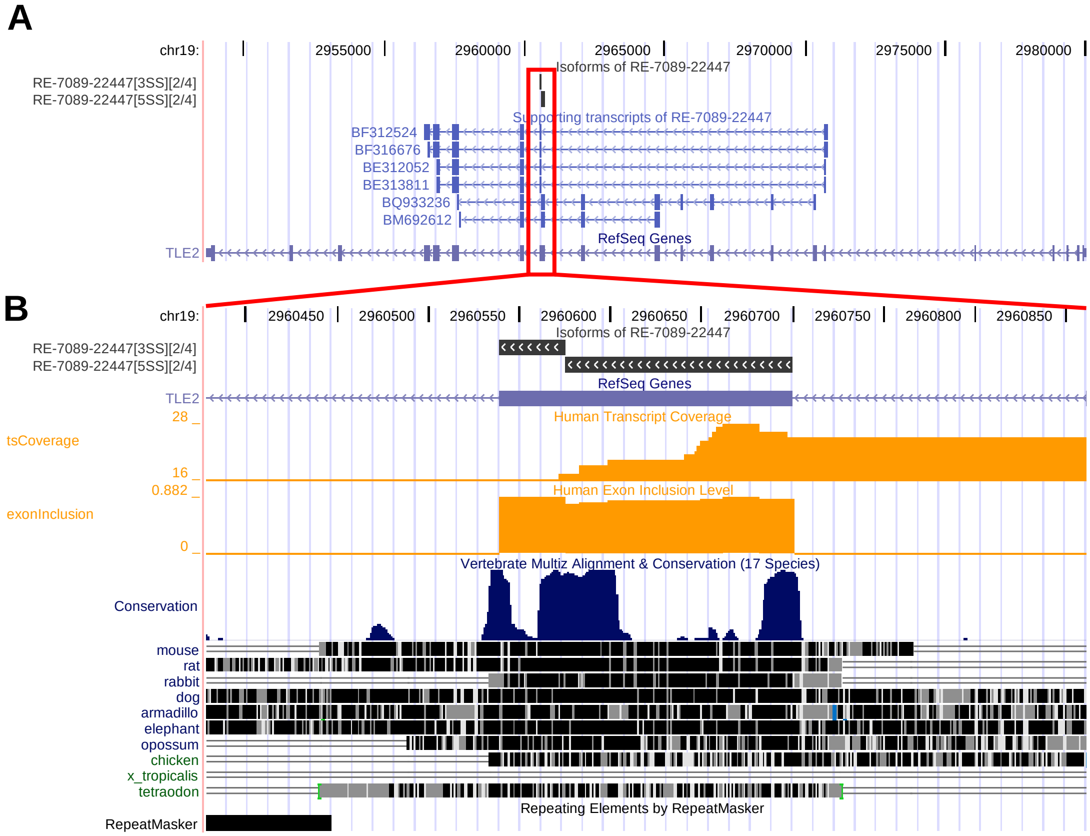
(**B**) Zoom-in view. See the legend of SI Fig. 5 for more details.



SI Figure 8: An example of a dual splice site (*C16orf24*) associated with intron retention.

(**A**) Zoom-out view. The evidence track is condensed due to the large number of supporting transcripts. Representative supporting transcripts can be seen in the UCSC known gene track.

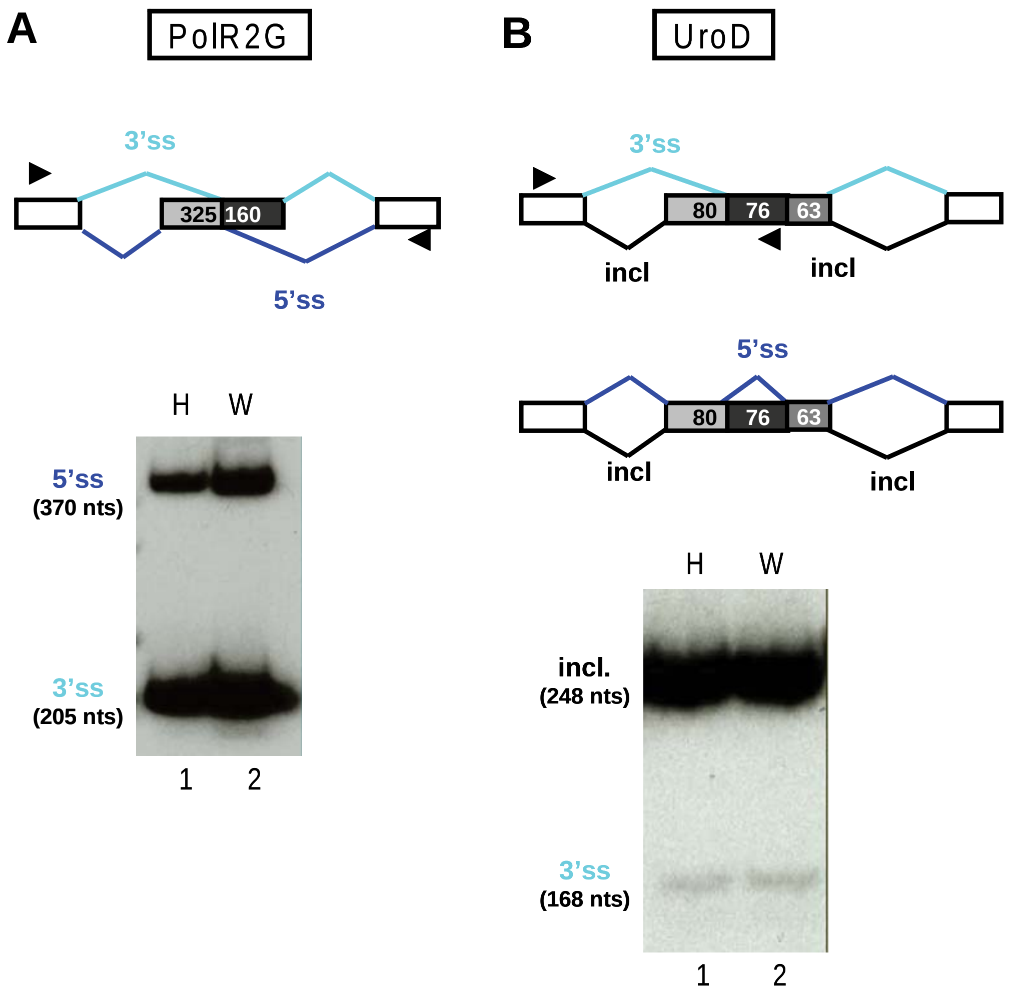
(**B**) Zoom-in view. See the legend of SI Fig. 5 for more details.



SI Figure 9: An example of a dual splice site (*SLE2*) associated with exon truncation.

(**A**) Zoom-out view.

(**B**) Zoom-in view. See the legend of SI Fig. 5 for more details.



SI Figure 10: Validation of additional dual splice sites in cells.

RT-PCR analysis of RNA from HeLa (H) or Weri-Rb1 (W) cells using primers specific to (**A**) *PolR2G* and (**B**) *UroD*. Diagrams are shown for the relevant regions of each gene, with the alternative exon sizes indicated. The sizes of the RT-PCR products are indicated next to each autoradiogram. In the case of *PolR2G*, the two alternative exons flanking the dual splice sites have different sizes, and therefore a single pair of primers is sufficient to distinguish the two isoforms. In the case of *UroD*, the splicing pattern is more complex. As shown in the diagram, besides the two alternative isoforms that use the dual splice site as a 5' and 3' splice site, respectively, the two alternative exons can be included simultaneously as a single larger exon (product labeled "incl"). This makes it difficult to choose a single pair of primers to unambiguously distinguish the two isoforms resulting from dual splicing. However, the existence of the 5' splice-site isoform, which is protein-coding, is virtually certain, because there are 324 supporting transcripts, including RefSeq evidence (SI Table 2). Therefore, in the RT-PCR analysis, we only used a primer set to validate the less abundant 3'-splice-site isoform (lanes 1 and 2).

Chapter 5

RNA landscape of evolution for optimal exon and intron discrimination

5.1 Abstract

Accurate pre-mRNA splicing requires primary splicing signals, including the splice sites, a polypyrimidine tract, and a branch site, and also other splicing-regulatory elements (SREs). The SREs include exonic (ESEs and ESSs) and intronic (ISEs and ISSs) splicing enhancers and silencers, which are typically located near the splice sites. However, it is unclear to what extent splicing-driven selective pressure constrains exonic and intronic sequences, especially those distant from the splice sites. Here we studied the distribution of SREs in human genes in terms of DNA strand-asymmetry patterns. Under a neutral evolution model, each mononucleotide or oligonucleotide should have a symmetric (Chargaff's second parity rule), or weakly asymmetric yet uniform, distribution throughout a pre-mRNA transcript. However, we found that large sets of unbiased, experimentally-determined SREs show a distinct strand-asymmetry pattern that is inconsistent with the neutral evolution model, and reflects their functional roles in splicing. ESEs are selected in exons and depleted in introns, and vice-versa for ESSs. Surprisingly, this trend extends into deep intronic sequences, accounting for one third of the genome. Selection is detectable even at the mononucleotide level, so that the asymmetric base compositions of exons and introns are predictive of ESEs and ESSs. We developed a method that effectively predicts SREs based on strand asymmetry, expanding the current catalog of SREs. Our results suggest that human genes have been optimized for exon and intron discrimination through an RNA landscape shaped during evolution.

5.2 Introduction

Most mammalian genes are split, with exons (~150 nt) separated by much longer introns (~3,000 nt). To produce a mature transcript from a pre-mRNA, introns are spliced out and exons are ligated by a large protein/snRNA complex, the spliceosome. Extensive efforts have been made to elucidate the splicing code, i.e., the combinations of *cis*-regulatory elements and *trans*-acting factors responsible for splicing efficiency and fidelity. Besides the degenerate splice-site motifs, which are necessary but not sufficient for specific exon and intron recognition, other sequence elements are required for both constitutive and alternative splicing (1, 2). Many splicing-regulatory elements (SREs) have been identified by experimental or computational approaches (3-10). Among them, two classes of well- studied SREs are ESEs recognized by SR proteins, and ESSs recognized by certain hnRNP proteins (1, 2). Adding further complexity, the effect of an SRE on splicing is often context-dependent. For example, an SR-protein-dependent ESE element, when present in an intron, can act as an ISS to repress splicing (11), whereas a number of ESSs, such as the GGG motif, are also potent ISEs (12). The combinatorial interactions of SR proteins and hnRNP proteins with their cognate SREs are an important aspect of splicing fidelity for most, if not all, exons and introns.

Several previous studies have focused on constitutively spliced exons and introns, and revealed a non-random distribution of SREs, which suggests that evolution has differentiated exons from introns for the purpose of splicing (4, 8, 9). More specifically, there is a higher density of ESEs in exons than introns, and vice-versa for ESSs. In addition, ESEs and ESSs are preferentially located in exonic and intronic sequences near the splice sites, respectively. These observations are consistent with results from comparative-genomics studies, which demonstrated that exonic and intronic sequences near the splice sites show a higher level of sequence conservation than sequences farther from the splice sites, especially for alternatively spliced exons (13).

Despite this progress, the understanding of the extent and pattern of functional constraints for accurate splicing of mammalian genes remains incomplete. An important limitation of previous studies is the lack of "completely neutral" sequences as controls to compare with real exons and introns, which prevents a rigorous assessment of selective forces that have enriched or depleted different classes of SREs in different regions. For the same reason, it has been difficult to prove or disprove splicing-coupled selection in sequences far from the splice sites, e.g., intronic sequences beyond several hundred nucleotides, although it is commonly believed that SREs are located near the splice sites (14).

On the other hand, neutral sequence evolution is reflected in DNA strand-asymmetry patterns, which may provide a powerful tool to evaluate and characterize the signatures of selection. According to Chargaff's second parity rule (PR2), the frequency of a mononucleotide or oligonucleotide should be (statistically) equal to that of its reverse complement on the same strand of a long genomic DNA (15, 16). PR2 has been validated in many organisms, from bacteria to mammals, and presumably reflects symmetric DNA mutations and repair (16). Exceptions to PR2, or DNA-strand asymmetry, do exist, reflecting different mechanisms in various organisms. In bacteria and vertebrates, strand asymmetry in gene regions is thought to arise from asymmetric but neutral transcription-coupled mutation (TCM) and repair (TCR) mechanisms (17). TCM and TCR have been invoked to explain the excess of guanine (G) + thymine (T) over adenosine (A) + cytosine (C) in the sense strand observed in mammals (18). However, stronger strand asymmetry in intronic sequences near the splice sites was also noted, and attributed to splicing-coupled selection (19).

Here we systematically investigate splicing-coupled selection in human constitutive exons and introns by characterizing the patterns of DNA-strand asymmetry of mononucleotides and oligonucleotides. This approach does not require neutral sequences as controls. Instead, we examine each exonic and intronic region separately, to see if SREs can be distinguished from random elements in terms of strand asymmetry, providing a hallmark of splicing-coupled selection. We provide evidence that the distributions of many known ESEs and ESSs differ from those of random elements in both exons and introns, including deep intronic sequences. The systematic bias and the pattern of SRE distribution cannot be explained by a neutral evolution model, suggesting that human genes have been optimized during evolution for discrimination between exons and introns, among other potential functional constraints.

5.3 Results

Patterns of mononucleotide strand asymmetry in exons and introns

To assess the selective pressure driven by pre-mRNA splicing fidelity and/or efficiency, we initially studied the mononucleotide strand asymmetry of human and mouse genes in five regions from constitutive internal exons and introns: the first (5'E) and last (3'E) 70 nucleotides of exons; the first (5'I) and last (3'I) 100 nucleotides of introns; and the middle 100 nucleotides (midLI) of long introns (≥3,000 nt) (Fig. 1A). Surprisingly, exons and introns show opposite strand asymmetry, as quantified by S_{TA} and S_{GC} (Fig. 1B and supporting information (SI) Table 1). T is more abundant than A, and G is more abundant than C in intronic regions, which is consistent with previous studies (18, 19). In contrast, T is less abundant than A, and there is only a slight excess of G over C in exons. The 5' and 3' extremities of introns generally have similar patterns, with an increased frequency of T and C (Fig. 1B). This nucleotide bias may partly reflect some longer-than-usual polypyrimidine tracts at the 3' extremity of some introns, but the underlying reason is less apparent at the 5' extremity.

The above observations indicate a more complicated landscape of strand asymmetry than can be explained by transcription-coupled mechanisms. Instead, the distinct asymmetry patterns of exons and introns could be due to protein-coding and/or splicing, whose signals are superimposed in exonic sequences. To separate these selective forces, which may have contributed to strand asymmetry in exons, we compared constitutive internal coding and 5'UTR exons, as well as the coding and 5'UTR portions of intronless genes. Notably, compared with coding exons, strand asymmetry in non-coding exons is very similar (Fig. 1C), with no or only moderate differences in either TA asymmetry (*p*=0.04 for human; *p*=0.02 for mouse; chi-square test, df=1; the same below, except where indicated) or GC asymmetry (*p*=0.58 for human; *p*=0.14 for mouse). In contrast, much weaker asymmetry, especially for S_{TA}, is observed in the coding portion of intronless genes, for which the effect of splicing is absent ($p<2.2\times10^{-16}$ for human and mouse). Importantly, strand asymmetry in the 5'UTR of intronless genes, in which protein-coding and splicing effects have presumably been separated, is barely detectable. The TA asymmetry is estimated to be -0.1% (*p*=0.8) and -1.0% (*p*=0.02), and GC asymmetry is estimated to be -0.7% (*p*=0.08) and 0.7% (*p*=0.07), for human and mouse, respectively (Fig. 1C). This observation contradicts the assumption that TCR is strongest immediately downstream of the transcriptional start site (17). Although these comparisons may have overlooked other potential differences between intron-containing and intronless genes, they support the notion that the observed strand asymmetry is strongly correlated with splicing-coupled selection. Interestingly, the pattern of strand asymmetry in lower organisms differs substantially from that of mammals (SI Fig. 5). In particular, yeast introns have strand asymmetry in the same direction as exons (SI Table 1). This pattern corroborates the observation that the yeast primary splicing signals are highly conserved among different introns, which often provides sufficient discrimination between exons and introns.

Non-random distribution of known SREs

We reasoned that if the landscape of strand asymmetry in exons and introns is associated with splicing-coupled selection, the bias of mononucleotides *per se* may not have a direct functional meaning. Rather, splicing factors, such as SR proteins and hnRNPs, could have preferences for certain sequence motifs, whose nature and frequency would determine the overall strand asymmetry in exons and introns. To evaluate splicing-coupled selection more directly, we analyzed the distribution of known and putative hexameric SREs in human exons and introns, in comparison with random hexamers. For each type of sequence (exon, 5’I, 3’I, and midLI), we divided all unique hexamers, including SREs, into three groups: those with positive (S>0), negative (S<0) or no (S=0) asymmetry, in which a hexamer is more, less, or equally frequent in the sense strand than in the antisense strand, respectively. Because all hexamers are part of reverse-complementary pairs (except self-complementary or palindromic ones), the number of hexamers with positive asymmetry has to be equal to the number with negative asymmetry, independently of the sequences under consideration. Our null hypothesis is the neutral-evolution model, under which SREs should be subject to the same selective pressure as random elements, so that the strand asymmetry of SREs should not differ from that of random elements. Alternatively, if the sequences are not neutral, and certain elements are enriched (depleted), more than half of those asymmetric elements should have positive (negative) asymmetry. Therefore, a systematic bias in the direction of strand asymmetry of SREs would provide direct evidence of splicing-coupled selection.

We first tested this hypothesis by examining the distribution of experimentally determined ESSs and ESEs. A panel of 103 ESS hexamers, dubbed FAS-hex3, was derived by cell-based selection from a library of random decamers engineered into an alternative exon in a fluorescent splicing reporter (8). These ESS hexamers do have a lower frequency in exons compared with flanking intronic sequences (8), but it was unclear if the distribution deviates from neutral evolution in exons or introns, or both. We found that the ESS hexamers show very biased strand asymmetries in both exons and introns, yet opposite in direction. As shown in Fig. 2A, 90 ESS hexamers (87%) have negative asymmetry in exons, implying that ESSs tend to be depleted in exons (p=3.2×10^{-14}). In contrast, in introns, especially in the 5’I and 3’I regions, most ESS hexamers have positive asymmetry (93 of 103 or 90% in both regions), implying that

ESSs tend to be enriched in introns (p=2.9×10^{-16}). Even in the midLI region, 70% (72 of 103) of ESS hexamers have positive asymmetry (p=5.3×10^{-5}), suggesting a role in repression of exon-like sequences (pseudo-exons) in introns. Therefore, the distribution of ESSs deviates from the prediction of the neutral-evolution model in both exons and introns, including deep intronic sequences, and is consistent with the role of ESSs in exon silencing.

We similarly studied a panel of 220 ESE hexamers, dubbed "Cooper ESEs", identified by *in vivo* functional SELEX experiments (3). The distribution of these ESEs is also significantly non-random, and has an opposite pattern compared with ESSs (Fig. 2A). Among the 219 non-palindromic hexamers, 169 (77%) ESE hexamers have positive asymmetry in exons (p=8.9×10^{-16}), whereas in the 5'I and 3'I regions, most (166 or 76%, p=2.2×10^{-14} for 5'I; 164 or 75%, p=1.8×10^{-13} for 3'I) have negative asymmetry. Again, even in the midLI region, 63% (137 of 218; one is absent in the midLI region, p=1.5×10^{-4}) have negative asymmetry. Similar results were also observed from two additional panels of experimentally determined ESEs: "Kole-ESEs" identified by *in vitro* functional SELEX experiments (7) and "literature ESEs" compiled in a survey of multiple studies (10) (SI Figs. 6 and 7). Therefore, ESEs tend to be enriched in exons and depleted in introns, including deep intronic sequences, which is consistent with their functional roles in exon recognition.

The skewed asymmetry of ESSs and ESEs in exons is not due to the depletion of in-frame stop codons. To demonstrate this, we separately examined the strand asymmetry of stop-codon-containing SREs and non-stop-codon-containing SREs, and found the same pattern of strand asymmetry for both groups (Fig. 2A). Interestingly, the frequencies of the three stop codons in ESSs, ESEs, and the termini of coding sequences (actual stop codons) are very different (Fig. 2B): UAG is much more frequent in ESSs (86%, p<2.2×10^{-16}), but almost absent in ESEs (p=0.06, moderate significance due to limited sample size; more significant results observed in SI Figs. 6B and 7B), compared to its use as a stop codon (24%). This bias likely reflects the similarity of UAG with the consensus motif (UAGGGA/U) of hnRNP A1 (20), which represses exon recognition and splicing when bound to exons. Taken together, the analyses of both ESSs and ESEs provide strong evidence that the distribution of SREs is selected to maximize splicing fidelity in both exons and introns, even for deep intronic sequences, which were previously assumed to be neutral (14).

Prediction of new SREs using strand asymmetry

The distinct landscape of strand asymmetry of known SREs also suggests a method for *de novo* SRE prediction. Instead of the four conventional categories of SREs (ESE, ESS, ISE, and ISS), we define two categories: exon-identity elements (EIEs), enriched in exons and important for exon recognition, and intron-identity elements (IIEs), enriched in introns and important for intron recognition. This definition reflects the functional overlap between ESEs and ISSs, which together roughly correspond to EIEs, and between ESSs and ISEs, which together roughly correspond to IIEs. In addition, this dual classification of elements may have a more natural correspondence with the two main categories of ubiquitous splicing-regulatory proteins, i.e., SR proteins and hnRNPs.

Overall, we predicted 1,131 hexamers with the strongest positive asymmetry in constitutive exons as EIEs (z=5, p=0.001, after Bonferroni correction for multiple testing) (Fig. 3). At the same significance level, we similarly predicted 569 and 568 hexamers with the strongest positive asymmetry in 5'I and 3'I sequences as IIEs, respectively. The 5' and 3' IIEs largely overlap, and their union gives 708 IIEs (Fig. 3). Among the EIEs, the hexamer GAAGAA, which is recognized by SF2/ASF and enhances exon recognition (21), is ranked third from the top (S=44%, z=40). AC-rich elements are also abundant among EIEs (3). In contrast, a number of top IIEs are U-rich elements, which can be recognized by several hnRNP proteins, such as hnRNP C (20).

To evaluate the method more quantitatively, we performed extensive comparisons of the predicted EIEs and IIEs with known SREs (3-5, 7-10), especially those determined by unbiased experimental approaches (3, 7, 8, 10). We found significant overlaps between EIEs and ESEs, and between IIEs and ESSs, respectively (Fig. 3B). In particular, 61% (63 of 103) of FAS-hex3 ESSs are predicted as IIEs, 3.5-fold greater than expected by chance ($p < 2.2\times10^{-16}$). Among them, five of the six (83%) representative ESS hexamers derived from clustering analysis (8) are predicted as IIEs. For Cooper ESEs (3), 50% (109 of 220) of the derived hexamers are predicted as EIEs (1.8-fold enrichment compared with random hexamers, $p < 1.3\times10^{-13}$). We note that comparisons with previous computationally-defined elements are likely biased, because such methods explicitly used the enrichment or depletion in exons (introns) to derive the elements. Nevertheless, the overlap between EIEs and Cooper ESEs, and that between IIEs and FAS-hex3 ESSs, which are unbiased, are among the largest in all the comparisons. In contrast, the overlap between ESEs and IIEs, and that between ESSs and EIEs, are significantly smaller than expected by chance (data not shown).

Next, we examined the strand-asymmetry patterns of predicted EIEs and IIEs to evaluate functional selection. As we did not use introns for predicting EIEs, our prediction method should not bias the strand asymmetry of EIEs in introns; a similar argument holds for IIEs in exons and midLI regions. However, we found significantly biased strand asymmetries for both EIEs and IIEs (SI Fig. 8), qualitatively similar to what we observed from known ESEs and ESSs, respectively (Fig. 2A). Therefore, these results provide an independent line of evidence that exons and introns—even intronic sequences distant from the splice sites—are under splicing-coupled selection.

Correlation between strand asymmetry of oligonucleotides and mononucleotides

We noticed that EIEs and IIEs have a strongly non-uniform base composition, with T>A and G>C in IIEs, and the opposite pattern in EIEs (Fig. 3A). This pattern is consistent with the compositional bias of overall exonic and intronic sequences (22), and with that of known ESSs (8). To understand the relationship between mononucleotide and oligonucleotide strand asymmetries, we asked if the base composition reflects only neutral evolution by examining the relationship between the observed strand asymmetry of hexamers and that expected from their base composition. Strikingly, ESEs and ESSs can be largely separated based on the strand asymmetry predicted from the base composition in exons and all three types of intronic regions (Fig. 4). This again suggests that the skewed base composition may be also constrained by splicing-coupled selection, probably because many SREs are degenerate and ubiquitous, and have nucleotide compositional biases.

However, we cannot exclude other selective pressures that might also cause mononucleotide asymmetry, especially in exons. Indeed, for coding exons, the three positions of codons have very different patterns of strand asymmetry, suggesting that the bias is in part related to protein-coding (SI Fig. 9). Importantly, at the four-fold degenerate (synonymous) sites (14), which are under the weakest selective pressure from the protein-coding perspective, we found that C>G and T>A (SI Fig. 10). This pattern is distinct from the overall pattern of coding exons and that of noncoding exons (Fig. 1 and SI Table 1). The excess of C over G is consistent with our model of splicing-coupled selection, although other interpretations have been proposed to relate this bias to RNA secondary structure (23). The excess of T over A cannot be readily explained by our model. However, we noticed that the abundance of A increases near the splice sites, where ESEs are more abundant (24). A similar position-dependent skewness has been recently found for certain amino acids, and is related to the enrichment of ESEs near splice sites (25).

5.4 Discussion

Detecting noncoding sequences under functional selection is an important step to decode the genetic information in the genome. In this study, we provide evidence for splicing-coupled selective forces, and characterize the resulting sequence patterns in human exons and introns, including deep intronic sequences. The widespread evidence of selection in multiple-exon genes, accounting for one third of the human genome, is surprising. Previous studies estimated that 5% of nucleotides in the genome are under evolutionary constraints, as deduced from multiple-species sequence alignments (26). In most cases, deep intronic sequences were assumed to be nearly neutral, unless significant cross-species conservation was detected. However, these alignment-based methods may fail to detect sequences under weak selection, due to the difficulties in precise alignment. In addition, these studies used four-fold degenerate sites or ancient repeats as a practical proxy for neutral sequences, which may also result in an underestimate of constrained sequences.

The widespread selection is consistent with, and provides further insight into, the current understanding of mechanisms that confer splicing fidelity. We have recently shown that alternative splicing events that represent evolutionary precursors or errors are prevalent in mammals, and weakly deleterious, so that a purifying selective force is discernible (27). Indeed, the distribution pattern of ESEs and ESSs, compared with that of random elements, cannot be explained by neutral evolution. The enrichment of ESEs (ISSs) in exons and their depletion in introns, together with the opposite pattern for ESSs (ISEs), maximizes the distinction between exon and intron identity, and therefore maximizes splicing fidelity. The same trend—albeit weaker in magnitude—in deep intronic sequences suggests selective pressure to suppress pseudo-exons. Therefore, the present genome has evolved into an optimal landscape to discriminate between exons and introns. Although the different densities of SREs in exons and introns were previously noted, earlier studies could not identify the exact pattern of selective constraints, due to the lack of a neutral model (4, 8, 9). In contrast, we employed the baseline from the strand-asymmetry pattern of random elements to gauge if SREs are more enriched or depleted than expected by chance. We note that the SREs we used for this purpose were originally derived from screens of random-sequence libraries inserted into the alternative exon of a splicing reporter. As far as we can tell, there is no apparent bias among these SREs due to the base composition, protein-coding or other characteristics of human genes. Therefore, the pattern of strand asymmetry of SREs we observed is unlikely to be artifactual.

An application of the characteristic strand-asymmetry landscape is to predict new SREs. We predicted elements with the strongest strand asymmetry in exons and introns as EIEs and IIEs, respectively. The number of hexamers showing significant asymmetry is considerably larger than the sets of SREs identified in several previous studies (4, 5, 8). According to comparisons with known ESEs and ESSs, our method is very effective in recovering many known elements. Therefore, many previously unknown elements are expected to be functional SREs as well, although further experimental validation will be required. However, the predictions could also include elements involved in other steps of post-transcriptional regulation. For example, elements with strongest asymmetry in 3'UTRs were recently used to predict microRNA targets (28). On the other hand, lack of asymmetry, e.g., for palindromic sequences, does not exclude a possible function in splicing regulation. Another potential caveat in this method is the assumption of symmetric neutral sequences to assign a significance value of strand asymmetry for each hexamer. This may represent an over-simplification, because background asymmetry might exist due to asymmetric, yet neutral mutation or repair processes. A solution to this problem is to control for low-order strand asymmetry (i.e., asymmetric base composition) using a Markov model. However, useful information might be lost in the process, as we observed that strand asymmetry estimated using merely base composition can largely distinguish between known ESEs and ESSs. As a proof of principle, here we used the simplest approach, before this issue can be addressed more rigorously in future studies. Although the significance level assigned to each hexamer might be biased, this does not affect the conclusion that the hexamers with the strongest asymmetry are more likely to be functional SREs, as observed in practice.

The correlation between higher-order strand asymmetry (e.g., hexamers) and that of low order (e.g., mononucleotides) can be at least partly explained by the degeneracy in the binding specificity of SR proteins and hnRNPs. As a general mechanism for splicing fidelity, the splicing machinery needs to have sufficient flexibility and robustness so that it can recognize signals embedded in various sequence contexts. This is especially important in coding exons, where splicing signals are superimposed on the more restrictive protein code. A direct consequence of the degeneracy of the binding motifs is that SREs are highly ubiquitous. Therefore, the higher-order constraints are also reflected in the base composition, because exonic (intronic) nucleotide substitutions towards EIEs (IIEs) are favored for the discrimination between exons and introns (29). However, we could not distinguish whether exonic and intronic sequences adapted to the specificity of the splicing machinery during early evolution or vice-versa. Given the considerable differences in both the exonic and intronic strand-asymmetry patterns, and in splicing-regulatory proteins, across eukaryotic species, a co-evolution scheme appears more likely, such that multiple selective forces and mutational processes can be reconciled to be compatible with the nearly optimal genetic code (30).

5.5 Materials and methods

Data compilation

Constitutive internal exons and introns for six species (human, mouse, rat, zebrafish, *D. melanogaster*, and *C. elegans*) were compiled from our database dbCASE (http://rulai.cshl.edu/dbCASE), which was based on high-quality transcripts (mRNA/EST) and genome alignment. The data were filtered to include only exons and introns flanked by AG/GT splice sites and supported by ≥ 4 transcripts. To exclude primary splicing signals, the first 1 nt and last 3 nt of exons, and the first 10 nt and last 30 nt of introns were removed. To avoid overlap of sequences, only exons ≥ 144 nt were used for 5'E and 3'E regions; similarly, only introns ≥ 240 nt were used for 5'I and 3'I regions (Fig. 1A). Repeat-masked sequences in different regions were then extracted. Alignments of yeast protein-coding genes were downloaded from the UCSC genome browser (assembly Oct. 2003, the SGD gene track), from which exons and introns were extracted. Introns that overlap with other genes were excluded. Nucleotides that overlap with primary splicing signals were also removed similarly.

The coding information of dbCASE constitutive exons was based on CDS annotations of RefSeq transcripts to identify coding and 5'UTR exons. To minimize contamination of 5'UTR exons by coding sequences, we further filtered the data by checking each putative noncoding exon against coding exons of all RefSeq and UCSC Known Gene exons. A putative noncoding exon was removed if there was any overlap with coding exons. Similarly, intronless genes were extracted according to the aligned RefSeq transcripts, followed by the exclusion of those overlapping with any other genes (e.g., embedded in the intronic region of another gene). Stop-codon usage was obtained from Codon Usage Database (31).

Experimentally determined ESEs and ESSs

Several previous studies identified ESEs or ESSs by screening a library of random sequences inserted into an alternative exon of a minigene as a splicing reporter, although technical details varied (3, 7, 8). The SREs identified by these studies represent a relatively unbiased sample of SREs, which are not restricted to a few specific splicing factors, and are therefore appropriate to characterize general distribution patterns of SREs. Another compilation of ESEs identified in separate experimental studies was also examined (10). Because the original SRE sequences are relatively long, they had to be converted into hexamers to calculate strand asymmetry. For the ESSs, 103 hexamers that appear at least three times among ESS decamers, dubbed FAS-hex3, were derived in the original study (8) and were used here. For the other three ESE datasets, the original sequences were converted into overlapping hexamers, resulting in 220 (Cooper ESEs), 386 (Kole ESEs), and 279 (literature ESEs) hexamers, respectively.

Calculation of strand asymmetry

For each type of sequence from exons, 5’I, 3’I or midLI regions, the strand asymmetry (skewness) of a mononucleotide or oligonucleotide (hexamer in particular), was calculated by

$$S=(N_s\text{-}N_{\mathrm{a}})/(\ N_s+N_{\mathrm{a}}), \qquad (1)$$

where N_{s} and N_{a} denote its total count in the sense and antisense strands of sequences, respectively (32). In particular, the mononucleotide TA asymmetry and GC asymmetry were denoted as S_{TA} and S_{GC}, respectively. At the mononucleotide level, we also calculated strand asymmetry for each nucleotide position, in the five types of regions (5’E, 3’E, 5’I, 3’I and midLI), to study the dependence on the distance of the position to the splice sites.

. The standard deviation of strand asymmetry was estimated by $2\sqrt{r(1-r)/N}$ using the binomial distribution, where r=(N_{s}+1)/(N+2) and N= N_{s} + N_{a}.

The expected strand asymmetry of a hexamer was also predicted from the base composition (f_A, f_C, f_G , f_T) of the sequences under consideration:

$$S^{\exp} = \left(\prod_{i=1}^{6} f_{B_{i,s}} - \prod_{i=1}^{6} f_{B_{i,a}}\right) \Big/ \left(\prod_{i=1}^{6} f_{B_{i,s}} + \prod_{i=1}^{6} f_{B_{i,a}}\right), \quad (2)$$

where $B_{i,s}$ and $B_{i,a}$ represent the base at position *i* of the hexamer in the sense and antisense strands, respectively.

Predicting EIEs and IIEs using strand asymmetry

To test the significance of the strand asymmetry, we made the simplifying assumption that under the neutral model, the sequences are symmetric, i.e., *r*=0.5, although strand asymmetry of base composition was observed. The reason for this assumption is that we found a correlation between SREs and base composition, and suspected that the base composition might have been skewed by selection (see Discussion). We tested the null hypothesis by a normal approximation, $z = (r - 0.5) \Big/ \sqrt{r(1-r)/N}$. A hexamer is predicted as an EIE if the z-score calculated using exon sequences is ≥5, which corresponds to the significance level *p*=0.001 after Bonferroni correction. Similarly, a hexamer is predicted as a 5'IIE or 3'IIE if the z-score calculated in 5'I or 3'I is ≥5. The two sets of IIEs largely overlap, and were pooled together.

Statistical analysis

The difference in strand asymmetry between two groups was tested by a chi-square test using the software R (http://www.R-project.org).

5.6 Acknowledgements

We thank Drs. Phillip Green, Xavier Roca, and Xiaowo Wang for valuable comments on the manuscript. This work was supported by NIH grant GM74688 (M.Q.Z. and A.R.K.).

5.7 References

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

2. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing *Annu. Rev. Biochem.* **72,** 291-336.

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

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

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

6. Smith, P. J., Zhang, C., Wang, J., Chew, S. L., Zhang, M. Q., & Krainer, A. R. (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers *Hum. Mol. Genet.* **15,** 2490-2508.

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

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

9. Zhang, X. H.-F. & Chasin, L. A. (2004) Computational definition of sequence motifs governing constitutive exon splicing *Genes Dev.* **18,** 1241-1250.

10. Zheng, Z.-M. (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression *J. Biomed. Sci.* **11,** 278-294.

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

12. McCullough, A. J. & Berget, S. M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection *Mol. Cell Biol.* **17,** 4562-4571.

13. Sorek, R. & Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse *Genome Res.* **13,** 1631-1637.

14. Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals *Nature Rev. Genet.* **7,** 98-108.

15. Lin, H. J. & Chargaff, E. (1967) On denaturation of deoxyribonucleic acid .2. Effects of concentration *Biochim. Biophys. Acta* **145,** 398-409.

16. Albrecht-Buehler, G. (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions *Proc. Natl. Acad. Sci. USA* **103,** 17828-17833.

17. Svejstrup, J. Q. (2002) Mechanisms of transcription-coupled DNA repair *Nature Rev. Mol. Cell Biol.* **3,** 21-29.

18. Green, P., Ewing, B., Miller, W., Thomas, P. J., Program, N. C. S., & Green, E. D. (2003) Transcription-associated mutational asymmetry in mammalian evolution *Nature Genet.* **33,** 514-517.

19. Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y., & Thermes, C. (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes *Nucleic Acids Res.* **32,** 4969-4978.

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

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

22. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome *Nature* **409,** 860-921.

23. Chamary, J. V. & Hurst, L. (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals *Genome Biol.* **6,** R75.

24. Fairbrother, W. G., Holste, D., Burge, C. B., & Sharp, P. A. (2004) Single nucleotide polymorphism - based validation of exonic splicing enhancers *PLoS Biol.* **2,** e268.

25. Parmley, J. L., Urrutia, A. O., Potrzebowski, L., Kaessmann, H., & Hurst, L. D. (2007) Splicing and the evolution of proteins in mammals *PLoS Biol.* **5,** e14.

26. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project *Nature* **447,** 799-816.

27. Zhang, C., Krainer, A. R., & Zhang, M. Q. (2007) Evolutionary impact of limited splicing fidelity in mammalian genes *Trends Genet* **23,** 484-488.

28. Cora, D., Di Cunto, F., Caselle, M., & Provero, P. (2007) Identification of candidate regulatory sequences in mammalian 3' UTRs by statistical analysis of oligonucleotide distributions *BMC Bioinformatics* **8,** 174.

29. Ke, S., Zhang, X. H. F., & Chasin, L. A. (2008) Positive selection acting on splicing motifs reflects compensatory evolution *Genome Res.***,** 10.1101/gr.070268.070107.

30. Itzkovitz, S. & Alon, U. (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences *Genome Res.* **17,** 405-412.

31. Nakamura, Y., Gojobori, T., & Ikemura, T. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000 *Nucleic Acids Res.* **28,** 292.

32. Baisnee, P.-F., Hampson, S., & Baldi, P. (2002) Why are complementary DNA strands symmetric? *Bioinformatics* **18,** 1021-1033.

5.8 Figures

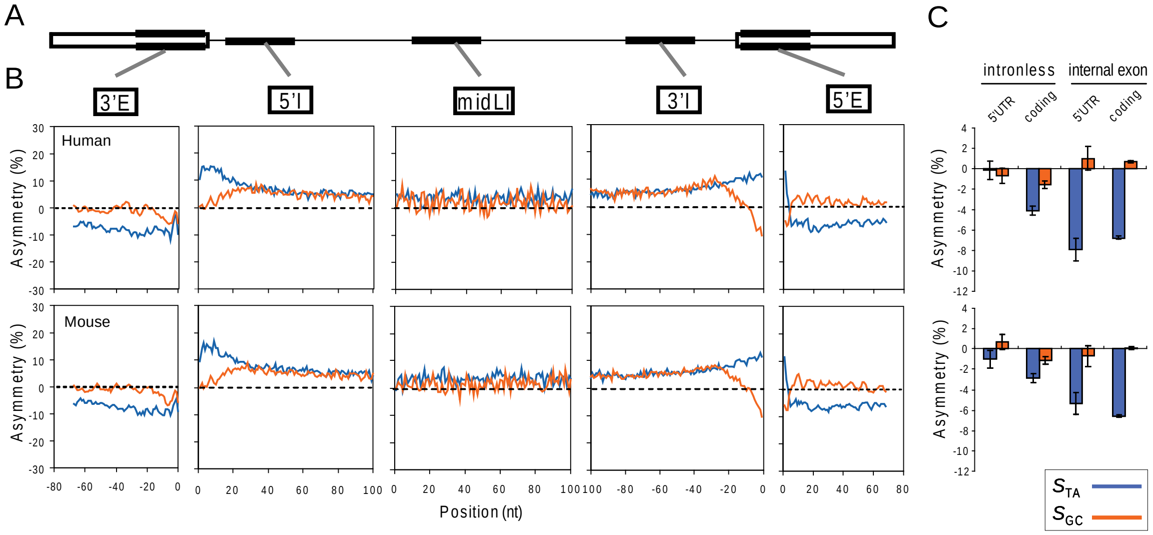


Figure 1: A landscape of mononucleotide strand asymmetry in exons and introns.

A. Diagram of five types of region analyzed in this study.

B. Strand asymmetry of human (top) and mouse (bottom) genes. TA and GC asymmetries are shown in blue and red, respectively. For intronic regions (5'I, midLI, and 3'I), strand asymmetry was calculated for each nucleotide position. For exonic regions (5'E and 3'E), strand asymmetry was calculated in sliding 3-nt windows, to smooth out the differences among the three positions of codons. Note that the coordinates in the abscissa are not relative to the splice sites, because nucleotides that are part of the consensus motifs were removed.

C. Strand asymmetry of coding and 5' UTR exons, and coding and 5' UTR portions of intronless genes for human (top) and mouse (bottom). Average strand asymmetry was calculated for each whole exon or region. Error bars represent the 95% confidence interval. The color-coding scheme is the same as in (**B**).

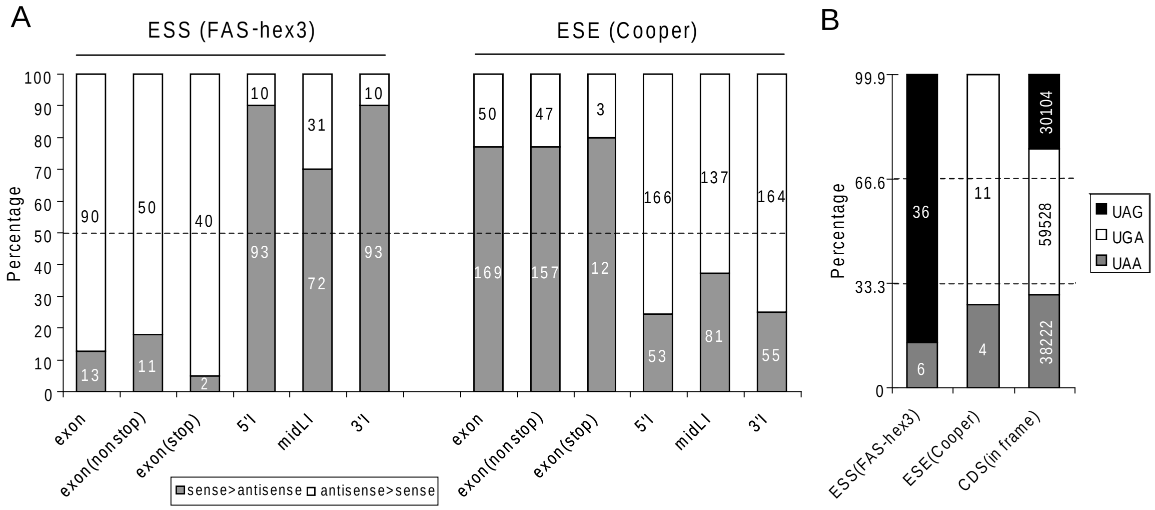


Figure 2: Distinct strand-asymmetry patterns of known SREs that discriminate between exons and introns.

A. The percentage of SREs with positive asymmetry (filled areas) and SREs with negative asymmetry (blank areas). In exons, SREs were also subdivided into two groups, depending on whether the hexamer comprises a stop codon or not, and the percentages were calculated separately for each group. Actual counts are shown inside each box.

B. The frequencies of the three stop codons in ESS hexamers, ESE hexamers, and coding sequences (CDS). Actual counts are shown inside each box.

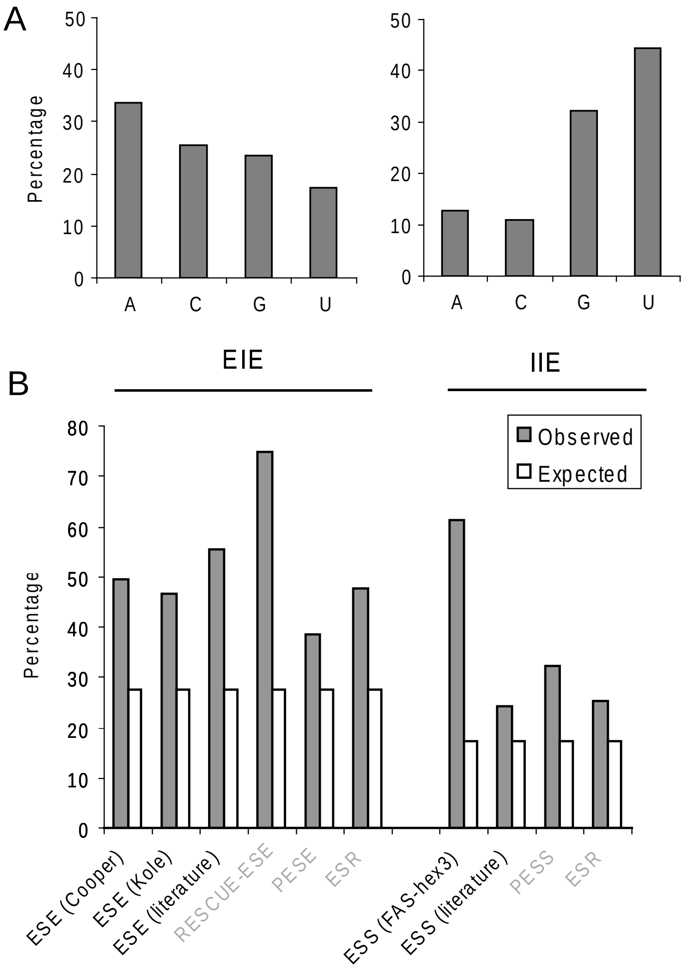
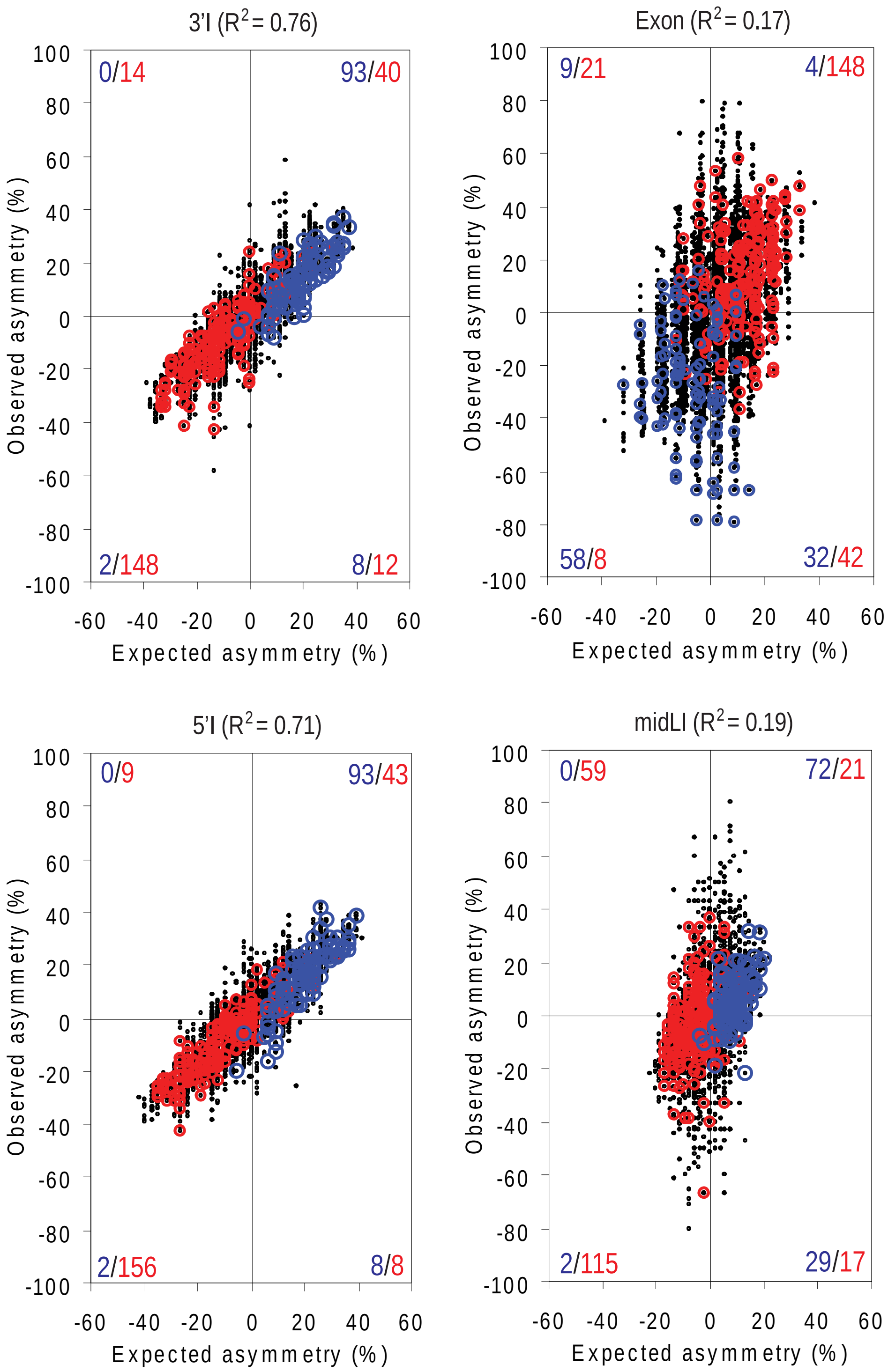


Figure 3: Predicted EIEs and IIEs using strand asymmetry.

A. Asymmetric base composition of EIEs and IIEs.

B. Overlap of predicted EIEs or IIEs with known SREs. The height of each bar represents the observed (filled bars) or expected (blank bars) percentages of previously reported SREs predicted here as EIEs or IIEs. In all the comparisons, the overlaps are statistically significant (*p*<0.007 in the worst case). Computationally-derived SREs, which might have implicit biases for this comparison, are labeled in gray letters.



(see legend at the next page)

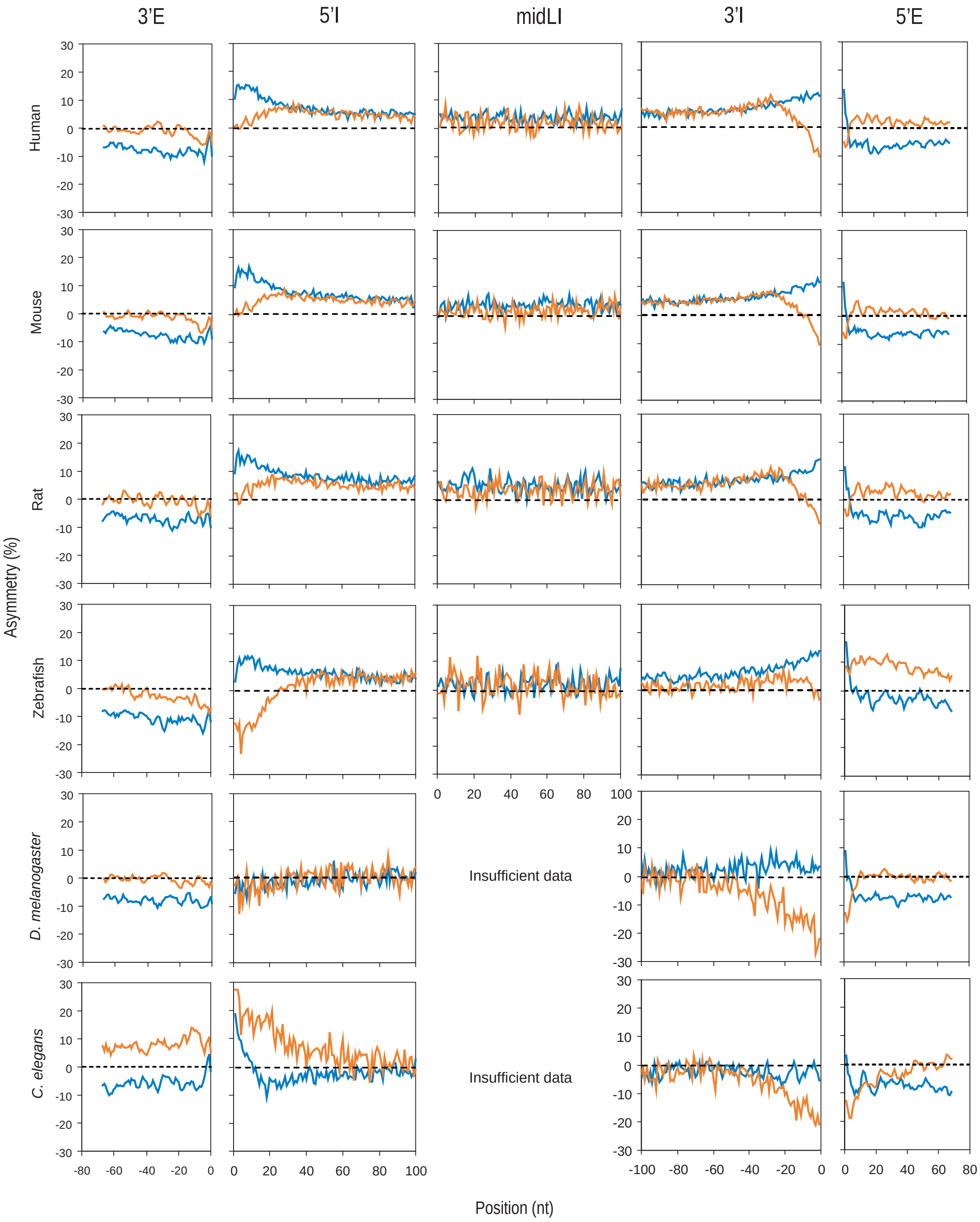
Figure 4: Correlation between the strand asymmetry of hexamers and their base composition.

For each panel, a black dot represents a hexamer. The ordinate shows the strand asymmetry of each hexamer calculated from its observed occurrences (high-order asymmetry). The abscissa shows the strand asymmetry of each hexamer expected from mononucleotide composition (low-order asymmetry). The squared Pearson correlation of the two values is indicated at the top. The FAS-hex3 ESSs and Cooper ESEs are overlaid and highlighted by blue and red circles, respectively. The number of ESSs or ESEs in each quadrant is also given in blue and red, respectively.

5.9 Supporting information

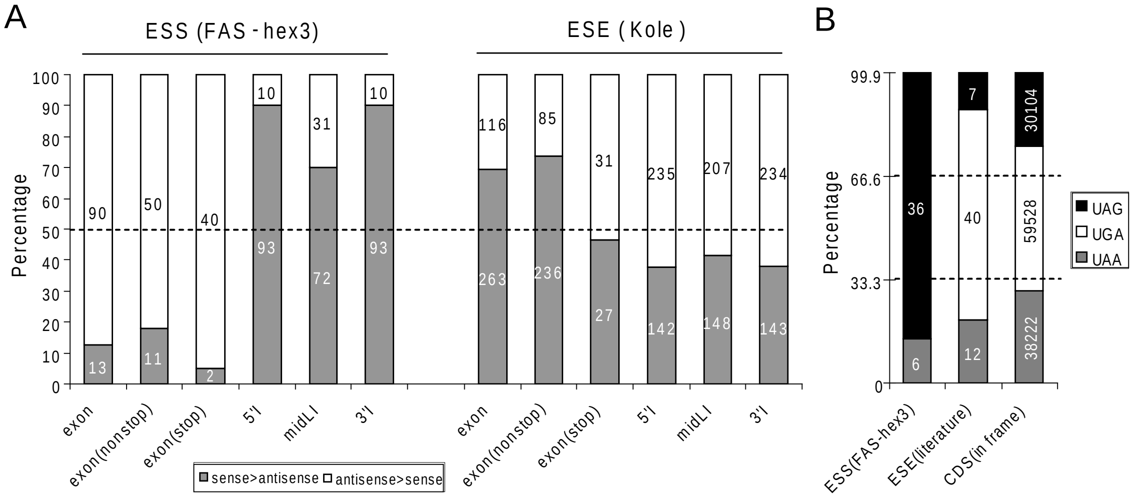
SI Table 1: Strand asymmetry of constitutive exons and introns, and coding and noncoding portions of intronless genes.

Dataset	Seq. Number	S_{TA} (%)	S_{GC} (%)
Exon (human)	27,351	-6.76	0.64
Exon (mouse)	36,009	-6.52	0.03
5’E (human)	9,565	-5.23	1.45
5’E (mouse)	12,852	-5.28	0.81
3’E (human)	9,565	-7.90	-1.28
3’E (mouse)	12,852	-7.61	-1.42
5’I (human)	26,820	7.39	4.69
5’I (mouse)	36,037	7.76	4.79
3’I (human)	26,820	6.84	4.72
3’I (mouse)	36,037	6.45	4.24
midLI (human)	8,632	3.77	1.94
midLI (mouse)	9,630	3.49	1.51
Coding exon (human)	26,130	-6.75	0.70
Coding exon (mouse)	34,397	-6.56	0.10
5’UTR exon (human)	642	-7.91	1.03
5’UTR exon (mouse)	765	-5.34	-0.68
Intronless coding region (human)	505	-4.08	-1.55
Intronless coding region (mouse)	484	-2.84	-1.10
Intronless 5’UTR (human)	522	-0.14	-0.66
Intronless 5’UTR (mouse)	524	-1.00	0.69
Exon (yeast)	553	-7.15	6.46
Intron (yeast)	258	-2.06	3.79



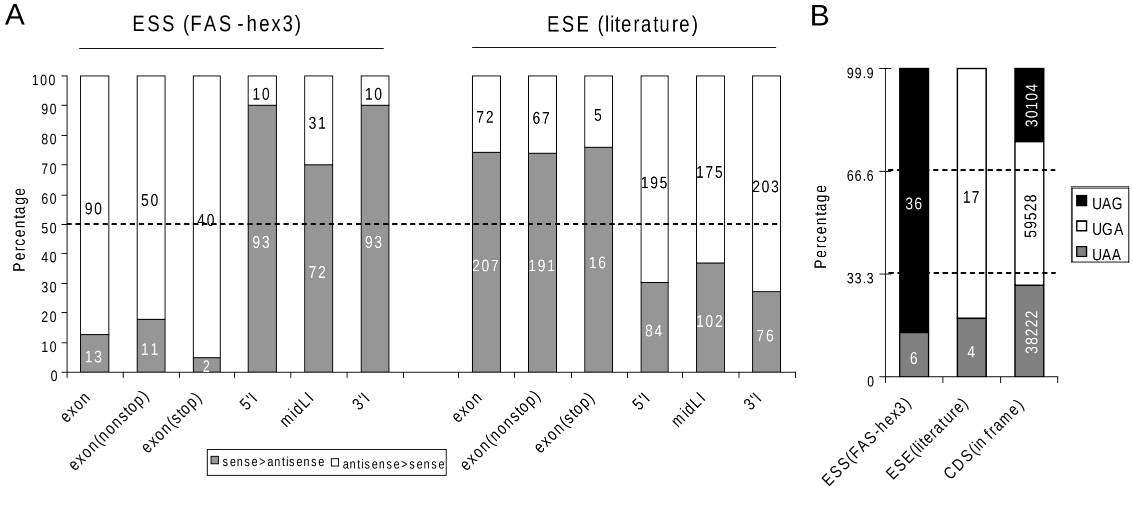
SI Figure 5: Landscape of strand asymmetry for six metazoan species.

See Fig. 1B in the main text for more details.



SI Figure 6: Distinct strand-asymmetry patterns of *in vitro* SELEX ESEs that discriminate between exons and introns.

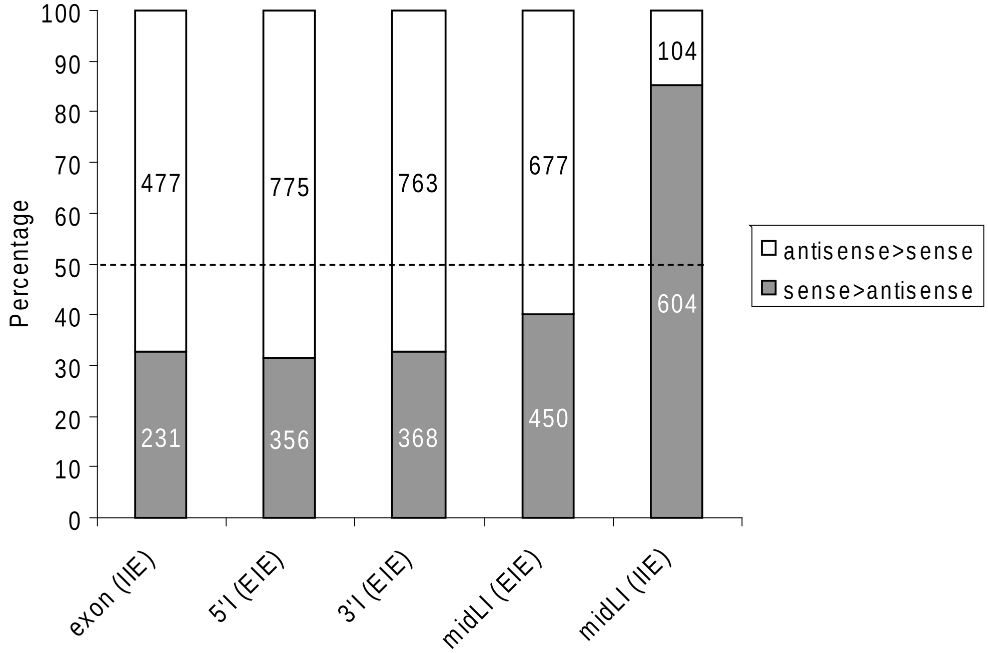
Data for ESSs are the same as in Fig. 2 in the main text. ESE hexamers (Kole ESEs) were derived from *in vitro* SELEX experiments. See Fig. 2 in the main text for more details.



SI Figure 7: Distinct strand-asymmetry patterns of published ESEs that discriminate between exons and introns.

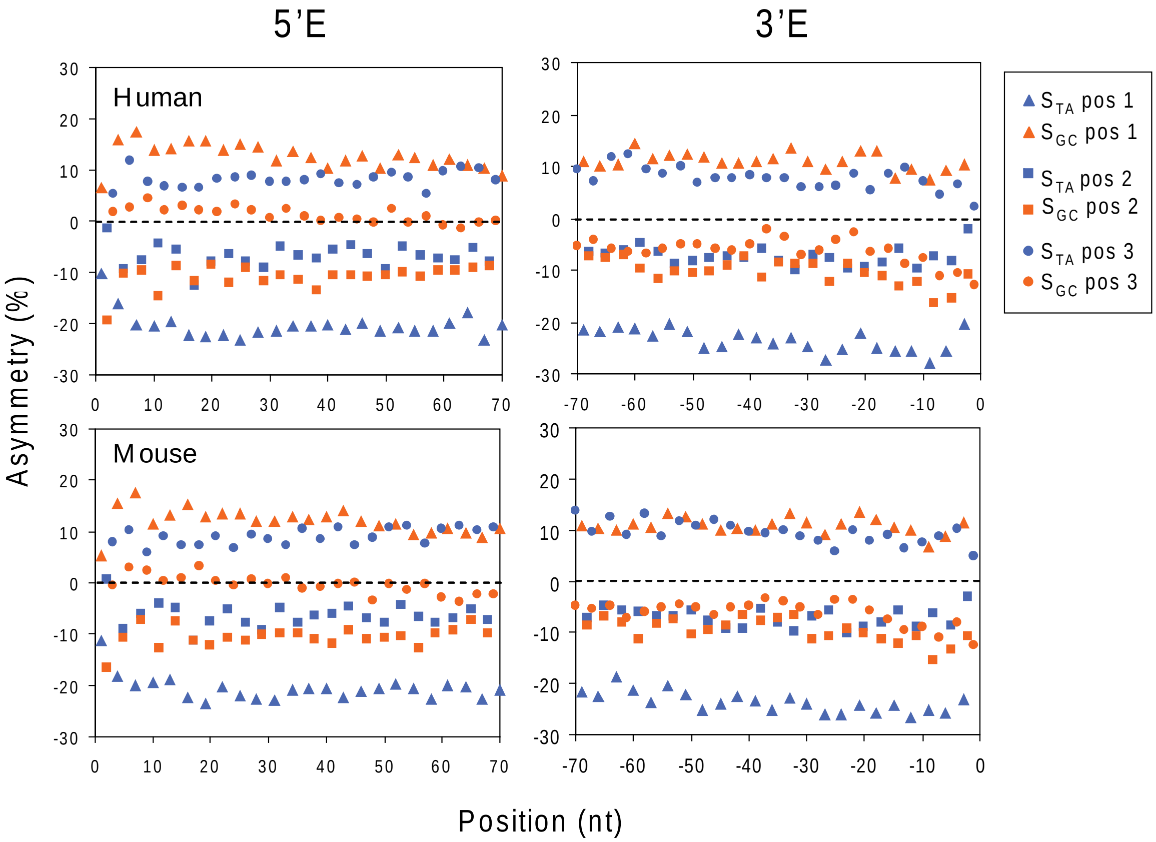
Data for ESSs are the same as in Fig. 2 in the main text. ESE sequences (literature ESEs) were determined experimentally in separate studies. See Fig. 2 in the main text for more details.

χ^2	85	155	138	46	353
p	<2.2e-16	<2.2e-16	<2.2e-16	1.4e-11	<2.2e-16



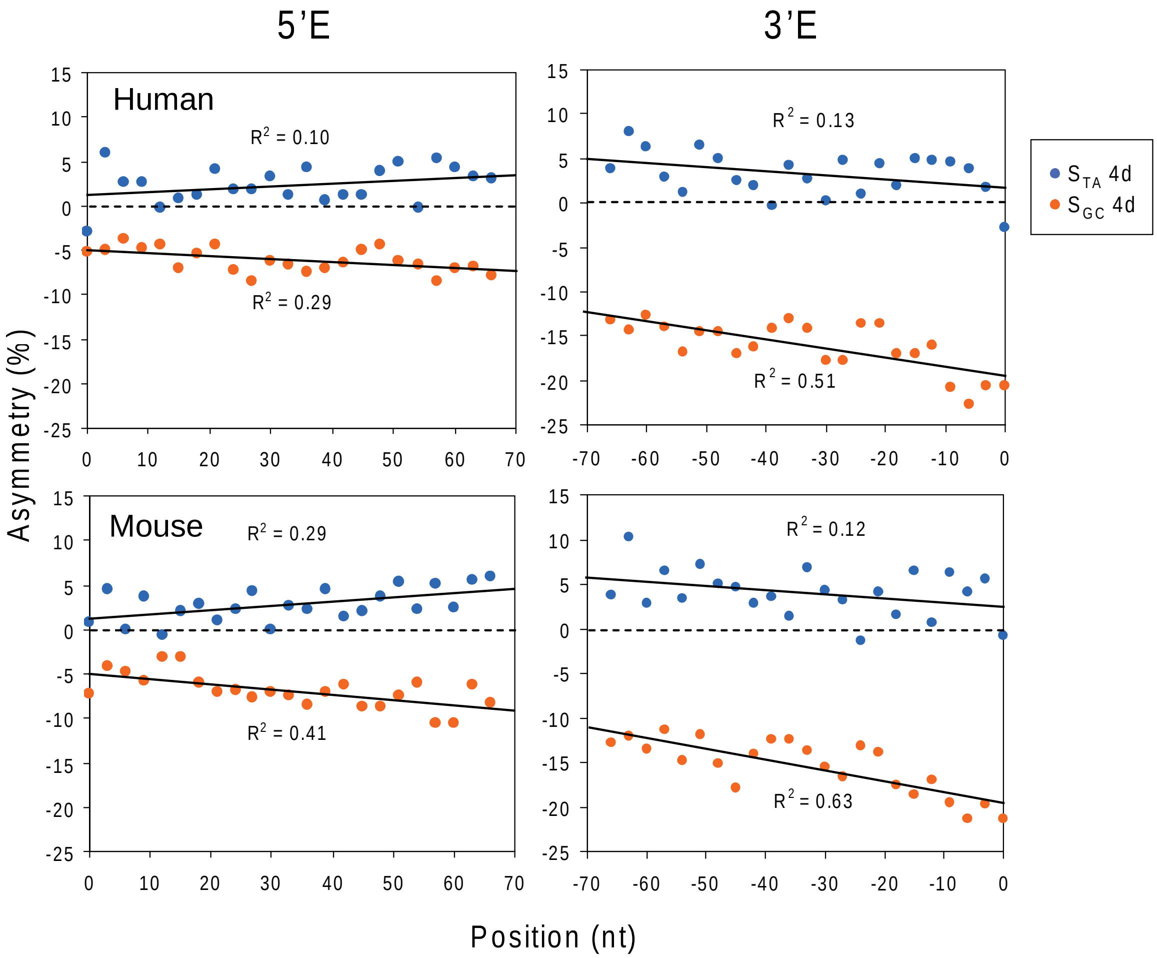
SI Figure 8: Distinct strand-asymmetry patterns of predicted EIEs and IIEs that discriminate between exons and introns.

The percentage of EIEs or IIEs with positive or negative asymmetry was calculated in different regions (exon, 5’I, 3’I and midLI), as indicated at the bottom. Actual counts are indicated inside each box. The significance of the deviation of each percentage from random was evaluated by a chi-square test, as shown at the top.



SI Figure 9: Different characteristics of strand asymmetry in the three positions of codons.

Coding exons were aligned by truncating nucleotides of incomplete codons at both ends. Strand asymmetry was calculated for each nucleotide position, similar to Fig. 1B in the main text. Strand asymmetry is shown by triangles, squares, and circles for the three positions, respectively. TA asymmetry is shown in blue and GC asymmetry is shown in red. See Fig. 1B in the main text for more details.



SI Figure 10: Correlation of strand asymmetry at four-fold degenerate sites with the distance to the splice sites.

The data are similar to SI Fig. 9, but only the strand asymmetry at the four-fold degenerate sites was calculated. In each panel, the squared Pearson correlation between the strand asymmetry and the distance to the splice sites is also indicated.

Chapter 6

Defining the splicing regulatory network of tissue-specific splicing factors Fox-1 and Fox-2

6.1 Abstract

The precise regulation of many alternative splicing (AS) events by specific splicing factors is an essential layer of post-transcriptional gene regulation to determine tissue types and developmental stages. However, the molecular basis of tissue-specific AS regulation and the properties of splicing-regulatory networks (SRNs) are only partly understood. Here we chose the brain- and muscle-specific splicing factor Fox-1 (A2BP1) and its paralog Fox-2 (RBM9) as a model system to predict their targets and define the SRNs. Fox-1/2 are conserved from worm to human, and specifically recognize the RNA element UGCAUG. We integrate Fox-1/2 binding specificity with phylogenetic conservation splicing- microarray data, and additional computational and experimental characterization. We predict thousands of Fox-1/2 targets with conserved binding sites at a false discovery rate (FDR) around 24%, including dozens validated experimentally, suggesting a surprisingly extensive regulatory network. The preferential position of the

binding sites differs among different types of AS, and determines either activation or repression of exon recognition. Many predicted targets are important for neuromuscular functions, and have been implicated in several genetic diseases, providing interesting candidates for further experimental investigation. We also identified instances of binding-site turnover (creation/loss) in different lineages and in different human populations, which likely reflect fine-tuning of gene-expression regulation during evolution.

6.2 Introduction

The sequencing of complete genomes has revealed that complex metazoans, including mammals, have only a moderately larger number of genes than unicellular yeast (1). A substantial amount of organismal complexity must have resulted from mechanisms of diversifying expression products, and temporal and spatial patterns, from a limited set of genes. Understanding how gene expression is orchestrated to determine developmental stages, specify cell types, and respond to external stimuli represents an important task in the post-genomic era (2). In the past decade, large-scale studies leveraged by high-throughput technologies and comparative genomics have started to provide profound insights into gene-regulatory networks. This is especially true for the initial step of gene expression, i.e., transcription. For example, transcriptional regulation can be very extensive, with a single transcription factor specifically regulating hundreds to thousands of targets (3-6).

Alternative splicing (AS), a process of removing introns from pre-mRNA transcripts and joining exons in different combinations, is an essential step of post-transcriptional gene expression regulation and a major source of proteomic diversity (7, 8). In mammals, as many as two-thirds of genes are alternatively spliced (9). The choice of exons and splice sites is largely determined by many RNA-binding protein, or splicing factors, which interact with *cis*-regulatory elements to activate or repress particular splicing events. Some splicing factors, including SR proteins and hnRNPs, are ubiquitously expressed, and likely important for most, if not all, constitutive or alternative splicing events (7, 8).

In addition, many other splicing factors have more restrictive and dynamic expression patterns, and play important roles in tissue-specific or developmentally regulated splicing of particular transcripts. However, the regulation and impact of these AS events remain poorly understood. A well-studied example is Sxl, Tra, Tra-2, and several other splicing factors, which regulate a cascade of AS events during *Drosophila* development, to determine the sex phenotype (10, 11). In mammals, splicing factors important for tissue-specific splicing include Nova-1/2 (12), PTB/nPTB (13, 14), Fox-1/2 (15), Muscleblind like (MBNL) (16) and CELF-family proteins (17), Hu proteins (18), TIA1/TIAR (19, 20), and probably many more that remain to be characterized. The identification of RNA targets for these factors is critical for understanding the splicing-regulatory networks (SRNs), but remains challenging; in most cases, only a handful of targets have been determined by separate experimental studies. Recently, the development of high-throughput technologies that monitor mRNA-isoform abundances and protein-RNA interactions, including splicing microarrays (9), RIP-chip (21) and CLIP assays, provided new opportunities to identify *in vivo* RNA targets and characterize SRNs genomewide. Indeed, these approaches have been used to study Nova-1/2 targets, which revealed important functions of co-regulated Nova-1/2 targets in the neuronal synapse and in axon guidance, as well as mechanisms by which Nova-1/2 activate or repress exon inclusion depending on the locations of their binding sites (22-24). However, applications of these high-throughput technologies to other splicing factors are still lacking.

Computational target prediction for specific splicing factors is even more challenging, largely due to the small size and degeneracy of splicing-factor binding motifs. An exception to this degeneracy is the hexanucleotide UGCAUG, which is an important intronic element for the splicing of several exons (25-27). A computational study further suggested that the element is enriched in the introns downstream of a set of neuron-specific exons (28). Recently, several groups identified the zebrafish and mammalian homologs of *Caenorhabditis elegans* fox-1 as the splicing factor recognizing the (U)GCAUG element (15, 29). In *C. elegans*, the *fox-1* gene is critical in the sex-determination pathway for X-chromosome dosage compensation (11). In mammals, *Fox-1* (also known as *A2BP1*) encodes an RNA-binding protein initially identified as an interacting partner of ataxin-2, and has at least one paralog, *Fox-2* (also known as *RBM9* or *Fxh*) (30). Both proteins are exclusively or preferentially expressed in brain, heart and skeletal muscle. In addition, mutations or abnormal expression of *Fox-1* has been found in patients with several genetic diseases, including epilepsy, mental retardation (31), autism (32-34) and heart disease (35). Fox-2 was also implicated in hormone signaling, as a corepressor of tamoxifen activation of the estrogen receptor (36). Therefore, Fox-1/2 are likely essential regulators for tissue-specific splicing, and systematic analysis of their targets may provide important insights into understanding the mechanisms of tissue-specific splicing regulation, the characteristics of SRNs, and their physiological roles.

In this study, we used Fox-1/2 as a model to define and characterize SRNs of tissue-specific splicing factors. We combined computational predictions from comparative genomics analysis, with experimental validation and characterization. Strikingly, our analysis revealed thousands of potential Fox-1/2 targets with binding sites highly conserved across vertebrate species. Fox-1/2 can activate or repress splicing depending on the locations of their binding sites, and contribute to more complex splicing patterns. Many of the predicted targets play important roles in neuromuscular functions and disorders. We also discuss the evolution of Fox-1/2 binding sites across different vertebrate lineages and among different human populations, and their potential phenotypic implications.

6.3 Results

Overview of the strategies to predict and characterize Fox-1/2 targets

Integrating tissue-specific splicing information from ESTs (37) and splicing microarrays (38-41) usually helps to improve the specificity of RNA target prediction. However, sensitivity is a serious concern, due to the low coverage and/or signal/noise ratio of these data sources. To characterize global features of SRNs, we sought to develop an effective method for genomewide Fox-1/2 target prediction, primarily based on the specific sequence motif UGCAUG, which we assumed to be necessary but not sufficient for regulation by Fox-1/2. To improve both specificity and sensitivity, we took advantage of the availability of 28 sequenced vertebrate genomes to perform a phylogenetic analysis, which is expected to effectively reduce false positive predictions (42). This strategy is based on the observation that the Fox-1/2 proteins are highly conserved, especially in the RNA-binding domain, across vertebrates, insects and worms [(43) and Fig. S1], as are several putative targets (44). Therefore, we began by predicting Fox-1/2 targets from all human internal exons with nearby UGCAUG elements that are significantly conserved in vertebrate species (Fig. 1).

More specifically, we compiled 204,305 human internal exons, with annotations of associated AS events, from high quality EST/cDNA-genome alignments [Materials and methods, see also (45)]. The 28-species multiz alignments of exonic sequences, and 200-nt of upstream and downstream intronic flanking (UIF and DIF) sequences, which total 1.9G-nt including orthologous sequences, were then extracted to search for Fox-1/2 binding sites in each species. The orthologous sites in the same alignment columns were grouped to define unique binding sites. For each of these unique sites, the conservation level was evaluated by a branch-length-score (BLS) method, which was adapted from previous studies with minor modifications (46). In this method, the conservation of a unique Fox-1/2 binding site is measured by the total branch length of the phylogenetic tree over which the site is conserved, normalized by the total branch length of the tree spanning all species. To determine the significance of conservation, we initially limited the search to 25,363 cassette exons. Fifty random motifs were generated by random permutations. The occurrences of these random motifs were searched to determine the distribution of BLS expected by chance, and thereby the statistical significance of observed BLS scores for Fox-1/2 sites. We then extended our search for conserved Fox-1/2 binding sites to all internal exons using the conservation thresholds determined from cassette exons. We expect that these predicted exons represent the conserved components of the Fox-1/2 SRNs. The predictions are then subject to various computational and experimental validation and characterization steps in order to reveal the underlying mechanisms of tissue-specific splicing regulation and their functional roles.

Comparative genomics analysis defines extensive Fox-1/2 SRNs with high specificity

We initially studied cassette exons, which is the most frequent form of AS in mammals (47). Without the constraint of cross-species conservation, the UGCAUG element is 1.4 fold and 1.6 fold more enriched in the UIF and DIF sequences, respectively, and slightly under-represented in exons, compared to random motifs with the same nucleotide compositions and controlled dinucleotide frequencies. Similar estimates were obtained by using the occurrence of the UGCAUG element on the antisense strand as an alternative control. This observation is consistent with and extends previous studies (28, 44, 48, 49). However, the analysis also suggests that target prediction relying only on the Fox-1/2 sequence motif has limited specificity.

We then explored how cross-species conservation can improve the specificity of Fox-1/2 target prediction. We first examined pairwise species conservation, i.e., the fraction of conserved sites between human and each of the other 27 vertebrate species, by treating intronic and exonic sites separately. The conservation rate of Fox-1/2 sites and random-motif sites decays exponentially with divergence from human (Fig. S2). Importantly, the rate of decay for Fox-1/2 binding sites is much slower than that of random sites. For example, 20% and 1.7% of human intronic Fox-1/2 sites are conserved in mouse and zebrafish, respectively, compared with 7.5% and 0.18% observed for random sites (Fig. S2A). Therefore, strong purifying selective pressure on the Fox-1/2 binding sites is manifested on a genomewide scale. Interestingly, no excess of the Fox-1/2 site conservation was found in the exonic region compared to random motifs, when only mammals were examined. However, we started to observe stronger selective pressure on Fox-1/2 sites in comparisons with non-mammalian vertebrates (Fig. S2B). Therefore, comparative analysis is not only able to improve the prediction of intronic binding sites, but also to predict to some extent exonic binding sites embedded in coding sequences.

Based on these observations, we adapted a BLS method to identify Fox-1/2 binding sites with significant conservation (46). As shown in Fig. 2 A and C, the conserved fraction of Fox-1/2 sites in UIF and DIF sequences is higher than that of random sites for the whole range of BLS thresholds. Consistent with the pairwise comparisons, the conserved fraction of Fox-1/2 sites in exonic sequences is less than that of random sites at lower thresholds, and becomes more enriched as the threshold increases (Fig. 2 B). In all three regions, the false discovery rate (FDR) decreases as the BLS threshold increases. For intronic sites, a BLS threshold of 0.22 achieves an FDR of 0.24 and 0.15 in UIF and DIF sequences, respectively (Fig. 2 A and C, indicated by an arrowhead in each panel). For exonic sites, a much more stringent threshold, i.e., BLS ≥ 0.8, is required to achieve comparable specificity (FDR = 0.24) (Fig. 2B, indicated by an arrowhead). As a tradeoff between sensitivity and specificity, we used these thresholds to predict Fox-1/2 target exons that are potentially functional and conserved during evolution.

Overall, we predicted 1,706 exons (including 1,457 nonoverlapping ones), from 1,103 genes, with at least one conserved Fox-1/2 binding site, compared with 407 exons expected by chance (overall FDR=24%). However, we note that higher specificity can be achieved by using more stringent BLS thresholds, especially for intronic sites, or by requiring the occurrence of multiple sites. For example, 192 exons have at least two sites in the same or different regions. With the same thresholds, only five exons are expected by chance (FDR=0.026).

The predicted target exons have a number of characteristics similar to known regulated tissue-specific exons. Among them, 757 exons (44.4%) are alternatively spliced, and this proportion is significantly larger than the overall fraction of AS exons in the human genome (25.7%, p=5×10^{-63}), or than the fraction of AS exons in all exons with conserved random sites (35.0%, p= 2×10^{-14}) (Fig. 3A). In addition, although other types of AS events are associated with predicted Fox-1/2 targets, cassette exons (48.6%) were significantly over-represented compared with the expected proportion estimated from all AS exons (29.6%, p=1.4×10^{-27}), or from AS exons with conserved random sites (35.8%, p=7.9×10^{-12}) (Fig. 3A). Importantly, a large proportion of AS exons predicted as Fox-1/2 targets have conserved AS patterns. For example, among the 544 cassette exons (including those with other types of AS and in the "multiple" category), 276 (50.7%) have conserved splicing patterns in mouse and/or rat, which is much higher than the overall conservation rate (10-20%) of AS events estimated previously [reviewed by (50)]. Furthermore, predicted target cassette exons show a significantly smaller size (75 nt vs

110 nt, median), and a higher preference for preserving the reading frame (67.5% vs 42%), compared with all cassette exons. We expect that the proportion of AS exons we observed is still an underestimate. Many AS exons with low EST coverage may be mistakenly classified as constitutive exons. Among the exons currently without evidence of AS, 83 exons (8.7%) have mouse and/or rat orthologous exons associated with AS events, and 176 exons (18.5%) are predicted as alternative conserved exons (ACEs) (51). However, we also note that predicted target exons that appear to be constitutively spliced have a higher FDR, compared with those that are AS exons (25.8% vs 18.8%).

In summary, comparative analysis of multiple genomes appear to be highly effective to predict functional Fox-1/2 targets. Our analysis suggests that thousands of exons and genes are potentially regulated by Fox-1/2 to generate tissue-specific mRNA and protein products. To the best of our knowledge, this represents the largest estimate of RNA targets that can be recognized by a single tissue-specific splicing factor. The extent of the SRN is surprising and comparable to the gene-regulatory network of certain master transcription factors (3-6), which was not previously appreciated.

Different types of AS events correlate with distinct patterns of Fox-1/2 motif distribution

Since all typical types of alternative exons and splice sites we studied are present in our predicted Fox-1/2 targets, we examined the distribution of Fox-1/2 binding sites separately for each type of AS exon. Interestingly, the positional preference of Fox-1/2 binding sites differs among different types of AS events. More specifically, for cassette exons, conserved Fox-1/2 binding sites are 1.75-fold more enriched in the intronic sequences downstream of the alternative exon (DIF region), compared with intronic sequences upstream of the alternative exon (UIF region) (Fig. 3B). This preferential location is significantly different from the distribution of conserved random sites, which are approximately equal in the two regions (p=1.8×10^{-8}). In contrast, there is no preference between the UIF and DIF regions for mutually exclusive exons or constitutive exons (Fig. 3 C and F). Of particular interest, for exons with alternative 5’ and 3’ splice sites, conserved Fox-1/2 sites tend to be more enriched in the intron involved in alternative splice-site selection (Fig. 3 D and E). This is particularly true for alternative 5' splice sites, for which the DIF region has 3.9-fold more putative binding sites than the UIF region (Fig. 3D). This preference is consistent with, but cannot be completely explained by, the generally higher level of sequence conservation in intronic sequences regulating alternative splice-site selection (52). Although random motifs are also more enriched in the UIF region for alternative 3' splice sites, no preference is observed for alternative 5' splice sites for random motifs. As another line of evidence, we examined the distribution of Fox-1/2 binding sites for all exons with alternative 5' or 3' splice-site selection, without requiring cross-species conservation. Again, putative Fox-1/2 binding sites are 1.2-fold more enriched in the DIF region than the UIF region for exons with alternative 5' splice sites, whereas a slightly greater enrichment in the UIF region is observed for exons with alternative 3' splice sites (p=0.0036). The preference for Fox-1/2 sites to be located near alternative splice sites suggests that Fox-1/2 may play an important role in the differential selection of alternative splice sites in a tissue-specific manner. Furthermore, the higher enrichment of putative binding sites near 5' splice sites may indicate that Fox-1/2 regulate AS more frequently via influencing 5' splice-site recognition.

Splicing patterns of Fox-1/2 targets across tissue suggest position-dependent and combinatorial regulation

We next asked if Fox-1/2 can enhance or repress splicing differently depending on the location of the presumptive binding sites. Studies of several exons, such as human *ATP5C1* (*F1γ*) exon 9, fibronectin EIIIB, *c-src* N1 exon, , *EWS* exon 4', suggested that Fox-1/2 binding sites in the intron downstream of the alternative exon usually enhance exon inclusion whereas upstream binding sites have the opposite effect (15, 53). However, whether this position-dependent regulation is generally true for Fox-1/2 is unclear. Other studies on a larger scale focusing on brain or muscle-specific exons found enrichment of the Fox-1/2 motif in the DIF region, but failed to find enrichment or depletion in the UIF region (38, 54). To understand the general mechanisms of Fox-1/2 mediated splicing regulation, we examined the splicing patterns of predicted Fox-1/2 targets in a panel of 47 tissues and cell lines, as measured by custom-designed splicing microarrays. This microarray platform includes both exon and exon-junction probes, which interrogate constitutively and alternatively spliced regions detected from EST/cDNA data (Castle et al, submitted), and can therefore monitor the abundance of both genes and individual mRNA isoforms. The presence of probes for each AS isoform allows more accurate measurement of AS patterns, compared with an earlier design, which had probes tiled only for one RefSeq transcript for each gene (9). Among the 544 cassette exons (including those in the "multiple" category in Fig. 3A), 234 exons are covered by the microarray for both the inclusion and skipping isoforms, and we therefore chose them for further analysis (Fig. 4A). For each exon, a splicing index was used to measure the change of exon inclusion level in each particular condition relative to a reference pool (24, 55). We also extracted the transcript abundance of *Fox-1* and *Fox-2* in the same tissue panels, as a proxy for protein levels. Consistent with previous observations, *Fox-1* and *Fox-2* are highly expressed in brain, heart and skeletal muscle, although *Fox-1* has a more restricted expression in these tissues (Fig. 4B). Overall, a majority (62%) of the cassette exons show a higher inclusion level in brain, heart and skeletal muscles, compared to other tissues (*p*=0.0004). This is consistent with the enrichment of conserved Fox-1/2 binding sites in the DIF region (1.6 folds), which is expected to enhance the inclusion of the upstream exon (Fig. 4C). To identify coregulated exons, we performed hierarchical clustering of both exons and tissues. This analysis successfully grouped brain, muscle and heart tissues in one cluster and other tissues in the other cluster, consistent with the expression pattern of Fox-1/2 (Fig. 4A and B).

In the case of exons, we obtained four clusters with different combinations of splicing patterns in brain and heart / skeletal muscle tissues: (i) exons with specific inclusion in both brain and heart/muscle tissues (denoted as B[+]M[+]); (ii) exons with specific inclusion only in brain tissues (denoted as B[+]M[-]); (iii) exons with specific skipping in brain and muscle/heart tissues (B[-]M[-]); and (iv) exons specifically skipped in brain tissues (B[-]M[+]) (Fig. 4A). We then compared the distribution of putative Fox-1/2 binding sites in different regions, to understand how they correlate with exon inclusion or skipping. We expect that exons mainly regulated by Fox-1/2 should have consistent splicing patterns in brain and heart/muscle, and belong to the B[+]M[+] or B[-]M[-] cluster, because Fox-1 and Fox-2 have high expression in these tissues. Therefore, focusing on these two clusters should help to infer the rules of Fox-1/2 mediated splicing by minimizing complications due to other factors. Interestingly, the B[+]M[+] and B[-]M[-] clusters have very distinct distributions of putative Fox-1/2 binding sites in different regions (Fig. 4D and F). More specifically, the B[+]M[+] cluster (Fig. 4D) shows a very strong tendency for the sites being located in the DIF region, rather than the UIF region (6-fold enrichment, p=0.007, compared with random motif sites as shown in Fig. 4H). This bias, with a magnitude much larger than previously observed (38, 54), clearly suggests that downstream binding sites are potent splicing enhancers in general. In contrast, for the B[-]M[-] cluster, we observed an opposite pattern of Fox-1/2 binding site distribution, with a 9-fold enrichment in the UIF region, rather than the DIF region (Fig. 4F) (p=0.0007, compared with random motif sites). This clearly suggests that upstream binding sites strongly repress exon inclusion. Therefore, our analysis provides strong evidence that the different effects of Fox-1/2 mediated splicing regulation generally depend on the locations of the binding sites.

We then studied the other two clusters, for which the splicing patterns differ between brain and heart/muscle tissues (Fig. 4 E and G). These complex patterns imply that in addition to Fox-1/2, other splicing factors may also play important roles in the tissue-specific splicing of these exons. For example, for the exons in the B[-]M[+] cluster (Fig. 4G), Fox-1/2 binding sites are significantly enriched in the DIF region (p=4×10^{-5}), to an extent similar to that observed in the B[+]M[+] cluster. However, these exons show very low inclusion in brain tissues. This could be explained by brain-specific repressors, which block the inclusion of these exons and counteract the enhancing effects of Fox-1/2. Alternatively, Fox-1/2 might be necessary but insufficient for the activation of these exons; other muscle/heart-specific coactivators might mediate exon inclusion together with Fox-1/2. We also note many other predicted targets have intricate splicing patterns, and may undergo more complex combinatorial regulation. Therefore, for some exons, Fox-1/2 are probably not the only determinants of tissue-specific splicing.

To confirm the splicing patterns observed from the splicing microarray data, we next performed semi-quantitative RT-PCR assays for several cassette exons. In all six cases we tested, cassette exons with conserved downstream intronic sites (*FMNL3*, *PTBP2* and *UAP1*) showed brain- and/or muscle/heart-specific exon inclusion, whereas those with only conserved upstream intronic sites (*PB1*, two exons from *MBNL1*) showed brain-and/or muscle/heart-specific skipping (Figs. 5, S3 and S4). For the *FMNL3* and *PTBP2* exons, the level of exon inclusion is similar between brain and heart/muscle, and belongs to the B[+]M[+] cluster observed from the microarray data. Similarly, the *PB1* exon inclusion is consistently low in brain and heart/muscle, and belongs to the B[-]M[-] cluster. For the *UAP1* exon and the two exons from *MBNL1*, it appears that Fox-1/2 activate or repress exon inclusion differently in brain compared to muscle/heart. This was especially true for the *UAP1* exon, which was included in muscle and heart, but predominantly skipped in brain, despite the presence of multiple downstream Fox-1/2 binding sites. Among these six tested exons, for five of them (83%, except for *FMNL3*), the tissue-splicing pattern determined by RT-PCR was consistent with the splicing microarray data. Therefore, the RT-PCR analysis suggests that that the splicing microarrays generally have high reliability. Importantly, they also support the idea that Fox1/2 alone is not always sufficient to determine the tissue-specific splicing pattern.

Overexpression and knockdown of Fox-1/2 alter the splicing of predicted Fox-1/2 targets

To further validate the predicted targets and confirm the position-dependent effect of Fox-1/2 binding sites, we next test the splicing of endogenous genes by semi-quantitative RT-PCR assays in the presence or absence of Fox-1/2 proteins. We first examined several cell lines using western blotting to see if Fox-1/2 are expressed. In all the cell lines we tested, including a few neuronal cell lines,, we found variable levels of Fox-2 protein, but not Fox-1 (data not shown). Among these cell lines, HeLa cells express a low level of Fox-2 (lane 2, Fig. 6A), which was reported to be sufficient for Fox-2-dependent splicing (53). Because RNAi in HeLa cells is very effective, we decided to use this cell line to test for alternative splicing of our predicted targets.

To compare with the standard HeLa cells, which express only Fox-2, we generated two other HeLa cell derivatives without Fox-1/2 and with only Fox-1 expression, respectively. Using a retroviral expression vector, we designed a short hairpin RNA (shRNA) to specifically knockdown endogenous Fox-2 expression (shFox-2) in HeLa cells (lane 1, Fig. 6A). The resulting stable transductant pool express neither Fox-1 nor Fox-2 proteins. We also generated a stable cell pool expressing only Fox-1 by co-transducing an shRNA against Fox-2 and a Fox-1 cDNA (shFox-2+Fox-1) (lane 3, Fig. 6A). We extracted total RNA from the three types of and performed RT-PCR analysis for predicted targets.

From the list of predicted target cassette exons, we chose to test genes with important cellular functions, such as transcription and RNA processing, and with links to genetic diseases, but others were selected at random. Among the 35 tested cassette exons with conserved downstream intronic sites, which are expressed in HeLa cells, 20 (57.1%) clearly gave a higher level of exon inclusion in the presence of Fox-1 or Fox-2 expression (Figs. 6B and S3, Table 1), whereas the rest did not show a discernible change; none of them gave a reduction in exon inclusion. These data clearly indicate the enhancer character of downstream intronic Fox-1/2 sites, and are consistent with previous studies (15, 53). Among the 22 test cassette exons with only conserved upstream intronic sites, 13 (59.1%) showed a clear change of inclusion level when Fox-1 or Fox-2 was expressed (Figs. 6C and S4, and Table 1). For most of these exons (10 of 11), Fox-1/2 expression repressed exon inclusion. However, in one case (*PLOD2*, see Fig. 6C), Fox-1/2 expression activated exon inclusion. We examined the entire length of the downstream intron and confirmed the absence of downstream sites. This result suggests that upstream intronic sites generally act as splicing silencers, with some interesting exceptions. More experiments are required to reveal the mechanistic differences among the upstream sites with different effects. Among the validated targets with conserved downstream intronic sites, two (*SFRS6* and *SULF1*) also have an upstream intronic site, which might also have silencing activity but not strong enough to counteract the enhancing effect of the downstream sites. Taken together, the RT-PCR validations strongly indicate that Fox-1/2 regulate the splicing of predicted targets depending on the locations of their binding sites in a predictable way.

Fox-1/2-mediated splicing regulation depends on the UGCAUG element

We next test if Fox-1/2 mediated splicing depends on the UGCAUG element. To this end, we generated two minigene constructs, one from the *FMNL3* gene (Fig. 7A) and the other from the *PB1* gene (Fig. 7C). The *FMNL3* minigene comprises the two constitutive exons flanking the cassette exons and both introns (Fig. 7A). Inclusion or skipping of the cassette exon likely results in the use of a different stop codon and polyA site. In the wild-type minigene, there are four conserved Fox-1/2 sites downstream of the cassette exon; the exon inclusion level greatly increased when Fox-1 was overexpressed, which recapitulates the splicing pattern of the endogenous gene (lanes 1 and 2, Fig. 7B). In contrast, Fox-1 mediated exon inclusion became much weaker or completely disappeared when two of the sites (lanes 3 and 4 with mutations in site 1 and 2; lanes 6 and 7 with mutations in site 3 and 4) or all four sites (lane 7 and 8) were mutated.

The *PB1* minigene is a chimeric construct consisting of the *PB1* cassette exon with partial flanking introns (~250-nt from the upstream and downstream introns, respectively) inserted into intron 1 of a human β-globin gene splicing reporter (Fig. 7C). There are three conserved Fox-1/2 sites upstream of the cassette exon. As shown in Fig. 7D, overexpression of Fox-1 strongly inhibited inclusion of the cassette exon (lanes 1 and 2). When we mutated one or more of the UGCAUG sites, the inhibitory effect of Fox-1 was reduced or eliminated (Fig. 7D lanes 3-10). Therefore, Fox-1/2- mediated alternative splicing of both *FMNL3* and *PB1* genes depends on the presence of UGCAUG elements. We also noticed that for these two cases, the UGCAUG element that is closest to a splice site appears to have a stronger effect than other more distal elements. These data also confirms our conclusion of the position-dependent effect of Fox-1/2 in regulating the splicing of the endogenous targets.

Predicted Fox-1/2 targets are enriched in genes important for neuromuscular functions

The large number of predicted targets raises the important question of how the Fox-1/2 SRNs are organized to perform cellular functions. To achieve a better understanding of these SRNs, we examined gene ontology (GO) terms enriched in the predicted Fox-1/2 target genes, in comparison with a control gene set derived from exons with a similar conservation level (Materials and methods). As shown in Table 2, many genes are involved in neuromuscular functions, including those related to cytoskeleton organization, ion channels, protein phosphorylation, muscle contraction, etc, which seems to be very consistent with the expression patterns of Fox-1/2. We also looked at the GO terms using orthologous genes in mouse, and found very similar annotations.

Several splicing factors known to be important regulators of brain- and/or muscle-specific splicing are also predicted as Fox-1/2 targets, including *Fox-1/2*, *PTBP1/2* (*PTB/nPTB*), *CUGBP1/2*, *NOVA1*, *ELAVL2*(*HuB*) and *MBNL1/2/3*. A previous study reported that Fox-1 can autoregulate its expression by repressing the inclusion of exon 6 (56). We predicted this exon; in addition, we also predicted four other exons in *Fox-1*, and one other exon in *Fox-2* as potential targets for auto-regulation. Among them are one of the mutually exclusive exons in *Fox-1* and its paralogous exon in *Fox-2*. The *Fox-1* exon, denoted as B40, is specifically included in brain, whereas the other mutually exclusive exon, denoted as M43, is specifically included in muscle (30). Therefore, the Fox-1/2-mediated alternative splicing of these two exons might be important in generating different isoforms of Fox-1 proteins in different tissues, which may in turn affect target-gene splicing differently. Importantly, the potential regulation of other tissue-specific splicing factors by Fox-1/2 implies that the Fox-1/2 SRNs are not limited to direct targets, but probably include a large number of indirect targets.

Consistent with the enrichment of neuromuscular genes, disruptions of our predicted Fox-1/2 target genes have been implicated in neurological, neurodegenerative, and sensory disorders, as well as heart disease and muscular dystrophy, as seen by examining genes with annotated phenotypes in the OMIM database (57). Therefore, our systematic results support several scattered observations reported in the literature (31-35). As an example, two neuroligin genes (*NLGN3* and *NLGN4X*) are mutated in patients with X-linked autism and Asperger syndrome (58). These two genes, and their paralog *NLGN2*, have a paralogous cassette exon with a very conserved downstream intronic Fox-1/2 binding site and show Fox-1/2-dependent splicing.. In addition, 15 predicted target genes, including *Fox-1* itself, show sporadic copy number variations in autistic patients, (32, 33) (personal communication, X. Zhao and J. Sebat). For complex genetic diseases, sporadic mutations can be found in many separate loci that lack apparent functional relationships. Therefore, placing the discrete disease-associated genes into a gene-regulatory network sheds light on common pathological mechanisms for these diseases. Interestingly, it appears that predicted Fox-1/2 targets are more likely to be disease genes, as 157 of 1103 predicted target genes (14.2%) are annotated in the OMIM database as disease genes, compared with a control proportion of 7.8% for all genes (p=8.3×10^{-14}), or 10.7% for genes with a comparable conservation level (p=0.0001) (Table 3). This reflects the potential pathological impact when conserved tissue-specific SRNs are dysregulated.

Creation and loss of Fox-1/2 binding sites may contribute to fine-tuning gene expression

The relatively large number of species included in our comparative analysis makes it possible to study not only the conservation of Fox-1/2 binding sites, but also the turnover (creation and loss) of the sites in specific lineages. Here we mainly focused on the intronic sites because of the mixed selective pressures due to protein coding versus splicing in exons. We estimate that ~17% of the binding sites are conserved at least in one of the five fish species we analyzed, including those in *UAP1*, Muscleblind like genes, *PBX1*, *NLGN3* and others. In contrast, ~19% of the sites are conserved only in mammals. Although these estimates are biased, due to the artificial enrichment of more conserved sites in our prediction, they nevertheless point to the evolutionary changes of Fox-1/2 splicing regulation, which may contribute to phenotypic differences across different species, or among different individuals in human populations, as illustrated in the examples below.

The first example is a 34-nt exon from *PTB* (exon 11) and *nPTB* (exon 10) (Fig. 8 A and C). The switch of expression from PTB to nPTB is important for the reprogramming of the splicing patterns of their target RNAs in developing neurons (59-61). Multiple regulators that control the PTB/nPTB switch have been recently identified, including two microRNAs that reduce *PTB* and *nPTB* transcripts in neuronal tissues and myoblasts, respectively (59, 61). At the splicing level, the inclusion of the 34-nt exon is critical for expression of the full-length functional products from both genes (60). In non-neuronal tissues, where PTB is highly expressed, *nPTB* exon inclusion is repressed by PTB, inducing nonsense-mediated mRNA decay of the truncated transcripts. In contrast, the inclusion level of the *nPTB* exon increases as the PTB expression is reduced in neuronal cells, resulting in an increased level of the full-length nPTB proteins. In addition to these known factors negatively regulating PTB/nPTB expression, we found that Fox-1/2 strongly activate the *nPTB* exon inclusion, presumably by interacting with the two downstream intronic binding sites; in contrast, the effect of Fox-1/2 on the paralogous *PTB* exon is more subtle (Fig. 8A). Further examination of the sequences near the cassette exons reveals that *PTB* has a T-to-C substitution at the first position of the binding site proximal to the 5' splice site, denoted as D-I. This mutation creates a CGCAUG element, which presumably has a much weaker binding affinity for Fox-1/2 (Fig. 8 A and C). We can further infer that the T-to-C mutation occurred in the last common ancestor of placental mammals, because an intact UGCAUG element is preserved in four non-mammalian vertebrates. However, in all placental mammals the site is lost, creating a CGCATG element. From these observations, the loss of the Fox-1/2 binding site and the creation of a conserved weak site very likely have resulted in different levels of *PTB* and *nPTBP* exon inclusion upon Fox-1/2 expression. An analogous difference between *PTB* and *nPTB* in response to a microRNA has been reported recently, although very different levels of regulation are involved (61). Taken together, the *PTB*/*nPTB* model system provides a good example of combinatorial regulation of gene expression at multiple levels, reflecting the fine- tuning of gene expression during evolution.

In the second example, we studied a 36-nt cassette exon from three Muscleblind like genes, *MBNL1*, *MBNL2* and *MBNL3*. All three exons are predicted and validated to be Fox-1/2 targets (Fig. 8 B and C). In the *MBNL1* and *MBNL2* exons, there are two Fox-1/2 binding sites: one overlapping with the polypyrimidine tract (-13 to -9, denoted as U-I) and the other in the exonic region (12 to 17, denoted as E-II). Both sites are conserved in almost all vertebrate species we analyzed, including fish. We expect that these two sites, when bound by Fox-1/2, may block the recognition of the 3' splice site. Indeed, overexpression of Fox-1 or Fox-2 reduces the inclusion isoform for both genes (Fig. 8B). Interestingly, the U-I site upstream of the *MBNL3* exon is polymorphic in human populations. More specifically, it overlaps with an A/G SNP (rs3736748) at the fourth position, resulting in two alleles UGC[A/G]UG, although the site is conserved in most vertebrate species. In addition, another site (U-III) is created further upstream from the exon (Fig. 8B). To our surprise, the *MBNL3* exon is predominantly included and the effect of Fox-1/2 expression is relatively weak, although the U-I site is intact in HeLa cells (Fig. 8B). Adding further complexity, we found that the allele frequency of the SNP differs dramatically in different populations, according to the HapMap data (χ^2=153, p=6×10^{-34}) (62, 63). Consequently, the Fox-1/2 binding site is intact in most of the African population (YRI), but is disrupted in most Asians (HCB/JPT), with Europeans (CEU) somewhere in between. This example provides a good model to study how genetic variations affect splicing regulation and result in phenotypic differences among individuals.

In these two examples, the paralogous intronic sequences, especially the Fox-1/2 binding sites, can still be aligned, despite considerable nucleotide substitutions. We found more examples belonging to this category, including another exon pair from *MBNL1/2*, *NLGN3/4*X/*4*Y, and *EBP41/41L2*. However, this is not always the case: in two pairs (or trios), one from *Fox-1/2* and the other from *ELAVL2/3/4* , the intronic sequences, including the putative Fox-1/2 binding sites, are very difficult to align. Since the Fox-1/2 sites in each paralog are significantly conserved across vertebrate species, the creation/loss of putative binding sites occurred very early after gene duplication, and was then fixed in the descendent species. Therefore, sequence divergence following gene duplication provided an independent way of producing genetic diversification, besides AS. Our results suggest that distinct protein products can be produced not only through direct amino-acid substitutions, but also through alterations of splicing patterns in the course of evolution.

6.4 Discussion

Extensiveness of Fox-1/2 SRNs

Eukaryotic gene expression is determined by multiple cellular machineries and their interactions in complex networks (2). Although the extensiveness and complexity of gene regulatory networks have been shown in a number of examples for transcriptional regulation (3-6), the mechanistic understanding of tissue-specific splicing and SRNs is very limited. So far, the best studied tissue-specific splicing factors in terms of regulatory networks are the neuronal splicing factors Nova1/2. These studies have led to two important insights: first, SRNs are highly organized and modular, so that genes co-regulated by the same splicing factor tend to be involved in related cellular functions (24); second, the effects of splicing factors on the alternative splicing of their target pre-mRNAs are highly predictable depending on the locations of their binding sites, which in principle allows elucidation of a "splicing code" (23).

Using the highly conserved and related brain-, heart- and muscle-specific splicing factors Fox-1 and Fox-2 as a model, we extended the current understanding of tissue-specific SRNs in several important aspects. We started from comprehensive computational predictions of Fox-1/2 targets based on their highly specific binding motif, UGCAUG, and comparative analysis of 28 vertebrate species. The methodology is highly effective and predicted thousands of target exons and genes with conserved Fox-1/2 binding sites. We estimate by statistical analysis that about 76% of the predicted targets are *bona fide* targets, and about 50-60% of them could be validated experimentally in HeLa cells. The validation rate in HeLa cells is somewhat lower than expected from the statistical estimate, probably due to more complex combinatorial regulation in tissues (see below). Nevertheless, these estimates suggest a high specificity of our prediction, which makes our analysis amenable to more detailed experimental follow-up. Importantly, the implied large number of *bona fide* Fox-1/2 targets suggests an unforeseen extensiveness of the SRNs, with a magnitude comparable to certain master transcription factors (3-6). This extensiveness was not apparent from previous studies, because the number of endogenous targets identified for an individual tissue-specific splicing factor was usually a few dozens or fewer (22, 24, 60). In addition, we focused only on the conserved components of the Fox-1/2 SRNs that can be predicted with high specificity and sensitivity. Many additional binding sites with a relatively low level of conservation might be also functional, as we observed by experimental validations (data not shown). Moreover, in some extreme cases, a functional Fox-1/2 binding site can be thousands of nucleotides away from the regulated exon (30); these also escaped from our predictions.

We expect that such extensive SRNs might not be unique for Fox-1/2, as previous approaches for identifying splicing-factor targets were limited in specificity and sensitivity. In addition, many other splicing factors are still poorly characterized. Our understanding of the mechanisms and impact of splicing regulation at the genome level will be greatly advanced with the development and application of new experimental and computational technologies with improved accuracies in detecting splicing factor–RNA interactions and splicing-isoform abundances.

Mechanisms of Fox-1/2-dependent exon activation and repression

The effect of splicing factors on splicing enhancement or silencing often depends on the location of the regulatory sequences they bind. This was reported both for ubiquitous splicing factors, such as SR proteins (64) and hnRNPs (65), as well as for tissue-specific splicing factors and Nova-1/2 in particular (23). However, whether similar mechanisms exist in other tissue-specific splicing factors, including Fox-1/2, has been unclear, largely due to the limited number of targets examined in previous studies (15, 29, 30, 38, 53, 54). Our comprehensive prediction of Fox-1/2 targets followed by experimental validation leads to a conclusive answer, at least in the case of Fox-1/2. Among the validated targets in HeLa cells, all tested alternative exons with downstream intronic binding sites are activated in the presence of Fox-1 or Fox-2, whereas most exons with upstream intronic or with exonic binding sites are repressed. This pattern is also consistent with our splicing microarray data and RT-PCR analysis in primary tissues. Therefore, of the opposite outcomes—splicing activation or repression—depending on the binding site locations may be a more general feature of tissue-specific splicing regulation.

Several features of Fox-1/2 binding-site distribution raise intriguing questions about the underlying regulatory mechanisms for how Fox-1/2 interact with the spliceosome to affect splicing. First, previous studies identified several Fox-1/2 targets with multiple binding sites in a repeated array nearby the alternative exon; whether a single site is sufficient for Fox-1/2 regulation is unclear. Our results suggest that only a relatively small proportion (11.3%) of predicted targets have multiple conserved binding sites, although this estimate may missed some exons with additional less conserved or distal binding sites. In a few cases, we confirmed that Fox-1/2 mediated splicing depends on a single binding site, as no other sites were found by an exhaustive sequence search in the complete flanking introns.

Second, for cassette exons, putative Fox-1/2 binding sites are generally more enriched in introns downstream of the alternative exon, with a peak around 30 nt from the exon; a smaller enrichment with a broader distribution was found in the upstream intron [(48) and data not shown]. Importantly, predicted targets associated with different types of AS events show distinct patterns of preferential binding site locations. For alternative 5' splice sites, putative Fox-1/2 binding sites have a strong preferential location in the DIF region, whereas a more moderate enrichment in the UIF region was observed for alternative 3' splice sites. The preferential enrichment at particular distances downstream of 5' splice sites suggests that Fox-1/2 might be more efficient in enhancing 5' splice-site recognition. In contrast, the mechanisms through which Fox-1/2 block exon recognition might be more heterogeneous. For example, it was reported that in the context of *hF1γ* gene exon 9, Fox-1 binding to the GCAUG element in intron 8 blocks pre-spliceosomal E-complex formation in intron 9, resulting in the skipping of exon 9 (66). We found cases (e.g., exons from Muscleblind like genes) in which the Fox-1/2 binding sites in the UIF region are very close to the downstream 3' splice site. In these cases, Fox-1/2 likely block the recognition of the intron preceding the alternative exon by interfering with binding of spliceosomal components that recognize the polypyrimidine tract and/or 3' splice site.

Third, for alternative exons with multiple Fox-1/2 binding sites, these sites are not equivalent in Fox-1/2-dependent activation or repression of exon inclusion. Rather, the sites closer to the splice sites appear to be more efficient than the distal sites, at least with the two minigenes we tested. One possible interpretation is that Fox-1/2 proteins bound to the proximal sites are more efficient at directly interacting with spliceosomal components.

Complex splicing patterns suggest potential combinatorial regulation

We noticed that the validation rate of predicted targets in HeLa cells is lower than the statistical predictions. Although we cannot rule out the possibility that the FDR based on permutations is an underestimate, another possibility is that Fox-1/2 alone are not always sufficient to affect the splicing pattern of a *bona fide* target pre-mRNA.

Both splicing-microarray analysis and RT-PCR validations suggest the existence of exons with complex splicing patterns that cannot be explained by Fox-1/2 regulation alone. Although the effect of Fox-1/2 on the splicing of these exons is not always observable because of the difficulty to identify the appropriate tissues or developmental stages, in some cases, we are able to demonstrate the requirement for other cooperatimg splicing factors in Fox-1/2-mediated splicing regulation. Our argument is based on the comparison of splicing patterns between brain and muscle/heart, in which Fox-1/2 are highly expressed. Using splicing microarrays, we identified clusters of cassette exons with inconsistent splicing patterns among these tissues, an indication of combinatorial regulation. This difference is especially pronounced for a cluster of exons with Fox-1/2 binding sites enriched in the introns downstream. These exons are predominantly included in muscle, as expected, but mostly skipped in brain tissues. One good example is the *UAP1* exon, with two downstream intronic binding sites. This exon is strongly activated by Fox-1/2 expression in HeLa cells, suggesting that it is a *bona fide* Fox-1/2 target. The low exon inclusion level in brain cannot be explained by variable expression levels of Fox-1/2, because it is even lower than that in thymus or tonsil, in which Fox-1/2 expression is very low. Therefore, there might be other splicing activators or repressors expressd and functional only in brain or heart/muscle, but not in both. A similar argument might hold for HeLa cells, which would help explain the lack of responsiveness of some exons to increased Fox-1/2 expression. A few splicing factors that interact or have the potential to interact, with Fox-1/2 have been reported recently. For example, several proteins in addition to Fox-1/2, including hnRNPs F/H and PTB/nPTB, are responsible for the neuron-specific splicing of the *c-src* N1 exon (53). The repressive activity of nPTB in such cases could explain why some exons are skipped despite the presence of downstream Fox-1/2 binding sites as potential enhancers. Alternatively, muscle-specific activators might also be important for the inclusion of these exons in heart and muscle. Very recently, one such factor, called sup-12, was identified in *C. elegans* by a genetic screen (67). sup-12 coordinately regulates tissue-specific splicing of the fibroblast growth factor receptor gene *egl-15*, by binding to a UGUGU element juxtaposed to the fox-1 binding sites. Because this protein shows a very high level of sequence conservation with the mammalian homologs (RBM38 and RBM24), it will be interesting to see if these mammalian homologs function similarly in cooperative splicing regulation with Fox-1/2.

Implications of Fox-1/2 SRNs for neuromuscular functions, disease, and evolution

This study extends previous observations and indicate that modularity may represent a more general feature of tissue-specific SRNs, with co-regulated genes sharing similar cellular functions (24). Such organization is key to the robustness of a biological system. Many of the predicted target genes are known to have important neuromuscular functions. For example, the list includes genes involved in muscle contraction, such as a number of myosin genes, *DMD*, titin (*TTN*) and tropomyosin 1 (*TPM1*). Several splicing factors known to be important for neuronal and/or muscle-specific splicing are also predicted as Fox-1/2 targets. The enrichment of genes with neuromuscular functions is consistent with the expression patterns of Fox-1/2. Not surprisingly, disruption of several of predicted target genes is implicated in various neuronal disorders, heart disease, and developmental defects. As genetic alterations in Fox-1 have also been reported in patients with some of these diseases, the Fox-1/ SRNs provide a good model to study the phenotypic effects of perturbing SRNs in *cis* or *trans* to obtain more mechanistic details.

On the other hand, splicing regulatory elements, including the Fox-1/2 binding motif, are generally short. The creation and loss of these elements by random mutations can readily occur during evolution. Not all of these mutations cause genetic diseases. Instead, some of the mutations might have only moderate effects and can therefore be tolerated. The inclusion of many species in our comparative study provides an opportunity to trace the history of each site, which may provide important information about phenotypic differences among different species or different human populations. We found that the creation and loss of putative Fox-1/2 binding sites in specific lineages are not rare events, despite the high binding specificity of these proteins. In two examples, we examined *PTB*, in which a site was likely lost in all mammalian species, and *MBNL3*, in which a site was likely lost in a majority of Asians while being preserved in a majority of Africans. Although further evidence is required, these observations are suggestive that tissue-specific SRNs might show considerable divergence between mammals and non-mammalian vertebrates, as well as among different human populations. In both cases, we compared not only orthologous sequences across many vertebrates, but also paralogous sequences. This analysis helps to understand how splicing regulation diverges after gene duplication.

In particular, the functional divergence of Fox-1 and Fox-2 paralogs likely has had profound effects on the SRNs. Several other splicing factors also have multiple paralogs, including SR and hnRNP A/B proteins, PTB/nPTB, Nova-1/2, Muscleblind like proteins, CELF family proteins, etc. In the case of PTB/nPTB, Nova-1/2, the paralogs have reciprocal expression patterns in different tissues or brain sections (60, 68). Fox-1 and Fox-2 have similarly high expression in brain and heart/muscle tissues, but Fox-2 also express in other tissues. However, because Fox-1 and Fox-2 recognize the same RNA element, it is impossible to distinguish Fox-1/2 targets through their predicted binding sites, as in the present study. Overexpression and knockdown of Fox-1 or Fox-2 individually in HeLa cells suggests that these two proteins have very similar effects in activating or repressing predicted targets. We expect that further insights will be gained by identification of the in vivo targets of Fox-1 and Fox-2 experimentally in the appropriate tissue types, as well as by determination of the mechanisms of action of these factors.

6.5 Materials and methods

Compilation of exons and AS events using splicing graphs

We built a database of classified alternative splicing events (dbCASE, http://rulai.cshl.edu/dbCASE) using high-quality transcripts (mRNA/EST) and genome alignment (coverage >85%, identity >95%), for human, mouse, rat and other model organisms (45). Briefly, transcripts from UniGene (ftp://ftp.ncbi.nih.gov/repository/UniGene/) and RefSeq (ftp://ftp.ncbi.nih.gov/refseq/release) (69) were aligned to genomic sequences using sim4 (70). The alignment of all transcripts to the same gene locus was then converted into a splicing graph, in which each splice site is represented by a node and each exon/intron is represented by an edge (52). Exons, introns, and typical types of AS events, including cassette exons, alternative 5' and 3' splice sites, and mutually exclusive exons, were detected by detected by analyzing subnetwork topologies. For this study, we mainly used 204,305 AG-GT internal exons, with associated annotations of AS events.

For each AS event in human, we also tried to identify the orthologous AS event in mouse and rat. This was done by mapping the genomic coordinates of the AS region in mouse or rat to the human genome using the tool liftOver obtained from the UCSC genome browser (http://genome.ucsc.edu). For example, for cassette exons, the alternative exons and the two flanking exons were used for the mapping. The mapped coordinates were then compared with the corresponding regions of human AS events.

Evaluation of motif site conservation

For each exon, we analyzed exonic sequences and 200-nt upstream/downstream intronic flanking (UIF/DIF) sequences. Multiple alignments of 28 vertebrate species (42) were extracted using the mafFrag program obtained from the UCSC genome browser.

To measure the level and significance of Fox-1/2 binding site conservation, a branch-length-score (BLS) approach was adapted (46). Briefly, the topology and branch lengths of the phylogenetic tree of 28 vertebrate species were obtained (42). For each block of multiple alignments, alignment gaps were removed. Fox-1/2 binding sites (UGCAUG) were then searched in each of the species using the gapless sequences. The sites were then mapped back into the multiple alignments to identify orthologous sites in the same alignment columns as unique sites, followed by the evaluation of conservation of these sites in the 28 vertebrate species. We allowed no movement of the sites in the assignment of orthologous sites, given the high-specificity and conservation of Fox-1/2 binding sites. However, in some instances, small insertions/deletions interrupt some sites, partly due to artifacts in sequence alignment (e.g. TGCATGG aligned with TGCAT-G); such indels wre tolerated. Therefore, our approach is more restrictive than the original approach (46), because we sought to trace the history of each individual site. For each unique site, the conservation was measured by the total branch length of the subtree over which the sites are present in the branches. The total branch length of the subtree was normalized using the total branch length of the phylogenetic tree of all 28 vertebrate species.

To determine the significance of motif site conservation, we estimate the null distribution of BLS using 50 random motifs generated by permutations. Random motifs containing CpG or GCAUG were avoided, because CpGs are underrepresented in vertebrate genomes and the GCAUG element is partial functional for Fox-1/2-mediated splicing. The same analysis was repeated for each of the random motifs and motif sites to calculate BLS scores. We tried different BLS thresholds from 0 to 1, with steps of 0.01 to determine an appropriate threshold for Fox-1/2 target prediction. For each threshold, a false discovery rate (FDR) was calculated by the ratio of the average number of sites with a BLS greater than the threshold for random motifs to that for the Fox-1/2 motif.

Experimental validation of the predicted Fox-1/2 targets

The human tissue total RNAs were purchased from Clonetech (Mountain View, CA). Two shRNAs against human Fox-2 were cloned in the MSCV retroviral vector as previously described (71). Human Fox-1 cDNA was cloned in the pWZL-hygro retroviral vector, expressing the Flag-tagged Fox-1 protein. To generate stable cell pools, HeLa cells were infected with MSCV (expressing the shRNA against Fox-2), or MSCV plus pWZL-hygro (expressing Fox-1 protein) vectors. We replaced the medium 24 h after infection, and 24 h later, infected cells were selected with puromycin (2 g ml^{-1}) for 72 h. In the case of double infection, cells were treated with hygromycin for 96 h after selection with puromycin. The effect of knockdown or overexpression was confirmed by western blotting using antibodies against human Fox-1 and Fox-2. Total RNAs were extracted from the stable cell pools using Trizol reagent (Invitrogen, Carlsbad, CA) and treated with DNase I. Reverse transcription was carried out using Superscript II reverse transcriptase as described from Invitrogen. Semi-quantitative PCR using Taq polymerase was performed by adding 0.1µl of [α-^{32}P]-dCTP to each 25µl reaction. The PCR reactions were run for 20-25 cycles depending on the abundance of the targets. The products were analyzed on a 6% native polyacrylamide gel.

The *FMNL3* minigene was cloned in the pcDNA3.1 vector. QuickChange PCR mutagenesis was carried out to generate the mutant constructs. Fugene 6 was used for transfection and RT-PCR analysis was done as above. The PB1 minigene was generated by inserting a PCR fragment, containing the cassette exon plus 243-nt upstream intronic region and 253-nt downstream intronic region, into intron 1 of human beta-globin gene using BglII and XhoI sites that were generated previously.

Splicing microarrays

We identified exons, exon junctions and AS events in the human genome by mapping RefSeqs, mRNAs, ESTs, and transcripts from patent databases to the genome. For each gene, 60-nt probes and 36-nt probes were optimized to monitor exons and exon junctions, respectively, printed on Agilent arrays. These arrays were used to monitor 47 diverse human tissues and cell lines in dye-swap replicates. Gene expression levels were estimated from probes monitoring constitutive exons and junctions. For each AS event, a proportional change of isoform abundances, relative to a reference pool, was then estimated using a previous method, with minor modifications (24, 55). More detailed information and data availability are described elsewhere (Castle et al. submitted).

Gene ontology (GO) term and OMIM phenotype analysis

The GO term analysis was performed using the online tool DAVID (72). DAVID gives a *p*-value, before and after multiple-test corrections, based on a modified hyper-geometric distribution. We used a background gene set controlling for the conservation level. More specifically, a gene was included in the control gene set if at least one of its exons has a consecutive hexanucleotide with a BLS greater than a specified threshold (BLS ≥0.22 for UIF and DIF sequences and BLS≥0.8 for exonic sequences). The OMIM phenotypes and associated genes were downloaded in Dec 2007 (57).

Statistical analysis

Fisher's exact test in the software R was used to evaluate the significance of two-by-two contingency tables (73).

Acknowledgements

We thank Scott Powers and David Mu for providing Fox-1 the antibody. This work was supported by NIH grant GM74688 (M.Q.Z. and A.R.K.).

6.6 Acknowledgements

We thank Scott Powers and David Mu for providing Fox-1 the antibody. This work was supported by NIH grant GM74688 (M.Q.Z. and A.R.K.).

6.7 References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome *Nature* **409,** 860-921.

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

3. Wei, C.-L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z*., et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome *Cell* **124,** 207-219.

4. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J*., et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs *Cell* **116,** 499-509.

5. Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions *Science* **316,** 1497-1502.

6. Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G., & Goodman, R. H. (2004) Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions *Cell* **119,** 1041-1054.

7. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing *Annu. Rev. Biochem.* **72,** 291-336.

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

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

10. Lopez, A. J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation *Annu. Rev. Genet.* **32,** 279-305.

11. Meyer, B. J. (2000) Sex in the worm: counting and compensating X-chromosome dose *Trends Genet.* **16,** 247-253.

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

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

14. Patton, J. G., Mayer, S. A., Tempst, P., & Nadal-Ginard, B. (1991) Characterization and molecular cloning of polypyrimidine tract-binding protein: a component of a complex necessary for pre-mRNA splicing *Genes Dev.* **5,** 1237-1251.

15. Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., & Inoue, K. (2003) A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG *EMBO J.* **22,** 905-912.

16. Ho, T., Charlet-B, N., Poulos, M., Singh, G., Swanson, M., & TA, C. (2004) Muscleblind proteins regulate alternative splicing *EMBO J.* **23,** 3103-3112.

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

18. Zhu, H., Hinman, M. N., Hasman, R. A., Mehta, P., & Lou, H. (2008) Regulation of neuron-specific alternative splicing of neurofibromatosis type 1 pre-mRNA *Mol. Cell Biol.* **28,** 1240-1251.

19. Del Gatto-Konczak, F., Bourgeois, C. F., Le Guiner, C., Kister, L., Gesnel, M.-C., Stevenin, J., & Breathnach, R. (2000) 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.

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

21. Keene, J. D., Komisarow, J. M., & Friedersdorf, M. B. (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts *Nat. Protocols* **1,** 302-307.

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

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

24. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse *Nat. Genet.* **37,** 844-852.

25. Huh, G. S. & Hynes, R. O. (1994) Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element *Genes Dev.* **8,** 1561-1574.

26. Lim, L. P. & Sharp, P. A. (1998) Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats *Mol. Cell. Biol.* **18,** 3900-3906.

27. Kawamoto, S. (1996) Neuron-specific alternative splicing of nonmuscle myosin II heavy chain-B pre-mRNA requires a cis-acting intron sequence *J. Biol. Chem.* **271,** 17613-17616.

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

29. Ponthier, J. L., Schluepen, C., Chen, W., Lersch, R. A., Gee, S. L., Hou, V. C., Lo, A. J., Short, S. A., Chasis, J. A., Winkelmann, J. C., *et al.* (2006) Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16 *J. Biol. Chem.* **281,** 12468-12474.

30. Nakahata, S. & Kawamoto, S. (2005) Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities *Nucleic Acids Res.* **33,** 2078-2089.

31. Bhalla, K., Phillips, H. A., Crawford, J., McKenzie, O. L. D., Mulley, J. C., Eyre, H., Gardner, A. E., Kremmidiotis, G., & Callen, D. F. (2004) The de novo chromosome 16 translocations of two patients with abnormal phenotypes (mental retardation and epilepsy) disrupt the *A2BP1* gene *J. Hum. Genet.* **49,** 308-311.

32. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007) Strong association of de novo copy number mutations with autism *Science* **316,** 445-449.

33. The Autism Genome Project Consortium (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements *Nat. Genet.* **39,** 319-328.

34. Martin, C., Duvall, J., Ilkin, Y., Simon, J., Arreaza, M., Wilkes, K., Alvarez-Retuerto, A., Whichello, A., Powell, C., Rao, K., *et al.* (2007) Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism *Am J Med Genet B Neuropsychiatr Genet* **144B,** 869-876.

35. Kaynak, B., von Heydebreck, A., Mebus, S., Seelow, D., Hennig, S., Vogel, J., Sperling, H.-P., Pregla, R., Alexi-Meskishvili, V., Hetzer, R., *et al.* (2003) Genome-wide array analysis of normal and malformed human hearts *Circulation* **107,** 2467-2474.

36. Norris, J. D., Fan, D., Sherk, A., & McDonnell, D. P. (2002) A negative coregulator for the human ER *Mol. Endocrinol.* **16,** 459-468.

37. Xu, Q. & Lee, C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences *Nucleic Acids Res.* **31,** 5635-5643.

38. Sugnet, C. W., Srinivasan, K., Clark, T. A., Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D*., et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays *PLoS Computat. Biol.* **2,** e4.

39. Hu, G. K., Madore, S. J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., & Wang, Y. (2001) Predicting splice variant from DNA chip expression data *Genome Res.* **11,** 1237-1245.

40. Clark, T., Schweitzer, A., Chen, T., Staples, M., Lu, G., Wang, H., Williams, A., & Blume, J. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays *Genome Biol.* **8,** R64.

41. Fagnani, M., Barash, Y., Ip, J., Misquitta, C., Pan, Q., Saltzman, A., Shai, O., Lee, L., Rozenhek, A., Mohammad, N*., et al.* (2007) Functional coordination of alternative splicing in the mammalian central nervous system *Genome Biol.* **8,** R108.

42. Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D*., et al.* (2007) 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser *Genome Res.* **17,** 1797-1808.

43. Kiehl, T.-R., Shibata, H., Vo, T., Huynh, D. P., & Pulst, S.-M. (2001) Identification and expression of a mouse ortholog of A2BP1 *Mamm. Genome* **12,** 595-601.

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

45. Zhang, C., Hastings, M. L., Krainer, A. R., & Zhang, M. Q. (2007) Dual-specificity splice sites function alternatively as 5' and 3' splice sites *Proc. Natl. Acad. Sci. USA* **104,** 15028-15033.

46. Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N*., et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures *Nature* **450,** 219-232.

47. Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V., & Muilu, J. (2004) ASD: the alternative splicing database *Nucleic Acids Res.* **32,** D64-69.

48. Yeo, G. W., Nostrand, E. L. V., & Liang, T. Y. (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements *PLoS Genet.* **3,** e85.

49. Voelker, R. B. & Berglund, J. A. (2007) A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing *Genome Res.* **17,** 1023-1033.

50. Blencowe, B. J. (2006) Alternative splicing: new insights from global analyses *Cell* **126,** 37-47.

51. Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T., & Burge, C. B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse *Proc. Natl. Acad. Sci. USA* **102,** 2850-2855.

52. Sugnet, C., Kent, W., Ares, M. J., & Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice in *Pac. Symp. Biocomput.*, pp. 66-77.

53. Underwood, J. G., Boutz, P. L., Dougherty, J. D., Stoilov, P., & Black, D. L. (2005) Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals *Mol. Cell. Biol.* **25,** 10005-10016.

54. Das, D., Clark, T. A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J. E*., et al.* (2007) A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing *Nucleic Acids Res.* **35,** 4845-4857.

55. Fehlbaum, P., Guihal, C., Bracco, L., & Cochet, O. (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants *Nucleic Acids Res.* **33,** e47.

56. Baraniak, A. P., Chen, J. R., & Garcia-Blanco, M. A. (2006) Fox-2 mediates epithelial cell-specific fibroblast growth factor receptor 2 exon choice *Mol. Cell. Biol.* **26,** 1209-1222.

57. McKusick, V. A. (1998) *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders* (Johns Hopkins University Press, Baltimore).

58. Jamain, S. (2003) Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism *Nat. Genet.* **34,** 27-29.

59. Boutz, P. L., Chawla, G., Stoilov, P., & Black, D. L. (2007) MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development *Genes Dev.* **21,** 71-84.

60. Boutz, P. L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., Jr., & Black, D. L. (2007) A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons *Genes Dev.* **21,** 1636-1652.

61. Makeyev, E. V., Zhang, J., Carrasco, M. A., & Maniatis, T. (2007) The microRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing *Mol. Cell* **27,** 435-448.

62. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs *Nature* **449,** 851-861.

63. The International HapMap Consortium (2005) A haplotype map of the human genome *Nature* **437,** 1299-1320.

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

65. Hung, L.-H., Heiner, M., Hui, J., Schreiner, S., Benes, V., & Bindereif, A. (2008) Diverse roles of hnRNP L in mammalian mRNA processing: A combined microarray and RNAi analysis *RNA* **14,** 284-296.

66. Fukumura, K., Kato, A., Jin, Y., Ideue, T., Hirose, T., Kataoka, N., Fujiwara, T., Sakamoto, H., & Inoue, K. (2007) Tissue-specific splicing regulator Fox-1 induces exon skipping by interfering E complex formation on the downstream intron of human F1{gamma} gene *Nucleic Acids Res.* **35,** 5303-5311.

67. Kuroyanagi, H., Ohno, G., Mitani, S., & Hagiwara, M. (2007) The Fox-1 family and SUP-12 coordinately regulate tissue-specific alternative splicing in vivo *Mol. Cell Biol.* **27,** 8612-8621.

68. Yang, Y. Y. L., Yin, G. L., & Darnell, R. B. (1998) The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia *Proc. Natl. Acad. Sci. USA* **95,** 13254-13259.

69. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins *Nucleic Acids Res.* **33,** D501-504.

70. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence *Genome Res.* **8,** 967-974.

71. Dickins, R. A., Hemann, M. T., Zilfou, J. T., Simpson, D. R., Ibarra, I., Hannon, G. J., & Lowe, S. W. (2005) Probing tumor phenotypes using stable and regulated synthetic microRNA precursors *Nat. Genet.* **37,** 1289-1295.

72. Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., & Lempicki, R. (2003) DAVID: database for annotation, visualization, and integrated discovery *Genome Biol* **4,** R60.

73. Ihaka, R. & Gentleman, R. (1996) R: A language for data analysis and graphics *J. Comput. Graph. Statist.* **5,** 299-314.

6.8 Tables and Figures

Table 1: Validated Fox-1/2 targets

CASE ID	Gene Symbol	Exon size	Sites UIF/Exon/DIF	Effect
CA-10006-98813-108289-108376-112247	*ABI1*	87	0,0,2	E
CA-351-151977-176401-176458-191342	*APP*	57	0,0,1	E
CA-1729-3143-33690-33717-34743	*DIAPH1*	27	0,0,2	E
CA-2037-185729-185976-186039-187932	*EPB41L2*	63	0,0,3	E
CA-84668-56874-69882-70097-70966	*FAM126A(1)*	215	0,0,3	E
CA-84668-56874-69882-70178-70966	*FAM126A(2)*	296	0,0,3	E
CA-91010-63327-63459-63574-64309	*FMNL3*	115	0,0,4	E
CA-57211-139436-141979-142025-144408	*GPR126*	46	0,0,3	E
CA-55605-115371-115854-115875-119792	*KIF21A*	21	0,0,2	E
CA-4254-67128-67844-67928-76324	*KITLG*	84	0,0,3	E
CA-54413-6346-11616-11676-13353	*NLGN3*	60	0,0,1	E
CA-4926-70633-71085-71127-72673	*NUMA1*	42	0,0,1	E
CA-4659-130706-131598-131766-132163	*PPP1R12A(1)*	168	0,0,2	E
CA-4659-130706-131634-131766-132163	*PPP1R12A(2)*	132	0,0,2	E
CA-5725-12146-13458-13492-13949	*PTBP1*	34	0,0,1	E
CA-58155-86118-87597-87631-88044	*PTBP2*	34	0,0,2	E
CA-6431-3632-4275-4543-4893	*SFRS6*	268	1,0,1	E
CA-23213-175221-177170-177204-194867	*SULF1*	34	1,0,2	E
CA-6675-31979-34199-34250-39259	*UAP1*	51	0,0,2	E
CA-5352-85380-86571-86634-87600	*PLOD2*	63	1,0,0	E*
CA-6885-42931-45556-45637-53787	*MAP3K7*	81	3,0,0	S
CA-4154-182734-190502-190538-191227	*MBNL1(1)*	36	1,1,0	S
CA-4154-182734-191227-191322-194231	*MBNL1(2)*	95	1,0,0	S
CA-4154-180500-181664-181718-182580	*MBNL1(3)*	54	1,0,0	S
CA-10150-138316-145816-145852-147139	*MBNL2*	36	1,1,0	S
CA-55796-56030-57992-58028-60418	*MBNL3*	36	1,1,0	S
CA-55193-127070-133957-134113-138019	*PB1*	156	3,0,0	S
CA-5087-255450-263372-263485-264837	*PBX1*	113	1,0,0	S
CA-55103-167532-170065-170143-171754	*RALGPS2*	78	1,0,0	S
CA-55700-20277-21696-21807-22997	*RPRC1*	111	1,0,0	S
CA-6733-265647-266020-266113-273872	*SRPK2*	93	1,0,0	S
CA-6926-6286-7192-7252-7513	*TBX3*	60	1,0,0	S

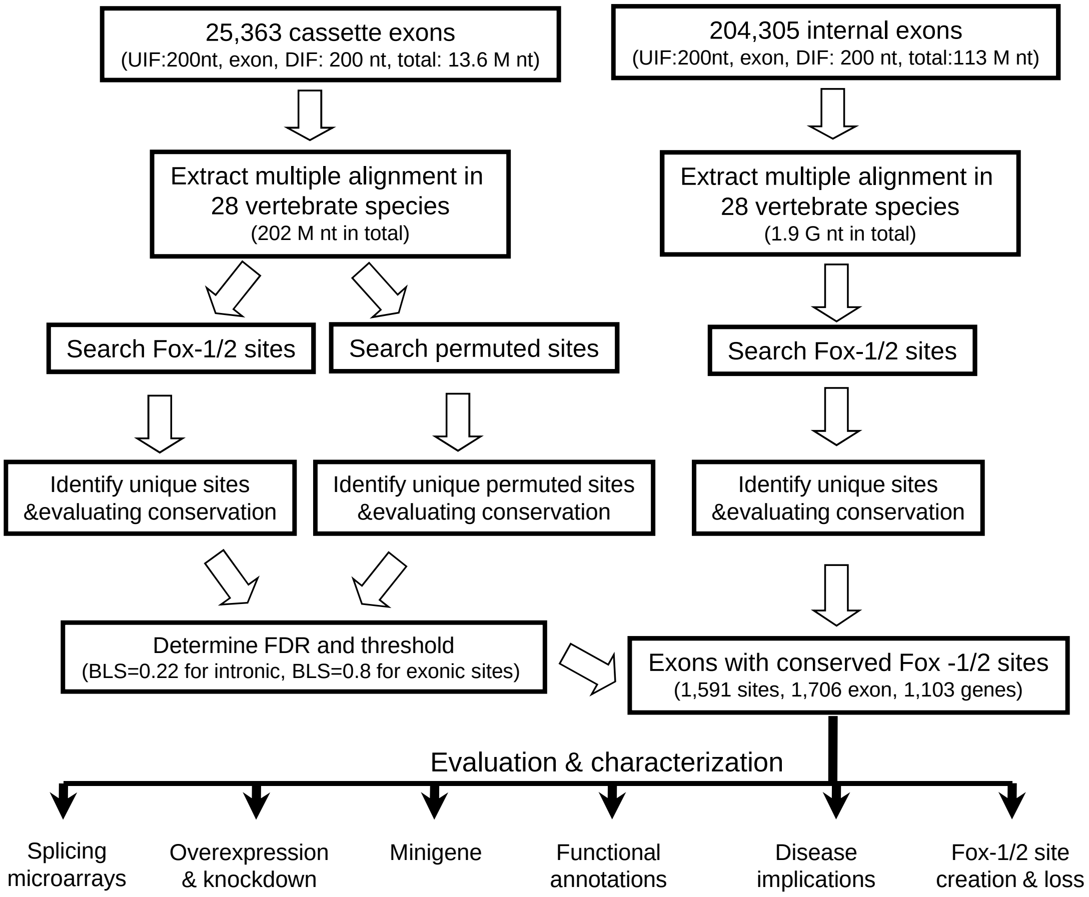
E, enhancer; S, silencer. When there are multiple cassette exons for a gene, they are distinguished by a number in parentheses. A cassette exon with conserved an upstream intronic Fox-1/2 binding site as enhancers was labeled by an asterisk. See gene structures and sequence conservation in Figs S3 and S4. More detailed information can be obtained from dbCASE by searching the CASE IDs.

Table 2: Representative Gene Ontology functions of predicted Fox-1/2 target genes

GO Term	Count	*p*-value	Fold Change	Benjamini
Biological Process				
cytoskeleton organization and biogenesis	68	1.1E-11	2.4	3.7E-08
actin filament-based process	37	6.5E-09	2.9	1.1E-05
potassium ion transport	31	1.7E-07	2.9	1.1E-04
metal ion transport	56	2.0E-07	2.1	1.1E-04
ion transport	91	3.0E-07	1.7	1.4E-04
cation transport	65	2.2E-06	1.8	9.0E-04
system development	61	2.6E-06	1.9	9.6E-04
nervous system development	59	8.3E-06	1.8	2.1E-03
muscle contraction	28	1.8E-05	2.5	4.3E-03
protein amino acid phosphorylation	70	4.2E-05	1.6	6.9E-03
Cellular Component				
cytoskeleton	123	2.5E-15	2.1	1.5E-12
actin cytoskeleton	51	1.2E-13	3.2	3.5E-11
non-membrane-bound organelle	148	7.1E-09	1.6	1.1E-06
myofibril	15	3.2E-08	5.9	3.9E-06
synapse	30	9.2E-08	3.0	9.4E-06
myosin	18	8.0E-07	4.0	7.0E-05
sarcomere	13	1.0E-06	5.5	6.7E-05
microtubule associated complex	24	1.7E-06	3.1	9.7E-05
striated muscle thick filament	9	9.1E-06	7.1	4.3E-04
A band	9	9.1E-06	7.1	4.3E-04
postsynaptic membrane	18	8.8E-05	2.9	3.6E-03
Golgi-associated vesicle membrane	9	5.3E-04	4.5	1.9E-02
Molecular Function				
cytoskeletal protein binding	79	4.4E-19	3.0	1.1E-15
actin binding	55	2.0E-13	3.0	1.6E-10
motor activity	38	1.4E-12	3.7	8.8E-10
calmodulin binding	37	2.4E-12	3.8	1.2E-09
ion channel activity	53	3.0E-07	2.1	1.0E-04
enzyme binding	35	2.3E-06	2.4	5.6E-04

Table 3: Comparison of genes documented in OMIM phenotypes

Dataset	Total Number	with OMIM phenotype
Predicted Fox-1/2 targets	1103	157 (14.2%)
all genes (with RefSeq)	22,910	1,784 (7.8%)
genes after controlling conservation	15,040	1,610 (10.7%)



25,363 cassette exons
(UIF:200nt, exon, DIF: 200 nt, total: 13.6 M nt)
204,305 internal exons
(UIF:200nt, exon, DIF: 200 nt, total:113 M nt)
Extract multiple alignment in 28 vertebrate species
(202 M nt in total)
Extract multiple alignment in 28 vertebrate species
(1.9 G nt in total)
Search Fox-1/2 sites
Search permuted sites
Search Fox-1/2 sites
Identify unique sites &evaluating conservation
Identify unique permuted sites &evaluating conservation
Identify unique sites &evaluating conservation
Determine FDR and threshold
(BLS=0.22 for intronic, BLS=0.8 for exonic sites)
Exons with conserved Fox -1/2 sites
(1,591 sites, 1,706 exon, 1,103 genes)
Evaluation & characterization
Splicing microarrays
Overexpression & knockdown
Minigene
Functional annotations
Disease implications
Fox-1/2 site creation & loss

Figure 1: Overview of strategies to define and characterize Fox-1/2-regulated targets.

Initially only cassette exons were used to determine the thresholds of Fox-1/2 binding-site conservation in upstream intronic flanking (UIF), exonic, and downstream intronic flanking (DIF) sequences. Multiple alignments for each of these regions were used to search putative Fox-1/2 binding sites (UGCAUG) in each of the species. The unique binding sites in the same alignment columns were then identified and their conservation levels were measured by branch-length scores (BLSs). The significance of each BLS was determined based on the null distribution of BLSs estimated with random motif sites. We then extracted sequences of the same regions for all human internal exons from dbCASE to search for conserved Fox-1/2 binding sites using the thresholds determined from cassette exons. Different types of computational and experimental analyses were then performed to validate and characterize predicted Fox-1/2 targets and the SRNs.

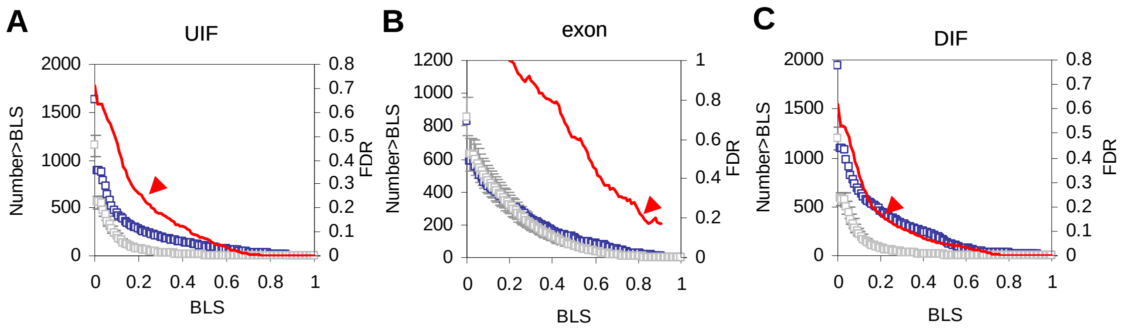


Figure 2: Comparative analysis accurately predicts Fox-1/2 targets.

Only sites present in human cassette exons are shown. The three panels are for UIF (**A**), exonic (**B**) and DIF (**C**) sequences, respectively. For each panel, the conserved fraction of Fox-1/2 binding sites and of random motif sites using varying thresholds of BLSs are shown in blue and gray, respectively (left axis). Error bars represent standard error of the mean. The corresponding FDR of prediction is shown in red (right axis). The thresholds used in the paper (0.22 for intronic sites and 0.8 for exonic sites) are indicated by an arrowhead.

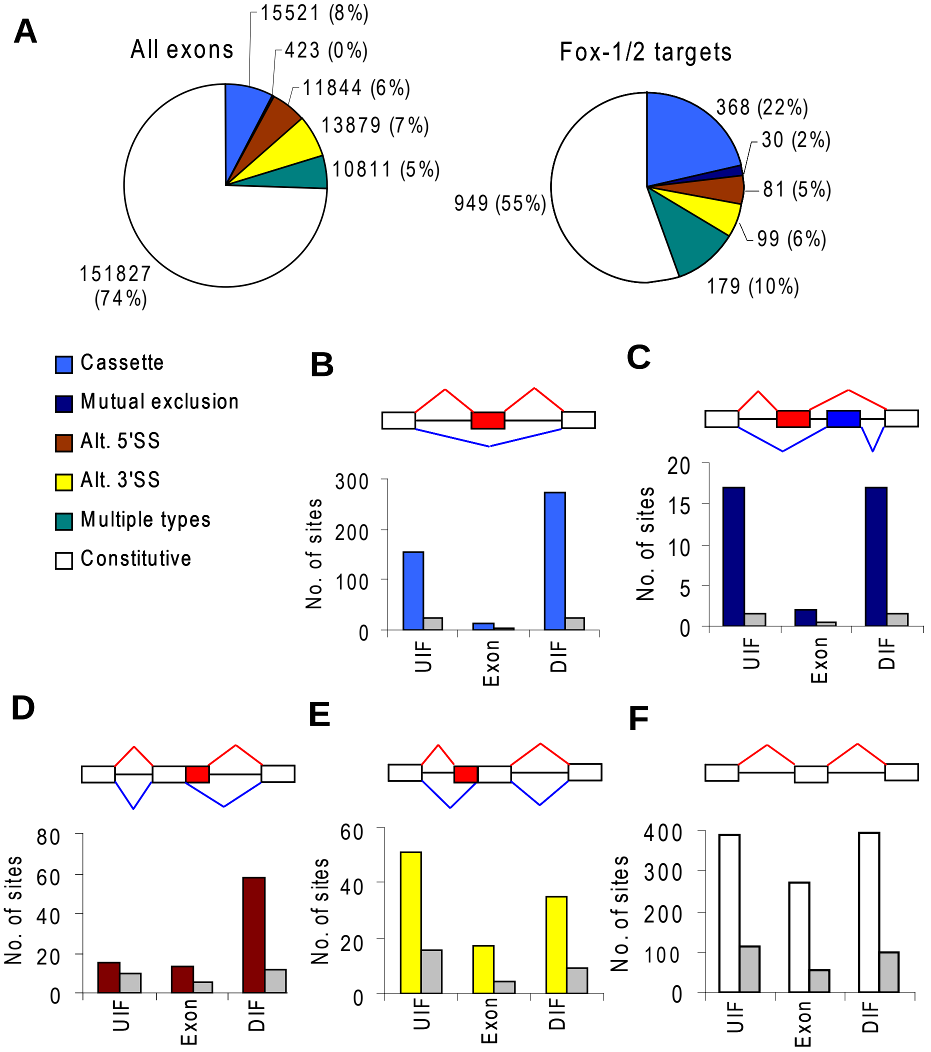


Figure 3: Different splicing patterns of predicted Fox-1/2 targets correlate with the locations of the Fox-1/2 binding sites.

(**A**) Proportions of different types of splicing patterns for all internal exons (left) and for predicted Fox-1/2 targets (right).

(**B-F**) Distribution of conserved Fox-1/2 binding sites in different regions for cassette exon (**B**) mutually-exclusive exons (**C**), alternative 5' splice sites (**D**), alternative 3' splice sites (**E**), and constitutive exons (**F**). In each panel, the splicing pattern is shown schematically above the histogram. The distribution of conserved Fox-1/2 sites is color-coded as in (**A**). The distribution of conserved random motif sites is shown in gray for comparison.

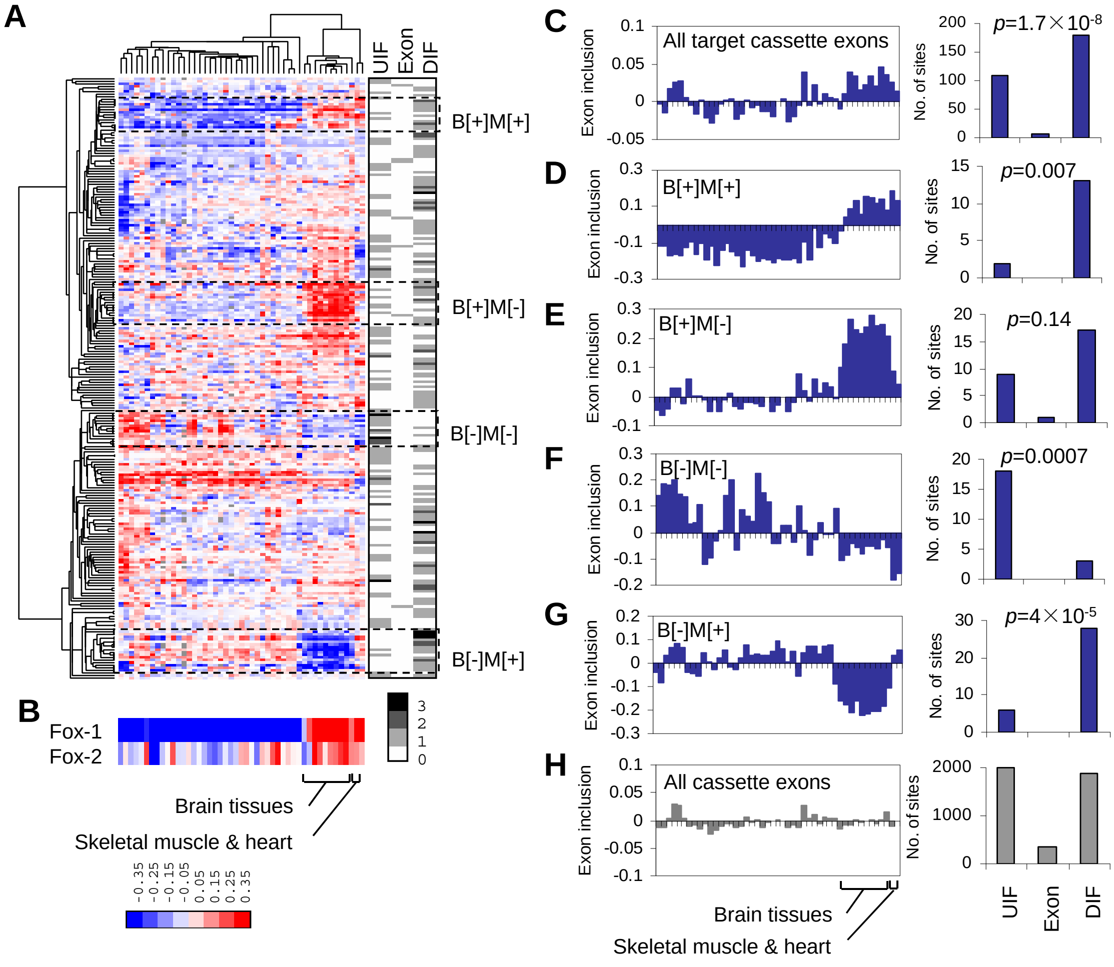


Figure 4: Splicing profiling of predicted Fox-1/2 targets shows position-dependent and complex modes of Fox-1/2-mediated splicing regulation.

(**A**) 234 cassette exons predicted as Fox-1/2 targets were analyzed as part of genomewide splicing microarrays in 47 human tissues and cell lines. The splicing index, or the proportional change of exon inclusion in each tissue relative to a reference tissue pool, was clustered by hierarchical clustering of both exons and tissues. The tissue cluster including brain tissues, heart and skeletal muscles is labeled. For each exon, the number of conserved Fox-1/2 binding sites in UIF, exonic and DIF sequences is shown on the right, in the same order as in the splicing heatmap. Four clusters of exons, with different combinations of splicing in brain and heart/muscle are labeled by dashed boxes. B[+]M[+]: high inclusion in both brain and heart/muscle; B[+]M[-]: high inclusion in brain and low inclusion in heart/muscle; B[-]M[-]: low inclusion in both brain and heart/muscle; B[-]M[+]: low inclusion in brain and high inclusion in heart/muscle.

(**B**) The expression pattern of Fox-1/2 monitored by the same arrays is displayed in the same order of tissues as in the splicing heatmap.

(**C-G**) Average expression profile (left) and distribution of conserved Fox-1/2 sites (right) for all 234 predicted targets (**C**) and exons belonging to the four clusters (**D-G**).

(**H**) As a control, the average profile of all cassette exons on the splicing microarrays is shown on the left. The distribution of random motif sites for cassette exons is shown on the right. This was used to test the enrichment/depletion of Fox-1/2 sites in different regions for all predicted targets or for targets in each of the clusters. The *p*-values from chi-square tests are also indicated in (**C-G**).

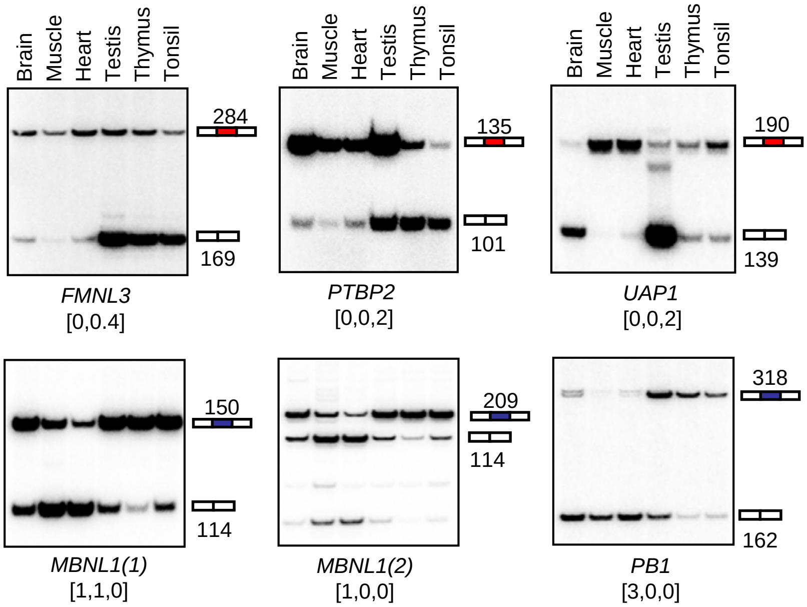


Figure 5: RT-PCR analysis of predicted cassette exons shows brain- and/or heart/muscle-specific splicing.

Exon inclusion level was measured in six human tissues by radioactive RT-PCR. Exons with downstream intronic Fox-1/2 binding sites are labeled in red, and exons with only upstream intronic Fox-1/2 binding sites are labeled in blue. The size of each PCR product is indicated.

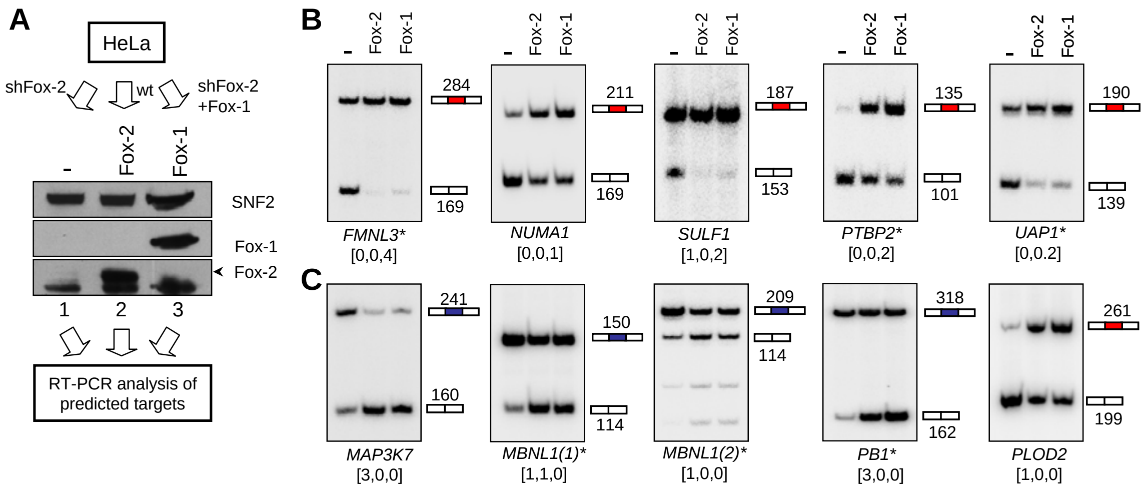


Figure 6: RT-PCR analysis validates predicted targets by Fox-1/2 overexpression and knockdown in HeLa cells.

(**A**) Schematic representation of experimental validation in control or transduced HeLa cells. Control HeLa cells express Fox-2 but not Fox-1. Two other transductant pools without Fox-1/2 expression or with only Fox-1 expressionwere generated by stable retroviral transduction with an shRNA against Fox-2 (shFox-2), or with a combination of shRNA against Fox-2 and stable transfection of Fox-1 cDNA (shFox-2 + Fox-1). The expression of Fox-1 or Fox-2 was confirmed by western blotting analysis using antibodies specific for each protein. Lane 1: HeLa cells with shRNA knockdown of Fox-2; lane 2: untreated HeLa cells; lane 3: HeLa cells with shRNA knockdown of Fox-2 and stable transduction of Fox-1 cDNA. A nonspecific band that crossreacts with the the Fox-2 antibody is indicated by an asterisk.

(**B and C**) Radioactive RT-PCR analysis of predicted Fox-1/2 targets with downstream intronic binding sites (**B**), or with only upstream intronic binding sites (**C**). All examples are cassette exons. For each exon, the gene symbol is shown below, together with the number of conserved Fox-1/2 binding sites in UIF, exonic, and DIF sequences. Exons activated by Fox-1/2 expression are shown in red, whereas exons repressed by Fox-1/2 expression are shown in blue. The size of the PCR products is also labeled. For some of the genes, indicated by an asterisk, the splicing pattern in tissues was also measure by RT-PCR and shown in Fig. 5.

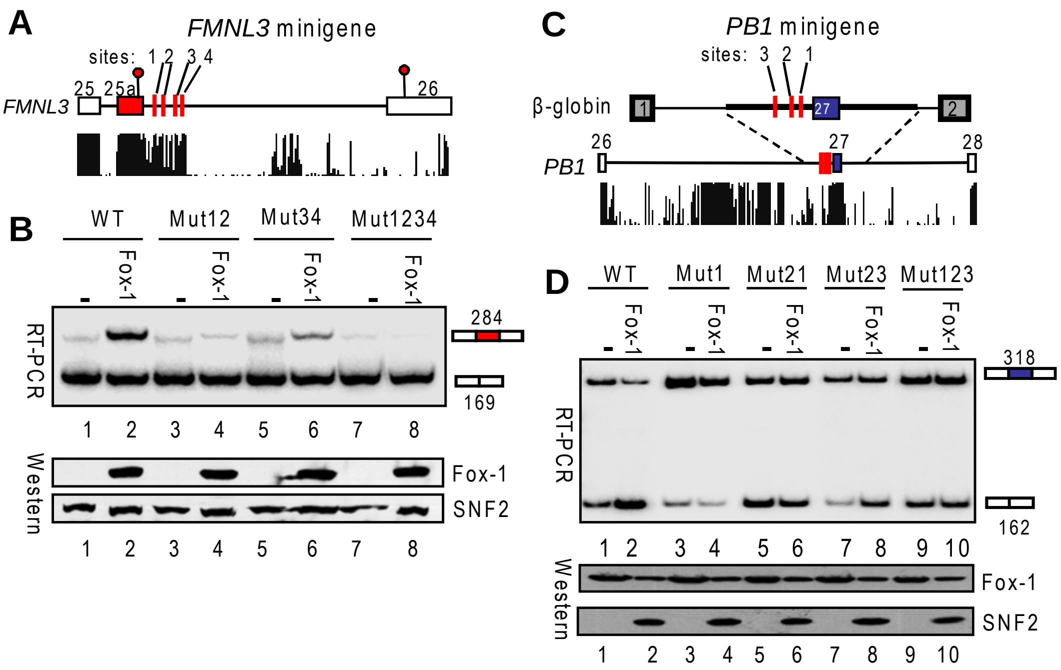


Figure 7: Fox-1/2-mediated splicing regulation depends on the UGCAUG elements.

(**A**) Schematic representation of the *FMNL3* minigene, which has four natural copies of putative Fox-1/2 binding sites (labeled 1 through 4) in DIF sequences. Different usage of stop codons due to alternative splicing is also indicated by red circles.

(**B**) Splicing of the *FMNL3* minigene cassette exon in the wild-type minigene, without or with Fox-1 protein, is shown in lanes 1 and 2, respectively. Lanes 3 to 8 show the splicing of the mutant minigenes. Mut12 (lanes 3 and 4) has mutations in sites 1 and 2, and similarly for Mut34 (lanes 5 and 6) and Mut 1234 (lanes 7 and 8). The expression level of Fox-1 was confirmed by Western blotting, as shown at the bottom.

(**C**) Schematic representation of the *PB1* minigene. The cassette exon, together with ~250 nt of UIF and DIF sequences, including three natural putative Fox-1/2 binding sites in the UIF region, were inserted into intron 1 of the human β-globin gene.

(**D**) Splicing of the *PB1* minigene. See the legend of (**B**) for more details.

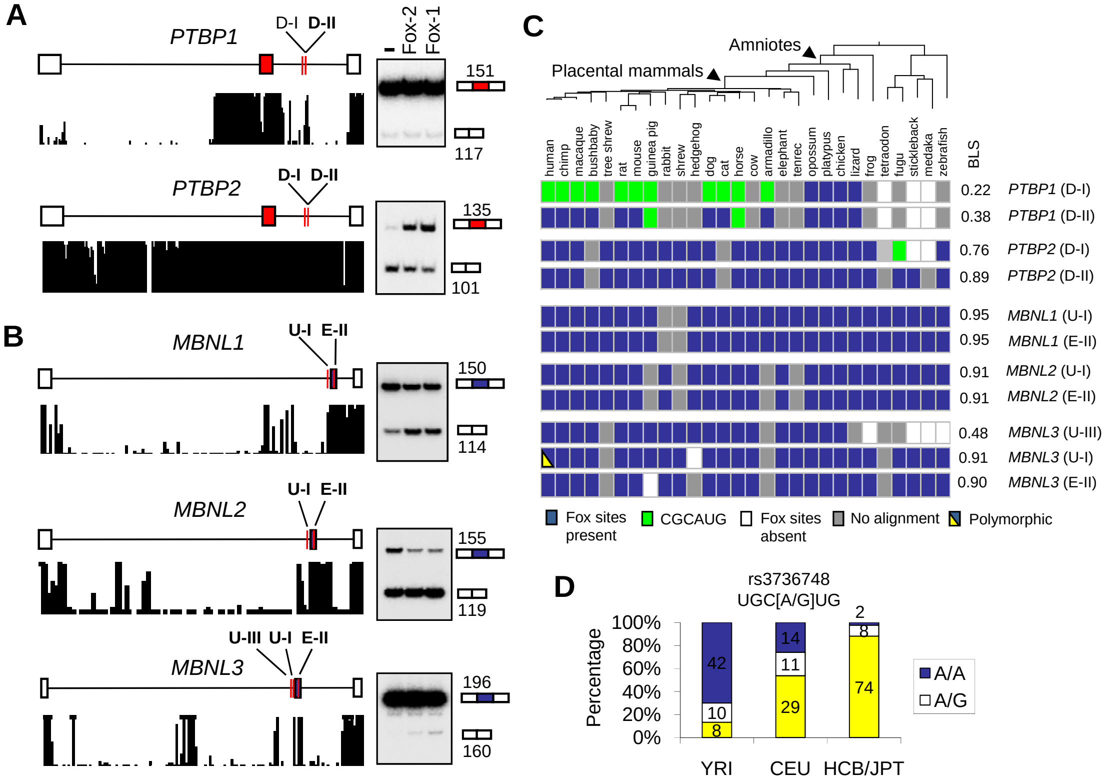


Figure 8: Creation and loss of Fox-1/2 binding sites reflect potential fine-tuning of gene expression after gene duplication.

(**A**) A 34-nt paralogous cassette exon from *PTBP1* (PTB) and *PTBP2* (nPTB). For each gene, the conservation pattern of the region is displayed under the schematic representation. The two downstream conserved putative Fox-1/2 binding sites (D-I and D-II) are labeled. Results of RT-PCR analysis are shown on the right for each exon.

(**B**) A 36-nt cassette exon from *MBNL1*, *MBNL2*, and *MBNL3*, shown similarly as in (**A**). The *MBNL1* and *MBNL2* exons each have two copies of the Fox-1/2 binding site, one in the UIF sequences close to the 3' splice site (U-I) and the other in the exon (E-II). The *MBNL3* exon has an additional site in the UIF sequences (U-III).

(**C**) The presence or absence of Fox-1/2 binding sites in 28 vertebrate species for the sites labeled in (**A** and **B**). The presence of each site in each species is color-coded and shown under the phylogenetic tree. The branch-length score (BLS) for each site is shown on the right. For the *PTB* exon, site D-I appears to be lost in placental mammals by a T-to-C mutation in the first position, resulting in a CGCAUG element, which is shown in green. For the *MBNL3* exon, site U-I is polymorphic in human and overlaps with an A/G SNP (rs3736748) in the fourth position.

(**D**) The allele frequency of the SNP rs3736748 in African Americans (YRI), Europeans (CEU) and Asians (HCB/JPT) was determined according to HapMap data. The A allele (blue) results in an intact Fox-1/2 binding site and the G allele (yellow) results in a disrupted site.

6.9 Supporting information

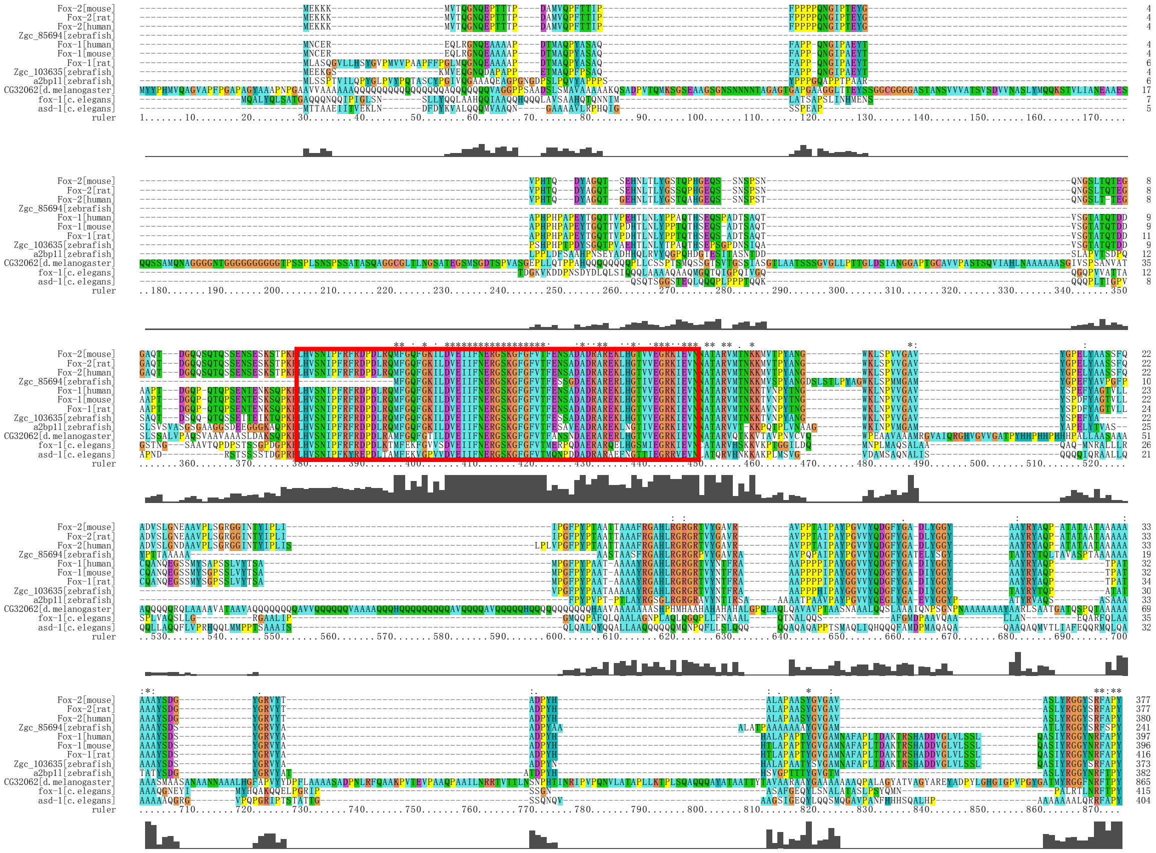


Figure S1: Multiple alignments of Fox-1/2 proteins.

One representative isoform is chosen for each protein. The alignments were obtained by Clustalw. The RRM is highlighted by a red box.

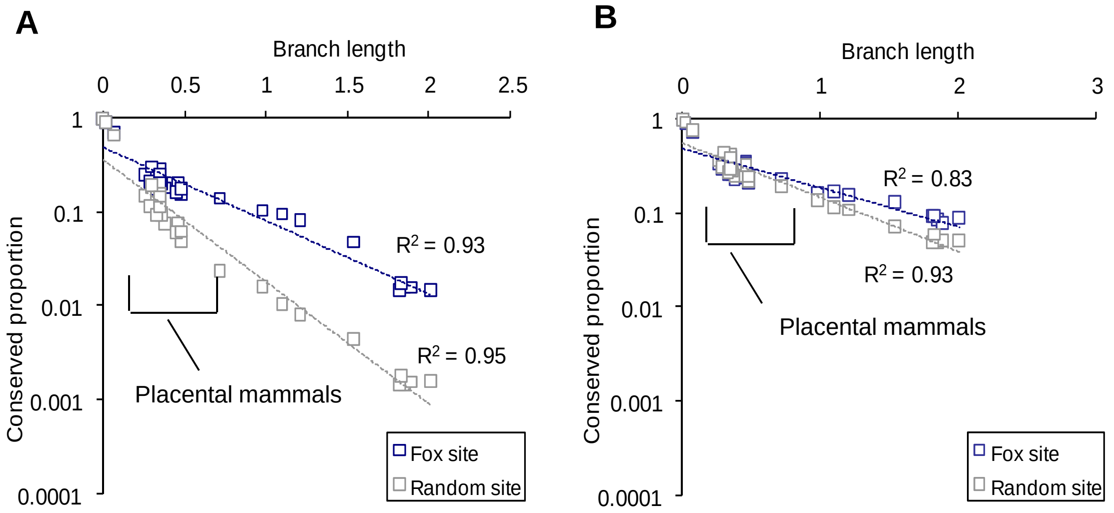


Figure S2: Pairwise conservation of Fox-1/2 binding sites between human and other species.

Exonic sites and intronic sites are treated separately in (A) and (B), respectively. Each blue (gray) point represents the fraction of human Fox-1/2 sites (random motif sites) conserved in each of the other 27 species. The pairwise conservation For both Fox-1/2 binding sites and random sites, the conserved fraction decays exponentially in terms of branch length. , The fitted line (in the log scale) and the squared Pearson correlation coefficient are given. Points corresponding to placental mammals are also labeled.

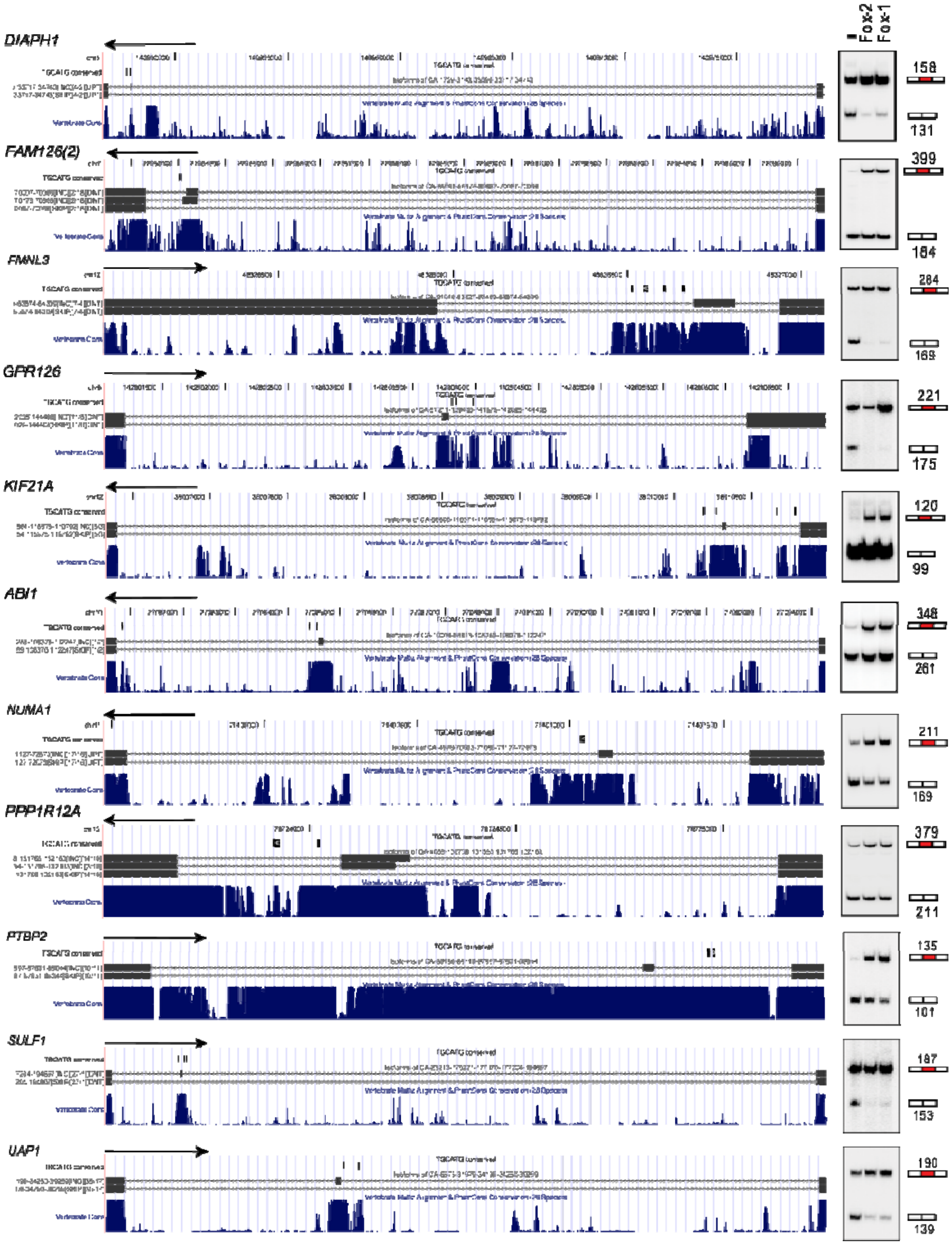


Figure S3: Examples of validated Fox-1/2 targets with conserved binding sites in the introns downstream of the cassette exons.

For each panel, the screenshot from the UCSC Genome Browser is shown on the left. The three tracks from top to bottom are the locations of conserved Fox-1/2 binding sites (for intronic sites, only those in 200-nt from each exon are shown), the inclusion and skipping isoforms of the cassette exon, and the sequence conservation among vertebrate species. The orientation of the gene is indicated by an arrow. In some cases, two cassette exons overlaps and share the same flanking constitutive exons, and they are therefore shown in the same panel. The result of RT-PCR analysis in HeLa cells is shown on the right. Exons activated by Fox-1 or Fox-2 expression are shown in red.

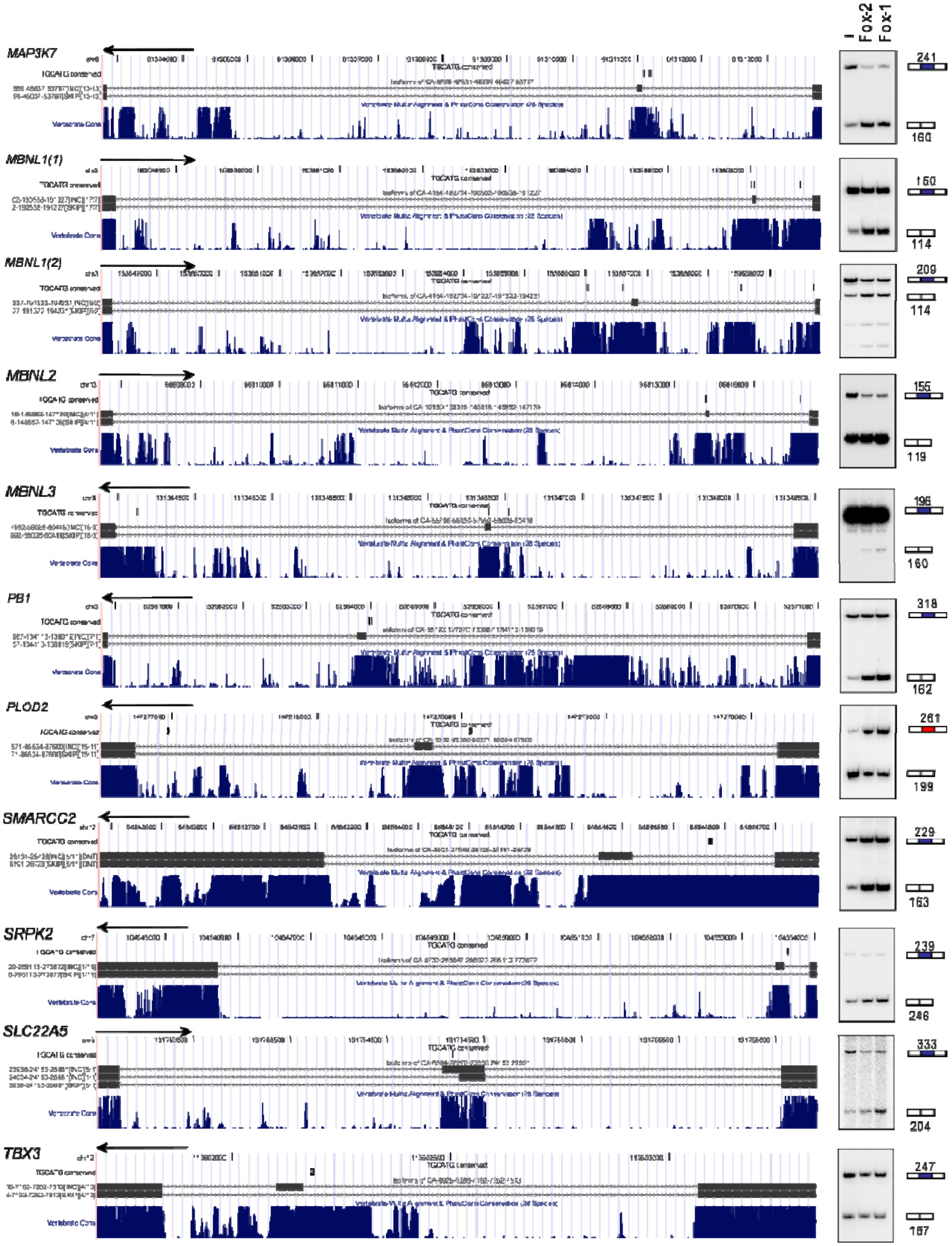


Figure S4: Examples of validated Fox-1/2 targets only with conserved binding sites in the introns upstream of the cassette exons.

Exons activated by Fox-1 or Fox-2 expression are shown in red. An exon (from *PLOD2*) repressed by Fox-1 or Fox-2 expression is shown in blue. See the legend of Fig. S3 for more details.

Chapter 7

Profiling alternatively spliced mRNA isoforms for prostate cancer classification

7.1 Abstract

Prostate cancer is one of the leading causes of cancer illness and death among men in the United States and world wide. There is an urgent need to discover good biomarkers for early clinical diagnosis and treatment. Previously, we developed an exon-junction microarray-based assay and profiled 1532 mRNA splice isoforms from 364 potential prostate cancer related genes in 38 prostate tissues. Here, we investigate the advantage of using splice isoforms, which couple transcriptional and splicing regulation, for cancer classification. As many as 464 splice isoforms from more than 200 genes are differentially regulated in tumors at a false discovery rate (FDR) of 0.05. Remarkably, about 30% of genes have isoforms that are called significant but do not exhibit differential expression at the overall mRNA level. A support vector machine (SVM) classifier trained on 128 signature isoforms can correctly predict 92% of the cases, which

outperforms the classifier using overall mRNA abundance by about 5%. It is also observed that the classification performance can be improved using multivariate variable selection methods, which take correlation among variables into account. These results demonstrate that profiling of splice isoforms is able to provide unique and important information which cannot be detected by conventional microarrays.

7.2 Introduction

Prostate cancer is the second leading cause of cancer illness and death among men in the United States and the third most common cancer world wide (1, 2). According to recent estimates, it accounts for 33% percent of new cancer incidences and six percent of cancer deaths in men world wide (2, 3). In 2002, the number of new incidences and deaths in the United States was approximately 189,000 and 30,200, respectively (2). The difficulty lies, at least partly, in the heterogeneous nature of the disease. Tumor growth is initially dependent on androgen levels, which stimulate cell proliferation and inhibit apoptosis via the androgen receptor (AR) pathway. The prostate-specific antigen (PSA) level has been a standard screening for early diagnosis; androgen ablation is a prevalent therapy to repress the development of androgen-dependent tumors. However, in many cases, this therapy eventually fails and patients die of the recurrent androgen independent prostate cancer (AIPC), a lethal form that progresses and metastasizes (see reviews in refs (4, 5)). Multiple pathways permit cancer cells to escape or bypass the control of the normal AR activation to up-regulate target genes abnormally (6). Although it has been reported that a number of genes are related to these pathways as well as other aspects of prostate cancer, there is still an urgent need for good biomarkers for early clinical diagnosis and treatment.

Microarray technologies developed in the last decade permit monitoring of mRNA abundance levels of tens of thousands of genes in parallel. The accuracy improvement and cost reduction have made them a routine approach in looking for genes that are differentially expressed between normal and tumor samples or between different tumor types/stages (7-14). In a recent study, Segal et al. summarized ~2000 array experiments and derived a panoramic view of activated/deactivated gene expression modules for various types of tumors (15).

Microarrays have also been employed in prostate cancer studies. Using cDNA arrays, Dhanasekaran et al. measured gene expression in 50 normal and neoplastic prostate specimens, as well as three prostate-cancer cell lines, and identified gene signatures characterizing androgen-dependent and AIPC samples (16). Nelson et al. (17) and DePrimo et al. (18) studied gene expression in the androgen treated LNCaP cell line, which was known to be highly androgen responsive. Lapointe et al. profiled 62 primary tumors and 41 normal specimens; three subclasses of tumors representing different tumor stages and risks of recurrence were obtained along with characteristic expression signatures (19). These studies demonstrated the potential of using microarray analyses in characterizing prostate cancer at the gene expression level.

While transcriptional regulation plays important roles within a cell, post-transcriptional regulation, such as alternative splicing, dramatically increases the diversity of the proteome. Alternative splicing also plays a critical role in gene expression regulation and human diseases (20, 21). It has been reported that about 15% of point mutations that cause human genetic diseases can alter splicing patterns (22). In particular, splicing aberrations have been characterized in a number of genes and tumor types (see review by Brinkman (23)).

In a previous work, we developed a microarray-based assay called RASL^{TM} (RNA-mediated Annealing, Selection, and Ligation), which can systematically monitor the abundances of unique splicing events (24). A modified version of the assay, the DASL$^{®}$ (cDNA-mediated Annealing, Selection, extension and Ligation) assay, offers additional robustness for analyzing highly degraded mRNAs, as well as an additional flexibility in probe design (25, 26). Different from other exon-junction arrays (27, 28), the DASL assay achieves high specificity and sensitivity due to the fact that both hybridization and ligation of a pair of oligos complementary to the 5' splice site of the upstream exon and the 3' splice site of the downstream exon are required (see ref (25) for details). In our recent study, this technology was applied to profile the abundances of ~1500 unique splice isoforms in prostate cancer cell lines, tumor specimens and normal control samples (29). This previous study led to two implications: (1) the splicing patterns were altered in a number of genes in response to androgen treatment in the LNCaP cell line; (2) a number of splice isoforms were differentially expressed in tumor samples. They prioritized a list of prostate cancer marker candidates for further investigations. In this study, we extend our previous work and perform a comprehensive analysis of using alternatively spliced isoforms to classify prostate cancer samples. Compared with our previous work, the focus of this study is to quantitatively compare isoform profiling and overall mRNA profiling for cancer classification, which has not been systematically investigated before. To be more specific, the contribution of this study lies in four key aspects: (1) Isoform-sensitive microarrays studies have been assumed to be able to provide more information for cancer classification than conventional microarray studies because isoform abundances couple both transcriptional regulation and splicing regulation. However, it has remained unclear how much unique information could be provided by isoform profiling. In this paper, this assumption is examined qualitatively for the first time through differential expression analysis. Further examinations for several genes are also described. (2) As in a number of other microarray studies (e.g. (16, 19)), hierarchical clustering has been used to segregate similar tissues. This approach was not able to obtain an unbiased estimation of the predictive power for new unknown samples. To assess the predictive power of isoform profiling and that of overall mRNA profiling, a support vector machine with recursive feature elimination (SVM-RFE) was employed to build prediction models and the prediction accuracies were compared. (3) Building a prediction model with a minimal subset of variables is one of the critical tasks in cancer classification. We compared two different variable selection methods for sample classification and examined whether the robustness of prediction can be improved by taking the correlation among isoforms into account during variable selection. (4) In our previous study, two smaller datasets generated in different batches were analyzed separately. The two lists of candidate markers selected from the two datasets had a relatively small overlap. To achieve more robust results, all analyses in this study were based on the larger combined dataset after careful normalizations.

7.3 Results

In our previous work (29), the two datasets of prostate tumors and normal samples were analyzed separately by hierarchical clustering because they were generated in two different batches and there were significant heterogeneities between them (data not shown). In both datasets, splice isoforms could be used to separate tumor samples and normal samples. However, the sample size in each dataset was limited and the overlap between the two lists of differentially expressed isoforms selected from the two datasets was relatively small. In this paper, the two datasets were combined after careful normalizations to achieve more robust results and statistical power (see Methods). The combined datasets included 22 cases of prostate tumors and 16 matched normal samples.

Splice isoforms reveal distinct signatures of prostate cancer

We first examined whether the global distinction between tumors and normal samples still exists in the combined dataset by unsupervised methods. As expected, tumors can be readily separated from normal samples by average-linkage hierarchical clustering (Figure 1 A and B, cluster C1 and C2) (30). Compared with cluster C2, the majority of tissues in cluster C1 are normal prostate and stroma, with the average tumor percentage being 8.2% ($p<0.0001$), and stromal percentage being 63.4% ($p<0.0001$). Of the three tumors segregated with normal samples in cluster C1, two have low tumor content. Additional analysis reveals that C2 cases in general have a significantly higher percentage of more advanced stages (Stage 3 or above) and more patients die of prostate cancer compared to C1 cases. Specifically, 100% of the cases in C1 were from patients with organ confined tumors (stage T2), whereas 50% of the cases in C2 were from metastasized patients (stage T3 tumors, $p<0.001$). At the time of analysis, none of the C1 patients died of prostate cancer while14% of the C2 patients died of prostate cancer. Interestingly, the cluster C2 enriched by tumors was further segregated into two sub-clusters, reflecting different percentage in tumor and stromal content (Mean tumor content in sub-cluster C2.1=47.9% v.s. C2.2=64.5%, $p=0.1$; Mean stromal content in C2.1=35.8% v.s. C2.2=20.5, $p=0.04$).

Singular value decomposition (SVD) was used to identify an orthogonal low dimensional space which preserves the maximal variation of the original high dimensional space. The first two principal components capture 17% and 9% of the total variation, respectively (Figure 1F). Remarkably, the first principal component alone shows a strong separation of tumor and normal samples. The clusters and sub-clusters derived from hierarchical clustering are also reflected in the 3D space spanned by the first three principal components (Figure 1G), which confirms the results of clustering.

Further examination of the gene clustering results shows distinct molecular signatures of different tissue clusters, including both well known marker genes and less studied marker candidates (Figure 1 C, D and E). Figure 1C shows isoforms up-regulated in cluster tumor sub-cluster C2.2, including isoforms from genes RPS2, XBP1, U1AF1 and ATP5A1, all of which were known to be up-regulated in tumors. Figure 1D shows isoforms down-regulated in normal tissues and up-regulated in tumor tissues, including isoforms from genes U2AF2, CLN3 and HPN. Figure 1E shows isoforms with high expression levels in normal tissues and down-regulated in tumor tissues, especially in sub-cluster C2.2. Several genes in this cluster are known to be involved in the TGF-beta signaling pathway, such as TGFB2, LTBP4 and TGFBR3.

Differentially expressed splice isoforms

A two sided t-test was used to identify genes with statistically significant changes in expression between tumors and normal samples. A false discovery rate (FDR) or q-value was calculated as described previously (31), to correct for multiple testing. As a result, 464 isoforms (30%) representing 222 genes (61%) are reported as being significant (q-value < 0.05) [see Additional file 1]. The high proportion of differentially expressed isoforms reflects the fact that the genes profiled are potentially related to prostate cancer according to existing evidence. Top isoforms among them include AMACR-2094, FGFR2-0101, FGFR2-0097, FGFR2-0098, CLU-0192, PGR-1162, etc.

Profiling of splice isoforms provides additional information to overall mRNA abundances

In theory, profiling individual splice isoforms can provide more information than profiling overall mRNA levels as in conventional microarrays. This is because isoform profiling detects the combinatorial effects of both transcriptional regulation and splicing regulation. Consider the simplest case of a gene with two alternatively spliced isoforms. If one isoform is up-regulated in tumors whereas the other is down-regulated, the overall mRNA abundance may not change. On the contrary, if the overall mRNA level is differentially expressed, there is at least one isoform exhibiting differential expression. However, how much additional information can be obtained for cancer classification by isoform profiling has not been systematically evaluated. To address this question, we compared individual isoforms and overall mRNAs for differential expression.

Due to the costs and array capacity, the original array design did not include probes targeting common regions of all isoforms. Therefore, the overall mRNA expression level can not be obtained directly. However, since the probed exon junctions target unique major isoforms and hybridization efficiencies of different probes are comparable (25), we reason that the overall expression level can be estimated by summing up the abundances of individual isoforms. To examine the validity of this idea, two well-known prostate cancer cell lines LnCaP and PC-3 were profiled using the same DASL assay (splicing array). For comparison, 107 genes were arbitrarily selected for gene expression profiling in the same cell lines (expression array). An independent oligo pool targeting common regions of all isoforms in each of the 107 genes were used in the expression array. Therefore, the log expression ratio of each gene in the two cell lines can be obtained from the estimation based on the splicing array and from the direct measurement in the expression array independently. To our satisfaction, the two quantities are highly correlated ($R^2 = 0.80$, p=2.2e-16), suggesting a reasonable accuracy of the estimation (Figure 2A).

Having validated the approach, the overall mRNA abundances of each gene in prostate tissues were estimated. A t-test was similarly applied to identify genes with significant differential expression in tumors at the overall mRNA level. In total, 159 genes (43.6%) are reported as being significant (q-value < 0.05). Again, the high proportion of significant genes reflects the fact that they are potentially relevant to prostate cancer according to previous studies. Strikingly, more genes are called significant by examining individual isoforms than by examining overall mRNAs (222 vs 159, p=0.001, chi-square test). Among the 159 genes that are called significant, 150 genes (94%) have at least one isoform that is reported as significant (Figure 2B). In contrast, only 68% of genes with significant isoforms can be detected at the overall mRNA level. The remaining 32% of the genes have significant isoforms but do not exhibit significant differential expression at the overall mRNA level. It is important to note that these genes represent the unique information that is provided by splice isoform sensitive microarrays and cannot be obtained from conventional microarrays.

From the perspective of isoforms, 78% of significant isoforms are from those genes that are also called significant whereas 22% of significant isoforms are from those genes that do not show overall mRNA differential expression (Figure 2D) [see Additional file 2 and 3]. Multiple testing has been appropriately accounted for, so the additional significant calls using splice isoforms are not due to the different stringencies of thresholds, but reflect additional information provided by including splicing regulation.

For many genes, only one isoform is specifically altered in tumors. In these cases, the addition of other isoforms to the total mRNA level simply introduces random noise. Notably, there are 14 genes with one isoform being up-regulated in tumors and another isoform being down-regulated. Among them, 3 genes are not significant at the overall mRNA level: CD44 (CD44-1404 vs CD44-1570), ITGB1 (ITGB1-0032 vs ITGB1-0033) and MAPT (MAPT-1060 vs MAPT-1061). CD44 is a multifunctional receptor involved in cell-cell interactions and cell trafficking. Deregulated expression of a number of variants is correlated with tumor metastasis (reviewed by (23)). ITGB1 is a protein involved in extra-cellular matrix interactions and is also related to many tumor types, including prostate cancer (22).

There are relatively fewer studies discussing the role of MAPT in cancer. MAPT encodes the microtubule-associated protein tau mainly expressed in the central nervous system. Mutations in the MAPT gene disrupt the normal binding of tau to tubulin. This in turn results in pathological deposits of hyperphosphorylated tau in the brain, which is a pathological hallmark of several neurodegenerative disorders (see review by Rademakers et al. (32)). Previously, Sangrajrang et al. found that MAPT was also expressed in the DU145 cell line using RT-PCR and the expression at the protein level was validated by Western blotting (33). The expression was elevated after estramustine treatment and the authors suggested that the protein may be positively related to drug resistance. This was consistent with a recent report demonstrating that the up-regulation of the protein tau was correlated to the decrease of paclitaxel sensitivity in breast cancer (34). In our data, MAPT-1060 (representing the skipping of exon 4A, numbered according to ref (32)) has a two fold increase in tumors relative to normal tissues(q-value=0.86%), whereas

MAPT-1061 (representing the inclusion of exon 4A) has a two fold decrease in tumors relative to normal tissues (q-value=0.16%). It is likely that exon 4A is uniquely skipped in prostate cancer cells. This hypothesis is further supported by the following evidence. Exon 4A harbors a C/T single nucleotide polymorphism (SNP) near the 5' splice site (Entrez SNP: rs17651549, contig position: 2715394). This SNP was assayed from 71 individuals and the C/T ratio is 0.886/0.114. In the major C allele, a putative exonic splicing enhancer (ESE) *cagccgg* encompassing the SNP is predicted by ESEfinder and resembles the specific RNA binding site of SF2/ASF, a critical serine rich (SR) protein that helps to recruit the splicing apparatus (score: 4.6, threshold: 1.956) (35). This putative ESE is disrupted in the minor T allele for all four SR proteins in ESEfinder including SF2/ASF, SC35, SRp40 and SRp55. However, further experimental studies and confirmation of the splicing alteration may be required to validate this hypothesis.

Profiling of splice isoforms improves predictive power

A robust prediction model to classify unknown samples is essential for early cancer detection and diagnosis. Having demonstrated that a large fraction of genes show differential expression at the splice isoform level but not at the overall mRNA level, a key question is how much additional predictive power can be achieved by isoform profiling. Another related problem is to select minimal subsets of variables with the best performance. Like many other types of tumors, a single molecular marker is usually not robust enough for prostate cancer detection, as is the case for the widely used PSA level for early stage screening. At the other extreme, including all variables from a genome-wide profiling is not justifiable either, due to the noise introduced by a huge number of uninformative variables and the difficulty in the interpretation of the resulting model.

A support vector machine (SVM) was used here to build the classifier because of its excellent performance in many previous studies with small sample sizes (36). An recursive feature elimination (RFE) algorithm was integrated as described previously with minor adaptations (37).

Leave-one-out cross validation (LOOCV) with external variable selection was used to give an unbiased evaluation of the prediction accuracy (see Methods for details). SVM-classifiers were built using the individual splice isoforms and estimated overall mRNA abundances. The results of LOOCV are shown in Figure 3A. For the classifiers using isoform abundances, the best performance, 35 correct predictions out of 38 samples (92%), is achieved when 128 isoforms are included for classification. For the classifiers using overall mRNA abundances, the best performance (87% correct predictions) is achieved when 32 genes are used. The additional information provided by splicing regulation gives rise to an improvement of about 5% in predictive power. Importantly, the difference persists in the whole range of different sizes of selected variable subsets, which is unlikely by random chance. With an independent method, this demonstrates that isoform profiling can provide valuable information for cancer classification. Also, the classification performance deteriorates when the subset of selected variables is too small in size (e.g., 4 variables). This is consistent with the previous observation that a robust cancer prediction model should use a reasonable number of molecular signatures (38).

Comparison of different variable selection methods

Both t-tests and SVM-RFE can generate lists of candidate markers. These two approaches represent univariate variable selection and multivariate variable selection, respectively. They have different assumptions and may characterize different yet overlapping perspectives of the molecular mechanisms underlying the data. For example, variables are assumed to be independent in a t-test but there is no assumption of independence in SVM-RFE. Comparing the multiple outputs of selected signatures by different methods may shed further insights into the data and the methods. Therefore, the two different variable selection approaches, t-test and SVM-RFE, were applied to select marker candidates and their performances in building linear SVM models were compared. The results of LOOCV are shown in Figure 3B. The best performance of t-test selection is achieved with a similar number of variables as SVM-RFE. Both methods result in an accuracy of 92%. The similar best performance by t-test and SVM-RFE is likely due to the distinct features of tumors and normal tissues. The information to classify the two groups is largely redundant. However, the curve of prediction accuracy by the SVM-RFE selection is smoother than that by the t-test selection as the size of selected variable subset decreases. This smaller variation suggests that SVM-RFE is more robust than t-test in variable selection for cancer classification.

The 128 isoforms selected by t-test (t-test128 list) and the 128 isoforms selected by SVM-RFE (svm128 list) share 42 isoforms (Table 2). The common list includes AMACR-2094, AMACR-2097, AMACR-2098, FGFR2-0099, FGFR2-0094, PGR-1166 and PGR-1555 among others. They may represent robust marker candidates. Significant isoforms in each list were further divided into two groups according to whether the corresponding genes also exhibit significant differential expression at the overall mRNA level. Interestingly, among those 86 isoforms included only in the svm128 list, 13 of the isoforms are in the category that the corresponding genes do not show significant differential expression at the overall mRNA level. In contrast, among the 86 isoforms included only in the t-test128 list, only 4 isoforms lie in this category. Therefore, SVM-RFE captures more information uniquely provided by considering splicing regulation (p=0.03, chi-square test). This demonstrates the advantage of a variable selection method taking the correlation between variables into account.

7.4 Discussion

The diagnosis and treatment of prostate cancer are fields with long histories. Various efforts have led to the progressive understanding of the disease. However, the present criteria of diagnosis and prognosis, as well as the approaches of treatment and surgery, are not sufficiently reliable. Previous gene expression profiling studies on prostate tumors and normal tissues demonstrated the feasibility in characterizing the molecular alterations at the overall mRNA transcript level. However, these transcriptome analyses were based on the old central dogma of “one gene, one mRNA”, which may underestimate the complexity of tumorigenesis (23).

Previously, we carried out a study of prostate cancer by exon-junction microarray-based assay and demonstrated the power of this integrated technology in detecting both transcriptional and splicing regulation (25, 29). In this paper, we present systematic analyses with the focus on using splice isoform profiling for prostate cancer classification. Isoform-sensitive microarrays have been used in several recent studies (24, 25, 27, 29, 39-43) (also see review by Lee and Roy (44)). These studies demonstrated that isoform-sensitive microarray is a reliable, high throughput approach to detecting splicing alterations in various tissues and conditions. Although more and more data are expected to be generated in the near future, the dataset used in this study is the only dataset currently available which screened a relatively large sample of cancer and normal tissues. As far as we know, this is the first systematic comparison of isoform-sensitive microarrays and conventional microarrays for cancer classification.

Previous studies have used a “splice index”, which is the fraction of each isoform, to remove the effect of transcriptional regulation (39, 40). This is not desired for cancer classification because as much information as possible should be incorporated. Therefore the abundance of each isoform, which couples both transcriptional regulation and splicing regulation, was used for classification. The performance was compared with that of using overall mRNA abundances. One has to note a caveat of the current DASL assay: it does not include probes complementary to the common regions of all mRNA transcripts for each gene due to the current limit in array capacity. Therefore, the overall mRNA level was estimated indirectly by summing up all the isoforms targeted. The estimation is not ideal due to the fact that not all isoforms were included in the array and the probes target splicing events that are not mutually exclusive in several cases. However, the estimation is reasonably good and highly correlated with the direct measurement by an expression array. Various other methods were tried to estimate the overall mRNA abundances, but the method used here is the most accurate and simplest.

Among the ~1500 isoforms from putative prostate cancer-related genes, a large fraction of them exhibit differential expression in cancer cells. Tumors and normal tissues can be readily separated by both unsupervised and supervised methods. By comparing individual isoforms and overall mRNAs for differential expression, we arrived at the conclusion that an isoform-sensitive microarray, which detects coupled transcription and splicing regulation, can provide about 30% more information than conventional microarrays. This value may still be underestimated due to the following reasons. The current DASL assay included only 364 genes potentially relevant with prostate cancer derived from previous studies. Till now, a large body of literature, especially those in the genomic scale, focused more on transcriptional regulation. Therefore, the selection of genes may be biased to those exhibiting aberrant transcriptional regulation.

The optimal prediction model was built by SVM with variable selection integrated, a powerful machine learning approach. With around 100 isoforms, the best classification performance can be achieved at a correct prediction rate of 92%. Compared with the optimal SVM classifier built with overall mRNA abundances, this represents an improvement of five percent. Therefore, both differential expression analysis and classification analysis quantitatively demonstrated the advantage of isoform-sensitive microarrays.

We also compared the effect of different variable selection approaches on classification performance. By taking the correlation between isoforms into account, isoforms selected by SVM-RFE are more robust for classification than isoforms selected by a t-test. Although univariate two-sample comparisons such as t-test are widely used to identify differentially expressed genes, the assumption of independence between genes or isoforms is not biologically justifiable. In cancer signal transduction pathways, a group of genes in the same pathway are interacting with each other; cross-talks often exist between pathways as well (C Jiang, personal communication). Variables are more convoluted in the DASL data due to the coupling of transcription and splicing. The multi-loci nature of the disease also makes it difficult to use a single or few molecular markers to build a sufficiently robust prediction model.

This study identified a number of known prostate cancer markers as well as less studied marker candidates, which span a wide spectrum of biological functional roles. Some are related to signal transduction (SIM2 and CDC42BPA), as well as extracellular matrix and cytoskeleton (CD44, MAPT and ILK). Others appear to be involved in epidermal differentiation and proliferation (KRT15, IGF1, PGR and HPN), cell growth and development (FGFR2), apoptosis (DBCCR1 and CLU), lipid metabolism (AMACR), etc. Very significantly, multiple isoforms from AMACR, a key player in catalyzing the isomerization of alpha-methyl-branched fatty acid and a recently reported good prostate cancer marker, show the strongest signal in our data (45). Several genes encoding splicing factors, such as U2AF1, U2AF2 and DHX34, also show significant differential expression. This is consistent with our observation that a large fraction of splicing factors are deregulated in tumors (C. Zhang et al, unpublished data).

Another interesting observation obtained by examining the panel of potential marker candidates selected by one or more methods is that a number of genes are normally expressed specifically in neuronal cells (such as MAPT, STAC, NELL2, etc). The relationship between abnormal expression of neuronal genes and tumors is not completely clear. However, it is believed that there is a link between diverse neurodegenerative diseases and cancers via the induction of antitumor immunity, known as paraneoplastic neurological degenerations (PND) (see review by Albert and Darnell (46)). Alternative splicing is also prevalent for neuronal genes.

7.5. Conclusions

Profiling of individual isoforms can provide unique and important additional insights into prostate cancer classification. Robust prediction models can be built with a subset of isoforms selected by multivariate variable selection method.

7.6 Methods

DASL assay

The DASL assay and array hybridization were described previously (25). In contrast to conventional microarrays which only measure the overall mRNA abundance of each gene, the most distinguishing feature of the DASL assay is that it permits the profiling of each individual mRNA splice isoform quantitatively. This technology has been shown to be highly sensitive, specific and reproducible ($R^2 > 0.99$ between replicates).

Tumor and normal tissue profiling

The array used in this study included 1532 isoforms from 364 genes. These genes, potentially related to prostate cancer, were selected from published literature, previous microarray data analysis, human genome anatomy projects and EST searching. All of them have known gene structures and alternative splicing patterns. Alternatively spliced exon junctions probed in the array were obtained by the alignment of mRNA transcripts/ESTs and the genome. They were manually annotated and are publicly available from the MAASE database (47, 48). In total, 22 cases of archived formalin fixed, paraffin embedded prostate tumors at different tumor stages and 16 adjacent normal matching samples from the UCSD prostate tumor bank were assayed, each with two replicates (Table 1). The detailed information about sample collection, preparation, RNA profiling experiment and probe quantification were described elsewhere (29). The raw data is available from the authors upon request.

Microarray data normalization and statistical analysis

Before further analysis, a $\log_2$ transformation was applied to raw intensities. Since the dataset was generated in two batches, heterogeneity between batches has to be removed. As a first step, each isoform (row) inside each batch was median-centered separately. Then, the two batches were combined and standardized to unit variance across each array (column) and isoform (row) as a whole. Finally, the two replicates of each tissue sample were averaged. In this way, each value in the data matrix represents the log expression ratio of an isoform in a particular sample with respect to a "common control" (15). The effect of normalization was examined by clustering the combined data using real expression values and null control probes, respectively. After normalization, there is no visible artificial distinction between the two batches.

To estimate the overall mRNA abundance of each gene, the intensities of all isoforms were summed. Then the same log transformation and normalization steps above were applied. Again, each normalized value represents the log expression ratio of mRNA abundance in a particular sample with respect to a "common control".

A two-sided t-test was used to select isoforms or genes with significant differential expression between tumors and normal tissues. To correct for the effect of multiple testing, false discovery rate (FDR) or q-value was calculated as described previously (31). A chi-square test was used to analyze the significance of frequency data.

Singular value decomposition

Singular value decomposition (SVD) is a standard mathematical transformation to find a set of orthogonal principal components (PCs) which explain as much variation as possible (49). The power of SVD has been shown in many fields as well as in microarray data analysis. Alter et al. and Holter et al. suggested that the first two PCs can characterize cell cycle phases of yeast genes(50, 51) . Liu et al. separated prostate and colon tumors from others with the first PC alone(52). In a similar spirit, SVD transformation was used in this study to reveal the “hidden” information underlying the original high dimensional dataset.

SVM-RFE

A linear support vector machine (SVM) optimizes a linear classifier $D(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i + b$ by maximizing the margin of support vectors from two classes, where $\mathbf{x}_i$ is the expression vector of a sample i and $\mathbf{w}$ is the vector of weighting coefficient, reflecting the contribution of each variable in classification (36). In the past few years, SVM has been developed and shown as a powerful tool for classification problems with a small sample size, such as microarray sample classification (e.g. ref (7)). SVM-RFE (RFE stands for recursive feature elimination) is a wrapper approach of variable selection, in which the predictive power of a subset of variables is measured collectively by the accuracy of the classification based on the subset in consideration (37, 53). Since an exhaustive search of the optimal subset is a combinatorial problem, a heuristic strategy must be applied. In SVM-RFE, variables are ranked by the weighting vector $\mathbf{w}$, by which a subset of variables with top ranks is selected. Then the weighting vector $\mathbf{w}$ is re-evaluated by optimizing a new classifier with the selected subset and a smaller subset is selected therein. This recursive procedure continues until the subset is small enough or the classification performance approaches some criteria. In this way, informative variables for classification are recursively selected (or uninformative variables are recursively eliminated). Details of the algorithm can be found in ref (37). Our implementation of SVM-RFE used SVMTorch for linear SVM model calculations (54). The default soft margin (C=100) was used.

Cross validation incorporating variable selection

Due to the limited sample size, leave-one-out cross validation (LOOCV) was used to evaluate the classification performance of SVM classifiers built with subsets of variables selected by t-test and SVM-RFE. In each round, one array (test set) is left out to test the classifier trained on the remaining arrays (training set). The classification performance is the percentage of correct predictions in all rounds. To get an unbiased result, in each round the variable selection step must be applied "externally", i.e. only on the training set, excluding the sample left out for validation (38). Therefore, the subsets of variables selected might be different from round to round. The number of times that a variable is selected reflects the robustness of the variable for classification. Therefore the final subset of variables can be selected by ordering the number of times that a variable is included in the selected subsets of all rounds.

7.7 Acknowledgements

We would like to thank Joanne Yeakley and Marina Bibikova for help generating the array data used in this study. We thank Dr. Jinhua Wang for helpful discussions during the project; we also thank Drs. Michael Wigler, Dustin Schones and Vladimir Jurukovski for critical reading of the manuscript. We would also like to thank anonymous reviews for helpful comments. This work was supported by grants from NIH to X.-D.F and M.Q.Z.

7.8 References

1. Parkin, D. M., Bray, F. I., & Devesa, S. S. (2001) Cancer burden in the year 2000. The global picture *Eur. J. Cancer* **37,** 4-66.

2. Jemal, A., Thomas, A., Murray, T., & Thun, M. (2002) Cancer statistics, 2002 *CA Cancer J Clin* **52,** 23-47.

3. Jemal, A., Murray, T., Samuels, A., Ghafoor, A., Ward, E., & Thun, M. J. (2003) Cancer statistics, 2003 *CA Cancer J Clin* **53,** 5-26.

4. Denmeade, S. R. & Isaacs, J. T. (2002) A history of prostate cancer treatment *Nat. Rev. Cancer* **2,** 389 -396.

5. Nelson, W. G., De Marzo, A. M., & Isaacs, W. B. (2003) Prostate Cancer *N Engl J Med* **349,** 366-381.

6. Feldman, B. J. & Feldman, D. (2001) The development of androgen-independent prostate cancer *Nat. Rev. Cancer* **1,** 34-45.

7. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P.*, et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures *Proc Natl Acad Sci USA* **98,** 15149-15154.

8. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S.*, et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications *Proc Natl Acad Sci USA* **98,** 10869-10874.

9. Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., & Patel, A. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling *Cancer Cell* **1,** 133-143.

10. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X.*, et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling *Nature* **403,** 503-511.

11. Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G.*, et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma *Nat. Med.* **8,** 816-824.

12. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M.*, et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses *Proc Natl Acad Sci USA* **98,** 13790-13795.

13. Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I.*, et al.* (2001)

Diversity of gene expression in adenocarcinoma of the lung *Proc Natl Acad Sci USA* **98,** 13784-13789.

14. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring *Science* **286,** 531-537.

15. Segal, E., Friedman, N., Koller, D., & Regev, A. (2004) A module map showing conditional activity of expression modules in cancer *Nat Genet* **36,** 1090-1098.

16. Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., & Chinnaiyan, A. M. (2001) Delineation of prognostic biomarkers in prostate cancer *Nature* **412,** 822-826.

17. Nelson, P. S., Clegg, N., Arnold, H., Ferguson, C., Bonham, M., White, J., Hood, L., & Lin, B. (2002) The program of androgen-responsive genes in neoplastic prostate epithelium *Proc Natl Acad Sci USA* **99,** 11890-11895.

18. DePrimo, S., Diehn, M., Nelson, J., Reiter, R., Matese, J., Fero, M., Tibshirani, R., Brown, P., & Brooks, J. (2002) Transcriptional programs activated by exposure of human prostate cancer cells to androgen *Genome Biol.* **3,** research0032.0031 - research0032.0012.

19. Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer *Proc Natl Acad Sci USA* **101,** 811-816.

20. Kan, Z., Rouchka, E. C., Gish, W. R., & States, D. J. (2001) Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs *Genome Res.* **11,** 889-900.

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

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

23. Brinkman, B. M. (2004) Splice variants as cancer biomarkers *Clin Biochem* **37,** 584-594.

24. Yeakley, J. M., Fan, J. B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S., & Fu, X. D. (2002) Profiling alternative splicing on fiber-optic arrays *Nat. Biotechnol.* **20,** 353-358.

25. Fan, J.-B., Yeakley, J. M., Bibikova, M., Chudin, E., Wickham, E., Chen, J., Doucet, D., Rigault, P., Zhang, B., Shen, R., *et al.* (2004) A Versatile Assay for High-Throughput Gene Expression Profiling on Universal Array Matrices *Genome Res.* **14,** 878-885.

26. Bibikova, M., Talantov, D., Chudin, E., Yeakley, J. M., Chen, J., Doucet, D., Wickham, E., Atkins, D., Barker, D., Chee, M., *et al.* (2004) Quantitative Gene Expression Profiling in Formalin-Fixed, Paraffin-Embedded Tissues Using Universal Bead Arrays *Am J Pathol* **165,** 1799-1807.

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

28. Clark, T. A., Sugnet, C. W., & Ares, M., Jr. (2002) Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays *Science* **296,** 907-910.

29. Li, H.-R., Wang-Rodriguez, J., Nair, T. M., Yeakley, J. M., Kwon, Y.-S., Bibikova, M., Zheng, C., Zhou, L., Zhang, K., Downs, T., *et al.* (2006) Two-dimensional Transcriptome Profiling: Identification of mRNA Isoform Signatures in Prostate Cancer from Archived Paraffin-embedded Cancer Specimens *Cancer Res* **in press**.

30. Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns *Proc Natl Acad Sci USA* **95,** 14863-14868.

31. Storey, J. D. & Tibshirani, R. (2003) Statistical significance for genomewide studies *Proc. Natl. Acad. Sci. U.S.A.* **100,** 9440-9445.

32. Rademakers, R., Cruts, M., & van Broechkoven, C. (2004) The role of tau (MAPT) in frontotemporal dementia and related tauopathies *Hum Mutat* **24,** 277-295.

33. Sangrajrang, S., Denoulet, P., Millot, G., Tatoud, R., Podgorniak, M. P., Tew, K. D., Calvo, F., & Fellous, A. (1998) Estramustine resistance correlates with tau over-expression in human prostatic carcinoma cells *Int J Cancer* **77,** 626-631.

34. Rouzier, R., Rajan, R., Wagner, P., Hess, K. R., Gold, D. L., Stec, J., Ayers, M., Ross, J. S., Zhang, P., Buchholz, T. A., *et al.* (2005) Microtubule-associated protein tau: A marker of paclitaxel sensitivity in breast cancer *Proc Natl Acad Sci USA* **102,** 8315-8320.

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

36. Vapnik, V. (1999) *The nature of statistical learning theory* (Springer-Verlag, New York).

37. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002) Gene selection for cancer classification using support vector machines *Machine Learning* **46,** 389-422.

38. Ambroise, C. & McLachlan, G. J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data *Proc. Natl. Acad. Sci. U.S.A.* **99,** 6562-6566.

39. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse *Nat Genet* **37,** 844-852.

40. Sugnet, C. W., Srinivasan, K., Clark, T. A., Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays *PLoS Computational Biology* **2,** e4.

41. Relogio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R. B., & Valcarcel, J. (2005) Alternative Splicing Microarrays Reveal Functional Expression of Neuron-specific Regulators in Hodgkin Lymphoma Cells *J. Biol. Chem.* **280,** 4779-4784.

42. Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., Morris, Q. D., *et al.* (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform *Molecular Cell* **16,** 929-941.

43. Fehlbaum, P., Guihal, C., Bracco, L., & Cochet, O. (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants *Nucl. Acids Res.* **33,** e47-.

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

45. Luo, J., Zha, S., Gage, W. R., Dunn, T. A., Hicks, J. L., Bennett, C. J., Ewing, C. M., Platz, E. A., Ferdinandusse, S., Wanders, R. J., *et al.* (2002) {alpha}-Methylacyl-CoA Racemase: A New Molecular Marker for Prostate Cancer *Cancer Res* **62,** 2220-2226.

46. Albert, M. L. & Darnell, R. B. (2004) Paraneoplastic neurological degenerations: keys to tumour immunity *Nat Rev Cancer* **4,** 36-44.

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

48. Zheng, C. L., Kwon, Y.-S., Li, H.-R., Zhang, K. U. I., Coutinho-Mansfield, G., Yang, C., Nair, T. M., Gribskov, M., & Fu, X.-D. (2005) MAASE: An alternative splicing database designed for supporting splicing microarray applications *RNA***,** rna.2650905.

49. Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computation* (Johns Hopkins Univ. Press, Baltimore).

50. Alter, O., Brown, P. O., & Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling *Proc Natl Acad Sci USA* **97,** 10101-10106.

51. Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., & Fedoroff, N. V. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity *Proc Natl Acad Sci USA* **97,** 8409-8414.

52. Liu, L., Hawkins, D. M., Ghosh, S., & Young, S. S. (2003) Robust singular value decomposition analysis of microarray data *Proc Natl Acad Sci USA* **100,** 13167-13172.

53. Xiong, M., Fang, X., & Zhao, J. (2001) Biomarker Identification by Feature Wrappers *Genome Res.* **11,** 1878-1887.

54. Collobert, R. & Bengio, S. (2001) SVMTorch: Support Vector Machines for Large-Scale Regression Problems *J Machine Learning Res* **1,** 143-160.

7.9 Tables and Figures

Table 1 Pathological information of tumor and normal prostate samples

ID	Age	Risk group	% tumor	BPH	Atrophy	Stroma	Inflam	PSA	Gleason	Stage
T5	67	low	50	0	0	20	0	8.48	5+4=9	T3bN1Mx
T21	74	Low	60	10	10	20	0	6.7	4+4=8	T2bNxMx
N22	74	Low	0	10	40	50	0	6.7		T2bNxMx
N30	55	Int	0	10	30	68	0	11.68		T2bN1Mx
N44	61	low	0	10	2	88	0	5.46		T2cNxMx
N46	74	High	0	45	20	35	0	8.06		T2aNxMx
N56	67	High	0	5	0	94	0	5.7		T2aN0Mx
T72	68	Int	70	0	0	30	0	8.27	4+3=7	T3bN1Mx
N77	66	Int	0	0	10	89	1	3.15		T2cNxMx
T78	66	Int	35	5	5	55	0	3.15	3+4=7	T2cNxMx
T84	60	high	70	5	0	25	0	9.99	4+5=9	T3bN0Mx
N85	66	Int	0	30	0	70	0	4.37		T3bN0Mx
T86	66	Int	90	5	0	5	0	4.37	4+4=8	T3bN0Mx
T87	61	High	25	45	5	25	0	2.23	4+3=7	T2bN0Mx
N88	61	High	0	10	30	60	0	2.23		T2bN0Mx
T107	68	Int	60	10	0	30	0	7.4	4+3=7	T2bNxMx
N109	67	Low	0	5	0	90	5	7		T2bNxMx
T110	67	Low	40	0	0	58	0	7	3+4=7	T2bNxMx
N113	70	Low	0	10	5	85	0	4.78		T3aNxMx
T114	70	Low	40	0	5	55	0	4.78	4+4=8	T3aNxMx
N121	50		0	30	2	68	0	0.22		
T122	67	Low	70	0	5	25	0	7	3+4=7	T2bNxMx
T123	78		80	0	0	20	0	17.7	5+5=10	NR
N133			0	25	5	75	0			
T147	78	Int	70	0	0	30	0	6.9	4+4=8	T2bNoMx
N148	67	Low	0	35	10	55	0	4.68		T2aNxMx
N155	70	Int	0	40	10	48	2	8.4		T2cNxMx
T167	72	Int	80	0	10	10	0	18	4+4=8	T2bNoMx
T174	83	high	70	5	0	25	0	15	5+4=9	T4
T177	67	Int	40	0	30	30	0	10.87	4+4=8	T2cNoMx
T189	77	N/A	70	0	0	0	30	2.51	5+5=10	T2bN2Mx
T192	61	Int	50	5	10	35	0	5.7	4+4=8	T3aNxMx
N196	73	low	0	40	5	55	0	4.59		T2bNxMx
T197	67	high	95	0	0	5	0	21.82	4+4=8	T3aN1Mx
T198	60		60	0	10	25	0	4.06	4+4=8	T3bNxMx
N201	64		0	20	5	45	0	UNK		T2bNxMx
T202	67	Int	90	0	5	5	0	12.34	4+4=8	T3bNxMx
T204	54	low	80	0	5	15	0	3.91	4+5=9	T3cNxMx

Table 2 Top prostate cancer marker candidates selected by both t-test and SVM-RFE.

Isoform ID	Normalized log2 expr	FDR	SVM-RFE freq.
ALDH1A2-0004	-1.21	1.3E-04	35
AMACR-2094	1.41	6.7E-05	38
AMACR-2097	1.08	9.2E-04	38
AMACR-2098	0.99	1.8E-03	17
ANXA2-0914	-1.04	1.8E-03	36
APBB3-0185	1.01	1.5E-03	38
BC008967-0877	-1.38	7.9E-05	26
C21ORF5-0239	1.24	6.0E-04	35
C7ORF24-0062	1.30	8.4E-05	17
CALCR-1180	1.05	5.2E-04	37
CCT8-0334	1.21	1.5E-04	32
CDC42BPA-1048	-1.19	6.0E-04	38
CDK7-0899	1.35	8.4E-05	37
CES1-0937	-1.34	7.9E-05	32
CLU-0197	-1.11	1.2E-03	38
EDNRB-1187	-1.24	4.7E-04	26
FGFR2-0094	-1.13	4.0E-04	19
FGFR2-0099	-1.03	7.7E-04	28
HEBP2-0472	1.08	7.8E-04	24
HSPD1-0152	1.10	1.8E-03	37
HSPD1-0154	1.17	2.8E-04	31
IGSF4-0722	0.72	2.1E-03	38
IMPDH2-0144	1.25	1.3E-04	34
IQGAP2-0234	1.17	5.6E-04	22
LAMR1-0523	1.20	1.3E-04	38
LTBP4-0746	-1.27	1.5E-04	33
LTBP4-0748	-1.10	1.4E-03	38
LYPLA1-0860	1.38	7.9E-05	35
NELL2-0805	-1.10	1.2E-03	24
PGR-1166	-1.16	4.0E-04	32
PGR-1555	0.85	7.5E-04	38
PPIB-0969	0.94	2.2E-03	34
PTS-0059	-1.07	2.2E-03	31
PYCR1-0058	1.28	4.1E-04	38
RING1-0217	-0.93	1.7E-03	22
SFRS10-1126	0.95	2.0E-03	34
SMPDL3B-2030	1.09	2.2E-04	38
STAC-1044	-1.31	7.9E-05	34
TGFB2-0085	-1.11	6.5E-04	38
TRIM29-1350	-1.29	1.5E-04	35
TRIM29-1353	-1.20	1.7E-04	34
TXNIP-1116	1.09	1.3E-03	38

Detail information of each isoform, such as the exon junction and probe design, can be accessed at the MAASE database (47); FDR is calculated using all 38 samples; SVM-RFE freq.: the number of times that an isoform is included in 38 selected subsets in leave-one-out cross validation.

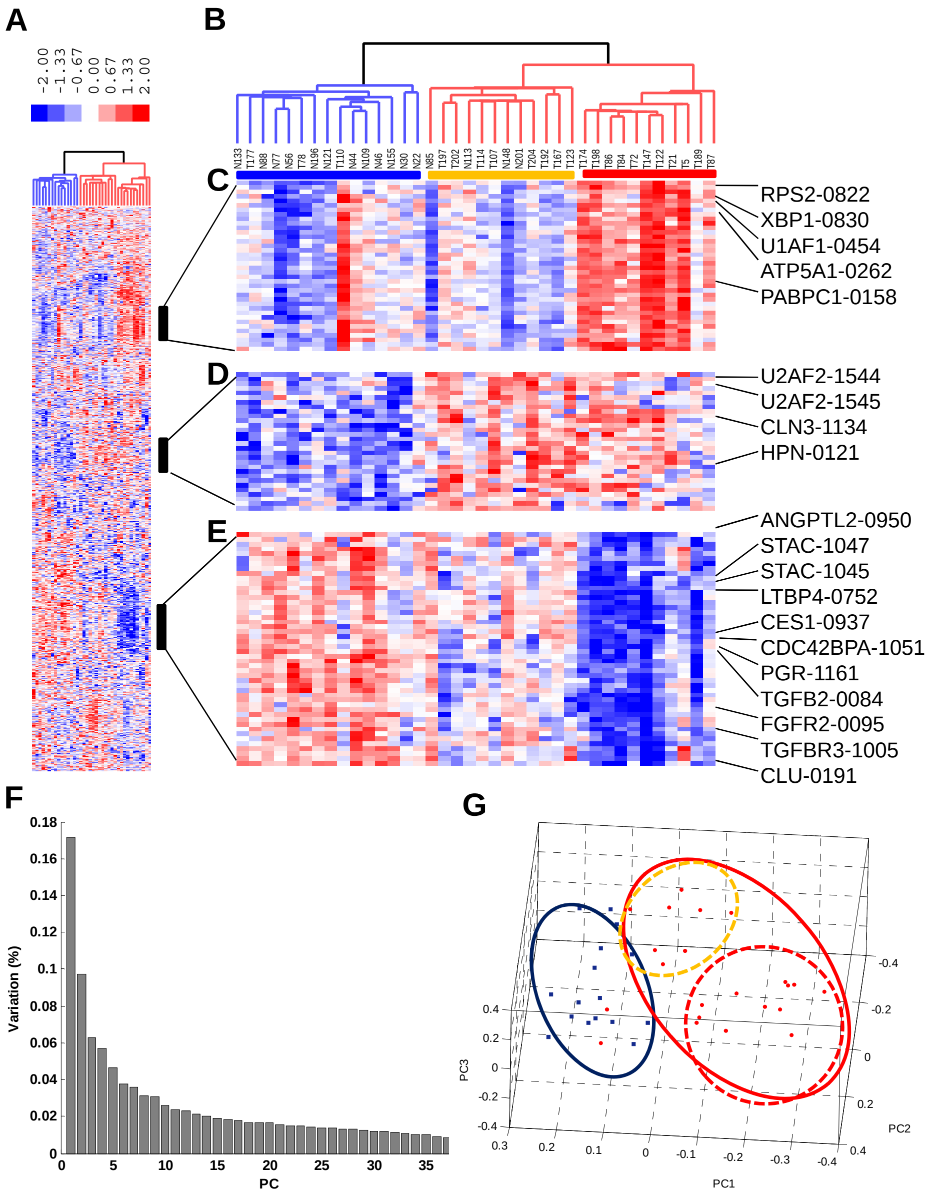


Figure 1: Prostate tumor and normal samples can be separated into distinct groups.

(A) A thumbnail overview of the result of the two-way average-linkage hierarchical clustering of 38 arrays (columns) and 1532 isoforms (rows), as described in ref (30). (**B)** Zoom-in view of the array clustering dendrogram. The two array clusters, C1 and C2, are enriched by normal samples and tumor samples, respectively. Cluster C2 is formed by two sub-clusters, reflecting differences in tumor percentage and stroma. **(C-E)** Isoform signatures up- or down-regulated in different array clusters. (**F and G)** The result of SVD. (**F**) The percentage of variation (y-axis) captured by each principal component (x-axis). (**G)** The low dimensional projection of arrays in the 3D space spanned by the first three principal components. SVD identified the same hierarchical structure as revealed by hierarchical clustering.

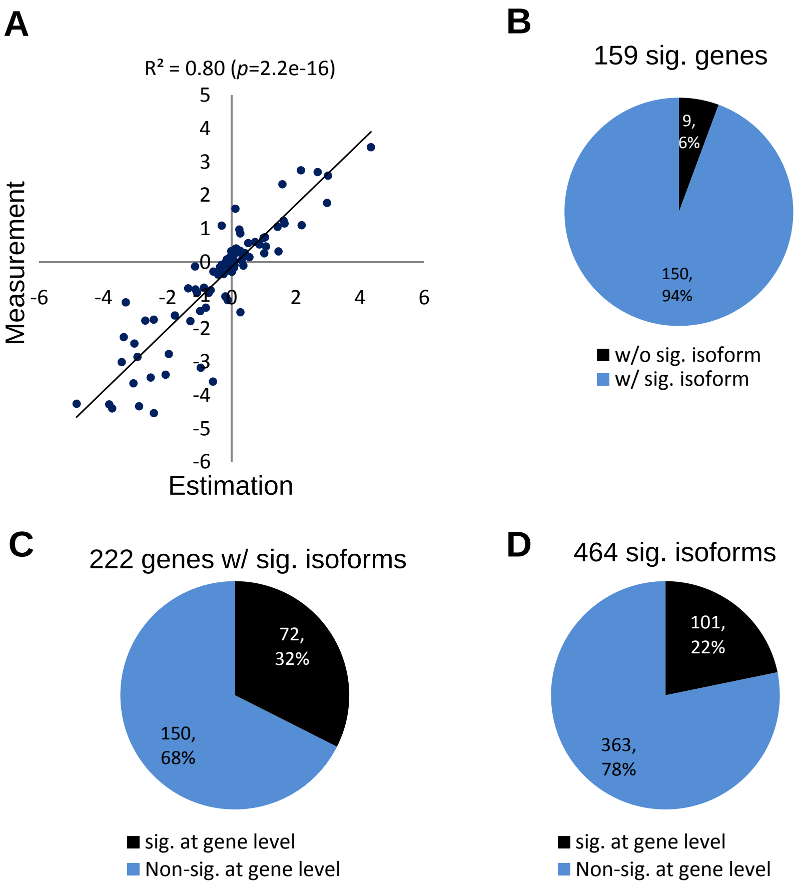


Figure 2: Profiling splice isoforms provides additional useful information for prostate cancer classification.

(A) The validity of estimating the overall mRNA abundance level from the isoform abundance level. The overall mRNA level was estimated by summing up the abundances of individual isoforms for each gene. The estimated mRNA abundances of 107 genes were compared with direct measurements by an independent expression microarray design (described in main text). Plotted are the scatter-plot of log expression ratios of these genes in two prostate cancer cell lines, LNCaP and PC-3. These two approaches show good agreement ($R^2 = 0.80$, p=2.2e-16). **(B)** 159 genes out of 364 profiled genes in the DASL assay exhibit differential expression between tumors and normal samples at the overall mRNA level (q-value=0.05). Most of them (92%) have isoforms with significant differential expression. (**C and D)** 464 isoforms from 222 genes are reported as being differentially expressed between tumors and normal tissues (q-value=0.05), which may be prostate cancer marker candidates. 32% of these genes (corresponding to 22% significant isoforms) do not show differential expression at the overall mRNA level, therefore can not be detected by conventional microarrays.

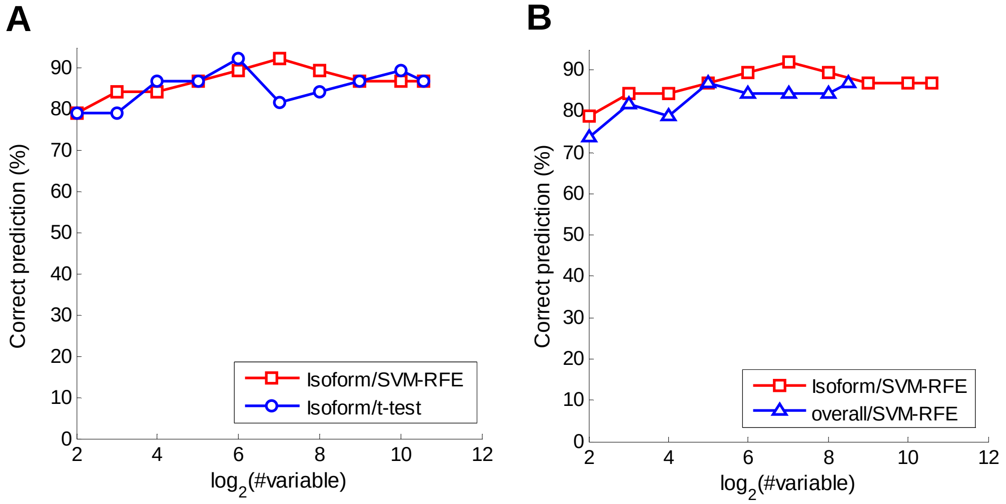


Figure 3: Prediction models built with linear SVM.

The performance is measured by leave-one-out cross validation. To get unbiased result, the variable selection and training are done in training arrays, which is completely independent with the testing array. (**A)** The comparison in classification performance of SVM-RFE selected variables using individual isoforms and the overall mRNAs. (**B)** The comparison in classification performance of variable subsets selected by SVM-RFE and t-test, using individual isoforms.

Chapter 8

Conclusions and perspectives

8.1 Conclusions and discussion

In summary, the major accomplishments of this dissertation include (i) the demonstration of the limited splicing fidelity, which produces widespread non-essential splicing variants, and the impact of purifying selective pressure to eliminate these low-abundance variants; (ii) discovery and characterization of dual-specificity splice sites; (iii) characterization of the pattern and magnitude of selective constraints for optimal exon and intron discrimination, accounting for one-third of the mammalian genomes; (iv) prediction of novel splicing-regulatory elements; (v) demonstration of the extensiveness and modular structure of the Fox-1/2 splicing-regulatory networks; (vi) demonstration of the importance of alternative splicing in cancer, and splicing microarrays in cancer classification and biomarker identification.

Besides these specific achievements described in this dissertation, my study has also raised a number of questions that can potentially lead to new findings. I and/or my collaborators are continuing pursuing some of these questions. For example, for the dual-specificity splice sites, it is provocative to explore the detailed molecular mechanism of the competition between the 5' and 3' splice-site isoforms. At what stage of the spliceosome assembly is the fate of splice-site identity committed? Answers to this question will potentially generate new insights into the mechanism of splice-site recognition and splicing-reaction catalysis. In a second direction, this study also points to a potential connection between dual-specificity splice sites and recursive splice sites, because they have similar motifs. Some dual-specificity splice sites might also function as recursive splice sites, resulting in the skipping of the exons, in which the dual splice sites reside. Indeed, one-third of dual-specificity splice sites have the skipping isoform, which is consistent with the idea. A simple strategy to experimentally test this hypothesis is to see (i) if the intermediate products can be detected and (ii) if mutations in one isoform of the dual-specificity splice site, which converts the dual site into a canonical 5' or 3' splice site, affect the skipping isoform.

A follow-up of the strand-asymmetry study is to experimentally validate some of the predicted EIEs and IIEs using biochemical approaches. This can be done by inserting the elements into the alternative exon of a minigene splicing reporter to see if the elements enhance or repress the inclusion of the exon. Alternatively, the elements can be also inserted into flanking intronic sequences to see if they have the opposite effect. Although multiple lines of evidence suggest that such context-dependent effect exist, not all of these elements necessarily do so. How SR proteins and hnRNPs interact with each other and the spliceosome to establish exon identity and intron identity? It would be very interesting to quantify the global positioning of these proteins (e.g., SF2/ASF and hnRNP A1) throughout the transcripts, using RIP-chip or CLIP technologies. Such experiment will help to answer an important question which was debated for years: is the recognition of exons and introns dominated by positive signals, negative signals, or both?

In the study of Fox-1/2 splicing-regulatory network, the current study is not able to distinguish Fox-1 and Fox-2 targets because they recognize the same sequence motif. HeLa cell was currently used for experimental validation and no apparent difference was observed so far. However, these two proteins have distinct expression patterns spatially (in different tissues) and temporally (in developing neurons). It would be very interesting to test how they regulate target splicing in neuronal tissues. In addition, the current study focuses on the highly conserved targets, which is necessary to achieve a high specificity, but has inevitably overlooked many less conserved, yet potentially important, targets.

8.2 Perspectives and future directions

The next few years will hopefully witness a big advance in the mechanistic understanding of splicing regulation at two levels: the general and more specific splicing code, and the phenotypic implications of splicing regulation. Investigations in these directions will be largely leveraged by the emergence of new RNA technologies, which are reviewed in Chapter 1.

A promising direction in future studies is the wide applications of splicing microarrays, CLIP and RIP, in combination with gene knockout or RNAi. Such experiments will allow the identification of targets genes for specific splicing factors in a genomewide scale. Initially, this can be done for the known important splicing factors, especially those expressed only in specific tissues. However, such studies are also very informative to identify exons and introns that are most susceptible to perturbations of SR proteins and hnRNPs, although these ubiquitous splicing factors regulate most, if not all, exons and introns. Such data will provide invaluable information to build splicing-regulatory networks. Comparisons of different conditions, i.e., between normal samples and disease samples, can help to understand splicing changes in diseases. Although isolated examples exist in previous studies, the global understanding of particular diseases at the splicing level remains very limited.

All the knowledge will be finally integrated to mathematical models to predict splicing outcomes from pre-mRNA sequences, expression of splicing factors, and external perturbations. Of particular interest are those models that only use the information accessible by the spliceosome, because several types of information, such as cross-species conservation, cannot be utilized by the cell although they are very useful for exon/intron prediction. Current efforts towards this goal considered the strength of primary splicing signals, and the general distribution of splicing-regulatory elements. Such models can be improved in several aspects. First, more splicing-regulatory elements have been identified by experimental and computational approaches, and inclusion of these elements will provide a more complete part list. Some of the elements are tissue-specific and functional only in particular conditions when the factors recognizing the elements are expressed. Second, current models usually do not consider RNA structure, which may affect the accessibility of splicing signals. However, computational prediction of RNA secondary structure remains a very challenging task. Third, current models of "splicing simulation" usually consider only single exons, although competitions among overlapping or nearby exons will greatly affect the splicing outcome. Therefore, it will be helpful to optimize the splicing of multiple exons of the same gene together. Lastly, inclusion of more information will also allow the development of richer and more realistic models that can detect interactions among different classes of splicing signals.