Skolkovo Institute of Science and Technology

TISSUE-SPECIFICITY AND REGULATION OF ABERRANT ALTERNATIVE SPLICING

Doctoral Thesis
by

ALEKSEI MIRONOV

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Assistant Professor, Dmitri D. Pervouchine

Moscow - 2022

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgment is made, and has not been submitted for any other degree.

Candidate (Aleksei Mironov)

Supervisor (Assistant Professor Dmitri D. Pervouchine)

Abstract

Alternative splicing plays a crucial role in the regulation of gene expression and expanding the diversity of eukaryotic proteomes. Besides the body of well-annotated alternative splicing events that substantially change the protein amino acid sequence, there is also a multitude of rarely expressed aberrant splice isoforms that lack functional characterization. However, aberrant splicing events may be important in specific physiological conditions and, therefore, their discovery and characterization is a prominent problem in bioinformatics. Recent development of high-throughput sequencing technologies has enabled the analysis of transcriptomes at unprecedented depth, thus opening new avenues to the characterization of aberrant splicing. This dissertation is devoted to two varieties of aberrant splicing events, the so-called tandem alternative splicing sites (TASS) and unproductive splicing events (USEs). Bioinformatic analysis that integrates several large-scale data sources, including transcriptome data of healthy human tissues provided by the Genotype-Tissue Expression (GTEx) Consortium, transcriptome response to the perturbation of RNA-binding proteins (RBPs), RBP footprinting assays, and other relevant data reveals unique tissue-specific properties of aberrant splicing events. The dissertation presents genome-wide catalogues of TASS and USEs, characterization of their tissue-specific and cell type-specific expression, and predictions of their regulation by RBPs.

Publications

Main author

1. A. Mironov, S. Denisov, A. Gress, O. V. Kalinina, and D. D. Pervouchine. An extended catalogue of tandem alternative splice sites in human tissue transcriptomes. PLoS Comput Biol, 17(4):e1008329, 04 2021

Co-author

1. S. Kalmykova, M. Kalinina, S. Denisov, A. Mironov, D. Skvortsov, R. Guigó, and D. Pervouchine. Conserved long-range base pairings are associated with pre-mRNA processing of human genes. Nat Commun, 12(1):2300, 04 2021

2. M. Sorokin, I. Kholodenko, D. Kalinovsky, T. Shamanskaya, I. Doronin, D. Konovalov, A. Mironov, D. Kuzmin, D. Nikitin, S. Deyev, A. Buzdin, and R. Kholodenko. RNA Sequencing-Based Identification of Ganglioside GD2-Positive Cancer Phenotype. Biomedicines, 8(6), May 2020

Conference presentations

1. A. Mironov, S. Denisov, A. Gress, O. V. Kalinina, and D. D. Pervouchine. An extended catalogue of tandem alternative splice sites in human tissue transcriptomes. MCCMB, Moscow, Russia, 2021

2. A. Mironov, S. Denisov, O. V. Kalinina, and D. D. Pervouchine. Functional annotation of splicing aberrations in non-coding RNA. EMBO | EMBL Symposium: The Non-Coding Genome, Heidelberg, Germany, 2019

3. A. Mironov, S. Denisov, O. V. Kalinina, and D. D. Pervouchine. Structural annotation of protein indels associated with splicing aberrations. MCCMB, Moscow, Russia, 2019

Contents

List of symbols, Abbreviations

AS Alternative Splicing

GTEx Genotype Tissue Expression project

mRNA messenger RNA

NMD Nonsense-mediated mRNA decay

PTC Premature termination codon

RBP RNA-binding protein

TASS Tandem alternative splicing sites

USE Unproductive splicing event

Cn Consensus nucleotide of a splice site

Nc Non-consensus nucleotide of a splice site

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

Genetic programs encoded in the DNA are executed in a process called gene expression. During this process, DNA is transcribed into RNA, which is then translated into proteins and eventually degraded. In eukaryotes, RNA is subject to various post-transcriptional modifications, including splicing, in which some parts of the RNA sequence (exons) are joined together while others (introns) are removed. The majority of eukaryotic genes undergo alternative splicing (AS), enabling generation of many diverse transcripts from the same gene. Studies have shown that AS is widely implicated in development, cell differentiation, tissue formation, stress response, and disease. However, besides protein-coding transcripts, which are characterized by high expression levels, strong evolutionary selection, and conservation, splicing machinery generates a multitude of rarely expressed alternatively spliced transcripts with unknown function. Several studies have attributed these weakly expressed transcripts to splicing aberrations. Nevertheless, such rare isoforms could be used by the cell only under specific conditions and, therefore, are as important as highly-expressed transcripts.

Current advances in genomic research and next-generation sequencing (NGS) have enabled large-scale transcriptome studies which simultaneously assess transcriptional activity in thousands of biological samples. One of them, the Genotype Tissue Expression (GTEx) project, has produced the largest to-date collection of RNA-seq experiments in healthy human tissues. The ENCODE consortium has created a large panel of shRNA-mediated knockdowns of more than two hundred

RNA-binding proteins (RBPs) and performed RNA-seq experiments to assess the responses of the cellular transcriptome to these perturbations. Based on these and other data sources, this dissertation presents a systematic assessment of tissue specificity and prediction of regulation of two types of aberrant AS events, the so-called tandem alternative splicing sites (TASS), which are characterized by close tandem arrangement of alternative splice sites, and unproductive splicing events (USEs), which generate substrates of the nonsense-mediated decay (NMD) pathway. In TASS, alternative donor (5') or acceptor (3') splice sites are located at a distance of several nucleotides from each other, and, as shown here, only one of them is predominantly expressed, while others usually originate from splicing noise. In USEs, the inclusion of an alternative splice isoform results in the incorporation of a premature termination codon (PTC) into the transcript, causing its degradation by NMD. Earlier, it was believed that the main role of NMD is to control the quality of splicing and prevent translation of truncated, dysfunctional proteins. However, it was found later that a coupling between AS and NMD (AS-NMD), referred to as unproductive splicing, is an important mechanism of gene expression regulation.

The main results of this dissertation are presented in two parts related to tandem alternative splicing sites (Chapter 5) and unproductive splicing (Chapter 6). Chapter 5 describes the catalogue of known and novel TASS along with the characterization of their tissue-specific expression, regulation, impact on protein structure, and evolutionary selection. Chapter 6 presents the analysis of tissue-specificity and regulation by RBPs across USEs in the human transcriptome. Chapter 2 contains the literature overview and delivers the background that is necessary for understanding the remaining chapters. Chapter 3 outlines thesis objectives. Chapter 4 contains the description of materials and methods. Chapters 7 and 8 are devoted to discussion of the results and conclusions.

Chapter 2

Background

2.1 Splicing

Widespread post-transcriptional RNA processing is one of the distinguishing features of eukaryotes compared to prokaryotes [1]. An important step in the RNA maturation is splicing, a process in which stretches of pre-mRNA called introns are excised, and the remaining sequences, called exons, are ligated together. All eukaryotic genomes contain introns, but their abundance and lengths vary greatly between species [2, 3]. An absolute majority of human protein-coding genes contain at least one intron and undergo splicing [4, 5], with the median number of introns in human intron-containing genes being about eight [3].

The correct definition of introns and exons is an important prerequisite for the proper execution of eukaryotic expression programs. It is underscored by the role of splicing errors in the pathophysiology of many diseases, such as cystic fibrosis [6], familial dysautonomia [7], tauopathy [8, 9], and many other hereditary diseases [10]. Mis-splicing can also contribute to the development of malignancies, for example, in myelodysplasia [11] and colorectal carcinoma [12].

The evolutionary benefits of having splicing in eukaryotic genes is a subject of many debates [13, 14]. Some researchers believe that introns invaded the genomes of emerging eukaryotes and behaved as selfish elements [15, 16, 17, 18]. Later, their evolution resulted in massive intron losses in primitive eukaryotes, while in complex eukaryotes a number of introns remained, which some authors associate with the

evolutionary advantages, including the possibility of alternative splicing and the regulation of gene expression [16, 13, 14].

During AS, introns of pre-mRNA are excised in many different ways and combinatorially increase the number of possible isoforms transcribed from the same gene [19]. Approximately 95% of mammalian genes are susceptible to AS [20], which strongly influences transcriptome diversity [21, 22] and provides additional layers of gene expression regulation [13].

Besides AS, the presence of introns *per se* affects gene expression. On the one hand, introns delay transcription elongation and pre-mRNA processing [23]. Accordingly, genes tend to contain fewer introns if they are involved in active cell division, as, for example, during early embryogenesis [24, 25, 26] or in the cellular response to stress [27]. On the other hand, introns can indirectly stimulate transcription by recruiting transcription and chromatin remodeling factors [28, 29], stimulating gene looping during transcription [30], or other mechanisms [31]. Thus, eukaryotic cells can fine-tune gene expression intensity by virtue of having larger or smaller number of introns.

Pre-mRNA splicing is performed via two transesterification reactions [32] catalyzed by a megadalton ribonucleoprotein complex called the spliceosome [33]. The spliceosome first recognizes conserved cis-regulatory elements in the pre-mRNA that determine the boundaries of introns and exons. Then, it enforces a conformational change in the pre-mRNA, in which splicing reactions become energetically favorable [33].

2.2 The molecular mechanism of splicing

In mammals, introns are primarily defined by four sequence elements in the pre-mRNA [32]: the 5'-splice site, also called the donor splice site, adenosine branch point, polypyrimidine tract that mostly consists of 15-20 pyrimidines, and the 3'-splice site, also called the acceptor splice site (Fig 2-1).

These elements control a two-step phosphoryl transfer mechanism of splicing (Fig 2-2). In the first reaction (branching), the 2'-hydroxyl group of the adenosine

Figure 2-1: **Mammalian primary splicing cis-elements [32].** The consensus nucleotide sequences are shown. N, R, and Y stand for any base, purine, and pyrimidine, respectively.

at the branch point attacks the phosphodiester group at the 5'-splice site forming a cleaved upstream exon and an intermediate complex consisting of a lariat intron and a downstream exon, in which the 5'-phosphate of the first nucleotide intron (G) is bound to the 2' oxygen branch point. In the second reaction (exon ligation), the exposed 3'-hydroxyl group of the upstream exon attacks the phosphodiester group at the 3'-splice site, ligating the upstream and the downstream exons and releasing the intron lariat.

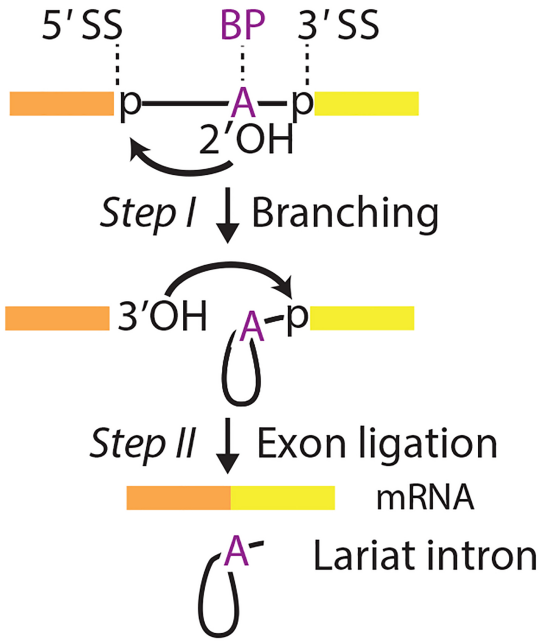

Figure 2-2: **A two-step phosphoryl transfer mechanism of splicing [32].** 5' SS, 3' SS and BP denote the 5'-splice site, the 3'-splice site, and the branch point, respectively.

Splicing is a relatively simple chemical process, yet it is regulated and catalyzed

by a very complex macromolecular machinery consisting of the major ribonucleoprotein complex (the spliceosome) and about 150 auxiliary regulatory proteins (splice factors) [33, 32, 34]. Most introns have a low sequence conservation [35, 36] and do not adopt any particular secondary or tertiary structures, instead depending on the spliceosome to align their reactive sites [37].

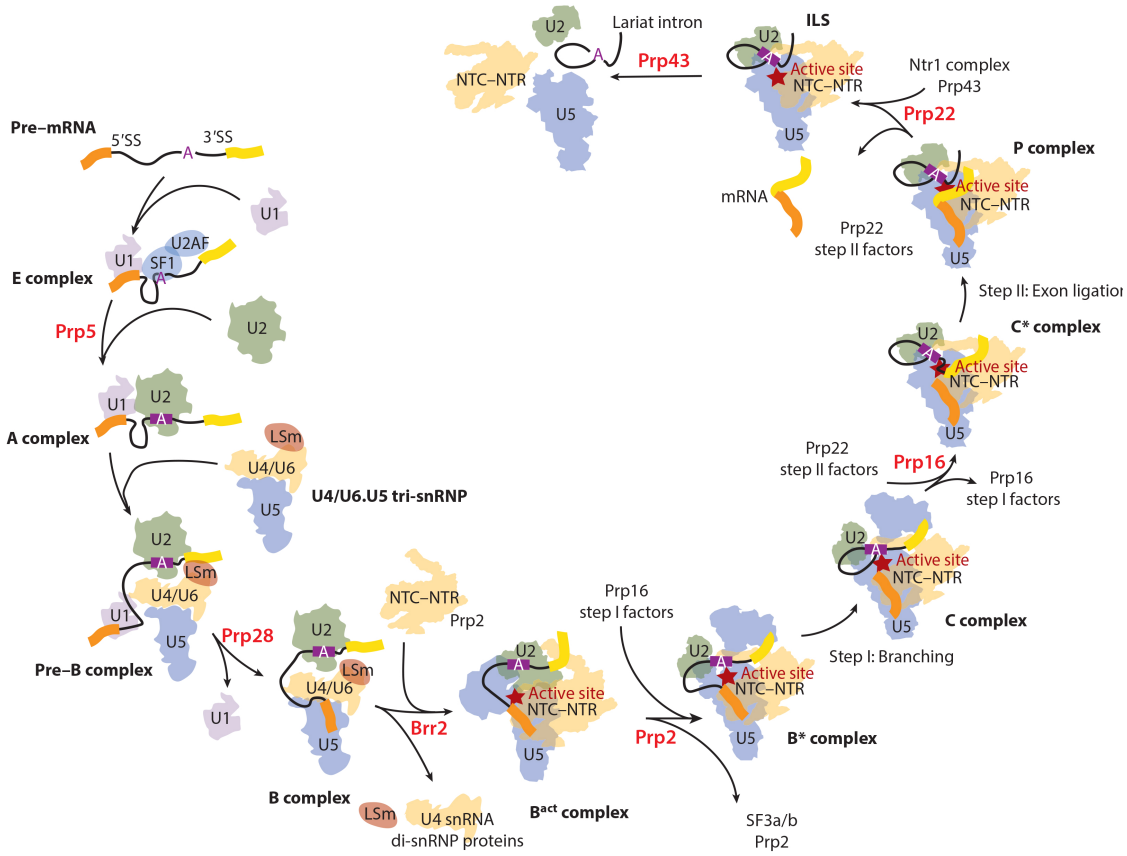

Figure 2-3: **Spliceosome assembly on a pre-RNA [32].**

Recent advances in cryo-electron microscopy allowed to visualize the spliceosome in a remarkable detail [37]. The spliceosome is composed of five subunits called small nuclear ribonucleoproteins (snRNPs, *U1*, *U2*, *U4*, *U5*, and *U6*) consisting of proteins and small nuclear RNAs (snRNAs). The interaction of snRNAs with proteins occurs in the cytoplasm and forms stable but inactive pre-snRNPs [33]. The pre-snRNPs are then re-imported back into the nucleus to be assembled dynamically on the transcribed mRNA. This property of performing the primary biogenesis in the remote compartment presumably represents a quality control mechanism, which

is not unique to the spliceosome and has also been observed in the maturation of miRNA [38], snoRNA [39], and ribosomal subunits [40].

Spliceosome assembly on the pre-RNA proceeds in a dynamic cascade of multiple protein-RNA and RNA-RNA interactions driving extensive structural and compositional rearrangements eventually shaping the splicing-prone conformation of the pre-RNA [37, 33, 32] (Fig 2-3). The rearrangements are mostly conducted by RNA helicases, including eight major factors (HGNC gene names are shown in parenthesis): DEAD-box helicases *PRP5* (*DDX46*), *PRP28* (*DDX23*), *UAP56* (*DDX39B*), Ski2-like helicase *BRR2* (*SNRNP200*), and DEAH-box helicases *PRP2* (*DHX16*), *PRP16* (*DHX38*), *PRP22* (*DHX8*), and *PRP43* (*DHX15*) [37, 41]. Each rearrangement is accompanied by a change in the composition of the associated splice factors and results in a formation of a new spliceosomal complex [37].

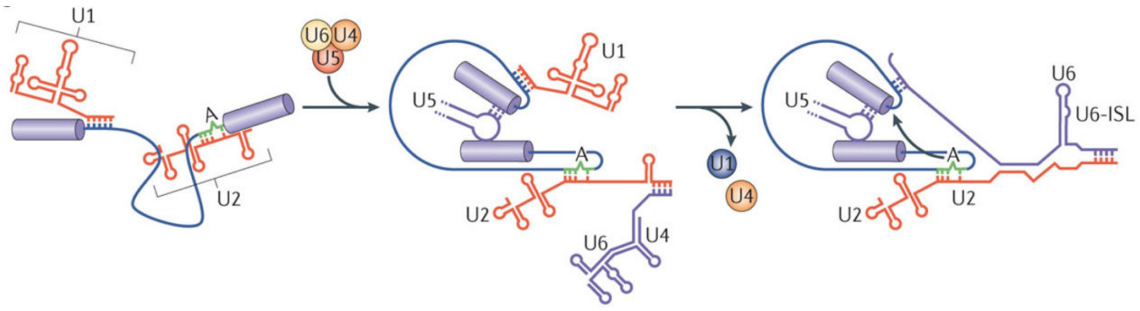

Figure 2-4: **The formation of the spliceosome catalytic core by snRNAs [33].** U6-ISL stands for the intramolecular stem-loop that is important for splicing catalysis.

The preparatory step of splicing reaction consists of six stages, by the end of which an active spliceosome is formed (B* complex). At the first stage, the *U1* snRNA interacts with the 5'-splice site with the help of SR proteins and other splice factors [42]. In these steps, the *SF1* binds the branch point, and *U2AF65/U2AF35* heterotrimer recognizes the polypyrimidine tract and 3'-splice site sequences, forming the E complex. Subsequently, according to the studies in yeast [33], *PRP5* and *UAP56* helicases catalyze the transition to the A complex, in which *SF1* and *U2AF* are displaced to allow *U2* snRNA to bind the branch site with the adenosine being bulged out (Fig 2-4, left) and interact with *U1* snRNP. At the third stage, the preassembled *U4/U6.U5* tri-snRNP is recruited to the A complex, forming the pre-B

complex in which the tri-snRNP is not stably bound [43] (Fig 2-4, middle).

In higher eukaryotes, the generation of the A and pre-B complexes is complex because introns are much longer than exons. For example, in humans and mice, the lengths of an intron is, on average, six times larger than the length of an exon, while in *Drosophila melanogaster* this ratio is around two [3]. Therefore, splice sites are mostly recognized in pairs across exons rather than introns through the interaction of *U1* and *U2* snRNPs located in different introns flanking the same exon [44, 45]. Recent structural studies demonstrated strong similarities in the E and A complexes formed across introns and exons but highlighted steric hindrance in the recruitment of the *U4/U6.U5* tri-snRNP to the A complex assembled around a short exon [46]. Some authors suggest that such hindrance makes the spliceosome stall at the pre-B stage and further remodel into an intron-spanning B complex involving the upstream 5'-splice site [47].

At the next stage, the pre-B complex transforms into a pre-catalytic spliceosome (B complex) mainly by the helicase *PRP28* [43], which disrupts the base-pairing between the *U1* and the 5'-splice site and allows the latter to base-pair with the *U6* snRNA. Next, RNA helicase *BRR2* unwinds the *U4/U6* snRNA duplex, which leads to the removal of the *U4* snRNA [41] and allows *U6* and *U2* snRNAs to fold together near the so-called internal stem-loop (ISL) in the *U6* snRNA, and, cradled by the NTC and NTR protein complexes [32], form the catalytic core of the spliceosome within the B^{act} complex. However, both the branch point adenosine and the 5'-splice site are blocked from the active site by the *SF3B* complex and *SF3A2* protein, respectively [32]. Finally, at the sixth stage, the *PRP2* helicase disrupts the interaction of the pre-RNA with the *SF3B* complex and transforms the B^{act} into the B* complex, although the exact mechanism of this rearrangement is currently unknown [32, 37]. As a result, the branch point adenosine and the 5'-splice site are docked into the catalytic core (Fig 2-4, right).

The B* complex catalyzes the first step of the splicing reaction and transforms into the C complex. The latter further rearranges into the C* complex with the help of the *PRP16* helicase and performs the second transesterification reaction, which yields the P (post-splicing) complex [48]. The release of the mRNA that generates

the intron lariat spliceosome (ILS) is catalyzed by the helicase *PRP22*. Finally, the ILS is disassembled by *PRP43* helicase, and the *U2*, *U5*, and *U6* snRNPs are released and recycled for additional rounds of splicing [37].

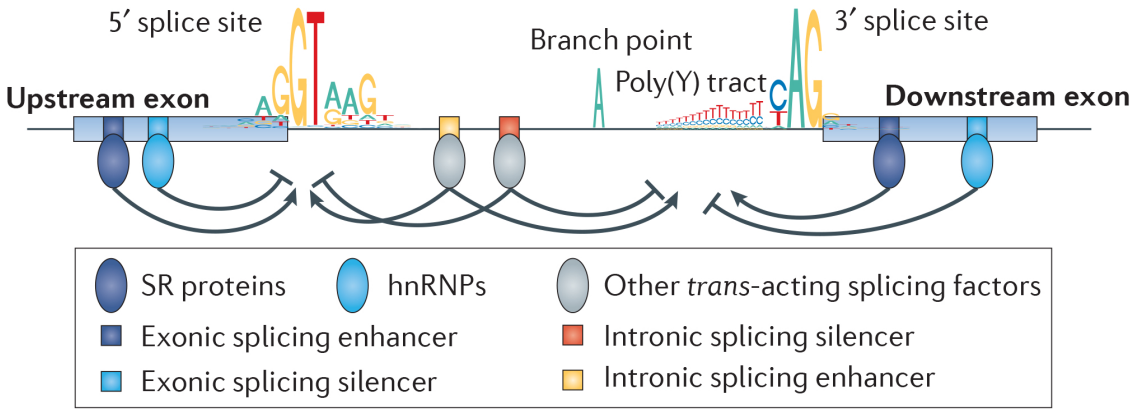

Figure 2-5: ***Cis*-elements and *trans*-factors of splicing [49].**

All stages of the spliceosome formation are coordinated by the *cis*-acting enhancer and silencer elements recognized by the *trans*-acting proteins that either stimulate or repress the spliceosome complex assembly (Fig 2-5).

Among the best studied are the so-called serine/arginine-rich (SR) proteins, which typically stimulate both constitutive and alternative splicing upon binding to exonic enhancer sequences [50, 19]. For example, *SRSF1* and *SRSF2* participate in the formation of the E complex, where they promote the recruitment of *U1* and *U2* snRNPs to the 5'- and 3'-splice sites, respectively [42, 51, 52]. In addition, their phosphorylation/dephosphorylation cycle within the spliceosome is required for the transition to the catalytic state [50]. Another important family of splicing factors are the so-called heterogeneous nuclear ribonucleoproteins (hnRNPs), which, in contrast to SR proteins, often repress spliceosomal assembly [53]. The known mechanisms of their action include the prevention of *U1* and *U2AF* binding to the pre-RNA at the initial stage of the E complex formation [54, 55, 56], as well as the prevention of the transition to the B complex [57].

The cryo-electron microscopy and spliceosome profiling experiments enables a remarkable progress in understanding the structural and compositional rearrangements of the spliceosome, revealing the formation of branching sites and open 3'-ends

of exons within catalytically active spliceosome *in vivo* [58].

Recent studies identified a number of non-canonical cases, including interrupted splicing events, in which the spliceosome stalls at the C complex, does not proceed to the P complex, and produces non-functional transcripts that are degraded later. They also detected recursive and nested splicing, in which the intron is excised in several consecutive splicing reactions. These findings identified at least some of the mechanisms behind the abundance of lowly expressed splice isoforms that show little evolutionary conservation and are often considered as splicing aberrations or splicing noise [59, 60]. However, aberrant splicing is inherently difficult to distinguish from regular alternative splicing, which is often considered as a major driver of eukaryotic proteome expansion [21].

2.3 Alternative splicing

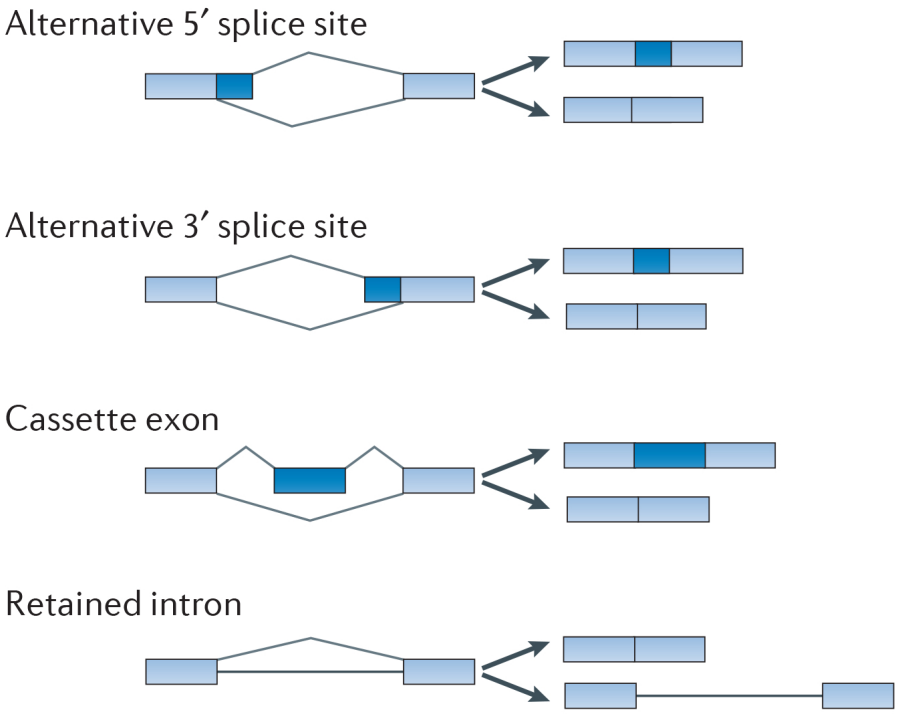

Figure 2-6: **Types of alternative splicing [49].**

Alternative splicing (AS) is a mechanism that allows pre-RNA to be processed into several different mRNAs via excision of different introns. The repertoire of AS

events is broad, with exon skipping being the most abundant, followed by alternative 3'-splice sites, alternative 5'-splice sites, and intron retention [61, 62] (Fig 2-6).

AS abundance is tightly correlated with the organism complexity [63]. Most eukaryotic genes undergo AS [64], but its functional role is still largely unknown, especially outside the context of cancer-specific AS [65, 66]. Also, a popular statement of a widespread extension of the proteome by AS has been recently debated, as the evidence of translation was obtained only for 2% to 37% of alternative isoforms [67, 68].

The degree of specificity to particular tissues, cell types, and developmental stages is often recognized as a proxy for AS events to be under regulation [69, 70], and such events are generally more evolutionarily conserved [71]. The examples of AS events observed during organ development and differentiation were found for mesenchymal stem cell [72] and myoblast [73] differentiation, development of neural tissues [74], heart [75], liver [76], testis [77], and other organs [66]. A recent study systematically assessed AS patterns across pre- and post-natal development of seven organs in six mammals and chicken [71]. It was found that brain tissues harbor the highest number of AS isoforms specific to particular development stages, and such isoforms are substantially more conserved between species than AS isoforms not associated with particular developmental stages. The authors of this study concluded that the interplay between AS and gene expression programs is fundamental to organ development, especially for the brain and heart. These results recapitulated previous discoveries, which emphasized the limited tissue-specificity of AS outside of the brain, heart, muscle, and testis tissues [78].

Tissue-specific AS relies on the competitive and synergistic interaction of tissue-specific and ubiquitously expressed RBPs [79, 80]. Only a small fraction of RBPs are tissue-specific, and most of them are differentially expressed in testis, brain, liver, or muscle tissues [81]. At that, RBPs that are expressed in the nervous system are particularly enriched with splice factors [82]. Notable examples of neural splice factors include *NOVA1/2*, *RBFOX1/2/3*, *PTBP1/2/3*, and *SRRM4*, knockout of which results in severe neurodevelopmental defects or lethal phenotypes [83]. However, the impact of specific neural splice variants activated by these RBPs is mostly

unknown.

Most splicing reactions occur co-transcriptionally [84, 85]. Therefore, in addition to the differential expression and activity of splice factors, AS is extensively affected by the RNA polymerase elongation rate [86], chromatin structure [87], DNA modifications [88], and other epigenetic factors [89]. Apart from that, the positions of exons tend to be aligned with those of the nucleosomes [87], which may influence exon definition.

2.4 Aberrant splicing

Despite many layers of regulation, AS is affected by noise and may result in stochastic fluctuations (aberrations) of alternative isoform abundance [90, 91]. Remarkably, less than 5% of AS events are conserved beyond mammals [64, 78]. This observation led to a hypothesis that a sizable fraction of AS represents non-regulated random events caused by the spliceosome errors [91, 60, 59, 92]. The spliceosome performs exon definition and accurate splice site selection during a complex, multi-step, dynamic assembly process in which each step is subject to regulation. Such flexibility yields remarkable diversity of produced alternative RNA isoforms, yet opens a large room for random fluctuations [93].

The primary driver of the fluctuations is the fuzziness of the splicing regulator sequences in the pre-mRNA, neither of which plays a dominant role [33]. For example, there are many cryptic splice sites throughout the transcriptome where splicing is not detectable despite seemingly active consensus sequences [94]. Moreover, the base-pairing between snRNA and the 5'-splice site sequence does not require full complementarity and may involve bulged nucleotides, thus predisposing the spliceosome to errors of splice site detection [95]. The outcome of the splicing reaction is determined by the competitive and synergistic influence of many diverse regulators, which adds plasticity to splicing regulation [96], but may also bring additional stochasticity, similar to the increase of transcriptional noise by the synergetic influence of transcription enhancers and silencers [97]. In addition, somatic or germline mutations and incorrect transcription of splicing regulatory elements also greatly

contribute to erroneous splicing [91].

Pickrell et al. studied low abundance isoforms using a set of *de novo* identified alternative 5'- and 3'-splice sites. It was estimated that in 0.7% of splicing reactions, the spliceosome erroneously picks an evolutionarily non-conserved splice site instead of choosing a conserved one [59]. The study also demonstrated a significant positive correlation between the level of splicing errors and intron length, yet no association with specific motifs. It further showed that the same motifs of splice factors enriched in the vicinity of constitutive splice sites were also found in the vicinity of non-conserved splice sites.

However, a sizable fraction of low abundance transcript isoforms is strictly regulated by splicing factors, thus indicating their possible implication in physiological processes. For example, *NOVA1* controls the inclusion of more than 200 NMD-inducing cryptic exons that are normally suppressed but activated during brain seizures [98]. Similarly, dozens of poorly conserved NMD-inducing cryptic exons are repressed by *TARDBP* [99] and *PTBP1/PTBP2* [100]. The loss of *TARDBP* is a hallmark of neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), and the genetic mechanism behind their pathophysiology involves cryptic splicing controlled by *TARDBP* [101].

2.5 Tandem alternative splice sites

TASS are a major subtype of alternative 5'- and 3'-splice sites characterized by a close tandem arrangement of splice sites [102, 69]. About 15–25% of mammalian genes possess TASS, and they occur ubiquitously throughout eukaryotes, in which alternative splicing is common [102]. Most TASS are believed to originate from splicing aberrations [103, 102]. Besides that, several TASS and their protein products were experimentally shown to be functionally involved in DNA binding affinity [104], subcellular localization [105], receptor binding specificity [106] and other molecular processes [102].

The outcome of AS of a frame-preserving TASS on the amino acid sequence encoded by the transcript is equivalent to that of a short genomic insertion or

deletion (indel). Indels cause broad genetic variation in the human population and impact human traits and diseases [107, 108]. For a different type of alternative splicing with a similar effect on amino acid sequence, alternative microexons, it has been demonstrated that the insertion of two amino acids may influence protein-protein interactions in the brain of autistic patients [109]. Structural analysis of frame-preserving genomic indels revealed that they predominantly adopt coil or disordered conformations [110]. Likewise, frame-preserving TASS with significant expression of multiple isoforms are overrepresented in the disordered protein regions and are evolutionarily unfavorable in structured protein regions [111].

The two most studied classes of TASS are the acceptor NAGNAGs separated by 3 nucleotides (nt) [69, 112, 113, 114] and the donor GYNNGYs separated by 4 nt [115]. In these TASS classes, AS is significantly influenced by the features of the cis-regulatory sequences, but much less is known about their function, tissue-specific expression, and regulation [115, 69, 116]. Recent genome-wide studies estimated that at least 43% of NAGNAGs and ∼20% of GYNNGYs are tissue-specific [115, 69]. It is believed that closely located TASS, such as NAGNAGs and GYNNGYs, originate from the inability of the spliceosome to distinguish between closely located cis-regulatory sequences, and, therefore, most TASS are attributed to splicing errors or noise [115, 117, 118]. However, it is not evident from the proteomic data what fraction of alternative splicing events and, in particular, of TASS splicing indeed lead to the changes in the protein amino acid sequence [67, 119, 120].

2.6 Nonsense-mediated decay

Nonsense mutations and frame-disrupting splicing errors give rise to transcripts with PTCs, which encode truncated, deleterious proteins. In eukaryotes, such transcripts are selectively degraded by the translation-dependent surveillance mechanism called the Nonsense-mediated decay (NMD) [122]. NMD is mediated by the RNA-dependent helicase and ATPase *UPF1* that binds to accessible mRNA molecules in the cytoplasm but is displaced from the protein-coding sequences by translating ribosomes [121, 123]. Two alternative pathways of NMD have been proposed: exon

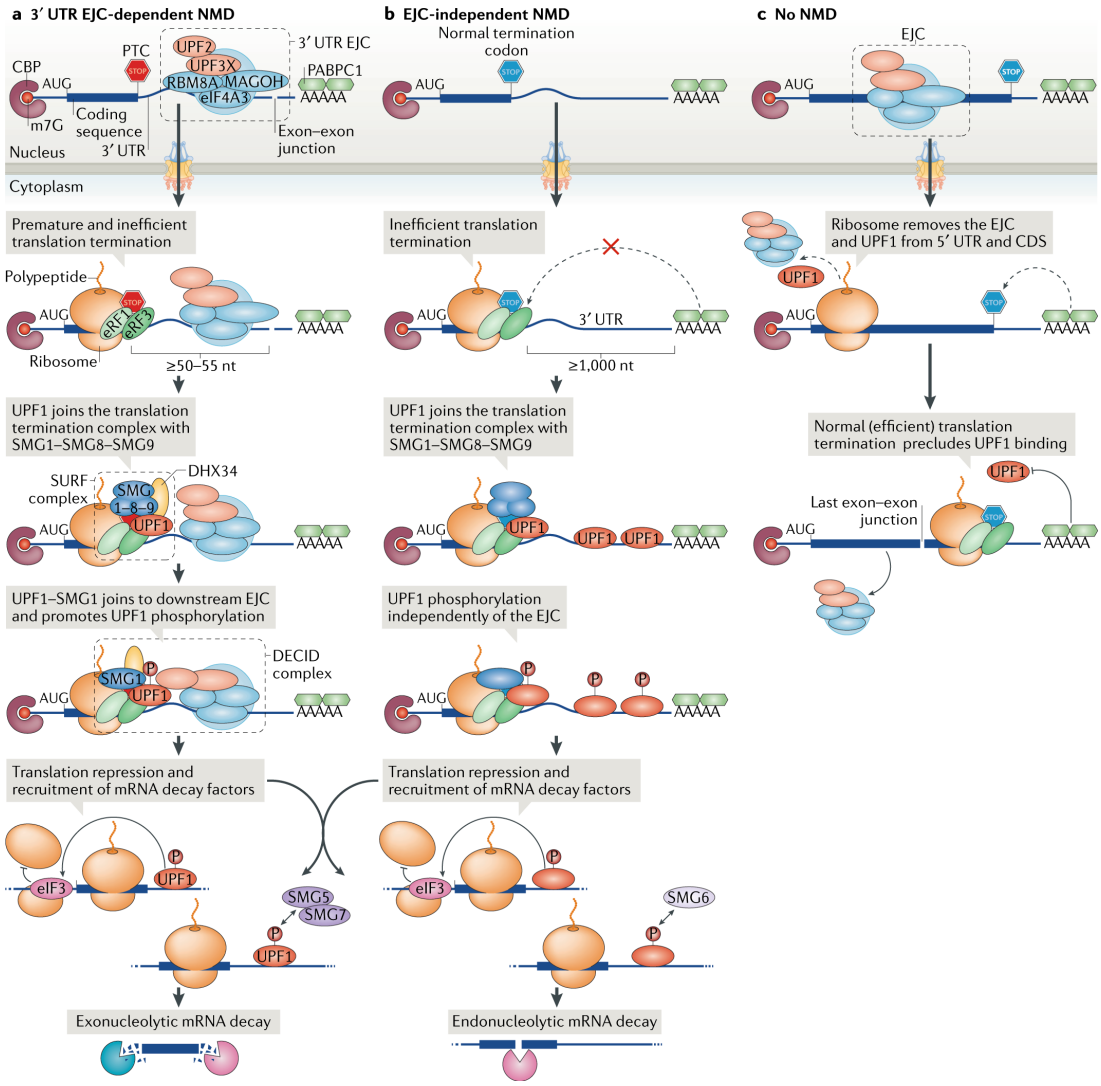

Figure 2-7: **The molecular mechanisms of NMD [121].** **a** EJC-mediated NMD. The presence of the EJC downstream from the stop codon makes the translation termination inefficient, presumably due to the interference of interaction between *PABPC1* and the translation termination complex. *UPF1* and co-factors join the translation termination complex and interact with EJC, which leads to the phosphorylation of *UPF1* and subsequent recruitment of mRNA decay factors. **b** EJC-independent NMD can be triggered at long 3' UTRs promoting the binding and phosphorylation of *UPF1*. **c** Normal translation termination is presumably promoted by the interaction of *PABPC1* with the translation termination complex, which precludes *UPF1* binding.

junction complex (EJC)-dependent NMD and EJC-independent NMD; both pathways implicate *UPF1* but require different co-factors and NMD-activating features of the targeted mRNA (Fig 2-7). At that, EJC-dependent NMD is recognized as more efficient than EJC-independent NMD [121].

In the EJC-dependent pathway, NMD distinguishes premature from normal translation termination by the presence of EJCs downstream from the stop codon [124, 125]. EJCs are deposited approximately 20–24 nt upstream of the exon-exon junctions during pre-mRNA splicing and later displaced physically from the mRNA sequence by translocating ribosomes [126]. EJCs that remain associated with the mRNA after the initial round of translation serve as indicators of whether the stop codon is premature or not because the latter is usually located in the last exon. The presence of exon-exon junctions at least 50–55 nucleotides downstream of the stop codon triggers NMD, the efficiency of which is increased with the distance to the nearest exon-exon junction and the abundance of exon-exon junctions located downstream [121]. In addition, the efficiency is modulated by the differential association of EJCs with auxiliary proteins, such as *SRSF1* [127], *RNPS1* [128], and other co-factors [121, 129]. Upon the translation termination, *UPF1* and *SMG1* kinase join the translation termination complex to form the so-called SURF complex [130]. The key NMD-activating event is the phosphorylation of *UPF1* by *SMG1* that is triggered by the interaction of the SURF complex with the downstream EJC, although the molecular mechanism of this interaction is currently not completely established [121].

In the normal termination of translation, the interaction of *UPF1* with the translation termination complex is outcompeted by the cytoplasmic poly(A)-binding protein 1 (*PABPC1*) bound to the poly(A) tail not far from the stop codon [131]. Conversely, a long 3'-untranslated region (UTR) may be stochastically bound by *UPF1*, which results in the repression of the interaction between *PABPC1* with the translation termination complex and subsequent EJC-independent *UPF1* phosphorylation, the molecular mechanism of which is currently unknown [121]. However, long 3' UTRs often contain multiple binding sites for RBPs that directly or indirectly inhibit NMD [132] contributing to a weak correlation between 3' UTR length and the efficiency of NMD [133, 134]. The steps following *UPF1* phosphorylation are the same in EJC-dependent and EJC-independent NMD. The phosphorylated *UPF1* recruits the endonuclease *SMG6* and other factors causing deadenylation and decapping, targeting the cleaved mRNA for degradation by cellular exonucleases [125].

2.7 Unproductive splicing

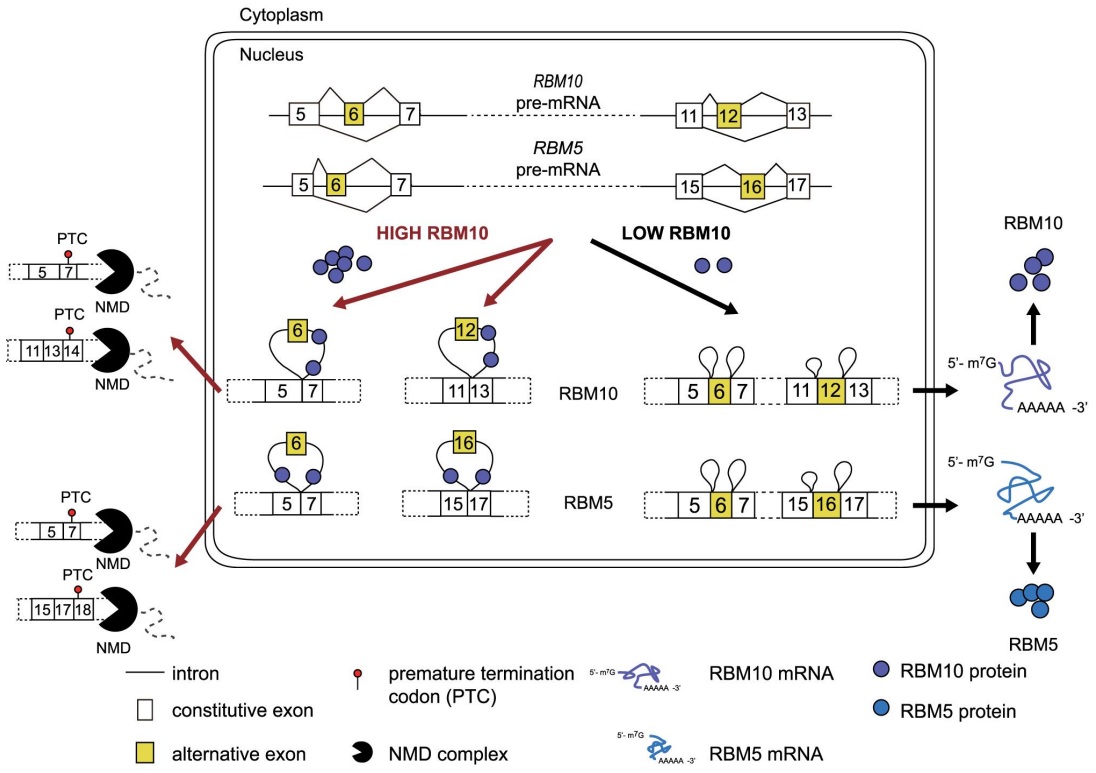

Figure 2-8: **The model of *RBM10* autoregulation and cross-regulation of *RBM5* via unproductive splicing [135].**

NMD serves not only as an mRNA quality control system but also as a regulatory homeostatic mechanism to maintain the abundance of a broad class of physiological transcripts [137, 138, 139, 140, 141]. In the mechanism, referred to as regulated unproductive splicing and translation (RUST) [142, 143] or simply unproductive splicing [144, 145], the cell employs AS to produce PTC-containing transcript isoforms in order to post-transcriptionally downregulate the expression level of the gene [146, 144]. For example, regulated skipping of alternative exons 6 or 12 in *RBM10* transcripts and exons 6 or 16 in *RBM5* transcripts leads to the repression of the expression of *RBM10* and *RBM5*, respectively [135] (Fig 2-8). Similarly, *PTBP1*-promoted selection of a weak alternative 5' splice site in the *HPS1* gene stabilizes its expression in all cell types except neurons and muscle cells [136] (Fig 2-9). Unproductive splicing plays an essential role in normal and disease conditions, including early embryonic development [139], granulocyte differentiation [147], sta-

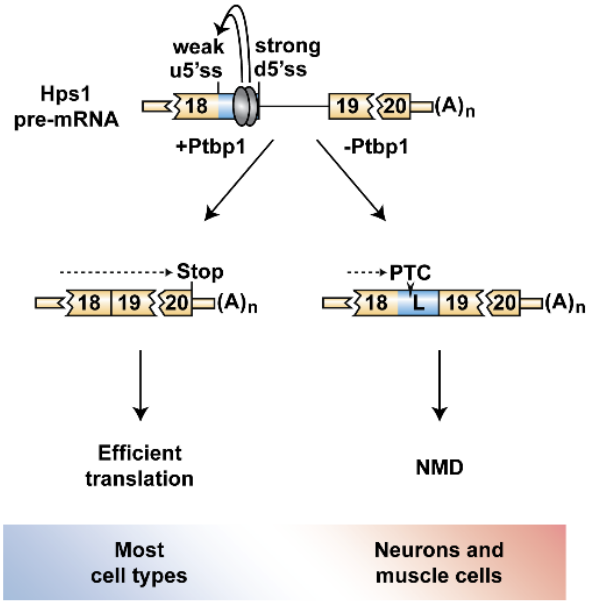

Figure 2-9: **The model of *HPS1* cross-regulation by *PTBP1* via unproductive splicing [136].**

bilization, and repression of oncogenic expression [148, 149].

The majority of USEs are described in genes encoding RBPs [150]. Many of them use unproductive splicing to autoregulate their expression levels in a negative feedback loop, in which the protein product binds its own pre-mRNA and causes alternative splicing to induce a PTC. This autoregulation takes place in almost all SR proteins [151], many hnRNP proteins [152, 153, 154, 155], spliceosome components [135, 156], and even in ribosomal proteins [157, 158]. Autoregulatory unproductive splicing is found in almost all eukaryotes studied to date and exhibits a high degree of evolutionary conservation [159, 160, 161].

Cross-regulatory unproductive splicing networks have a different hierarchical organization compared to transcriptional networks, with a few master regulators and many more regulatory connections among RBPs than between RBPs and other genes [143]. These connections have been characterized for many splicing factors, particularly for SR proteins, which coordinate their expression in a dense unproductive splicing network [151]. Cross-regulatory circuits among paralogs such as *PTBP1/PTBP2* [162, 152], *SRSF3/SRSF7* [163, 164], *RBM10/RBM5* [135], *RB-*

FOX2/RBFOX3 [165], *hnRNPD/ hnRNPDL* [155], and *hnRNPL/hnRNPLL* [166] are particularly abundant, but unproductive splicing also extends beyond RBPs and shapes the transcriptional landscape in other gene classes [150]. Remarkably, the relationship between the expression of the NMD isoform and the mRNA or protein levels in many cases is complex due to indirect connectivities in the regulatory network [154, 167, 168, 169]. Tissue specificity of unproductive splicing has been studied only for a handful of cases [150], including the regulation of neural-specific expression of the postsynaptic proteins *DLG4* and *GABBR1* that is controlled by *PTBP1* and *PTBP2* [170, 171, 172], which are also discussed here.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Thesis Objectives

The main goal of this project is to identify tissue-specific aberrant splicing events, including TASS and USEs, and predict mechanisms of their regulation. The aims of this project are subdivided into the following groups:

- genome-wide identification of human TASS and characterization of their tissue-specific expression;
- characterization of TASS evolutionary signatures and the impact of TASS on protein structure;
- prediction of TASS regulation by RBPs;
- characterization of tissue-specific expression and regulation of experimentally validated USEs;
- prediction of novel tissue-specific USEs and mechanisms of their regulation by RBPs;

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Materials and Methods

Throughout this work, the GRCh37 (hg19) primary assembly of the human genome sequence is used. It was downloaded from the UCSC genome browser [173].

4.1 Tandem alternative splice sites

4.1.1 The catalogue of TASS

The annotated splice sites

To identify the annotated splice sites, the internal boundaries of non-terminal exons were extracted from the comprehensive annotation of the GENCODE database v19 [174] and from UCSC RefSeq database [175]. The union of these sets yielded 569,694 annotated splice sites (Table A.1, A).

Expressed splice sites

The RNA-seq data from 8,548 samples in the GTEx Consortium v7 data was analyzed as before [176]. Short reads were mapped to the human genome using STAR aligner v2.4.2a in two-pass mode by the data providers allowing for the identification of both annotated and *de novo* splice junctions [177]. Split reads supporting splice junctions were extracted using the IPSA package with the default settings [178] (Shannon entropy threshold 1.5 bit). At least three split reads in at least two samples from different tissues were required to call the presence of a splice site.

Samples of EBV-transformed lymphocytes and transformed fibroblasts and samples with aberrantly high number of split reads were excluded. Only split reads with the canonical GT/AG dinucleotides were considered. Germline polymorphisms (SNPs, deletions and insertions) located within the splice site or within 35 nt of adjacent exonic regions were identified. Splice sites that were expressed exclusively in the samples, in which a polymorphism was present but absent in the other samples, were excluded to avoid split read misalignment caused by the discrepancy between the reference genome and the individual genotypes. This filtration removed 1.15% of expressed splice sites that were supported by 0.3% of the total number of split reads. As a result, 794,646 expressed splice sites were obtained (Table A.1, A).

Cryptic splice sites

The SpliceAI software [179] was used to scan the canonical transcriptome sequences and select splice sites with splice probability score greater than 0.1. According to the data provided by SpliceAI authors, at least 95% of exons having percent-spliced-in (Ψ) value (see below) below 0.1 are flanked by splice sites that fall below this score threshold. Splice sites that were previously called expressed or annotated were excluded resulting in a list of 607,639 cryptic splice sites (Table A.1, A).

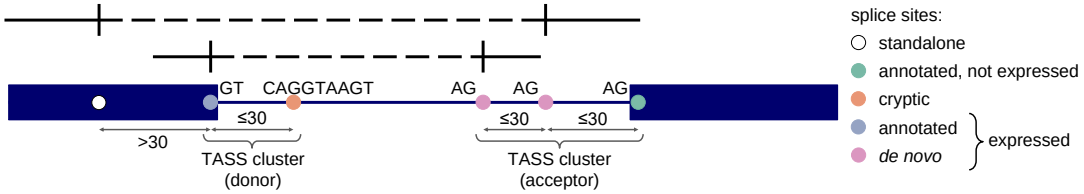

Figure 4-1: **TASS clusters.** Splice sites are categorized as annotated (GENCODE and Refseq), *de novo* (inferred from RNA-seq) or cryptic (detected by SpliceAI). TASS clusters consist of splice sites of the same type (donor or acceptor) such that each two consecutive ones are within 30 nt from each other.

4.1.2 TASS clusters

A TASS cluster was defined as a set of at least two splice sites of the same type (either donor or acceptor) such that each two successive splice sites are within 30 nt from each other (Fig 4-1). The number of splice sites in a TASS cluster will be referred to as cluster size. According to this definition, each splice site can belong either to a TASS cluster of size two or larger, or be a standalone splice site.

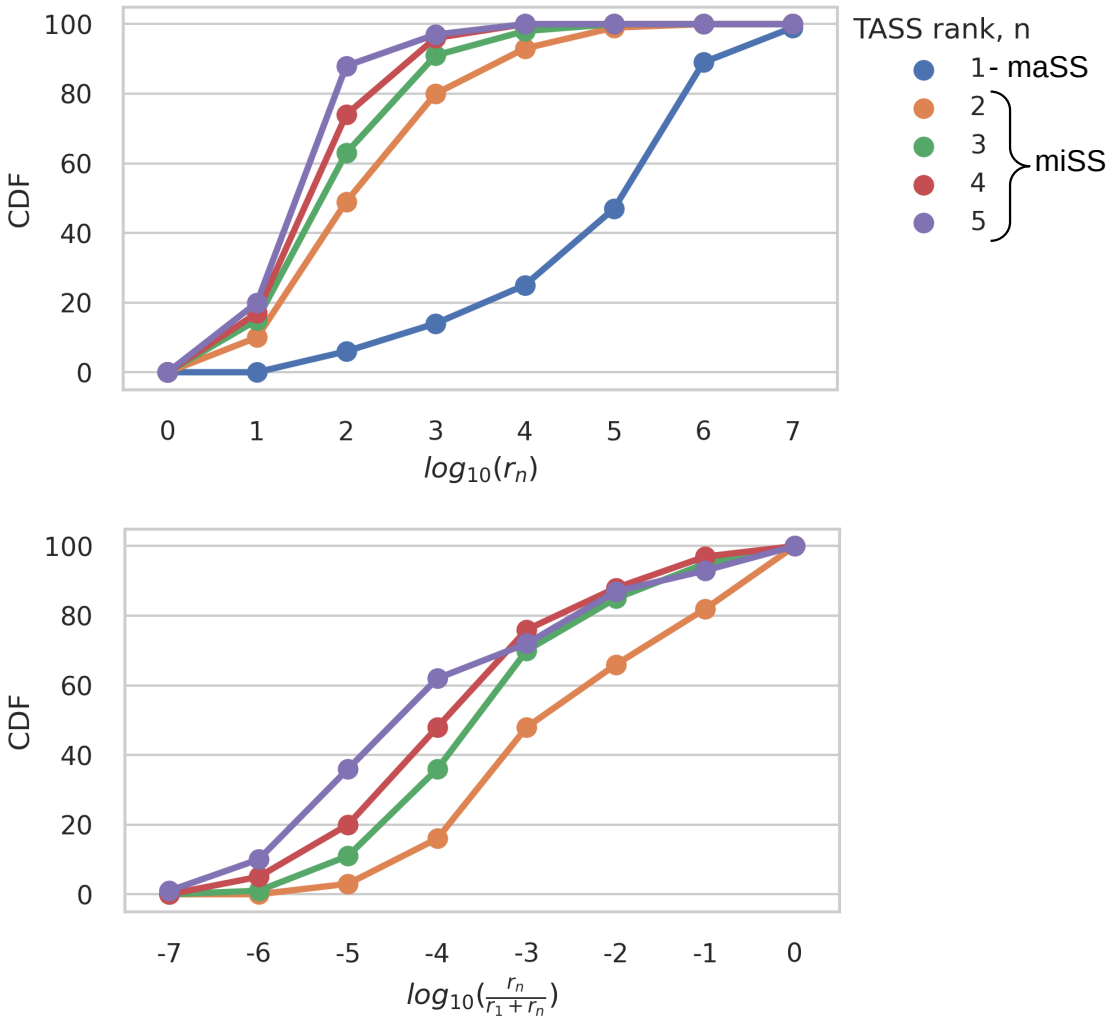

Figure 4-2: **maSS and miSS.** The expression of the major splice sites (maSS , i.e. rank 1) and minor splice sites (miSS, i.e. rank two or higher). Top: the cumulative distribution of r_n, the number of split reads supporting maSS and miSS. Bottom: the cumulative distribution of r_n relative to the sum of r_n and r_1.

A TASS cluster and all its constituent splice sites were categorized as coding

if the cluster contained at least one non-terminal boundary of a coding exon, and non-coding otherwise. Thus, non-coding splice sites are located in UTRs of protein-coding transcripts or in other transcript types such as long non-coding RNA. Splice sites were ranked based on the total number of supporting split reads. The splice site strength was assessed by MaxEntScan software [180], which computes a similarity metric between the splice site sequence and the consensus sequence. The higher MaxEnt scores correspond to splice site sequences that are closer to the consensus.

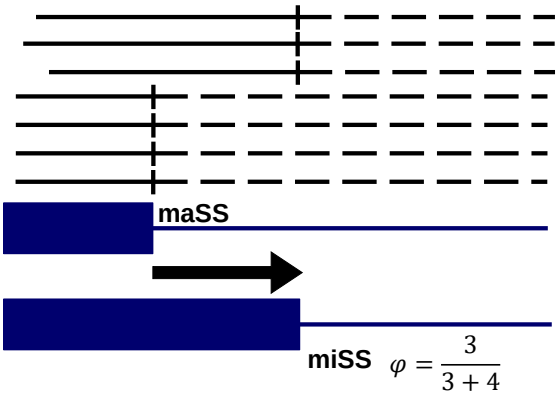

Figure 4-3: **The definition of the ϕ value exemplified.** A hypothetical maSS is supported by 4 split reads, while a hypothetical miSS is supported by 3 split reads, resulting in the ϕ value of 3/7.

4.1.3 Major and minor splice sites

After pooling together the read counts from all 8,548 GTEx samples, the splice sites within each TASS cluster were ranked by the number of supporting reads (Fig 4-2, top). The dominating splice site (rank 1, also referred to as major splice site, or maSS) is expressed at a substantially higher level compared to splice sites of rank 2 or higher (referred to as minor splice sites, or miSS); within TASS clusters, miSS are expressed several orders of magnitude weaker relative to maSS (Fig 4-2, bottom).

To quantify the relative usage of a miSS, I introduced the metric ϕ that takes into account only one end of the split read. It is defined as the number of split reads supporting a miSS as a fraction of the combined number of split reads supporting miSS and maSS. In contrast to the conventional percent-spliced-in (PSI, Ψ) metric

for exons [178], ϕ measures the expression of a miSS relative to that of the corresponding maSS and takes into account only one end of the supporting split read (Fig 4-3).

4.1.4 Response of TASS clusters to NMD inactivation

To assess the response of TASS clusters to the inactivation of NMD, I used RNA-seq data from the experiments on co-depletion of *UPF1* and *XRN1*, two key components of the NMD pathway [181]. Short reads were mapped to the human genome using STAR aligner v2.4.2a with the default settings. The read support of splice sites was called by IPSA pipeline as before (see section 4.1.1). TASS in which the major splice site was supported by less than 10 reads were discarded. The response of a miSS to NMD inactivation was measured by $\phi_{KD} - \phi_C$, where ϕ_{KD} is the relative expression in KD conditions and ϕ_C is the relative expression in the control.

4.1.5 Expression of miSS in human tissues

Significantly expressed (significant) miSS

The number of reads supporting a splice site can be used for presence/absence calls; however it depends on the local read coverage in the surrounding genomic region and on the total number of reads in the sample [182, 183]. A good proxy for these confounding factors is the number of reads supporting the corresponding maSS. Therefore, I quantified the expression of miSS relative to maSS and selected miSS that are expressed at significantly high level at the given maSS expression level, separately in each tissue. Since the number of reads often exhibits an excess of zeros, I treated the total number of reads supporting a miSS (r_{miSS}) in each tissue as a zero-inflated Poisson random variable with the parameters $(\hat{\pi}(r_{maSS}), \hat{\lambda}(r_{maSS}))$ which depend on the number of reads supporting the corresponding maSS (r_{maSS}), i.e.

$$\hat{\lambda} = a_0 r_{maSS}^{a_1} \tag{4.1}$$

$$\hat{\pi} = \text{logit}^{-1}(b_0 + b_1 r_{maSS}). \tag{4.2}$$

I estimated the parameters a_0, a_1, b_0, and b_1 separately in each tissue using zero-inflated Poisson (ZIP) regression model [184], computed the expected value of r_{miSS} for each miSS given the value of r_{maSS}, and assigned a P-value for each miSS as follows:

$$\text{P-value} = 1 - (CDF_{Poisson}(r_{min}, \hat{\lambda})(1 - \hat{\pi}) + \hat{\pi}). \tag{4.3}$$

To account for multiple testing, I converted the matrix of P-values for all miSS in all tissues to a linear array and estimated the false discovery rate by the Q-value method [185]. A miSS was called as significantly expressed (or, shortly, significant) if it had the Q-value below 5% and ϕ value greater than 0.05 in at least one tissue.

Tissue-specific miSS

The level of expression of a miSS relative to its corresponding maSS is reflected by the ϕ metric. To identify tissue-specific miSS among significantly expressed miSS, I analyzed the variability of the ϕ metric between and within tissues using the following linear regression model. For each significant miSS individually, r_{miSS} was modelled as a function of r_{maSS} by the equation

$$r_{miSS} = a_0 r_{maSS} + \sum_t a_t D_t r_{maSS}, \tag{4.4}$$

where D_t is a dummy variable corresponding to the tissue t. The slope a_t in this model can be interpreted as the change of the miSS relative usage in tissue t with respect to the tissue average, i.e.,

$$\widehat{\phi}_{tissue-average} = \frac{\hat{a}_0}{1+\hat{a}_0} \tag{4.5}$$

$$\widehat{\phi}_t = \frac{\hat{a}_0 + \hat{a}_t}{1+\hat{a}_0+\hat{a}_t} \tag{4.6}$$

$$\Delta\widehat{\phi}_t = \frac{\hat{a}_t}{(1+\hat{a}_0+\hat{a}_t)(1+\hat{a}_0)}. \tag{4.7}$$

The significance of tissue-specific changes of ϕ represented by a_t can also be estimated using this linear model. This allows assigning P-values (and Q-values) to a_t for each miSS in each tissue. In order to filter out significant, but not substantial changes of tissue-specific miSS expression, I required the Q-value corresponding to a_t be below 5% and the absolute value of $\Delta\hat{\phi}_t$ be above 5%; a miSS satisfying these conditions was called tissue-specific in the tissue t. A miSS was called tissue-specific if it was specific in at least one tissue. Additionally, the sign of a_t allows to distinguish upregulation ($a_t > 0$) or downregulation ($a_t < 0$) of a miSS in the tissue t.

4.1.6 Regulation of miSS by RBP

RNA-seq data from the experiments on the depletion of 248 RBPs in two human cell lines (K562 and HepG2) were downloaded from ENCODE portal website in BAM format [186]. Short reads were mapped to the human genome using STAR aligner v2.4.0k [177]. Out of 248 RBPs, I left only those for which eight samples were present: two KD and two control samples for each of the two analyzed cell lines. Additionally, I required the presence of at least one publicly available enhanced crosslinking and immunoprecipitation (eCLIP) experiment [187] for each RBP. This confined the list of potential regulators to 103 RBPs (Table A.2).

I used rMATS-turbo v.4.1.0 [188] in novelSS mode to identify both novel and annotated alternative splicing events between KD and control samples for each RBP in each cell line. The minimum intron length and the maximum exon length were set to 10 and 1000, respectively. Since the definition of Ψ value for alternative donor and acceptor splice sites in rMATS pipeline corresponds to the definition of ϕ value,

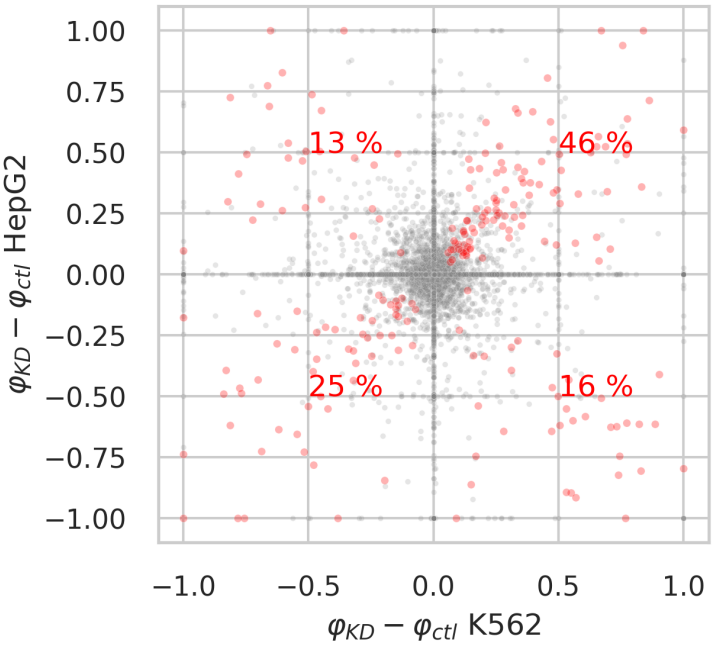

Figure 4-4: **The response of a miSS to inactivation of an RBP in K562 (x-axis) and HepG2 (y-axis) cell lines.** Fractions of significant miSS-RBP pairs in each quadrant are shown (the fractions are summed to 100%).

I used the rMATS output to directly extract $\Delta\phi_{KD}$ values and P-values. I obtained Q-values for 9,303 significantly expressed miSS in each RBP and cell line. In order to filter out significant, but not substantial changes of miSS expression between KD and control samples, I required the Q-value be below 5% and the absolute value of $\Delta\phi_{KD}$ be above 0.05 in both HepG2 and K562 cell lines. As a result, I obtained 221 significant RBP-miSS pairs, of which 65 pairs (29%) showed a discordant response to KD between cell lines (Fig 4-4). These cases were excluded, and 156 RBP-miSS pairs (101 pairs with $\Delta\phi_{KD} > 0$ and 55 pairs with $\Delta\phi_{KD} < 0$) were kept for downstream analysis of miSS-RBP-tissue triples.

The gene read counts data was downloaded from GTEx (v7) portal on 08/05/2020 [189] and processed by DESeq2 package using apeglm shrinkage correction [190]. Differential expression analysis was done for each tissue against all other tissues. The P-values for 103 RBPs in each tissue were adjusted for FDR using Q-value [185]. An RBP was classified as tissue-specific if the Q-value in the corresponding tissue was below 5% and the absolute value of $\log_2$ fold change was larger than 0.5. A

tissue-specific RBP was considered upregulated in tissue t ($\Delta RBP_t > 0$) if the $\log_2$ fold change value was positive and downregulated ($\Delta RBP_t < 0$) if the $\log_2$ fold change value was negative. As a result, I obtained 1,115 RBP-tissue pairs (388 upregulated pairs and 727 downregulated pairs).

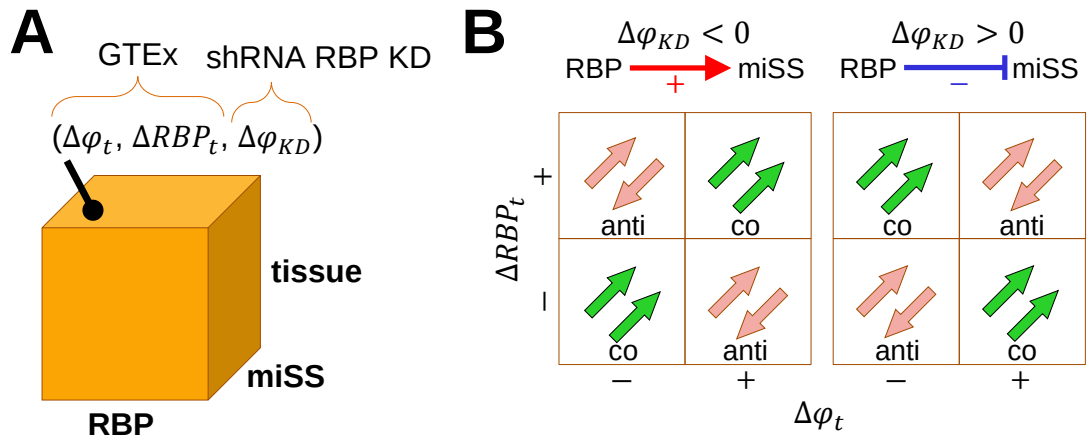

Figure 4-5: **Characterization of miSS-RBP-tissue triples (A)** Each miSS-RBP-tissue triple is characterized by three metrics: ($\Delta\phi_t$, the change of miSS relative usage in tissue t; ΔRBP_t, the change of the RBP expression in tissue t; and $\Delta\phi_{KD}$, the change of miSS relative usage upon inactivation of RBP by shRNA-KD. **(B)** The response of miSS to RBP inactivation defines activating ($\Delta\phi_{KD} < 0$, red) and repressing ($\Delta\phi_{KD} > 0$, blue) regulation, which together with other metrics define co-directed and anti-directed triples.

I obtained 14,005 miSS-tissue pairs (6,265 upregulated pairs and 7,740 downregulated pairs) in the analysis of tissue-specific expression of miSS. The intersection of this set with RBP-tissue pairs and RBP-miSS pairs resulted in a list of 256 miSS-RBP-tissue triples that were characterized by three parameters, $\Delta\phi_t$, ΔRBP_t, and $\Delta\phi_{KD}$, where $\Delta\phi_t$ is the change of the miSS relative usage in the tissue t, ΔRBP_t is the change of the RBP expression in the tissue t, and $\Delta\phi_{KD}$ is the response of miSS to RBP inactivation by shRNA-KD (Fig 4-5, A). I classified a miSS-RBP-tissue triple as co-directed if the correlation between RBP and miSS expression was concordant with the expected direction of miSS expression changes from shRNA-KD (e.g., if $\Delta\phi_t > 0$, $\Delta RBP_t > 0$ and $\Delta\phi_{KD} < 0$) and anti-directed otherwise (Fig 4-5, B). That is, in co-directed triples the direction of regulation in the observed correlation and in the shRNA-KD coincide, and in anti-directed triples they

are opposite.

The eCLIP peaks, which were called from the raw data by the producers, were downloaded from ENCODE data repository in bed format [191, 192]. The peaks in two immortalized human cell lines, K562 and HepG2, were filtered by the condition $\log FC \geq 3$ and P-value < 0.001 as recommended [187]. Since the agreement between peaks in the two replicates was moderate (the median Jaccard distance 25% and 28% in K562 and HepG2, respectively), I took the union of peaks between the two replicates in two cell lines, and then pooled the resulting peaks. The presence of eCLIP peaks was assessed in the ± 20 nt vicinity of a miSS position.

I downloaded the *PTBP1* overexpression data [193] (2 full-length *PTBP1* overexpression samples, 4 control samples) from NCBI SRA archive in FASTQ format under the accession number SRP059242. As before, short reads were mapped to the human genome using STAR aligner v2.4.2a with the default settings. I used rMATS-turbo v.4.1.0 with the same approach as I used for shRNA-KD data to infer $\Delta\phi_{PTBP1-OE}$ values and associated P-values and Q-values for 9,303 significantly expressed miSS.

4.1.7 Expression and regulation of miSS in primary cells

Primary cell transcriptome data (94 RNA-seq experiments) from 19 tissues of origin were downloaded from ENCODE portal website in BAM format [194, 186]. Each sample was assigned to one of the 9 cell types (mesenchymal smooth muscle cells, endothelial cells, epithelial cells, cardiomyocytes, fibroblasts, melanocytes, stem cells, preadipocytes, skeletal muscle cells) according to metadata. Short reads were mapped to the human genome by the data providers using STAR aligner v2.3.1z [177]. The read support of splice sites was called by IPSA pipeline as before (see section 4.1.1). The identification of cell-type-specific and tissue-of-origin-specific miSS was done using linear regression as before (see section 4.1.5). The ϕ values were calculated for 9,303 significantly expressed miSS in each sample requiring at least one of r_{miSS} and r_{maSS} values to be greater than 20 for positive ϕ values and substituting the ϕ values with zero otherwise. Pearson correlation coefficient was used as a measure of similarity of miSS expression profiles. Gene expression pro-

files were assessed by the data providers using RSEM v.1.2.19 [195]. Read counts were library size-corrected using the `DESeq2` package [196]. From the gene set, I selected 103 RBPs introduced before (see section 4.1.6). MiSS-RBP-cell type triples and miSS-RBP-tissue triples were obtained as before by merging PROMO miSS and RBP expression data with the responses of miSS to shRNA-KD of RBP. A miSS-RBP-cell type (miSS-RBP-tissue) triple was defined co-directed if the sign of the Spearman correlation coefficient of miSS expression and RBP expression within samples of this cell type (tissue) coincided with the expected sign of correlation inferred from shRNA-KD of RBP. For example, if the miSS is upregulated in the KD of RBP, the expected sign of the correlation would be negative.

4.1.8 Evidence of miSS translation in Ribo-Seq data

The global aggregate track of Ribo-Seq profiling, which tabulates the total number of footprint reads that align to the A-site of the elongating ribosome, was downloaded in BIGWIG format from GWIPS-viz Ribo-Seq genome browser [197]. It was intersected with TASS coordinates to obtain position-wise Ribo-Seq signal for miSS and maSS. The analysis was carried out on the intronic miSS in TASS clusters of size two. For each miSS, the relative Ribo-Seq support was calculated as

$$RS = \frac{\#reads_{miSS}}{\#reads_{miSS} + \#reads_{maSS}}, \tag{4.8}$$

where $\#reads_{miSS}$ and $\#reads_{maSS}$ are the number of Ribo-Seq reads supporting the first exonic nucleotide of miSS and maSS, respectively. Higher values of RS indicate stronger evidence of translation.

4.1.9 Structural annotation of miSS

All amino acids that are lost or gained due to using miSS instead of maSS were structurally annotated with respect to their spatial location in protein three-dimensional structure using StructMAn [198]. As a control, all amino acids in all isoforms of the human proteome were structurally annotated as well. Briefly, the procedure of structural annotation consists in mapping a particular amino acid onto all experi-

mentally resolved three-dimensional structures of proteins that are homologous to a given human isoform. The mapping is done by means of pairwise alignment of the respective protein sequences. Then, the spatial location of the corresponding amino acid residue in the structure is analyzed in terms of proximity to other interaction partners (other proteins, nucleic acids, ligands, metal ions) and propensity to be exposed to the solvent or be buried in the protein core. Such annotations from different homologous proteins are then combined while taking into account sequence similarity between the query human isoform and the proteins with the resolved structures, alignment coverage, and the quality of the experimental structure. This procedure resulted in structural annotations for 23,095,050 amino acids from 88,573 protein isoforms.

Further, a correspondence between 86,647 UniProt protein identifiers and 106,403 ENSEMBL transcripts identifiers was established, discarding 3,194 transcripts that had ambiguous mappings [199]. Custom scripts were used to map 23,095,050 amino acids within structural annotation of UniProt entries to the human transcriptome and, furthermore, to the human genome using ENSEMBL transcript annotation. This procedure yielded 17,093,614 non-redundant genomic positions since some UniProt entries correspond to alternative isoforms of the same protein, and thus some amino acids from different entries can map to the same nucleotide in the genome. At that, positions that had ambiguous structural annotation from different transcripts were discarded.

Unlike maSS and exonic miSS, most of the intronic miSS are located outside of ENSEMBL transcripts and thus can not be directly classified based on the structural annotation. However, the structural annotation of exonic miSS coincides with that of the respective maSS in most cases (Fig A-1). I therefore assumed that the short distance between maSS and miSS allows to assign the structural annotation of the first exonic nucleotide of a maSS to all miSS including miSS located in introns. Under this assumption, the structural annotation was defined for 6,879 out of 12,667 frame-preserving expressed miSS in the coding regions.

Protein coordinates of predicted short linear motifs (SLiMs) were extracted from the ELM database [200] and mapped to genomic coordinates as described above.

Regions between maSS and miSS (miSS indels) were compared with nearby exonic regions defined as the regions of the same length as miSS indels but located in the adjacent exons on the distance equal to the indel length. A SLiM is recognized to overlap with a particular region (miSS indel of nearby exonic region) if its genomic projection overlaps at least one exonic nucleotide.

Homology modeling of 3D protein structures was performed in I-TASSER web server with default parameters [201]. I used the FoldX Stability tool [202] to assess the total energy required to fold the proteins from their unfolded state with default parameters. Computational alanine-scanning mutagenesis was performed in BAlaS web server [203] with default parameters.

4.1.10 Evolutionary selection of miSS

Splice sites of annotated human transcripts were extracted from the comprehensive annotation of the human transcriptome (GENCODE v19 and NCBI RefSeq) using custom scripts [204, 175]. Internal boundaries of non-terminal exons (excluding splice sites overlapping with TASS clusters) were classified as constitutive splice sites if they were used as splice sites in all annotated transcripts. Position weight matrices were used to build consensus sequences for donor and acceptor constitutive splice sites as described in [205, 206]. Orthologs of the annotated human splice sites were identified in multiple sequence alignment of 46 vertebrate genomes with the human genome (GRCh37), which was downloaded from the UCSC Genome Browser in MAF format [173]. The alignments with marmoset and galago (bushbaby) genomes were extracted from MAF, and the alignment blocks were concatenated. The genomic sequence of splice sites in the common ancestor of human and marmoset with galago as an outgroup was reconstructed by parsimony [207]. Only splice sites with the canonical GT/AG dinucleotides in all three genomes were considered. The analysis was further confined to TASS clusters of size two, in which only intronic miSS were considered to avoid the confounding effect of selection acting on the coding sequence in exonic miSS. This procedure resulted in 34,550 TASS (17,275 maSS and 17,275 miSS) in the coding regions.

To estimate the strength of evolutionary selection acting on the consensus (Cn)

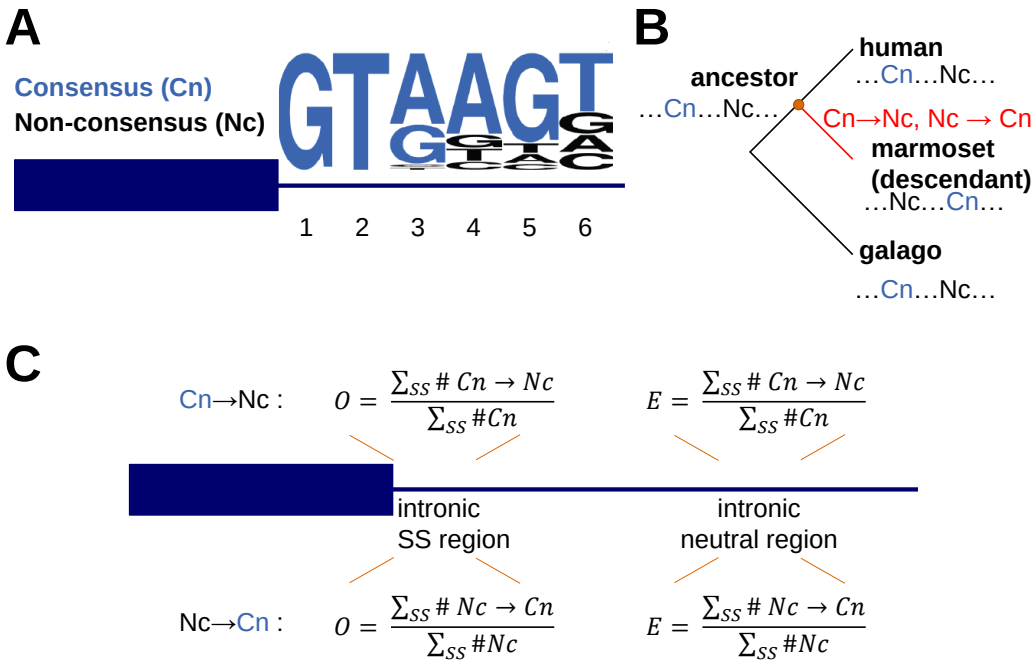

Figure 4-6: **Evolutionary selection of miSS. (A)** The definition of the consensus (Cn) and non-consensus (Nc) nucleotide variants in the donor splice site. The definition for acceptor splice site is similar. **(B)** The evolutionary tree used to reconstruct the ancestral sequence of human and marmoset. **(C)** The computation of O and E statistics.

and non-consensus (Nc) nucleotides (Fig 4-6, A), a previously developed method was used with several modifications [208]. The frequency of Cn-to-Nc (or Nc-to-Cn) substitutions at different positions relative to the splice site (observed, O) were compared with the background frequencies of the corresponding substitutions in neutrally-evolving intronic regions (expected, E) (Fig 4-6, B, C). The ratio of observed to expected (O/E) equal to one indicates neutral evolution (no selection); $O/E > 1$ indicates positive selection; $O/E < 1$ indicates negative selection. At that, only intronic positions from the positions +3 to +6 for the donor splice sites and positions from -24 to -3 for the acceptor splice sites were considered (the canonical GT/AG dinucleotides were excluded as they were required to be conserved). The substitution counts were summed over all positions in these ranges. Furthermore, splice sites from the human genome were mapped onto the ancestral genome us-

ing MAF alignments but the substitutions were analyzed in the marmoset lineage, where the substitutions process goes independently from the human lineage (Fig 4-6, B). This approach mitigates the systematic underrepresentation of Cn-to-Nc substitutions and the overrepresentation of Nc-to-Cn substitutions in the human lineage leading to artificial signs of strong positive and negative selection in cryptic and not significant miSS (Fig A-2) [208]. Constitutive splice sites were matched to maSS (miSS) by the ancestral strength using random sampling from the set of constitutive splice sites without replacement and requiring the strength difference not to be larger than 0.01.

4.1.11 Allele frequencies of SNPs in the vicinity of miSS

Germline SNPs located within [-35 nt, +6 nt] for the donor miSS and within [-21 nt, +35 nt] for the acceptor miSS were identified in GTEx and 1000 Genomes [108] data. Allele frequencies of the SNPs were obtained using vcftools [209] and custom scripts. For comparison of allele frequencies between different miSS categories, the maximum allele frequency of SNPs related to each of the miSS was calculated.

4.1.12 Mixture model for the estimation of the fraction of noisy miSS

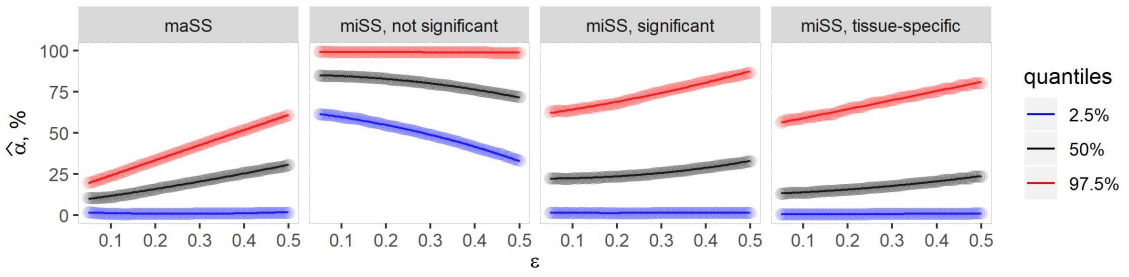

Figure 4-7: **Estimation of the 95% confidence interval of α for different expression categories of miSS.**

The mixture model to estimate the fraction of noisy splice sites was constructed as follows. Denote by k the size of the sample of interest (tissue-specific miSS,

significant non-tissue-specific miSS, non-significantly expressed miSS, or maSS). It is assumed that the sample of interest is a mixture of two subsamples, αk splice sites from the negative set (cryptic splice sites, which demonstrate no evidence of selection, Fig A-2) and $(1-\alpha)k$ splice sites from the positive set (all constitutive splice sites). For every α in the range from 0 to 1 with the step 0.0033, I sample randomly αk elements from the negative set and $(1-\alpha)k$ elements from the positive set 300 times and construct the joint frequency distribution of α and O/E. To obtain the marginal (conditional) distribution corresponding to the observed value of O/E in the actual set of interest, I use an infinitesimal margin ϵ to compute the empirical probability density in $(O/E-\epsilon, O/E+\epsilon)$, and take the limit $\epsilon \to 0$ using the linear regression model $p = \beta_0 + \beta_1 + \epsilon$. The quantiles were calculated for every ϵ in the range from 0.025 to 0.5 with the step 0.005 (Fig 4-7). The interval estimates of α are inferred from the 2.5% and 97.5% quantiles.

4.2 Unproductive splicing

4.2.1 Unproductive splicing events

In this part, the GRCh37 (hg19) assembly of the human genome was used along with the comprehensive gene annotation (v35lift37) from GENCODE [174]. The coordinates of exons, introns, and stop codons were extracted from transcripts labeled as protein-coding and NMD to identify AS event types shown in Fig 4-8. Among them, USEs were characterized by PTCs that were located within or downstream of the AS event, with the exception of poison exons and intron detention that can trigger NMD by inducing splice junctions in 3' UTR. As a result, I obtained a stringent set of 5,309 USEs (Table A.3) and 38,396 AS events that didn't generate an NMD isoform.

4.2.2 Quantification of AS

STAR v2.4.2a alignments [177] of short reads in 8551 samples from the GTEx Consortium v7 [176] (excluding samples from transformed cells and testis) were obtained

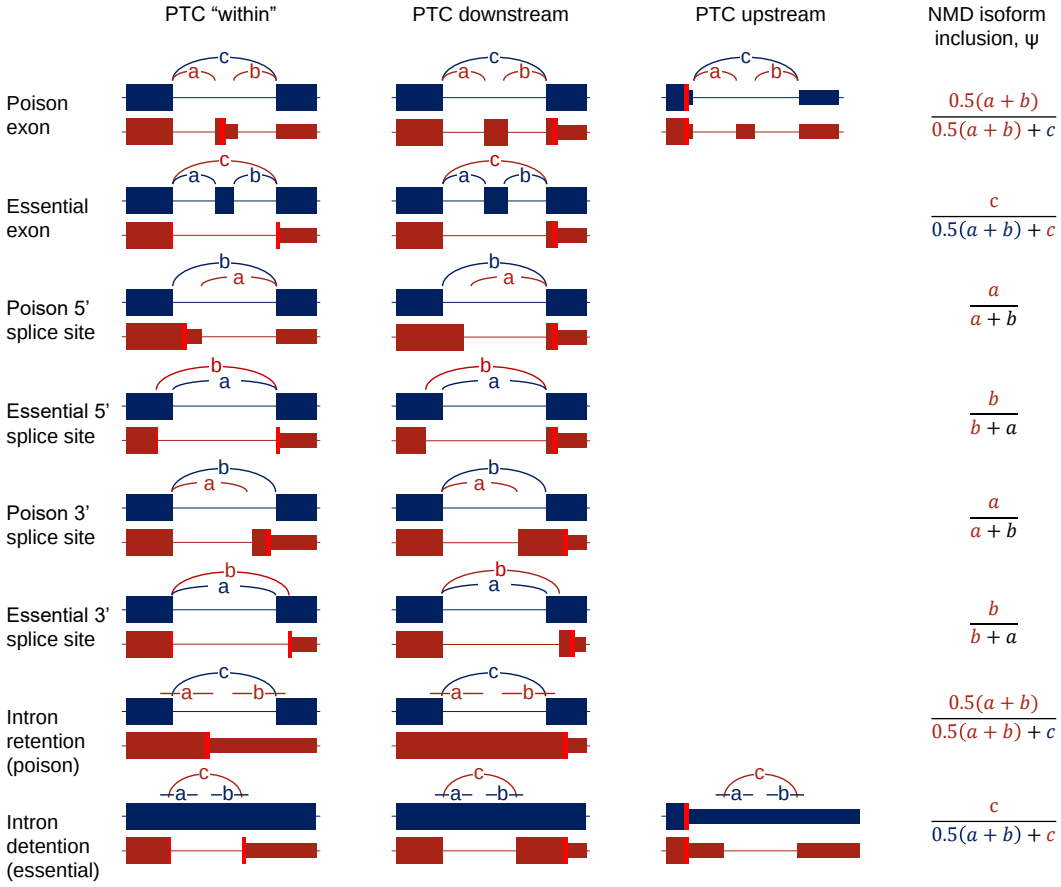

Figure 4-8: **USE classes.** An USE is referred to as poison/essential if the NMD isoform (red) represents an insertion/deletion compared to the protein-coding isoform (blue). PTCs are colored in bright red.

from dbGaP website in BAM format under the accession number phs000424/GRU. Split reads supporting splice junctions as well as non-split reads supporting intron retention were extracted from BAM files using the IPSA package with the default settings [178].

In the GTEx samples, the splicing rate was estimated from split-read counts computed by the IPSA pipeline using the ψ (percent-spliced-in, PSI) metric, defined as the number of reads supporting the NMD isoform as a fraction of the combined number of reads supporting the NMD isoform and the protein-coding isoform (Fig 4-8). For uniformity, the ψ metric was defined with respect to the NMD isoform, i.e., $\psi = 1$ for a poison exon assumes that it is 100% included, but $\psi = 1$ for an essential exon assumes that it is 100% skipped. Only ψ values with the denominator of at

least 15 were considered reliable. Further, I discarded AS events if the lower and the upper quartiles of the ψ distribution in GTEx samples coincided, or if ψ values were missing in more than 80% of samples of a particular tissue.

The results of RBP perturbation experiments followed by RNA-seq (Table A.4), including knockdown (KD), knockout (KO), and overexpression (OE), and the results of NMD inactivation experiments (Table A.5) were downloaded from the ENCODE portal and the SRA archive in BAM and FASTQ formats [210]. Short reads in FASTQ files were mapped to the GRCh37 assembly of the human genome using STAR-2.7.7a aligner in two-pass mode with the following options:
--outSJfilterOverhangMin 20 8 20 20
--alignSJDBoverhangMin 8
--outFilterMismatchNmax 999
--outFilterMismatchNoverReadLmax 0.1.
In differential splicing analysis, BAM files were processed with rMATS v.4.1.1 [188] with the following parameters: --novelSS --mil 10 --mel 10000 --libType fr-unstranded --variable-read-length. Only JC output files were used, i.e. differential splicing analysis did not involve reads mapped within exons. In NMD inactivation experiments, I adjusted rMATS P-values for testing multiple USEs using the Benjamini-Hochberg method (FDR) and recognized an USE as targeted by NMD if $\Delta\psi$ was significantly greater than zero at FDR < 0.05.

4.2.3 Gene expression quantification and analysis

In the GTEx samples, the local gene expression level of an USE was defined as the denominator of its ψ ratio, which reflects the abundance of short reads in the vicinity of the AS event and allows estimation of gene expression changes that are independent of AS acting globally in the gene, however at the cost of higher stochasticity. In contrast, the global gene expression level was defined as the total number of reads mapped to a gene that harbors an USE (counts obtained from the GTEx portal). Both local and global gene expression values were normalized according to the DESeq2 methodology with a pseudocount of 8 [196]. Namely, each row of the expression matrix, whose columns correspond to samples and rows correspond

to genes, was divided by the row median. The size factor sf_k of the sample k was estimated as the median of the k-th column. Each gene expression value was divided by sf_k and $\log_2$-transformed. The resulting expression values were centered by subtracting the median over all samples. The normalized local and global gene expression levels are denoted by e_l and e_g, respectively.

To evaluate differential gene expression in the RBP perturbation and NMD inactivation experiments, the read counts were extracted from the respective BAM files using RNASeQC utility [211] and processed with DESeq2 package [196] with apeglm shrinkage correction [190]. In each RBP perturbation experiment, I checked that RBP expression indeed changed upon perturbation in the intended direction, i.e. decreased upon KD or KO and increased upon OE; otherwise, the experiment was excluded from the analysis.

4.2.4 Unproductive splicing and gene expression

To detect the decrease of gene expression that was accompanied by the activation of the NMD isoform in GTEx, I compared gene expression levels in the 25% of samples with the highest ψ value (the upper quartile) vs. the 25% of samples with the lowest ψ value (the lower quartile) for each USE. Namely, I estimated the medians of e_l and e_g in these sample groups and compared them using the Mann-Whitney sum of ranks test with continuity correction. The statistical significance of the differences in e_l and e_g between the two groups (denoted as Δe_l and Δe_g) was interpreted by z-scores associated to the U-statistic of the test.

Next, I determined the features that distinguish USEs from protein-coding AS events. First, I selected the events, in which both e_l and e_g were positive (positive set) and those, in which both e_l and e_g were negative (negative set). In most cases, the negative set was substantially larger among USEs compared to protein-coding AS events (Fig A-3, A). Next, I compared the distributions of five parameters in the negative set for USEs and protein-coding AS events using the positive set as a control (Fig A-3, B-I). I found that the most remarkable feature that distinguished USEs and protein-coding AS events was the z-score of the global gene expression level. Additionally, I checked that gene expression levels don't show a significant

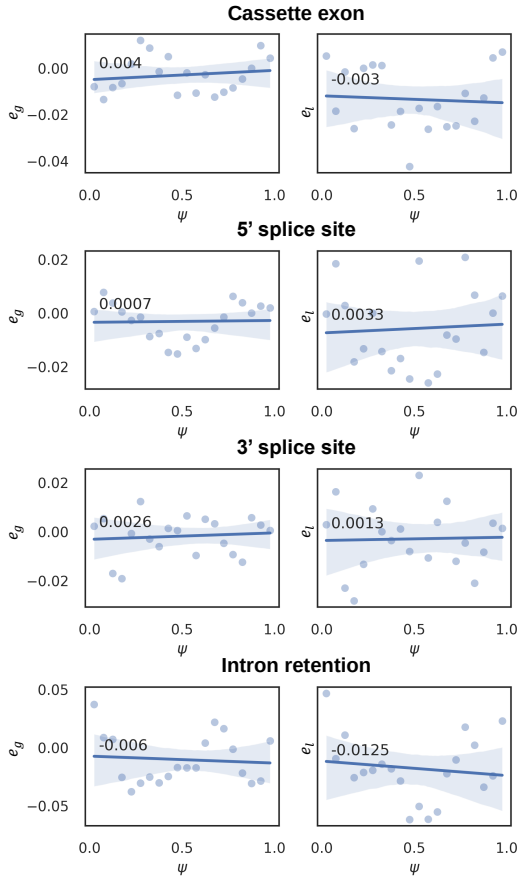

Figure 4-9: **The dependence of the median e_g and e_l values on ψ in protein-coding AS events.** The scatter plots represent median e_g (left) and e_l (right) values in genes hosting AS events, which were subdivided into 20 bins of equal size according to their values. None of the AS classes shows a significant dependence of median e_g and e_l values on ψ (FDR ≥ 0.05).

dependence on ψ value for AS events in protein-coding genes (Fig 4-9).

4.2.5 Tissue specificity of USEs

Each USE was characterized by three parameters, ψ, e_l, and e_g. In each sample, I computed the deviations of these three metrics from their respective medians across all GTEx samples (pooled medians) and performed a sign test ($H_0 : p = 0.5$) in each tissue to statistically evaluate the number of positive and negative deviations. As a result, I obtained P-values adjusted for testing of multiple tissues using the Benjamini-Hochberg false discovery rate (FDR) and the sign of the deviation of the

tissue median from the pooled median for ψ, e_l, and e_g. After discarding the tissues, in which e_l and e_g had opposite signs, I selected tissues, in which the deviations of ψ, e_l, and e_g were significant (FDR < 0.05) and categorized them into two groups by ψ and e_g having the same or the opposite signs. One-sided departures from the null hypothesis that deviations with the same and the opposite signs are equally likely were assessed using a binomial test, and the resulting P-values were adjusted for testing multiple USEs using the Benjamini-Hochberg FDR. USEs with FDR < 0.1 were considered tissue-specific.

4.2.6 Identification of regulators in tissue-specific USEs

The assessment of differential splicing by rMATS was carried out separately for each RBP perturbation experiment (Table A.4). For uniformity, I used $\Delta\psi$ values reflecting the difference in the NMD isoform splicing rates between the high RBP condition and the low RBP condition (Control vs. KD, Control vs. KO, and OE vs. control). Only ψ estimates with the read support of at least 15 and $|\Delta\psi| \geq 0.1$ were considered. Consequently, each USE was characterized by $\Delta\psi$ values, one for each perturbation experiment, and their respective P-values corrected for multiple testing using the Benjamini-Hochberg FDR. Similarly, each USE was characterized by the global gene expression changes Δe_g, one for each perturbation experiment, and their associated P-values adjusted for multiple testing.

To account for the fact that some RBPs had only one perturbation experiment, while others had many, the values of $\Delta\psi$ and Δe_g were converted to scores ± 1.5, ± 1, and 0, where the sign corresponds to the sign of $\Delta\psi$ and Δe_g, and the absolute values 1.5, 1, and 0 correspond to significant differences (FDR < 0.05), insignificant differences (FDR ≥ 0.05), and discarded values, respectively. These scores were averaged over all perturbation experiments for each RBP-USE pair. An RBP was recognized as a candidate regulator of an USE if the average $\Delta\psi$ and Δe_g scores had opposite signs. Additionally, I tested whether the tissue-specific expression profile of the predicted regulator was consistent with that of the USE using the same strategy as in the analysis of tissue specificity, but comparing ψ changes with the regulator expression changes instead of the changes in expression of the host gene.

4.2.7 RBP footprinting data

The complete POSTAR3 database of processed crosslinking and immunoprecipitation (CLIP) experiments in tabular format was kindly provided by the authors [212]. An RBP-USE pair was considered as having CLIP support in the gene if the RBP had at least one peak obtained by Piranha or eCLIP pipelines within the gene that harbors the USE, the boundaries of which were determined by the GENCODE annotation. An RBP-USE pair was further considered as having local CLIP support if a Piranha or eCLIP peak occurred within the intron containing the USE that was extended by 20 nt into the flanking exons.

4.2.8 Proteomic data

The data on protein expression in human tissues, which was measured by the precursor intensity in mass spectrometry (MS1-level), was downloaded from the Proteomics DB portal [213]. To test for negative associations between the NMD isoform splicing rate and the expression of the host gene product at the protein level, I matched the tissues from Proteomics DB to GTEx tissues (Table A.6), computed the median ψ values for Proteomics DB tissues, and sorted the tissues in ascending order by the median ψ. The one-sided Jonckheere trend test was applied to check whether protein expression levels follow a descending trend. Rejection of the null hypothesis in such a test indicated a negative association between ψ and the protein expression level.

4.3 Statistical analysis

The data were analyzed using python version 3.8.2 and R statistics software version 3.6.3. Non-parametric tests were performed with the scipy.stats python package using normal approximation with continuity correction. MW denotes Mann-Whitney sum of ranks test. Error bars in all figures and the numbers after the $\pm$ sign represent 95% confidence intervals. Jonckheere trend test was performed with the clinfun R package. Levels of significance 0.05, 0.01, 0.001 are denoted as *, **, ***.

Chapter 5

Tandem alternative splice sites

Tandem alternative splice sites (TASS) is a special class of alternative splicing events that are characterized by a close tandem arrangement of splice sites. The biggest existing catalogue of TASS, TASSDB2, is based on the evidence of transcript isoforms from expressed sequence tags (EST) [214]. The advances of high-throughput sequencing technology open new possibilities to identify and characterize TASS [215]. Here, I revisit the catalogue of TASS by analyzing a large array of RNA-seq samples from the Genotype Tissue Expression (GTEx) project [176]. I substantially extend the existing catalogue of TASS, characterize their common genomic features, and systematically describe a large set of TASS that have functional signatures such as evolutionary selection, tissue-specificity, impact on protein structure, and regulation by RBPs. While it is believed that the expression of TASS primarily originates from splicing noise, here I show that a number of previously unknown TASS may have important physiological functions and estimate the proportion of noisy splicing of TASS. The TASS catalogue is available through a track hub for the Genome Browser (see appendix A.1).

5.1 The catalogue of TASS

In order to identify TASS, I combined three sources of data. First, I extracted the annotated splice sites from GENCODE and NCBI RefSeq human transcriptome annotations [204, 175]. This resulted in a list of ∼570k splice sites, which

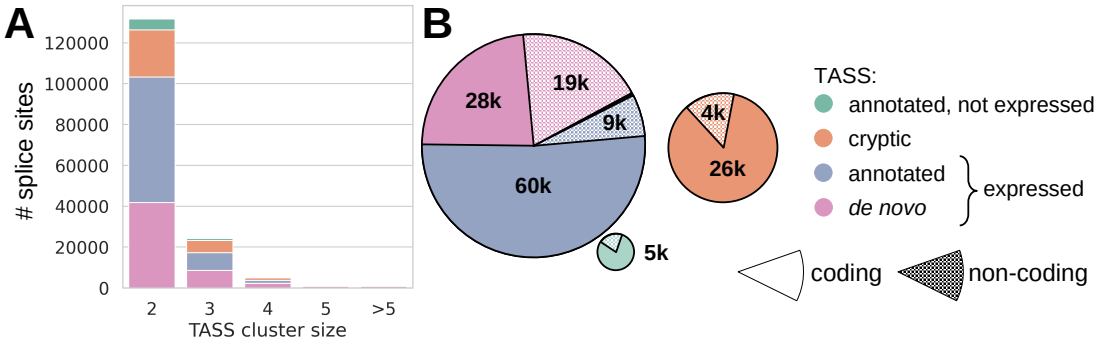

Figure 5-1: **The abundance of TASS. (A)** The TASS cluster size distribution. **(B)** The number of annotated, *de novo* and cryptic TASS in coding and non-coding regions.

will be referred to as annotated. Next, I identified donor and acceptor splice sites in split read alignments from RNA-seq experiments in the Genotype Tissue Expression Project (GTEx) [176] by pooling together its 8,548 samples. This resulted in a list of ∼800k splice sites, which will be referred to as expressed. A splice site may belong to both these categories, i.e. be annotated and expressed, or be annotated and not expressed, or be expressed and not annotated (the latter are referred to as *de novo*). Third, I scanned the human genome sequence with SpliceAI software [179] and selected splice sites with SpliceAI score greater than 0.1 excluding splice sites that were previously called expressed or annotated. This resulted in a list of ∼600k sequences that are similar to splice sites, but have no evidence of expression or annotation and will therefore be referred to as cryptic. The combined list of splice sites from all three sources contained approximately one million unique splice sites (Table A.1, A). According to the presented definition (see chapter 4), ∼177k of them were found to be located in TASS clusters (Table A.1, B).

About 99% of splice sites in TASS clusters are located in clusters of size 2, 3, 4 and 5 (Fig 5-1, A). In what follows, I confined the analysis to TASS clusters consisting of five or fewer splice sites and having at least one expressed splice site (Table A.1, C). This way I obtained ∼151k splice sites; among them ∼69k (46%) expressed annotated splice sites, ∼47k (31%) expressed *de novo* splice sites, ∼5k (3%) annotated splice sites that are not expressed, and ∼30k (20%) cryptic splice sites (Fig 5-1, B). I categorized a TASS cluster as coding if it contained at least one non-terminal

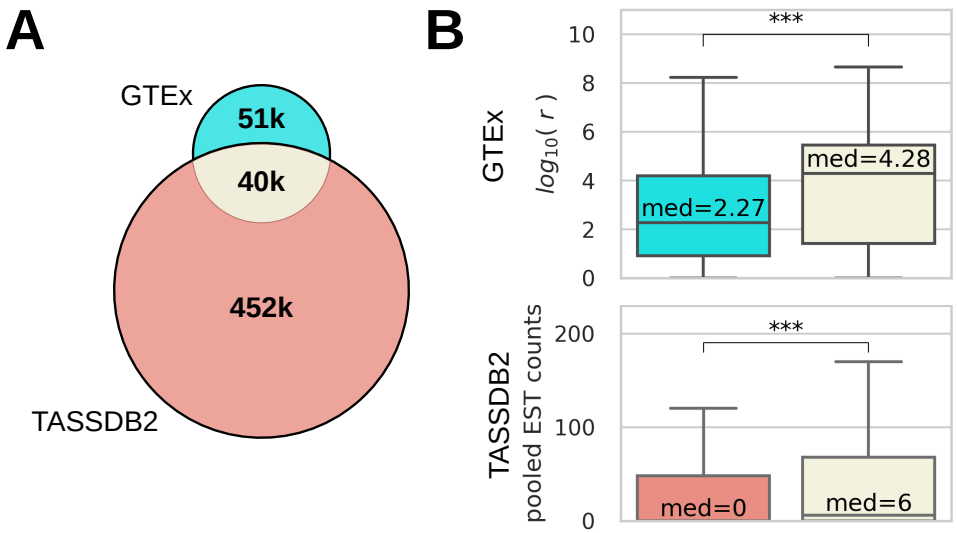

Figure 5-2: **A comparison of the TASS catalogue with the TASSDB2 database.** **(A)** The catalogue of TASS and TASSDB2 database. Only TASS separated by 2–12 nt were counted to match the TASSDB2 content. **(B)** The distributions of the total number of read counts in GTEx (top) and the total number of EST counts provided in TASSDB2 (bottom) for the three TASS categories in panel (A).

boundary of an annotated protein-coding exon, and non-coding otherwise. 87% of annotated TASS and 60% of *de novo* TASS were coding. This set extends TASSDB2 database, which is limited to TASS separated by 2–12 nt [214], by ∼51k TASS, which are absent from TASSDB2 (Fig 5-2, A). The newfound TASS are less expressed than TASS that are common to TASSDB2; however, the TASS from TASSDB2 that were not identified by my analysis are expressed at a significantly lower level (Fig 5-2, B).

Table 5.1: **The abundance and split read support of annotated and *de novo* TASS**

	expressed TASS		**% of split reads supporting TASS**
	number	%	
total	115,912	100%	100%
annotated	69,330	59.81%	99.83%
de novo	46,582	40.19%	0.17%

Although more than a third of the expressed TASS are *de novo* (Fig 5-1, B), their split read support is much weaker than that of the annotated TASS (Table 5.1).

Table 5.2: **The fractions of annotated and *de novo* splice sites among maSS and miSS.**

	de novo	annotated	total
maSS	9,043 (13%)	61,130(87%)	70,173
miSS	37,539 (82%)	8,200(18%)	45,739

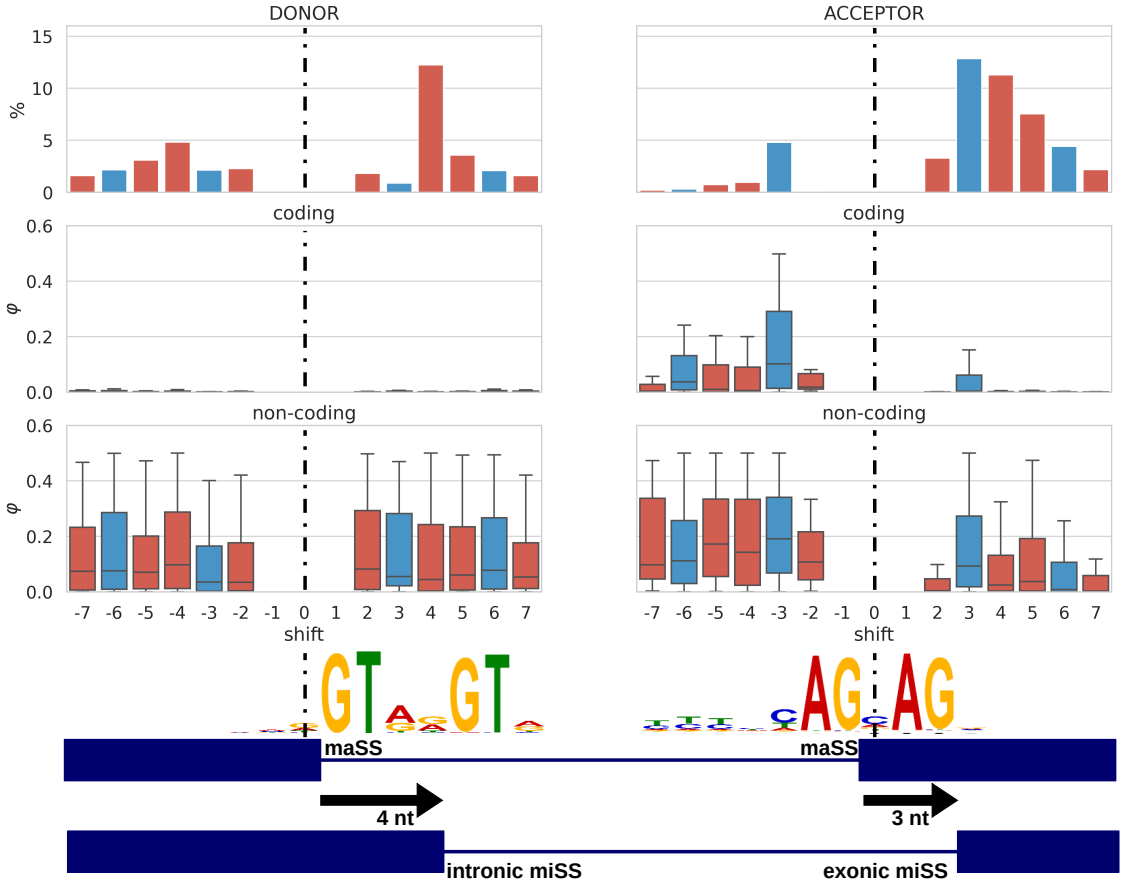

Figure 5-3: **The characterization of shifts in TASS.** The distribution of TASS shifts, i.e., the positions of miSS relative to maSS (top) for miSS of rank 2. The relative usage of miSS (ϕ) in coding (middle) and non-coding regions (bottom). The logo chart of miSS sequences of +4 donor shifts and of +3 acceptor shifts are shown. Frame-preserving (frame-disrupting) shifts are colored blue (red).

Accordingly, I classified TASS as either maSS or miSS according to their expression rank (see chapter 4) and found that among the total of 45,739 expressed miSS, the majority (82%) were not annotated, unlike the expressed maSS, 87% of which were annotated (Table 5.2).

Since each miSS is associated with a uniquely defined maSS within a TASS

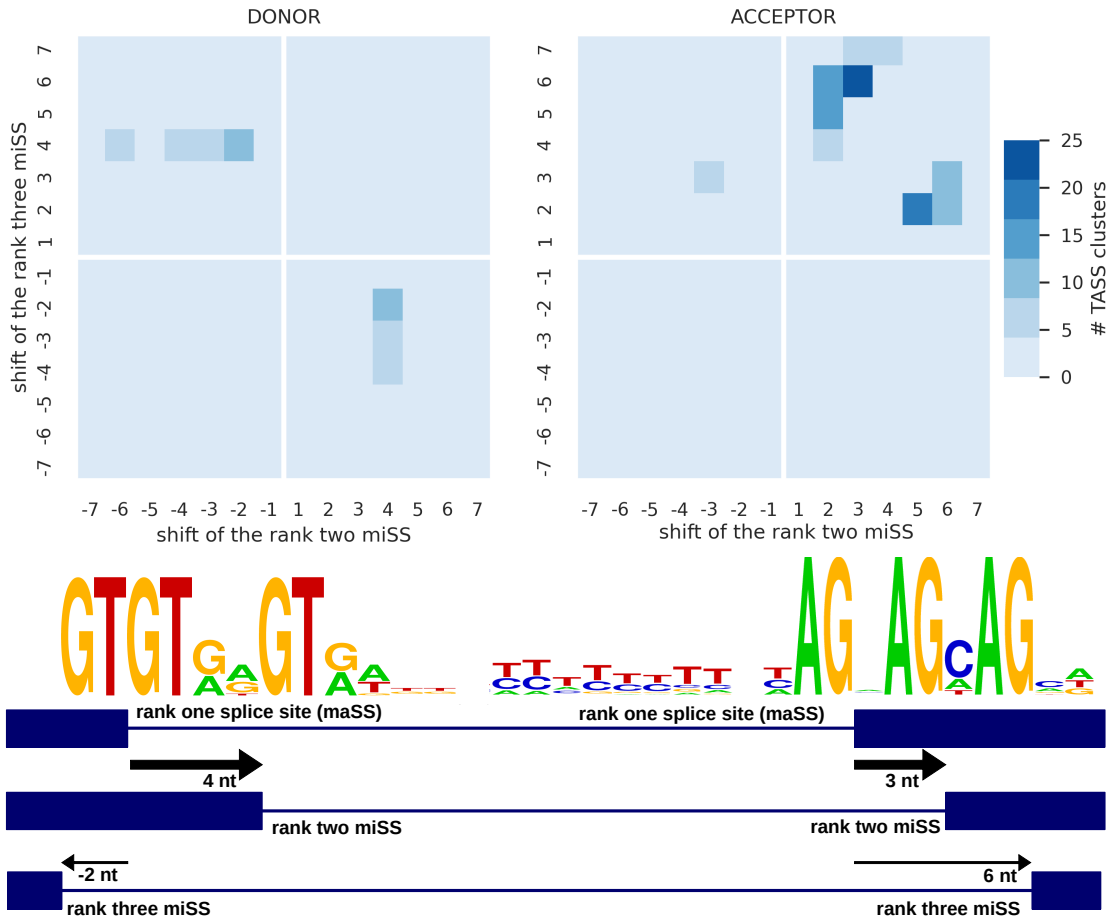

Figure 5-4: **TASS clusters of size three.** A TASS cluster of size three is characterized by two shift values, one for the rank two miSS relative to maSS, and the other for the rank three miSS relative to maSS. The top panel shows the joint distribution of rank two miSS shift (x-axis) and rank three miSS shift (y-axis) for donors (left) and (acceptors). The bottom panel shows LOGO charts of miSS sequences corresponding to shifts of +4 and -2 for the donor splice site, and +3 and +6 shifts for the acceptor splice site.

cluster (see chapter 4), the position of the miSS relative to the position of the maSS, which will be referred to as shift, is defined uniquely. Positive shifts correspond to miSS located downstream of the maSS in a gene, while miSS with negative shifts are located upstream of the maSS. Consistent with previous observations [59], the distribution of shift values for miSS in TASS clusters of size 2 reveals that the most frequent shifts among donor miSS are ±4 nt, while acceptor miSS are most frequently shifted by ±3 nt (Fig 5-3, top). These characteristic shifts likely arise from splice site consensus sequences, e.g. NAGNAG acceptor and GYNNGY donor

splice sites [115, 216]. Donor miSS in TASS clusters of size larger than two are often separated by an even number of nucleotides, while acceptor miSS are often separated by a multiple of 3 nt. For example, rank two and rank three donor miSS are often located 2 or 4 nt from the maSS and tend to have shifts with opposite signs, while rank two and rank three acceptor miSS are often separated by 3 nt and tend to be located downstream of the maSS (Fig 5-4). In what follows, I refer to a miSS that is located outside of the exon as intronic, and exonic otherwise (Fig 5-3, bottom). Intronic miSS correspond to insertions, while exonic miSS correspond to deletions.

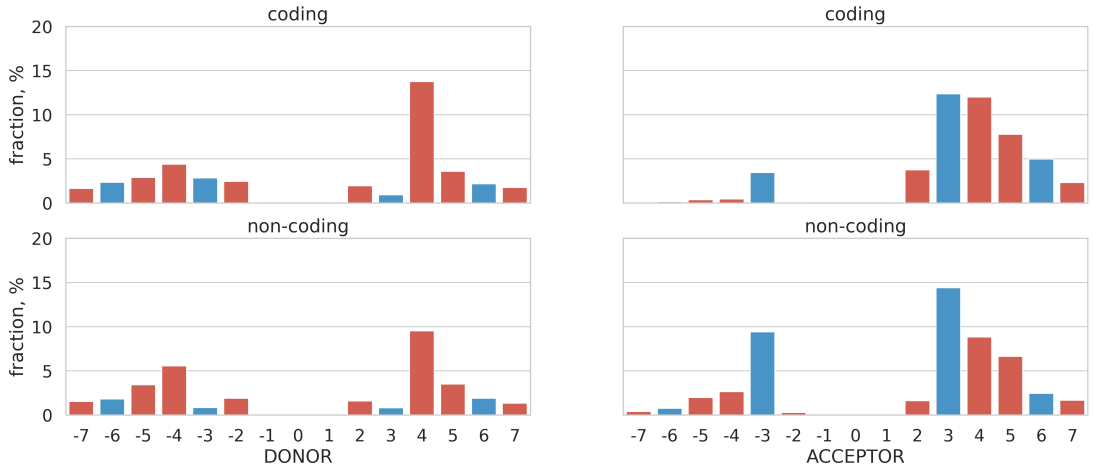

Figure 5-5: **Shift frequencies in coding vs. non-coding regions.** See Figure 5-3, top for comparison. Frame-preserving (frame-disrupting) shifts are colored blue (red).

In the coding regions, I expect the distance between miSS and maSS to be a multiple of three in order to preserve the reading frame. Indeed, shifts by a multiple of 3 nt are the most frequent among coding acceptor miSS; however a considerable proportion of shifts by not a multiple of 3 nt also occur in both donor and acceptor miSS (Fig 5-5, top). I therefore asked whether these frame-disrupting shifts are actually expressed, and found that, in spite of their high frequency, the relative expression of miSS in the coding regions, as measured by the ϕ metric, is still dominated by shifts that are multiple of 3 nt (Fig 5-3, middle), while the relative expression of miSS in the non-coding regions doesn't depend on the shift (Fig 5-3, bottom). Consistent with this, frame-disrupting miSS in the coding regions are significantly

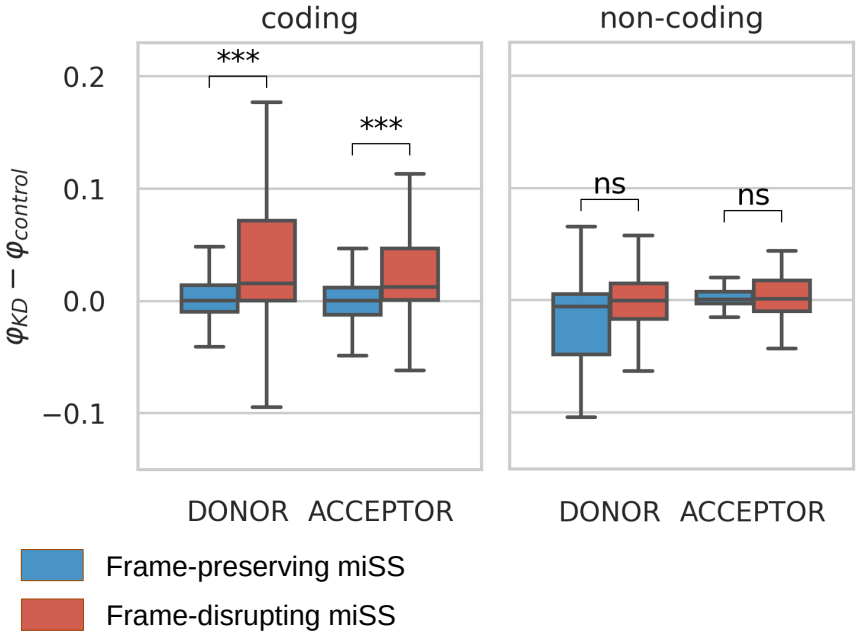

Figure 5-6: **The change of miSS relative usage ($\phi_{KD} - \phi_{control}$) upon NMD inactivation.**

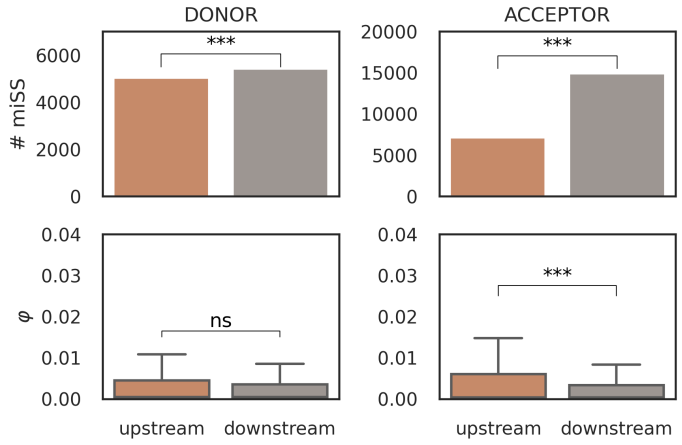

Figure 5-7: **The abundance and relative expression of upstream vs. downstream shifts.**

upregulated after the inactivation of the nonsense mediated decay (NMD) pathway by co-depletion of two its major components, *UPF1* and *XRN1* [181], while no such upregulation is observed in non-coding regions (Fig 5-6). This indicates that the broad positional repertoire of frame-disrupting shifts in coding TASS is efficiently

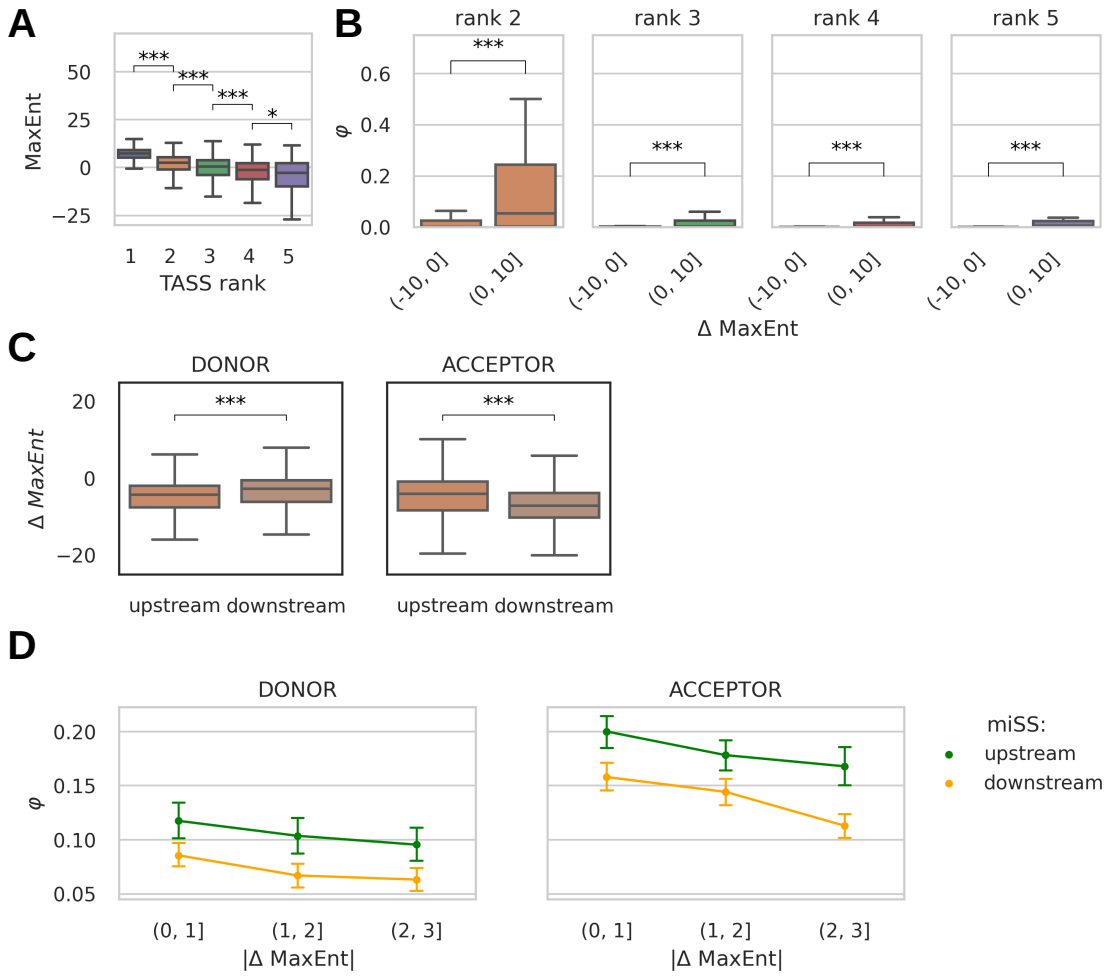

Figure 5-8: **The strength of TASS consensus sequences.** **(A)** According to MaxEnt scores, maSS (i.e., rank one sites) are on average stronger than miSS (i.e., rank 2, 3, 4, 5). **(B)** Within each rank group, the relative usage of a miSS (ϕ) generally increases with increasing ΔMaxEnt value, its strength relative to that of the maSS. **(C)** The distribution of ΔMaxEnt values for upstream and downstream shifts. **(D)** The relative usage of a miSS (ϕ) as a function of the difference of TASS strengths. The upstream miSS are used more frequently when the splice sites are nearly of the same strength.

suppressed by NMD.

I next asked whether the expression patterns systematically differ for miSS located upstream and downstream of the maSS. In the coding regions, the acceptor miSS are more often shifted downstream, while miSS located upstream tend to be expressed stronger than miSS located downstream (Fig 5-3; Fig 5-7). These patterns

are in line with previously observed avoidance of AG dinucleotides upstream of the expressed acceptor splice sites [217] and the proposed splice junction wobbling mechanism, in which the upstream acceptor splice site is usually expressed stronger than the downstream one [116]. However, the expression difference can also be explained by subtle, yet systematic differences in splice site strengths as miSS are on average weaker than maSS (Fig 5-8, A), the strength of a miSS relative to the strength of maSS is correlated with its relative expression (Fig 5-8, B), and the upstream acceptor miSS tend to be on average stronger than the downstream acceptor miSS (Fig 5-8, C, right). Nonetheless, the upstream miSS are expressed stronger than the downstream miSS even for miSS that are similar in strength to their corresponding maSS (Fig 5-8, D).

5.2 Expression of miSS in human tissues

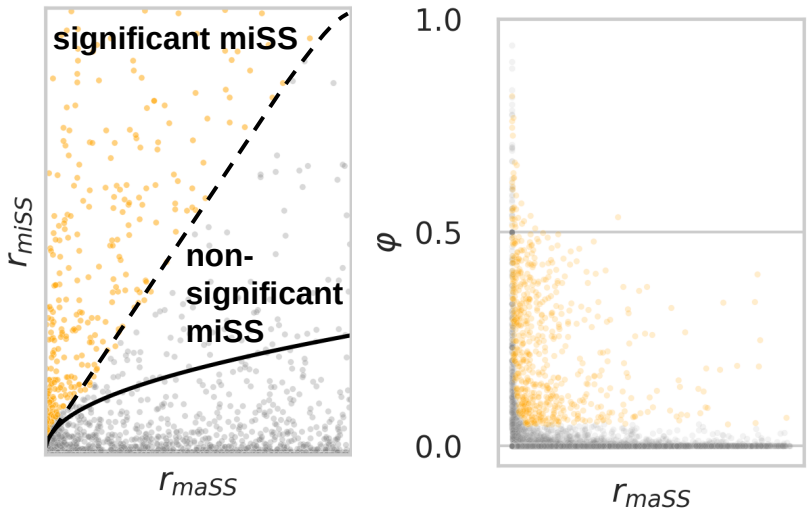

Figure 5-9: **Identification of significantly expressed miSS.** Zero-inflated Poisson model of miSS expression relative to maSS enables identification of significantly expressed miSS (left). Each dot represents a miSS. The expected values of r_{miSS} are shown by the solid curve. The FDR cutoff of 5% is shown by the dashed curve. The ϕ value decreases with increase of r_{maSS} (right).

Tissue specificity is commonly considered as a proxy for a splicing event to be

under regulation [69, 70]. To assess the tissue-specific expression of miSS, I calculated the ϕ_t metric, i.e. the relative expression of miSS with respect to maSS, by aggregating GTEx samples within each tissue t. However, different tissues are represented by a different number of individuals and, consequently, TASS in different tissues have different read support. To account for this difference and for the dependence of the relative expression of miSS on the gene expression level [183, 115], I constructed a zero-inflated Poisson linear model that describes the dependence of miSS-specific read counts (r_{miSS}) on maSS-specific read counts (r_{maSS}). Using this model, I estimated the statistical significance of miSS expression and selected significantly expressed miSS using Q-values with 5% threshold [185] and additionally required ϕ_t value to be above 0.05 (Fig 5-9). In what follows, I shortly refer to these significantly expressed miSS as significant.

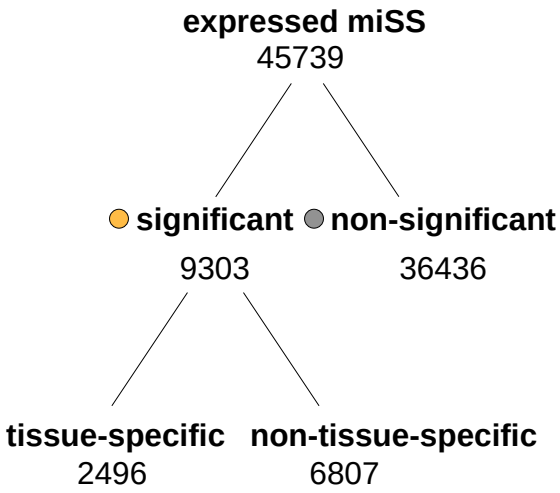

Figure 5-10: **The classification of expressed miSS.**

Out of 45,739 expressed miSS, 9,303 (20%) were significantly expressed in at least one tissue (Fig 5-10). To identify tissue-specific miSS among significant miSS, I built a linear model with dummy variables corresponding to each tissue (see chapter 4). A miSS was called tissue-specific if the slope of the dummy variable corresponding to at least one tissue was statistically discernible from zero (Q-value $< 5\%$), i.e.

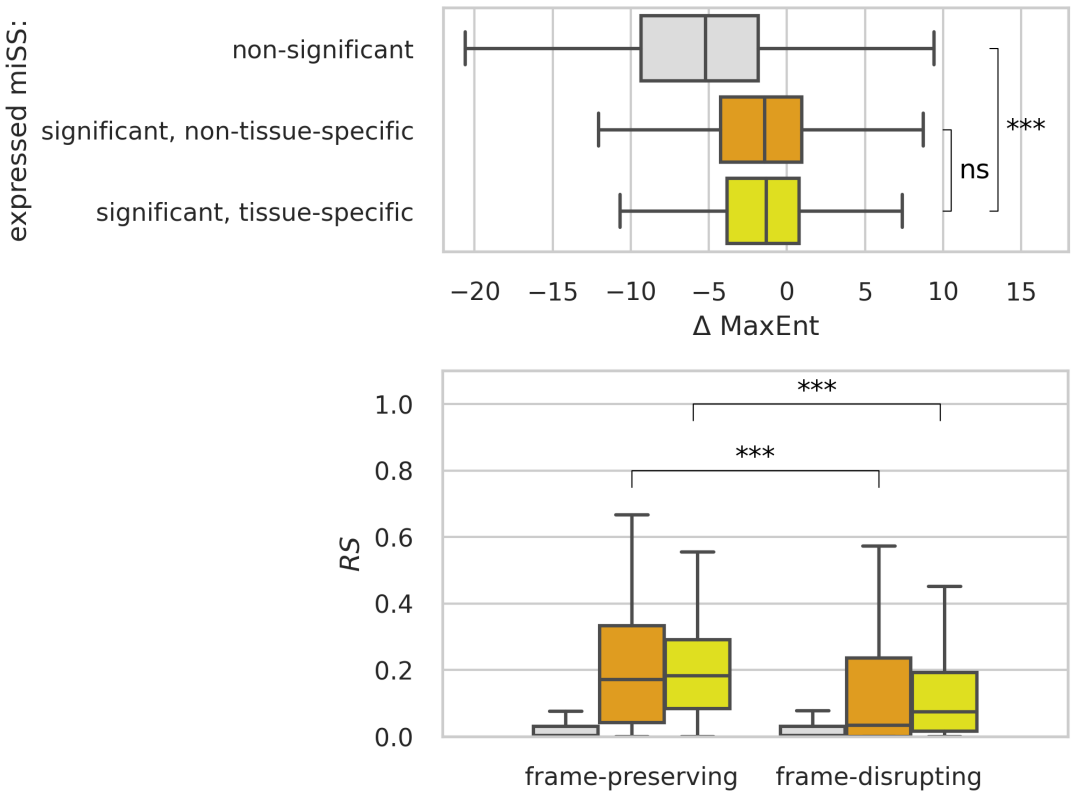

Figure 5-11: **Splice site strength and RiboSeq support of significantly expressed and tissue-specific miSS.** The distribution of ΔMaxEnt values for miSS in different expression categories (top). The distribution of *RS* (RiboSeq support) values for miSS of different expression categories in protein-coding regions (bottom).

the proportion of reads supporting a tissue-specific miSS deviates significantly from the average across tissues. To account not only for significant, but also for substantial changes, I additionally required the absolute value of $\Delta\phi_t$ be above 0.05, which resulted in a conservative list of 2,496 tissue-specific miSS (Fig 5-10). Among these miSS, 234 (9%) became maSS in at least one tissue. In the coding regions, tissue-specific miSS preserve the reading frame more often than non-tissue-specific miSS do (Table A.7); they also have on average stronger consensus sequences and, among the latter, frame-preserving miSS have stronger evidence of translation according to Ribo-Seq data (Fig 5-11). The intronic regions nearby tissue-specific and significantly expressed miSS tend to be more conserved evolutionarily as compared to those nearby not significant miSS (Fig 5-12, A), with frame-disrupting

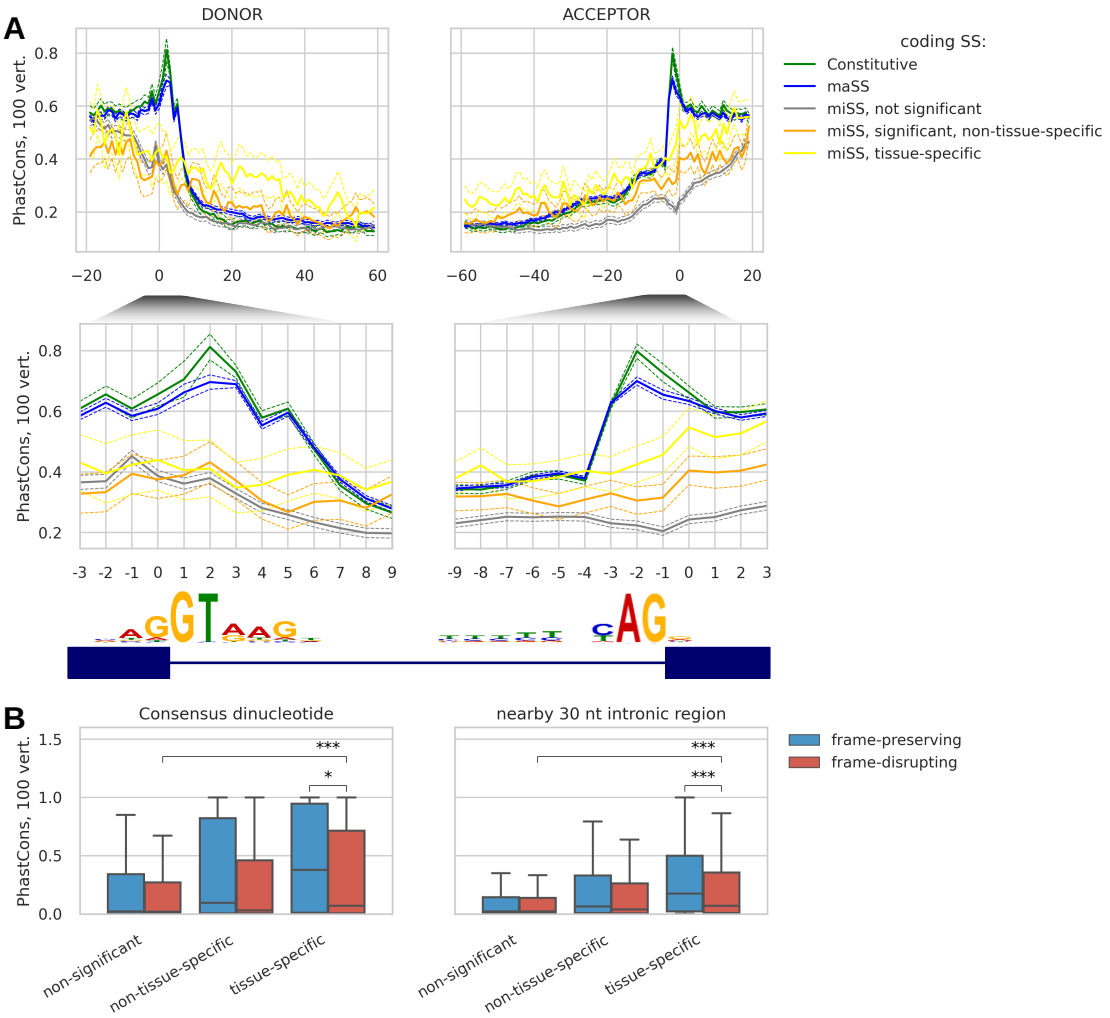

Figure 5-12: **Evolutionary conservation of miSS. (A)** The average PhastCons scores (100 vertebrates) for positions near the miSS in different expression categories. **(B)** The distribution of average PhastCons scores (100 vertebrates) at the consensus dinucleotides of splice sites (left) and average PhastCons scores of the adjacent 30 nt intronic regions (right).

miSS being significantly less conserved than frame-preserving miSS (Fig 5-12, B). The distribution of shifts in tissue-specific and other significantly expressed miSS strongly differs from that in non-significantly expressed miSS: among donor miSS, the fraction of +4 shifts is almost two times lower in significantly expressed miSS than in non-significant ones, among acceptor miSS ±3 nt shifts are dominating in significantly expressed miSS while the fraction of other shift variants is lower than in non-significantly expressed miSS.

I checked if there is an enrichment of specific Gene Ontology (GO) categories

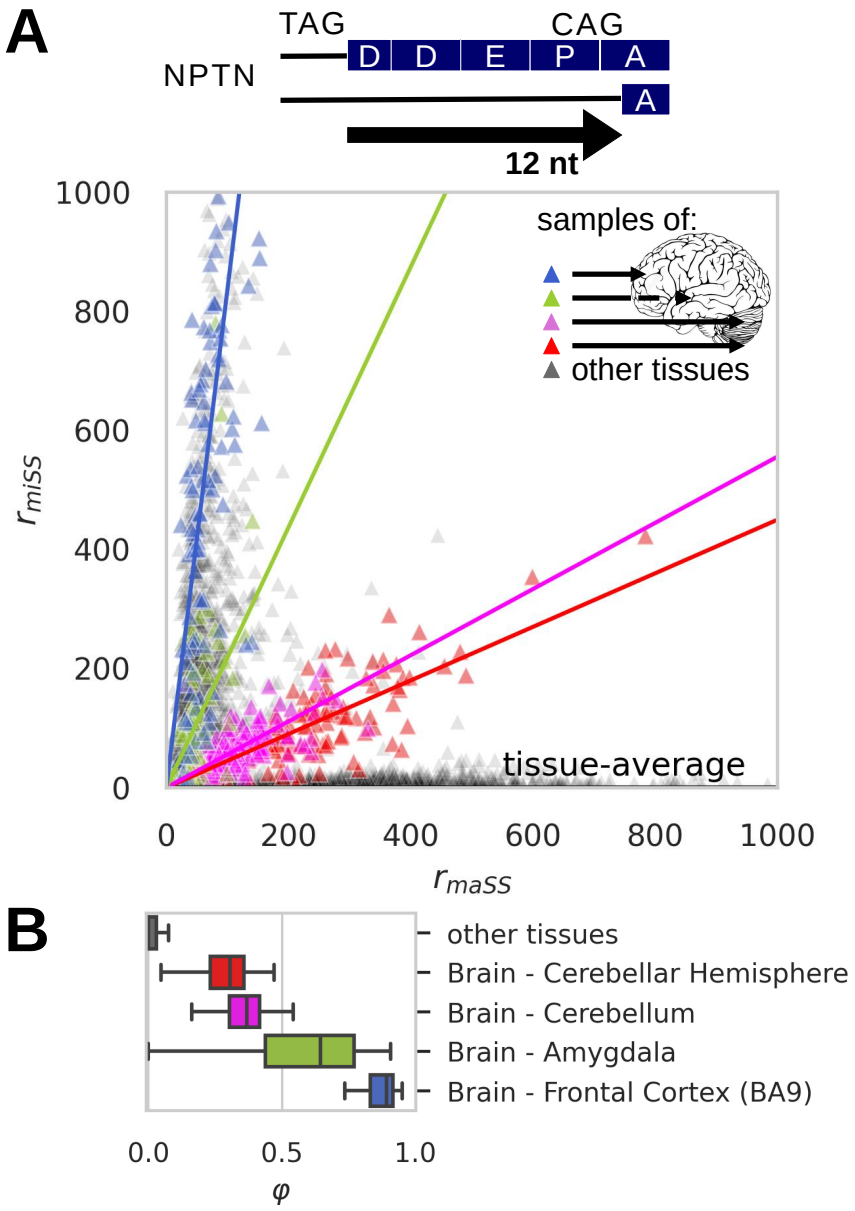

Figure 5-13: **Tissue-specific miSS in the *NPTN* gene. (A)** The indel caused by *NPTN* miSS results in the deletion of DDEP motif from the aminoacid sequence (top). Tissue-specific expression of a miSS in the *NPTN* gene (bottom). Each dot represents a sample, i.e. one tissue in one individual. Tissue specificity is estimated by a linear model with dummy variables corresponding to tissues. **(B)** The distribution of *NPTN* miSS ϕ values in selected tissues.

in genes harboring tissue-specific miSS compared with genes not having such miSS. GO-analysis in GOrilla [218] found a strong enrichment in nuclear localization and chromatin organization process (Table A.8). However, I found that the majority of

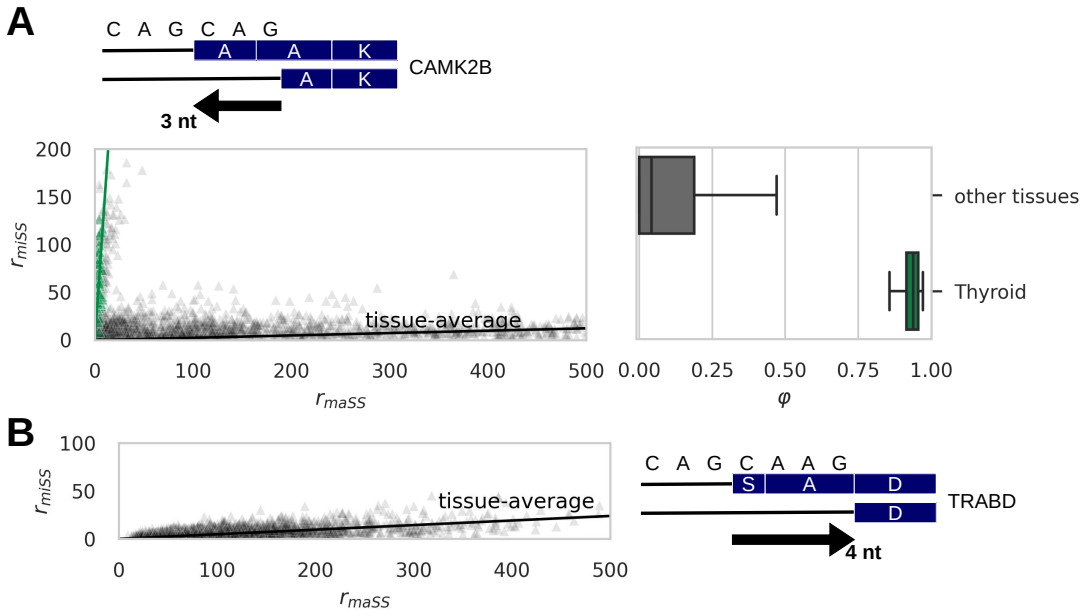

Figure 5-14: **Examples of tissue-specific and non-tissue-specific miSS. (A)** A thyroid-specific miSS in exon 14 of the gene *CAMK2B*. The miSS becomes a maSS in thyroid as its ϕ value exceeds 0.5. **(B)** The miSS in exon 8 of the *TRABD* gene is non-tissue-specific.

genes having tissue-specific miSS also have non-significant miSS (Fig A-4, A), and such genes tend to contain more exons in them (Fig A-4, B), creating a selection bias for GO-analysis [219]. To eliminate the bias, for each gene having tissue-specific miSS, I randomly selected a gene harboring no tissue-specific miSS but having the same number of exons (Fig A-4, C). I used this subset as a background for the analysis of GO-enrichment in GOrilla [218] and DAVID [220] web servers and obtained no significantly enriched categories with a false discovery rate below 0.1, possibly indicating the versatility of gene expression regulation via tissue-specific tandem alternative splice sites.

One notable example of a tissue-specific miSS is in the exon 7 of *NPTN* gene, which encodes neuroplastin, an obligatory subunit of Ca^{2+}-ATPase, required for neurite outgrowth, the formation of synapses, and synaptic plasticity [221, 222]. The slope of the linear model has a distinct pattern of variation across tissues, and moreover within brain subregions (Fig 5-13, A). Brain-specific expression of the acceptor miSS instead of maSS in *NPTN* leads to the deletion of Asp-Asp-Glu-

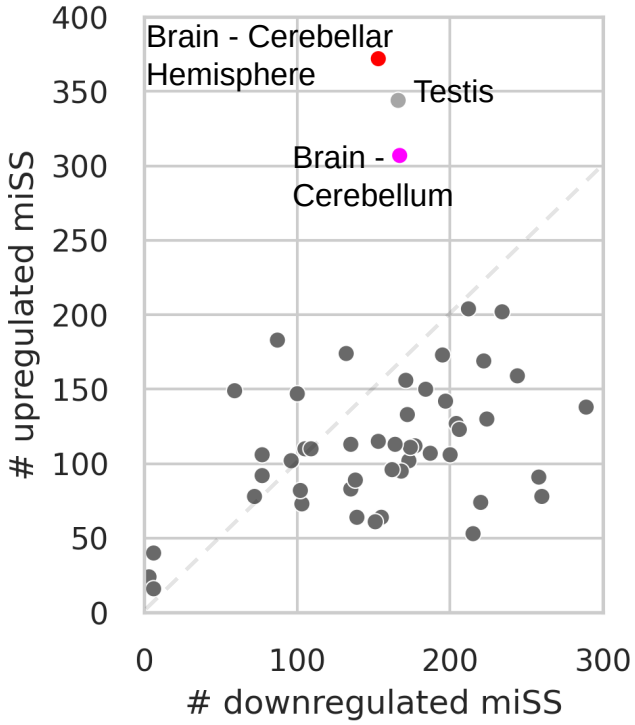

Figure 5-15: **Tissue-specificity patterns of miSS.** The number of tissue-specific (up- or downregulated) miSS in each tissue.

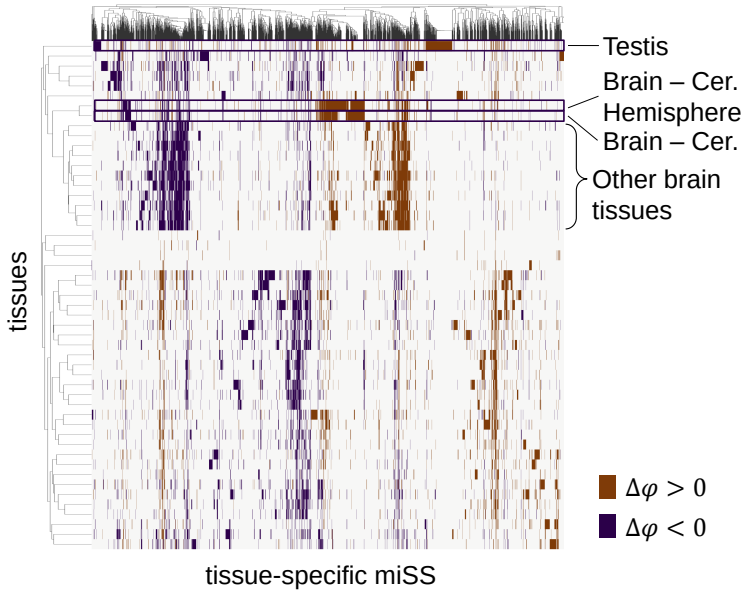

Figure 5-16: **Clustering of tissue-specific miSS and tissues based on ϕ values.**

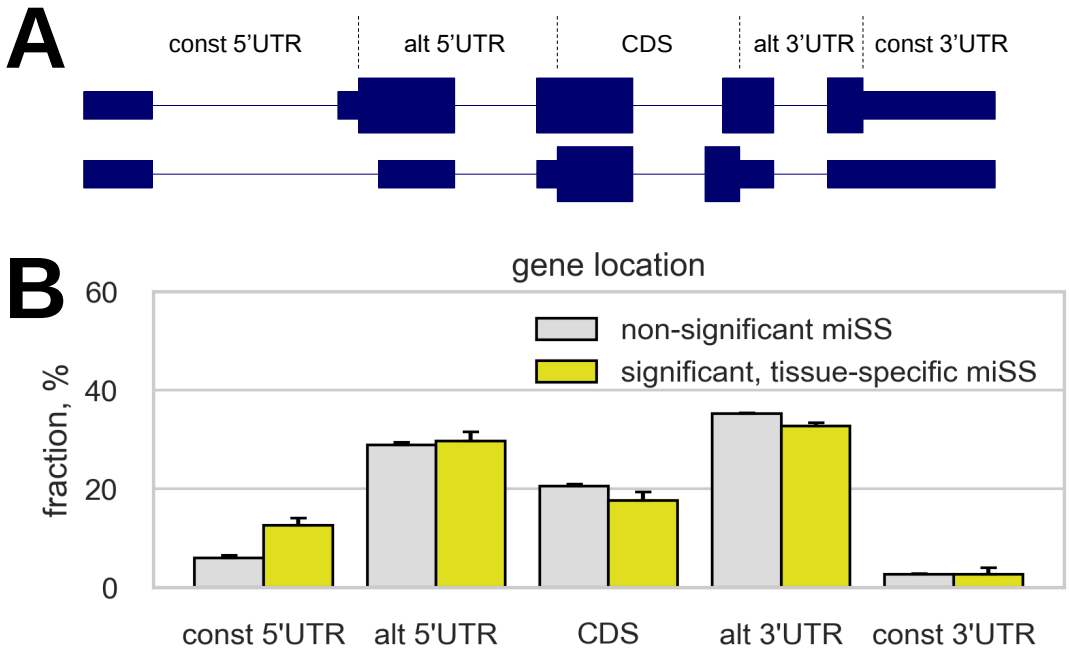

Figure 5-17: **miSS within UTRs. A** The segmentation of genes into five categories based on GENCODE transcript annotation. Constitutive (const) 5'UTR and 3'UTR regions are annotated as untranslated in all transcripts, while alternative (alt) 5'UTR and 3'UTR regions overlap with protein-coding exons in some transcripts. **B** The fractions of tissue-specific and non-significant miSS in gene location categories.

Pro (DDEP) sequence from the canonical protein isoform (Fig 5-13, B). Two more examples of tissue-specific and non-tissue-specific miSS are provided in (Fig 5-14, A) and (Fig 5-14, B).

Tissues differ by the number of tissue-specific miSS and by the proportion of miSS that are upregulated or downregulated. The sign of the slope in the linear model that describes the dependence of r_{miSS} on r_{maSS} allows to distinguish up- and downregulation. In agreement with previous reports on alternative splicing [223], a number of tissues including testis, cerebellum, and cerebellar hemisphere harbor the largest number of tissue-specific miSS (Fig 5-15, Table A.9). The testis and the brain have a distinguished large set of miSS with almost exclusive expression in these tissues that set them apart statistically from the other tissues (Fig 5-16).

I further found that while 77% tissue-specific miSS belong to protein-coding TASS clusters (Table A.7), the majority of them are located within regions anno-

tated as protein-coding in some transcripts and untranslated in others (alternative UTRs, Fig 5-17, A, B). In addition, tissue-specific miSS are enriched in constitutive 5'UTRs in comparison to non-significant miSS (Fig 5-17, B), suggesting a participation in gene expression regulation via splicing in 5'UTRs [224, 225].

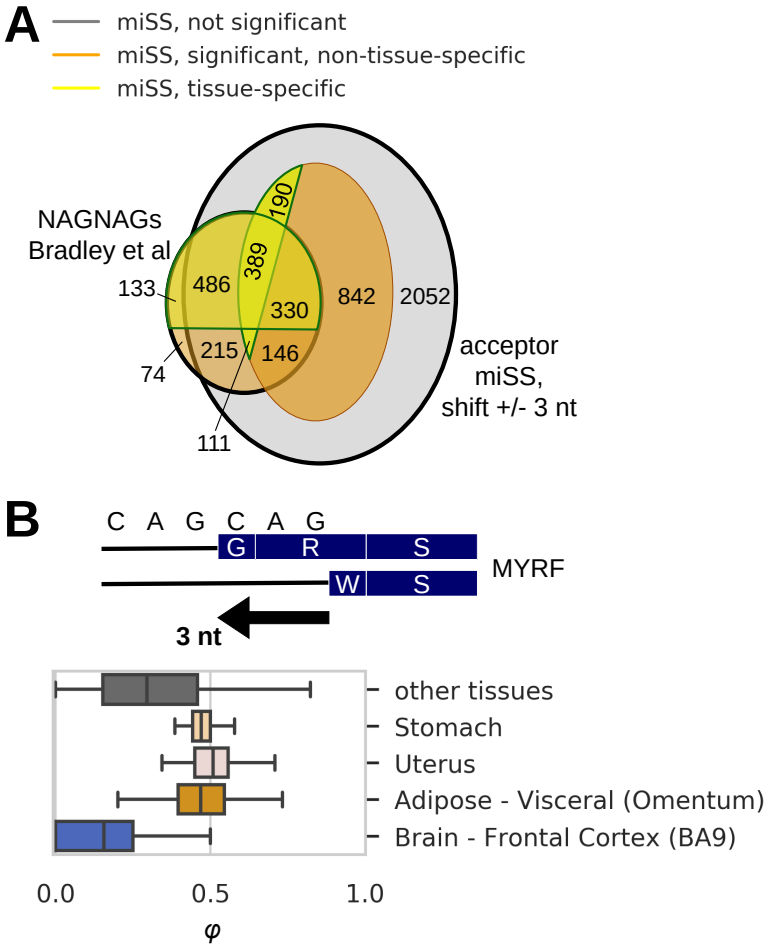

Figure 5-18: **NAGNAGs. (A)** The intersection of the acceptor miSS located ±3 nt from the maSS with the list of NAGNAGs provided by Bradley et al [69]. **(B)** A NAGNAG acceptor splice site in the exon 20 of the *MYRF* gene. The upstream NAG is upregulated in the stomach, uterus, adipose tissues and downregulated in the brain.

A special class of TASS are the so-called NAGNAG acceptor splice sites, i.e., alternative acceptor sites that are located 3 bp apart from each other [216]. According to the current reports, they are found in 30% of human genes and appear to be functional in at least 5% of cases [112]. Here, I identified an extended set

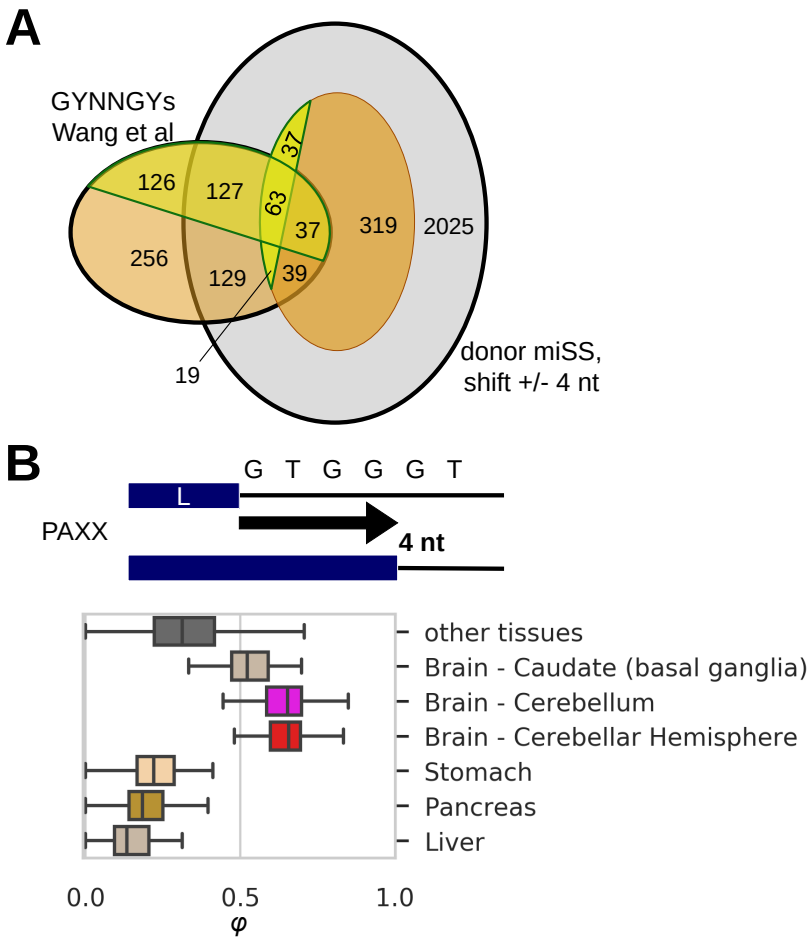

Figure 5-19: **GYNNGYs. (A)** The intersection of the donor miSS located ±4 nt from maSS with the list of GYNNGYs provided by Wang et al [80]. **(B)** A GYNNGY donor splice site in the exon 2 of the *PAXX* gene. The downstream GY is upregulated in the brain and downregulated in the stomach, pancreas, and liver tissues.

of 4,761 expressed acceptor miSS, of which 690 are tissue specific, that are located ±3 nt from maSS (Fig 5-18, A) which reconfirms 89% of 1,884 alternatively spliced and 29% of 1,338 tissue-specific NAGNAGs reported by Bradley et al [69]. Furthermore, I identified 190 tissue-specific NAGNAGs that are not present in the previous lists [69]. Among them there is a NAGNAG acceptor splice site in the exon 20 of the *MYRF* gene, which encodes a transcription factor that is required for central nervous system myelination. The upstream NAG is upregulated in stomach, uterus and adipose tissues and downregulated in brain tissues (Fig 5-18, B). Similarly, I identified an extended set of 2,794 expressed GYNNGY donor splice sites, i.e., alternative

donor splice sites that are located 4 bp apart from each other (Fig 5-19, A). This set reconfirms 52% of 796 GYNNGY donor splice sites reported by Wang et al [115]. Additionally, I identified 37 novel tissue-specific GYNNGYs including a donor splice site in the exon 2 of the *PAXX* gene (Fig 5-19, B), the product of which plays an essential role in the non-homologous end joining pathway of DNA double-strand break repair [226]. Unlike NAGNAGs, alternative splicing at GYNNGYs disrupts the reading frame and is expected to generate NMD-reactive isoforms [115]. Indeed, GYNNGY miSS along with other frame-disrupting miSS are significantly upregulated after inactivation of the nonsense mediated decay (NMD) pathway by the co-depletion of two major NMD components, *UPF1* and *XRN1* [181] (Fig 5-20).

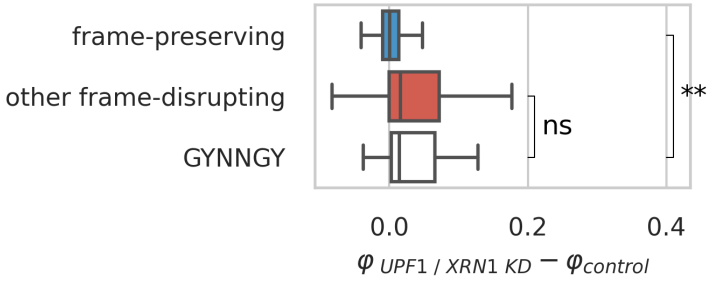

Figure 5-20: **The response of GYNNGY miSS to NMD inactivation.**

The expression of splicing factors is functionally associated with tissue-specific patterns of alternative splicing [79]. In order to identify the potential regulatory targets of splicing factors among miSS, I analyzed the data on shRNA depletion of 103 RBPs followed by RNA-seq and compared it with tissue-specific expression of miSS and RBP [210]. My strategy was to identify tissues with significant up- or downregulation of a miSS responding to the inactivation of a splicing factor with the same signature of tissue-specific expression.

To this end, I identified miSS that are up- or downregulated upon RBP inactivation by shRNA-KD (RBP-miSS pairs) and matched them with the list of tissue-specific miSS (miSS-tissue pairs) and the list of tissue-specifically expressed RBP (RBP-tissue pairs). The intersection of these lists resulted in 256 miSS-RBP-tissue triples. Each triple was classified as co-directed or anti-directed according to the rules shown in Fig 4-5 (see chapter 4).

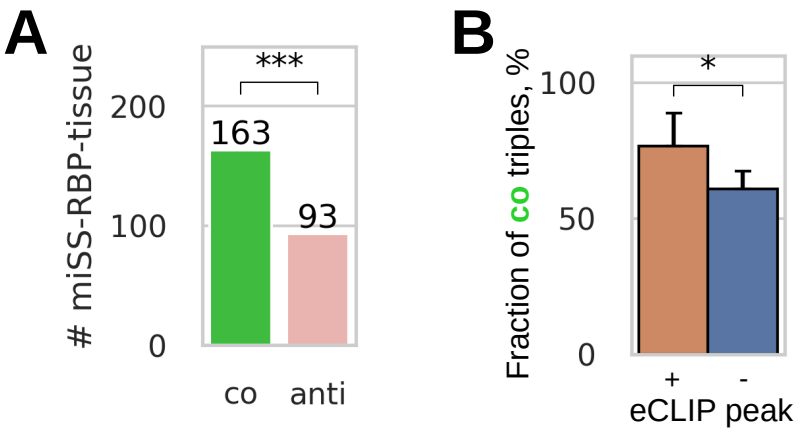

Figure 5-21: **The abundance of co-directed and anti-directed miSS-RBP-tissue triples. (A)** The number of co-directed triples is significantly greater than the number of anti-directed triples. **(B)** The fraction of co-directed miSS-RBP-tissue triples ($\#co/(\#co + \#anti)$) is significantly greater among triples supported by an eCLIP peak of the RBP near miSS as compared to non-supported triples.

In order to obtain a stringent list of regulatory targets, I applied 5% FDR threshold correcting for testing in multiple tissues, multiple RBPs, and multiple miSS, and additionally required that miSS relative usage and RBP expression change not only significantly, but also substantially ($|\Delta\phi_t| > 0.05$, $|\Delta\phi_{KD}| > 0.05$, and $|\Delta RBP_t| > 0.5$. As a result, I obtained 163 co-directed and 93 anti-directed miSS-RBP-tissue triples, an uneven proportion that is unlikely to be due to pure chance alone (Fig 5-21, A). Next, I compared these predictions to the footprinting of RBP by the eCLIP method [187] and found that co-directed triples are significantly more abundant among miSS-RBP-tissue triples that are supported by an eCLIP peak (Fig 5-21, B). I summarized the data on co-directed and anti-directed miSS-RBP-tissue triples in (Table A.10). I identified six miSS-RBP candidate pairs with tissue-specific splicing regulation that is co-directed with RBP expression and supported by an eCLIP peak (Table A.11). A notable example is the downregulation of the acceptor miSS in exon 6 of the *QKI* gene by *PTBP1* in muscle and cardiac tissues (Fig 5-22), which is consistent with previous reports on the coregulation of alternative splicing by *QKI* and *PTBP1* during muscle cell differentiation [227].

To further investigate potential involvement of *PTBP1* in the regulation of alternative usage of other miSS, I analyzed *PTBP1* overexpression data [193] and

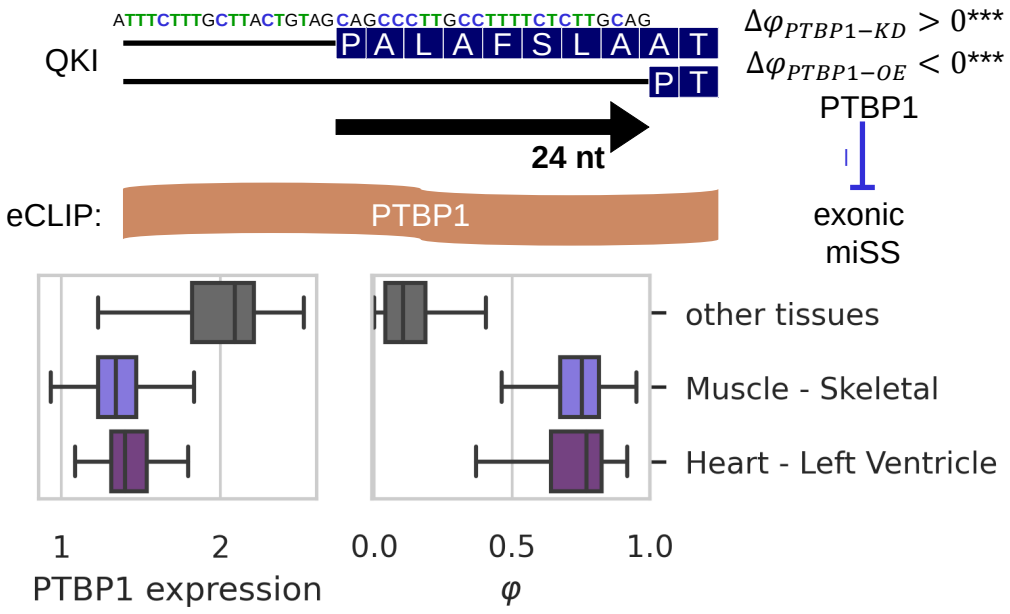

Figure 5-22: ***PTBP1* regulates a miSS in the *QKI* gene.** A deletion of eight aminoacids in the *QKI* gene caused by the exonic miSS overlapping with *PTBP1* eCLIP peak (top). In muscles and heart, the expression of *PTBP1* ($\log_2 TPM$) is suppressed, while the relative usage of the miSS is promoted (bottom). The miSS is activated in response to *PTBP1* inactivation and is inhibited in response to *PTBP1* overexpression, suggesting downregulation by *PTBP1*.

identified 50 events, in which the response to *PTBP1* overexpression or the response to shRNA-KD of *PTBP1* was statistically significant (Q-value $< 5\%$) and substantial ($|\Delta\phi| > 0.05$) (Table A.12). As expected, the miSS responses to *PTBP1* overexpression and inactivation by shRNA-KD were negatively correlated (Fig 5-23, A). I also found that miSS located proximally (shift < 5) and distally (shift ≥ 5) with respect to maSS responded differently to *PTBP1* perturbations. *PTBP1* stimulated the expression of proximal miSS and suppressed the expression of distal miSS (Fig 5-23, B) suggesting a different mode of regulation for polypyrimidine tracts overlapping the TASS region. Such a coordinated expression of TASS has been previously reported in *C. elegans*, in which the expression of proximal and distal TASS is remarkably coordinated between germ-line and somatic cells [228], and in human and murine

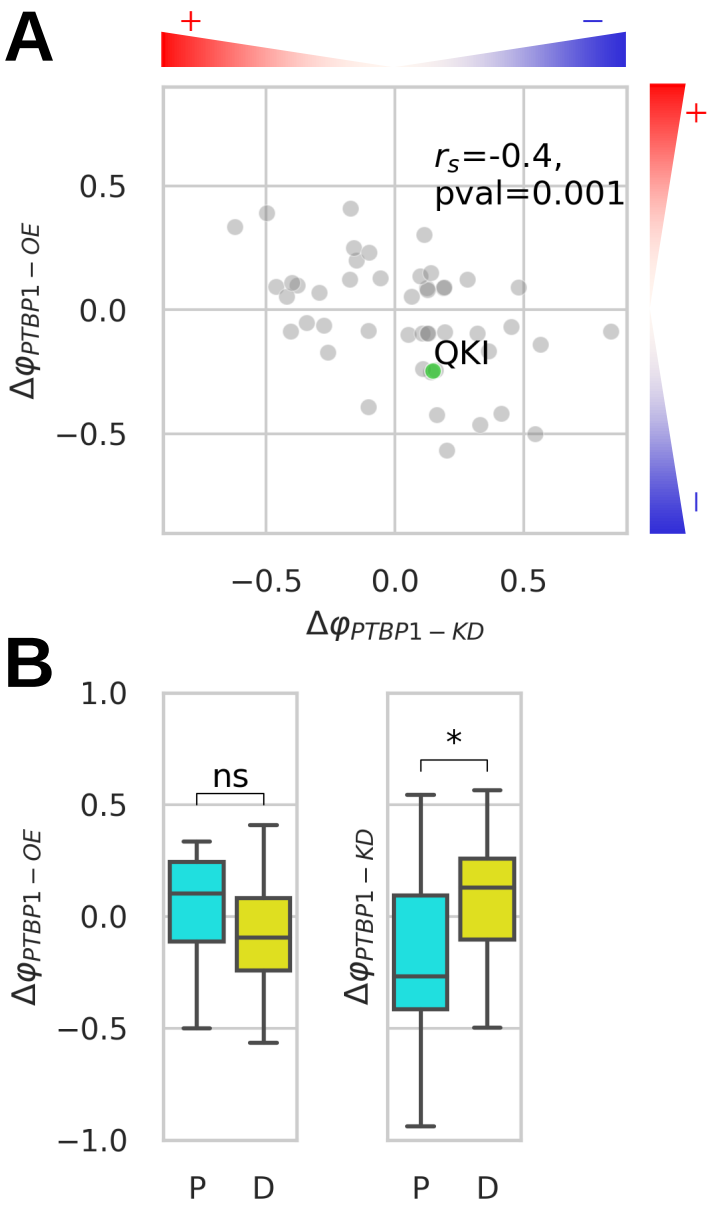

Figure 5-23: **miSS controlled by *PTBP1*.** **(A)** The responses of miSS to *PTBP1* overexpression and inactivation by shRNA-KD are negatively correlated. Shown are 50 miSS, in which the response to *PTBP1* overexpression or the response to shRNA-KD of *PTBP1* is statistically significant (Q-value $< 5\%$) and substantial ($|\Delta\phi| > 0.05$). **(B)** The responses of proximal (shift$<$5 nt) and distal (shift$\geq$5 nt) miSS to *PTBP1* overexpression and inactivation by shRNA-KD are opposite suggesting a different mode of regulation for polypyrimidine tracts overlapping the TASS region.

samples, in which the NAGNAG isoforms showed a remarkable co-regulatory pattern [114]. It suggests that *PTBP1* could be one of the master regulators that govern

such coordinated changes across cell types.

5.3 Expression of miSS in cell types

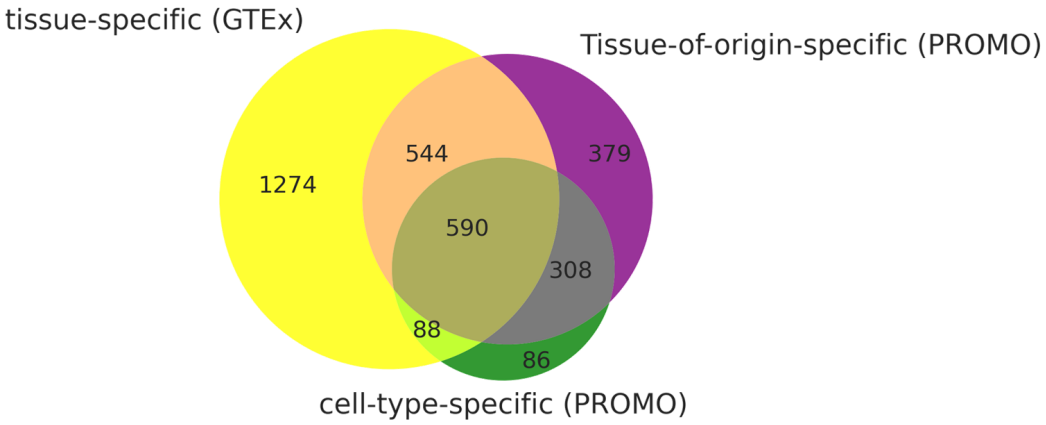

Figure 5-24: **The intersection of tissue-specific miSS identified using GTEx data with cell-type-specific miSS and tissue-of-origin-specific miSS identified using PROMO cells data.**

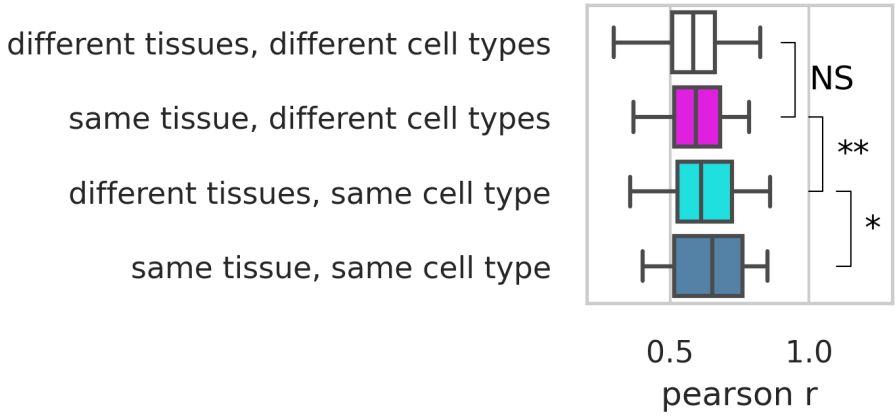

Figure 5-25: **The similarity of miSS expression profiles measured by the Pearson correlation coefficient r in the same or different tissues of origin vs. the same or different cell type.**

Tissue-specific alternative splicing originates from that of the constituent cell types, in which TASS splicing programs can also be functionally distinct. To dissect

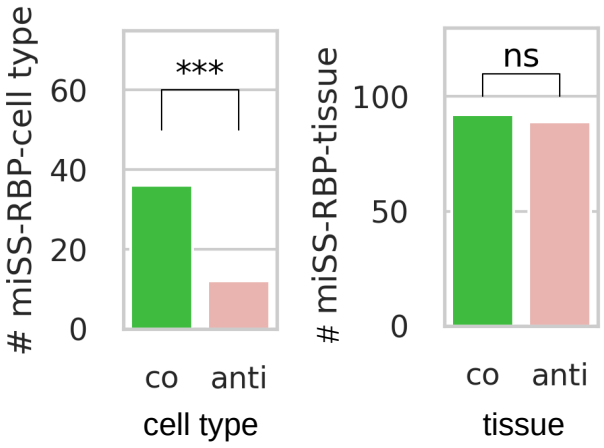

Figure 5-26: **The abundance of co-directed triples significantly exceeds the abundance of anti-directed triples for the association of miSS-RBP-cell type (left), while there is no significant difference for the association of miSS-RBP-tissue (right).**

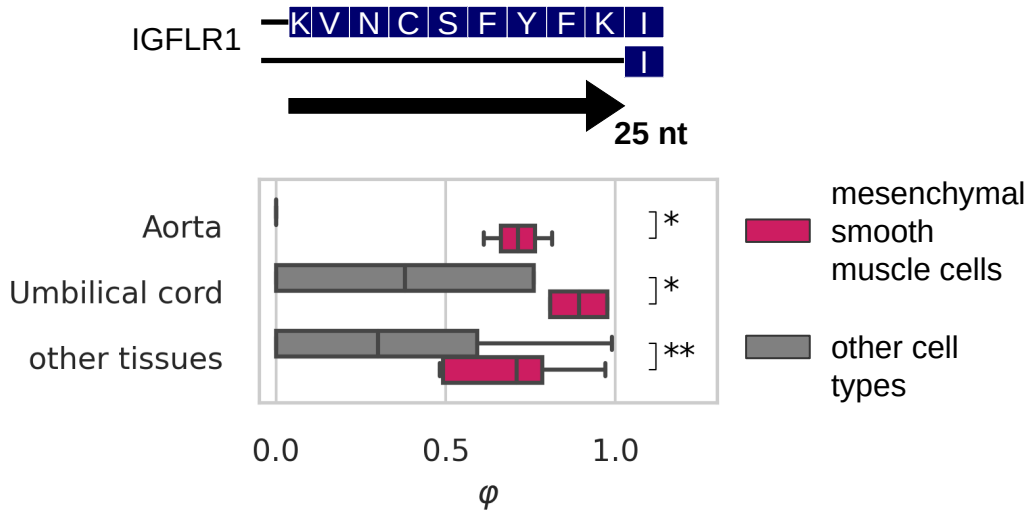

Figure 5-27: **The expression of an acceptor miSS in exon 2 of the *IGFLR1* gene is upregulated in mesenchymal smooth muscle cells regardless of the tissue-of-origin.**

the cell-type-specific expression of TASS, I analyzed RNA-seq data for primary cells from different locations in the human body [194]. Using the same methodology as in the analysis of tissue-specific TASS, I identified 1,821 tissue-of-origin-specific and 1,072 cell-type-specific miSS among significantly expressed miSS (Fig 5-24).

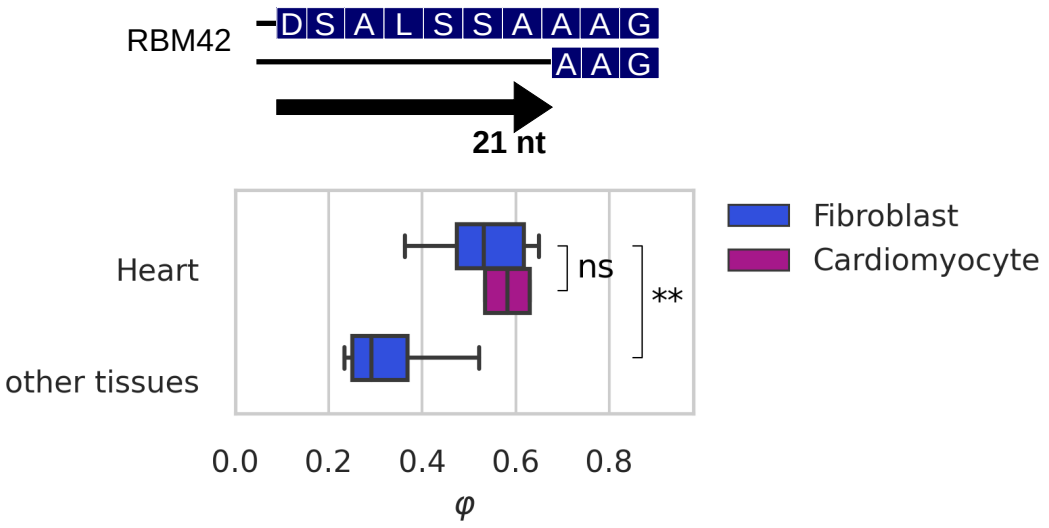

Figure 5-28: **The expression of an acceptor miSS in exon 6 of the *RBM42* gene is upregulated in both heart fibroblasts and heart cardiomyocytes, but not in fibroblasts from other tissues.**

Using Pearson correlation as a similarity measure, I found that miSS expression profiles were more similar for the same cell type from different tissues than for different cell types from the same tissue (Fig 5-25). Furthermore, the proportion of co-directed instances among miSS-RBP-cell-type triples is significantly larger than it is among miSS-RBP-tissue triples (Fig 5-26). On the one hand, it suggests that miSS expression is governed more by the cell type than by the tissue. It is the case, for instance, for exon 2 of the *IGFLR1* gene, which is upregulated in mesenchymal smooth muscle cells regardless of the tissue (Fig 5-27). On the other hand, the expression of some miSS depends on the tissue of origin regardless of the cell type, e.g., a miSS in exon 6 of the *RBM42* gene is upregulated in both heart fibroblasts and heart cardiomyocytes, but not in any cell type from other tissues (Fig 5-28).

5.4 Structural annotation of miSS

Alternative splicing of frame-preserving TASS results in mRNA isoforms that translate into proteins with only a few amino acids difference. It was reported earlier that alternative splicing tends to affect intrinsically disordered protein regions [229], and

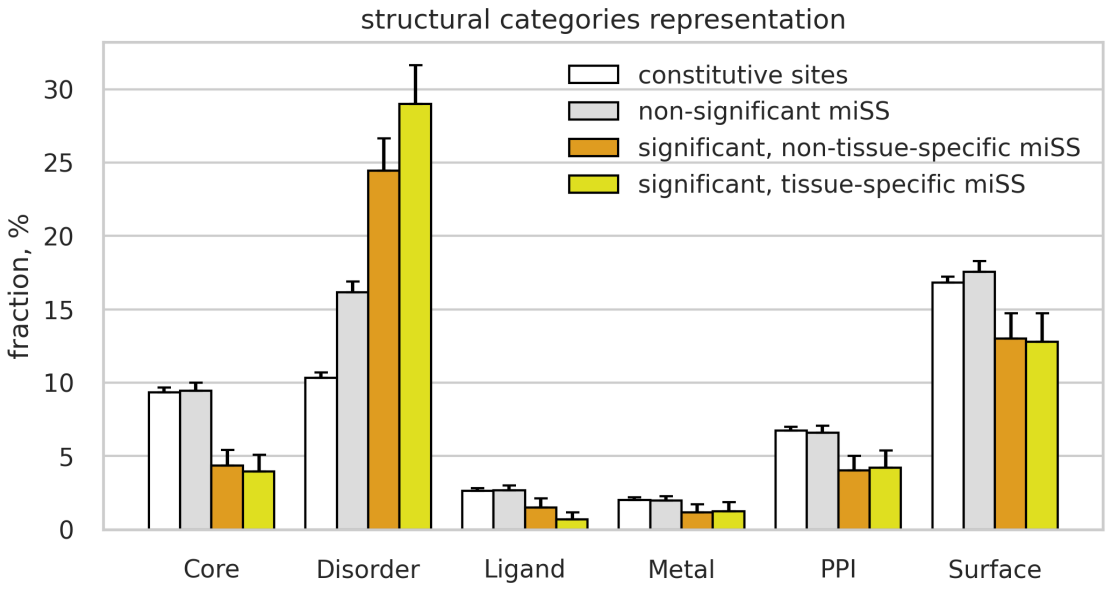

Figure 5-29: **The proportion of miSS in genomic regions corresponding to protein structural categories.**

that TASS with significant support from ESTs and mRNAs (506 such splice sites in total) are further overrepresented within regions lacking a defined structure [111].

I analyzed the structural annotation of the human proteome (see chapter 4) and found that significantly expressed miSS preferentially affect disordered protein regions, and tissue-specific miSS are found in disordered regions even more frequently (Fisher exact test, p-value$<$0.01, Fig 5-29). Within disordered protein regions, indels that are caused by significantly expressed miSS and their nearby exonic regions are enriched with SLiMs, short sequence segments that often mediate protein interactions playing important functional roles in physiological processes and disease states [230, 231, 232] (Fig 5-30, A). Furthermore, protein sequences of indels and nearby exonic regions for tissue-specific miSS are significantly enriched with methylation sites, one of the most frequent post-translational modifications (PTMs) from the dbPTM database [233] (Fig 5-30, B). Interestingly, I found that nucleotide sequences of indels caused by tissue-specific miSS in disordered protein regions are more conserved evolutionarily compared to those of non-tissue-specific miSS, further supporting the enrichment of functional regulatory sites such as SLiMs or PTMs (Fig 5-30, C).

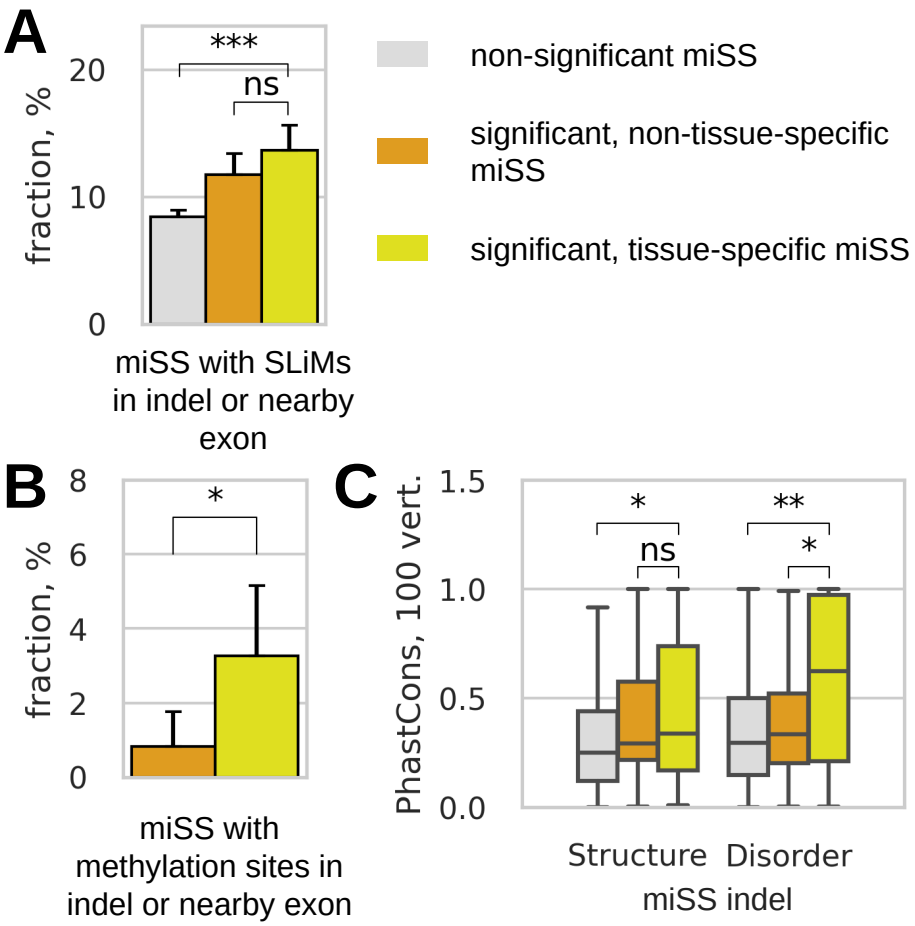

Figure 5-30: **Protein-level characterization of miSS indels.** **(A)** The proportion of miSS overlapping occurrences of predicted SLiMs. **(B)** The proportion of frame-preserving miSS in genomic regions corresponding to protein methylation sites from the dbPTM database. **(C)** The distribution of the average PhastCons conservation score (100 vertebrates) in the genomic regions between miSS and maSS.

A notable example of a SLiM within indel that is caused by miSS is located in the *PICALM* gene, the product of which modulates autophagy through binding to ubiquitin-like *LC3* protein [234, 235] (Fig 5-31). The expression of the short isoform lacking 15 nt at the acceptor splice site results in the deletion of Phe-Asp-Glu-Leu (FDEL) sequence, which represents a canonical LIR (*LC3*-interacting region) motif [236]. This motif interacts with *LC3* protein family members to mediate processes involved in selective autophagy. This miSS is slightly upregulated in whole blood and downregulated in brain tissues consistently with a possible role in physiological

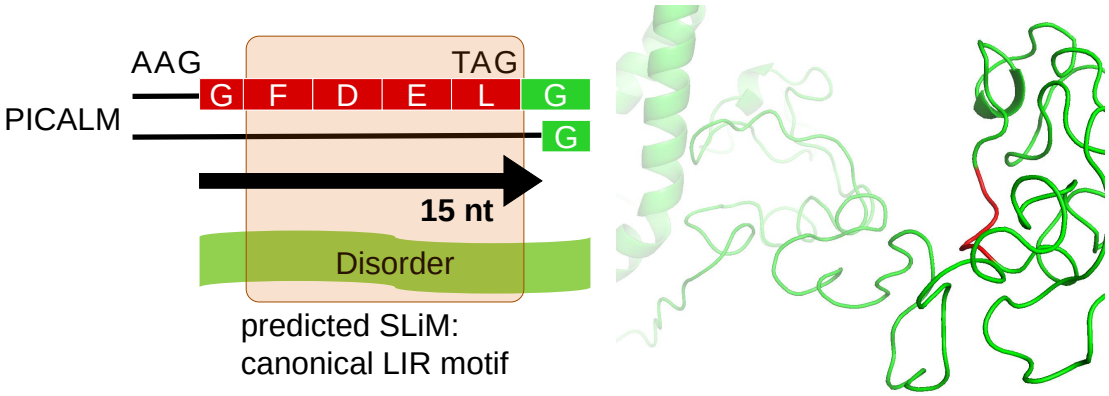

Figure 5-31: **The expression of an acceptor miSS in the predicted disordered region in the *PICALM* gene.** The expression of a miSS results in the deletion of five amino acids containing a predicted canonical LIR motif. The maSS-expressing structure of PICALM was modelled with I-TASSER (green); miSS indel is shown in red.

regulation of autophagy.

Despite the enrichment of tissue-specific miSS in disordered regions, a sizable fraction of them (more than 10%) still correspond to functional structural categories such as protein core, sites of protein-protein interactions (PPI), ligand-binding, or metal-binding pockets (Fig 5-29). I therefore looked further into particular cases to discover novel functional miSS. For example, a shift of the donor splice site by 6 nt in the exon 17 of *PUM1* gene results in a deletion of two amino acids (Fig 5-32). This miSS is upregulated in skin, thyroid, adrenal glands, vagina, uterus, ovary, and testis, but downregulated in almost all brain tissues. Only the structure of the miSS-expressing isoform of *PUM1* is accessible in PDB (PDB ID: 1M8X) [237], in which the deletion site maps to an alpha helix. I modelled the structure of maSS-expressing isoform using the I-TASSER web server [201] and found that the alpha helix is preserved, and only its raster is shifted by two amino acids into the preceding loop. The residues in this part of the helix become more hydrophobic, which may influence the overall helix or protein stability. To confirm this, I estimated the stability of the proteins corresponding to the two isoforms using FoldX [202]. The

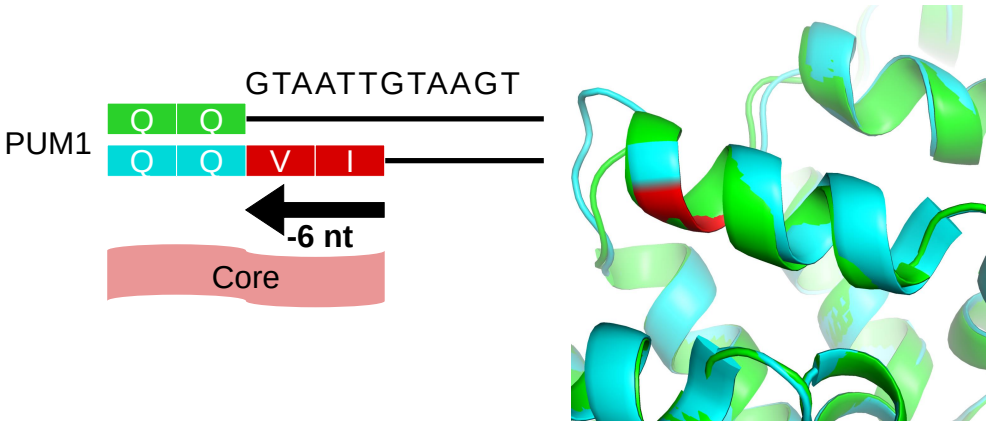

Figure 5-32: **The expression of a donor miSS in the *PUM1* gene.** The expression of a miSS results in the deletion of two amino acids from the core. The miSS-expressing structure is accessible at PDB (green, PDB ID: 1M8X); the maSS-expressing structure was modelled (cyan) and aligned to the miSS-expressing structure with I-TASSER.

difference of +13 kcal/mol between the estimated free energies of the minor and the major isoforms indicates that the minor isoform is less stable due to deletion of two hydrophobic residues.

The expression of the miSS in exon 10 of *ANAPC5* gene results in a 13 amino acids deletion from a protein interaction region (Fig 5-33). I modelled the interaction of these 13 amino acids with the adjacent protein structures using the computational alanine-scanning mutagenesis (CASM) in BAlaS [203]. I found 58 residues (49 residues in the *ANAPC5* protein and 9 residues in the *ANAPC15* protein), which, when mutated to alanine, cause a positive change in the energy of interaction with the 13 amino acid miSS indel region. The miSS is expressed concurrently with the maSS except for the brain tissues, in which the miSS is significantly downregulated. This may indicate the role of the miSS in various pathways in which *ANAPC5* is involved as an important component of the cyclosome [238, 239].

In order to visualize structural classes associated with TASS, I created a track hub supplement for the Genome Browser [240]. The hub consists of three tracks:

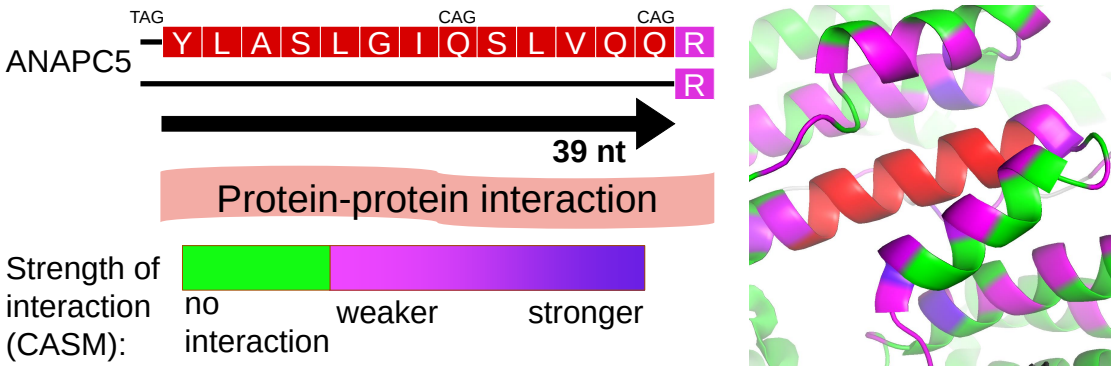

Figure 5-33: **The expression of an acceptor miSS in the *ANAPC5* gene.** The expression of a miSS results in the deletion of 13 amino acids involved in protein-protein interactions. An intermediate splice site (middle CAG) is not significantly expressed. The maSS-expressing structure along with the interacting proteins is accessible at PDB (green, PDB ID: 6TM5); the miSS indel is shown in red. Computational alanine scanning mutagenesis (CASM) in BAlaS [203] identified residues of the neighboring proteins that contribute to the free energy of the interaction with the miSS indel region. The strength of the interaction (the positive change of the energy of interaction) is shown by the gradient color.

location of TASS indels, structural annotation of a nearby region, and tissue specific expression of selected TASS (Fig A-5). The catalogue of expressed miSS is also available in the table format (Table A.13).

5.5 Evolutionary selection and conservation of miSS

In order to measure the strength of evolutionary selection acting on significantly expressed and tissue-specific miSS, and to evaluate how it compares with the evolutionary selection acting on maSS and splice sites outside TASS clusters, I applied a previously developed test for selection on splice sites [208] with several modifications (see chapter 4).

In the coding regions, the strength of negative selection acting to preserve Cn nucleotides in significantly expressed and tissue-specific miSS is comparable to that in maSS and in constitutive splice sites, while no statistically discernible negative selection was detected in miSS that are not significantly expressed (Fig 5-34, A,

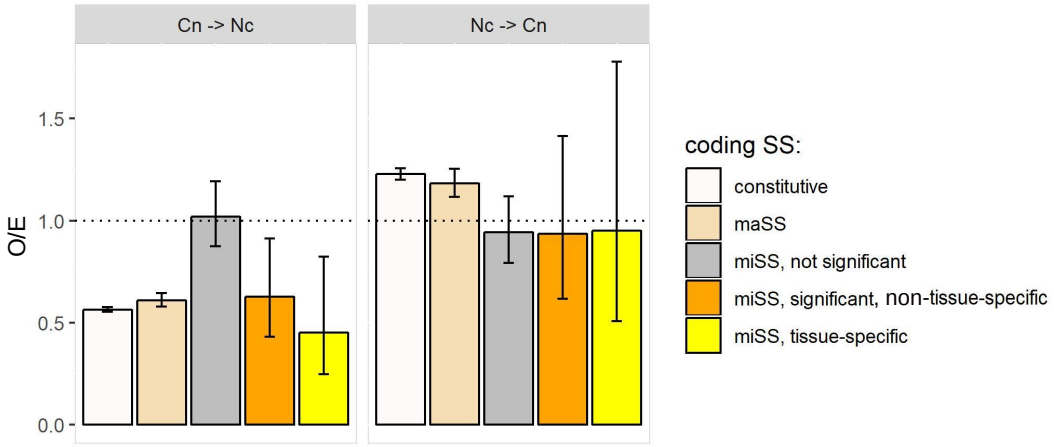

Figure 5-34: **The strength of the selection acting on miSS.** The strength of the selection, defined as the ratio of the observed (O) to the expected (E) number of substitutions, in selected categories of splice sites in the coding regions. The neutral expectation ($O/E = 1$) is marked by a dashed line. The error bars denote confidence intervals for the ratio of two binomial proportions based on likelihood scores [241].

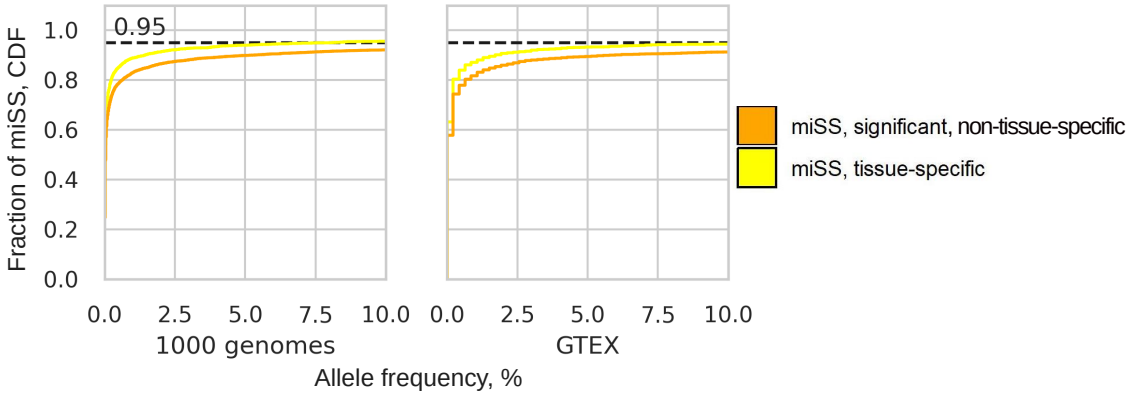

Figure 5-35: **Allele frequencies of SNPs nearby miSS.** The cumulative distribution of allele frequencies of SNPs in splice site consensus sequences and in 35-nt exonic regions adjacent to splice sites from GTEx and 1000 Genomes projects.

left). In contrast, the strength of positive selection, i.e., the O/E ratio for substitutions that create Cn nucleotides, is not significantly different from 1 in all miSS regardless of their expression, while a significant positive selection was detected in maSS and constitutive splice sites (Fig 5-34, A, right). This indicates that the evolutionary selection may preserve the suboptimal state of significantly expressed and tissue-specific miSS relative to its corresponding maSS. Interestingly, I found

that tissue-specific miSS have slightly lower allele frequency of single nucleotide polymorphisms in their splice site sequences and nearby exonic regions compared to non-tissue-specific miSS (Fig 5-35) indicating stronger negative selection acting on the nucleotide sequences of tissue-specific miSS in short-term evolutionary processes [108, 242].

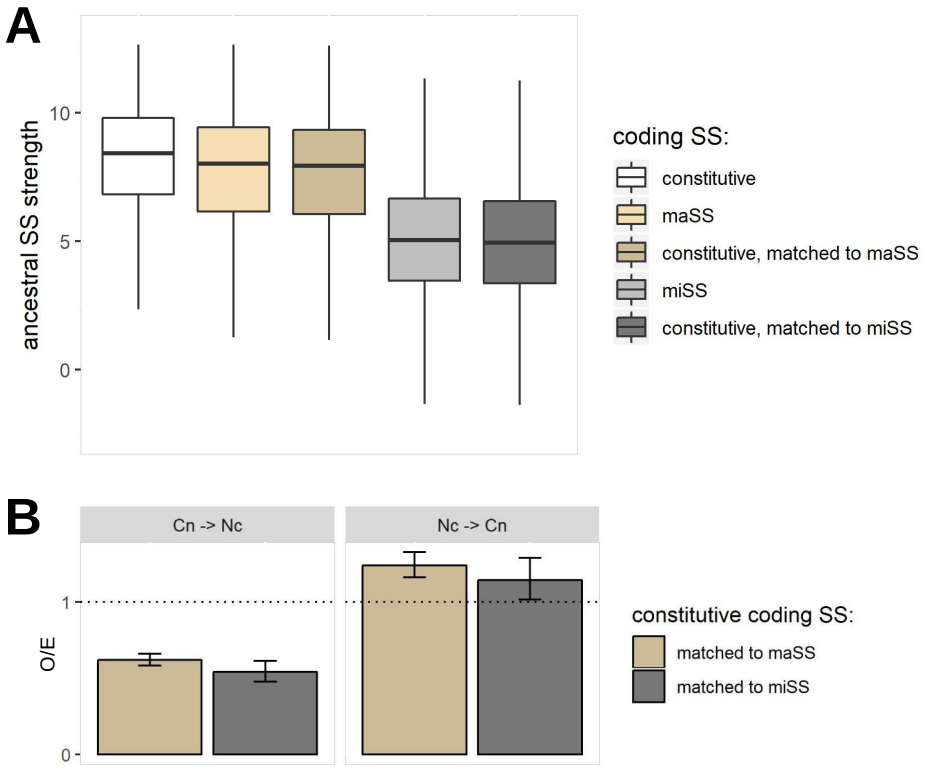

Figure 5-36: **Association between the splice site strength and the selection.** **(A)** The distribution of ancestral strength for different splice site categories. **(B)** The strength of the selection acting on constitutive coding splice sites matched to miSS and maSS by the ancestral splice site strengths.

It was shown previously that the strength of the consensus sequence impacts evolutionary selection acting on a splice site [243, 244]. Indeed, the comparison of ancestral consensus sequences showed that constitutive splice sites and maSS have similar strengths in the ancestral genome (ancestral strengths), while miSS are considerably weaker (MW-test, p-value$< 10^{-15}$, Fig 5-36, A). To control for the influence of the splice site strength on the evolutionary selection, I sampled constitutive splice sites matching them by the ancestral strength with maSS and

with miSS (Fig 5-36, B). However, despite a considerable difference in strengths, I observed no significant difference in evolutionary selection between constitutive splice sites that were matched to maSS and to miSS, indicating that the observed difference in selection acting on miSS and maSS is not due to weaker consensus sequences of miSS.

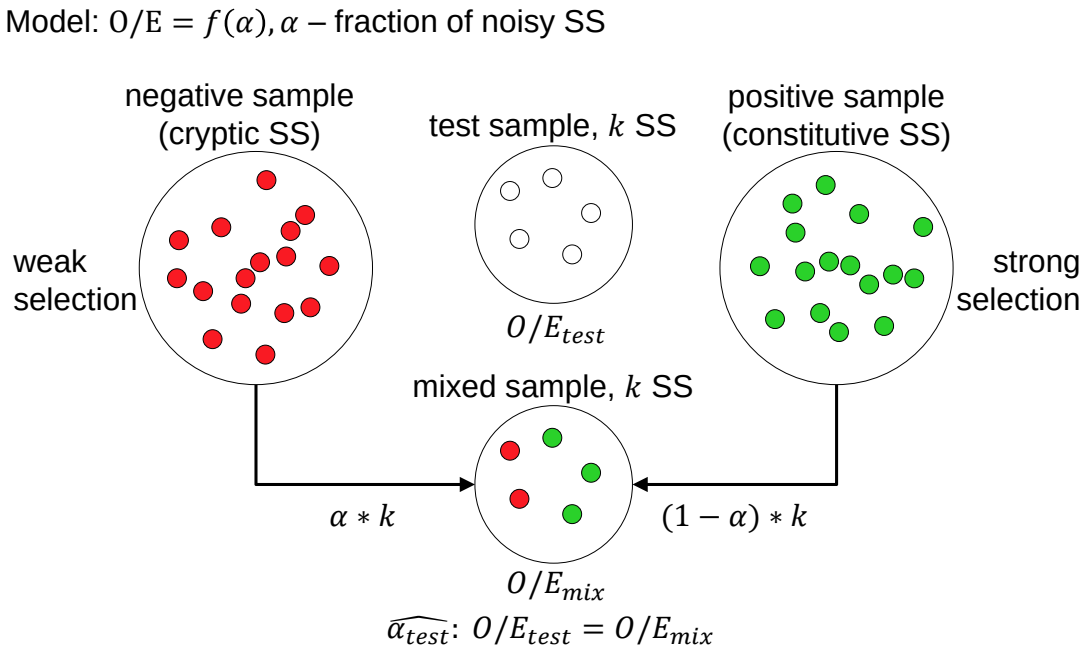

Figure 5-37: **The mixture model for the estimation of the fraction of noisy splice sites** The fraction of noisy splice sites (α) is estimated based on the O/E ratio. A test sample of size k is modelled as a mixture of αk purely noisy (cryptic) splice sites and $(1-\alpha)k$ purely functional constitutive splice sites.

The difference in evolutionary selection between significantly expressed and other miSS could arise from the difference in the fraction of noisy splice sites in these miSS categories. To estimate the fraction of noisy splice sites (α), I constructed a mixture model (Fig 5-37), in which I combined αk splice sites from the negative set of cryptic splice sites and $(1-\alpha)k$ splice sites from the positive constitutive set and measured the strength of evolutionary selection in the combined sample for all values of α. Using this model, I constructed the joint distributions of α and O/E values of Cn-to-Nc substitutions for maSS, significantly expressed non-tissue-specific miSS,

tissue-specific miSS and the rest of miSS (Fig 5-38, left). From these distributions, I estimated 95% confidence intervals for the values of α that correspond to the actual O/E values in the observed samples (Fig 5-38, right; Fig 4-7). The resulting estimates for the fraction of noisy splice sites among maSS, significantly expressed non-tissue-specific miSS, tissue-specific miSS, and the rest of the miSS are <15%, <60%, <54% and >63%, respectively, indicating that at least 46% of tissue-specific miSS are statistically discernible from noise.

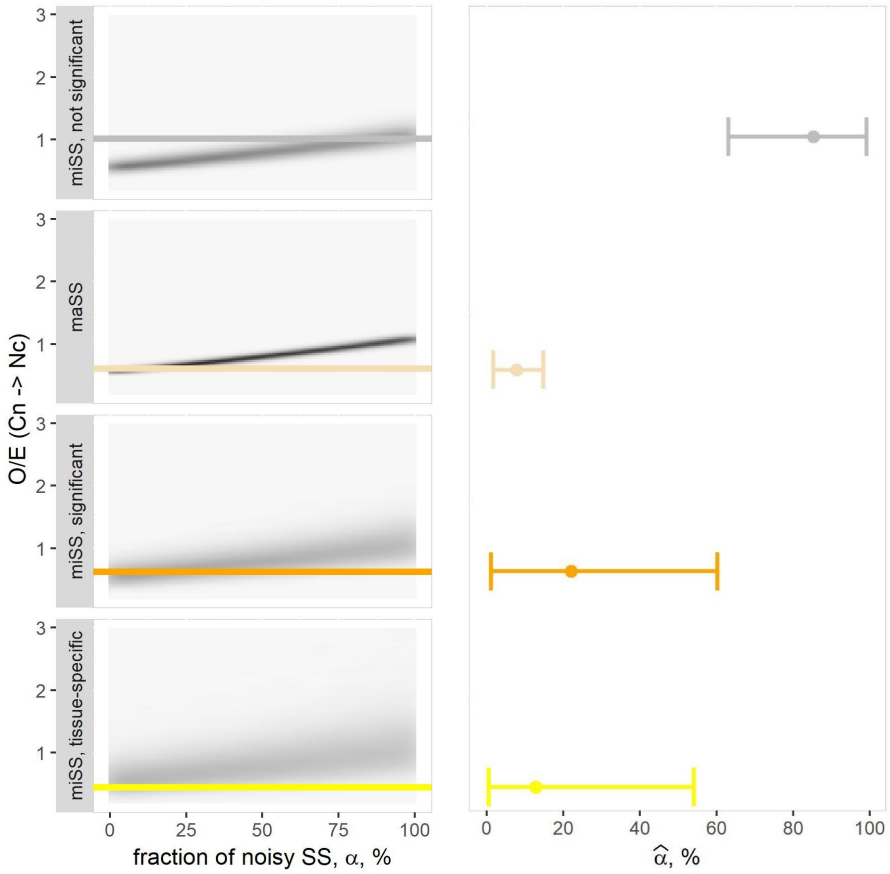

Figure 5-38: **The estimation of the fraction of noisy splice sites.** The estimation of the fraction of noisy splice sites (α) from the observed values of O/E. Bootstrapped joint distribution of O/E and α values (left). The 2.5%, 50% and 97.5% quantiles of the estimated α (right).

Chapter 6

Unproductive splicing

Unproductive splicing is a process, in which gene expression is regulated by alternative splicing producing transcripts with premature termination codons (PTC), which are degraded by the NMD pathway. Despite the annotation of thousands of NMD isoforms in databases, only a few dozen USEs have been experimentally verified, and just a handful have known roles in tissue-specific regulation of gene expression. Here, I perform a systematic analysis of human tissue transcriptomes from the GTEx project to detect concordant tissue-specific changes in alternative splicing of NMD isoforms and gene expression levels, and combine it with the data from ENCODE consortium and SRA archive on the transcriptome response to the perturbation of expression of a large panel of RBPs, RBP footprinting assays, and related proteomic data. I perform an exhaustive literature search to catalog experimentally validated USEs, and show for a number of remarkable cases that, indeed, the regulation of unproductive splicing has strong tissue-specific signatures on the level of splicing, gene expression, and protein expression. These cases include previously proposed models of brain-specific cross-regulation of gene expression via *PTBP1*-controlled unproductive splicing in *PTBP2*, *DLG4* and *GABBR1* genes. In addition, I identify 2831 novel USEs including 568 events, in which NMD isoform inclusion is significantly associated with the downregulation of gene expression. Among latter, this association is manifested tissue-specifically in 86 cases, with cerebellum tissue harboring most of the NMD-promoting events and skeletal muscle tissue harboring most of the NMD-inhibiting events. In more than a half of the cases, I was able to predict at least one

tissue-specifically expressed RBP responsible for such regulation from RBP perturbation experiments. I present a high-confidence set of 31 novel predicted examples of regulated USEs, including *PTBP1*-controlled events in the *DCLK2*, *IQGAP1*, and *ACSF3* genes, which may be responsible for the tissue-specific inhibition of their host gene expression. These results represent an invaluable resource for molecular biologists and bioinformaticians studying unproductive splicing.

6.1 Poison and essential USEs

In what follows, I consider four major types of AS events, namely cassette (skipped) exons, alternative 5'- and 3'-splice sites, and retained introns. They are classified according to the poison-essential dichotomy as it was done earlier [245] (Fig 4-8). An USE is referred to as poison (essential) if the NMD isoform represents an insertion (deletion) with respect to the protein-coding isoform. For instance, poison exons are normally skipped, but they trigger degradation by NMD when included in the mRNA. Numerous poison exons have been described in genes encoding SR proteins [144]. The reciprocal case is essential exons, which are normally included in the mRNA, but induce a PTC when skipped. Examples of essential exons are found in genes encoding hnRNP proteins, including *PTBP1* [152], *PTBP2* [153], and *FUS* [154]. Similarly, an alternative splice site or intron is referred to as poison (essential) if PTC occurs as a result of insertion (deletion) of a mRNA fragment. Examples of these USEs have also been described [136, 245, 164, 246].

The characterization of USEs and their regulation was based on the following strategy (Fig 6-1). In cross-regulation, I expect a negative association between ψ (see chapter 4) and gene expression level. Therefore, an RBP that inhibits the NMD isoform must upregulate the expression level, and an increase in the expression level of such RBP must be accompanied by an increase in the expression level of the target gene and a decrease in ψ. Conversely, an RBP that activates the NMD isoform must downregulate the expression level, and an increase in the expression level of such RBP will lead to a decrease in the expression level of the target gene and an increase in ψ. In auto-regulation, I consider only negative feedback loops, in which an RBP

activates the NMD isoform, increases ψ, and downregulates its own expression level.

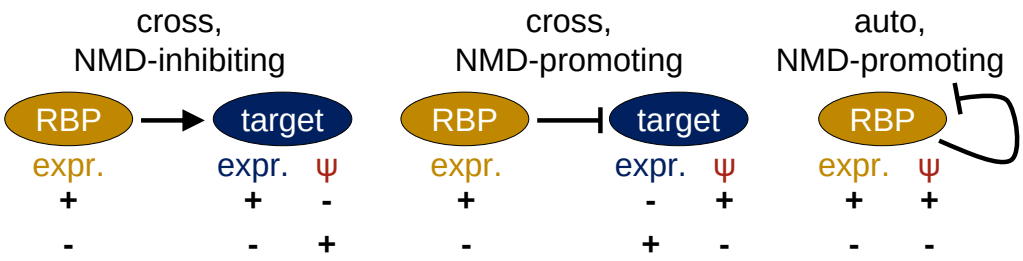

Figure 6-1: **The expected changes of AS and gene expression level of the target caused by high (+) and low (−) expression levels of the regulator (RBP).**

6.2 Validated and annotated USEs

I performed an exhaustive literature search to catalog experimentally validated USEs (Table A.14). In particular, I collected information on experimental outcomes and indicated whether the USE is reactive to NMD inactivation, whether RBP perturbations affect the NMD isoform inclusion or gene expression on protein and mRNA levels, and whether there is evidence of binding of the targeted mRNA by RBPs. I obtained 237 RBP-USE pairs containing 48 RBPs and 57 USEs, including 203 cross-regulatory and 34 auto-regulatory cases for which the experimental validation was reported in human cell lines, mouse cell lines, or tissues. In what follows, these USEs will be referred to as validated. A snapshot of these findings is shown in Fig 6-2 and also in Fig 6-3, which illustrates a particularly dense subnetwork for SR proteins.

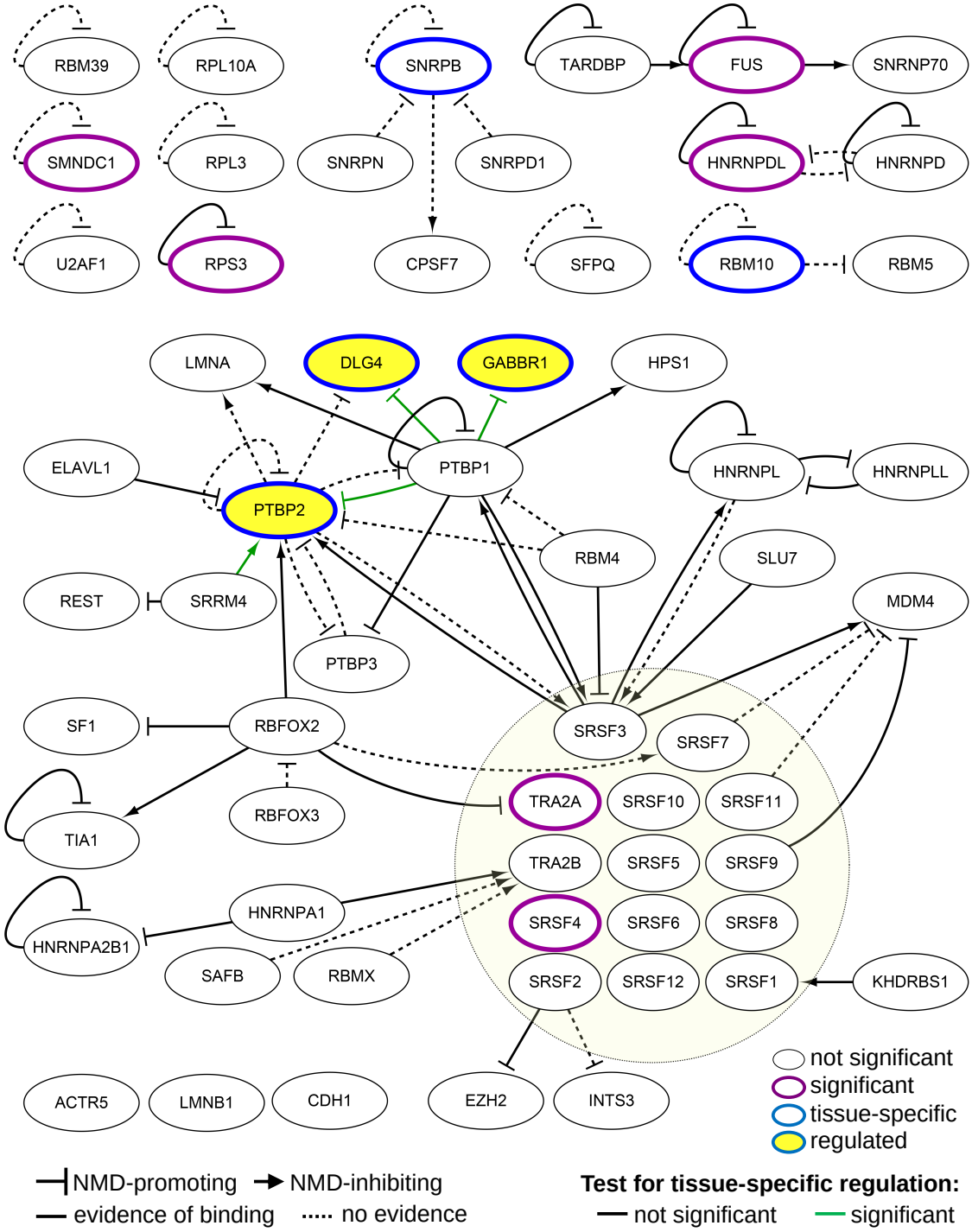

Figure 6-2: **A regulatory network of validated USEs.** A subnetwork of SR proteins is exempt to Fig 6-3. The nodes represent USEs listed in Table A.17. The edges represent NMD-promoting and NMD-inhibiting regulatory connections. The color code of the nodes represents the classification of USEs as significant or tissue-specific (Table 6.1). The color code of the edges represent the result of the tissue-specificity test.

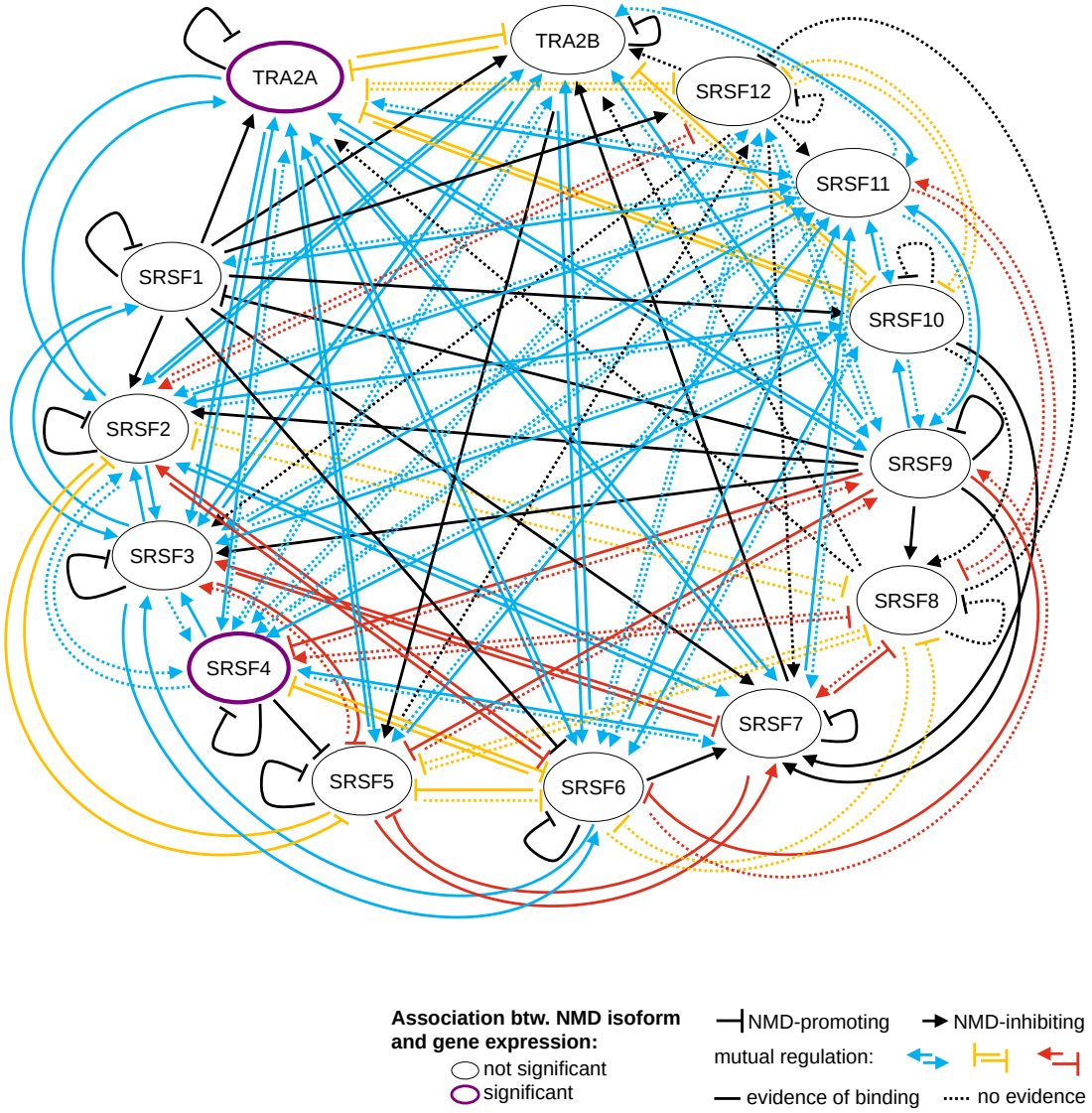

Figure 6-3: **A regulatory subnetwork of SR proteins.** A regulatory subnetwork of SR proteins according to the data from Table A.17. The legend is as in Fig 6-2.

Among four considered AS types (Fig 4-8), I identified 5,309 AS events that switch from protein-coding to NMD isoform (Table A.3). These USEs are referred to as annotated. As in previous works [245], I used publicly available NMD inactivation experiments followed by RNA-seq to check whether the annotated USEs indeed produce NMD-sensitive transcripts. In accordance with the literature, 36 of 45 (80%) validated USEs with sufficient read support demonstrated a significant increase of ψ upon NMD inactivation in at least one experiment. However, only 1,435 of 3,196 (45%) annotated USEs did so, indicating that the evoked response

to NMD inactivation may not always be a good predictor of unproductive splicing. Therefore, I treated the reactivity to NMD inactivation as an additional evidence and didn't use it as a constraint.

6.3 Association of USEs with gene expression

Table 6.1: The number of unproductive splicing events.

USE group	Validated	Annotated	Total
All	48	2,831	2,879
Significant	11	568	579
Tissue-specific	5	86	91
Regulated	3	47	50
CLIP in the gene	3	31	34
Local CLIP support	3	14	17

To study the association between unproductive splicing and gene expression at the mRNA level, I analyzed RNA-seq data from about 8.5 thousand GTEx samples, first disregarding their tissue attribution. Samples from testis were excluded since the action of the NMD system in this tissue is different from that in other tissues [247, 248, 249]. After the removal of the USEs with low variability and low read support, 2,831 out of 5,309 annotated cases and 48 out of 57 validated USEs remained (Table 6.1). To detect concordant changes between the NMD isoform splicing rate and the gene expression level, I characterized each USE by $\psi_H - \psi_L$, the difference of median ψ values between the upper and the lower quartiles of the ψ distribution, Δe_g and Δe_l, the difference of medians of the global and local expression levels, and their respective z-scores (see chapter 4). A comparison with AS events that don't generate NMD isoforms revealed a set of 579 USEs with z-score below -5, which I refer to as significant (Table A.15). Among them, there were 11 validated cases, most of which showed a remarkable magnitude of ψ changes (Fig 6-4, A).

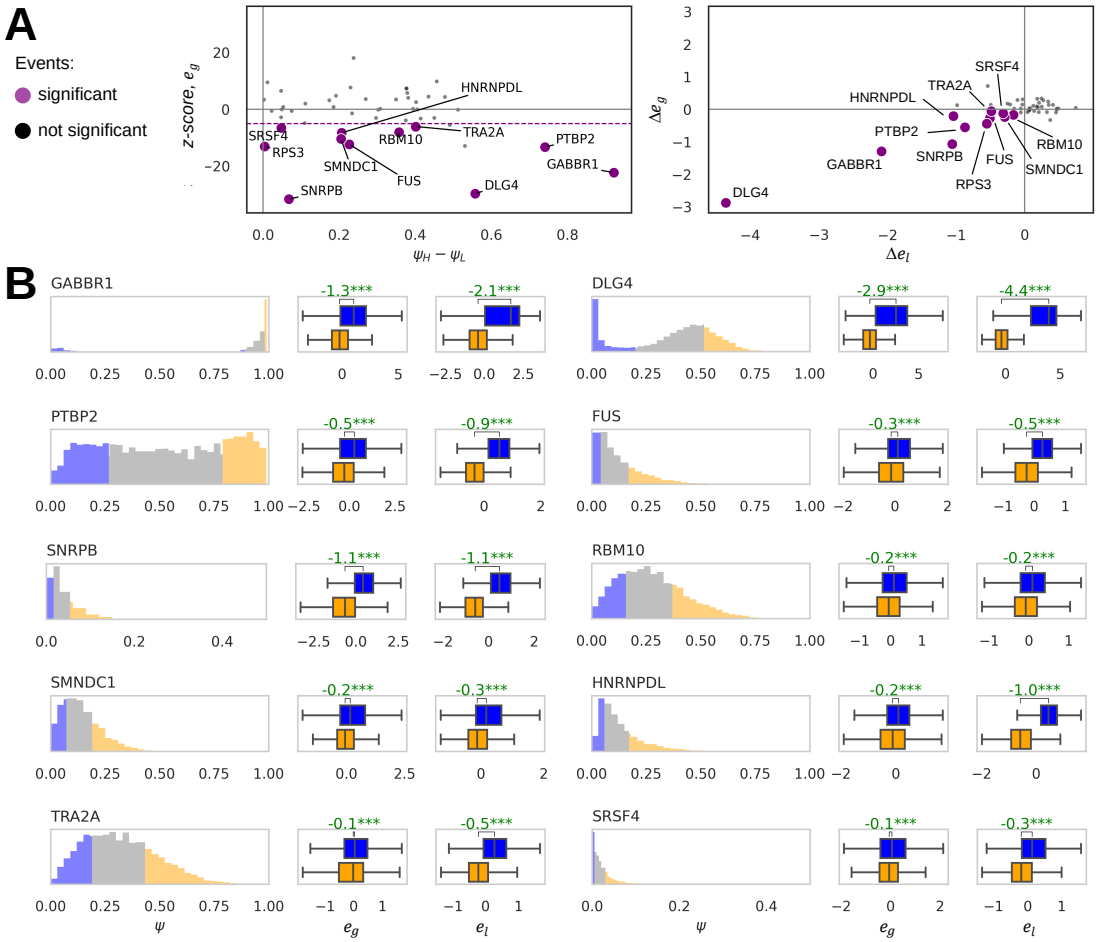

Figure 6-4: **Significance of validated USEs. (A)** Significant USEs are characterized by the z-score of e_g below -5, and negative Δe_g and Δe_l values. $\psi_H - \psi_L$ denotes the difference of the medians in the upper and the lower quartiles of the ψ distribution. **(B)** The distribution of ψ values for selected USEs from panel (A). The top 25% and the bottom 25% of ψ values are colored in orange and blue, respectively. The box plots represent the distribution of e_g and e_l in the two groups. Shown in green are Δe_g and Δe_l values. Asterisks (***) denote a statistically discernible difference at the 0.1% significance level.

The strongest association was observed for the well-documented *PTBP1* targets such as *GABBR1* and *DLG4*, in which $\psi_H - \psi_L$ exceeded 50% and was accompanied by more than a twofold decrease of the expression level (Fig 6-4, B). Other examples include USEs in *RBM10*, an important component of the spliceosome that is associated with genetic diseases and cancer [250, 251, 252], and *TRA2A*, a member of the SR protein family also known as an oncogene [253, 254, 255]. In *RBM10*, skipping of

the essential exon 6 leads to a decrease in expression, which promotes proliferation in lung adenocarcinoma [256, 135]. In *TRA2A*, suppression of expression occurs through the stimulation of the poison exon 2 in its mRNA by the *TRA2A* gene product and its paralog *TRA2B* [151].

6.4 Tissue-specific regulation of USEs

To explore tissue-specificity of the association between splicing rate and gene expression for significant USEs, I estimated the deviations of the median ψ value and the median e_g in each tissue from the respective pooled medians and compared the number of tissues, in which these deviations had opposite signs, with the number of tissues, in which they had the same sign (see chapter 4). Under the null hypothesis that deviations with the same and the opposite signs are equally likely, this procedure yielded 91 USEs with significantly negative tissue-specific associations (Table A.16). In what follows, I refer to these USEs as tissue-specific (Table 6.1).

On the other hand, a number of large-scale functional assays have assessed the response of the cellular transcriptome to perturbations of RBP expression levels [210]. I collected a panel of 419 such experiments followed by RNA-seq for 248 RBPs (Table A.4). For each RBP, I assessed the concordance of changes in the NMD isoform splicing rate and changes in the target gene expression using multiple experiments and selected RBP-USE pairs, in which these changes had opposite signs (see chapter 4). In 124 out of 203 validated cross-regulatory cases, for which this assessment was possible, I identified 30 RBP-USE pairs with the expected regulatory outcome and only one pair, in which the direction of the regulation was opposite to that reported in the literature (Table A.17). In application to tissue-specific USEs, it yielded at least one regulator in 74 cases, in 50 of which tissue-specific expression of the regulator was accompanied by tissue-specific ψ changes. In what follows, I refer to these USEs as regulated (Table 6.1).

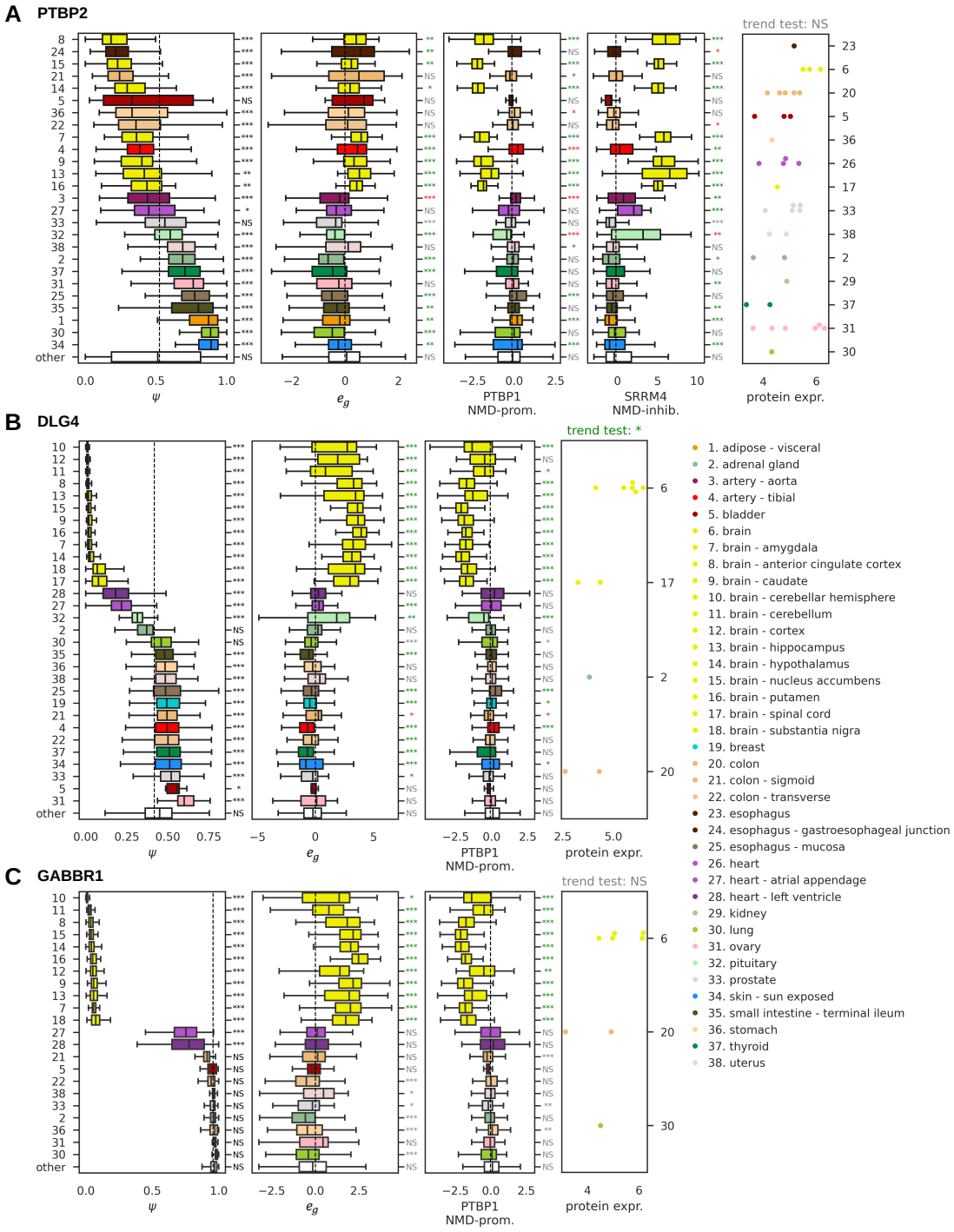

Figure 6-5: **Examples of validated tissue-specific USEs.** Panels (A), (B), and (C) show the results for *PTBP2*, *DLG4*, and *GABBR1* genes, respectively. In each case, boxplots show (left to right) the distribution of ψ, global gene expression level (e_g), the expression levels of regulators, and the protein expression level. Statistically significant deviation from the pooled median are marked with asterisks (***FDR $<$ 0.001; **FDR $<$ 0.01; *FDR $<$ 0.05; NS not significant). The color of an asterisk shows whether the deviation is in the expected (green) or the opposite direction. Tissues are denoted by numbers and sorted in ascending ψ order.

To further confine the list of candidates, I required that the RBP bind the target pre-mRNA. To check this, I used high-throughput crosslinking and immunoprecipitation (CLIP) experiments provided in the POSTAR3 database [212] and found evidence of binding by at least one predicted regulator for 34 regulated USEs, including 17 cases with footprints in the immediate proximity of the USE (Table 6.1, Table A.18).

This analysis reconfirmed brain-specific USEs in the *PTBP2*, *DLG4*, and *GABBR1* genes, in which the NMD isoform is induced by *PTBP1*-mediated skipping of an essential exon (Fig 6-5). These USEs are known to upregulate gene expression by inhibiting the NMD isoform expression in the brain and activating it in the other tissues [153, 172, 170]. In accordance with this, I observed a statistically significant association with brain-specific expression of *PTBP1* in all three cases and a significant association with the expression of *SRRM4* for *PTBP2*. To further confirm the observed patterns, I used proteomic data to demonstrate that protein expression is increased in the brain, with a particularly strong trend in the case of *DLG4*.

I next considered 34 regulated USEs with CLIP support in the gene and clustered them according to their AS rates (Fig 6-6). The clustering revealed three USE groups of approximately equal sizes, which were characterized by a decreased ψ in the brain (cluster 1), a decreased ψ in skeletal muscle and heart (cluster 2), and an increased ψ in the brain (cluster 3). These USEs form a regulatory network, in which *PTBP1* controls the expression of 15 targets (Fig 6-7). In cluster 1, *PTBP1* stimulates brain-specific gene expression by activating the NMD isoform in non-neural tissues, as it also does in the case of *PTBP2*, *GABBR1*, and *DLG4*.

Figure 6-6: **Clustering diagram of 34 regulated USEs with CLIP support in the gene.** Clustering diagram of 34 regulated USEs with CLIP support in the gene based on tissue-specific deviations of ψ. The clusters are characterized by a decreased ψ in the brain (cluster 1), a decreased ψ in skeletal muscle and heart (cluster 2), and an increased ψ in the brain (cluster 3).

A previously unknown example of USE regulation was found in the *DCLK2* gene, the product of which is necessary for the development of the hippocampus and the regulation of dendrite remodeling [257, 258]. I predict that *PTBP1* stimulates the inclusion of the poison exon 16, which leads to the suppression of *DCLK2* expression

with a significant effect observed both at the mRNA and protein levels (Fig 6-8, A). In this case, the presence of CLIP peaks of *PTBP1* both upstream and downstream of the poison exon indicates a non-canonical role of *PTBP1* as a splicing activator, consistently with the reports on the dual action of *PTBP1* that depends on its binding position relative to the exon [259].

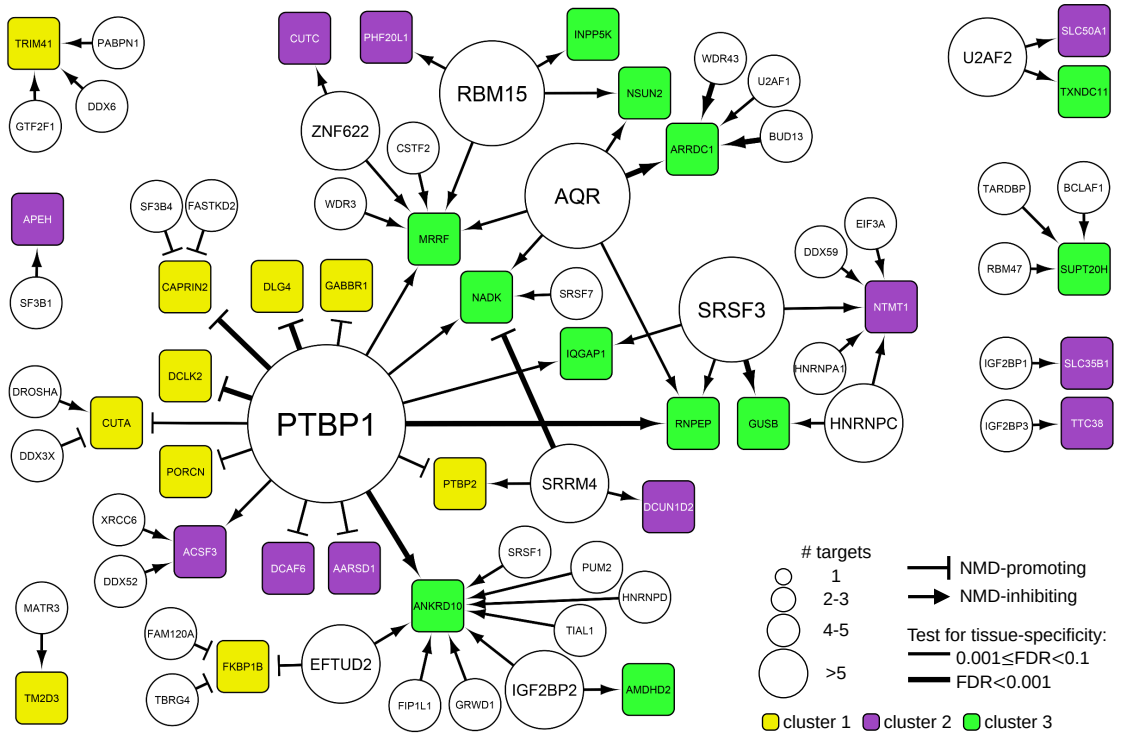

Figure 6-7: **The predicted network of regulated USEs with CLIP support in the gene.** The nodes and edges are as in Fig 6-2. The nodes from clusters 1-3 are labeled with the same colors as in Fig 6-6.

PTBP1 may cooperate with other regulators such as *SRSF3*, *RBM15* and *AQR* to inhibit NMD isoforms in genes from cluster 3, e.g. *IQGAP1* (Fig 6-8, B). The expression of *IQGAP1* is maintained at low levels in healthy brain tissues, but in brain tumors it is increased [260]. I confirm a brain-specific decrease of *IQGAP1* expression that is associated with increased poison exon 29 inclusion and reduced expression of *PTBP1* and *SRSF3*, most remarkably in the cerebellum. *PTBP1* may also regulate cluster 2 targets such as *ACSF3*, where it promotes the inclusion of

an essential exon together with *DDX52* and *XRCC6* (Fig 6-8, C). The product of *ACSF3* is necessary for mitochondrial metabolism, and hence it is upregulated in tissues where cells are rich in mitochondria, including liver, muscle, and heart [261, 262], where *PTBP1*, *DDX52*, and *XRCC6* are upregulated.

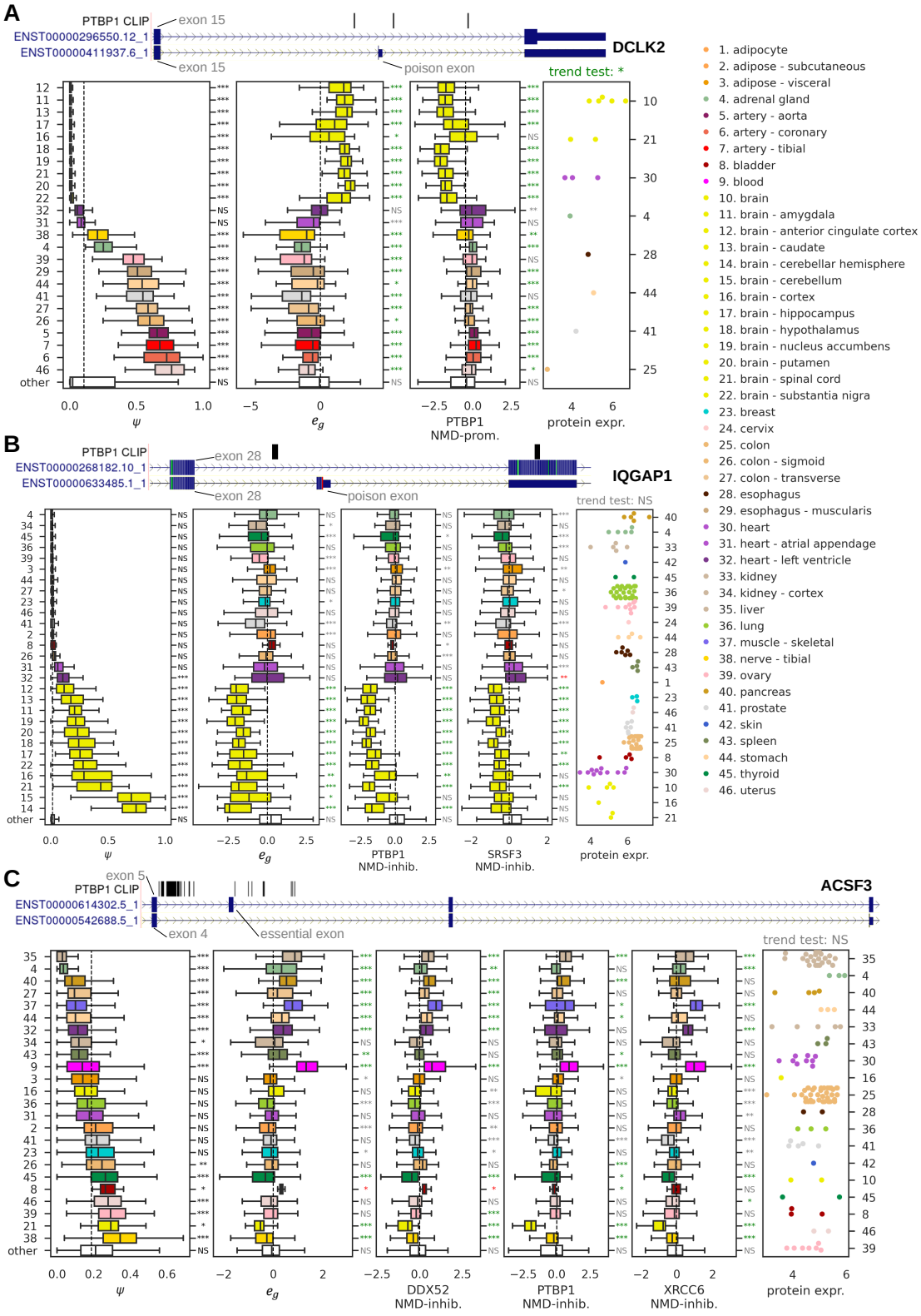

Figure 6-8: **Examples of novel tissue-specific USEs.** Panels (A), (B), and (C) show the results for *DCLK2*, *IQGAP1*, and *ACSF3* genes, respectively. The ideograms in each panel show the USE and CLIP peaks of *PTBP1*. The rest of the legend is as in Fig 6-5.

6.5 Tissues with frequent USE regulation

I next asked whether unproductive splicing is more active in some tissues than in the others. Concordant changes between ψ and gene expression level could arise by pure chance due to large fractions of upregulated NMD isoforms and downregulated genes. For instance, tibial nerve, thyroid, and prostate are characterized by simultaneous upregulation of NMD isoforms and downregulation of expression in many genes, while in skeletal muscle and heart, the pattern is the opposite (Fig 6-9).

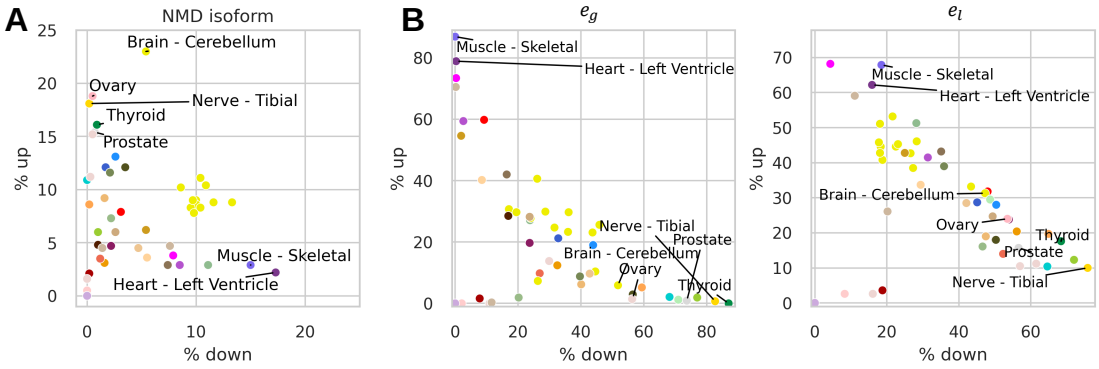

Figure 6-9: **Characterization of tissues by the number of up- and downregulated USEs and genes. (A)** The fraction of significant USEs that are upregulated (Y-axis) and downregulated in a tissue (X-axis). The color code for tissues is as in (Fig 6-8). **(B)** The fraction of significant USEs in which gene expression (e_g,left, and e_l, right) is upregulated (Y-axis) or downregulated (X-axis) in a tissue.

To address this, I performed a χ^2-test for association between up- and downregulation of the NMD isoform and up- and downregulation of gene expression level in each tissue and compared the observed and the expected number of USEs (Fig 6-10). The observed frequencies of negative concordant changes between NMD isoform splicing rate and gene expression level significantly exceeded the expected frequencies in cerebellum, most brain subregions, and ovary, but not in thyroid, prostate, or tibial nerve. These findings indicate a particularly strong association between unproductive splicing and gene expression in the brain and support the importance of NMD in brain development and disease [263]. In support of this finding, the functional enrichment analysis in DAVID [220] web-server detected the overrepresentation of brain-specific genes within those having at least one tissue-specific USE,

although the effect is not statistically significant (FDR>0.1, Table A.19).

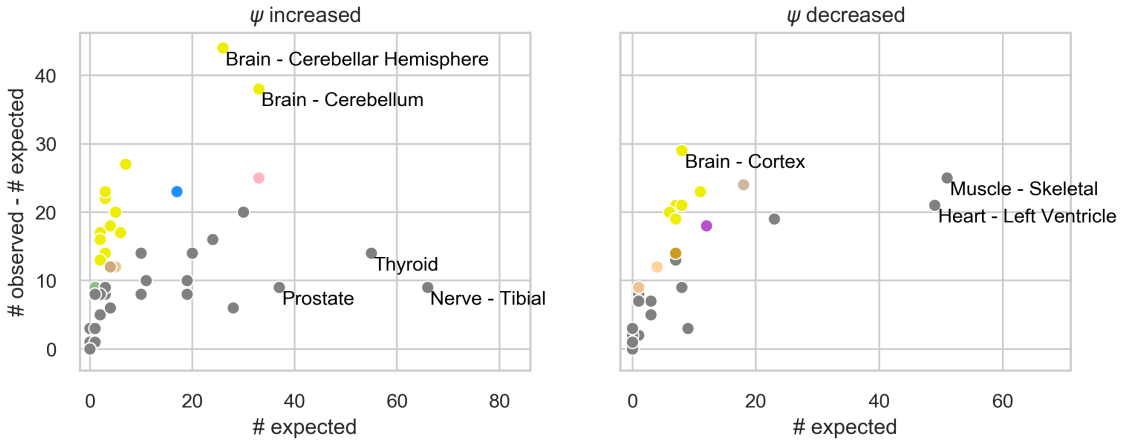

Figure 6-10: **The abundance of tissue-specific USEs.** The excess of observed tissue-specific USEs vs. the number of expected tissue-specific USEs with positive (left) and negative (right) deviations of ψ from the pooled median. Tissues with FDR < 0.05 are colored as in Fig 6-8. Tissues with no significant excess of observed tissue-specific (FDR ≥ 0.05) are colored in gray.

Chapter 7

Discussion

7.1 Regulation of tissue-specific TASS

Increasing amounts of high-throughput RNA-seq data have uncovered the expanding landscape of human alternative splicing [264]. Presented here is the most complete up-to-date catalogue of 45,739 miSS, of which 9,303 are significantly expressed in human tissues according to GTEx data. It significantly extends the TASSDB2 database constructed based on the evidence from ESTs [214] by adding ∼18k miSS (Fig A-6, A), which are enriched with significantly expressed miSS despite being weakly expressed on average (Fig A-6, B, C). It also adds data on specific TASS classes such as NAGNAGs [69] and GYNNGYs [115]. On the one hand, the number of detected TASS is reaching a plateau with increasing the number of GTEx samples up to 8,548 (Fig A-7) indicating that this catalogue is close to complete. On the other hand, a substantial fraction of TASS are noisy (more than 63% among miSS that are not significantly expressed) reflecting the natural tradeoff between sensitivity and specificity of bioinformatics approach.

While the majority of miSS in the coding regions are located downstream of their respective maSS, the upstream miSS tend to be expressed stronger, i.e., the spliceosome tends to systematically choose a miSS that is located upstream. This pattern likely results from the linear scanning mechanism, in which the spliceosome traverses the pre-mRNA in the 5' to 3' direction so that tandem splice sites follow first come-first served principle [265, 266]. The relative expression is the strongest

for the frame-preserving acceptor miSS in support of the observation that transcript isoforms with frame-disrupting miSS are suppressed by NMD. We therefore expected that frame-disrupting miSS would be rare among significantly expressed and tissue-specific miSS. However we observe that almost a half of tissue-specific coding miSS disrupt the reading frame (Table A.7). Furthermore, frame-disrupting tissue-specific miSS are more conserved than non-significantly expressed miSS (Fig 5-12, B) indicating a potential function such as, for example, fine-tuning of gene expression levels via NMD [245, 146, 159].

The analysis of tissue-specificity recapitulated patterns previously observed in the analysis of alternative splicing [223], such as high abundance of tissue-specific miSS in testis, cerebellum, and other brain tissues (Table A.9). At that, the frequency of brain-specific events might be underestimated because samples representing brain tissues make up a higher fraction than any other tissue [176] introducing the bias in brain-specific ϕ and expression levels towards median values calculated across all samples.

Previous reports indicated that strongly expressed miSS located at a distance of 3, 6, 9 nt from the maSS in coding regions, which are to a large extent equivalent to the significantly expressed miSS introduced here, are overrepresented in disordered protein regions [111]. The evolutionary selection against alternatively spliced NAGNAGs in protein-coding genes is stronger in structured regions than in disordered regions [111]. Here, we extended this result by showing that tissue-specific miSS are even more enriched in disordered protein regions than other significantly expressed miSS. Furthermore, we showed that tissue-specific miSS are associated with SLiMs and post translational modification sites. While there is no positive selection for Cn nucleotides among neither significantly expressed nor other miSS, we observed a strong negative selection acting to preserve Cn nucleotides in the former (Fig 5-34). This finding can be explained by the tendency of functional miSS to preserve the suboptimal state relative to maSS, i.e functional miSS are evolutionarily conserved and maintain their Cn nucleotides, but they also do not harbor more Cn nucleotides not to outcompete maSS. Furthermore, we showed a tendency of many tissue-specific miSS to be regulated by RBPs, e.g., the miSS in the exon 6 of the

QKI gene is likely regulated by *PTBP1* (Fig 5-22). All these findings are indicative of a functional role of at least a proportion of miSS.

It has been demonstrated that cell-type-specific alternative splicing within the same tissue may affect a large fraction of multi-exon genes and govern cell fate and tissue development [267, 268]. For example, pairs of exons in genes *GRIA1* and *GRIA2* follow a strict mutually-exclusive pattern between different neuronal types [269]. In this study, we examined miSS expression in primary cells from different tissues and identified hundreds of cell-type-specifically expressed miSS. While the comparison of expression profiles suggested that miSS expression and its regulation by RBPs depends to a greater degree on the cell type than on the tissue of origin, both local (cell type) and global (tissue) gene expression environments can contribute to the specificity of miSS usage (Fig 5-27,Fig 5-28).

The observations made in this manuscript are based on the analysis of RNA-seq data from the GTEx project [176]. It is hence worthwhile to address the question which proportion of these alternative splicing events translate to the protein level. Direct measurement of this proportion by, for example, shotgun proteomics is not instructive for many reasons, including limited coverage and low sensitivity of such experiments [270], as well as the fact that the cleavage site consensus of a widely used trypsin protease overlaps with the amino acid sequence induced by the splice site consensus, thus producing non-informative peptides [271]. This question has been debated in the literature [67, 119, 120]. On the one hand, proteomics data support the expression of a single predominant protein isoform for most human genes [67]. On the other hand, ribosome profiling suggests translation of alternative isoforms [272], and experimental studies demonstrate the functional importance of alternative splicing in modulation of protein-protein interactions [273]. Our study adds to this debate in that we have collected multiple lines of evidence that support expression on the protein level and functional importance of TASS-related isoforms. Our estimate of significantly expressed miSS largely exceeds the conservative estimate of proteomics-supported alternative splicing events [67]. Our analysis of Ribo-Seq experiments supports their expression, and in many cases this expression is tissue-specific. We also showed that significantly expressed miSS, as well as maSS,

are under negative selection pressure. Finally, our analysis confirms that sites in protein sequence that correspond to TASS events are depleted from structured protein regions, just as for alternative splicing events in general [229, 274], which also suggests their non-neutral evolution and hence functionality on the protein level. In line with previous research [274], we demonstrated that when located in disordered protein regions, TASS-associated events often affect sites of post-translational modification.

7.2 Regulation of tissue-specific unproductive splicing

A coordinated interaction between AS and NMD that leads to the degradation of specific mRNAs, which constitutes unproductive splicing, is a widespread phenomenon that has been documented in almost all eukaryotes [275, 276, 146, 150]. Compared to other post-transcriptional regulatory mechanisms such as endogenous RNA interference [277, 278, 279, 280] and the control of mRNA stability by RNA modifications [281, 282, 283], unproductive splicing appears to act pervasively at the transcriptome level, as evidenced by the fact that nearly a third of human protein-coding genes have at least one annotated NMD transcript isoform [284].

The examples of *PTBP2*, *DLG4*, and *GABBR1* demonstrate as a proof of principle that tissue-specific unproductive splicing can be inferred from the transcriptomic data. Although we observed a significant association between NMD isoform splicing rate and gene expression level in many cases (Fig 6-4, B), the majority of validated USEs exhibit no convincing evidence of such association. In part, this lack of response is explained by a bias in observing splicing changes that are masked by rapid degradation of unproductive isoforms by the NMD pathway and constant influx of pre-mRNAs due to ongoing transcription [161]. The unproductive splicing could also be inactive in fully differentiated tissues and operate only in specific conditions such as differentiation [285, 147], neurogenesis [172, 153], or hypoxia [286]. In fact, most splicing factors demonstrate only minor differences in expression between tissues [81] in comparison to KD or OE experiments, in which the validated USEs

have been originally described. The regulatory potential of unproductive splicing requires a relatively large dynamic range of AS changes, which may be restricted in the mature tissues. All these factors inherently limit the sensitivity of the proposed method.

A fundamental challenge in studying unproductive splicing is the existence of feedback loops and indirect connections in the network. For instance, in the case of autoregulation, the product of a gene that harbors a poison exon counteracts its own upregulation by promoting exon inclusion, thus leading to the downregulation of gene expression level and, consequently, to suppression of the NMD isoform that was initially upregulated. Splicing regulatory networks may contain circuits that serve as proxies and interfere with direct interactions, e.g., *PTBP1* upregulates *SRSF3*, which in turn upregulates *PTBP2*, but *PTBP1* itself directly downregulates *PTBP2* (Fig 6-2). As a result, the sign and the magnitude of the association between NMD isoform splicing rate and RBP expression may vary depending on the connectivity in the network potentially leading to both false positive and false negative predictions.

The methodology outlined here aims at the discovery of tissue-specific USEs. The core part of it is based on a statistically robust test that compares gene expression levels in the extremes of ψ distribution rather than on the analysis of correlations, thus estimating not only significant but also substantial and explicit differences. It doesn't depend on the tissue attribution in GTEx and can be applied to other panels of RNA-seq experiments. The other parts measuring the response of the transcriptome to the perturbations of RBP expression levels in combination with RBP footprinting provide the evidence of causality for the proposed regulatory circuits and can be used independently of the first part. Modular organization of this workflow allows discovery of regulated USEs based on their splicing and expression signatures in other transcriptomic datasets, although not all aspects of these signatures are currently well understood.

For instance, the response of the validated USEs to the inactivation of NMD system components and perturbations of RBP expression levels in cell lines doesn't always agree with the signatures that are observed in tissues. In some cases, the activation of unproductive splicing in tissues is concordant with the changes in

mRNA expression levels but inconsistent with protein expression levels and vice versa. The analysis of RBP perturbation experiments consistently demonstrated nearly the same number of positive and negative associations between ψ and e_g in the validated USEs implying the existence of positive feedback loops, the physiological relevance of which is debatable [245]. Further analysis of the molecular mechanisms underlying unproductive splicing and larger volumes of multi-omics data will be required to address these concerns in future studies.

Chapter 8

Conclusion

Large panels of transcriptomic data represent an invaluable resource for studying gene expression at the level of alternative splicing. This dissertation aimed to study two classes of aberrant splicing events, which represent large fractions of alternative isoforms in the human transcriptome, TASS and USEs. The main conclusions of this project are as follows:

- While the majority of aberrant splice isoforms represent splicing noise, there is still a substantial fraction of events showing evidence of tissue-specific regulation by RBPs and other features indicating their functional importance.

- A large number of tissue-specific TASS affect structured protein regions and may adjust protein-protein interactions or modify the stability of the protein core, including TASS in *PUM1* and *ANAPC5* genes.

- A systematic analysis has demonstrated for the first time the involvement of unproductive splicing in the regulatory circuits that define tissue-specificity of gene expression.

- I confirmed tissue-specific patterns that were previously proposed for USEs in *PTBP2*, *DLG4*, and *GABBR1* and identified several dozen novel USEs having evidence of tissue-specific regulation through unproductive splicing by trans-regulators.

- Many USEs are specific to the brain or skeletal muscle tissues and are likely controlled by PTBP1, e.g. those in *DCLK2*, *IQGAP1* and *ACSF3*.

These results greatly expand the current knowledge on the function of alternatively spliced transcripts and represent a useful resource for molecular biologists and bioinformaticians studying alternative splicing and gene expression regulatory pathways.

Bibliography

[1] A. Herbert and A. Rich. RNA processing and the evolution of eukaryotes. *Nat Genet*, 21(3):265–269, Mar 1999.

[2] M. Deutsch and M. Long. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*, 27(15):3219–3228, Aug 1999.

[3] X. Hong, D. G. Scofield, and M. Lynch. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol*, 23(12):2392–2404, Dec 2006.

[4] E. A. Grzybowska. Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem Biophys Res Commun*, 424(1):1–6, Jul 2012.

[5] R. Jorquera, C. González, P. Clausen, B. Petersen, and D. S. Holmes. Improved ontology for eukaryotic single-exon coding sequences in biological databases. *Database (Oxford)*, 2018:1–6, Jan 2018.

[6] F. Pagani, E. Buratti, C. Stuani, M. Romano, E. Zuccato, M. Niksic, L. Giglio, D. Faraguna, and F. E. Baralle. Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element. *J Biol Chem*, 275(28):21041–21047, Jul 2000.

[7] R. Sinha, Y. J. Kim, T. Nomakuchi, K. Sahashi, Y. Hua, F. Rigo, C. F. Bennett, and A. R. Krainer. Antisense oligonucleotides correct the familial dysautonomia splicing defect in IKBKAP transgenic mice. *Nucleic Acids Res*, 46(10):4833–4844, Jun 2018.

[8] K. R. Brunden, J. Q. Trojanowski, and V. M. Lee. Advances in tau-focused drug discovery for Alzheimer's disease and related tauopathies. *Nat Rev Drug Discov*, 8(10):783–793, Oct 2009.

[9] J. L. Chen, P. Zhang, M. Abe, H. Aikawa, L. Zhang, A. J. Frank, T. Zembryski, C. Hubbs, H. Park, J. Withka, C. Steppan, L. Rogers, S. Cabral, M. Pettersson, T. T. Wager, M. A. Fountain, G. Rumbaugh, J. L. Childs-Disney, and M. D. Disney. Design, Optimization, and Study of Small Molecules That Target Tau Pre-mRNA and Affect Splicing. *J Am Chem Soc*, 142(19):8706–8727, May 2020.

[10] M. M. Scotti and M. S. Swanson. RNA mis-splicing in disease. *Nat Rev Genet*, 17(1):19–32, Jan 2016.

[11] K. Yoshida, M. Sanada, Y. Shiraishi, D. Nowak, Y. Nagata, R. Yamamoto, Y. Sato, A. Sato-Otsubo, A. Kon, M. Nagasaki, G. Chalkidis, Y. Suzuki, M. Shiosaka, R. Kawahata, T. Yamaguchi, M. Otsu, N. Obara, M. Sakata-Yanagimoto, K. Ishiyama, H. Mori, F. Nolte, W. K. Hofmann, S. Miyawaki, S. Sugano, C. Haferlach, H. P. Koeffler, L. Y. Shih, T. Haferlach, S. Chiba, H. Nakauchi, S. Miyano, and S. Ogawa. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, 478(7367):64–69, Sep 2011.

[12] A. S. Adler, M. L. McCleland, S. Yee, M. Yaylaoglu, S. Hussain, E. Cosino, G. Quinones, Z. Modrusan, S. Seshagiri, E. Torres, V. S. Chopra, B. Haley, Z. Zhang, E. M. Blackwood, M. Singh, M. Junttila, J. P. Stephan, J. Liu, G. Pau, E. R. Fearon, Z. Jiang, and R. Firestein. An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth. *Genes Dev*, 28(10):1068–1084, May 2014.

[13] B. S. Jo and S. S. Choi. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform*, 13(4):112–118, Dec 2015.

[14] M. Chorev and L. Carmel. The function of introns. *Front Genet*, 3:55, 2012.

[15] I. B. Rogozin, L. Carmel, M. Csuros, and E. V. Koonin. Origin and evolution of spliceosomal introns. *Biol Direct*, 7:11, Apr 2012.

[16] M. Csuros, I. B. Rogozin, and E. V. Koonin. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol*, 7(9):e1002150, Sep 2011.

[17] H. Keren, G. Lev-Maor, and G. Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*, 11(5):345–355, May 2010.

[18] M. Lynch. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A*, 99(9):6118–6123, Apr 2002.

[19] J. Ule and B. J. Blencowe. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol Cell*, 76(2):329–345, Oct 2019.

[20] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol*, 14(3):153–165, Mar 2013.

[21] T. W. Nilsen and B. R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, Jan 2010.

[22] Y. Liu, M. Gonzàlez-Porta, S. Santos, A. Brazma, J. C. Marioni, R. Aebersold, A. R. Venkitaraman, and V. O. Wickramasinghe. Impact of Alternative Splicing on the Human Proteome. *Cell Rep*, 20(5):1229–1241, Aug 2017.

[23] P. Heyn, A. T. Kalinka, P. Tomancak, and K. M. Neugebauer. Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences. *Bioessays*, 37(2):148–154, Feb 2015.

[24] L. G. Guilgur, P. Prudêncio, D. Sobral, D. Liszekova, A. Rosa, and R. G. Martinho. Requirement for highly efficient pre-mRNA splicing during Drosophila early embryonic development. *Elife*, 3:e02181, Apr 2014.

[25] J. K. Biedler, W. Hu, H. Tae, and Z. Tu. Identification of early zygotic genes in the yellow fever mosquito Aedes aegypti and discovery of a motif involved in early zygotic genome activation. *PLoS One*, 7(3):e33933, 2012.

[26] P. Heyn, M. Kircher, A. Dahl, J. Kelso, P. Tomancak, A. T. Kalinka, and K. M. Neugebauer. The earliest transcribed zygotic genes are short, newly evolved, and different across species. *Cell Rep*, 6(2):285–292, Jan 2014.

[27] D. C. Jeffares, C. J. Penkett, and J. Bähler. Rapidly regulated genes are intron poor. *Trends Genet*, 24(8):375–378, Aug 2008.

[28] N. I. Bieberstein, F. Carrillo Oesterreich, K. Straube, and K. M. Neugebauer. First exon length controls active chromatin signatures and transcription. *Cell Rep*, 2(1):62–68, Jul 2012.

[29] A. B. Rose, T. Elfersi, G. Parra, and I. Korf. Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *Plant Cell*, 20(3):543–551, Mar 2008.

[30] A. M. Moabbi, N. Agarwal, B. El Kaderi, and A. Ansari. Role for gene looping in intron-mediated enhancement of transcription. *Proc Natl Acad Sci U S A*, 109(22):8505–8510, May 2012.

[31] O. Shaul. How introns enhance gene expression. *Int J Biochem Cell Biol*, 91(Pt B):145–155, Oct 2017.

[32] M. E. Wilkinson, C. Charenton, and K. Nagai. RNA Splicing by the Spliceosome. *Annu Rev Biochem*, 89:359–388, Jun 2020.

[33] A. G. Matera and Z. Wang. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol*, 15(2):108–121, Feb 2014.

[34] Y. Shi. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat Rev Mol Cell Biol*, 18(11):655–670, Nov 2017.

[35] A. E. Vinogradov. "Genome design" model: evidence from conserved intronic sequence in human-mouse comparison. *Genome Res*, 16(3):347–354, Mar 2006.

[36] S. G. Park, S. Hannenhalli, and S. S. Choi. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. *BMC Genomics*, 15:526, Jun 2014.

[37] B. Kastner, C. L. Will, H. Stark, and R. Lührmann. Structural Insights into Nuclear pre-mRNA Splicing in Higher Eukaryotes. *Cold Spring Harb Perspect Biol*, 11(11), Nov 2019.

[38] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front Endocrinol (Lausanne)*, 9:402, 2018.

[39] S. Massenet, E. Bertrand, and C. Verheggen. Assembly and trafficking of box C/D and H/ACA snoRNPs. *RNA Biol*, 14(6):680–692, Jun 2017.

[40] D. Hernandez-Verdun, P. Roussel, M. Thiry, V. Sirri, and D. L. Lafontaine. The nucleolus: structure/function relationship in RNA metabolism. *Wiley Interdiscip Rev RNA*, 1(3):415–431, 2010.

[41] O. Cordin and J. D. Beggs. RNA helicases in splicing. *RNA Biol*, 10(1):83–95, Jan 2013.

[42] S. Cho, A. Hoang, R. Sinha, X. Y. Zhong, X. D. Fu, A. R. Krainer, and G. Ghosh. Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A*, 108(20):8233–8238, May 2011.

[43] C. Boesler, N. Rigo, M. M. Anokhina, M. J. Tauchert, D. E. Agafonov, B. Kastner, H. Urlaub, R. Ficner, C. L. Will, and R. Lührmann. A spliceosome intermediate with loosely associated tri-snRNP accumulates in the absence of Prp28 ATPase activity. *Nat Commun*, 7:11997, Jul 2016.

[44] K. L. Fox-Walsh, Y. Dou, B. J. Lam, S. P. Hung, P. F. Baldi, and K. J. Hertel. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A*, 102(45):16176–16181, Nov 2005.

[45] X. Xiao, Z. Wang, M. Jang, and C. B. Burge. Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci U S A*, 104(47):18583–18588, Nov 2007.

[46] X. Li, S. Liu, L. Zhang, A. Issaian, R. C. Hill, S. Espinosa, S. Shi, Y. Cui, K. Kappel, R. Das, K. C. Hansen, Z. H. Zhou, and R. Zhao. A unified mechanism for intron and exon definition and back-splicing. *Nature*, 573(7774):375–380, Sep 2019.

[47] M. Schneider, C. L. Will, M. Anokhina, J. Tazi, H. Urlaub, and R. Lührmann. Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. *Mol Cell*, 38(2):223–235, Apr 2010.

[48] D. S. Horowitz. The mechanism of the second step of pre-mRNA splicing. *Wiley Interdiscip Rev RNA*, 3(3):331–350, 2012.

[49] H. Dvinge, E. Kim, O. Abdel-Wahab, and R. K. Bradley. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer*, 16(7):413–430, Jul 2016.

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

[51] J. D. Kohtz, S. F. Jamison, C. L. Will, P. Zuo, R. Lührmann, M. A. Garcia-Blanco, and J. L. Manley. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature*, 368(6467):119–124, Mar 1994.

[52] X. D. Fu and T. Maniatis. The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc Natl Acad Sci U S A*, 89(5):1725–1729, Mar 1992.

[53] A. Busch and K. J. Hertel. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA*, 3(1):1–12, 2012.

[54] J. M. Izquierdo, N. Majos, S. Bonnal, C. Martínez, R. Castelo, R. Guigó, D. Bilbao, and J. Valcárcel. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell*, 19(4):475–484, Aug 2005.

[55] S. Sharma, A. M. Falick, and D. L. Black. Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of U2AF and the prespliceosomal E complex. *Mol Cell*, 19(4):485–496, Aug 2005.

[56] E. J. Wagner and M. A. Garcia-Blanco. Polypyrimidine tract binding protein antagonizes exon definition. *Mol Cell Biol*, 21(10):3281–3288, May 2001.

[57] A. E. House and K. W. Lynch. An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat Struct Mol Biol*, 13(10):937–944, Oct 2006.

[58] J. E. Burke, A. D. Longhurst, D. Merkurjev, J. Sales-Lee, B. Rao, J. J. Moresco, J. R. Yates, J. J. Li, and H. D. Madhani. Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. *Cell*, 173(4):1014–1030, May 2018.

[59] J. K. Pickrell, A. A. Pai, Y. Gilad, and J. K. Pritchard. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet*, 6(12):e1001236, Dec 2010.

[60] E. Melamud and J. Moult. Stochastic noise in splicing machinery. *Nucleic Acids Res*, 37(14):4873–4886, Aug 2009.

[61] E. R. Gamazon and B. E. Stranger. Genomics of alternative splicing: evolution, development and pathophysiology. *Hum Genet*, 133(6):679–687, Jun 2014.

[62] M. Sammeth, S. Foissac, and R. Guigó. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*, 4(8):e1000147, Aug 2008.

[63] P. Yang, D. Wang, and L. Kang. Alternative splicing level related to intron size and organism complexity. *BMC Genomics*, 22(1):853, Nov 2021.

[64] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, Dec 2012.

[65] A. Kahraman, M. Buljan, and K. Vitting-Seerup. Editorial: Alternative Splicing in Health and Disease. *Front Mol Biosci*, 9:878668, 2022.

[66] F. E. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*, 18(7):437–451, Jul 2017.

[67] M. L. Tress, F. Abascal, and A. Valencia. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci*, 42(2):98–110, Feb 2017.

[68] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014.

[69] R. K. Bradley, J. Merkin, N. J. Lambert, and C. B. Burge. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol*, 10(1):e1001229, Jan 2012.

[70] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[71] P. V. Mazin, P. Khaitovich, M. Cardoso-Moreira, and H. Kaessmann. Alternative splicing during mammalian organ development. *Nat Genet*, 53(6):925–934, Jun 2021.

[72] J. W. Park, S. Fu, B. Huang, and R. H. Xu. Alternative splicing in mesenchymal stem cell differentiation. *Stem Cells*, 38(10):1229–1240, Oct 2020.

[73] R. K. Singh, Z. Xia, C. S. Bland, A. Kalsotra, M. A. Scavuzzo, T. Curk, J. Ule, W. Li, and T. A. Cooper. Rbfox2-coordinated alternative splicing of Mef2d and Rock2 controls myoblast fusion during myogenesis. *Mol Cell*, 55(4):592–603, Aug 2014.

[74] C. K. Vuong, D. L. Black, and S. Zheng. The neurogenetics of alternative splicing. *Nat Rev Neurosci*, 17(5):265–281, May 2016.

[75] M. M. van den Hoogenhof, Y. M. Pinto, and E. E. Creemers. RNA Splicing: Regulation and Dysregulation in the Heart. *Circ Res*, 118(3):454–468, Feb 2016.

[76] A. Bhate, D. J. Parker, T. W. Bebee, J. Ahn, W. Arif, E. H. Rashan, S. Chorghade, A. Chau, J. H. Lee, S. Anakk, R. P. Carstens, X. Xiao, and A. Kalsotra. ESRP2 controls an adult splicing programme in hepatocytes to support postnatal liver maturation. *Nat Commun*, 6:8768, Nov 2015.

[77] D. D. Licatalosi. Roles of RNA-binding Proteins and Post-transcriptional Regulation in Driving Male Germ Cell Development in the Mouse. *Adv Exp Med Biol*, 907:123–151, 2016.

[78] J. Merkin, C. Russell, P. Chen, and C. B. Burge. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, 338(6114):1593–1599, Dec 2012.

[79] A. R. Grosso, A. Q. Gomes, N. L. Barbosa-Morais, S. Caldeira, N. P. Thorne, G. Grech, M. von Lindern, and M. Carmo-Fonseca. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res*, 36(15):4823–4832, Sep 2008.

[80] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[81] S. Gerstberger, M. Hafner, M. Ascano, and T. Tuschl. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv Exp Med Biol*, 825:1–55, 2014.

[82] Q. Li, J. A. Lee, and D. L. Black. Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci*, 8(11):819–831, Nov 2007.

[83] B. Raj and B. J. Blencowe. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron*, 87(1):14–27, Jul 2015.

[84] H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigó. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res*, 22(9):1616–1625, Sep 2012.

[85] D. L. Bentley. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet*, 15(3):163–175, Mar 2014.

[86] I. Jonkers, H. Kwak, and J. T. Lis. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*, 3:e02407, Apr 2014.

[87] S. Schwartz, E. Meshorer, and G. Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–995, Sep 2009.

[88] M. Gutierrez-Arcelus, H. Ongen, T. Lappalainen, S. B. Montgomery, A. Buil, A. Yurovsky, J. Bryois, I. Padioleau, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, T. Giger, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, 11(1):e1004958, Jan 2015.

[89] N. H. Gehring and J. Y. Roignant. Anything but Ordinary - Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends Genet*, 37(4):355–372, Apr 2021.

[90] Z. Zhang, D. Xin, P. Wang, L. Zhou, L. Hu, X. Kong, and L. D. Hurst. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol*, 7:23, May 2009.

[91] S. N. Hsu and K. J. Hertel. Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA Biol*, 6(5):526–530, 2009.

[92] N. Stepankiw, M. Raghavan, E. A. Fogarty, A. Grimson, and J. A. Pleiss. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res*, 43(17):8488–8501, Sep 2015.

[93] Y. Wan and D. R. Larson. Splicing heterogeneity: separating signal from noise. *Genome Biol*, 19(1):86, Jul 2018.

[94] Y. Kapustin, E. Chan, R. Sarkar, F. Wong, I. Vorechovsky, R. M. Winston, T. Tatusova, and N. J. Dibb. Cryptic splice sites and split genes. *Nucleic Acids Res*, 39(14):5837–5844, Aug 2011.

[95] X. Roca, M. Akerman, H. Gaus, A. Berdeja, C. F. Bennett, and A. R. Krainer. Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. *Genes Dev*, 26(10):1098–1109, May 2012.

[96] D. E. Egecioglu and G. Chanfreau. Proofreading and spellchecking: a two-tier strategy for pre-mRNA splicing quality control. *RNA*, 17(3):383–389, Mar 2011.

[97] L. Wolf, O. K. Silander, and E. van Nimwegen. Expression noise facilitates the evolution of gene regulation. *Elife*, 4, Jun 2015.

[98] T. Eom, C. Zhang, H. Wang, K. Lay, J. Fak, J. L. Noebels, and R. B. Darnell. NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *Elife*, 2:e00178, Jan 2013.

[99] J. Humphrey, W. Emmett, P. Fratta, A. M. Isaacs, and V. Plagnol. Quantitative analysis of cryptic splicing associated with TDP-43 depletion. *BMC Med Genomics*, 10(1):38, May 2017.

[100] J. P. Ling, R. Chhabra, J. D. Merran, P. M. Schaughency, S. J. Wheelan, J. L. Corden, and P. C. Wong. PTBP1 and PTBP2 Repress Nonconserved Cryptic Exons. *Cell Rep*, 17(1):104–113, Sep 2016.

[101] X. R. Ma, M. Prudencio, Y. Koike, S. C. Vatsavayai, G. Kim, F. Harbinski, A. Briner, C. M. Rodriguez, C. Guo, T. Akiyama, H. B. Schmidt, B. B. Cummings, D. W. Wyatt, K. Kurylo, G. Miller, S. Mekhoubad, N. Sallee, G. Mekonnen, L. Ganser, J. D. Rubien, K. Jansen-West, C. N. Cook, S. Pickles, B. Oskarsson, N. R. Graff-Radford, B. F. Boeve, D. S. Knopman, R. C. Petersen, D. W. Dickson, J. Shorter, S. Myong, E. M. Green, W. W. Seeley, L. Petrucelli, and A. D. Gitler. TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. *Nature*, 603(7899):124–130, Mar 2022.

[102] M. Hiller and M. Platzer. Widespread and subtle: alternative splicing at short-distance tandem sites. *Trends Genet*, 24(5):246–255, May 2008.

[103] K. Szafranski and M. Kramer. It's a bit over, is that ok? The subtle surplus from tandem alternative splicing. *RNA Biol*, 12(2):115–122, 2015.

[104] Z. Kozmik, T. Czerny, and M. Busslinger. Alternatively spliced insertions in the paired domain restrict the DNA sequence specificity of Pax6 and Pax8. *EMBO J*, 16(22):6793–6803, Nov 1997.

[105] K. Tadokoro, M. Yamazaki-Inoue, M. Tachibana, M. Fujishiro, K. Nagao, M. Toyoda, M. Ozaki, M. Ono, N. Miki, T. Miyashita, and M. Yamada. Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. *J Hum Genet*, 50(8):382–394, 2005.

[106] M. Yan, L. C. Wang, S. G. Hymowitz, S. Schilbach, J. Lee, A. Goddard, A. M. de Vos, W. Q. Gao, and V. M. Dixit. Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science*, 290(5491):523–527, Oct 2000.

[107] J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet*, 19(R2):R131–136, Oct 2010.

[108] A. Auton, L. D. Brooks, R. M. Durbin, and et al Garrison. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015.

[109] M. Irimia, R. J. Weatheritt, J. D. Ellis, N. N. Parikshak, T. Gonatopoulos-Pournatzis, M. Babor, M. Quesnel-Vallières, J. Tapial, B. Raj, D. O'Hanlon, M. Barrios-Rodiles, M. J. Sternberg, S. P. Cordes, F. P. Roth, J. L. Wrana, D. H. Geschwind, and B. J. Blencowe. A highly conserved program of neuronal

microexons is misregulated in autistic brains. *Cell*, 159(7):1511–1523, Dec 2014.

[110] M. Lin, S. Whitmire, J. Chen, A. Farrel, X. Shi, and J. T. Guo. Effects of short indels on protein structure and function in human genomes. *Sci Rep*, 7(1):9313, Aug 2017.

[111] M. Hiller, K. Szafranski, K. Huse, R. Backofen, and M. Platzer. Selection against tandem splice sites affecting structured protein regions. *BMC Evol Biol*, 8:89, Mar 2008.

[112] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, and M. Platzer. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet*, 36(12):1255–1257, Dec 2004.

[113] R. Sinha, S. Nikolajewa, K. Szafranski, M. Hiller, N. Jahn, K. Huse, M. Platzer, and R. Backofen. Accurate prediction of NAGNAG alternative splicing. *Nucleic Acids Res*, 37(11):3569–3579, Jun 2009.

[114] K. Szafranski, C. Fritsch, F. Schumann, L. Siebel, R. Sinha, J. Hampe, M. Hiller, C. Englert, K. Huse, and M. Platzer. Physiological state co-regulates thousands of mammalian mRNA splicing events at tandem splice sites and alternative exons. *Nucleic Acids Res*, 42(14):8895–8904, Aug 2014.

[115] M. Wang, P. Zhang, Y. Shu, F. Yuan, Y. Zhang, Y. Zhou, M. Jiang, Y. Zhu, L. Hu, X. Kong, and Z. Zhang. Alternative splicing at GYNNGY 5' splice sites: more noise, less regulation. *Nucleic Acids Res*, 42(22):13969–13980, Dec 2014.

[116] K. W. Tsai, W. C. Chan, C. N. Hsu, and W. C. Lin. Sequence features involved in the mechanism of 3' splice junction wobbling. *BMC Mol Biol*, 11:34, May 2010.

[117] T. M. Chern, E. van Nimwegen, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and M. Zavolan. A simple physical model predicts small exon length variations. *PLoS Genet*, 2(4):e45, Apr 2006.

[118] Y. Dou, K. L. Fox-Walsh, P. F. Baldi, and K. J. Hertel. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*, 12(12):2047–2056, Dec 2006.

[119] M. L. Tress, F. Abascal, and A. Valencia. Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem Sci*, 42(6):408–410, Jun 2017.

[120] B. J. Blencowe. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem Sci*, 42(6):407–408, Jun 2017.

[121] T. Kurosaki, M. W. Popp, and L. E. Maquat. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol*, 20(7):406–420, Jul 2019.

[122] S. Lykke-Andersen and T. H. Jensen. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol*, 16(11):665–677, Nov 2015.

[123] D. Zünd, A. R. Gruber, M. Zavolan, and O. Mühlemann. Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3' UTRs. *Nat Struct Mol Biol*, 20(8):936–943, Aug 2013.

[124] E. D. Karousis, S. Nasif, and O. Mühlemann. Nonsense-mediated mRNA decay: novel mechanistic insights and biological impact. *Wiley Interdiscip Rev RNA*, 7(5):661–682, Sep 2016.

[125] M. W. Popp and L. E. Maquat. Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. *Cell*, 165(6):1319–1322, Jun 2016.

[126] H. Le Hir, E. Izaurralde, L. E. Maquat, and M. J. Moore. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J*, 19(24):6860–6869, Dec 2000.

[127] I. Aznarez, T. T. Nomakuchi, J. Tetenbaum-Novatt, M. A. Rahman, O. Fregoso, H. Rees, and A. R. Krainer. Mechanism of Nonsense-Mediated mRNA Decay Stimulation by Splicing Factor SRSF1. *Cell Rep*, 23(7):2186–2198, May 2018.

[128] M. H. Viegas, N. H. Gehring, S. Breit, M. W. Hentze, and A. E. Kulozik. The abundance of RNPS1, a protein component of the exon junction complex, can determine the variability in efficiency of the Nonsense Mediated Decay pathway. *Nucleic Acids Res*, 35(13):4542–4551, 2007.

[129] G. Nogueira, R. Fernandes, J. F. García-Moreno, and L. Romão. Nonsense-mediated RNA decay and its bipolar function in cancer. *Mol Cancer*, 20(1):72, Apr 2021.

[130] I. Kashima, A. Yamashita, N. Izumi, N. Kataoka, R. Morishita, S. Hoshino, M. Ohno, G. Dreyfuss, and S. Ohno. Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes Dev*, 20(3):355–367, Feb 2006.

[131] G. Singh, I. Rebbapragada, and J. Lykke-Andersen. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol*, 6(4):e111, Apr 2008.

[132] K. G. Toma, I. Rebbapragada, S. Durand, and J. Lykke-Andersen. Identification of elements in human long 3' UTRs that inhibit nonsense-mediated decay. *RNA*, 21(5):887–897, May 2015.

[133] J. A. Hurt, A. D. Robertson, and C. B. Burge. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res*, 23(10):1636–1650, Oct 2013.

[134] H. Tani, N. Imamachi, K. A. Salam, R. Mizutani, K. Ijiri, T. Irie, T. Yada, Y. Suzuki, and N. Akimitsu. Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability. *RNA Biol*, 9(11):1370–1379, Nov 2012.

[135] Y. Sun, Y. Bao, W. Han, F. Song, X. Shen, J. Zhao, J. Zuo, D. Saffen, W. Chen, Z. Wang, X. You, and Y. Wang. Autoregulation of RBM10 and cross-regulation of RBM10/RBM5 via alternative splicing-coupled nonsense-mediated decay. *Nucleic Acids Res*, 45(14):8524–8540, Aug 2017.

[136] F. M. Hamid and E. V. Makeyev. Regulation of mRNA abundance by polypyrimidine tract-binding protein-controlled alternate 5' splice site choice. *PLoS Genet*, 10(11):e1004771, Nov 2014.

[137] S. M. Medghalchi, P. A. Frischmeyer, J. T. Mendell, A. G. Kelly, A. M. Lawler, and H. C. Dietz. Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum Mol Genet*, 10(2):99–105, Jan 2001.

[138] J. Weischenfeldt, I. Damgaard, D. Bryder, K. Theilgaard-Mönch, L. A. Thoren, F. C. Nielsen, S. E. Jacobsen, C. Nerlov, and B. T. Porse. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev*, 22(10):1381–1396, May 2008.

[139] D. R. McIlwain, Q. Pan, P. T. Reilly, A. J. Elia, S. McCracken, A. C. Wakeham, A. Itie-Youten, B. J. Blencowe, and T. W. Mak. Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proc Natl Acad Sci U S A*, 107(27):12186–12191, Jul 2010.

[140] T. Li, Y. Shi, P. Wang, L. M. Guachalla, B. Sun, T. Joerss, Y. S. Chen, M. Groth, A. Krueger, M. Platzer, Y. G. Yang, K. L. Rudolph, and Z. Q. Wang. Smg6/Est1 licenses embryonic stem cell differentiation via nonsense-mediated mRNA decay. *EMBO J*, 34(12):1630–1647, Jun 2015.

[141] C. H. Lou, A. Shao, E. Y. Shum, J. L. Espinoza, L. Huang, R. Karam, and M. F. Wilkinson. Posttranscriptional control of the stem cell and neurogenic programs by the nonsense-mediated RNA decay pathway. *Cell Rep*, 6(4):748–764, Feb 2014.

[142] B. P. Lewis, R. E. Green, and S. E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A*, 100(1):189–192, Jan 2003.

[143] Anna Desai, Zhiqiang Hu, Courtney E. French, James P. B. Lloyd, and Steven E. Brenner. Networks of splice factor regulation by unproductive splicing coupled with nonsense mediated mRNA decay. *BiorXiv*, May 2020.

[144] L. F. Lareau, M. Inada, R. E. Green, J. C. Wengrod, and S. E. Brenner. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929, Apr 2007.

[145] S. A. Filichkin and T. C. Mockler. Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. *Biol Direct*, 7:20, Jul 2012.

[146] J. Z. Ni, L. Grate, J. P. Donohue, C. Preston, N. Nobida, G. O'Brien, L. Shiue, T. A. Clark, J. E. Blume, and M. Ares. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev*, 21(6):708–718, Mar 2007.

[147] J. J. Wong, W. Ritchie, O. A. Ebner, M. Selbach, J. W. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T. L. Khoo, C. G. Bailey, J. Holst, and J. E. Rasko. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, 154(3):583–595, Aug 2013.

[148] M. Dewaele, T. Tabaglio, K. Willekens, M. Bezzi, S. X. Teo, D. H. Low, C. M. Koh, F. Rambow, M. Fiers, A. Rogiers, E. Radaelli, M. Al-Haddawi, S. Y. Tan, E. Hermans, F. Amant, H. Yan, M. Lakshmanan, R. C. Koumar, S. T. Lim, F. A. Derheimer, R. M. Campbell, Z. Bonday, V. Tergaonkar, M. Shackleton, C. Blattner, J. C. Marine, and E. Guccione. Antisense oligonucleotide-mediated MDM4 exon 6 skipping impairs tumor growth. *J Clin Invest*, 126(1):68–84, Jan 2016.

[149] J. Barbier, M. Dutertre, D. Bittencourt, G. Sanchez, L. Gratadou, P. de la Grange, and D. Auboeuf. Regulation of H-ras splice variant expression by cross talk between the p53 and nonsense-mediated mRNA decay pathways. *Mol Cell Biol*, 27(20):7315–7333, Oct 2007.

[150] J. F. García-Moreno and L. Romão. Perspective in Alternative Splicing Coupled to Nonsense-Mediated mRNA Decay. *Int J Mol Sci*, 21(24), Dec 2020.

[151] N. K. Leclair, M. Brugiolo, L. Urbanski, S. C. Lawson, K. Thakar, M. Yurieva, J. George, J. T. Hinson, A. Cheng, B. R. Graveley, and O. Anczuków. Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol Cell*, 80(4):648–665, Nov 2020.

[152] R. Spellman, M. Llorian, and C. W. Smith. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol Cell*, 27(3):420–434, Aug 2007.

[153] P. L. Boutz, P. Stoilov, Q. Li, C. H. Lin, G. Chawla, K. Ostrow, L. Shiue, M. Ares, and D. L. Black. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev*, 21(13):1636–1652, Jul 2007.

[154] Y. Zhou, S. Liu, G. Liu, A. Oztürk, and G. G. Hicks. ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation. *PLoS Genet*, 9(10):e1003895, Oct 2013.

[155] K. Kemmerer, S. Fischer, and J. E. Weigand. Auto- and cross-regulation of the hnRNPs D and DL. *RNA*, 24(3):324–331, Mar 2018.

[156] A. L. Saltzman, Q. Pan, and B. J. Blencowe. Regulation of alternative splicing by the core spliceosomal machinery. *Genes Dev*, 25(4):373–384, Feb 2011.

[157] M. Cuccurese, G. Russo, A. Russo, and C. Pietropaolo. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res*, 33(18):5965–5977, 2005.

[158] S. Takei, M. Togo-Ohno, Y. Suzuki, and H. Kuroyanagi. Evolutionarily conserved autoregulation of alternative pre-mRNA splicing by ribosomal protein L10a. *Nucleic Acids Res*, 44(12):5585–5596, Jul 2016.

[159] L. F. Lareau and S. E. Brenner. Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol Biol Evol*, 32(4):1072–1079, Apr 2015.

[160] M. Kalyna, C. G. Simpson, N. H. Syed, D. Lewandowska, Y. Marquez, B. Kusenda, J. Marshall, J. Fuller, L. Cardle, J. McNicol, H. Q. Dinh, A. Barta, and J. W. Brown. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res*, 40(6):2454–2469, Mar 2012.

[161] C. Kovalak, S. Donovan, A. A. Bicknell, M. Metkar, and M. J. Moore. Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved alternative splicing nonsense-mediated decay pathways. *Genome Biol*, 22(1):132, May 2021.

[162] M. C. Wollerton, C. Gooding, E. J. Wagner, M. A. Garcia-Blanco, and C. W. Smith. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell*, 13(1):91–100, Jan 2004.

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

[164] M. L. Anko, M. Muller-McNicoll, H. Brandl, T. Curk, C. Gorup, I. Henry, J. Ule, and K. M. Neugebauer. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol*, 13(3):R17, 2012.

[165] M. Jangi, P. L. Boutz, P. Paul, and P. A. Sharp. Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev*, 28(6):637–651, Mar 2014.

[166] O. Rossbach, L. H. Hung, S. Schreiner, I. Grishina, M. Heiner, J. Hui, and A. Bindereif. Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Mol Cell Biol*, 29(6):1442–1451, Mar 2009.

[167] S. Sun, Z. Zhang, R. Sinha, R. Karni, and A. R. Krainer. SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol*, 17(3):306–312, Mar 2010.

[168] M. Jiménez, R. Urtasun, M. Elizalde, M. Azkona, M. U. Latasa, I. Uriarte, M. Arechederra, D. Alignani, M. Bárcena-Varela, G. Álvarez Sola, L. Colyn, E. Santamaría, B. Sangro, C. Rodriguez-Ortigosa, M. G. Fernández-Barrena, M. A. Ávila, and C. Berasain. Splicing events in the control of genome integrity: role of SLU7 and truncated SRSF3 proteins. *Nucleic Acids Res*, 47(7):3450–3466, Apr 2019.

[169] Y. Nakano, M. C. Kelly, A. U. Rehman, E. T. Boger, R. J. Morell, M. W. Kelley, T. B. Friedman, and B. Bánfi. Defects in the Alternative Splicing-Dependent Regulation of REST Cause Deafness. *Cell*, 174(3):536–548, Jul 2018.

[170] S. Zheng, E. E. Gray, G. Chawla, B. T. Porse, T. J. O'Dell, and D. L. Black. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nat Neurosci*, 15(3):381–388, Jan 2012.

[171] S. Zheng. Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression. *Int J Dev Neurosci*, 55:102–108, Dec 2016.

[172] E. V. Makeyev, J. Zhang, M. A. Carrasco, and T. Maniatis. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell*, 27(3):435–448, Aug 2007.

[173] M. Haeussler, A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper, G. P. Barber, D. Haussler, R. M. Kuhn, and W. J. Kent. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*, 47(D1):D853–D858, Jan 2019.

[174] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, and P. Flicek. GENCODE 2021. *Nucleic Acids Res*, 49(D1):D916–D923, Jan 2021.

[175] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733–745, Jan 2016.

[176] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigó. Human genomics. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, May 2015.

[177] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[178] D. D. Pervouchine, D. G. Knowles, and R. Guigó. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics*, 29(2):273–274, Jan 2013.

[179] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K. Farh. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176(3):535–548, Jan 2019.

[180] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11(2-3):377–394, 2004.

[181] S. Lykke-Andersen, Y. Chen, B. R. Ardal, B. Lilje, J. Waage, A. Sandelin, and T. H. Jensen. Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes Dev*, 28(22):2498–2517, Nov 2014.

[182] L. Wang, S. Wang, and W. Li. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185, Aug 2012.

[183] B. Saudemont, A. Popa, J. L. Parmley, V. Rocher, C. Blugeon, A. Necsulea, E. Meyer, and L. Duret. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol*, 18(1):208, Oct 2017.

[184] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27(8):48192, 2008.

[185] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, Aug 2003.

[186] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*, 46(D1):D794–D801, Jan 2018.

[187] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, 13(6):508–514, Jun 2016.

[188] S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, 111(51):E5593–5601, Dec 2014.

[189] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. Kyung, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 45(6):580–585, Jun 2013.

[190] A. Zhu, J. G. Ibrahim, and M. I. Love. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, 35(12):2084–2092, Jun 2019.

[191] I. Dunham, A. Kundaje, S. F. Aldred, and Collins et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[192] C. A. Sloan, E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, and J. M. Cherry. ENCODE data at the ENCODE portal. *Nucleic Acids Res*, 44(D1):D726–732, Jan 2016.

[193] S. Gueroussov, T. Gonatopoulos-Pournatzis, M. Irimia, B. Raj, Z. Y. Lin, A. C. Gingras, and B. J. Blencowe. An alternative splicing event amplifies evolutionary differences between vertebrates. *Science*, 349(6250):868–873, Aug 2015.

[194] A. Breschi, M. Muñoz-Aguirre, V. Wucher, C. A. Davis, D. Garrido-Martín, S. Djebali, J. Gillis, D. D. Pervouchine, A. Vlasova, A. Dobin, C. Zaleski, J. Drenkow, C. Danyko, A. Scavelli, F. Reverter, M. P. Snyder, T. R. Gingeras, and R. Guigó. A limited set of transcriptional programs define major cell types. *Genome Res*, 30(7):1047–1059, Jul 2020.

[195] B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, Aug 2011.

[196] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12):550, 2014.

[197] A. M. Michel, G. Fox, A. M Kiran, C. De Bo, P. B. O'Connor, S. M. Heaphy, J. P. Mullan, C. A. Donohue, D. G. Higgins, and P. V. Baranov. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res*, 42(Database issue):D859–864, Jan 2014.

[198] A. Gress, V. Ramensky, J. Büch, A. Keller, and O. V. Kalinina. StructMAn: annotation of single-nucleotide polymorphisms in the structural context. *Nucleic Acids Res*, 44(W1):W463–468, Jul 2016.

[199] No authors listed. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47(D1):D506–D515, Jan 2019.

[200] M. Kumar, S. Michael, J. Alvarado-Valverde, B. Mészáros, H. Sámano-Sánchez, A. Zeke, L. Dobson, T. Lazar, M. Örd, A. Nagpal, N. Farahi, M. Käser, R. Kraleti, N. E. Davey, R. Pancsa, L. B. Chemes, and T. J. Gibson. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res*, 50(D1):D497–D508, Jan 2022.

[201] J. Yang and Y. Zhang. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*, 43(W1):W174–181, Jul 2015.

[202] J. Delgado, L. G. Radusky, D. Cianferoni, and L. Serrano. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, Oct 2019.

[203] C. W. Wood, A. A. Ibarra, G. J. Bartlett, A. J. Wilson, D. N. Woolfson, and R. B. Sessions. BAlaS: fast, interactive and accessible computational alanine-scanning using BudeAlaScan. *Bioinformatics*, 36(9):2917–2919, May 2020.

[204] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22(9):1760–1774, Sep 2012.

[205] S. Stamm, J. Zhu, K. Nakai, P. Stoilov, O. Stoss, and M. Q. Zhang. An alternative-exon database and its statistical analysis. *DNA Cell Biol*, 19(12):739–756, Dec 2000.

[206] S. Denisov, G. Bazykin, A. Favorov, A. Mironov, and M. Gelfand. Correlated Evolution of Nucleotide Positions within Splice Sites in Mammals. *PLoS One*, 10(12):e0144388, 2015.

[207] James S. Farris. Methods for computing wagner trees. *Systematic Biology*, 19(1):83–92, March 1970.

[208] S. V. Denisov, G. A. Bazykin, R. Sutormin, A. V. Favorov, A. A. Mironov, M. S. Gelfand, and A. S. Kondrashov. Weak negative and positive selection and the drift load at splice sites. *Genome Biol Evol*, 6(6):1437–1447, May 2014.

[209] P. Danecek, A. Auton, G. Abecasis, and Albers et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug 2011.

[210] E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J. Y. Chen, N. A. L. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L. P. B. Bouvrette, J. Bergalet, M. O. Duff, K. E. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, M. P. McGurk, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B. A. Yee, B. Zhou, A. L. Louie, S. Aigner, X. D. Fu, E. Lécuyer, C. B. Burge, B. R. Graveley, and G. W. Yeo. A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719, Jul 2020.

[211] A. Graubert, F. Aguet, A. Ravi, K. G. Ardlie, and G. Getz. RNA-SeQC 2: Efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics*, Mar 2021.

[212] W. Zhao, S. Zhang, Y. Zhu, X. Xi, P. Bao, Z. Ma, T. H. Kapral, S. Chen, B. Zagrovic, Y. T. Yang, and Z. J. Lu. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res*, 50(D1):D287–D294, Jan 2022.

[213] L. Lautenbacher, P. Samaras, J. Muller, A. Grafberger, M. Shraideh, J. Rank, S. T. Fuchs, T. K. Schmidt, M. The, C. Dallago, H. Wittges, B. Rost, H. Krcmar, B. Kuster, and M. Wilhelm. ProteomicsDB: toward a FAIR open-source resource for life-science research. *Nucleic Acids Res*, 50(D1):D1541–D1552, Jan 2022.

[214] R. Sinha, T. Lenser, N. Jahn, U. Gausmann, S. Friedel, K. Szafranski, K. Huse, P. Rosenstiel, J. Hampe, S. Schuster, M. Hiller, R. Backofen, and M. Platzer. TassDB2 - A comprehensive database of subtle alternative splicing events. *BMC Bioinformatics*, 11:216, Apr 2010.

[215] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, Dec 2008.

[216] A. Busch and K. J. Hertel. Extensive regulation of NAGNAG alternative splicing: new tricks for the spliceosome? *Genome Biol*, 13(2):143, Feb 2012.

[217] M. S. Gelfand. Statistical analysis of mammalian pre-mRNA splicing sites. *Nucleic Acids Res*, 17(15):6369–6382, Aug 1989.

[218] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, Feb 2009.

[219] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*, 11(2):R14, 2010.

[220] B. T. Sherman, M. Hao, J. Qiu, X. Jiao, M. W. Baseler, H. C. Lane, T. Imamichi, and W. Chang. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*, Mar 2022.

[221] D. Gong, X. Chi, K. Ren, G. Huang, G. Zhou, N. Yan, J. Lei, and Q. Zhou. Structure of the human plasma membrane Ca2+-ATPase 1 in complex with its obligatory subunit neuroplastin. *Nat Commun*, 9(1):3623, Sep 2018.

[222] P. W. Beesley, R. Herrera-Molina, K. H. Smalla, and C. Seidenbecher. The Neuroplastin adhesion molecules: key regulators of neuronal plasticity and synaptic function. *J Neurochem*, 131(3):268–283, Nov 2014.

[223] Q. Xu, B. Modrek, and C. Lee. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*, 30(17):3754–3766, Sep 2002.

[224] C. S. Lim, S. J. T Wardell, T. Kleffmann, and C. M. Brown. The exon-intron gene structure upstream of the initiation codon predicts translation efficiency. *Nucleic Acids Res*, 46(9):4575–4591, 05 2018.

[225] C. Cenik, A. Derti, J. C. Mellor, G. F. Berriz, and F. P. Roth. Genome-wide functional analysis of human 5' untranslated region introns. *Genome Biol*, 11(3):R29, 2010.

[226] A. Craxton, D. Munnur, R. Jukes-Jones, G. Skalka, C. Langlais, K. Cain, and M. Malewicz. PAXX and its paralogs synergistically direct DNA polymerase Î» activity in DNA repair. *Nat Commun*, 9(1):3877, Sep 2018.

[227] M. P. Hall, R. J. Nagel, W. S. Fagg, L. Shiue, M. S. Cline, R. J. Perriman, J. P. Donohue, and M. Ares. Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation. *RNA*, 19(5):627–638, May 2013.

[228] J. M. Ragle, S. Katzman, T. F. Akers, S. Barberan-Soler, and A. M. Zahler. Coordinated tissue-specific regulation of adjacent alternative 3' splice sites in C. elegans. *Genome Res*, 25(7):982–994, Jul 2015.

[229] P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, and A. K. Dunker. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A*, 103(22):8390–8395, May 2006.

[230] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. Attributes of short linear motifs. *Mol Biosyst*, 8(1):268–281, Jan 2012.

[231] K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, and N. E. Davey. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev*, 114(13):6733–6778, Jul 2014.

[232] B. Uyar, R. J. Weatheritt, H. Dinkel, N. E. Davey, and T. J. Gibson. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Mol Biosyst*, 10(10):2626–2642, Oct 2014.

[233] K. Y. Huang, T. Y. Lee, H. J. Kao, C. T. Ma, C. C. Lee, T. H. Lin, W. C. Chang, and H. D. Huang. dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Res*, 47(D1):D298–D308, Jan 2019.

[234] Y. Tian, J. C. Chang, E. Y. Fan, M. Flajolet, and P. Greengard. Adaptor complex AP2/PICALM, through interaction with LC3, targets Alzheimer's APP-CTF for terminal degradation via autophagy. *Proc Natl Acad Sci U S A*, 110(42):17071–17076, Oct 2013.

[235] K. Moreau, A. Fleming, S. Imarisio, A. Lopez Ramirez, J. L. Mercer, M. Jimenez-Sanchez, C. F. Bento, C. Puri, E. Zavodszky, F. Siddiqi, C. P. Lavau, M. Betton, C. J. O'Kane, D. S. Wechsler, and D. C. Rubinsztein. PICALM modulates autophagy activity and tau accumulation. *Nat Commun*, 5:4998, Sep 2014.

[236] T. Johansen and T. Lamark. Selective Autophagy: ATG8 Family Proteins, LIR Motifs and Cargo Receptors. *J Mol Biol*, 432(1):80–103, Jan 2020.

[237] X. Wang, P. D. Zamore, and T. M. Hall. Crystal structure of a Pumilio homology domain. *Mol Cell*, 7(4):855–865, Apr 2001.

[238] V. Bobo-Jiménez, M. Delgado-Esteban, J. Angibaud, I. Sánchez-Morán, A. de la Fuente, J. Yajeya, U. V. Nägerl, J. Castillo, J. P. Bolaños, and A. Almeida. APC/CCdh1-Rock2 pathway controls dendritic integrity and memory. *Proc Natl Acad Sci U S A*, 114(17):4513–4518, Apr 2017.

[239] M. Delgado-Esteban, I. García-Higuera, C. Maestre, S. Moreno, and A. Almeida. APC/C-Cdh1 coordinates neurogenesis and cortical size during development. *Nat Commun*, 4:2879, 2013.

[240] B. J. Raney, T. R. Dreszer, G. P. Barber, H. Clawson, P. A. Fujita, T. Wang, N. Nguyen, B. Paten, A. S. Zweig, D. Karolchik, and W. J. Kent. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, 30(7):1003–1005, Apr 2014.

[241] Biometrical Journal. Confidence limits for the ratio of two binomial proportions based on likelihood scores: Non-iterative method. *Jun-Mo Nam*, 37(3):375–379, 1995.

[242] N. Chen, I. Juric, E. J. Cosgrove, R. Bowman, J. W. Fitzpatrick, S. J. Schoech, A. G. Clark, and G. Coop. Allele frequency dynamics in a pedigreed natural population. *Proc Natl Acad Sci U S A*, 116(6):2158–2164, Feb 2019.

[243] M. Irimia, S. W. Roy, D. E. Neafsey, J. F. Abril, J. Garcia-Fernandez, and E. V. Koonin. Complex selection on 5' splice sites in intron-rich organisms. *Genome Res*, 19(11):2021–2027, Nov 2009.

[244] P. Razeto-Barry, J. Díaz, and R. A. Vásquez. The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity. *Genetics*, 191(2):523–534, Jun 2012.

[245] D. Pervouchine, Y. Popov, A. Berry, B. Borsari, A. Frankish, and R. Guigó. Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *Nucleic Acids Res*, 47(10):5293–5306, Jun 2019.

[246] P. L. Boutz, A. Bhutkar, and P. A. Sharp. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev*, 29(1):63–80, Jan 2015.

[247] E. Y. Shum, S. H. Jones, A. Shao, J. Dumdie, M. D. Krause, W. K. Chan, C. H. Lou, J. L. Espinoza, H. W. Song, M. H. Phan, M. Ramaiah, L. Huang, J. R. McCarrey, K. J. Peterson, D. G. De Rooij, H. Cook-Andersen, and M. F. Wilkinson. The Antagonistic Gene Paralogs Upf3a and Upf3b Govern Nonsense-Mediated RNA Decay. *Cell*, 165(2):382–395, Apr 2016.

[248] C. C. MacDonald and P. N. Grozdanov. Nonsense in the testis: multiple roles for nonsense-mediated decay revealed in male reproduction. *Biol Reprod*, 96(5):939–947, May 2017.

[249] A. B. Zetoune, S. Fontanière, D. Magnin, O. Anczuków, M. Buisson, C. X. Zhang, and S. Mazoyer. Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC Genet*, 9:83, Dec 2008.

[250] A. Hegele, A. Kamburov, A. Grossmann, C. Sourlis, S. Wowro, M. Weimann, C. L. Will, V. Pena, R. Lührmann, and U. Stelzl. Dynamic protein-protein interaction wiring of the human spliceosome. *Mol Cell*, 45(4):567–580, Feb 2012.

[251] J. J. Johnston, J. K. Teer, P. F. Cherukuri, N. F. Hansen, S. K. Loftus, K. Chong, J. C. Mullikin, and L. G. Biesecker. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet*, 86(5):743–748, May 2010.

[252] J. H. Jung, H. Lee, B. Cao, P. Liao, S. X. Zeng, and H. Lu. RNA-binding motif protein 10 induces apoptosis and suppresses proliferation by activating p53. *Oncogene*, 39(5):1031–1040, Jan 2020.

[253] X. Zhao, Q. Chen, Y. Cai, D. Chen, M. Bei, H. Dong, and J. Xu. TRA2A Binds With LncRNA MALAT1 To Promote Esophageal Cancer Progression By Regulating EZH2/Beta-catenin Pathway. *J Cancer*, 12(16):4883–4890, 2021.

[254] Y. Tan, X. Hu, Y. Deng, P. Yuan, Y. Xie, and J. Wang. TRA2A promotes proliferation, migration, invasion and epithelial mesenchymal transition of glioma cells. *Brain Res Bull*, 143:138–144, Oct 2018.

[255] W. Xu, H. Huang, L. Yu, and L. Cao. Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC). *Med Oncol*, 32(4):96, Apr 2015.

[256] J. Zhao, Y. Sun, Y. Huang, F. Song, Z. Huang, Y. Bao, J. Zuo, D. Saffen, Z. Shao, W. Liu, and Y. Wang. Functional analysis reveals that RBM10 mutations contribute to lung adenocarcinoma pathogenesis by deregulating splicing. *Sci Rep*, 7:40488, Jan 2017.

[257] G. Kerjan, H. Koizumi, E. B. Han, C. M. Dubé, S. N. Djakovic, G. N. Patrick, T. Z. Baram, S. F. Heinemann, and J. G. Gleeson. Mice lacking doublecortin

and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures. *Proc Natl Acad Sci U S A*, 106(16):6766–6771, Apr 2009.

[258] E. Shin, Y. Kashiwagi, T. Kuriu, H. Iwasaki, T. Tanaka, H. Koizumi, J. G. Gleeson, and S. Okabe. Doublecortin-like kinase enhances dendritic remodelling and negatively regulates synapse maturation. *Nat Commun*, 4:1440, 2013.

[259] F. M. Hamid and E. V. Makeyev. A mechanism underlying position-specific regulation of alternative splicing. *Nucleic Acids Res*, 45(21):12455–12468, Dec 2017.

[260] K. L. McDonald, M. G. O'Sullivan, J. F. Parkinson, J. M. Shaw, C. A. Payne, J. M. Brewer, L. Young, D. J. Reader, H. T. Wheeler, R. J. Cook, M. T. Biggs, N. S. Little, C. Teo, G. Stone, and B. G. Robinson. IQGAP1 and IGFBP2: valuable biomarkers for determining prognosis in glioma patients. *J Neuropathol Exp Neurol*, 66(5):405–417, May 2007.

[261] C. E. Bowman, S. Rodriguez, E. S. Selen Alpergin, M. G. Acoba, L. Zhao, T. Hartung, S. M. Claypool, P. A. Watkins, and M. J. Wolfgang. The Mammalian Malonyl-CoA Synthetase ACSF3 Is Required for Mitochondrial Protein Malonylation and Metabolic Efficiency. *Cell Chem Biol*, 24(6):673–684, Jun 2017.

[262] J. M. Ellis, C. E. Bowman, and M. J. Wolfgang. Metabolic and tissue-specific regulation of acyl-CoA metabolism. *PLoS One*, 10(3):e0116587, 2015.

[263] S. R. Jaffrey and M. F. Wilkinson. Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nat Rev Neurosci*, 19(12):715–728, Dec 2018.

[264] E. Park, Z. Pan, Z. Zhang, L. Lin, and Y. Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet*, 102(1):11–26, Jan 2018.

[265] K. Chua and R. Reed. An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Mol Cell Biol*, 21(5):1509–1514, Mar 2001.

[266] M. Mikl, A. Hamburg, Y. Pilpel, and E. Segal. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat Commun*, 10(1):4572, Oct 2019.

[267] X. Zhang, M. H. Chen, X. Wu, A. Kodani, J. Fan, R. Doan, M. Ozawa, J. Ma, N. Yoshida, J. F. Reiter, D. L. Black, P. V. Kharchenko, P. A. Sharp, and C. A. Walsh. Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. *Cell*, 166(5):1147–1162, Aug 2016.

[268] P. Wu, D. Zhou, W. Lin, Y. Li, H. Wei, X. Qian, Y. Jiang, and F. He. Cell-type-resolved alternative splicing patterns in mouse liver. *DNA Res*, Jan 2018.

[269] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. M. Sunkin, M. Hawrylycz, C. Koch, and H. Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci*, 19(2):335–346, Feb 2016.

[270] H. L. Röst, L. Malmström, and R. Aebersold. Reproducible quantitative proteotype data matrices for systems biology. *Mol Biol Cell*, 26(22):3926–3931, Nov 2015.

[271] X. Wang, S. G. Codreanu, B. Wen, K. Li, M. C. Chambers, D. C. Liebler, and B. Zhang. Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. *Mol Cell Proteomics*, 17(3):422–430, Mar 2018.

[272] R. J. Weatheritt, T. Sterne-Weiler, and B. J. Blencowe. The ribosome-engaged landscape of alternative splicing. *Nat Struct Mol Biol*, 23(12):1117–1123, Dec 2016.

[273] J. D. Ellis, M. Barrios-Rodiles, R. Colak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim, J. L. Wrana, and B. J. Blencowe. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell*, 46(6):884–892, Jun 2012.

[274] M. Buljan, G. Chalancon, A. K. Dunker, A. Bateman, S. Balaji, M. Fuxreiter, and M. M. Babu. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr Opin Struct Biol*, 23(3):443–450, Jun 2013.

[275] Q. M. Mitrovich and P. Anderson. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. *Genes Dev*, 14(17):2173–2184, Sep 2000.

[276] K. D. Hansen, L. F. Lareau, M. Blanchette, R. E. Green, Q. Meng, J. Rehwinkel, F. L. Gallusser, E. Izaurralde, D. C. Rio, S. Dudoit, and S. E. Brenner. Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. *PLoS Genet*, 5(6):e1000525, Jun 2009.

[277] M. Wight and A. Werner. The functions of natural antisense transcripts. *Essays Biochem*, 54:91–101, 2013.

[278] Z. Guo, M. Maki, R. Ding, Y. Yang, B. Zhang, and L. Xiong. Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci Rep*, 4:5150, Jun 2014.

[279] T. Watanabe, Y. Totoki, A. Toyoda, M. Kaneda, S. Kuramochi-Miyagawa, Y. Obata, H. Chiba, Y. Kohara, T. Kono, T. Nakano, M. A. Surani, Y. Sakaki, and H. Sasaki. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453(7194):539–543, May 2008.

[280] O. H. Tam, A. A. Aravin, P. Stein, A. Girard, E. P. Murchison, S. Cheloufi, E. Hodges, M. Anger, R. Sachidanandam, R. M. Schultz, and G. J. Hannon. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–538, May 2008.

[281] J. Yu, M. Chen, H. Huang, J. Zhu, H. Song, J. Zhu, J. Park, and S. J. Ji. Dynamic m6A modification regulates local translation of mRNA in axons. *Nucleic Acids Res*, 46(3):1412–1423, Feb 2018.

[282] S. Ke, A. Pandya-Jones, Y. Saito, J. J. Fak, C. B. Vågbø, S. Geula, J. H. Hanna, D. L. Black, J. E. Darnell, and R. B. Darnell. A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes Dev*, 31(10):990–1006, May 2017.

[283] K. Licht, M. Hartl, F. Amman, D. Anrather, M. P. Janisiw, and M. F. Jantsch. Inosine induces context-dependent recoding and translational stalling. *Nucleic Acids Res*, 47(1):3–14, Jan 2019.

[284] J. M. Mudge, A. Frankish, and J. Harrow. Functional transcriptomics in the post-ENCODE era. *Genome Res*, 23(12):1961–1973, Dec 2013.

[285] E. Kim, J. O. Ilagan, Y. Liang, G. M. Daubner, S. C. Lee, A. Ramakrishnan, Y. Li, Y. R. Chung, J. B. Micol, M. E. Murphy, H. Cho, M. K. Kim, A. S. Zebari, S. Aumann, C. Y. Park, S. Buonamici, P. G. Smith, H. J. Deeg, C. Lobry, I. Aifantis, Y. Modis, F. H. Allain, S. Halene, R. K. Bradley, and O. Abdel-Wahab. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell*, 27(5):617–630, May 2015.

[286] L. B. Gardner. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol*, 28(11):3729–3741, Jun 2008.

Appendix A

Supplementary materials

A.1 Supplementary information

The TASS catalogue is available through a track hub for the UCSC Genome Browser (https://raw.githubusercontent.com/magmir71/trackhubs/master/TASShub.txt). To visualize it, copy and paste the link into the form at http://genome.ucsc.edu/cgi-bin/hgHubConnect#unlistedHubs.

A.2 Supplementary figures

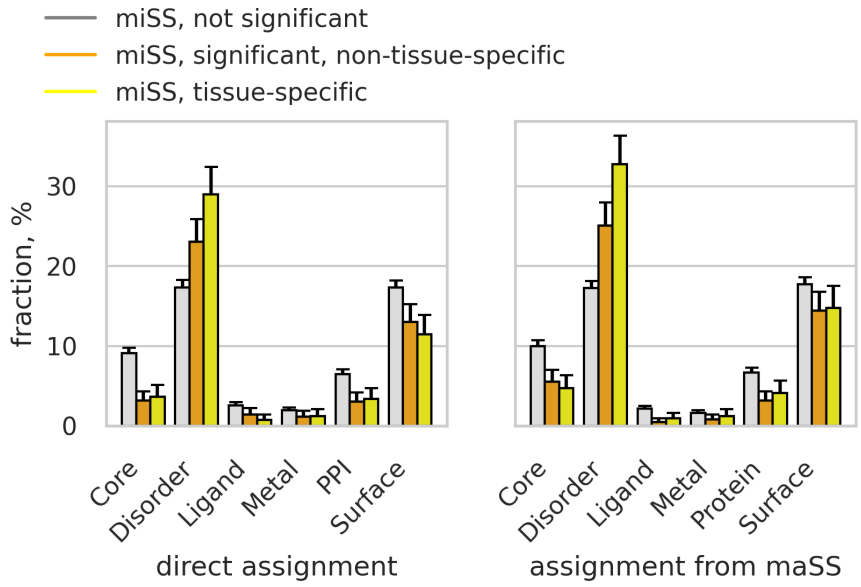

Figure A-1: **The comparison of the structural annotation assigned directly to miSS (left) or from the structural annotation of the corresponding maSS (right).** Only exonic miSS and corresponding maSS are considered.

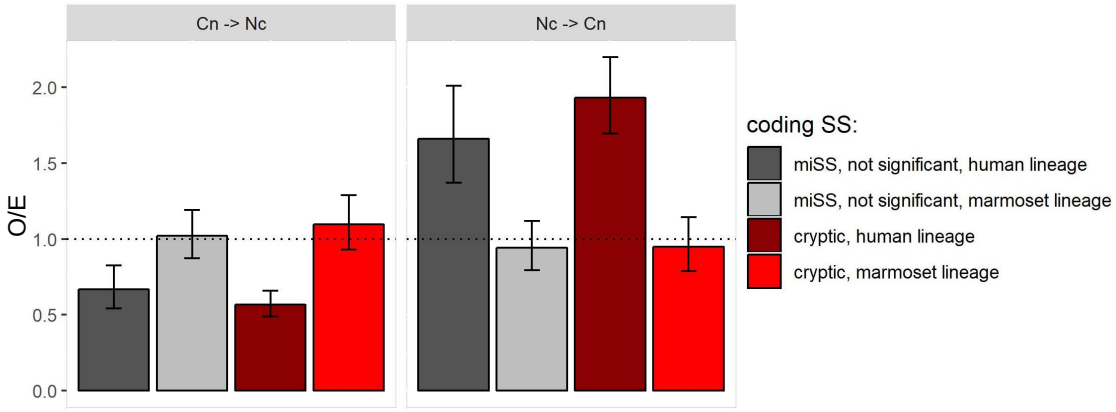

Figure A-2: **The selection of cryptic and not significant miSS in coding regions for marmoset and human genomes.**

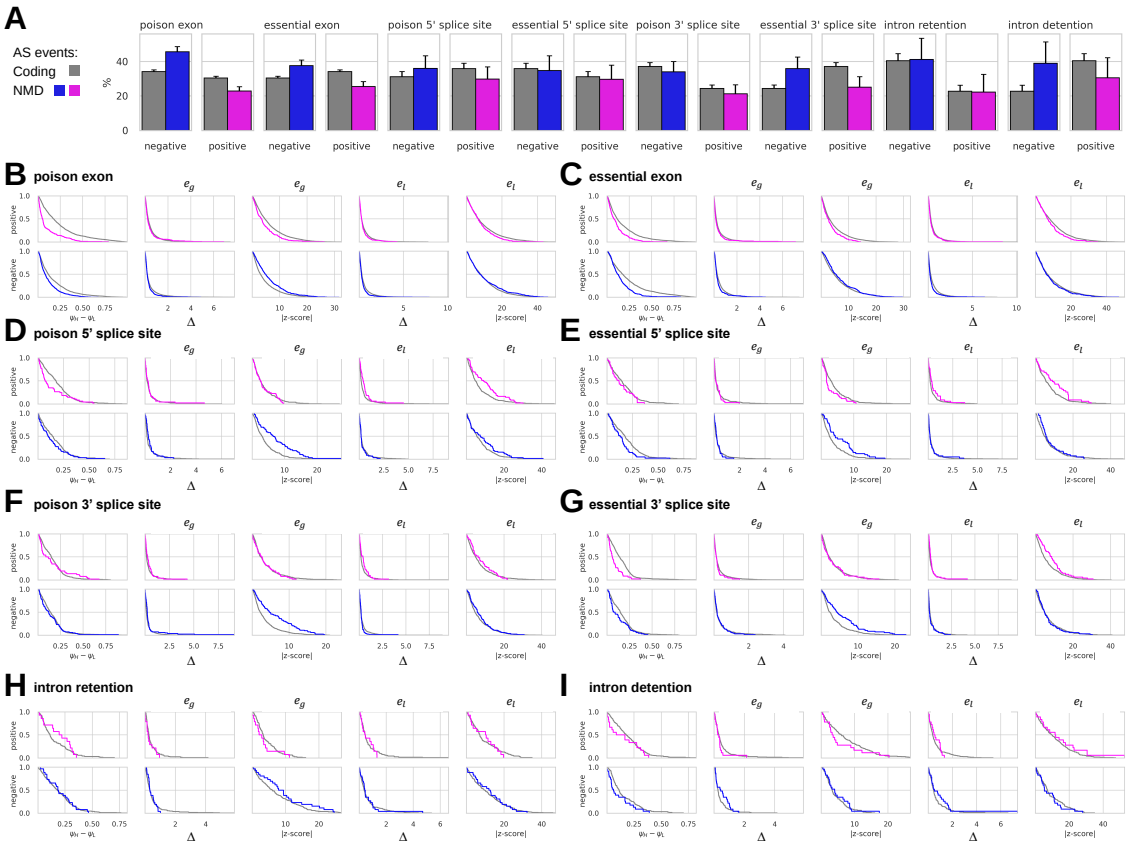

Figure A-3: **The features discriminating USEs and protein-coding AS events. (A)** The fraction of USEs and protein-coding AS events with the same sign of Δe_g and Δe_l (positive, $\Delta e_g > 0$ and $\Delta e_l > 0$, and negative, $\Delta e_g < 0$ and $\Delta e_l < 0$). Error bars represent 95% confidence intervals. **(B-I)** The distribution of $\psi_H - \psi_L$, Δe_g, z-score of Δe_g, Δe_l, z-score of Δe_l in events from the positive and negative sets (see panel A). The empirical cumulative distribution function (CDF) is shown as 1-CDF.

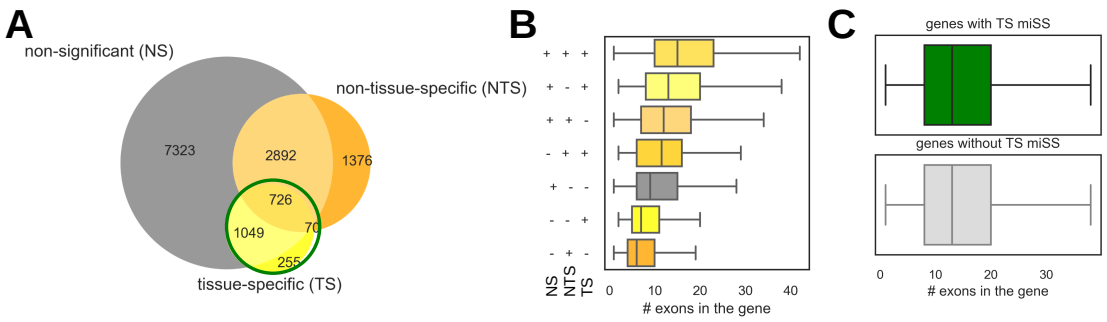

Figure A-4: **Genes containing tissue-specific miSS. (A)** Genes were characterized as having at least one non-significant (NS), significant, but non-tissue-specific (NTS), or tissue-specific (TS) miSS. **(B)** Genes having miSS belonging to different categories tend to contain more exons. **(C)** Matching genes having at least one TS miSS with genes having only NS or NTS miSS by the number of exons in the gene.

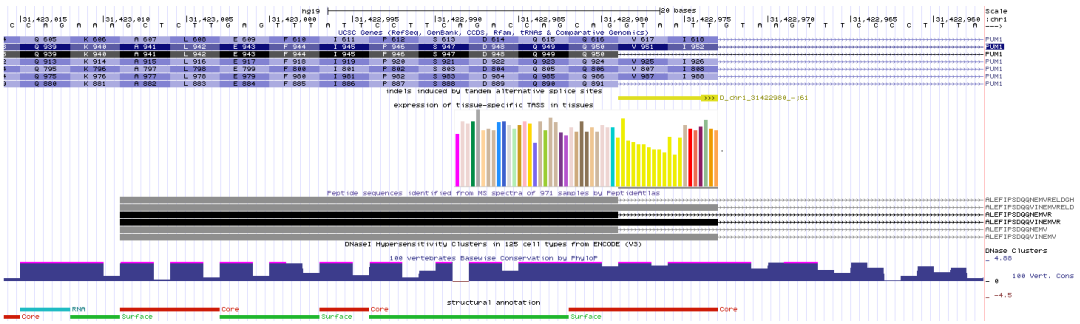

Figure A-5: **An example snapshot of the representation of the comprehensive catalogue of human TASS with a Genome Browser track hub.**

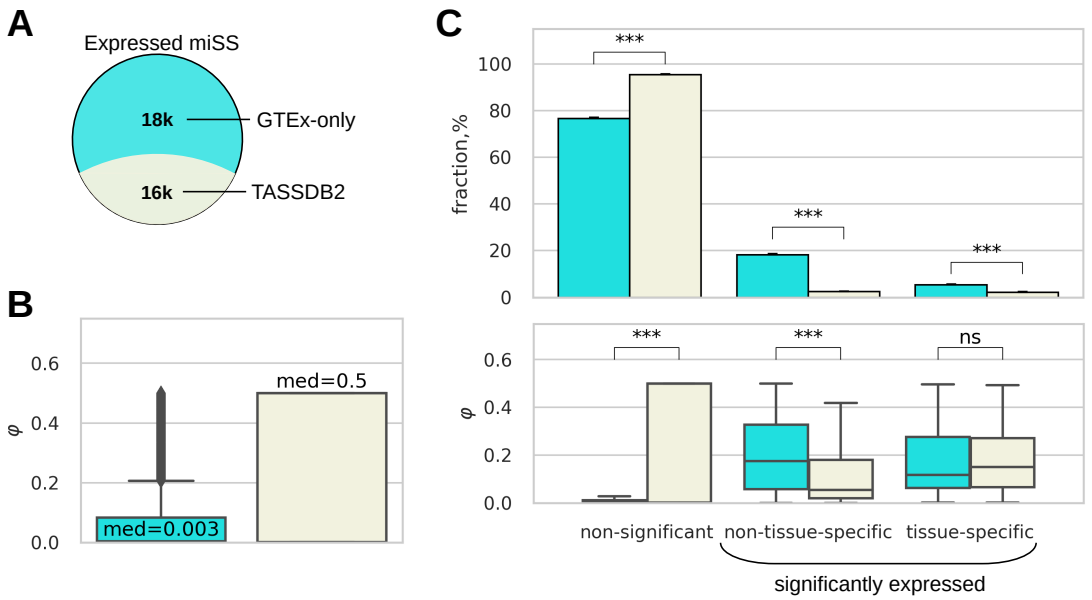

Figure A-6: **The constructed miSS catalogue extends the TASSDB2 database.** **(A)** The intersection of the set of expressed miSS with TASSDB2. **(B)** miSS not contained in TASSDB2 have on average lower ϕ values than miSS in TASSDB2. **(C)** miSS not contained in TASSDB2 are enriched with tissue-specific and non-tissue-specific significantly expressed miSS (top); within these categories they have similar or higher ϕ values compared with miSS in TASSDB2 (bottom).

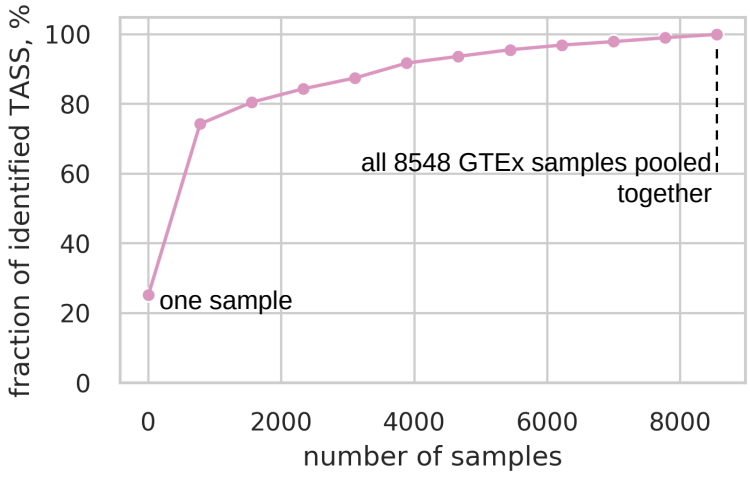

Figure A-7: **The dependence of the fraction of identified TASS on the number of considered samples.**

A.3 Supplementary tables

Online materials are available at https://doi.org/10.5281/zenodo.7097854

Table A.1: **Summary statistics at different filtration steps of the TASS catalogue.**

See Online materials.

Table A.2: **Accession codes for samples of shRNA RNP KD and eCLIP.**

See Online materials.

Table A.3: **Annotated USEs.** AS events that switch the protein-coding isoforms to NMD isoforms according to the GENCODE annotation. The position of the PTC in the NMD isoform is shown the column "stop_rel_position". The column "psi_ctl" shows the average value of ψ in the control. The column "max_delta_psi" shows the maximum value of $\Delta\psi$ observed in the NMD inactivation experiments. The last six columns show the values of $\Delta\psi$ in six analyzed NMD inactivation experiments. The names of these columns contain the data source, the cell line, and the experiment. NE and NS labels stand for not expressed and not significant values, respectively.

See Online materials.

Table A.4: **The list of RBP perturbation experiments and their accession numbers.**

See Online materials.

Table A.5: **The list of NMD inactivation experiments and their accession numbers.**

Data source	**Cell lines**	**Experiment**	**List of identifiers**	**Annotation**
ENCODE	K562	UPF1 KD	ENCFF243KYH ENCFF632FEO ENCFF055IEE ENCFF823LBA	exp exp ctl ctl
ENCODE	HepG2	UPF1 KD	ENCFF850ORL ENCFF819BTX ENCFF168UDP ENCFF321ZFN	exp exp ctl ctl
SRP090916	54-1	UPF1 siRNA	SRR4361751 SRR4361752	ctl exp
SRP063462	Hela	UPF1 shRNA	SRR2300536 SRR2300795 SRR2301041 SRR2301042	ctl exp ctl exp
SRP063493	Nalm-6	CHX treatment	SRR2313096 SRR2313097 SRR2313098 SRR2313090 SRR2313091 SRR2313092	exp exp exp ctl ctl ctl
SRP041788	Hek293	UPF1 siRNA + XRN1 siRNA	SRR1275416 SRR127541	exp ctl

Table A.6: **The correspondence between Proteomics DB tissues and GTEx tissues (SMTSD).**

See Online materials.

Table A.7: **Characteristics of miSS in different expression categories.** TS and non-TS stand for "tissue-specific" and "non-tissue-specific"

miSS	**non-significant**		**non-TS**		**TS**	
	%	#	%	#	%	#
coding %	78%	28,530	41%	2,770	77%	1,931
annotated %	10%	3,804	37%	2,542	74%	1,854
frame-preserving, coding %	35%	10,059	53%	1,467	59%	1,141

Table A.8: **GO-enrichment analysis of tissue-specific miSS.** Genes having at least one tissue-specific miSS were compared with genes having no tissue-specific miSS using GOrilla web server [218].

See Online materials.

Table A.9: **Abundance of tissue-specific miSS in tissues.**

tissue	**# tissue-specific miSS**	**# upregulated miSS**	**# downregulated miSS**
Brain - Cerebellar Hemisphere	525	372	153
Testis	510	344	166
Brain - Cerebellum	474	307	167
Brain - Nucleus accumbens (basal ganglia)	436	202	234
Muscle - Skeletal	427	138	289
Brain - Frontal Cortex (BA9)	416	204	212
Skin - Sun Exposed (Lower leg)	403	159	244
Brain - Cortex	391	169	222
Brain - Anterior cingulate cortex (BA24)	368	173	195
Brain - Caudate (basal ganglia)	354	130	224
Whole Blood	349	91	258
Brain - Spinal cord (cervical c-1)	339	142	197
Heart - Left Ventricle	338	78	260
Brain - Hypothalamus	334	150	184
Adipose - Subcutaneous	331	127	204
Brain - Hippocampus	329	123	206

Continued on next page

Nerve - Tibial	327	156	171
Pituitary	306	174	132
Esophagus - Mucosa	306	106	200
Thyroid	305	133	172
Brain - Amygdala	294	107	187
Heart - Atrial Appendage	294	74	220
Brain - Putamen (basal ganglia)	289	112	177
Brain - Substantia nigra	285	111	174
Skin - Not Sun Exposed (Suprapubic)	277	113	164
Artery - Tibial	275	102	173
Ovary	270	183	87
Pancreas	268	53	215
Adrenal Gland	268	115	153
Liver	263	95	168
Lung	258	96	162
Breast - Mammary Tissue	248	113	135
Small Intestine - Terminal Ileum	247	147	100
Esophagus - Muscularis	227	89	138
Adipose - Visceral (Omentum)	219	64	155
Prostate	219	110	109
Colon - Transverse	218	83	135
Colon - Sigmoid	215	110	105
Artery - Aorta	212	61	151
Uterus	208	149	59
Stomach	203	64	139
Spleen	198	102	96

Continued on next page

Esophagus - Gastroesophageal Junction	184	82	102
Kidney - Cortex	183	106	77
Artery - Coronary	176	73	103
Vagina	169	92	77
Minor Salivary Gland	150	78	72
Bladder	46	40	6
Cervix - Endocervix	27	24	3
Cervix - Ectocervix	25	23	2
Fallopian Tube	22	16	6

Table A.10: **miSS-RBP-tissue triples.**

See Online materials.

Table A.11: **Predicted cases of miSS regulation by RBP with eCLIP support.**

miSS	**gene name**	**RBP**	**shift**	**coding region**	**RBP action on miSS**
A_chr6_163984476_+	QKI	PTBP1	24	coding	suppression
A_chrX_102933579_-	MORF4L2	U2AF1	-51	non-coding	activation
D_chr7_99063734_-	PTCD1	EFTUD2	18	coding	suppression
A_chrX_102933579_-	MORF4L2	U2AF2	-51	non-coding	activation
D_chr10_70098399_+	HNRNPH3	SRSF1	-45	coding	suppression
A_chr12_123003598_-	RSRC2	U2AF2	-22	coding	suppression

Table A.12: **miSS reactive to *PTBP1* KD and OE.**

See Online materials.

Table A.13: **Expressed miSS.**

See Online materials.

Table A.14: **Validated USEs.** The list of USEs and their regulators, for which the experimental validation was reported in human or mouse cell lines or tissue specimens. The table key is the USE position in the gene for a particular organism, characterized by the type of AS event, the position of the PTC relative to the AS event, and the set of regulators and their mode of action (NMD-inhibiting/NMD-promoting) reported in the literature. The next column section presents the description of the carried out experiments and their results, followed by a short functional annotation of the USE and the targeted gene, manually curated genomic positions of splice junctions and intron retention sites, and links to the UCSC Genome Browser for USEs. Literature resources are cited as PubMed IDs (PMID).

See Online materials.

Table A.15: **Significant USEs.** The classification of validated and annotated USEs as significant or not significant. The values of $\psi_H - \psi_L$, Δe_g, $z - score$ of Δe_g, Δe_l, $z - score$ of Δe_l are listed.

See Online materials.

Table A.16: **Tissue-specific USEs.** For each USE, shown are deviations of the ψ from the pooled median, the deviation of e_g and e_l from the respective pooled medians. The column "Effect" indicates whether the test for tissue-specificity in a tissue gave the expected (negative) or the opposite (positive) association.

See Online materials.

Table A.17: **Regulation of the validated RBP-USE pairs.** The columns "DS" and "DE" represent the difference in NMD isoform inclusion and gene expression level under the condition of high RBP expression vs. low RBP expression, averaged over experiments. NS and NT labels denote not significant and not tested cases, respectively. The last three columns list the evidence of RBP binding from CLIP data and literature citations.

See Online materials.

Table A.18: **Regulation of tissue-specific RBP-USE pairs.** Shown are all RBP-USE pairs that were predicted from RBP perturbation assays for tissue-specific USEs. The columns "# expected" and "# opposite" show the number of tissues with the expected and the opposite associations between ψ and RBP expression. The last two columns show the evidence of RBP binding from CLIP data.

See Online materials.

Table A.19: **GO-analysis of tissue-specific USEs.** The output of functional annotation chart analysis of genes containing tissue-specific USEs vs. genes containing only non-tissue-specific USEs.

See Online materials.