1 Chapter: Introduction

DNA stores two predominant classes of information; (1) The sequences that serve as template strand to transcribe diverse types of functional RNAs, including mRNAs that are later translated into proteins; (2) Regulatory instructions to determine when and where RNAs are transcribed (Hood and Galas, 2003). The heredity units of this genetic information are called genes, and in complex multicellular organisms, they have variable expression patterns that largely define the molecular environment of different cell-types, enabling the definition of specialized cellular phenotypes with a single genome.

In eukaryotes, gene expression is a multi-step process which can be regulated at different levels. It begins with activation of promoter and enhancer sequences that control the transcription of a particular gene. Then, transcription initiation complexes bind to the gene promoters, recruiting transcript elongation factors that initiate the transcription. While the nascent transcripts are forming, there are a series of co-transcriptional events that occur before RNA synthesis is complete. For most genes, these pre-mRNA processing events include 5′ capping, splicing and 3′ polyadenylation. After these processes, mature mRNA molecules are ready to be exported to the cytoplasm, to become a template for protein synthesis, or to directly perform their roles as non-coding RNAs.

Both mRNA capping and polyadenylation correspond to pre-mRNA end processing events that are essential to produce mature mRNA molecules. During mRNA capping a guanosine residue is added to the 5’ mRNA end, forming an atypical 5′-5′ triphosphate bound (different from the regular 3’- 5’ triphosphate bond present between other mRNA nucleotides). Methylation of this guanosine residue at its N^7 position leads to the formation of a minimal CAP 5’ structure (CAP_0), but in higher eukaryotes further 2’-O-methylation methylations of the first and second transcribed nucleotides can give rise to extended CAP structures known as CAP_1 and CAP_2 (Leung and Amarasinghe, 2016; Wei et al., 1975). On the other hand, the process of

polyadenylation takes place at the 3′ end of the nascent pre-mRNA transcripts. Polyadenylation factors first cleave pre-mRNA transcripts and then synthesise a poly(A) tail. Both CAP and poly(A) tails are bound by proteins that promote the circularization and stability of mRNAs and therefore regulation of these pre-mRNA processing steps can have a deep impact on gene expression (Wells et al., 1998; Wilusz et al., 2001). While long poly(A) tails (>25 nt) promote mRNA stabilization, short poly(A) tails are often targets for uridylation, which triggers decapping and mRNA decay by 5′ exonuclease activity (Chang et al., 2014; Morgan et al., 2017; Rissland and Norbury, 2009).

In the pre-mRNA, non-coding RNA sequences (introns) are excised while the remaining RNA sequences (exons) are re-joined through a two-step transesterification reaction. This process is known as splicing, and it has a major determinant role of the mature mRNA sequence composition. The presence of introns can be detected in bacteria or eukaryotic organelles, however, they are most commonly present in eukaryotic nuclei. In both bacteria and eukaryotes, splicing is enabled by RNA structures that catalyze the two consecutive transesterification reactions. However, nuclear pre-mRNA splicing is carried out by the spliceosome, a large ribonucleoprotein complex which orchestrates the exon excision of all introns across the transcriptome, as opposed to bacterial introns that have their own catalytic activity which enables their removal from pre-mRNA transcripts.

The information contained in the resultant mRNA sequence highly depends on pre-mRNA processing regulation. Alternative polyadenylation can lead to mRNAs with different 3′ UTR length, which can have a direct repercussion over mRNA stability (Tian and Manley, 2017). At the same time, 5′ decapping and recapping cycles have been observed and their regulation could lead to fine control of transcriptome diversity (Trotman and Schoenberg, 2019). However, the majority of pre-mRNA sequence re-arrangements occurs during splicing, which can be regulated to generate a different selection of exonic sequences, having an enormous potential to regulate the sequences that are going to remain as mature mRNA sequences.

1.1 Splicing; a pivotal step of eukaryotic RNA-processing

Higher eukaryotic pre-mRNA often contains non-coding sequences known as introns. These are precisely removed during RNA splicing, which consists of two consecutive transesterification reactions (Fig 1.1). During the first transesterification reaction, a 2′ hydroxyl group (OH) from an adenosine residue, known as the branch point, performs a nucleophilic attack over the phosphate group from 3′-5′ phosphodiester bonds that connect 5′ exon-intron junctions. This initiates a bimolecular nucleophilic substitution (S_N2), in which 3′-5′ phosphodiester bonds at the 5′ss are broken while 2′-5′ phosphodiester bonds are formed between the branch point and the 3′ intronic ends, generating a lariat intermediary. During the second transesterification reaction, the same type of nucleophilic substitution (S_N2) takes place, forming new 3′-5′ phosphodiester bonds between 5′ and 3′ exons while breaking the 3′-5′ phosphodiester bond that connects 3′ intron-exon junctions. Thus, secondary product lariats are released, and these are thought to subsequently be rapidly degraded.

Figure 1.1: Splicing mechanism. Splicing occurs through two consecutive S_N2 transesterification reactions that lead to the branching and exon ligation. Br.A indicates the branch site. Yellow arrows represent electron movement during the nucleophilic attack, showing the corresponding intermediate state and highlighting the leaving group in red. Schematic was taken from (Lee and Rio, 2015)

1.1.1 Spliceosomal machineries

Introns are thought to have emerged during evolution through the invasion of mobile genetic elements in bacterial genes, which originally gave rise to a class of self-catalytic introns, known as Group II introns (Novikova and Belfort, 2017; Papasaikas and Valcárcel, 2016). Group II introns are still present in bacterial and eukaryotic organelle genes, but their presence has not been detected in the eukaryotic nucleus (Lambowitz and Zimmerly, 2011). Instead, nuclear eukaryotic splicing is enabled by spliceosomes. Both self-catalytic and spliceosomal splicing occur through analogue chemical mechanisms that involve two transesterification reactions. Since the catalytic RNA-structures that are present in Group II and spliceosomal introns are remarkably similar, self-catalytic splicing is thought to be the evolutionary ancestor of spliceosomal splicing.

Spliceosomes are some of the most complex molecular machines in eukaryotic cells and they are formed by more than a hundred proteins (~350 in human cells) and five small nuclear ribonucleoproteins (snRNPs). Eukaryotic cells often have two active parallel spliceosomal complexes, which differ mainly in their abundance and molecular composition. The most abundant spliceosome is known as the major spliceosome, while the less abundant is known as the minor spliceosome. Even though most of the protein components are shared between major and minor spliceosomes, U5 is the only snRNP shared between the two spliceosomes; U1, U2, U3 and U6 are exclusively part of major spliceosomes, while U11, U12, U4atac and U6atac are specific to minor spliceosomes.

Spliceosomal snRNPs are key components for splicing catalysis because they mediate RNA-RNA interaction between pre-mRNA and spliceosomes to allow for the precise recognition and processing of splice junctions. Due to differences in sequence composition between major and minor spliceosomal snRNPs, different types of splicing signals are processed by the two spliceosomes. Since minor spliceosome snRNPs are about 100-fold less abundant than the major snRNPs, introns that are processed by the major spliceosome (U2-type introns) are more

abundant and more efficiently removed than introns removed from the minor spliceosome (U12-type introns). Computational analyses of U2 and U12-type introns have shown that the loss of the minor spliceosome has occurred on several occasions through eukaryotic evolution (Bartschat and Samuelsson, 2010; Lin et al., 2010).

1.1.2 Canonical nuclear eukaryotic splicing

The precise recognition of splice sites relies on early spliceosome assembly over pre-mRNA intron-exon boundaries, which is primarily driven by RNA-RNA interactions between spliceosomal snRNPs and specific pre-mRNA consensus sequences. Among the core sequences that drive splice site recognition are 5′/3′ consensus sequences (sometimes referred to as splice donor and acceptor sites), branch sites and polypyrimidine tracts. Given that gene architecture can be very different across eukaryotic species, different spliceosomal mechanisms have evolved to adapt the spliceosome assembly over exon/intron junctions (De Conti et al., 2013). Moreover, splice site recognition can be influenced by the presence of RNA *cis-acting* sequence elements that are often bound by proteins that promote or inhibit spliceosomal assembly, having a direct impact on splicing efficiency (Matlin et al., 2005).

1.1.3 Core spliceosomal splicing signals

Precise intron removal relies on the recognition of consensus splice site sequences located at exon/intron junctions (Fig 1.2a). Within splice site consensus sequences, the 5′ and 3′ intronic ends are the most conserved regions. In U2-type intron, nearly all 5′ and 3′ intronic ends (~99%) correspond to GT-AG dinucleotides (Burset, 2000; Parada et al., 2014). In contrast, U12-type introns can be efficiently processed having GT-AG or AT-AC as terminal dinucleotides, and their splice site's consensus motifs have higher information content than U2-type introns, evidencing the relevance of splicing dinucleotide context for U12-type introns (Burge et al., 1999).

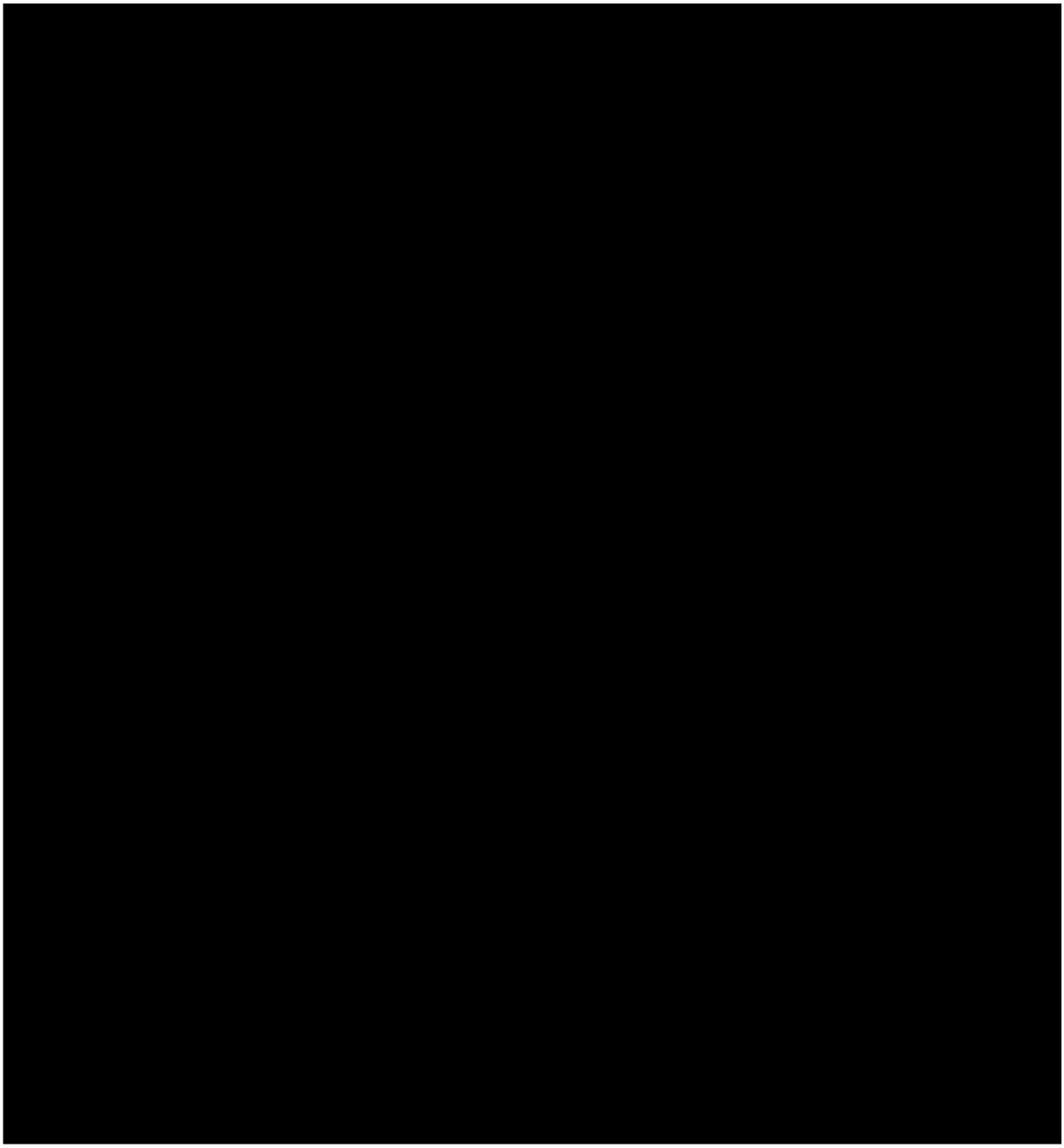

Figure 1.2: Spliceosomal core signals and assembly. A. Splicing consensus core signals of U2 and U12-dependent introns. Size of the letters is proportional to the positional frequency of nucleotides across 5′/3′ splice sites and branch sites. Notice that while U2-dependent introns have GT-AG dinucleotides, U12-dependent introns can have either AT-AC or GT-AG dinucleotides. Schematic taken from (Padgett, 2012) B. Co-transcriptional spliceosomal snRNPs assembly leads to the formation of different complexes. Initial recognition of splice sites results in the assembly of Complex E, which only through several re-arrangements and snRNP exchanges forms an activated complex B* that in turn catalyzes the first transesterification reaction. Further structural rearrangements lead to the formation of complex C, which catalyzes the second transesterification reaction that results in the formation of a post-spliceosomal complex (complex P) that is disassembled and to release the splicing products and recycle the snRNPs. Additional proteins that are involved during this process were omitted. Schematic was adapted from (Matera and Wang, 2014).

Branch sites also have consensus motifs around the adenosine residue that provide a free hydroxyl group for the first transesterification reaction (Fig 1.2a). For introns processed by the major spliceosome, the consensus sequences around the branch sites are highly degenerate and hence less conserved, whereas splice site sequences are highly conserved across U12-dependent introns (Levine and Durbin, 2001). Thus, the computational prediction of branch sites is imprecise in higher eukaryotes where intronic regions can span several kilobases, many of which in humans range between 10^2-10^3 kilobases. In fact, only through recently developed sequencing technologies, has it become possible to obtain a detailed map of an active splicing branch point in the human transcriptome (Bitton et al., 2014; Stepankiw et al., 2015). The analysis of these data suggests that most human introns can have multiple branch points, which means that there is often competition to react with a single 5′ splice site, and some of these branch points are frequently used in a tissue-specific manner (Pineda and Bradley, 2018).

Between the branch site and intron 3′ ends, U2-type introns have a polypyrimidine tract (spanning around 15-20 nucleotides in humans) which is directly recognized by the major spliceosome (Schellenberg et al., 2008). Although polypyrimidine tracts are absent in U12-type introns, their recognition serves as a key regulatory step during early spliceosome assembly. *In vitro* mutations of polypyrimidine tracts, splice sites or branch points have been shown to have a detrimental effect on splicing efficiency. In addition, mutations of these canonical splicing signals could account for about 10% of the heritable human disorders (Padgett, 2012). For example, mutations that disrupt or create splice sites at the laminin A locus can lead to multiple types of diseases, ranging from muscular dystrophy to premature ageing syndromes (Scotti and Swanson, 2016).

1.1.4 Spliceosome assembly and catalysis.

The assembly of spliceosomal components over nascent splice sites on the pre-mRNA molecule is a stepwise process which is highly conserved across eukaryotes (Fig 1.2b). Assembling both the major and minor spliceosome start with the recruitment of snRNPs to 5′ and 3′ splice sites, which are subsequently

rearranged to catalyze splicing thought analogous mechanisms. During the early assembly of the major spliceosome, 5′ss and 3′ss are precisely recognized by U1 and U2 snRNPs respectively, forming the complex E, which is the earliest spliceosomal complex that is committed to splicing. This initial step is largely driven by base-pairing interactions between the consensus sequences located at 5'ss and branch sites, and the corresponding U1 and U2 snRNPs, but it is also supported by additional protein factors, such as SF1 and U2AF heterodimers (U2AF65/U2AF35) that bind to the branch site and polypyrimidine tract, respectively. Once complex E is formed, it undergoes ATP-dependent rearrangements which promote the interaction between U1 and U2 snRNPs, leading to complex A formation. Then, recruitment of U4/U5·U6 tri-snRNPs to the 5′ss leads to the formation of the pre-catalytic B complex, which after the removal of U1 and U4 snRNPs and recruitment of additional protein factors, gets to its active form (complex B*) and catalyzes the first transesterification reaction. Lastly, snRNP rearrangements promote the formation and activation of complex C, which enables the catalysis of the second transesterification reaction necessary to complete the splicing process. All these result in the formation of exon-exon junctions across transcripts and the release of lariat RNAs as secondary products. Once splicing is completed, the spliceosome is released from the splice junction. However, some proteins that form part of the B and C complexes are deposited 24 nucleotides upstream of exon-exon junctions, forming what is known as exon-junction complex (EJC), which promotes stability of mRNAs, and is also involved in mRNA transport and translation (Le Hir et al., 2016).

1.1.5 Precise recognition of splice sites

The spliceosome has evolved to recognize *bona fide* splice sites across different eukaryotic transcriptomes. However, the splice sites that are recognized by the machinery do not always reassemble the consensus sequence, particularly in higher eukaryotes, which have weaker splice sites. To be able to recognize these weak splice sites, the spliceosome recognizes additional features that complete the missing information when splice sites deviate from the consensus sequence.

1.1.5.1 Cis-acting regulatory elements

One of the features that provide additional information for splice site recognition is cis-acting regulatory elements that can enhance or inhibit splice site recognition. These RNA sequence elements can be located within introns or exons and they are involved in defining both constitutive and alternative exons (Matlin et al., 2005). As a general rule, exonic splicing enhancers are bound by factors belonging to the SR protein family, while splicing exonic and intronic silencers are bound by heterogeneous nuclear ribonucleoproteins (hnRNPs). Both SR and hnRNPs have specific RNA-binding domains that allow them to bind pre-mRNA sequences and influence the formation of E and A complexes during early spliceosome assembly.

1.1.5.2 Intron and exon definition

Additional mechanisms of spliceosomal assembly enable further specificity to recognize splice sites. The principle of these mechanisms is to recognize two contiguous splice sites simultaneously, thereby dramatically decreasing the chance of spurious splice site recognition. To achieve this, spliceosomal complexes can assemble in two different ways; (1) Over introns, promoting cross-intron interaction of spliceosomal particles, which is known as intron definition, and (2) over exons, promoting spliceosomal interactions across exons, known as exon definition.

The modalities of spliceosomal assembly are highly influenced by the gene architecture found in eukaryotes. As a general rule, lower eukaryotic genes are characterized by large exons, interrupted by small introns, whereas higher eukaryotic genes tend to have the opposite pattern. This means that in lower eukaryotes the distances between intronic ends are short enough to allow intron definition. In contrast, higher eukaryotes are characterized by presenting relatively small exons (~120 nt long in mammals) and introns that can span hundreds to several hundred thousands of nucleotides (Ast, 2004). Given this gene architecture, the spliceosome assembly is more likely to form cross-exons rather than cross-introns, as exon splice ends are considerably further from each other. The proposal of exon definition during the early ’90s (Robberson et al., 1990), was fundamental to our understanding regarding how splice sites are recognised by the spliceosome in higher eukaryotic

organisms. Moreover, since the first and last exons are not flanked by splice sites on both sides, 5' CAP and 3' polyA structures are also involved in the respective definition of these exons.

Even though exon definition is dominant in higher eukaryotes, long and short introns co-exist in their genomes (McCullough and Berget, 1997). This is possible because small introns in higher eukaryotes (< 250 nt) can be defined by forming cross-intron spliceosomal assemblies. Moreover, artificial expansion experiments of short introns in vertebrates have shown that the splicing machinery can adapt to either assembly by intron or exon definition, depending on intron size (Sterner et al., 1996). By contrast, the expansion of small introns in *S. pombe* and *D. melanogaster* abolishes their splicing, suggesting that lower eukaryotic organisms cannot perform efficient exon definition to initiate early spliceosome assembly (Guo et al., 1993; McCullough and Berget, 1997; Mount et al., 1992; Talerico and Berget, 1994).

1.2 Widespread alternative splicing expands transcriptome and proteome diversity in vertebrates

In vertebrates, nearly all multi-exonic transcripts undergo alternative splicing, affecting approximately 95% of multi-exonic human genes (Pan et al., 2008; Wang et al., 2008). These alternative splicing events affect ~40% of human exons, which are involved in a range of different types of alternative splicing events (Zhang and Chasin, 2006). The most common type of alternative splicing in humans is the alternative inclusion of full exons, known as cassette exons (Bradley et al., 2012; Zhang and Chasin, 2006). There are three other basic types of alternative splicing: alternative 5′ss selection, alternative 3'ss selection and intron retention (Fig 1.3). All of these determine the inclusion of sequences that can have an impact on protein production or mRNA stability.

Each one of the different types of alternative splicing leads to the expansion of transcriptome diversity by generating isoforms with different combinations of splice site selection from a single gene. Alternative splicing can result in the coexistence of multiple isoforms from a single gene, and these can be in different concentrations

across different tissues. For instance, the GluN1 subunit of the NMDA receptor is encoded by a single gene, but it has eight different annotated isoforms with alternatively included exons that have an impact on GluN1 subcellular trafficking, receptor gating and pharmacological properties of NMDA receptors (Paoletti et al., 2013; Rumbaugh et al., 2000; Vance et al., 2012). GluN1 isoforms have overlapping expression patterns, but their relative proportions vary across neuronal tissues (Paoletti et al., 2013). Moreover, there are metazoan genes such as *slo*, *neurexin* and *Dscam* that can produce on the order of hundred to hundred thousand different mRNA isoforms through complex regulation of their splice site selections (Graveley, 2001). The recent development of high throughput screens based on CRISPR-based technologies has enabled genome-wide interrogation of exon exclusion events, evidencing widespread alternative splicing effects over cellular processes and allowing for deeper understanding of some of the mechanisms underlying alternative splicing regulation (Gonatopoulos-Pournatzis et al., 2018, 2020; Thomas et al., 2020).

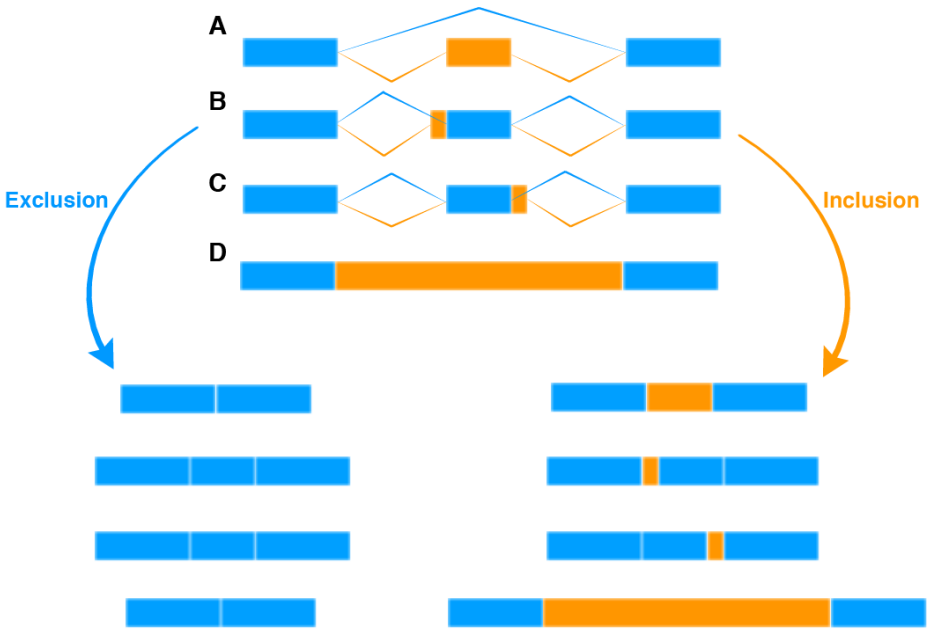

Figure 1.3: Alternative splicing event types. There are four basic types of alternative splicing types. A. Cassette exon inclusion/skipping. B. Alternative 5′ splice site selection. C. Alternative 3′ splice site selection. D. Intron retention. Alternatively included mRNA segments are coloured in orange.

1.2.1 Unproductive splicing events

The examples above demonstrate the great potential of alternative splicing to diversify the transcriptome and proteome in eukaryotic genomes. However, not all the generated isoforms lead to stable mRNAs. Alternative splicing is also coupled with cytoplasmic mRNA degradation by a pathway known as nonsense-mediated decay (NMD) (Lewis et al., 2003; Popp and Maquat, 2013). NMD takes place in the cytoplasm, but is highly determined by the EJCs that are deposited after splicing along the nascent mRNA transcripts. When ribosomes bind to mRNA during the pioneer round of translation, they displace EJCs that are on their path, and if after disassembly there are any EJCs still bound to the mRNA, NMD is triggered. Since premature stop codons (PTCs) incorporated by alternative splicing favour ribosome disassembly, EJCs downstream PTCs are not removed (unless the EJC is covering an exon-exon junction that is located ≤ 50-55 nt downstream the PTC) and as consequence NMD is triggered.

Transcriptome-wide studies indicate that around one in three alternative splicing events in human and mouse results in isoforms that are predicted to be targeted by NMD (Lewis et al., 2003; Pan, 2006; Weischenfeldt et al., 2012). Intron retention is one of the main alternative splicing events that lead to NMD, affecting as many as three-quarters of the multi-exonic genes in mammals (Braunschweig et al., 2014). Additionally, the inclusion of alternative exons or 5'/3' alternative splices site processing can directly incorporate a PTC, or it can induce frameshifts that can ultimately result in a PTC inclusion and degradation by NMD. Deep RNA-seq analyses have evidenced a large fraction of unannotated splice sites processed in very low proportions, which in part is believed to be attributed to stochastic mis-splicing events that result in isoforms that are degraded by NMD (Pertea et al., 2018; Pickrell et al., 2010; SEQC/MAQC-III Consortium, 2014). Since these splice sites are mostly not conserved between species, they are often considered part of the transcriptional noise that is generated by stochastic splicing errors. The measurement of splicing noise across RNA-seq samples have been used to

estimate the splicing error rate to be around 0.7% in normal human cells, but it could be higher in some types of cancers associated with higher splice site diversity (Kahles et al., 2018; Pickrell et al., 2010)

The systematic analysis of alternative splicing events that lead to NMD targeting have also uncovered highly conserved alternative splicing events coupled to NMD. One proposed function of these splicing events is to provide mechanism to downregulate gene expression, and this mechanism of gene expression regulation was termed regulated unproductive splicing and translation (RUST) (Lareau et al., 2004; Lewis et al., 2003; Lykke-Andersen and Jensen, 2015; McGlincy and Smith, 2008; Nickless et al., 2017). Among these events, the inclusion of exons that directly or indirectly introduce PTC is known as poison exons and they affect the gene expression of several splicing factors and other RNA-binding proteins (Desai et al., 2020; Lareau et al., 2007; Ni et al., 2007; Saltzman et al., 2008). Since RUST is a negative feedback loop mechanism for maintaining homeostatic protein levels for some splicing factors, mutations that abolish NMD pathway (such as UPF2 mutations) indirectly affect a wide range of splicing events (Ni et al., 2007; Weischenfeldt et al., 2012). Moreover, neuron-specific expression of certain genes is enforced by RUST (also referred to as AS-NMD), which have a key role during neuronal differentiation (Zhang et al., 2016; Zheng, 2016).

One-fifth of the conserved cassette exons between *H. sapiens* and *M. musculus* are predicted to be poison exons (Baek and Green, 2005). In fact, many ultraconserved and highly conserved elements identified across vertebrate genomes are associated with poison exons (Lareau et al., 2007; Ni et al., 2007). Recent CRISPR-Cas9-based screening have functionally interrogated highly conserved poison exons, and it has been reported that many are essential for cell growth and tumour suppression (Thomas et al., 2020).

1.2.2 Global assessment of alternative splicing and its impact over the proteome diversity

Technological advances in nucleic acid sequencing have led to the development of high throughput massively parallel RNA sequencing methods, commonly known as RNA-seq (Wang et al., 2009). The continuous development of RNA-seq methods and bioinformatics approaches to carry out the data analysis have enabled the characterization and quantification of alternative splicing events with unprecedented resolution (Engström et al., 2013; Lagarde et al., 2016; Mortazavi et al., 2008), positioning alternative splicing as a key RNA processing step to enhance transcriptome diversity.

However, given the substantial amount of isoforms that are degraded by the NMD pathway, it is reasonable to ask how much impact alternative splicing has in terms of proteome diversity and function. Despite numerous examples where alternative splicing plays a key role to regulate protein function, the vast majority of systematic evaluations of alternative splicing have been done over the transcriptome level, with limited evidence from proteomics data (Lee and Ji, 2017).

1.2.2.1 Mass-spectrometry based assays: Futile alternative splicing events or lack of sensitivity?

Analyses of publicly available proteomics data from eight large-scale proteomics experiments using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) have found 282 splicing events in human proteins (Abascal et al., 2015), which contrasts with the more than 100,000 distinct alternative splicing events that transcriptome-wide analyses have reported (Pan et al., 2008; Wang et al., 2008). Since the detection of alternative splicing events using proteomics may not be as sensitive as the transcriptomics approaches, the extent to which alternative splicing impacts proteome diversity remains a matter of debate (Blencowe, 2017; Tress et al., 2017a, 2017b). Even though a significant fraction of the splicing events observed could lead to RUST (therefore only having an impact over transcript diversity), more recent proteomic analyses have identified an increasing number of

genes affected by alternative splicing at the protein level (Lau et al., 2019; Wright et al., 2016).

From the experimental point of view, one of the technical challenges that proteome analyses have to overcome in order to have a sensitive detection of alternative splicing events is to provide enough peptide coverage across gene bodies to detect splice junctions (Aebersold et al., 2018; Blencowe, 2017). Peptides that span exon-exon junctions are critical to distinguish isoforms and identify alternative splicing events. Recent reports have shown underrepresentation of junction-spanning peptides in publicly available proteomics data due to a bias in fragmentation patterns resulting from trypsinization during the sample preparation (Wang et al., 2018). Despite these technical issues, recent integrative analysis of transcriptomic and proteomic datasets have shown consistent alternative splicing changes after U5 snRNP depletion, demonstrating that changes in alternative splicing contribute to both proteomic composition and diversity in humans (Liu et al., 2017). Thus, further development of experimental and bioinformatic approaches may enable researchers to overcome technical issues of proteomics analyses and corroborate or dispute the extensive alternative splicing events reported at the transcriptome level.

1.2.2.2 Alternative splicing events rewire protein interaction networks across tissues

Transcriptome profiling of vertebrates has unveiled distinguishable patterns of alternative splicing across tissues. Since tissue-specific cassette exons have a strong bias in their length to be a multiple of three (symmetric exons), their differential inclusion is less likely to trigger degradation by NMD (Baek and Green, 2005; Lewis et al., 2003). While several of these individual tissue-specific alternative splicing events have been associated with functional roles in development and cellular functions, less is known about the impact of their coordinated splicing events across tissues. Systematic analysis of tissue-specific exons showed that they are significantly enriched for disordered protein domains, which are often part of protein binding domains (Buljan et al., 2012; Ellis et al., 2012). These analyses showed that genes with tissue-specific exon inclusion are associated with more binding partners

and that they occupy central positions in protein-protein interaction (PPI) networks, suggesting that alternative splicing may have a major role in modulating and shaping PPI networks across tissues. To test this hypothesis, Yang and colleagues performed large scale protein binding profiling experiments of full-length alternatively spliced isoforms (Yang et al., 2016). Their results showed that the majority of alternative splicing events tested changed more than 50% of the protein interaction partners, providing evidence of transcriptome-wide effects of alternative splicing over PPI networks.

1.3 Fine-tuned control of alternative splicing

The mechanisms that lead to tissue-specific splicing patterns are mainly driven by the recognition of *cis*-regulatory elements (introduced in section 1.1.5.1). These elements are bound by *trans*-acting RNA-binding proteins (RBPs), which can promote or inhibit the formation of E and A complexes, that ultimately determine the commitment of the spliceosomal machinery to carry out splicing (Matlin et al., 2005). Thus, the expression patterns and activity of these *trans*-acting RBPs can strongly drive the tissue-specific alternative splicing patterns that are observed in RNA-seq experiments.

1.3.1 Features associated with alternative splicing events

To have precise control of an alternative splicing event, having regulatory elements that can be bound by RBPs is not the only requirement. In addition, the activity of the regulators must have a significant effect on spliceosome assembly. If the splicing signals and context of a given exon lead to a near-optimal recognition by the spliceosome, then it is likely that this exon will be constitutively recognized. In fact, ~60% of human exons are constitutively spliced (Zhang and Chasin, 2006). The presence of several features associated with the splice site sequence composition and intron-exon structures have been shown to be characteristic of alternative splicing events.

1.3.1.1 Splice site strength

One characteristic feature of alternative splicing events is their sub-optimal recognition by the spliceosome. This is in part due to their weaker splice sites in comparison with constitutively processed exons (Ast, 2004; Carmel et al., 2004; Stamm et al., 1994). Deviations from the splice site consensus sequences increase the free energy of U1 binding, making splice site recognition less efficient (Carmel et al., 2004). This makes the splice site recognition be conditioned by the action of regulatory elements that can promote or prevent splicing of a given weak splice site (Luco et al., 2011). Computational analyses of orthologous alternative and constitutive exons between mouse, rat and human show that alternative splicing sites are under selection to be weak (Garg and Green, 2007). Moreover, the weakening of alternative splice sites has been proposed as an evolutionary mechanism by which constitutive alternative splice sites can become alternative (Ast, 2004).

1.3.1.2 Gene-architecture effect on alternative splicing

The gene architecture of eukaryotes has an impact on alternative splicing. In lower eukaryotes, where intron definition is the dominant spliceosomal assembly mechanism, intron retention is the most prevalent type of alternative splicing (Keren et al., 2010; Kim et al., 2008). Conversely, in higher eukaryotes, where exon definition is the most common splicing assembly modality, the most common alternative splicing event corresponds to differential inclusion of cassette exons (De Conti et al., 2013). Experiments show that increase of mammalian intron size leads to exon skipping, which is supported by a computational analysis that shows that exon skipping is more likely to occur when the exons are flanked by long introns (Fox-Walsh et al., 2005; Kim et al., 2007; Sterner et al., 1996). At the same time, experimental expansion of vertebrate exons results in exon skipping (as the exon definition is blocked), but when the same enlarged exons are situated in between short flanking introns, are included again (Sterner et al., 1996).

Evolutionary analyses across 17 vertebrate genomes have shown an expansion of intron sizes though vertebrate evolution, where mammals have significantly longer

introns than their vertebrate ancestors, with primates having the longest intron sizes (Gelfman et al., 2012). As predicted by the intron expansion experiments discussed above, the expansion of intron size is correlated with the number of alternative splicing events observed across vertebrates, with primates displaying the largest proportion of alternative splicing events (Barbosa-Morais et al., 2012). Moreover, the strength of the splice sites has an effect on intron expansion through vertebrate evolution, where the presence of weak splice sites restricts the intronic expansion, demonstrating that both splice site sequences and gene architecture are important factors that modulate splice site recognition (Gelfman et al., 2012).

1.3.1.3 Epigenetic modulation

The epigenetic context also has an incidence over exon definition and alternative splicing. Genome-wide mapping of nucleosome positioning shows an enrichment of nucleosomes over exons, which is a conserved trend from plants to mammals and possibly favoured by higher exonic GC-content (Andersson et al., 2009; Gaffney et al., 2012; Li et al., 2018; Luco et al., 2011; Nahkuri et al., 2009; Schwartz et al., 2009; Tilgner et al., 2009; Tillo and Hughes, 2009). Since the length of DNA wrapped around nucleosomes (~147 nt) resembles the average exon size, nucleosome positioning has been proposed to have a role in exon definition. This model is supported by the observation that exons flanked by long introns have higher enrichment of nucleosomes than exons flanked by short introns (Spies et al., 2009). As mentioned above, splicing recognition of exons flanked by long introns tends to be more inefficient and is associated with alternative splicing. Thus nucleosome positioning may contribute to exon recognition of intrinsically inefficient splice sites. This hypothesis is also supported by the pronounced enrichment of nucleosomes at weak splice sites and deceased nucleosome occupancy at pseudoexons (Tilgner et al., 2009).

During transcription, RNA polymerase II slows down upon the encounter of nucleosomes and their positioning over exons might have a kinetic effect on splicing (Hodges et al., 2009; Keren et al., 2010; Luco et al., 2011). Slowing the elongation rate of RNA polymerase II leads to higher inclusion rates of exons (Kadener, 2001;

de la Mata et al., 2003; Nogués et al., 2002). Thus, nucleosomes can act as 'speed bumps', giving more time to RNA polymerase II to recruit splicing factors that allow an efficient recognition of splice sites (Keren et al., 2010; Luco et al., 2011). Moreover, nucleosomes that are positioned over exons are often subject to histone modifications, which can promote the recruitment of additional regulatory *trans*-acting factors, providing an additional regulatory layer of alternative splicing control (Andersson et al., 2009; Luco et al., 2011).

1.3.1.4 Effect of secondary structures

As the pre-mRNA is being generated, the formation of RNA structures influences alternative splicing by diverse mechanisms (Jin et al., 2011). RNA secondary structure analyses have demonstrated this association with alternative splicing events (Shepard and Hertel, 2008). Local RNA structure formation can have an impact on splicing by restricting the accessibility of core splicing signals (Buratti and Baralle, 2004; McManus and Graveley, 2011). In addition, RNA secondary structures can modulate the activity of cis-regulatory elements by conditioning the binding of splicing factors (Buratti et al., 2004; McManus and Graveley, 2011). For example, RNA secondary structure formation can restrict the accessibility of MBNL1 and RBFOX2 binding sites (Taliaferro et al., 2016). Given that the analysis of RBP crosslinking immunoprecipitation (CLIP)-seq data shows that most occurrences of consensus RBP binding motifs are not bound *in-vivo*, RNA structures may provide additional contextual features beyond the primary motif sequences (Taliaferro et al., 2016; Van Nostrand et al., 2016).

The formation of RNA structures can also enhance RBP regulatory range by bringing distal regulatory elements in close proximity with their exon targets (Lewis et al., 2017a). This can be particularly important for RBFOX2 regulated exons since more than half of RBFOX2-binding sites are found over 500 nt away from any annotated exons (Lovci et al., 2013). Moreover, the formation of long-range RNA secondary structures can bring in contact with regulatory elements that are even further apart. The best-characterized example can be found in *D. melanogaster* for the DSCAM

gene, where RNA-RNA interactions regulate the selection of exons within arrays of mutually exclusive exons (Graveley, 2005; Yang et al., 2011).

RNA secondary structures may also have direct effects over exon skipping events by a mechanism known as “looping-out”, in which inter-intronic base-pairing RNA interactions can loop out exons to promote their skipping (Jin et al., 2011). This mechanism is supported by the enrichment of conserved complementary sequences present in intronic sequences flanking exon skipping events (Miriami et al., 2003). Moreover, the artificial introduction of self-complementary regions across exons suppresses exon inclusion in yeast, suggesting a cause-effect relationship between RNA-structure and exon skipping (Howe and Ares, 1997). The expansion of these self-complementary regions through primate evolution is related to primate-specific retrotransposons, called Alu elements, which are enriched in alternative exons flanking regions, suggesting regulatory roles over alternative splicing (Lev-Maor et al., 2008).

1.4 Non-canonical splicing feature effects over alternative splicing

As discussed above, features that lead to suboptimal recognition of splice sites are often associated with alternative splicing events. For example, weak splice sites or unusual exon-intron structures are often targets of regulatory features, enabling fine-tuned regulation of alternative splicing events. However, there are several more extreme examples of this phenomenon, which involves splicing signals or gene structures that defy the canonical exon definition model.

Splice site signals or splicing mechanisms that do not fit the classical model of splicing recognition are known as non-canonical splicing events. In the following section, I will be discussing different types of non-canonical splicing events, most of which are reviewed by Sibley and colleagues (Sibley et al., 2016).

1.4.1 Unusual splice sites

1.4.1.1 A minority group of introns is processed by a dedicated parallel spliceosomal machinery

The first class of splice sites to be considered non-canonical, generally corresponding to AT-AC introns, are processed by the minor spliceosome (see section 1.1.3). They correspond to around ~0.35% of human splice sites, which is a much smaller frequency in comparison with the amount of splice sites processed by the major spliceosome (~99%) (Burset, 2000; Parada et al., 2014; Patel and Steitz, 2003; Tarn and Steitz, 1996; Verma et al., 2018). Yet, they are processed by parallel spliceosomal machinery, known as the minor spliceosome, in which the catalytic core is based in a dedicated set of snRNPs, including U11, U12, U4atac and U6atac, plus U5 snRNP that is the only common in both spliceosomes (Patel and Steitz, 2003; Tarn and Steitz, 1996). The minor spliceosome processes both AT-AC and GT-AG intron, but unlike U2-dependent introns, U12-dependent introns splicing is slower and does not depend on the presence of long polypyrimidine tracts as for U2-dependent introns.

1.4.1.2 Non-canonical splice sites

Despite the fact that the recognition of GT-AG/AT-AC dinucleotides is context-dependent, disruption of canonical dinucleotides have abolishing effects over splicing efficiency, leading to the accumulation of intermediary splicing products and cryptic splice site activation (Aebi et al., 1986; Montell et al., 1982). Even though there are strong restrictive rules regarding dinucleotide composition, exceptions to dinucleotide spliceosomal rules have been detected. The most common deviation is GC-AG introns, which are usually processed by the major spliceosomes and often involved in alternative splicing events (Jackson, 1991; Shapiro and Senapathy, 1987; Thanaraj and Clark, 2001). The systematic analysis of expressed sequence tags (EST), full-length cDNA and RNA-seq have identified additional variants of the dinucleotide rules (Burset, 2000; Parada et al., 2014; Sibley et al., 2016). During my previous work, I analysed RNA-seq data to provide a *bona fide* annotation of non-canonical splice sites. Since most of the raw detected introns were not

biologically meaningful, we developed a systematic set of filters to generate a high confidence list of non-canonical splice sites in the human genome (Parada et al., 2014). As expected by their weak splice site nature, the number of non-canonical U2 and U12-dependent introns is limited, but they are highly involved in alternative splicing (Parada et al., 2014; Szafranski et al., 2007). Moreover, the presence of non-canonical splice sites is often compensated by cis-regulatory elements that enable the recognition by the spliceosomal complexes (Brackenridge, 2003; Parada et al., 2014).

1.4.1.2.1 XBP1 intron is the only known nuclear intron that is not processed by the spliceosome

The only nuclear RNA that is known to be processed by non-spliceosomal machinery is the one present at XBP1. In metazoans, as part of the unfolded protein response pathway, the non-canonical splice sites of XBP1 are recognized and processed in the cytoplasm by IRE1α (Cox and Walter, 1996). Efforts to discover novel non-spliceosomal splice sites in humans using RNA-seq data have been discouraged by the presence of RT-artefacts during the cDNA reverse transcription necessary for most RNA-seq technologies (Parada et al., 2014). Even though recent RNA-seq analyses in plants suggest the presence of novel nuclear non-spliceosomal introns their artifactual origin cannot be discarded (Pucker and Brockington, 2018). Newly developed technologies are enabling the direct sequencing of single RNA molecules (Garalde et al., 2018), which might open new opportunities for the systematic search for nuclear non-spliceosomal introns.

1.4.1.3 Cryptic-splice sites

The spliceosome is able to discriminate against suboptimal splice sites due to mechanisms that promote splicing fidelity, such as exon definition and activity of DEAD/H-box ATPases (De Conti et al., 2013; Semlow and Staley, 2012). However, since vertebrates tend to have long introns, for example in humans most of them range is between 10^5-10^6 nt long (Coelho and Smith, 2014), the splicing machinery is prone to errors and processing of suboptimal substrates. This group of sub-optimally recognizing splice sites are known as cryptic splice sites (Sibley et al., 2016).

Recognition of cryptic splice sites can lead to the introduction of whole exons (cryptic exons) or additional 5′/3′ alternative splice sites, and they often promote the inclusion of a PTC and mRNA degradation by NMD. Several surveillance mechanisms that disfavour the recognition of cryptic splice sites have been described (Boehm et al., 2018; Ehrmann et al., 2019; Zarnack et al., 2013). However, mutations can lead to activation of cryptic splice sites which have been linked to cancer and other genetics diseases (DeBoever et al., 2015; Singh and Cooper, 2012).

1.4.1.4 U2AF65 independent splicing

While non-canonical splice sites are predicted to cause inefficient splice site recognition, their processing still depends on the effective recognition of splice site signals by the spliceosomal ribonucleic protein complexes. However, in some exceptional cases, the recognition of core splicing signals can be bypassed. For example, even though $U2AF^{65}$ is thought to be part of the core spliceosomal machinery, a subgroup of zebrafish introns can undergo $U2AF^{65}$-independent splicing. The recognition of most intron branch sites is carried out by the U2AF complex, in which $U2AF^{65}$ is a key subunit that has been shown to be sufficient and necessary for the splicing of some introns (Guth et al., 1999; Ruskin et al., 1988; Smith and Valcárcel, 2000). Lin and collaborators identified a set of highly stable secondary structures that enable $U2AF^{65}$-independent splicing. These are hairpin-like structures formed by Intronic repeats AC and GT, respectively positioned at 5′ and 3′ intronic ends and can promote accurate splice definition regardless of the absence of polypyrimidine tract sequences (Lin et al., 2016).

1.4.2 Non-canonical intron-exon structures

The exon recognition model was originally proposed to explain how relatively small exons are recognized from much longer intronic sequences, which in humans cover around 23% of the entire genome (Sibley et al., 2016). Even though exon definition is the most common spliceosomal assembly across vertebrate genomes, some vertebrate gene structures favour intron definition (Gelfman et al., 2012). Particularly, some vertebrate small introns can lead to intron definition when their flanking exons are medium or large size, evidencing that in some vertebrates spliceosome

assembly is able to adapt to different exon-intron structures (De Conti et al., 2013; Lim and Burge, 2001; Sterner et al., 1996).

Both exon and intron definition mechanisms involve simultaneous recognition of 5′ and 3′ splice sites, which is thought to be an evolutionary adaptation to avoid the recognition of spurious splice sites. However, since the spliceosomes correspond to large macromolecular structures whose molecular mass is estimated to be ~2.5 MDa and given its physical dimensions it has been predicted to span between 85-113-nt linearized RNA (Behzadnia et al., 2007; Sasaki-Haraguchi et al., 2012; Wahl et al., 2009). Even though the presence of intron and exons that are smaller than 65 nt are rare, their existence in vertebrate annotation databases suggests that additional mechanisms exist to enable spliceosome assembly around extremely close splice sites.

1.4.2.1 Analysis of short and ultra-short introns

Even though short introns are relatively common in invertebrates, in mammals they represent a minority group (Lim and Burge, 2001). Since the intron length varies across eukaryotes, Lim and Burge fit lognormal mixture models to identify populations of small introns relative to the different intron size distributions of humans and four other eukaryotes. Based on the lognormal mixture models they defined a cutoff to extract groups of short introns relative to their species-specific size distribution (134 nt for humans). In addition to finding the core splicing signals associated with U2-type introns, short introns were also found to have an enrichment of G triplets (GGG), which are well known to be associated with intronic splicing enhancers (Lim and Burge, 2001; McCullough and Berget, 1997, 2000). This suggests the presence of a compensatory mechanism that allows the recognition of short introns.

Further analyses have focused on a group of introns with even shorter sizes: ultra-short introns, which in humans are defined as introns 65 nt or shorter (Sasaki-Haraguchi et al., 2012; Shimada et al., 2015). Since the size of these introns is predicted to be smaller than the amount of RNA that is spanned by the spliceosome (85-113-nt), the processing of ultra-short introns defies the standard

intron definition model (Behzadnia et al., 2007; Sasaki-Haraguchi et al., 2012; Wahl et al., 2009). Despite the theoretical constraints of ultra-short intron processing, Sasaki-Haraguchi and colleagues found ultra-short introns annotated in human transcript databases. Through RT-PCR and minigene analyses they demonstrated that the removal of these introns was dependent on spliceosomal activity and strongly depends on the presence of G-rich intronic enhancer sequences.

Even though further bioinformatic analyses and RT-PCR experiments have identified possible shorter introns (< 43 nt) in the human transcriptome, their detection is often associated with non-canonical splice sites that do not reassemble U2-type or U12-type core splicing signals (Sasaki-Haraguchi et al., 2012). Among these, XBP1 is a well established 26-nt non-spliceosomal intron that is removed by the endonuclease activity of IRE1α. Potentially novel ultra-short introns have been detected in RNA-seq data, and they are mostly associated with non-canonical splice sites and strong secondary structures. Since intramolecular RT template switching is also a well-known source of spurious intron detection in transcriptomic data (Cocquet et al., 2006; Houseley and Tollervey, 2010; Mader et al., 2001; Parada et al., 2014; Roy and Irimia, 2008), more evidence is needed to confirm or refute the existence of ultra-short microexons shorter than 43-nt, particularly those lacking spliceosomal signals.

1.4.2.2 Microexons

Since exon definition is the most frequent spliceosomal assembly mechanism across vertebrates, the length of exons is also a critical feature that affects splicing. Manipulation of exon sizes has indicated that extension or shortening of exons is detrimental to splicing efficiency due to interference with the spliceosomal exon definition. However, extremely short exons, known as microexons (≤ 30) have been reported (Beachy et al., 1985; Cooper and Ordahl, 1985; Santoni et al., 1989; Small et al., 1988; Volfovsky et al., 2003). A subgroup of microexons has been identified to have strong neuronal-specific inclusion patterns (Irimia et al., 2014; Li et al., 2015). The neuronal regulation of microexons is dynamic and has the most highly conserved network of alternative splicing events currently described in vertebrates.

Flanking intronic regions of neuronal microexons are often associated with strongly conserved regions, which largely correspond to cis-regulatory elements that are essential for their recognition.

The regulation of microexon alternative splicing events is closely related to their size. In experiments where microexon sequences have been expanded, they lose their tissue-specific alternative splicing patterns (Black, 1991). Thus, the size of microexons and their effect on exon definition might be another example by which sub-optimal recognition of splicing features are related to tissue-specific alternative splicing events.

1.4.2.3 Recursive splicing

In vertebrates, introns tend to often be an order of magnitude bigger than exonic sequences. The removal of some long intronic regions has been shown to be the result of the splicing of several smaller introns through a process known as recursive splicing. Recursive splicing often involves the processing of 3′ and 5′ splice sites that are next to each other, denoted as recursive splicing (RS) sites. Since the recognition of adjacent splice sites from RS sites does not promote the inclusion of extra exonic sequences, these splice sites are often described as 0-length exons. One of the possible mechanisms to avoid steric hindrance during RS site processing involves downstream recognition of cryptic 5′ splice sites (Sibley et al., 2015). The initial recognition of both RS 3′ splice site and downstream 5′ leads to the definition of a longer exon (RS-exon) which enables spliceosomal processing (Blazquez et al., 2018; Sibley et al., 2016).

1.5 Non-canonical nucleic acid structures

Initial understanding of DNA structure gave fundamental insights into how genetic information flows inside the cell and across generations (Watson and Crick, 1953). The canonical and most common DNA structure found in living systems corresponds to a right-handed double helix, known as B-DNA. Even though B-DNA is the most stable structure under physiological conditions, other alternative DNA structures have also been characterized. These non-B DNA secondary structures include

Z-DNA, hairpins, cruciforms, slipped structures, intramolecular triplexes (H-DNA) and G-quadruplexes (Bochman et al., 2012; Kaushik et al., 2016). Even though most of the DNA segments are structured as the canonical B-DNA conformation, some sequences (here referred to as non-B DNA motifs) are more likely to form alternative structures under favourable conditions. Alternative conformations of the DNA are often formed as a by-product of biological processes, such as transcription, replication, recombination and DNA repair, which can lead to transient conformational changes or long term stabilization of alternative DNA structures (Kouzine et al., 2017; Wang and Vasquez, 2017). Non-B DNA motifs generate local distortions of the B-DNA structure and promote the formation of single-stranded DNA, which is vulnerable to damage (Pannunzio and Lieber, 2018). To prevent this, a number of helicases are involved in non-B DNA structure destabilization (or unwinding). The understanding of the dynamic conformational changes of B-DNA structures is key to identifying sources of genome instability (Georgakopoulos-Soares et al., 2018; Zhao et al., 2010).

Some non-B DNA structures are not only associated with genome instability and recurrent mutations, but they also play a role in gene expression regulation. For example, G-quadruplexes are enriched in promoters and nucleosome depleted regions, suggesting an active role in gene expression regulation (Hänsel-Hertsch et al., 2016; Huppert and Balasubramanian, 2007). Since Non-B DNA structures represent deviations from the B-DNA substrate that RNA pol II uses as a template, elongation rates during transcription can be affected by the presence of non-B DNA structures, which may have a kinetic impact on alternative splicing (Nieto Moreno et al., 2015). However, little is known regarding how non-B DNA structures can impact alternative splicing or other RNA processing events.

1.5.1 G-quadruplex formation

Among non-B DNA structures, G-quadruplexes influence over genomic instability and gene expression have been one of the most studied (Fay et al., 2017). G-quadruplex formation is driven by the inherent propensity of guanines to self-assemble (in the presence of monovalent cations) into planar structures known

as G-quartets (Bang, 1910; Gellert et al., 1962). Each G-quartet is composed of four guanine nucleotides that interact with each other through cyclic Hoogsteen hydrogen-bonds (Fig 1.4a). The presence of runs of guanines (G-tracts) in either DNA or RNA may lead to the formation of consecutive G-quartets that can stack with each other to form G-quadruplexes (G4s) structures (Fig 1.4a-b). Ultimately, the formation of a G4 can modulate gene expression at different stages, not only having an effect on gene expression levels, but also on RNA processing events.

Diverse computational and experimental evidence indicates that G4s formed at the DNA level (DNA G4) are enriched at promoters and have an impact on their activity. Moreover, an increasing amount of evidence suggests an important role of G4s formed at the RNA level (RNA G4). During DNA replication and RNA transcription, helicase activity is required for DNA and RNA G4 unwinding, therefore G4 formation may have an impact over DNA/RNA polymerization kinetics. In fact, recent genome-wide DNA polymerization speed measurements indicate a global effect of G4s and other non-B DNA structures on DNA polymerization and mutation rates (Guiblet et al., 2018). On the other hand, the genome-wide *in-vivo* formation of RNA G4s is a matter of debate and putative effects over gene expression have just recently begun to be systematically explored (Biffi et al., 2014; Guo and Bartel, 2016; Kwok et al., 2018). RNA G4s may favour or block the binding of RBPs and their formation has been related to splicing, 3′ processing, transcription termination, RNA localization and translation regulation (Fay et al., 2017).

One of the first exemplary cases of G4-mediated regulation of alternative splicing was found in the hTERT gene, which encodes for the catalytic subunit of the telomerase enzyme, and one of its exon skipping events is promoted by the stabilization of intronic G4s (Gomez et al., 2004). Gomez and colleagues hypothesized that G4 formation can prevent RBP binding to intronic enhancers, leading to exon skipping. However, based on different functional assays, G4 formation has also been proposed to promote RBP binding to splicing enhancers (Didiot et al., 2008; Marcel et al., 2011; Ribeiro et al., 2015). Since G4-dependent splicing events were often demonstrated by introducing mutations at G4 motifs, it was unclear from these results whether the G4 or the linear form of these G-rich

sequences act as a splicing enhancer. To disentangle these effects, Huand and colleagues showed that mutations that prevent intronic G4 formation but keep G tracts intact, led to exon exclusion of an alternative exon in the CD44 gene (Huang et al., 2017). Since CD44 intronic G4 motif sequence can be bound by two RBPs that have the opposite effect on CD44 exon exclusion, RNA G4 formation may function as a switch to promote one RBP binding over the other (Bartys et al., 2019). However, the genome-wide effect of RNA G4 formation over splicing factor binding remains unclear.

The implementation of dual-colour splicing reporters to perform high-throughput screening of chemical compounds that can regulate alternative splicing in a G4 dependent manner has made it possible to identify two small molecules, emetine and cephaeline, that disrupt G4 formation (Zhang et al., 2019a). Genome-wide evaluation of emetine effects on alternative splicing showed substantial alternative splicing changes after treatment, with nearly 60% being exon skipping events.

Figure 1.4: G-quartet and G-quadruplex structure. A. Hoogsteen bonding between four guanines results in a planar G-quartet formation, which is stabilized by metal cations (M+) such as potassium cations. B. G-quadruplex structure formation by stacking of three G-quartets with intervening single-stranded loops. C.

Consecutive G-tracts are separated by 1-7 bp of intervening sequence (loops). Adapted from (Capra et al., 2010).

1.5.2 R-loop formation

During transcription, dynamic hybrid structures between DNA and nascent RNA transcripts can be formed (Crossley et al., 2019). These RNA-DNA hybrid structures are collectively known as R-loops and can be favoured depending on the structural DNA context. Formation and/or stabilization of R loops is particularly favourable when the non-template strand is G-rich, but it can also be promoted by DNA supercoiling, the presence of DNA nicks, and the formation of G-quartets (Duquette et al., 2004; Santos-Pereira and Aguilera, 2015). The continuous activity of DNA/RNA helicases and ribonucleases H (RNAse H1 and H2) release R-loop structures (Santos-Pereira and Aguilera, 2015). Interestingly, R-loops and G4s were both found to be unwound by a common helicase in humans (DHX9) (Chakraborty and Grosse, 2011). This helicase activity is important to avoid single-stranded DNA damage and to preserve genomic stability.

Similarly to G4s, R-loop detection is enriched at promoters, where their formation has been shown to have a kinetic effect on transcription, leading to RNA pol II pausing (Chen et al., 2017). The impact of R-loop formation, as well as the formation of G4s and other non-canonical nucleic acid structures, impacts transcript elongation rates and can have a kinetic repercussion on co-transcriptional events involved in RNA processing, such as alternative splicing (Dujardin et al., 2013; Nieto Moreno et al., 2015). Moreover, the formation of R-loops and other non-B DNA structures can originate due to mis-splicing events. For example, mutations of alternative splicing factors can lead to R-loop accumulation, which may have strong implications for genomic stability and be relevant in the context of cancer pathogenesis (Li and Manley, 2005; Nguyen et al., 2019).

1.6 Deciphering the non-canonical splicing code and its implications in tissue-specific splicing

1.6.1 Transcriptomic revolution

The revolutionary development of sequencing technologies has enabled deep transcriptome exploration, providing a precise landscape of gene expression patterns across tissues, cell types and organism populations. The first sequencing technologies were largely based on experimental procedures initially developed by Frederick Sanger. Further improvements of these sequencing technologies allowed for systematic sequencing of cDNA libraries to generate expressed sequence tags (EST) or full-length cDNA, largely driven by different international consortia (Okazaki et al., 2002; Strausberg et al., 2002).

The public availability of ESTs and full-length mRNA sequences allowed for initial cataloguing of alternative splicing events. Despite the fact that microarrays enabled the first genome-wide assessments of gene expression and alternative splicing, they were only able to quantify genes or alternate splicing events that were previously known. It was the development of next-generation sequencing technologies (NGS) that allowed the discovery and quantification of transcripts to be performed in a single experiment. The main improvement of NGS technologies over the classic Sanger sequencing methods was the robust generation of cell-free sequencing libraries that enabled a massive parallel sequence of short DNA fragments (Shendure and Ji, 2008). While the sequencing of genomic DNA enabled the characterization of entire genomes, the massive parallel sequencing of cDNA libraries (RNA-seq) revolutionized the way to assess gene expression and alternative splicing.

However, in order to enable the accurate and systematic evaluation of alternative splicing events using RNA-seq data, diverse data analysis methodologies were developed including read-mapping, splice junction discovery and quantitative assessments of gene expression and alternative splicing. After more than a decade since RNA-seq was developed, alternative splicing analytical methods are still being

advanced, and the detection and quantification of non-canonical splicing events still represent a major challenge as they are often excluded from standard RNA-seq analyses (Sibley et al., 2016; Stark et al., 2019).

1.6.2 Alternative splicing tissue-specific code

Transcriptome profiling of multiple vertebrate tissues using RNA-seq has expanded our genome-wide understanding of tissue-specific alternative splicing events (Barash et al., 2010; Barbosa-Morais et al., 2012). The quantitative assessment of 3,665 cassette exon inclusion events across 27 murine tissues made it possible to build a predictive model to identify cis-regulatory elements, providing a first glance of the so-called “splicing code” (Barash et al., 2010). These studies demonstrated that the sequence contained within flanking intronic regions was enough to build a strong predictive model of tissue-specific alternative splicing and has inspired the development of different machine learning approaches to study tissue-specific alternative splicing (Barash et al., 2010; Leung et al., 2014; Zhang et al., 2019b). Moreover, the use of splicing code models has unveiled a catalogue of disease-causing variants, suggesting an important role of these cis-regulatory elements regarding the homostatic equilibrium of cellular identity and function (Xiong et al., 2015).

1.6.2.1 Canonical and non-canonical neuronal splicing code

Among major vertebrate tissues, neuronal tissues have the most distinctive alternative splicing patterns, with the biggest set of tissue-specific cassette exons (GTEx Consortium, 2015; Melé et al., 2015; Tapial et al., 2017; Yeo et al., 2004a). Most of the neuronal alternative splicing events are established during neuronal differentiation, where dramatic alternative splicing changes can be observed (Su et al., 2018; Vuong et al., 2016).

1.6.2.1.1 Sequence motif code

Neuronal alternative exons are characterized by having weak splice sites (Fig 1.5a), which means that additional regulatory factors can have a large influence on their inclusion (Coelho and Smith, 2014). During embryonic development, RBPs have a

combinatorial effect over neuronal splicing (Vuong et al., 2016). Dynamic changes on RBP gene expression generate a different molecular context for alternative splicing, which leads to a dynamic and conserved network of alternative splicing events during vertebrate brain development (Barbosa-Morais et al., 2012; Irimia et al., 2014; Torres-Méndez et al., 2019; Vuong et al., 2016; Weyn-Vanhentenryck et al., 2018). Immediate intronic flanking regions of neuronal cassette exons have a high concentration of cis-regulatory that are binding sites (Fig 1.5a). For example, downstream intronic regions of neuronal cassette exons often contain binding sites of neuro-oncological ventral antigen 2 (NOVA), serine/arginine repetitive matrix protein 4 (SRRM4), RNA-binding protein fox proteins (RBFOX), while both upstream and downstream intronic flanking regions can contain motifs for polypyrimidine tract binding (PTB) binding. The combinatorial effect of PTB1 binding (which repress neuronal exon definition in non-neuronal tissues) and the binding of NOVA, SRRM4, RBFOX1 (that promote neuronal exon inclusion in neurons) enables a neuron-specific selection of exons.

1.6.2.1.2 Architectural code

However, primary sequence motifs are not the only important feature in the determination of neuronal splicing. Splicing code analyses suggest that exon-intron architectural features are also key determinants of neuronal alternative splicing (Fig 1.5a). Cassette exons that are alternatively included in neurons tend to be short and symmetrical (non-frameshifting) (Barash et al., 2010; Coelho and Smith, 2014). This observation was strongly supported by previous well-studied neuronal alternative splicing events that involve microexons. For example, SRC is a non-receptor tyrosine kinase that is expressed across vertebrate brains and its activity during development is critically regulated by the inclusion of a microexon that encodes between 5-6 aa (conserved 6 aa sequence across chicken, rodents and humans and 5 aa long in some amphibians such as *Xenopus laevis*) (Collett and Steele, 1992; Levy et al., 1987; Martinez et al., 1987). Even though several cis-regulatory sequences have been found to promote its neuronal splicing pattern (Fig 1.5b), early experimental manipulation of N1 *SRC* exon have demonstrated that length extension

results in the abolition of its neuronal pattern, meaning that exon size itself can be a part of the neuronal splicing code (Black, 1991).

Deeper analyses of microexons have shown that SRRM4, RPBOX1 and PTB1 contribute to the selective inclusion of microexons in the brain (Gonatopoulos-Pournatzis and Blencowe, 2020; Irimia et al., 2014; Li et al., 2015). Even though these RBPs regulate a major fraction of neuronal alternative splicing, microexons are the most dynamic and conserved sub-group of neuronal exons. Microexon alternative splicing patterns are highly conserved across vertebrates and their differential inclusion is predicted to have different protein-regulatory properties. Microexons residues overlap significantly more with surface protein domains and are enriched in charged residues, suggesting that microexon inclusion could regulate protein interactions (Irimia et al., 2014). Recent mutational analysis implementing CRISPR-Cas9 screenings have enabled a genome-wide interrogation of splicing networks that are involved in microexon splicing (Gonatopoulos-Pournatzis et al., 2018). The CRISPR-Cas9 screening results and additional siRNA lead Gonatopoulos-Pournatzis and colleagues to identify intronic splicing enhancers at upstream intronic microexon regions bound by SRMM4 and two novel microexon co-activators. Together these factors may contribute to overcoming the steric hindrance issues related to microexon definition and contribute to the neuronal-specific alternative splicing patterns observed for microexons. Moreover, the neuronal code of microexons corresponds to the most conserved network of alternative splicing currently described, some being conserved since at least 600 million years of evolutionary time (Irimia et al., 2014; Torres-Méndez et al., 2019).

1.6.2.1.3 The RNA structural code

Another feature associated with the neuronal splicing code is the formation of RNA secondary structures (Barash et al., 2010; Coelho and Smith, 2014). Even though this RNA structural code has been less explored, it is known that the effects of cis-regulatory elements can be modulated by the presence of RNA structures in nascent transcripts. One particular example where secondary structures play a role in neuronal splicing definition can be found at RBFOX regulated exons, where the majority RBFOX binding sites are located within distal intronic regions and

secondary structures play a key role to enable their regulatory effect over exon definition (Lovci et al., 2013). Lovci and colleagues explored the role of RNA secondary structures over RBFOX mediated splicing regulation and they found that long-range RNA-RNA base-pairing interactions form RNA bridges that are necessary for the regulatory effects of distal RBOX binding sites (Lovci et al., 2013).

In addition, a non-canonical splicing mechanism called back-splicing is favoured by the presence of complementary intronic sequences that can form secondary RNA structures. During back-splicing, the second nucleophilic attack is performed over an upstream 3′ splice leading to circular RNA products, which are particularly abundant in the brain. Moreover, circRNA production is upregulated during neuronal differentiation, and a subset gets highly enriched in synaptic compartments (Rybak-Wolf et al., 2015; You et al., 2015). RNA structures that favour back-splicing are often derived from complementary intronic sequences associated with ALU elements (Jeck et al., 2013).

All of this suggests that RNA structures can play an important role in the definition of canonical and non-canonical splicing. However, the contribution of non-canonical DNA and RNA structures over neuronal splicing remains almost completely unexplored.

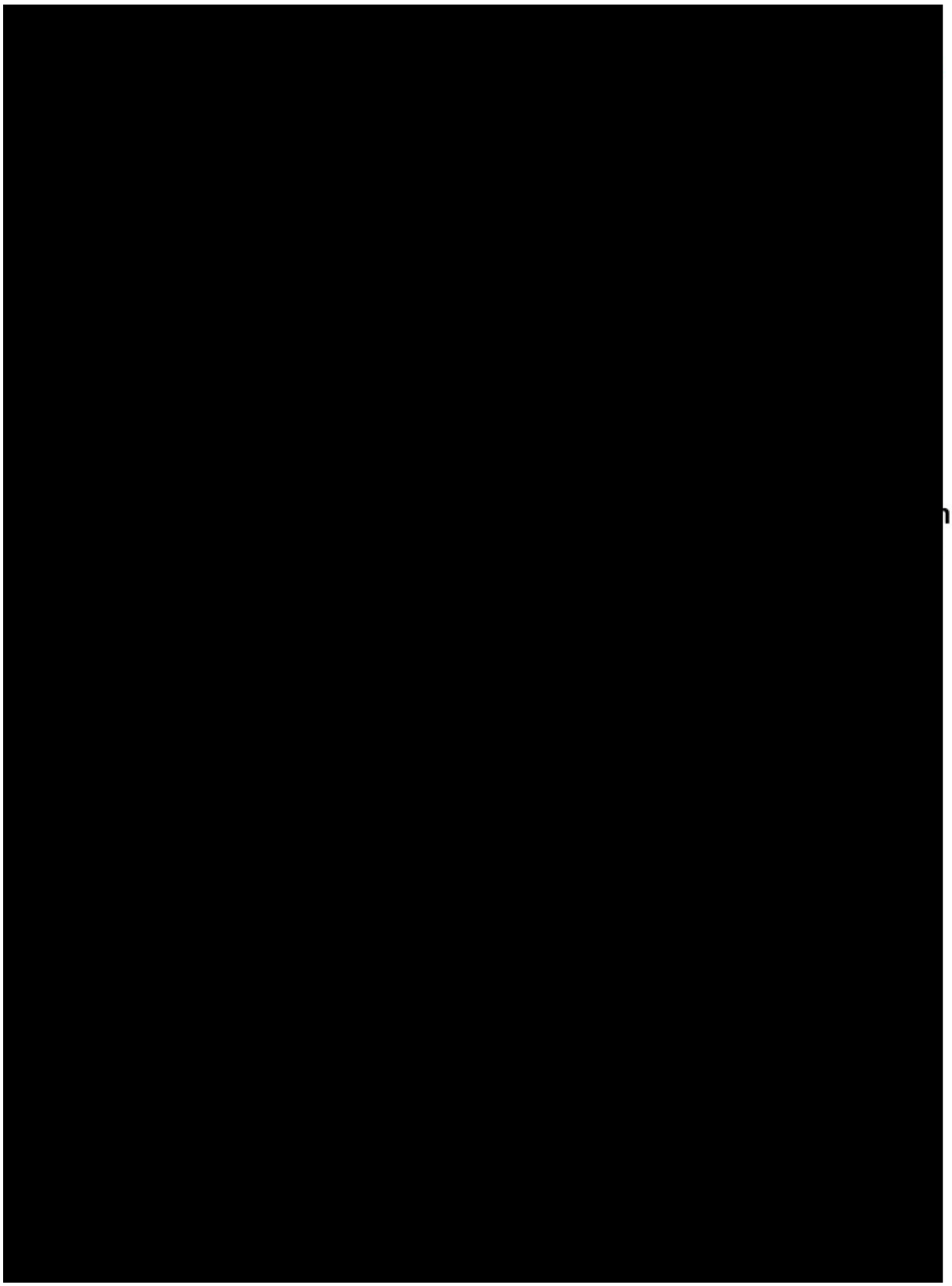

Figure 1.5: Neuronal splicing code. A. Schematic summary of the splicing code results obtained by Barash and colleagues (Barash et al., 2010). Features associated with neuronal cassette exon skipping (top) and exclusion (bottom) are shown. Different vertical columns coloured in light blue, orange and green enclose sequence features that were significantly found to be associated with cassette exons that are differentially included in the central nervous system (CNS). The colour of the

letters indicate enrichment (red) or depletion (blue), while the font size corresponds to the respective level of enrichment or depletion. Black edges connecting the different features indicate co-association, where its thickness indicates different levels of co-association significance. B. Extensive experimental data identify different cis-regulatory sequences that control N1-*SRC* microexon splicing. In non-neuronal cells, N1 exon definition is disfavored by PTB binding at both flanking intronic regions. While in neurons, PTB expression is replaced by a nPTB paralog, which together with other RBPs (shown at bottom) promote exon definition. Newly identified cis-regulatory elements and protein factors that regulate N1 and other neuronal microexons are coloured in grey and displayed with dashed lines. This figure was adapted from Coelho and Smith and updated with some new protein factors that were suggested by Gonatopoulos-Pournatzis and collaborators (Coelho and Smith, 2014; Gonatopoulos-Pournatzis et al., 2018).

1.6.3 Non-canonical splicing detection and quantification using RNA-seq data

The first transcriptome-wide alternative splicing analyses were based on public ESTs. Even though sequenced EST segments are strongly biased to the 3′ end of transcripts, these analyses demonstrated the benefits of transcriptome sequencing. The initial analyses of EST and cDNA sequences not only enabled genome-wide characterization of alternative splicing events but also uncover certain aspects that did not fit into the standard model of vertebrate splicing, such as the presence of non-canonical splice sites and microexons (Burset, 2000; Volfovsky et al., 2003).

The development of RNA-seq sequencing technologies enabled a deep exploration and annotation of the transcriptome across model organisms and other species. However, the aim of annotating splice junctions using RNA-seq data challenged the bioinformatic alignment algorithms, because widely used RNA-seq platforms generate shorter reads than ESTs and other Sanger based sequence technologies. This pushed the bioinformatic field to develop novel algorithms and strategies to perform spliced alignments (Engström et al., 2013). To perform efficient splice junction detection, different assumptions are made by spliced aligners, which involves splice site sequences, exon/intron sizes and splice site usage. For instance, Tophat (Trapnell et al., 2009) initially only detected GT-AG splice junctions to reduce the probability of spurious splice site detection, enabling the detection of the great majority of splice junctions, but completely ignoring some U12-type splice sites and

other non-canonical splice sites. Thus, the progressive expansion of our canonical splicing model has had a direct impact on the way RNA-seq analyses are performed to study splicing, but at the same time, the exploration of non-canonical splicing events represent a constant source of bioinformatics challenges.

The development of “seed and extension” algorithms such as GSNAP (Wu and Nacu, 2010) or MapSplice (Wang et al., 2010b) enabled the genome-wide detection of non-canonical splice sites. The fundamental heuristic principle of these strategies is to map short read sub-segments, called alignment seeds, to the genome and then extend the resultant alignment by dynamic programming algorithms (i.e. Smith-Waterman). However the seed size requirements, to perform significant genome seed mapping, limited the ability to discover novel microexons. As most bioinformatics tools are designed to detect and quantify canonical splicing, efficient detection and quantification of alternative splicing events required further development of bioinformatics and experimental approaches.

To perform efficient splice junction detection, different assumptions can be made about how splicing normally takes place. For instance, Tophat (Trapnell et al., 2009) initially only detected GT-AG splice junctions to reduce the probability of spurious splice site detection, enabling the detection of the great majority of splice junctions, but completely ignoring some U12-type splice sites and other non-canonical splice sites. Thus, the progressive expansion of our canonical splicing model has had a direct impact on the way RNA-seq analyses are performed to study splicing, but at the same time, the exploration of non-canonical splicing events represents a constant source of bioinformatics challenges.

1.6.3.1 Identification of neuronal non-canonical splicing events using RNA-seq data

As discussed above (see section 1.6.2.1) diverse non-canonical splicing events are strongly associated with the neuronal splicing code. However, their detection and quantification have required the development of novel bioinformatics approaches. For instance, the detection of circRNAs, recursive splicing and microexons have required the development of new strategies for RNA-seq alignment. Despite their importance to the understanding of neuronal transcriptomics dynamics, the detection

and quantification of these alternative splicing events are mostly excluded from standard RNA-seq analyses.

1.6.3.1.1 Recursive splicing detection

Neuronal transcripts tend to have longer introns than transcripts coming from other tissues (Sibley et al., 2015; Thakurela et al., 2013). These long intronic sequences favour exon definition over intron definition, where spliceosomal particles are first assembled cross-exon to promote the formation of the pre-initiation complex (Ast, 2004; Hollander et al., 2016). However, in order to complete the splicing process, splice sites need to get close so that the second transesterification reaction can take place. Therefore different mechanisms have been proposed to promote the splicing of very long introns. Recursive splicing has the potential to break down the processing of long introns into smaller intron splicing steps, and therefore the fine mapping of RS-sites has been of great interest to understand the dynamics of the neuronal transcriptome (Blazquez et al., 2018; Dye et al., 2006; Hollander et al., 2016; Pai et al., 2018; Sibley et al., 2015).

The first strategies that were able to achieve a comprehensive mapping of RS-sites greatly relied on the quantification of intronic reads. The observation of “saw-tooth” coverage patterns spanning intronic regions in total RNA-seq samples led to the discovery of RS sites near low coverage valleys (Fig 1.6a). Further systematic detection of intronic “saw-tooth” coverage patterns across introns led researchers to find 197 RS sites in *D. melanogaster* and 11 in humans (Duff et al., 2015; Sibley et al., 2015) Novel computational methods applied to nascent RNA-seq data (which provides a larger fraction of intronic reads than regular RNA-seq) have enabled the expansion of the catalogue of recursively spliced introns by 4-fold in *D. melanogaster (Pai et al., 2018)*.

1.6.3.1.2 Identification of circRNAs

Even though the initial detection of circRNAs was reported a long time ago, they have been considered to be mainly produced by mis-splicing events (Cocquerelle et al., 1993). Given that in most analyses only the junction in one transcriptional direction is detected, backsplicing events are normally ignored. Development of

bioinformatics methods to accurately detect and quantify back-splicing events have been critical to have a transcriptome-wide perspective of the biogenesis and possible roles of circRNAs.

The detection of circRNAs using RNA-seq analyses relies on the detection of back splice junctions, which unlike linear splice junctions, connect a downstream 5′ splice site, with an upstream 3′ splice site (Fig 1.6b). A wide range of methods has been developed to enable the systematic detection of circRNAs using RNA-seq (Cooper et al., 2018; Zeng et al., 2017). These bioinformatic developments have enabled circRNAs to be profiled across different tissues, where a clear enrichment in neuronal tissues have been observed, and also the detection possible cis-regulatory RNA structures that may have a role in their circRNA biogenesis through fine-tuning control of back splicing.

1.6.3.1.3 Discovery and quantification of microexons

During RNA-seq analysis, reads are normally mapped to a reference genome using splice-aware alignment software, such as STAR or HISAT2. During this process, splice junctions can be detected when a read map spans two alignment blocks, separated by a long gap, which normally corresponds to intronic sequences. However, in the case of microexons, reads tend to span the whole microexon so aligners have to find three alignment blocks to successfully map exon-microexon-exon junctions (EEEJ). Most of the conventional RNA-seq aligners cannot efficiently do this while the reads are being mapped to the genome (Li et al., 2015; Wu et al., 2013).

The efficient discovery and quantification of microexons have required the development of specialized multi-step computational workflows (Irimia et al., 2014; Li et al., 2015). These methods use the annotated exon-exon junction sequences to guide the discovery of microexons. These tools enable the discovery of internal microexon sequences based on reads that are frequently misaligned by conventional mapping algorithms (Fig 1.6c, more details about these methods will be given at section 2.1.1). The development of these computational workflows to detect microexons in RNA-seq data enabled the genome-wide detection of microexons

across different vertebrates and the identification of a neuronal microexon subgroup that have strong brain-specific patterns (Irimia et al., 2014; Li et al., 2015). Moreover, these neuronal microexons were shown to be dysregulated in brains of individuals with autism spectrum disorders, suggesting the existence of a coordinated microexon splicing program whose dysregulation may lead to psychiatric diseases (Irimia et al., 2014).

Figure 1.6: Development of novel bioinformatic methods have enabled the detection of different non-canonical splicing events. **A.** Sibley and collaborators developed a novel bioinformatic approach for RS site detection across vertebrate introns (Sibley et al., 2016). Since recursive splicing is a multi-step intron removal process, introns containing RS tend to be associated with saw-tooth patterns in the intronic read density. Then through sequence analysis, they detect RS sites that in some cases are associated with novel splice junctions. **B.** Splice-aware RNA-seq

aligners are able to identify reads that come from exon-exon junctions, by detecting two consecutive blocks of alignments. However, common alignment algorithms are unable to correctly align spliced reads that come from exon-exon junctions originated through back-splicing. Instead, several bioinformatics tools have been developed to detect and quantify backsplicing (Zeng et al., 2017). These tools can map reads coming from back exon-exon junctions, which otherwise would be unmapped or partially mapped (soft-clipped) by the conventional RNA-seq aligners. **C**. Conventional RNA-seq aligners often fail to map reads that span microexons. The development of computational methods enabled the discovery of internal microexon using RNA-seq data. To identify novel microexons, these methods try to find reads with an unmapped section that can be reallocated inside the intronic sequences (Irimia et al., 2014; Li et al., 2015). With these novel approaches of RNA-seq analysis, cis-regulatory elements were found to be associated with microexon inclusion, including binding sites of some RBPs such as SRMM4. Irimia and collaborators showed that SRMM4 is downregulated in brain samples taken from ASD patients. This figure was adapted from (Sibley et al., 2016).

1.6.3.2 Genome-wide evaluation of non-canonical RNA-structures effects over alternative splicing

The first genome-wide evaluation of non-canonical alternative splicing events was carried out in human and mouse transcriptomes, for which the authors showed correlations between alternative splicing and the bioinformatic prediction of non-B DNA structures (Tsai et al., 2014). By implementing a logistic regression model, Tsai and colleagues found a significant correlation of alternative splicing events with the presence of different predicted non-B DNA motifs. They also found that among other non-B DNA motifs evaluated, the presence of G4 motifs was highly correlated with exon skipping events. While these analyses were only based on *in-silico* prediction of non-B DNA motifs, some experimental approaches have been developed for the experimental detection of some of these structures.

1.6.3.2.1 Genome-wide detection of G4 formation and its impact over alternative splicing modulation

Initial low-throughput detection of G4 formation was based on biophysical and biochemical methods, which provided evidence to support the *in vitro* and *in vivo* formation of G4 at the DNA level (Kwok and Merrick, 2017; Lam et al., 2013). Only through recent developments of novel sequencing-based approaches, a genome-wide evaluation of G4 formation was possible (Chambers et al., 2015; Hänsel-Hertsch et al., 2016, 2017; Kwok and Merrick, 2017; Kwok et al., 2016;

Marsico et al., 2019a). First genome-wide detection was based on chromatin immunoprecipitation (ChIP) of G-quadruplexes, using antibodies that were able to recognize G4 structures in DNA. These experiments were able to detect between 700-1000 G4s that were formed from G4 motifs (Hänsel-Hertsch et al., 2016; Lam et al., 2013). Additional assays were based on the assumption that G4 formation leads to polymerase stalling, which can be detected through different sequencing-based strategies. These set of experimental methods are called G4-seq and have provided a comprehensive experimental identification of DNA sequences that form G-quadruplexes *in vitro* (Chambers et al., 2015; Kwok and Merrick, 2017; Marsico et al., 2019a). Moreover, G4-seq technologies could also be adapted to detect sequences motifs that lead to G4 formation at the RNA level (rG4-seq) (Kwok et al., 2016).

Even though the *in-vivo* formation of RNA G4s is still a matter of debate (Biffi et al., 2014; Guo and Bartel, 2016), recent studies suggest that RNA G4 formation can modulate *in vitro* RBP binding to mRNA molecules (Benhalevy et al., 2017). However, since many proteins have affinities for G-rich sequences, such as G4 motifs, it is still unclear whether RBP binding is driven by linear G-rich sequences or G4 formation (Fay et al., 2017). Huang and collaborators showed that ribonucleoprotein F (hnRNPF) binding sites are enriched in G4 motifs and mutations that destroy G4 forming capacity while maintaining G-content, can abrogate exon inclusion, by interfering with hnRNPF binding (Huang et al., 2017) (mentioned in section 1.5.1). Previous experimental evidence suggested that G4 formation and hnRNP F/H binding are mutually exclusive events (Dominguez et al., 2010; Samatanga et al., 2013). Thus, the effects of G4 formation on RNA-binding proteins is currently not well understood, but effects of G4 formation on alternative splicing have been repeatedly suggested by different research groups (Gomez et al., 2004; Hastings and Krainer, 2001; Huang et al., 2017; Marcel et al., 2011; Tsai et al., 2014; Weldon et al., 2018; Zhang et al., 2019a).

1.7 Research aims

In this thesis, I report on computational analyses to study two populations of alternative exons defined by their non-canonical splicing features: microexons and G4-flanked exons. For this purpose I pursued the following objectives:

1. Develop, MicroExonator, a novel computational workflow designed to improve the detection and quantification of microexons using RNA-seq data.
    a. Implement the different computational steps in a unified, user-friendly pipeline using state of the art computational strategies to ensure reproducibility and scalability of the analyses.
    b. Perform simulation-based approaches to benchmark against other computational methods used for microexon discovery.
    c. Enable integration of MicroExonator results with downstream alternative splicing analysis.
    d. Use MicroExonator to study microexon alternative splicing events across mouse development and neuronal subcellular types.
2. Characterize different non-canonical DNA and RNA sequence structures associated with alternative splicing.
    a. Calculate the enrichment of different non-B DNA motifs across human splice sites.
    b. Analyse G4-seq data to evaluate *in vitro* G4 formation across splice sites of different species.
    c. Perform a detailed characterization of G4 enrichment at splice sites to evaluate their association with:
        i. Splice site strength
        ii. Template and non-template strands
        iii. Intron/exon structures
3. Evaluate the association of microexons and G4-flanked exons with dynamic alternative splicing changes induced by neuronal depolarization.