Computational analyses of non-canonical architectural and structural features associated with alternative splicing

UNIVERSITY OF CAMBRIDGE

Submitted for the degree of Doctor of Philosophy

by

Guillermo Eduardo Parada González

University of Cambridge

Wellcome Sanger Institute

Homerton College

September 2019

Preface

The dissertation is submitted for the degree of Doctor of Philosophy.

I declare this is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated in the text. This document, in whole or in parts, has not been submitted for any other degree or diploma.

It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Guillermo Eduardo Parada González
Cambridge, UK.
July 2020

Computational analyses of non-canonical architectural and structural features associated with alternative splicing

Guillermo Eduardo Parada González

Summary

Splicing of nuclear introns is catalysed by the spliceosome, one of the most complex macromolecular machines currently known. Even though the canonical splicing signals that drive the precise recognition of splice sites are well-characterised, recent advances in transcriptome profiling technologies and computational method development have enabled widespread identification of non-canonical splicing features. Non-canonical splicing is highly associated with dynamic splicing regulation, and occurs most prevalently in neuronal tissues. In this present work, I have investigated two types of non-canonical features that are related to atypical exon-intron structures and DNA/RNA conformations.

First, I studied a group of extremely small exons, known as microexons (≤30 nucleotides), which were shown to be part of an evolutionarily conserved network of neuronal alternative splicing events that play essential roles in neuronal development. Since standard RNA-seq tools cannot efficiently detect microexon splice sites, I developed MicroExonator, a novel pipeline for reproducible de novo discovery and quantification of microexons. As a proof of principle, I analysed microexon alternative inclusion patterns across 289 RNA-seq samples coming from eighteen different tissues across a wide range of mouse embryonic and adult stages. I detected 2,938 microexons, 343 of which are differentially spliced throughout mouse embryonic development, including 35 that are not present in mouse transcript annotation databases. Unsupervised clustering of microexons alone segregates brain tissues by developmental time and further analysis suggest a key function for microexon inclusion in axon growth and synapse formation. Moreover, I developed a module to adapt MicroExonator splicing analysis to single-cell RNA-seq samples that I used to analyse data from the mouse visual cortex. As a result, I found 39 microexons that are differentially included between glutamatergic and gabaergic neurons, fifteen of which are found in genes that encode synaptic proteins.

The second type of non-canonical features that I studied are sequences associated with non-B DNA structures and possibly atypical RNA conformations. I analysed the enrichment of different non-B DNA motifs across splice site sequences. The strongest and most consistent enrichments were found for G-quadruplex motifs, which are enriched ~3-fold both upstream and downstream of splice junctions. Further analysis of G4-seq experiments corroborated the enriched motifs detected at splice sites leads to *in-vitro* G-quadruplex formation. Moreover, enrichment analyses of G-quadruplex motifs and G4-seq experiments across multiple species suggest that the association of G-quadruples to splice sites is a property restricted to mammals and birds. Interestingly, I found stronger enrichment of G-quadruplexes associated with weak splice sites, suggesting that they could function as cis-regulatory elements of alternative splicing events.

Finally, to explore if microexons and exons flanked by intronic G-quadruplexes were involved in dynamic splicing changes, I analyse alternative splicing events induced by depolarisation treatments in human and mouse neurons. I found a widespread cassette exon skipping response after neuronal depolarization, which was particularly enriched in microexons and exons flanked by G-quadruplexes motifs. Taken together, these results suggest that non-canonical splicing features are an important regulatory mechanism of alternative splicing. Further characterisation of non-canonical splicing might provide a better understanding of fine-tuned alternative splicing mechanisms, in particular in the context of neuronal development and heterogeneity.

This thesis is dedicated to my parents, both Mabel and Guillermo, who have devoted a big part of their life to raise me as the man who I am today. They not only provided me unconditional love and patience, but also the opportunities that enabled me to find my passion in life. I would like to thank my grandmother Mercedes for sparking my scientific curiosity at a very young age by buying my fun scientific books when I was a child. Finally to my grandfather Guillermo, now resting in peace, the first academic of our family, with whom I had very passionate discussions about science and life that I will never forget.

Acknowledgements

I want to thank Dr Martin Hemberg and Prof Eric Miska for supervising me during my PhD. I much appreciate the support they have provided me during this period, which enabled me to take full advantage of my privileged position as a PhD student of this prestigious institution. Members from both Hemberg and Miska lab contributed significantly towards my initial learning and development. Particularly, I would like to thank Dr Tallulah Andrews, Dr Vladimir Kiselev and Dr Tomás Di Domenico. I also want to sincerely thank Dr Ilias Georgakopoulos-Soares, the first graduated PhD from Hemberg lab, for all of his support, collaboration and friendship. I also want to acknowledge Dr Sarah Teichmann and prof Chris Smith for their critical feedback of my doctoral research, as part of my thesis committee, and also to Prof Chris Ponting and Dr Jen Harrow for accepting to read and evaluate this thesis.

I am really grateful to my parents, Mabel and Guillermo, for their unconditional love and support, and to Isabel, my beloved girlfriend, for her much needed support during the toughest moments of PhD. I also thank all other family members and friends from Chile, particularly the CDP group, Joaquin, Jacqueline and Maria Jose, who were always there for me. I would also like to thank my dear friends from Cambridge, particularly Dr Jenkinks, Dr Fryer, Dr Singh, as well as future doctors Eijsbouts, De Jonghe and Kosałka. My experience as a PhD student would not have been the same without them.

Finally, I would like to acknowledge other people who played an important role to motivate me in my early years to pursue a career as a researcher. I thank Dr Eduardo Ravanal, who was my high school biology teacher, for teaching me the basic concepts of molecular biology and pushing me towards academia. I also thank Prof Katia Gysling for all her support during my undergraduate training, which enabled me to successfully get a position as a PhD student of this university. Finally, I would like to thank Dr Roberto Munita for the training he provided during my undergraduate research, which sparked my interest in computational biology.

List of figures

List of abbreviations

AG	Adrenal gland
CaMK	Calmodulin-dependent protein kinase
CaRRE	CAMK IV-responsive RNA element
cDNA	Complementary DNA
CE	Core exon
CRISPR	Clustered regularly interspaced short palindromic repeats
CSG	Contiguous splice graph
DNA	Deoxyribonucleic acid
DPC	Days post conception
EEEJ	Exon-microexon-exon junctions
EJC	Exon junction complex
EPI	Epiblast stem cell
EST	Expressed sequence tag
G4	G-quadruplex
GO	Gene ontology
GTF	Gene transfer format
hnRNP	Heterogeneous nuclear ribonucleoprotein
mESC	Mouse embryonic stem cell
mRNA	Messenger RNA
NMD	Nonsense-mediated decay
NMDA	N-Methyl-D-aspartic acid
PCR	Polymerase chain reaction
PDS	Pyridostatin
PPCA	Probabilistic principal component analysis
PPI	Protein-protein interaction network
PSI	Percent spliced in
PTC	Premature stop codon
RBP	RNA-binding protein
RNA	Ribonucleic acid
RT	Reverse transcriptase
RUST	Regulated unproductive splicing and translation
SKM	Skeletal muscle
snRNP	Small nuclear ribonucleoprotein
ss	Splice site
UTR	Untranslated region

Index

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.

[redacted]

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).

[redacted]

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).

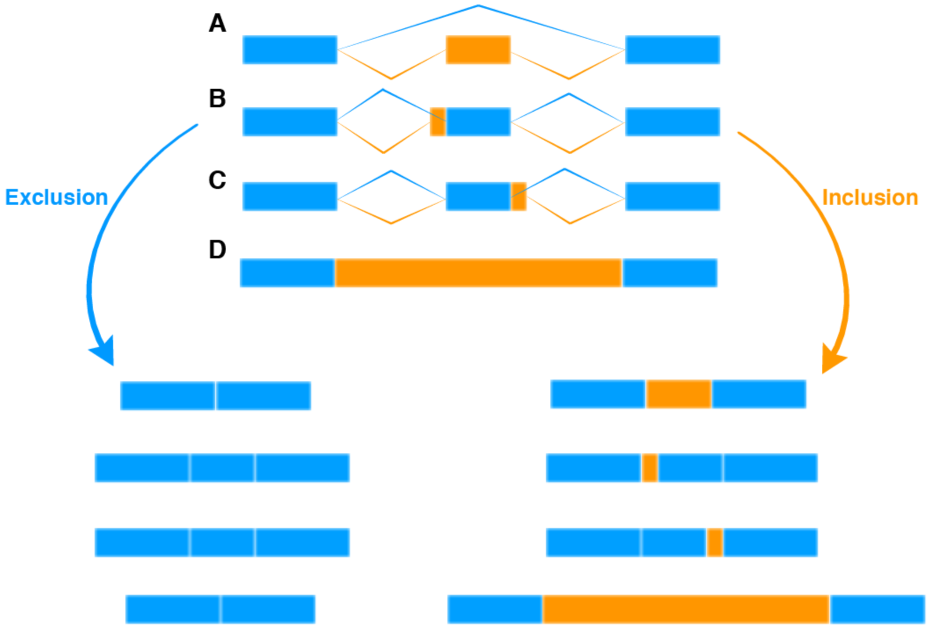

Figure 1.3: Alternative splicing event types. There are four basic types of alternative splicing types. A. Cassette exon inclusion/skipping. B. Alternative 5′ splice site selection. C. Alternative 3′ splice site selection. D. Intron retention. Alternatively included mRNA segments are coloured in orange.

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.

[redacted]

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.

[redacted]

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).

[redacted]

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.

2 Chapter II: Reproducible RNA-seq processing for detection and quantification of microexons

Collaboration note

Most of the work presented in this chapter are results that will be published as a separate manuscript in a peer reviewed journal. While I conceived the core idea of the initial computational analyses with Roberto Munita[1], the development of the software was exclusively performed by me. Close communication and interaction with Ilias Georgakopoulos-Soares[2] was beneficial for this project's development, who also tested the software in collaboration with Veronika Kedlian[3].

2.1 Introduction

The initial report of microexons dates back in 1985 for the *Ubx* gene which in Drosophila was found to contain two 5 nt microexons (Beachy et al., 1985). This discovery was followed by several other reports of constitutive and alternative microexons discovered in various vertebrate genes (Cooper and Ordahl, 1985; Santoni et al., 1989; Small et al., 1988; Ustianenko et al., 2017). Even though some of these microexons were found to be tissue-specific or regulated through brain development (Santoni et al., 1989; Small et al., 1988), systematic analyses of

[1] Former Ph.D. student at Department of Cellular and Molecular Biology, Pontificia Universidad Católica de Chile and current postdoctoral fellow at the Division of Molecular Hematology (DMH), Lund University.

[2] Former Ph.D. student at the Sanger Institute, co-supervised by Serena Nik-Zainal and Martin Hemberg. Current postdoctoral fellow at Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco.

[3] Current Ph.D. student at the Sanger Institute, supervised by Sarah Teichmann.

microexons were obstructed by technical difficulties associated with their detection in mRNA sequences.

Initial gene annotation of model organisms was extensively carried out by mapping of expressed sequence tags (EST) and other cloned cDNA sequences (Dias Neto et al., 2000; Okazaki et al., 2002). However, the correct alignment of these cDNA sequences was acknowledged to be particularly challenging in the presence of microexons (Florea et al., 1998). The development of an algorithm to correct cDNA alignments, allowed for the detection of 224 previously unknown microexons across human, *Caenorhabditis elegans* and *Drosophila melanogaster* (Volfovsky et al., 2003). Further development of this strategy was directly implemented by GMAP, an EST/cDNA alignment tool, which also incorporated a statistical model to avoid reporting spurious microexons (Wu and Watanabe, 2005).

2.1.1 Computational methods for discovery and quantification of microexons using RNA-seq data

The advent of high throughput RNA sequencing technologies (RNA-seq) provided an unprecedented opportunity to explore the transcriptome. However, widely used RNA-seq platforms, such as Illumina, generate RNA sequencing reads that are shorter (50-150 nt) than the average EST length. Two main strategies have been developed for short RNA-seq read mapping; (1) Exon-first approach, in which reads are first mapped through ungapped alignment, enabling the mapping of reads within exonic regions. Subsequently, only unmapped reads undergo a second round of spliced read mapping. (2) Seed-extend approach, in which the read alignment process is subdivided into units of ungapped alignments, often referred as alignments seeds, and only seeds successfully mapped to the genome are extended (Garber et al., 2011).

However, the alignment of reads that span microexons has been identified as a particularly hard problem, which can prevent the correct alignment of reads unless the aligners have strategies implemented to align to microexons reads (Wu and Watanabe, 2005). Among the RNA-seq aligners that have proven to be more sensitive to microexon detection, there is Olego (Wu et al., 2013), which combines exon-first and seed-extend approaches to perform RNA-seq alignments. In a first step, exonic reads are mapped using an approach similar to BWA (Li and Durbin, 2009), and during the second step, unmapped reads are split into alignment seeds to discover splice junctions through the seed-extend approach. The feature that makes Olego particularly sensitive to detect microexons is the use of small seeds during this last step to find splice junctions, which enabled Wu and collaborators to identify 1,665 microexons in mouse retina RNA-seq samples, 37.8% of which were not annotated, suggesting great discovery potential of RNA-seq analyses.

Systematic discovery and quantitative analyses of microexons using RNA-seq data have been performed by the implementation of pipelines that integrate multiple alignment steps. VAST-TOOLS (Irimia et al., 2014; Tapial et al., 2017) is a multi-module analysis pipeline that can quantify alternative splicing events measured as the “percent spliced-in” (PSI), which corresponds to the percent of the transcript that undergoes a particular splicing event (e.g. cassette exon inclusion). Irimia and colleagues developed a module to discover microexons using RNA-seq data which was based on bowtie alignments to an extensive library of possible exon-microexon-exon junctions (EEEJ) and then many of the discovered microexons were deposited in VASTDB. However, this module to discover microexon is currently unpublished and the public version of VAST-TOOLS is just restricted to quantify alternative splicing events that are annotated on VastDB, a comprehensive and curated database of splice sites (Tapial et al., 2017). Thus, microexon analyses with VAST-TOOLS are not suitable to discover and quantify microexons that are only included under certain experimental conditions (such as disease models) or even perform analyses in genome assemblies that are not included in VAST-TOOLS.

Li and collaborators developed a computational method called Augmented Transcript Mapping, ATMap (Li et al., 2015), which can discover novel microexons using RNA-seq data. ATMap first maps RNA-seq reads to annotated transcripts using Stampy (Lunter and Goodson, 2011). Then, alignments are processed to identify insertions at splice sites, which can be re-aligned into the intronic spaces flanked by canonical dinucleotides. Even though ATMap strategy was shown to be more sensitive for microexon discovery than traditional RNA-seq mappers, this software has not been released to the public domain. Thus, even though these multi-step computational methods were proven to be very sensitive in the hands of their own developers, no one in the community has been able to use them.

2.1.2 Reproducible bioinformatics analysis using workflow manager platforms

Computational workflows to discover microexons have proven to be an effective way to tailor RNA-seq processing steps in a way that favours sensitive and specific discovery and quantification of microexon alternative splicing events (Irimia et al., 2014; Li et al., 2015). Both, ATMap and VAST-TOOLS microexon module, rely on multiple steps that are performed by software which was developed by third party academic groups. These computational software are often deposited in public repositories, such as GitHub, where multiple versions of a single bioinformatic tool may be released over time. Since a combination of different software versions across the software repositories that a given pipeline needs often leads to different results, reproducibility of workflow based methods is an important challenge.

A diverse range of workflow management systems (WMS) have been developed over time, but only a few have been consistently used by large communities of computational biologists (Di Tommaso et al., 2017; Goecks et al., 2010; Köster and Rahmann, 2012; Larsonneur et al., 2018; Leipzig, 2017; Wang and Peng, 2019). Different WMS have been designed to enhance bioinformatic reproducibility, however their design has been oriented to solve different needs. For example, some WMS are oriented towards enhancing the accessibility of computational tools for biologists with limited experience in bioinformatics. Galaxy (Goecks et al., 2010) and Taverna (Wolstencroft et al., 2013) provide web-based interfaces to build computational workflows without the need of any software installation or command-line execution. On the other hand, command-line based WMS, such as Nextflow (Di Tommaso et al., 2017) and Snakemake (Köster and Rahmann, 2012), enable the design of scalable computational pipelines that can work on a standard laptop as well as high-performance computing systems (HPCS) and cloud environments. Nextflow and Snakemake enable the implementation of virtual environments and cloud containers that can fully ensure bioinformatic reproducibility.

2.1.3 Computational environments

Since bioinformatic workflows that depend on different combinations of software versions might limit the number of compatible workflows that can be used on a single computational environment, the use of environment managers has become essential for routine use of computational workflows. Conda (https://conda.io) is an open source package repository in which each computational software is available as relocatable binaries. This allows the dynamic building of isolated software without allowing system-wide administrator privileges and enables fine control of package versions. Within the computational biology community, the Bioconda project (Grüning et al., 2018) greatly expanded the bioinformatic tools available as Conda software packages from various language ecosystems such as Python, R, Perl, Java, C/C++ and Julia.

Snakemake enables a direct integration with Conda, which not only allows users to run and develop multiple computational workflows on a single workstation, but also allows the usage of different versions of software for the different steps of a single workflow. Each process within a Snakemake workflow is defined as a *rule* which contains the instruction to process input files and produce specific output files. Each *rule* can be assigned to its own conda environment, thereby enabling the use of software that would otherwise be incompatible. The fine control of the environment together with the extensive documentation have resulted in Snakemake being one of the most extensively used WMSs by the computational biology community.

2.2 Results

2.2.1 Development of a reproducible bioinformatic workflow to discover and quantify microexons in RNA-seq data

MicroExonator is a computational workflow that integrates several existing software packages with custom python and R scripts to perform discovery and quantification of microexons using RNA-seq data. MicroExonator can analyse RNA-seq data stored locally, but it can also fetch any RNA-seq datasets deposited in the NCBI Short Read Archive or other web-based repositories. As microexon annotations remain incomplete and sometimes inconsistent across different transcript annotations, MicroExonator can incorporate prior information from multiple databases such as RefSeq (Pruitt et al., 2014), GENCODE (Harrow et al., 2006), ENSEMBL (Hubbard et al., 2002), UCSC (Hsu et al., 2006) or VastDB (Tapial et al., 2017). To discover putative novel microexons, reads are first mapped using BWA-MEM (Li and Durbin, 2009) to a reference library of splice junction sequences. Misaligned reads are then searched for insertions located at exon-exon junctions. Detected insertions are retained if they can be successfully mapped to the corresponding intronic region with flanking canonical U2-type splicing dinucleotides (Sheth et al., 2006), decreasing the chances of spurious mapping by incrementing the length of the sequence that is aligned inside the intron (Fig 2.1a). To maximise the number of reads that can be assigned to each splice site, annotated and putative novel microexon sequences are integrated as part of the initial splice tags where they were detected. Reads are re-aligned with Bowtie, performing a fast but sensitive mapping of reads which is further processed to quantify PSI microexon values and perform quantitative filters (Fig 2.1b).

MicroExonator employs several filters to remove spurious matches to intronic sequences which may arise due to sequencing errors (Wu and Watanabe, 2005). To illustrate these filters I ran the initial mapping steps over RNA-seq from mouse corresponding to 289 RNA-seq samples from 18 different murine tissues and 1,657 single cells from mice visual cortex (Sloan et al., 2016; Tasic et al., 2016;

Weyn-Vanhentenryck et al., 2018). Given the large amount of spurious intronic matches that can introduce false positive microexon detection, MicroExonator implements a series of filters to provide a high confidence list of microexons (Fig 2.2a-b). As a first filtering step only those insertions that can be detected in a minimum number of independent samples (i.e. technical or biological replicates, three samples is set as default) are considered. Additionally, MicroExonator scores the sequence context of the detected canonical splice sites to measure the strength of their upstream and downstream splice junctions as quantified by a splicing strength score (Parada et al., 2014), and a Gaussian mixture model is used to exclude matches that have low U2 splice-site score values. (Fig 2.1b). Finally, MicroExonator integrates the splicing strength, probability of spurious intronic matching, and genomic conservation scores, in an adaptive filtering function to remove low confidence candidates. This final filtering step leads to a high quality list of microexons, where microexons that have a high probability of spuriously matching (generally microexon of 4 nt or shorter) are excluded (Fig 2.2c, Fig 2.2d). Further technical specifications and usability are included in the Appendix section.

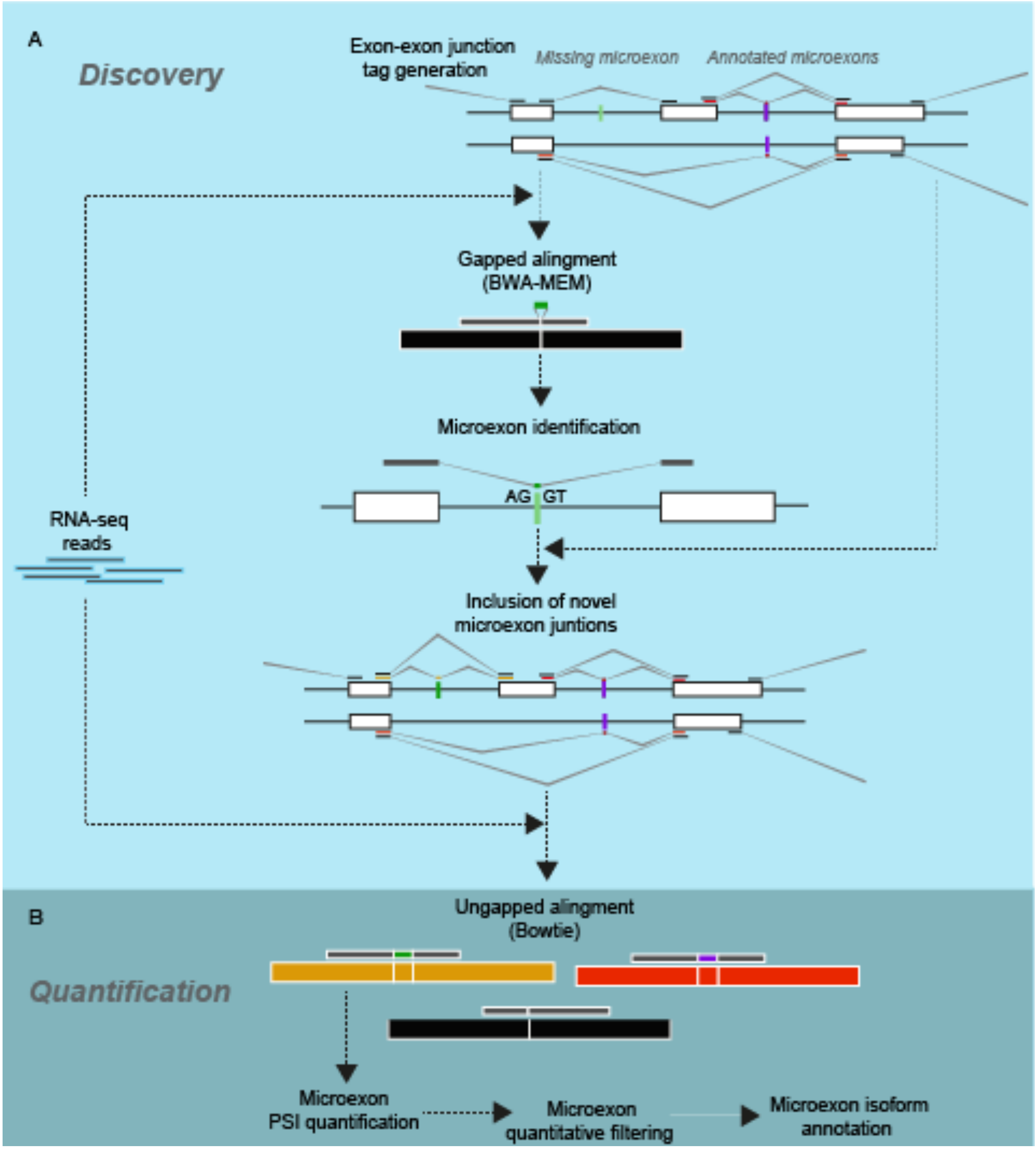
A
Discovery
Exon-exon junction tag generation
Missing microexon
Annotated microexons
Gapped alingment (BWA-MEM)
Microexon identification
AG
GT
RNA-seq reads
Inclusion of novel microexon juntions
B
Quantification
Ungapped alingment (Bowtie)
Microexon PSI quantification
Microexon quantitative filtering
Microexon isoform annotation

Figure 2.1: Overview of the MicroExonator workflow. A. To discover unannotated microexons, RNA-seq reads are aligned with BWA-MEM to the annotated splice junctions. The resulting alignments are post-processed to identify insertions at splice sites. Inserted sequences are tried to be mapped inside the corresponding introns with flanking GT-AG splice sites. **B.** Both putative novel and annotated microexons are quantified and filtered to produce a final list of microexons into transcript models which can be used for downstream analysis.

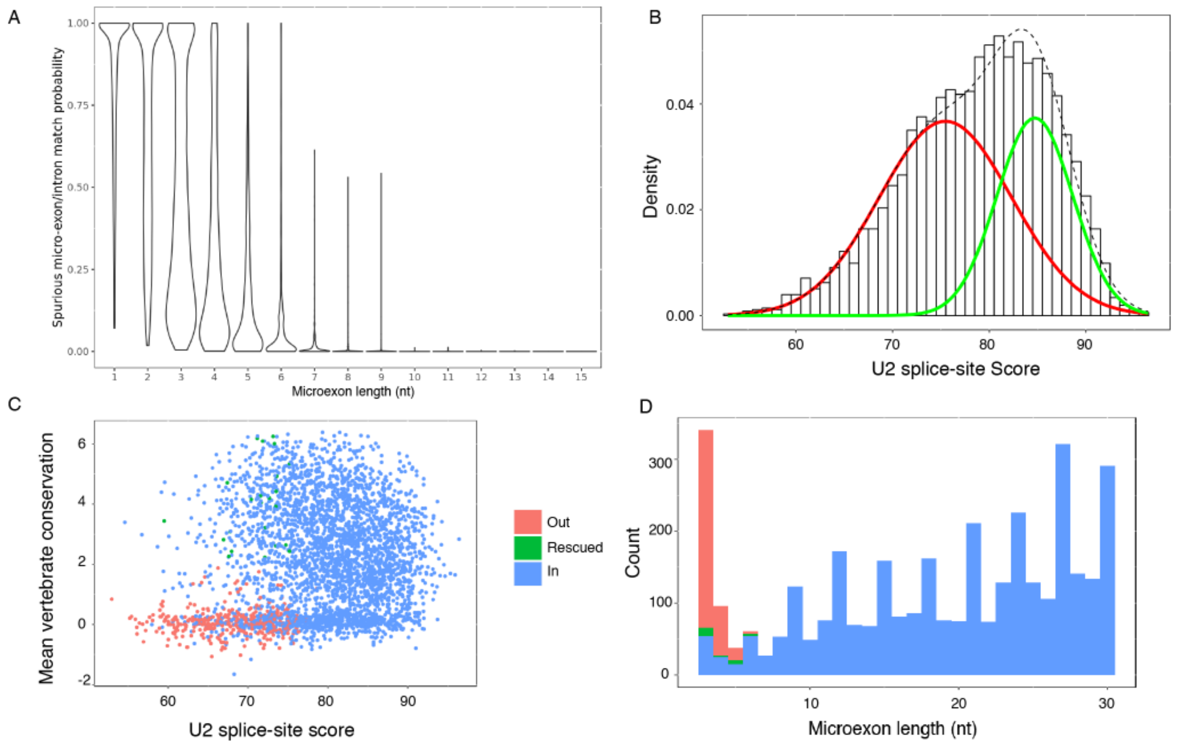

Figure 2.2: Quantitative microexon exon filtering. A. Probability of microexon spurious matches was calculated taking into account microexon length, splice site canonical dinucleotides and the size of the introns in which each microexon was discovered (see 2.3 Methods section) . **B.** A two component Gaussian mixture is used to fit the U2 consensus splicing score distribution. Lower U2 splice-site score gaussian curve (red line) is assumed to fit the distribution of spurious microexons, whereas true microexon distribution of U2 splice-site scores should be represented by a gaussian curve with higher U2 splice-site score (green line). **C.** Distribution of U2 splice-site score and mean vertebrate conservation values (phyloP score over microexons and their dinucleotides) for the total amount of candidate microexons before the final filters. Red dots represent microexons that were filtered out, blue dots microexons that were kept in the final high confident list of microexons and green dots microexons that were initially filtered out, but were rescued due to high conservation values (phyloP score ≥ 2 is used as the default value). **D.** Proportion of microexons that were filtered out, kept or rescued across different microexon sizes.

2.2.2 Benchmarking of computational methods for microexon discovery

To compare MicroExonator with other methods I incorporated a set of synthetic microexons into the GENCODE gene annotation (Fig 2.3a-b). The microexon sizes were drawn from the previously reported distributions (Irimia et al., 2014; Li et al., 2015) with greater abundance of in-frame microexons (Fig 2.3c). Moreover, I modified a copy of the mouse reference genome to replace the flanking intronic region of simulated microexons with sequences extracted from annotated splice sites. To simulate spurious microexons and evaluate their impact over the specificity of microexon discovery protocols, I randomly incorporated insertions across splice junctions. As these inserted sequences have the potential to map to intronic spaces, only microexon discovery protocols that have modules to statistically differentiate microexons from spurious matches are expected to perform well in my simulation.

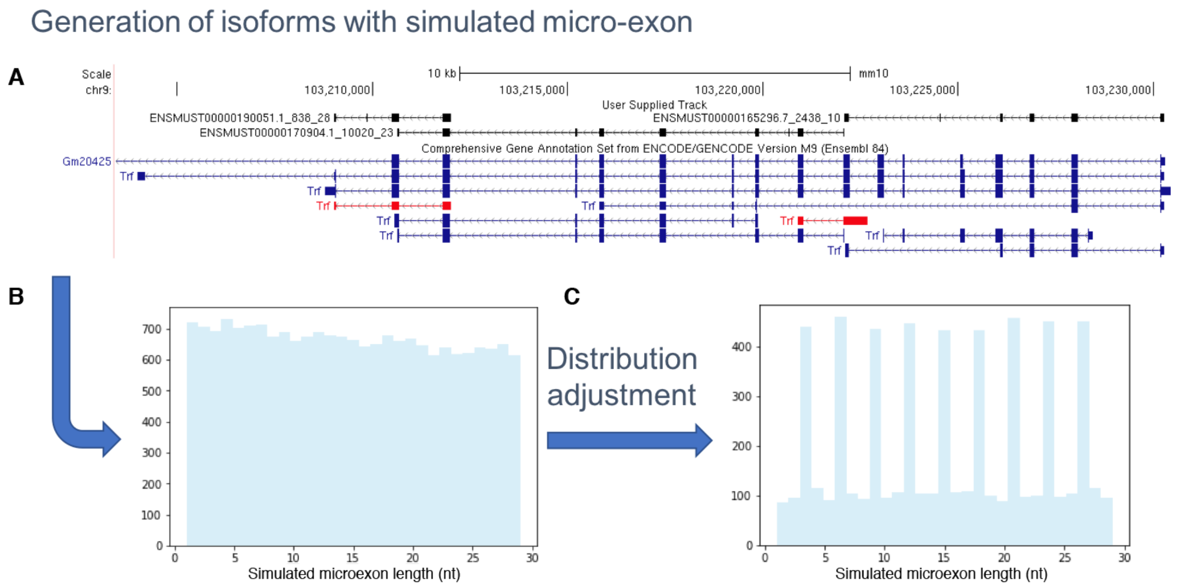

Figure 2.3: Ground truth generation for the assessment of microexon exon discovery modules. A. UCSC image showing the new isoforms generated by the insertion of simulated microexons. **B.** Raw size distribution of simulated microexons. **C.** Distribution adjustment to expected symmetric/asymmetric microexon proportions.

I used Polyester (Frazee et al., 2015) to simulate reads with a standard Illumina sequencing error rate and processed them using either MicroExonator, HISAT2 (Kim et al., 2015), STAR (Dobin et al., 2013), or Olego (Wu et al., 2013). The results show that the microexon filtering steps allow MicroExonator to distinguish simulated microexons from spurious microexons with a sensitivity >80% for all microexon lengths (Fig 2.4a-c). Even though all four aligners could detect a significant fraction of the simulated microexons, they are all limited in their ability to discover very short microexons; STAR's sensitivity drastically declines for microexons <10 nt, while the sensitivity of HISAT2 and Olego drops for microexons <8 nt. Moreover, the direct output of STAR and HISAT2's do not represent a reliable source of microexons, as they have low specificity. Using the default parameters results in a false discovery rate (FDR) of 43.0% and 33.3%, respectively. Olego had the highest specificity (FDR = 13.0%) of the other mappers, while MicroExonator achieves an FDR of 9.8%. Since MicroExonator's false discovery events are concentrated in the shortest microexons, discarding microexons <3 nt or <4 nt reduces the FDR to 2.4% and 0.75%, respectively.

The simulations also allow us to calculate the ground truth percent spliced in (PSI) values for the microexons, and to quantity how frequently a splice junction is incorporated in a transcript. MicroExonator is the only method that has low PSI errors for microexons <10 nt (Fig 2.4d). Even though MicroExonator's error rates are slightly higher for microexons >10 nt, they are still comparable to other methods. Taken together, these results show that MicroExonator is more accurate for annotating and quantifying microexons from RNA-seq data compared to conventional RNA-seq aligners.

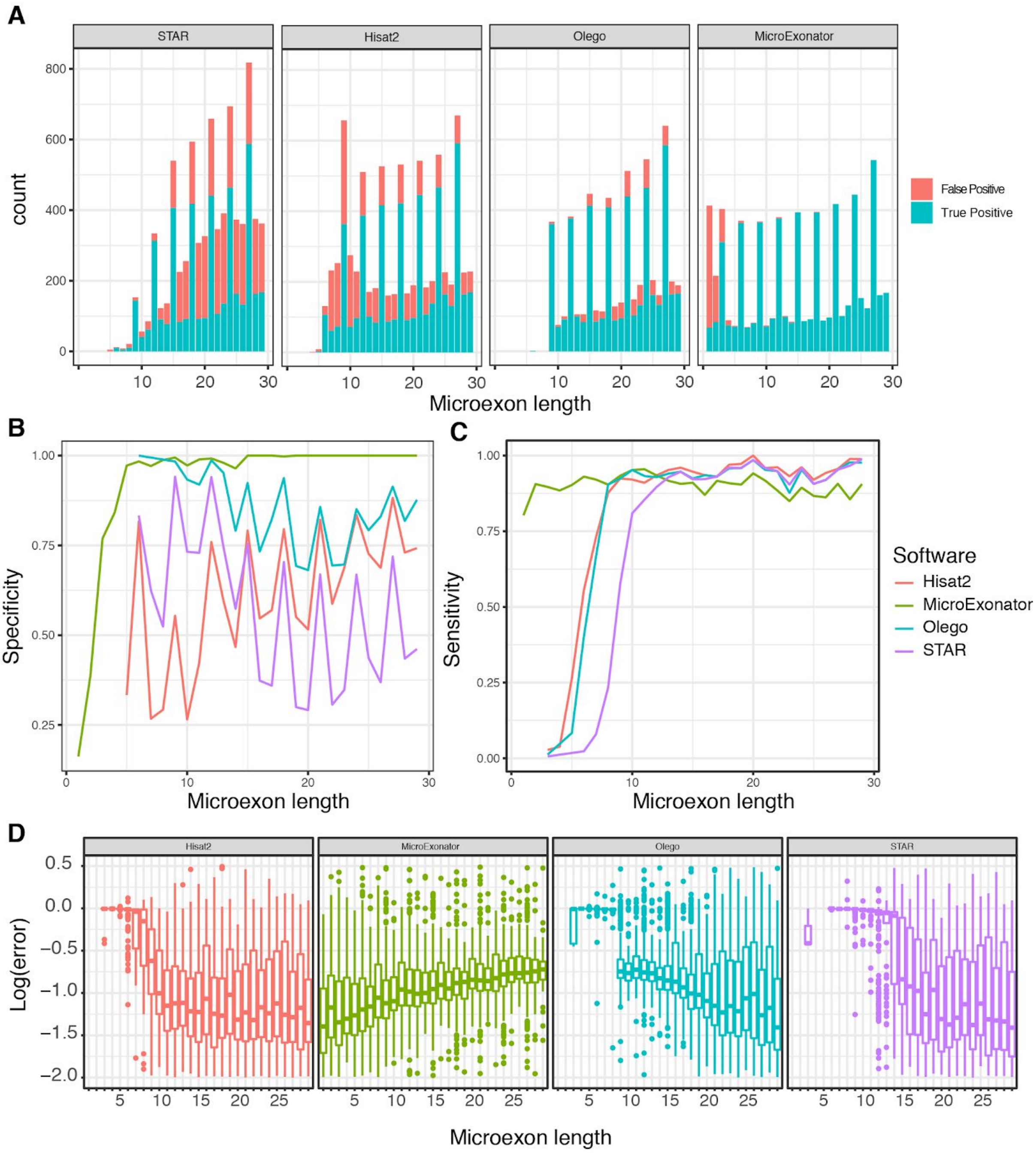

Figure 2.4: Evaluation of microexon discovery performance of RNA-seq aligners and MicroExonator using synthetic data. A. Size distribution of simulated microexons that were detected by the different software. **B-C.** Specificity and sensitivity of detected simulated microexons using multiple available tools for evaluation. **D.** Log10 error PSI values show the accuracy of the microexon quantification.

2.3 Methods

2.3.1 Annotation guided microexon discovery using RNA-seq data

MicroExonator was implemented over the Snakemake workflow engine (Köster and Rahmann, 2012), to facilitate a reproducible processing of a large number of RNA-seq samples. During an initial discovery module, MicroExonator uses annotated splice junctions supplied by the user (a gene model annotation file can be provided in GTF or BED format) to find novel microexons. RNA-seq reads are first mapped to a library of reference splice junction tags using BWA-MEM (Li and Durbin, 2009) with a configuration that enhances deletion detection (bwa mem -O 6,2 -L 25). The library of splice junction tags consists of annotated splice junctions between exons ≥ 30 nt and spanning introns ≥80 nt. For each splice junction, a reference sequence tag is generated by taking 100 nt upstream and downstream from the corresponding transcript sequence. Splice junction alignments are processed to extract read insertions with anchors ≥8 nt that map to exon-exon junction coordinates. Inserted sequences are then re-aligned inside the corresponding intronic sequence, but only matches flanked by canonical splice site dinucleotides (GT-AG) are retained (Fig 2.1a). The obtained reads are re-mapped to the reference genome using HISAT2 (Kim et al., 2017). A preliminary list of microexon candidates is generated based on reads whose insertions are aligned to the intronic spaces with no mismatches (soft clipping alignments are ignored). To further avoid misalignment artifacts, reads containing putative microexon sequences are mapped to the genome using HISAT2. Reads that map with higher mapping scores to the genome than the microexon junctions are discarded.

MicroExonator is currently available at GitHub (https://github.com/hemberg-lab/MicroExonator), where all the code and instructions on how to use it are available. Additional technical specifications and usability can be found in the appendix.

2.3.2 Quantification of microexon inclusion

In a subsequent quantification module, novel microexon candidates are integrated with the provided gene annotation to generate a second library of splice junctions tags, where putative novel loci from the discovery phase and annotated microexons are integrated at the middle of the tag sequences (Fig 2.1b). Reads are aligned again to this expanded library of splice junction tags using Bowtie (Langmead et al., 2009), which performs a fast ungapped alignment allowing for 2 mismatches (bowtie -v 2 -S). Reads that map to splice junction tags are also mapped to the reference genome using Bowtie also allowing two mismatches. Reads that could only fully map to a single splice junction tag but no other location are counted towards novel or annotated microexons.

2.3.3 Filtering of spurious intronic matches

MicroExonator uses a series of filters to distinguish real splicing events from spurious matches. For a random sequence of length L_s, where all four nucleotides have the same frequency, the probability of at least one spurious match inside an intron with flanking GT-AG dinucleotides can be calculated as:

Equation I. $$P_s = 1 - (1 - \frac{1}{4^{L_S+4}})^K$$

Where K is the number of k-mers of length $L_s + 4$ that are possible to extract from an intron of length L_i. K can be calculated as K = L_i - ($L_S + 4$). As only intronic matches that are flanked by canonical dinucleotides (4 nt) are allowed, the length of the sequence that is searched inside the intron corresponds to $L_s + 4$. Additionally, splice site signals are evaluated by measuring how well they match the canonical splicing motif as defined by the U2 position frequency matrices (Sheth et al., 2006). I call this U2-splice score or splice strength (normalized to range from 1 to 100), and it is used to build a two component Gaussian mixture model (Figure 2.2b).

Microexons shorter than 3nt cannot be identified with high specificity, and thus they are reported as a separate list. Microexons that are 3nt or longer, are prioritised according to a score (M_s) that is determined from the Gaussian mixture model probability and other parameters that are relevant for distinguishing real microexons from sequencing errors and other artefacts. The score is computed as:

Equation II. $$M_s = 1 - \frac{1 - P_S P_{U2}}{n}$$

Where P_{U2} is the probability of an intronic match, given a U2 Score, to belong to the component with higher U2 Score from the resultant gaussian mixture model and n is the number of intronic matches. During the final filter, microexons are prioritised according to M_s values.

An adaptive threshold to filter microexons by M_s values is calculated after every MicroExonator run. For this purpose, a linear model is used to fit the number of detected microexons as a function of their length, using different M_s values ranging between 0 and 1. MicroExonator suggests the M_s threshold under which the minimal residual standard deviation sum is obtained. A html report file is automatically generated at the end of every MicroExonator run, and it contains a plot of the variation of the sum of residual standard deviation values under different M_s thresholds.

By default, MicroExonator uses the suggested M_s score to filter out low scoring microexons, but the threshold can be set manually by the user. If conservation data (e.g. phyloP/PhastCons) is provided, then all low scored microexons that exceed a user-defined conservation threshold (default value = 2) are also included in the high confidence list of microexons and flagged as “rescued”.

2.3.4 RNA-seq simulation

I used simulations to evaluate the performance of different methods for microexon discovery and quantification. I used Polyester (Frazee et al., 2015) to simulate

RNA-seq reads from modified mouse GENCODE gene models (V11). To generate true positive microexons, I inserted a set of randomly selected sequences with a length of 1 to 30 nucleotides inside annotated introns longer than 80 nts (Fig 2.3a-b). At the same time, to simulate the splice site sequence distribution, I replaced splice site sequences from the simulated microexons with annotated mouse splice site sequences. In addition, to simulate spurious microexon matching (false positive microexons), I randomly included a set of insertions corresponding to intronic sequences at exon-exon junctions that were not flanked by canonical splicing sequences. The insertion rates and lengths were simulated parameters extracted from real RNA-seq experiments from postnatal forebrain samples. Taken together, our simulations provide a realistic set of false positive microexons that emulates real RNA-seq experiment condition as closely as possible.

3 Chapter III: Microexon quantitative analyses across mouse brain development and visual cortex

Collaboration note

All of the work shown in this Chapter will be appended to some results from Chapter II to publish a manuscript under preparation, in which I will be the leading author and Eric Miska and Martin Hemberg will be the corresponding authors. Moreover, the results and figures here presented partially correspond to the current last version of the results that we are planning to submit within a month from this thesis submission date. I produced all the code necessary to carry the complete data analysis and data visualization here presented, but the overall product was only possible to collaborative efforts carried out by Roberto Munita, Ilias Georgakopoulos-Soares, Hugo Fernandez[4], Emmanouil Metzakopian[5], Maria Estela Andres[6].

3.1 Introduction

Even though the first reports of the highly neuron-specific microexon inclusion events date several decades ago (Santoni et al., 1989; Wiestler and Walter, 1988), only recent genome-wide analyses of microexons have enabled to uncover the landscape of neuronal microexon splicing events (Irimia et al., 2014; Li et al., 2015). These analyses have shown that microexons are a highly conserved and regulated network of neuronal events, which modify a wide range of neuronal proteins involved in neurogenesis and axonogenesis, synapse, kinase activity, vesicle transport and cytoskeleton regulation.

Quantitative analyses of microexon inclusion have shown clusters of coordinated microexon inclusion events that are progressively included through *in vitro* neuronal

[4] Postdoctoral fellow at UK Dementia Research Institute, Department of Clinical Neurosciences, University of Cambridge. Cambridge, UK

[5] Group leader at UK Dementia Research Institute, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

[6] Group leader at the Department of Cellular and Molecular Biology, Faculty of Biological Sciences, Pontificia Universidad Católica de Chile, Santiago, Chile.

differentiation (Irimia et al., 2014). These microexon splicing events are largely regulated by the combinatorial effects of a range of RBPs, such as SRRM4, RPBOX1 and PTB1 (Gonatopoulos-Pournatzis et al., 2018; Irimia et al., 2014; Li et al., 2015). Experimental knockdown and overexpression of SRRM4 have been shown to have large effects over neuronal cassette exon inclusion and have functional consequences for neurite outgrowth and interfere with neuronal differentiation process (Calarco et al., 2009; Raj et al., 2011, 2014). The generation of knockout SRRM4 mice showed that the loss of this protein factor results from impairments of the central and peripheral nervous systems, affecting neurite outgrowth, cortical layering and axon guidance (Quesnel-Vallières et al., 2015). Even though SRRM4 affects a wide range of alternative splicing events, microexons are the main group that is affected by SRRM4 absence and the re-establishment of *wild-type* PSI levels of a single microexon at *UNC13B* gene was shown to be sufficient to rescue a neuritogenesis defects induced by SRRM4 absence (Quesnel-Vallières et al., 2015). Together these results demonstrate the key role of microexon inclusion for normal neuritogenesis .

Since clusters of microexons have been shown to be progressively included during *in vitro* neuronal differentiation (Irimia et al., 2014), I hypothesized that there are groups of microexons that are differentially included during mouse embryonic development and that they have a wide range of effects on neuronal protein functions. The massive amount of data that is currently available in public repositories enabled unprecedented access to transcriptome complexity, however microexons cannot be efficiently detected when standard tools to process RNA-seq data are used. Therefore the processing of raw available RNA-seq experiments using methods that can reliably identify and quantify microexons are necessary to explore their tissue-specific patterns and dynamic splicing changes during developmental time. Moreover, since neuronal tissues such as the brain cortex are particularly diverse in terms of cell-types, the integration of scRNA-seq data has the potential to provide a detailed map of microexon splicing changes across brain cell-types. Thus, in this chapter I used MicroExonator to process a large set of bulk and single cell RNA-seq experiments in order to explore microexon inclusion patterns during mouse embryonic development and across cortical neuronal subtypes.

3.2 Results

3.2.1 Microexon inclusion changes dramatically over mouse embryonic development

To investigate how microexon inclusion patterns change during mouse development, I analysed 271 RNA-seq datasets generated by the ENCODE consortium (ENCODE Project Consortium, 2004). These RNA-seq data originate from 17 different tissues, (including forebrain, hindbrain, midbrain, neural tube, adrenal gland, heart, and skeletal muscle) across 7 different embryonic stages (ranging from E10.5 to E16.5), early postnatal (P0) and early adulthood (8 weeks). In addition, I analysed 18 RNA-seq experiments from mouse cortex across nine different time points; embryonic development (E.14.5 and E16.5), early postnatal (P4, P7, P17, P30), and older (4 months and 21 months) (Weyn-Vanhentenryck et al., 2018). Using the annotations provided by GENCODE and VastDB I detected 2,966 microexons in total, and I quantified their inclusion by calculating PSI values for each mouse sample. As some microexons were detected in lowly expressed genes, I only retained microexons whose inclusion or exclusion was supported by >4 reads in >10% of the samples, and this resulted in 2,557 microexons.

To characterize the splicing patterns I performed dimensionality reduction using probabilistic principal component analysis (PPCA) (Roweis, 1998; Tipping and Bishop, 1999), and I identified three components that together explain 79.4% of the total PSI variance across samples (Fig 3.1a-b). The first principal component (PC1) accounts for 56.9% of PSI variance and strongly correlates with embryonic developmental stage of neuronal samples measured as days post conception (DPC) between E10.5 and E14.5, suggesting a strong coordination of microexon splicing during brain embryonic development (Fig 3.1c). PC2 explains 16.7% of PSI variability and is exclusively related with muscular-specific microexon inclusion patterns that were detected in heart and skeletal muscle, suggesting muscle-specific microexon splicing patterns (Fig 3.1a). Finally, PC3 explains 6.2% of PSI variability and it is related to microexon alternative splicing changes in whole cortex postnatal samples, suggesting that microexon neuronal splicing keeps changing after birth, but to a lesser extent than during embryonic development (Fig 3.1b).

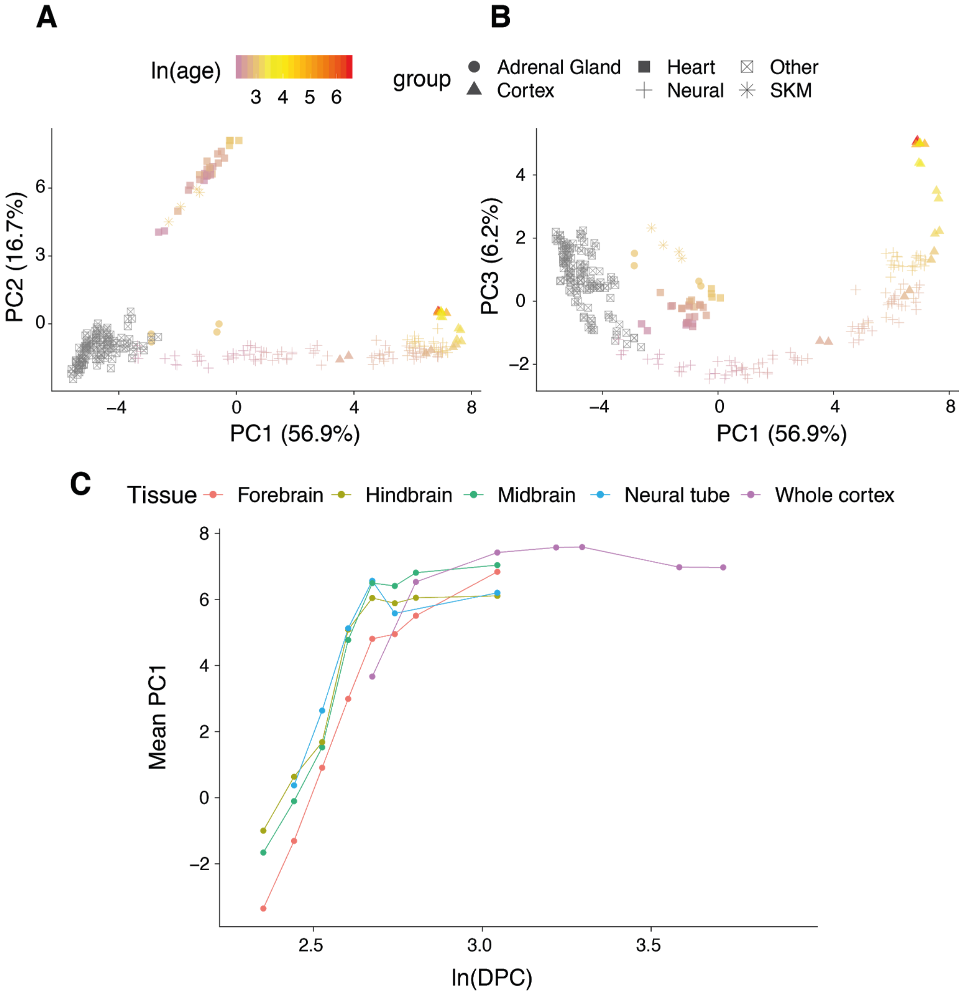

Figure 3.1. Microexon inclusion through mouse embryonic development. A. Dimensionality reduction using probabilistic principal component analysis of microexon PSI values across mouse embryonic and postnatal samples reveals correlation with developmental time for PC1. PC2 separates heart and SKM from other tissues. **B.** PC3 is correlated with developmental time of the postnatal brain samples. **C.** PC1 correspondence with embryonic developmental time, here expressed as log days post conception.

To further investigate tissue-specific microexon changes throughout development I performed biclustering of microexon PSI values from the different embryonic samples, and I obtained 24 microexon and 17 sample clusters (Fig 3.2a). Each of the sample clusters represents a combination of well defined subsets of tissues and embryonic states (Fig 3.2b). For example, samples corresponding to brain, heart, skeletal muscles (SKM) and adrenal gland (AG) form separate groups, with the only exception being E10.5 brain samples which clustered together with embryonic facial prominence and limb from E10.5 to E12.0. Consistent with the dimensionality reduction analysis, samples from the brain cluster preferentially by developmental time rather than by neuronal tissue, suggesting that microexon alternative splicing changes are greater between developmental stages than between brain regions.

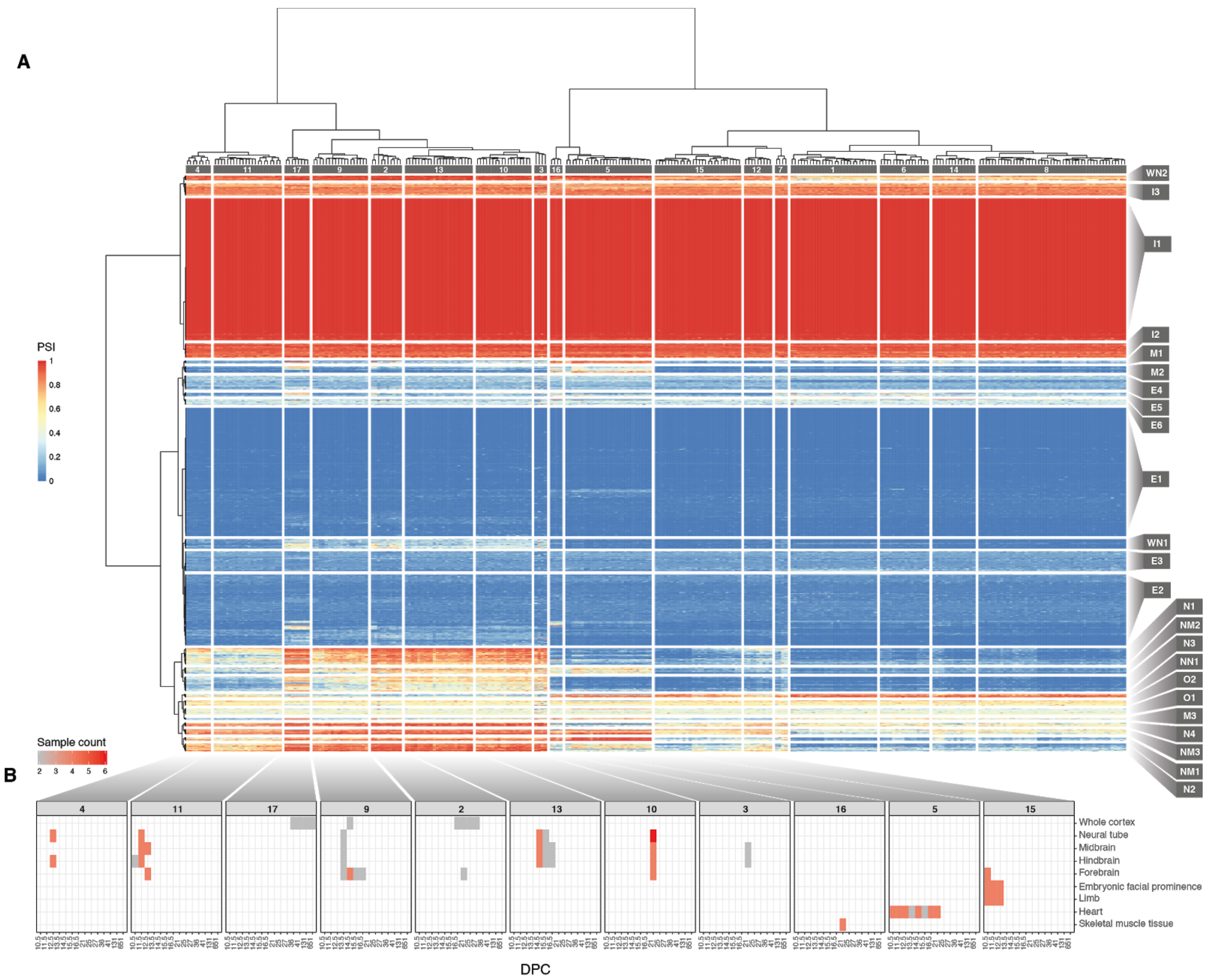

Figure 3.2. Microexon PSI biclustering A. Heatmap showing microexon inclusion patterns across analysed RNA-seq samples. Rows correspond to microexons and columns to RNA-seq samples. Blue to red colour scale represents PSI values. Microexon cluster names are shown on labels displayed at the right side. **B.** Tissue type and developmental stage composition from sample clusters containing neuronal samples or samples associated with high microexon inclusion.

The 24 microexon clusters were further analysed by dividing them into eight main categories based on the loading factors of the first two components from the PPCA (Figure 3.3). Assuming that PC1 and PC2 represent variance that can be associated with brain and muscle respectively, loading factors can be used as a proxy to evaluate the tissue-specificity behavior of microexon clusters. Following this logic, microexon clusters that have high mean loading factors (>0.03) for PC1 and PC2, were considered as neuromuscular (NM1-3). Clusters that have high loading for either PC1 or PC2 were considered as neuronal (N1-4) and muscular (M1-3), respectively. The remaining microexon clusters correspond to microexon that mostly have PSI inclusion levels that do not change across tissues. Thus, those microexon clusters that have an average PSI value lower than ⅓ were classified as Excluded (E1-6), while microexon clusters with a mean PSI value greater than ⅔ were classified as Included (I1-2). Only two clusters did not match any of the classification criteria mentioned above, so they were labeled as Other (O1-2). The number ID given to each microexon cluster corresponds to ranks computed based on PC1 or PC2 mean loading factors. By this way, N1 corresponds to the neuronal microexon cluster with the highest mean loading factor for PC1, while M1 is the muscular microexon cluster with highest mean loading factor for PC2.

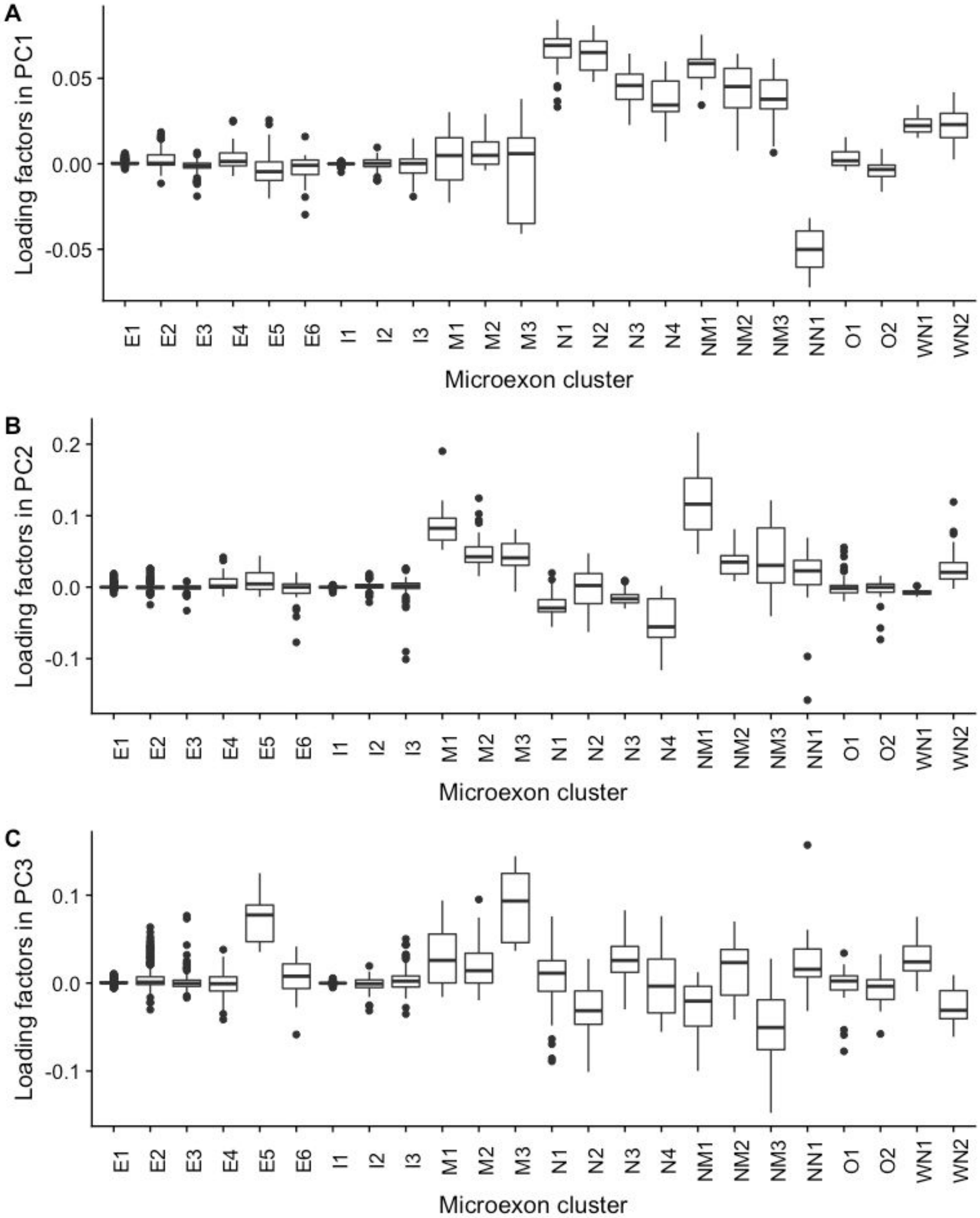

Figure 3.3: PPCA loading factors across microexon clusters. A-C. Letters in the x-axis denote the different microexon clusters.

Studies of standard alternative exons have shown that they typically have weaker splice signals than constitutive ones, and that they are less likely to disrupt the reading frame (Keren et al., 2010). Thus, I measured the splice site strengths as defined by the average U2 score of microexon flanking splice sites and the fraction of microexons that preserve the reading frame for each cluster (Fig 3.2d). As expected, the included clusters exhibit the strongest splicing signals, while the excluded clusters have the weakest splice sites, suggesting that constitutive inclusion of microexons relies on strong splicing signals. Moreover, the excluded clusters have a lower fraction of in-frame events, implying that they are likely to be more disruptive to gene function. Interestingly, neuronal, muscular and some neuromuscular clusters have almost as weak splice sites as the excluded clusters, but the fraction of in-frame events is on average 79.2%. This is considerably higher than the in-frame fractions for longer cassette exons (overall 43.2% and developmentally regulated 68.7%) (Weyn-Vanhentenryck et al., 2018). On the other hand, non-neuronal clusters have high U2 scores and also the highest in-frame microexon fraction. The in-frame fraction of each microexon cluster is strongly correlated with the conservation of the coding sequence (Pearson correlation = 0.86, p-value < 10^{-7}, Fig 3.3e), which implies that microexon clusters with higher conservation tend to preserve the protein frame.

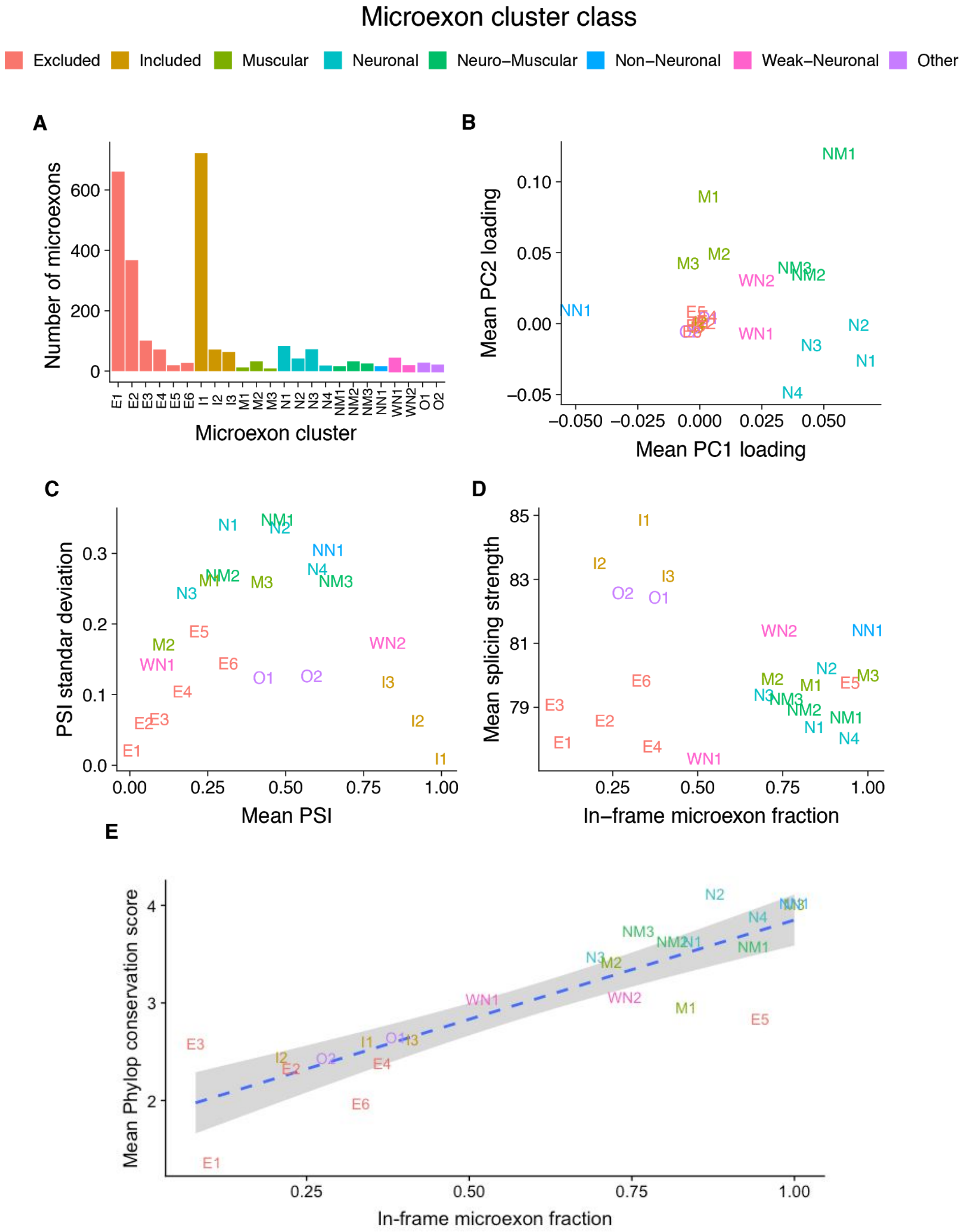

Figure 3.4: Inclusion properties of microexon clusters. **A.** Number of microexons belonging to each cluster. **B**. Mean loading factors across each cluster for PC1 and PC2. **C**. Mean and standard deviation of PSI values across microexon clusters. **D.** Mean U2 scores and in-frame fraction across microexon clusters. **E.** Relationship between genomic conservation and fraction of in-frame microexons for different microexon clusters.

Since microexons were previously shown to be progressively included during *in vitro* neuronal differentiation (Irimia et al., 2014), I hypothesized that neuronal and neuromuscular microexon clusters are progressively included throughout mouse embryonic brain development. Since PC1 strongly correlates with the developmental time from the samples (Fig 3.1), it can be considered as a proxy for early neuronal developmental time. To display how microexon PSI values relate to PC1, I calculated the average PSI value for microexon across tissue clusters and then I sorted these values according to the mean PC1 values of each tissue cluster (Fig 3.5). As expected, microexon clusters with higher mean PC1 loading factor values show greater mean PSI variability tissue clusters (Fig 3.3, Fig 3.5). Moreover, neuronal and neuromuscular clusters show a progresive increase of mean PSI inclusion values between tissue cluster number 11 and 2, which correspond to a range of tissue clusters that consist of neuronal samples extracted from increasingly older embryos (Fig 3.2, Fig 3.5). Moreover, across this same range of tissue clusters, non-neuronal microexons (NN1) show decreasing mean PSI values, which is in accordance with the negative loading factor values that were observed for this cluster. All these analyses suggest that there are groups of microexons that are progressively included at different rates during mouse embryonic development, while there is a minority group of microexon which follows the opposite trend.

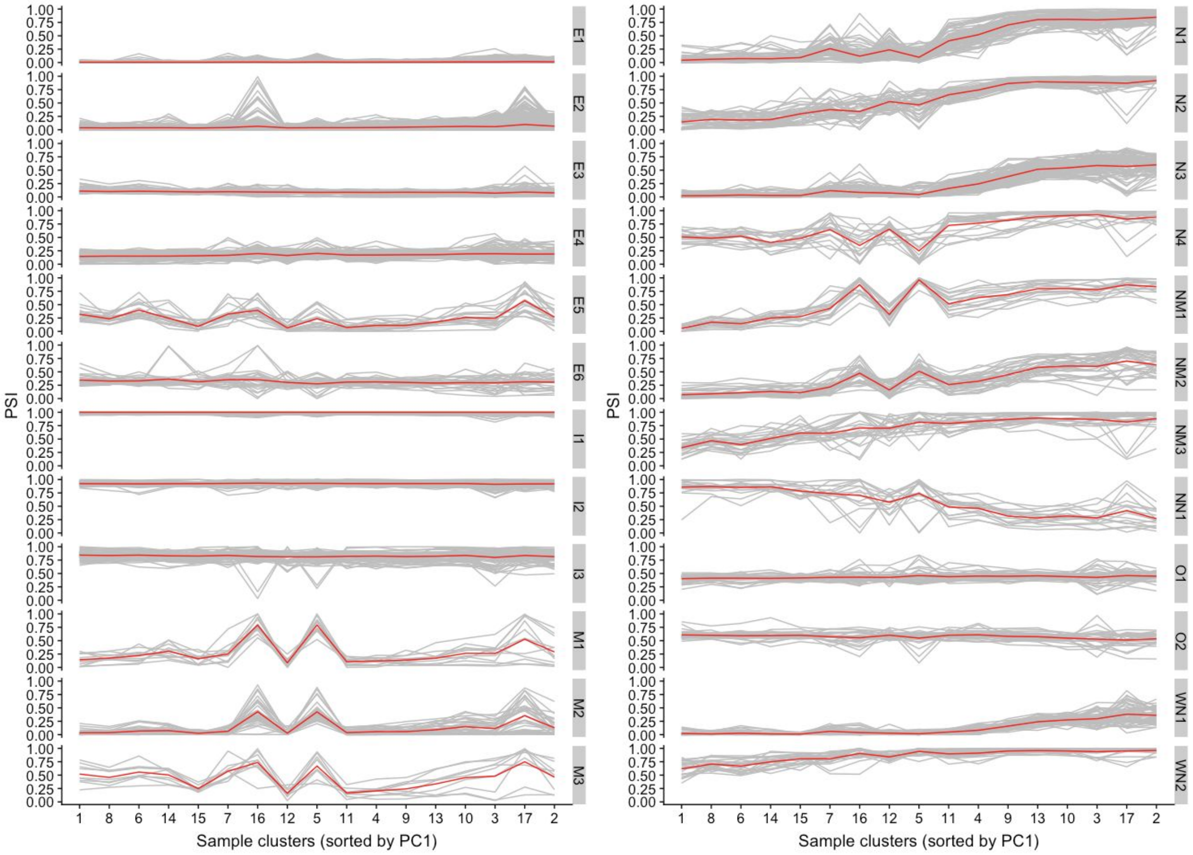

Figure 3.5: Microexon PSI values across all identified microexon clusters. Grey lines correspond to individual microexons, while red lines denote the average PSI value for a given microexon cluster across sample clusters.

In order to compare the PSI variation across mouse embryonic brain development, I defined the group of tissue clusters with lowest absolute values of mean PC1 loading factors (C1, C6 and C8) as baseline for null neuronal microexon inclusion (negative control). As expected, the contrast of the mean PSI values between these values and neuronal and muscular samples revealed distinct patterns across neuronal, neuro-muscular and non-neuronal clusters (Fig 3.6).

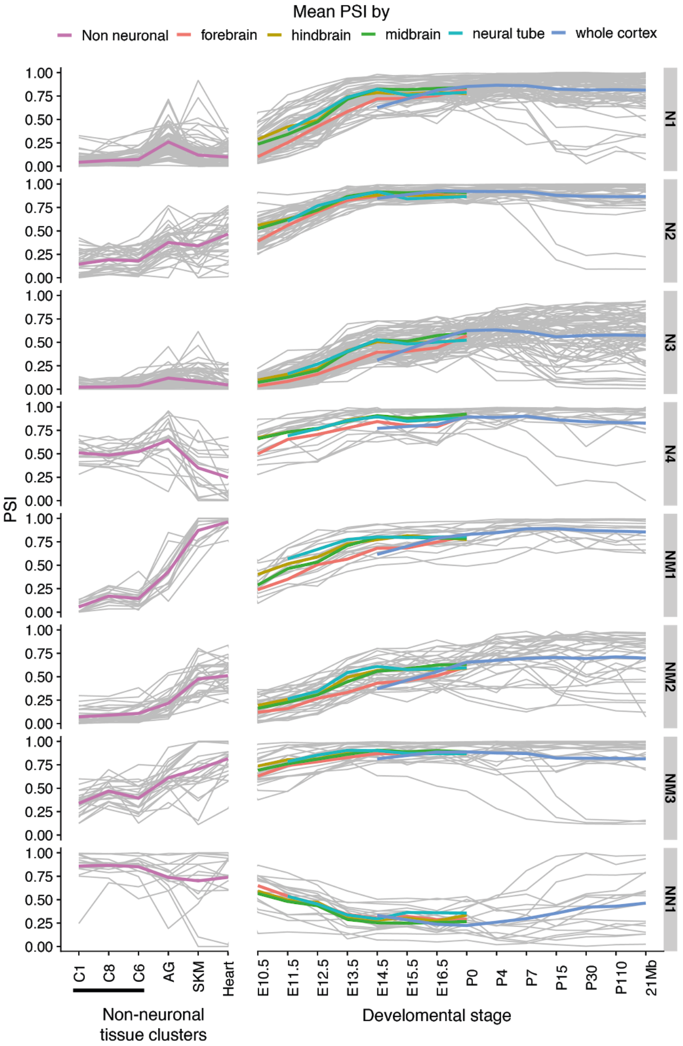

Figure 3.6: Mean PSI values across neuronal and neuromuscular microexons. Each grey line represents mean PSI values for a microexon across all samples from a tissue cluster or neuronal developmental stage (x-axis).

To quantitatively assess alternative splicing across different sample sets, I integrated Whippet (Sterne-Weiler et al., 2018), which provides a module for quantifying splicing events (whippet-quant) and a statistical framework to assess alternative splicing events (whippet-delta). Given an input gene annotation file, Whippet builds contiguous splice graphs (CSGs) to represent each transcript. In a CSG nodes represent non-overlapping exonic sequences, while edges represent splice junctions or contiguous exonic regions (Fig 3.7). Since the reads are directly mapped to the CSG, Whippet enables a fast annotation-oriented quantification of splicing events. Thus, I integrated Whippet as an optional microexon re-quantification module downstream of the MicroExonator discovery module. For this purpose, MicroExonator integrates the final list of high confidence microexons into the gene annotation file and generates a gene transfer file (GTF) which enables Whippet to quantify annotated and novel microexons, in addition to other alternative splicing events. To incorporate MicroExonator quantification results into Whippet's statistical framework, PSI values for microexon splicing nodes are replaced by the ones obtained by MicroExonator. In a later step, both Whippet and MicroExonator based quantifications are used to assess alternative microexon inclusion across the given set of comparisons.

[redacted]

Figure 3.7: An overview of Whippet's computational workflow to quantify alternative splicing events. A. Illustration of Whippet's Node assignment given an example gene annotation with two isoforms. **B.** Representation of the CSG model that would be built given the example gene annotation provided above. **C.** Transcriptome indexing from CSGs generated for each annotated gene. **D.** Read alignment to the indexed transcriptome. E. Alternative splicing quantification through node PSI estimation, which takes into account the full set of RNA-seq reads aligned to edges that connect or exclude the corresponding splicing nodes. This figure was taken from Sterne-Weiler *et. al* 2018.

The implementation of the Whippet quantification module enabled the systematic assessment of microexon alternative splicing events across mouse embryonic brain development. RNA-seq samples from midbrain, hindbrain and neural tube (MHN) were grouped by their correspondent developmental stage and compared with the previously defined negative control using whippet-delta. The evaluation of microexon alternative splicing events detected using both Whippet and MicroExonator, shows an increasing number of inclusion events throughout mouse embryonic brain development (Fig 3.8a-b), which is consistent with the gradual inclusion of neuronal microexons observed in Fig 3.6. High correlation values can be observed between delta PSI values for microexon splicing nodes quantified with Whippet and MicroExonator (Fig 3.8c). However, correlation values obtained across different microexon splicing node types differ substantially (Fig 3.8d). While most microexon microexons splicing nodes are CE type (here referred as mCE) and flanked by strictly intronic regions, some CE are also flanked by AA or AD splicing nodes that represent the inclusion of a longer microexon (mAA or mAD) or an exon longer than 30 nt (AA or AD). Correlation between delta PSI values calculated using Whippet and MicroExonator is highest for mCE, mAA and mAD splicing nodes (Fig 3.8d). By contrast, microexons that are flanked by exonic (CE_mAA / CE_mAD) or microexonic splicing nodes (mCE_mAA / mCE_mAD) had significantly lower correlation. These splicing nodes are frequently derived from complex alternative splicing events where microexons could be completely skipped or included in a shorter form. Whippet was reported to perform particularly well for complex alternative splicing events (Sterne-Weiler et al., 2018). However, Whippet PSI measurements for mCE_mAA and mCE_mAD splicing nodes are highly correlated with their corresponding mAA and mAD nodes (Fig 3.9e), suggesting that their measurements may not be independent under Whippet quantification model. Some of these highly correlated microexon pairs exhibit lower correlations when quantified by MicroExonator, suggesting active competition between shorter and longer microexons. Since the MicroExonator quantification module is only based on the relative number of spliced reads that represent each set of splicing paths that are compatible or incompatible with microexon inclusion, it was able to disentangle the inclusion on microexon associated to alternative 5′/3′ splice sites. Competition of short and longer forms of microexons have already been reported to have a key role for LAR-RTP protein function in synaptic adhesion, thus a precise quantification of these events might contribute to deeper understanding of neuronal microexon splicing (Won and Kim, 2018; Yamagata et al., 2015a).

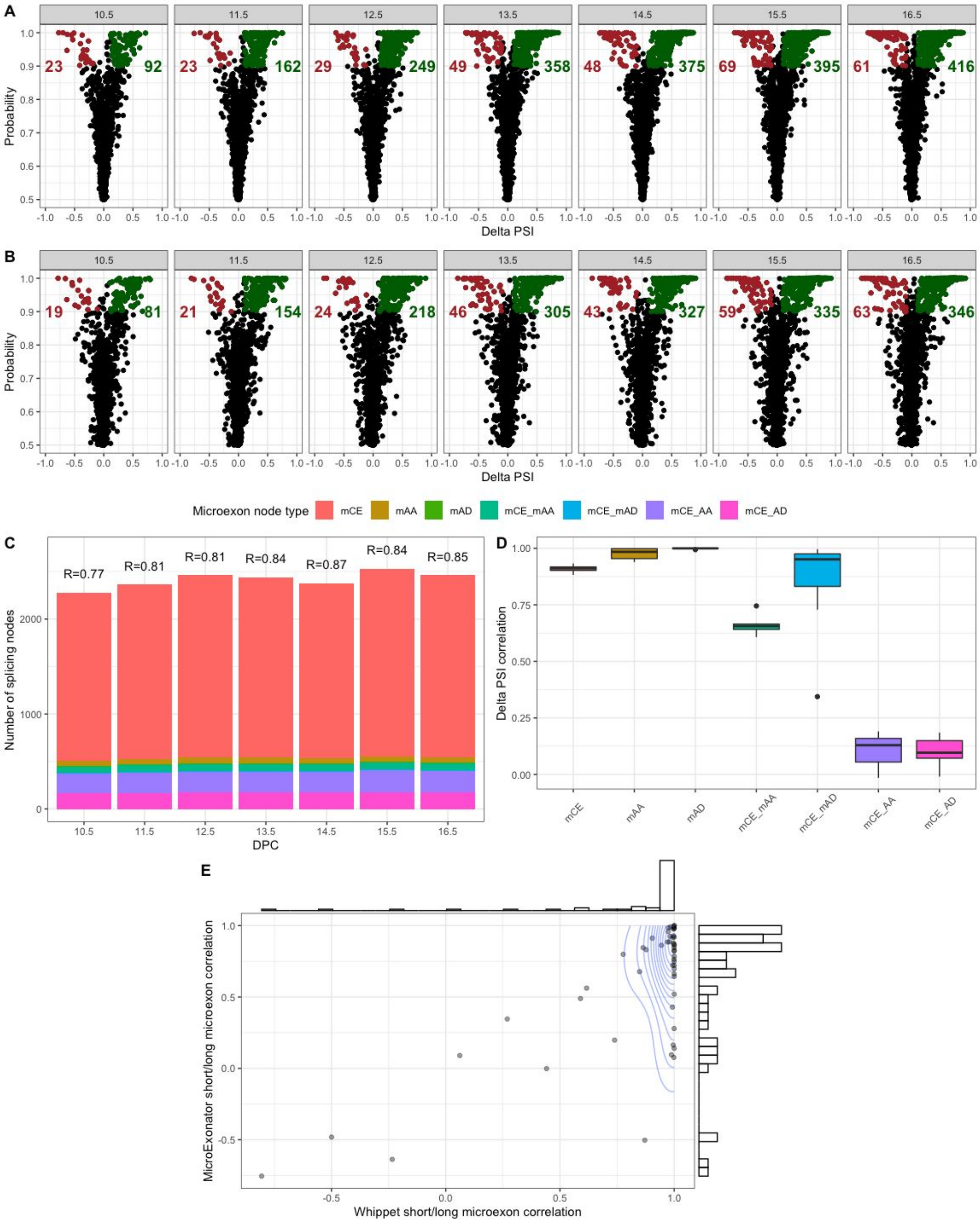

Figure 3.8: Differential inclusion analysis performed MicroExonator and Whippet quantification outputs show similar trends. A-B. Volcano plots showing the distribution of delta PSI values of microexon splicing nodes and their corresponding probability of being differentially included across MHN samples coming from different developmental stages (E10.5-E16.5). Delta PSI measurements were calculated using Whippet (A) or MicroExonator (B) microexon inclusion quantification. Alternatively included splicing nodes are highlighted in red (excluded) and green (included). Coloured numbers indicate the corresponding quantity of each group of differentially included splicing nodes. **C.** Abundance of splicing nodes quantified across the different comparisons classified according to the different classes mentioned above. Number on top indicate Perason's correlation index values (R). **D.** Correlation between mCE_AA / mCE_AD and their flanking mAA / mAD splicing nodes on Whippet and MicroExonator quantification.

I found 422 microexons that were consistently detected as differentially included on both Whippet and MicroExonator splice node quantification across at least one of the MHN comparisons performed against the defined base group. Interestingly, 323 of these microexon changes are maintained for all subsequent stages once they have been observed, meaning that they correspond to stable transcriptome signatures that are acquired during embryonic mouse brain development (Appendix - Table I). The distribution of the developmental stages when these sustained microexon changes started to be detected differed. While some microexon clusters showed early changes (N1 and N2), other clusters started to be differentially included later on (N3, NM1 and NM2) (Fig 3.9a). As forebrain tends to show delayed microexon inclusion compared to midbrain, hindbrain and neural tube (Fig 3.1c, 3.6), I pooled forebrain samples between E10.5 and postnatal (P0) and compared samples grouped by developmental stage with the non-neuronal control sample group. I found 401 microexons that were differentially included during at least one forebrain developmental stage, with 258 that were sustained through all later developmental stages (Fig 3.9b). While all the observed microexon changes across neuronal and neuromuscular clusters correspond to inclusion events, microexons from the non-neuronal cluster (NN1) only correspond to exclusion (Fig 3.9a-b).

In agreement with previous studies (Irimia et al., 2014; Li et al., 2015) I also found strong inclusion patterns associated with heart and SKM. In addition, I found microexon inclusion patterns associated with AG samples (Fig 3.1a-b, 3.6). Compared with the set of non-neuronal control samples, I found 81, 109 and 58 microexons to be differentially included in heart, SKM and AG respectively (Fig 3.9c). Most neuronal and neuromuscular microexon clusters show distinct microexon inclusion patterns compared to controls, whereas non-neuronal clusters were associated with microexon inclusion events in heart or exclusion events in SKM samples (Fig 3.9c).

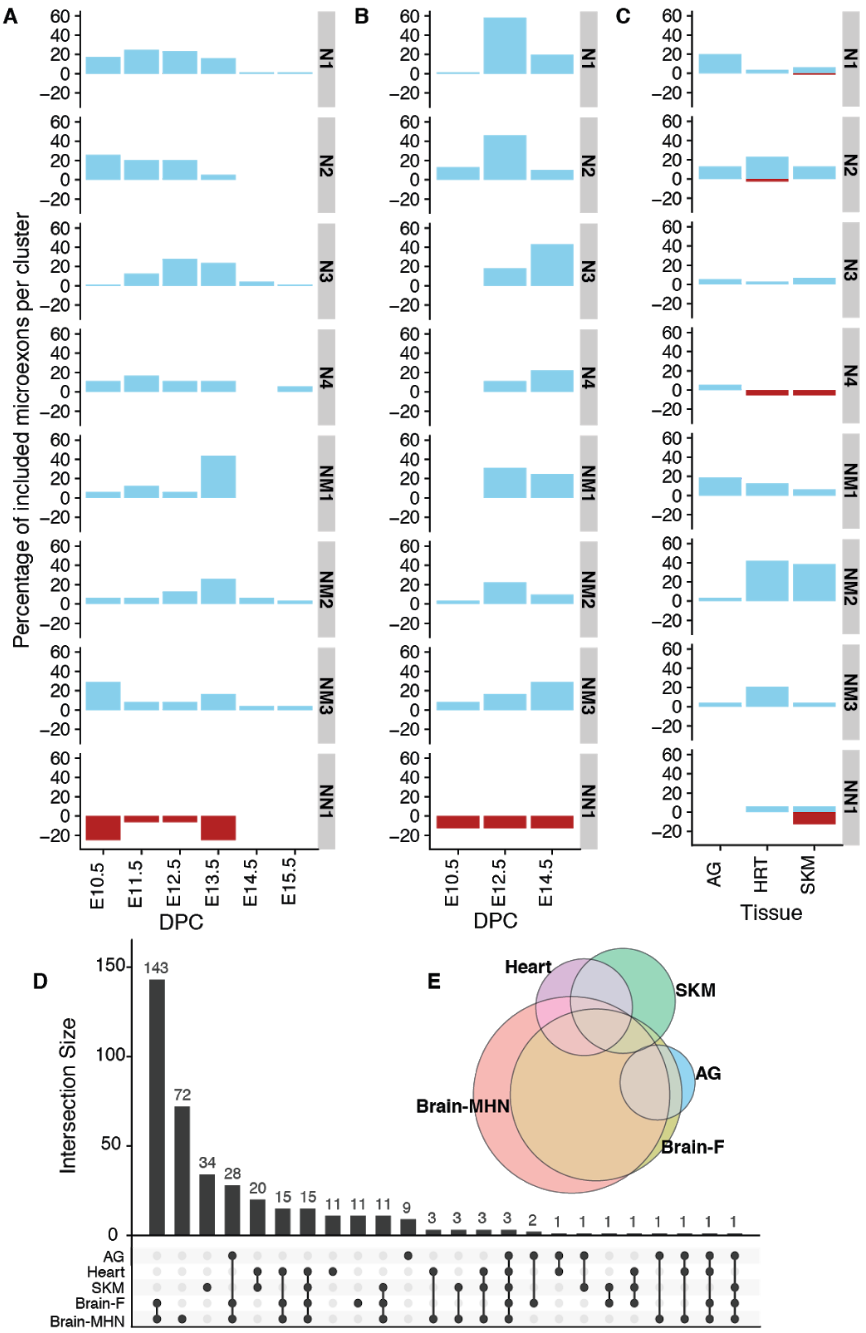

Figure 3.9: Differential inclusion analysis of microexons. A-C. Alternative microexons detected between non-neuronal tissue samples and midbrain, hindbrain and neural tube (F); forebrain (G); adrenal gland (AG), heart (HRT) and skeletal muscle (SKM) (H). Microexon splicing changes are represented as the percentage of microexons corresponding to each microexon cluster, where microexon inclusion fractions are represented with blue bars and exclusion events with upside down red bars. **D.** Intersection between microexon sets that were differentially included across sample groups. The vertical bars show the number of microexons corresponding to combinations indicated by the connected dots below. **E.** Area-proportional Euler diagram representing the most abundant intersections between differentially included microexon sets.

The set of microexons that were differentially included across the different tissue groups (brain-MHN, forebrain, heart, SKM and AG) overlap. Closer inspection reveals high concordance between the set of microexons associated with sustained changes in inclusion across MHN and forebrain samples. Surprisingly, I found a significant overlap of alternatively included microexons that have concordant patterns in AG and neuronal samples (hypergeometric test p-value < 10^{-30}). Nearly all of the AG microexons are also found in neuronal samples (Fig 3.9d-e), but in AG I observed lower PSI values (Fig 3.10). I hypothesize that the mixture between neuronal and non-neuronal isoforms found in AG is due to the chromaffin cells in the adrenal medulla which are derived from the neural crest and share fundamental properties with neurons (Bornstein et al., 2012; Shtukmaster et al., 2013).

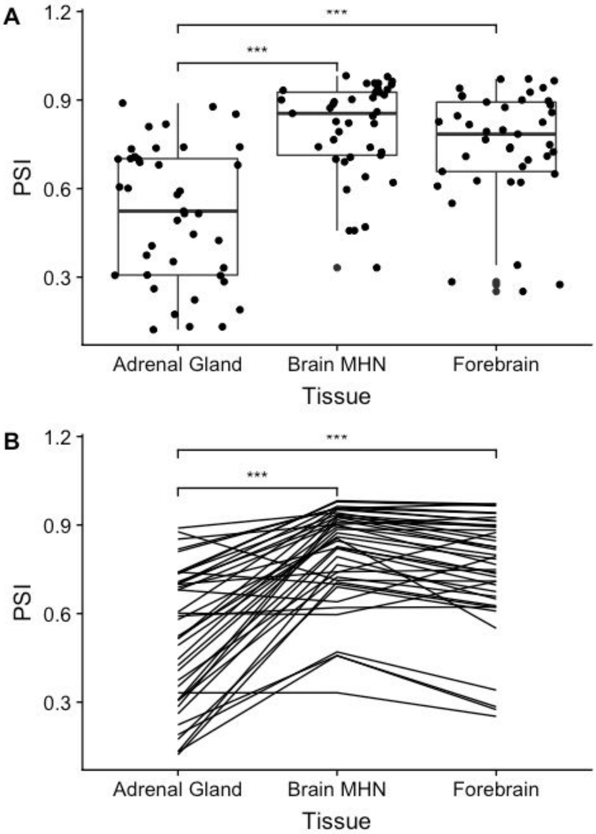

Figure 3.10: Differences in PSI score between adrenal gland, brain MHN and forebrain tissues. A. Shown by box-plots superimposed with jittered dot-plots. **B.** Shown by line-plots. Statistical differences were assessed by Wilcoxon test while correcting for multiple comparisons. Significant p-values are denoted by * (>0.05), ** (>0.01) and *** (>0.001).

3.2.2 Microexon alternative splicing is coordinated throughout embryonic development

Based on *in vitro* studies of neuronal differentiation, it has been proposed that microexons are an integral part of a highly conserved alternative splicing network (Irimia et al., 2014). Our analysis of mouse embryonic data (Fig 3.6) shows that most microexons remain included once their splicing status has changed. To explore possible functional consequences of these splicing changes I analyzed the interactions between the proteins which contain microexons by constructing tissue specific protein-protein interaction (PPI) networks for brain, heart, SKM and AG using STRING (Szklarczyk et al., 2017). For all four PPI networks the degree of connectivity was significantly higher than expected by chance given the same number of nodes (p-value<10^{-16}). On average, there were 2.7-fold more connections than expected by chance, with brain having the largest number of connections (Appendix - Table II). Next, I considered the gene ontology (GO) terms and pathways associated with the PPI networks (Fabregat et al., 2018). The Reactome pathways that showed a significant enrichment, include parts of molecular complexes that are involved in membrane trafficking pathways, e.g. “ER to Golgi anterograde transport”, “Clathrin-mediated endocytosis”, “Golgi associated vesicle biogenesis”, “Intra-Golgi and retrograde Golgi-to-ER traffic” and “Lysosome vesicle biogenesis” (Fig 3.11a-b). I also found a distinct cluster that is annotated as part of “Protein-protein interactions at synapses" (Fig 3.11d). This group includes presynaptic proteins, e.g. liprins (*PPFIA1*, *PPFIA2* and *PPFIA4*), protein tyrosine phosphatase receptors (*PTPRF*, *PTPRD* and *PTPRS*) and neurexins (*NRXN1* and *NRXN3*), which are involved in trans-synaptic interactions with multiple postsynaptic proteins, having a key role in synaptic adhesion and synapse organization. The interactions of these proteins have been shown to be highly regulated by alternative splicing (Takahashi and Craig, 2013), and our results reveal that many of these events occur towards the end of embryonic development (Fig 3.11f).

In agreement with previous reports that have highlighted the importance of microexons for axonal and neurite outgrowth (Ohnishi et al., 2017; Quesnel-Vallières et al., 2015), I detected 18 proteins in the PPI network that are annotated as part of the “Axon guidance” Reactome pathway. These proteins are found in the center of the network and they are connected with the domains involved with membrane trafficking and transsynaptic protein-protein interactions (Fig 3.11a-e). For two of the proteins associated with this pathway, the non-receptor tyrosine kinase protein SRC and L1 cell adhesion molecule (L1cam), microexon inclusion is known to play a key role in neuritogenesis (Kamiguchi and Lemmon, 1998; Keenan et al., 2017), but the importance of microexons in other proteins in this pathway remains poorly characterised. At early developmental stages (E10.5-E11.5) I found several microexon alternative splicing events in genes associated with “membrane trafficking” pathways concentrated. A subset, “clathrin mediated endocytosis” is associated with microexon changes in the later stages, as most events became significant only after E12.5 (Fig 3.11g). Similarly, “axon guidance” microexon changes mostly occur at E11.5, in particular the microexon alternative splicing events for proteins that interact with L1cam.

Since microexon inclusion occurs in several waves (Fig 3.6), I hypothesized that the temporal dynamics would be reflected in the topology of the PPI network. To quantitatively evaluate the position of each gene in the network, I calculated several centrality measures. The result is not straightforward to interpret since several of the central nodes feature more than one microexon inclusion event (e.g. *SYNJ1*, *ANK3* and *DCTN2*), which sometimes emerge at different embryonic stages . Nevertheless, the results show that L1cam and 6 out of 10 of its interactors are amongst the 15% of nodes with highest eigencentraly and that *SRC* has the highest harmonic centrality and betweenness. An investigation of genes corresponding to some of the most relevant GO terms revealed that proteins located at more central positions of the network (measured as eigencentrality[7]), have microexons that are included

[7] Eigencentrality, also known as eigenvector centrally, is a measure of node centrality that is computed based on the eigenvectors of the adjacency matrix. This method assigns higher centrality to nodes that are more connected, particularly to those that are also connected with other highly central nodes.

earlier in mouse embryonic brain development (Kruskal-Wallis rank sum, p-values < 0.05) (Fig 3.11h-i).

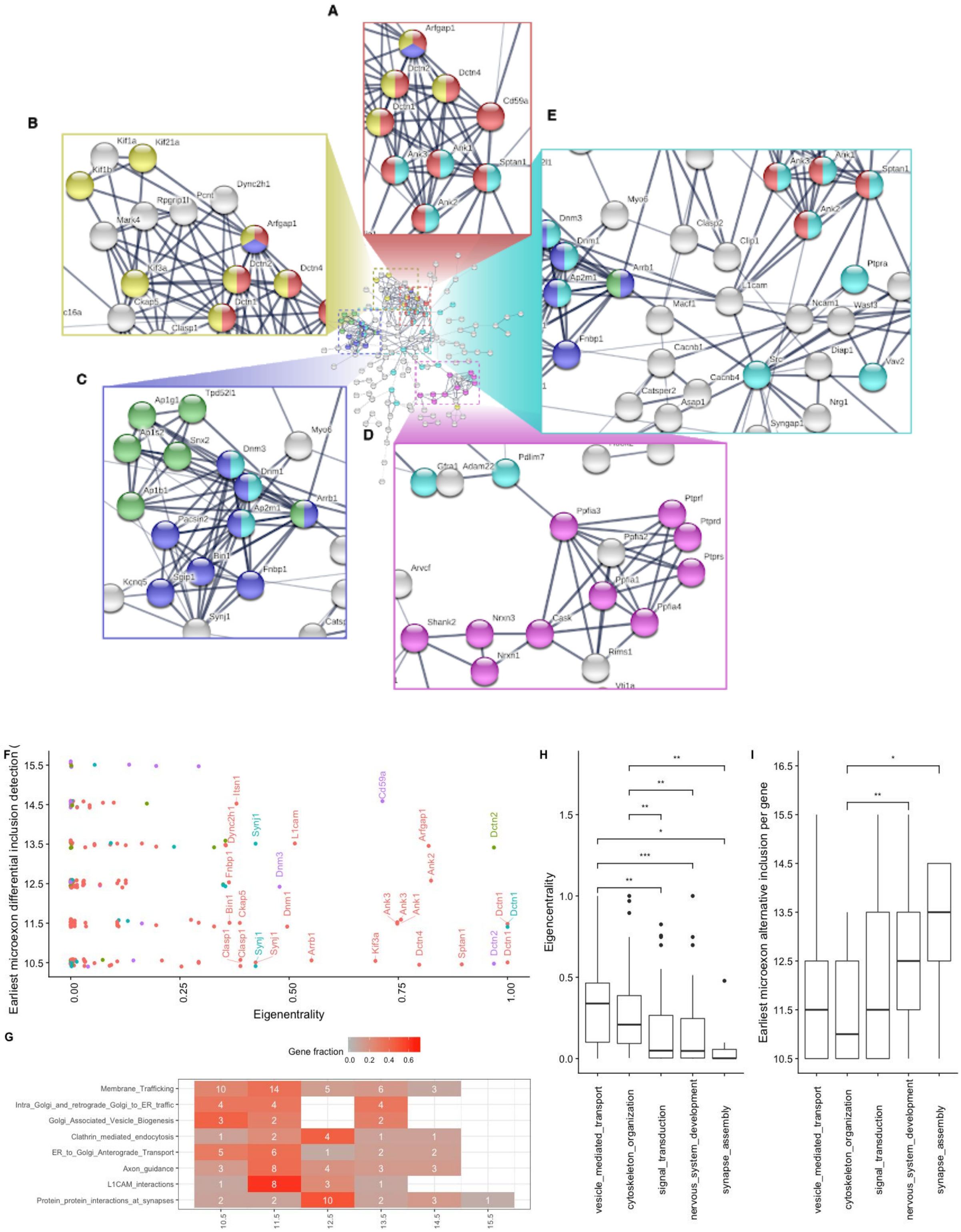

Figure 3.11: Microexon protein-protein interaction network. A-E) PPI network using as input genes that have microexons that are differentially included across mouse embryonic brain development. Colours represent different Reactome pathways that were enriched on the network; Axon guidance (light blue), Protein-protein interactions at synapses (pink), ER to Golgi anterograde transport (red), Clathrin-mediated endocytosis (dark blue), Golgi associated vesicle biogenesis (green), Intra-Golgi and retrograde Golgi-to-ER traffic (yellow). F) Eigencentrality calculated for each gene node in relation to the developmental stage at which each microexon was included. G) Effect of microexon alternative splicing over different Reactome pathways. Counts indicate the number of microexons that start to be differentially included at each developmental stage for different Reactome pathways that were significant after taking the whole genome as background. H-I) Eigencentrality and earliest developmental stage at which each gene is affected by differential microexon inclusion show differences across some of the GO categories that were significantly enriched after gene background correction. Statistical differences were assessed by Wilcoxon test while correcting for multiple comparisons. Significant p-values are denoted by * (>0.05), ** (>0.01) and *** (>0.001).

3.2.3 MicroExonator enables the identification of novel neuronal microexons

Of the 343 microexons that were differentially included across brain development, 90 were not consistently annotated between GENCODE and VastDB. I found 26 neuronal microexons that are only annotated in GENCODE, and 33 neuronal microexons that are not annotated in GENCODE, but are present in VastDB. Despite the fact that the mouse genome is comprehensively annotated, I found 35 neuronal microexons that are not annotated in GENCODE nor VastDB. Due to the high sensitivity and specificity demonstrated in simulations (Fig 2.4), I expect that all 34 microexons >4 nts are true positives.

To validate one of the novel microexons, I focused on the Dctn2 gene (eigencentrality of 0.76), where I detected two adjacent differentially included microexons of length 9 and 6 nts (Fig 3.4a). Neither of these microexons are annotated in GENCODE, but the 9-nt microexon are annotated in VastDB (MmuEX0013953). Interestingly, the downstream 6-nt microexon that was discovered by MicroExonator is validated by spliced ESTs (Benson et al., 2004). I detected differential inclusion of the 6-nt Dctn2 microexon from E10.5 in MHN samples, whereas in forebrain it is differentially included from E12.5 (Fig 3.4b).

Hugo Fernandez performed qRT-PCR experiments to assess the inclusion of the Dctn2 6-nt microexon during a mESC to neuron differentiation protocol using one set of primers that were designed to amplify Dctn2 isoforms with 6-nt microexon inclusion and another set to amplify total Dctn2 isoforms. After normalizing the qRT-PCR values using dilution series of neuronal samples, I calculated the ratio of 6-nt inclusion across mESC, EPI cells and differentiated neurons at two different stages (Fig 3.4c). The inclusion ratios from the qRT-PCR measurements indicate that the Dctn2 6-nt microexon is included through *in vitro* differentiation of mESC to neuron, consistent with our findings during embryonic development for this microexon. These results show that the alternative splicing quantification provided by MicroExonator can identify novel microexons, even for model organisms that are well annotated.

3.2.4 Identification of microexons in zebrafish brain.

To demonstrate how MicroExonator can be applied to species with less complete annotation, I analyzed 23 RNA-seq samples from zebrafish brain (Park and Belden, 2018). I found 1,882 microexons, of which 23.8% are not found in the ENSEMBL gene annotation. I used liftover (Hinrichs et al., 2006) to assess whether some of these microexons are evolutionarily conserved microexons in mouse, and I successfully mapped 401 zebrafish microexons. Of these, 85% mapped directly to a previously identified mouse microexon, and most of the remaining 15% mapped to longer exons. Mapping the microexons in the other direction, 617 out of the 2,938 that were identified from the mouse development data mapped to the zebrafish genome and 49.7% of those in return mapped to a zebrafish microexon. By integrating these results I obtained a total of 402 microexon pairs that are found in both zebrafish and mouse. Since 90.3% of the pairs had identical length in both species, they are highly likely to correspond to evolutionarily conserved microexons. I calculated the percentage of conserved microexons between mouse and zebrafish for each mouse microexon cluster, and I found that microexon clusters involved in neuronal regulation (Neuronal, Neuro-muscular, Non-Neuronal and Weak-Neuronal clusters) have a significantly higher degree of conservation than the other microexon clusters (two-sided Wilcoxon test, p-value < 0.01, Fig 3.4d).

To compare the microexon annotation between mouse and zebrafish, I calculated the number of conserved microexons between these two species that are missing in mouse or zebrafish gene transcript annotation. While only 6.9% of these exons are missing from the mouse transcript annotation provided by GENCODE, 16.1% are missing from the ENSEMBL zebrafish transcript annotation. Moreover, the largest fraction of conserved microexons that are missing in zebrafish transcript annotation corresponds to neuronal microexons (Fig 3.4e).

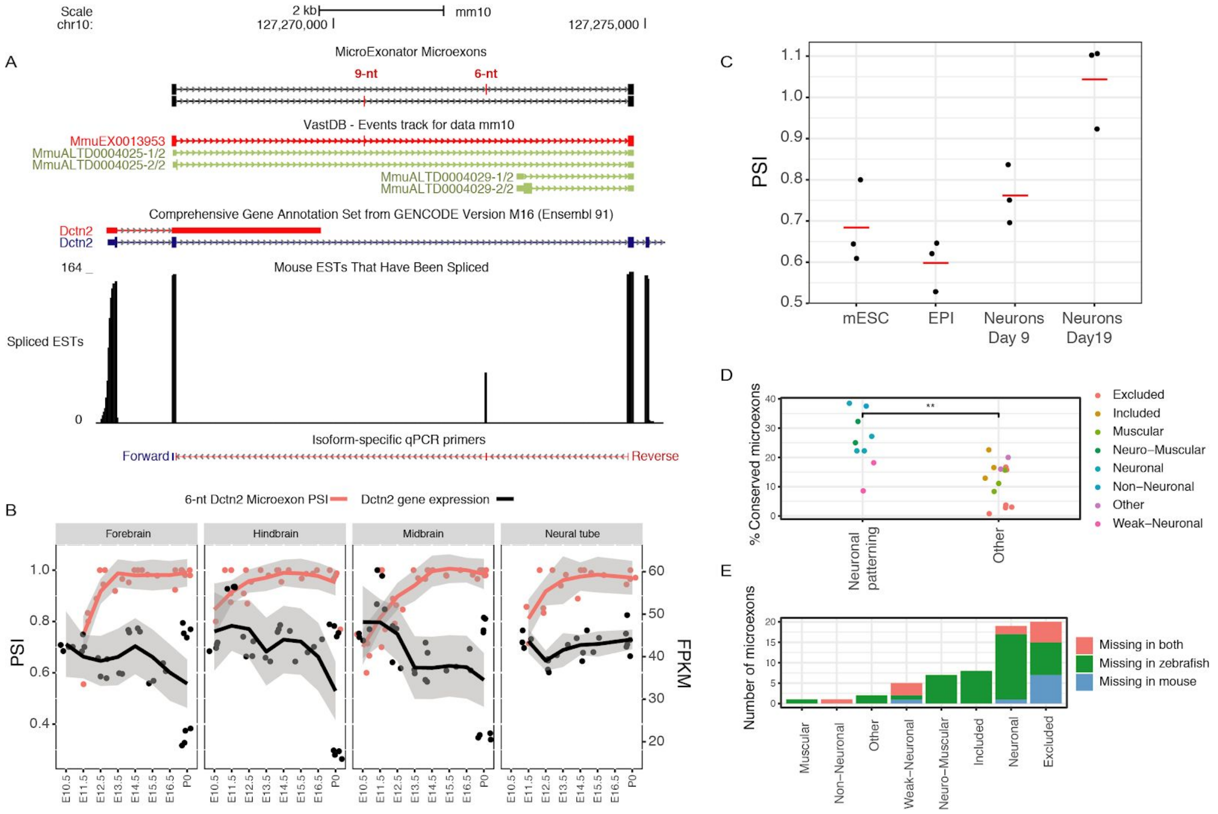

Figure 3.12: Discovery of novel microexons in mouse and zebrafish. A. Alternative Dctn2 microexons that are inconsistently annotated in mouse GENCODE and VastDB annotations. **B.** Novel 6-nt Dctn2 microexon shows a progresive inclusion through mouse embryonic development. **C.** PSI values calculated from normalized RT-PCR measurements show a gradual inclusion of the 6-nt Dctn2 microexon though *in vitro* neuronal mESC to neuron differentiation. **D.** Microexon clusters that exhibit neuronal patterning have higher conservation percentage between mouse and zebrafish than the other microexon clusters. Every dot

corresponds to a different microexon cluster and the colour indicates its type. ** denotes p-value<0.01 calculated for a two-sided Wilcoxon test. **E.** Number of conserved microexons between mouse and zebrafish that are missing from their transcript annotation.

3.2.5 Cell type specific microexon inclusion in mouse visual cortex.

Our analysis of neuronal development suggested that the main difference in microexon inclusion is between time points rather than tissues. However, these data do not reflect the diversity of cell types within neuronal tissues, and since the neural cortex is one of the most diverse tissues in the murine body, I hypothesized that microexon inclusion patterns may vary amongst different subcellular types that can be found in the adult mouse neuronal cortex. Full length scRNA-seq experiments using the SMART-seq2 protocol have enabled the identification of two main neuronal classes, glutamatergic and GABA-ergic neurons, and seven non-neuronal cell-types (Tasic et al., 2016). Despite the 3′ bias previously reported for SMART-seq2 protocol, these scRNA-seq experiments enabled Tasic and co-workers to evaluate exon usage and identify alternative splicing events across cell-types. Thus, I developed a downstream module of MicroExonator to perform alternative splicing analysis of microexons using full-length single-cell data. I used it to identify microexon alternative splicing events between GABA-ergic and glutamatergic neurons defined by Tasic et al., 2016, containing 739 and 764 cells, respectively.

I first ran the microexon discovery module with an expanded annotation, which included the microexons discovered from our previous analyses. This yielded 2,344 microexons that were included in at least one cell. Next, I used Whippet to quantify the PSI of the microexons detected by MicroExonator for each cell. Since alternative splicing analysis heavily relies on the number of splice junction reads detected, the sparsity of read coverage scRNA-seq is a technical challenge that needs to be overcome in order to reliably identify alternative splicing events. Thus, for each neuronal type I systematically pooled GABA-ergic or glutamatergic neurons into pseudo-bulk groups of 15 cells, which were subsequently quantified by Whippet using an indexed transcriptome that considers all the novel microexons identified during the analysis. The analysis of pseudo bulk PSI values identified a total of 39 differentially included microexons, 20 of which were also identified from the single cell PSI values (Fig 3.13). Moreover, all of these steps were implemented as an optional extension of the core snakemake workflow of MicroExonator, which means it can be used to identify alternative splicing events between other groups of cells profiled using full length scRNA-seq protocols.

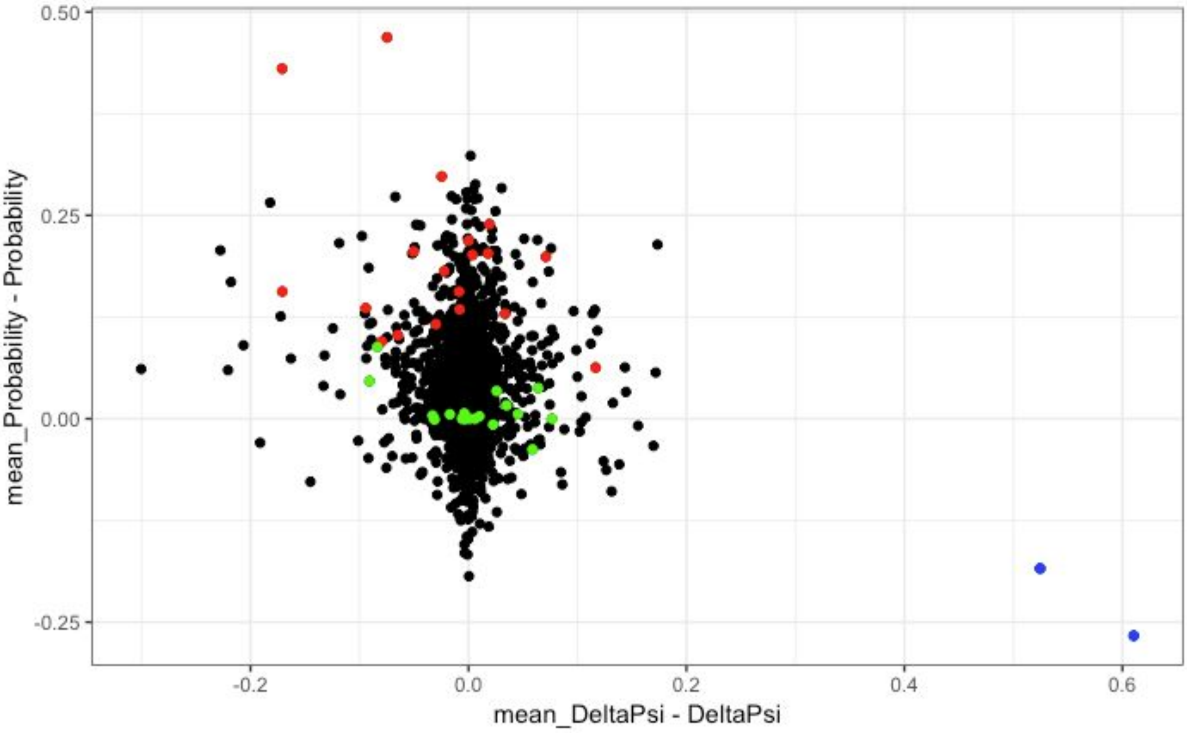

Figure 3.13: Differences between unpooled and pooled methodologies to assess microexon splicing changes in single cell data. For each microexon, different delta PSI and probability values were obtained while GABAergic and glutamatergic data were processed through pseudo pooling (pooled) or standard analysis (unpooled). To get an idea of the consistency of the results across these two methodologies. Since the pseudo pooling process was done ten times to avoid random arrangement effects, the comparison between pooled and unpooled strategies can be measured as the mean of the pooled results (delta PSI and probability) minus the ones obtained by the unpooled approach.

Among the genes that contain differentially included microexons between GABA-ergic and glutamatergic neurons is a group of eleven genes that encode for proteins that localize at synaptic compartments. I found seven presynaptic proteins, two postsynaptic proteins and two proteins that have been observed at both locations (Fig 3.14a). For example, the type IIa RPTPs subfamily of proteins undergoes tissue-specific alternative splicing that determines the inclusion of four short-peptide inserts, known as mini-exon peptides (meA-meD) (Pulido et al., 1995a, 1995b; Takahashi and Craig, 2013). While meB comprises four residues (ELRE) and is encoded by a single microexon, meA has three possible variants that can form as a result of the combinatorial inclusion of two microexons; meA3 (ESI), meA6 (GGTPIR) and meA9 (ESIGGTPIR) (Yamagata et al., 2015a). Our analysis shows a consistent inclusion of meB in both GABA-ergic and glutamatergic neurons. However, I detected cell type specific rearrangement of meA microexons which promotes inclusion of meA9 in glutamatergic neurons, while in GABA-ergic neurons meA variants are mostly excluded (Fig 3.14b). Alternative splicing of meA/B microexons are key to determining the selective trans-synaptic binding of PTPδ to postsynaptic proteins, which is a major determinant of synaptic organization (Takahashi and Craig, 2013). In addition, I found other alternatively spliced microexons in genes that are involved in synaptic cell-adhesion, e.g. Gabrg2, Nrxn1 and Nrxn3 (Südhof, 2017; Takahashi and Craig, 2013). The microexon inclusion in these genes is variable across the core clusters, sometimes showing stark differences between GABA-ergic and glutamatergic neuron subtypes (Fig 3.15). These results suggest that microexon inclusion is not only coordinated at the tissue-type level, but that it is also finely tuned across neuronal cell-types, and these differences may be of importance for determining neuronal identity.

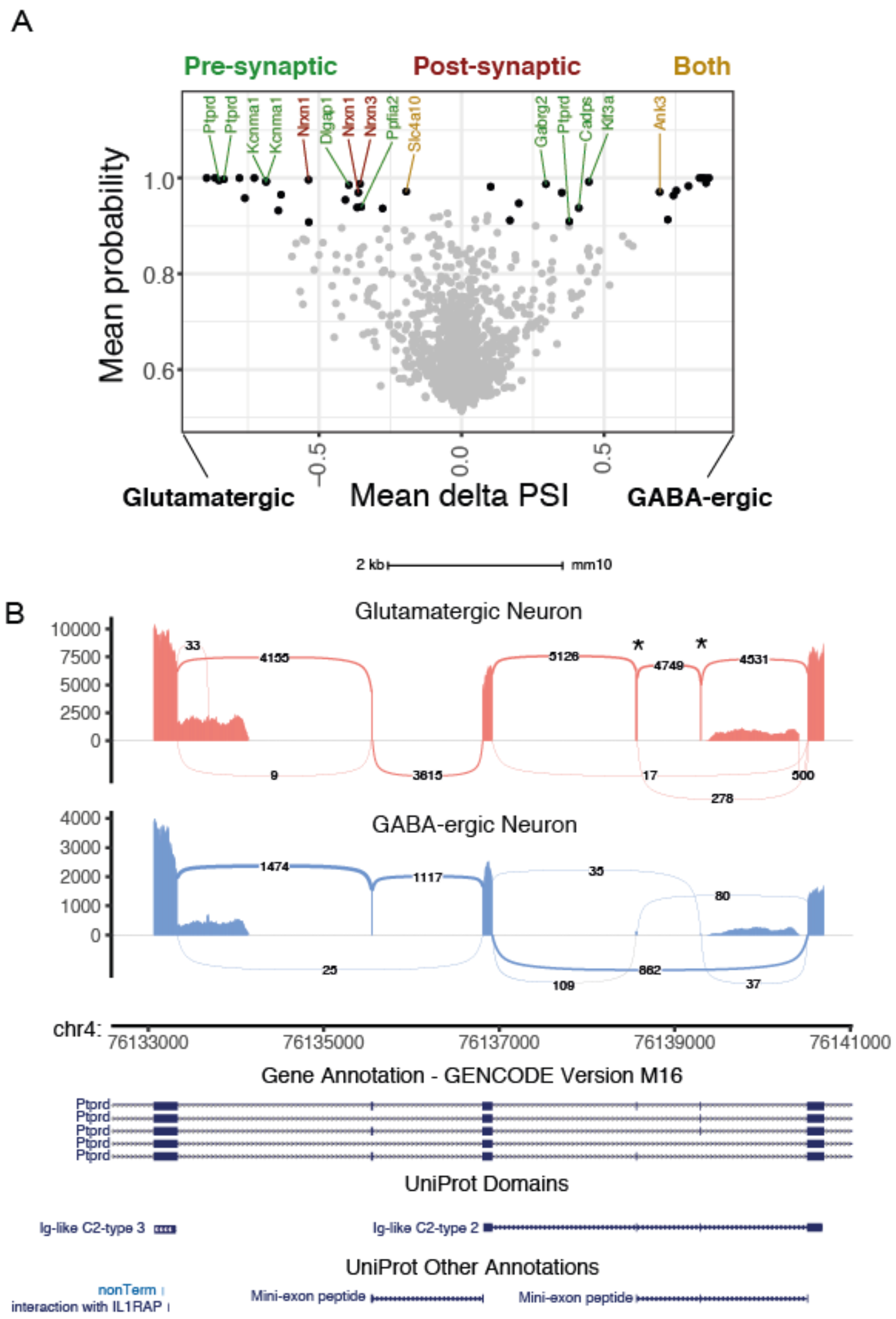

Figure 3.14: Differential alternative splicing analysis of microexons between glutamatergic and GABA-ergic neurons. A. Volcano plot showing an overview of the alternatively included microexons between glutamatergic and GABA-ergic neurons. Differentially included microexons are highlighted in black. Detected synaptic proteins containing cell-type specific microexons are labeled with different colours depending on their sub-synaptic localization. **B.** Sashimi plot showing PTPδ microexons that determine the inclusion of meA/B mini-exon peptides. Numbers indicate the amount of splice reads that support each splice junction and * denote microexons that were detected as differentially included between glutamatergic and GABA-erigic neurons.

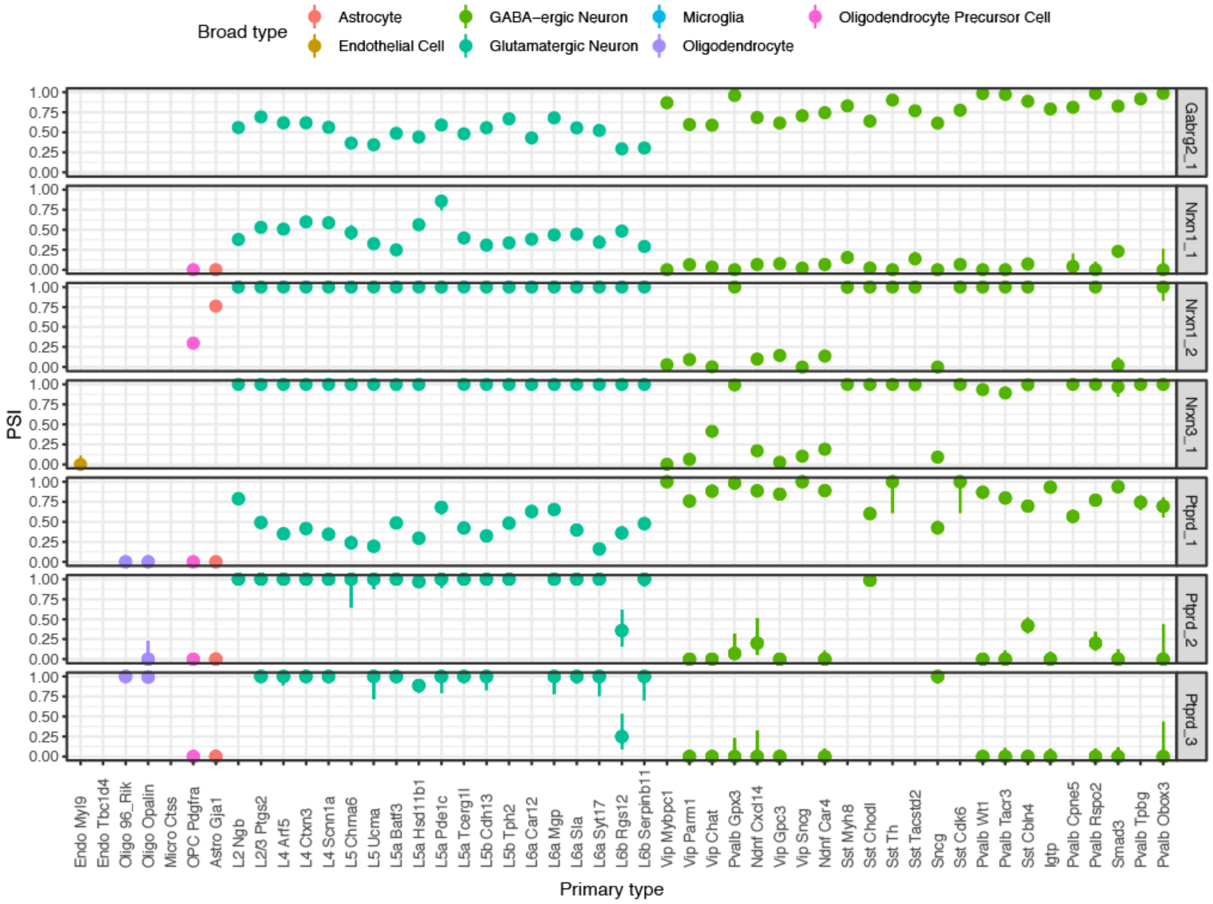

Figure 3.15: Microexon inclusion patterns at synaptic proteins across all core clusters of proteins involved in trans-synaptic interactions. Each panel shows the inclusion pattering of microexons that were found differentially included between GABA-ergic and glutamatergic neurons. Colours indicate the different broad types which each cell-type belong to.

3.3 Methods

3.3.1 Microexon analyses across mouse development using bulk RNA-seq data

As a proof of principle, I applied MicroExonator to 283 RNA-seq datasets, obtained from the ENCODE project (Sloan et al., 2016), corresponding to embryonic and postnatal tissue samples coming from 17 different tissues. For the sequence, I used mm10 mouse genome assembly, obtained from UCSC genome browser database (Karolchik et al., 2003), and as source of annotated splice junctions I used the union of GENCODE Release M16 (Harrow et al., 2006) and VastDB (Tapial et al., 2017). I quantified novel and annotated microexons, Percent of spliced-in (PSI) values, by using MicroExonator's built-in scripts or by using Whippet, which provides a one-step approach for quantify splicing at the splicing node level (Sterne-Weiler et al., 2018). Bi-clustering of samples and microexons were performed applying Ward's minimum variance criterion implemented in R (Müllner and Others, 2013; Murtagh and Legendre, 2014) over a MicroExonator distance matrix where the similarity of the samples was calculated from the PSI values . Moreover, PSI values were also used to perform PPCA using ppca function from pcaMethods R library (Stacklies et al., 2007).

The obtained PPCA loading factors were used to systematically classify microexon clusters. Assuming that PC1 and PC2 are related with variance observed at brain and muscle respectively, loading factors can be used as a proxy to evaluate the tissue-specificity behavior of a given microexon inclusion. Thus, microexons that have high loading factors (>0.03) for PC1 and PC2, were considered as neuromuscular (NM1-3). The ones that only have high loading factors for either PC1 or PC2 were considered as neuronal (N1-4) and muscular (M1-3) respectively. Additionally, one microexon cluster was found with a significant negative loading factor over PC1 (lesser than -0.03), which I considered to be non-neuronal (NN1). I also found microexon clusters that have a consistent inclusion (I1-7) or exclusion

(E1-5) pattern across all the samples and to perform differential microexon inclusion analyses I grouped sample files according to the bi-clustering results.

To perform differential microexon inclusion analyses I grouped sample files according to the bi-clustering results. For each alternative microexon cluster (N1-5, NM1-3 and NN1), baseline and signal sample sets were defined. I quantified splicing nodes using Whippet quantification module (whippet-quant.jl) and I supplied MicroExonator output as input to the Whippet differential inclusion module (whippet-delta.jl). I used both MicroExonator and Whippet quantification to assess changes in microexon inclusion between every signal cluster and its corresponding baseline cluster array. Across the different comparisons, I only considered as significant those microexons which have >0.9 probability of being differentially included and >=0.1 delta PSI values. To further avoid quantification errors, I only selected those microexons that were detected as differentially included using both MicroExonator and Whippet quantification. For each signal cluster I calculated differentially included microexons enrichment using Pearson's chi-squared test with Yates' continuity correction (Yates, 1934). Differentially included microexons were classified accordingly with the tissue composition of signal clusters in which they were found to be differentially included.

I further analyzed the sets of genes that have microexon differentially included in brain, SKM, heart or adrenal gland by building a protein-protein interaction network using STRING (Szklarczyk et al., 2017).

3.3.2 Neuronal mouse dopamine neuron preparation and RT-PCR validations

Mouse embryonic stem cells (mESC) were differentiated into dopamine neurons as previously described (Metzakopian et al., 2015). Briefly, mESCs were first differentiated into Epiblast stem cells (EPI) using fibronectin coated plates and N2B27 basal media (composed of Neurobasal media, DMEM/F12, B27 and N2 supplements, L-glutamine and 2-Mercaptoethanol) supplemented with FGF2 (10mg/ml) and Activin A (25mg/ml). After three passages, EPI were differentiated into dopaminergic neurons using plates collated with poly-L-lysine (0.01%) and

Laminin (10ng/ml) and N2B7 media supplemented with PD0325901 (1mM) for 48hours (Day 0 to Day 2). 3 days later (Day 5), N2B27 media was supplemented with Shh agonist SAG (100nM) and Fgf8 (100ng/ml) for 4 days. Media was then changed to N2B27 media supplemented with BDNF (10ng/ml), GDNF (10ng/ml) and ascorbic acid (200nM) from Day 9 onwards. During neuronal differentiation cells were passaged at Day 3 and Day 9. Cells were collected for qRT-PCR analysis at several stages: mESC, EPI, Day 9 neurons and Day 19 neurons. RNA extraction was performed using the RNeasy Mini Kit (Qiagen) and samples analysed with a QuantStudio 5 PCR system (Thermo Fisher Scientific). These experimental details were provided by Hugo Fernandez, who performed these experiments.

3.3.3 Systematic microexon identification in Zebrafish brain

RNA-seq experiments for the Zebrafish brain tissues across different time points were obtained from (Park et al. 2018) with GEO accession code GSM2971317. Microexon detecting and quantification was performed with MicroExonator using default parameters and taking Ensembl gene predictions 95 and danRer11 genome assembly as an input. To perform a comparative analysis between mouse and zebrafish microexons, I performed a batch coordinate conversion using the liftOver script from USCS utilities (Karolchik et al., 2003), which provides an straightforward, conservative and non-exhaustive way to find conserved microexons.

3.3.4 Single cell analyses

I applied MicroExonator to single cell data from mouse visual cortex (Tasic et al., 2016). In addition to MicroExonator PSI quantification, I also computed PSI using Whippet for all microexons found in this dataset. To compare microexon inclusion rates, I used Whippet to perform an iterative quantification and inclusion analysis. I pooled data coming from 2 neuronal cell-types: GABA-ergic and glutamatergic in pseudo-bulk groups of 5 cells (or fewer for the last group), and repeated this process 10 times. During each iteration, splicing node PSIs values were calculated using whippet-quant.jl. Both single-cell and pseudo-bulk were used to assess differential inclusion of splicing nodes using whippet-delta.jl to obtain average delta

PSI and probability values. Only those that had at least 0.9 mean probability and a mean delta PSI value within single cell delta PSI value +/- 0.25, were considered as significant. Sashimi plots were generated by adapting ggsashimi's code (Garrido-Martín et al., 2018) to display the total number of reads that is supported by each splice site. The read counts were further processed to calculate splice site usage rates.

4 Chapter IV: Analysis of non-B DNA motifs across splice sites

Collaboration note

The work described in this chapter is currently published as pre-print in biorxiv (Georgakopoulos-Soares et al.) and under review in Nature Communications. Both Ilias Georgakopoulos-Soares and I conceived the project, and contributed equally to the manuscript and bioinformatic analyses. While Ilias Georgakopoulos-Soares quantified non-B DNA motifs across the genome, I processed splice sites and took the lead compiling an R notebook that generates all of the figures presented here. All of our work was done under the supervision of Martin Hemberg and Eric Miska.

4.1 Introduction

Nucleic acid oligomers can adopt different conformations. In the case of DNA the most frequent structure formed *in vivo* corresponds to B-DNA, a right handed double helix, which is considered the canonical DNA structure. However, different sequence contexts, environments and biological processes, such as replication and transcription, can favour the formation of non-canonical DNA structures collectively known as non-B DNA. More than 20 non-canonical secondary structures have been previously reported for DNA, including G-quadruplexes (G4s), hairpins, cruciforms and triplexes (Ghosh and Bansal, 2003).

Sequences that predispose DNA to non-canonical conformations are known as non-B DNA motifs and they have been associated to different sources of genome instability, such as translocations and double strand breaks (Bacolla et al., 2016; Georgakopoulos-Soares et al., 2018). However, some of them can also have regulatory roles of gene expression. In particular, G4s have been shown to be enriched in promoters and nucleosome depleted regions, and some of them have been found to have important gene regulation roles (Hänsel-Hertsch et al., 2016; Huppert and Balasubramanian, 2007). For example, a G4 in the promoter of the oncogene *MYC* acts as a repressor (Hurley et al., 2006; Siddiqui-Jain et al., 2002; Yang and Hurley, 2006). Similarly, a G4 in the promoter of the proto-oncogene *KRAS* has a negative effect on expression levels (Cogoi and Xodo, 2006).

Since many non-B DNA motifs can also lead to similar secondary structures at the RNA level, their formation has the potential to affect mRNA processing (Bevilacqua et al., 2016a; Kwok and Merrick, 2017; Uzilov and Underwood, 2016). Since the presence of an extra 2′-hydroxyl group on RNA molecules promotes additional intramolecular interactions within RNA G4s, G4s are more stable in RNA than DNA molecules (Fay et al., 2017; Zhang et al., 2010). However, the impact of non-canonical DNA and RNA structures over alternative splicing remains only partially understood (Buratti and Baralle, 2004; Warf and Berglund, 2010) and although a role of G4s in splicing has been suggested (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), the extent of G4 impact on alternative splicing remains to be explored. In this chapter I describe a systematic characterization of non-B DNA motifs across splice sites and an exploration of their implications for alternative splicing modulation.

4.2 Results

4.2.1 Genome wide analysis of non-B DNA motifs across splice sites

To investigate the contribution of non-canonical secondary structures to splice site definition, we systematically explored the distribution of seven known non-B DNA motifs. Since the secondary structures can form both at the DNA and RNA level (Bevilacqua et al., 2016b; Biffi et al., 2013; Strobel et al., 2018) and it is plausible that DNA structures could have an impact, both strands were considered for this initial analysis. In order to characterize the distribution of non-B DNA motifs across splice sites, we calculated the enrichment of non-B DNA motif occurrences across splice site flanking regions. The enrichment profiles varied substantially across the different non-B DNA motif categories (Fig 4.1), with the highest enrichment found for G4s, both at the 3′ss (2.44-fold) and the 5′ss (4.06-fold). High enrichment of short tandem repeats was also observed, but that was expected since a subset of them overlap with intronic polypyrimidine tracts which are known to be part of the core splicing signal (Coolidge et al., 1997; Dominski and Kole, 1991). By contrast, the enrichment patterns for G4s or H-DNA motifs cannot be explained by the distribution of known splicing signals.

Alternative splice sites are often associated with weak splice sites, which allow them to be modulated by cis-regulatory elemens (Ast, 2004). Thus, if non-B DNA structures function as splicing cis-regulatory elements, their enrichments may vary across exons flanked by weak and strong splice sites. To investigate the association of non-B DNA motifs and splice site strength, we measured the splice strength using publicly available position weight matrices (Sheth et al., 2006), and divided the splice sites into quartiles based on how well they match the position weight matrix. We found that several non-B DNA motifs are not evenly enriched across splice site quartiles (Fig 4.2). While some non-B DNA motifs had higher enrichment at strong splice sites (splice site quartile 4, Q4), G4s were more enriched at weak splice sites, suggesting that they may be associated with alternative splicing events.

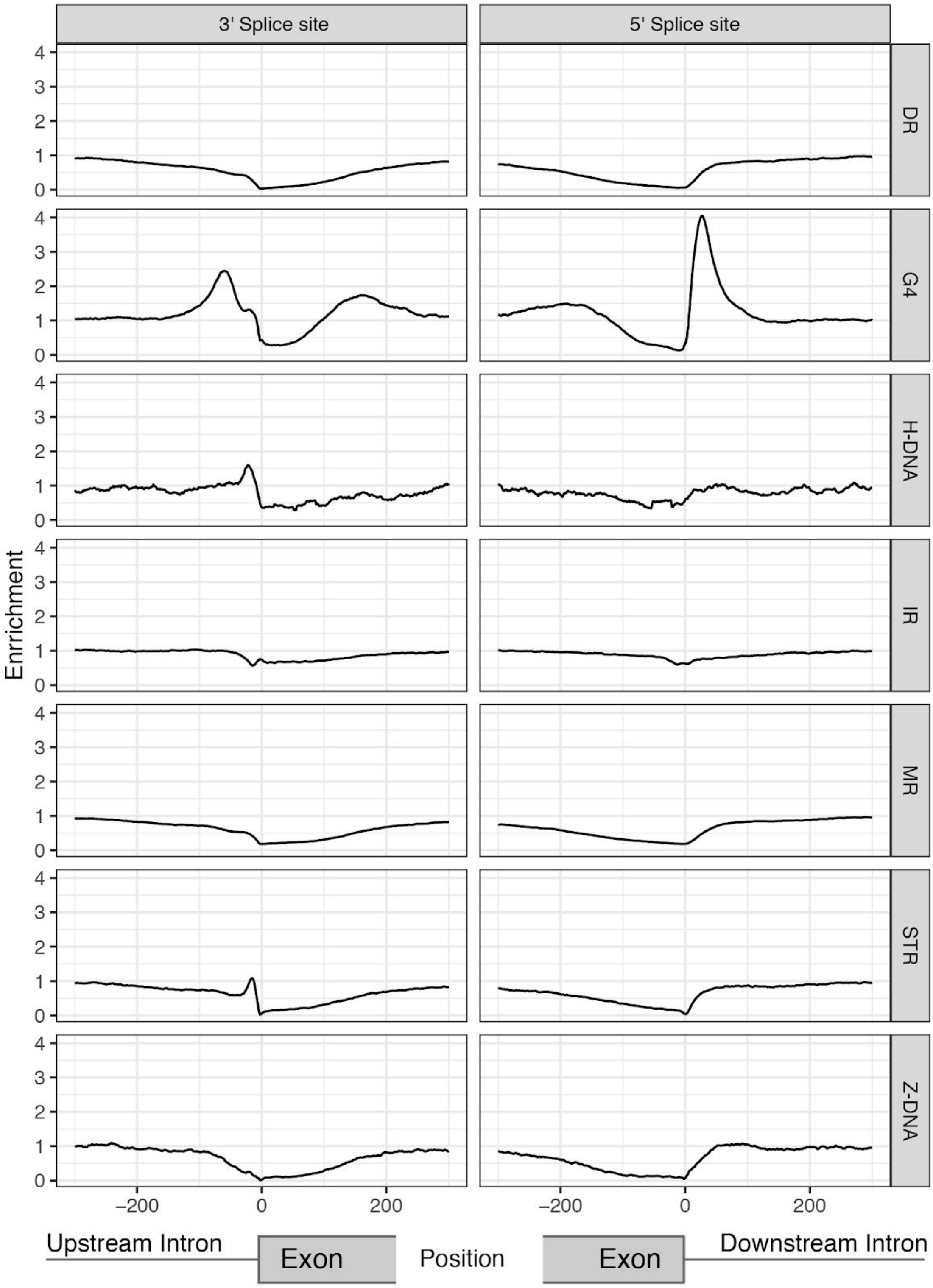

Figure 4.1: Landscape of non-B DNA motifs across human splice sites. Distribution of non-B DNA motifs relative to splice sites. Seven non-B DNA motifs are shown, namely direct repeats (DRs), G-quadruplexes (G4s), H-DNA, inverted repeats (IRs), mirror repeats (MRs), short tandem repeats (STRs) and Z-DNA. Enrichment was calculated as the occurrences of a non-B DNA motif at a given position over the median number of occurrences of that motif across a 1kB window each side from the splice site.

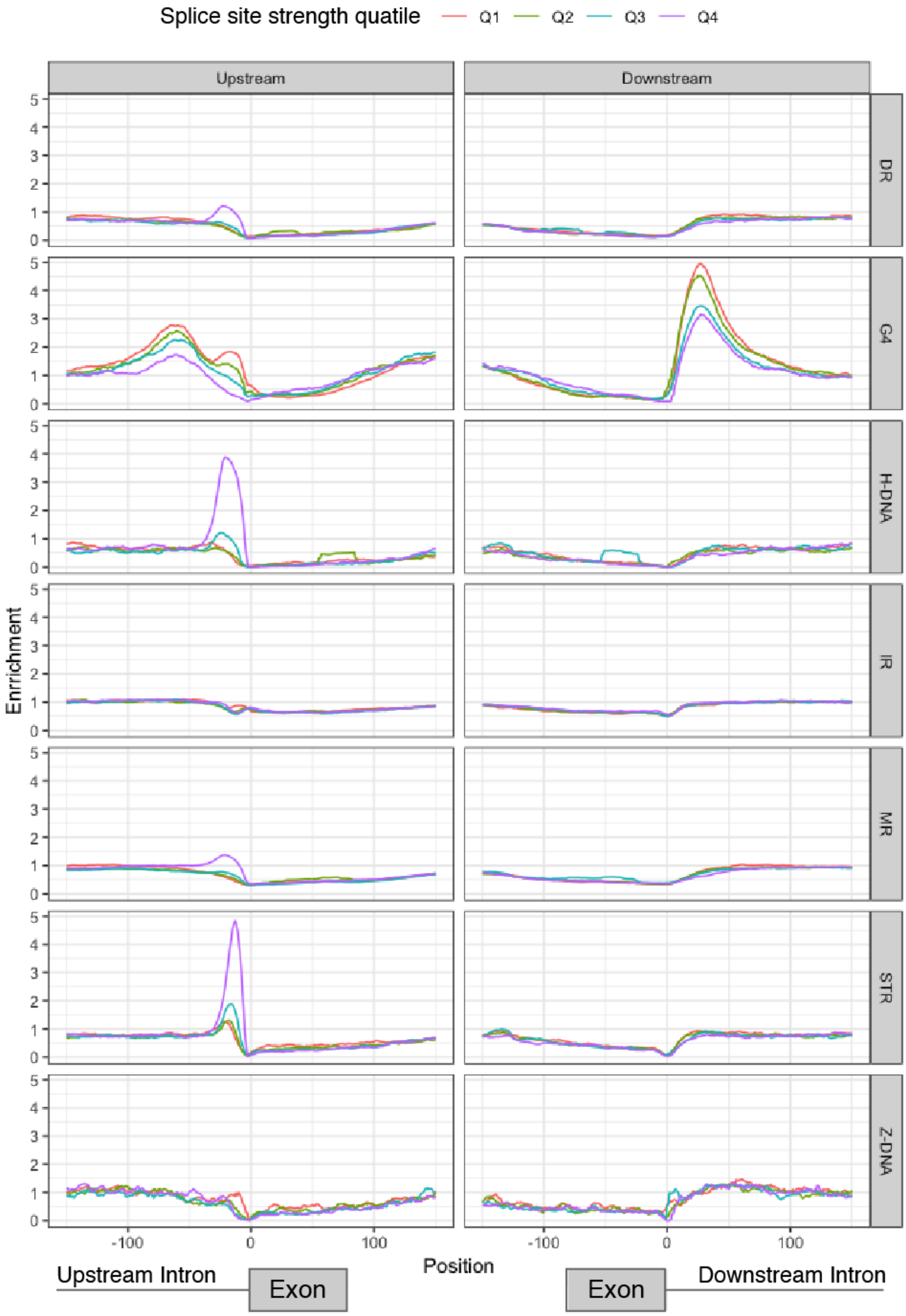

Figure 4.2: Non-B DNA motif enrichment varies with splice strength. Enrichment of non-B DNA motifs across splice site strength quartiles. Seven non-B DNA motifs are shown, namely direct repeats (DRs), G-quadruplexes (G4s), H-DNA, inverted repeats (IRs), mirror repeats (MRs), short tandem repeats (STRs) and Z-DNA.

4.2.2 G4 enrichment analyses of splice sites

Since G4s had the strongest enrichment and they were particularly enriched at weak splice sites (Fig 4.1, 4.2), we investigated the distribution of G4s across splice sites in further detail. To control for the effect of the nucleotide composition of splice sites in the distribution of the GC-rich G4s, we shuffled the 100 nt window each side of the splice site while controlling for dinucleotide content. Comparing the observed frequency to the median from 1,000 permutations we observed a corrected 2.53-fold and 2.73-fold enrichment for the frequency of G4s at the 3′ss and 5′ss, respectively (p-value<0.001 in both 3′ss and 5′ss), indicating that the G4 patterns are not driven by the sequence composition of splice sites. Since G4 motif enrichment was highest across intronic regions that are proximal to splice junctions, we count the number of splice junctions that have at least a G4 motif in close proximity. Within 100 nt of each splice junction we identified 19,987 and 20,088 G4s at the 3′ss and 5′ss, respectively. In total, 31% of human genes contain a G4 motif near at least one splice site within a distance of 100 bp. G4 motifs were found within 100 nt for 8.79% and 8.83% of the 3′ss and 5′ss, respectively. The reported G4 motif frequencies are likely a conservative estimate since we do not take into account intermolecular G4s or G4s that do not adhere to the consensus motif (G□3N1-7G□3N1-7G□3N1-7G□3), (Huppert and Balasubramanian, 2007; Kikin et al., 2006; Varizhuk et al., 2017).

Since G4 formation and template DNA promote polymerase stalling, polymerase stop assays have been implemented to detect G4 formation *in vitro* (Weitzmann et al., 1996). Moreover, since polymerase stalling was found to affect base calling, Chambers and collaborators were able to develop a genome-wide DNA G4 formation assay based on high-throughput DNA sequencing, which they called G4-seq (Chambers et al., 2015). During this assay G4 are stabilised *in vitro* using K+ or pyridostatin (PDS), and G4 formation is inferred by the detection of mismatches induced by base calling errors. Improved versions of the G4-seq protocol have been used to generate maps of G4 formatting sequences across different species, demonstrating a strong enrichment at gene promoter regions and 5′ UTR for human, mouse and *Trypanosoma* (Marsico et al., 2019b). Thus, we analysed these publicly available G4-seq data to corroborate our findings regarding G4 motif enrichments around splice sites.

We first measured the distribution of G4s relative to the splice sites for HEK-293T cells (a human cell line) in Pyridostatin (PDS) and K$^+$ treatments from (Marsico et al., 2019b). In both conditions, we observed an enrichment of G4-seq peaks relative to the 3′ss and 5′ss, but with a more pronounced G4 enrichment in PDS treatment compared to K$^+$ treatment (Fig 4.3a). The majority of G4 positions derived from G4-seq peaks in K$^+$ and PDS treatments did not overlap consensus G4 motifs (Fig 4.3b), which could be explained by the wider range of G4 structures that can be detected though G4-seq, such as noncanonical long loop and bulged structures that cannot be detected with our conservetive G4 motif definition (Chambers et al., 2015). From all G4 motifs that we predicted *in silico*, at least 66.31% and 66.27% were detected at G4-seq experiments (under K$^+$ and PDS condition) for the 3′ and 5′ splice sites respectively. We also found 20.88% and 21.05% of overlapping peaks between G4-seq experiments for 3′ and 5′ splice sites respectively.

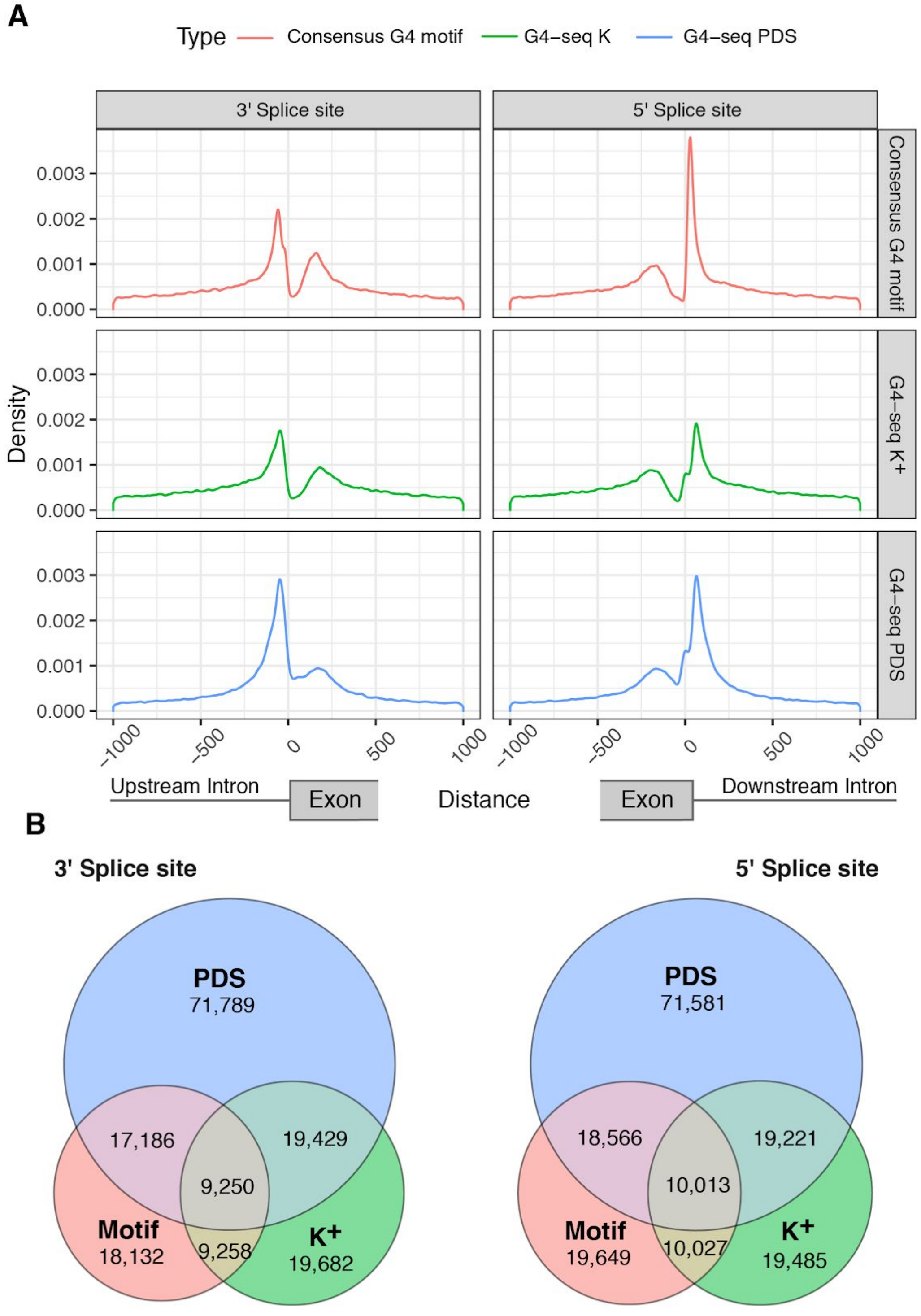

Figure 4.3: Analysis of high-resolution G4-seq data validates in-silico enrichment of G4 motifs A. Distance between nearest G4 motif / G4-seq peak and a splice site separately for 3′ / 5′ splice sites. **B.** Venn diagrams for the occurrences of G4s within 100 nt of the 3′ss (upstream) and 5′ss (downstream) using the consensus G4 motif, the K$^+$ treatment G4-seq derived G4 peaks and the PDS treatment G4-seq derived G4 peaks (Marsico et al., 2019b) and reporting the overlapping G4s between them.

4.2.2.1 G4s are enriched at weak splice sites

Weak splice sites are highly involved in alternative splicing and often contain additional regulatory elements (Erkelenz et al., 2018; Parada et al., 2014; Sibley et al., 2016). To explore the distribution of G4s across weak and strong splice sites, we calculated a splicing strength score for all internal exons based on splice site position weight matrices (Parada et al., 2014; Sheth et al., 2006). We grouped splice sites into four quantiles based on the splicing strength scores, and explored the enrichment levels of G4-seq peaks for each quantile separately. We found an inverse relationship between the calculated splicing strength score and G4 enrichment, with the weakest splice sites having the highest enrichment of G4s both at the 3′ss and the 5′ss with 2.77-fold and 4.95-fold enrichment, respectively (Fig 4.4a-b). For both mouse and human, the splicing strength scores for splice junctions with a G4 are significantly lower than for splice junctions without a G4 (Mann-Whitney U, p-value<0.001).

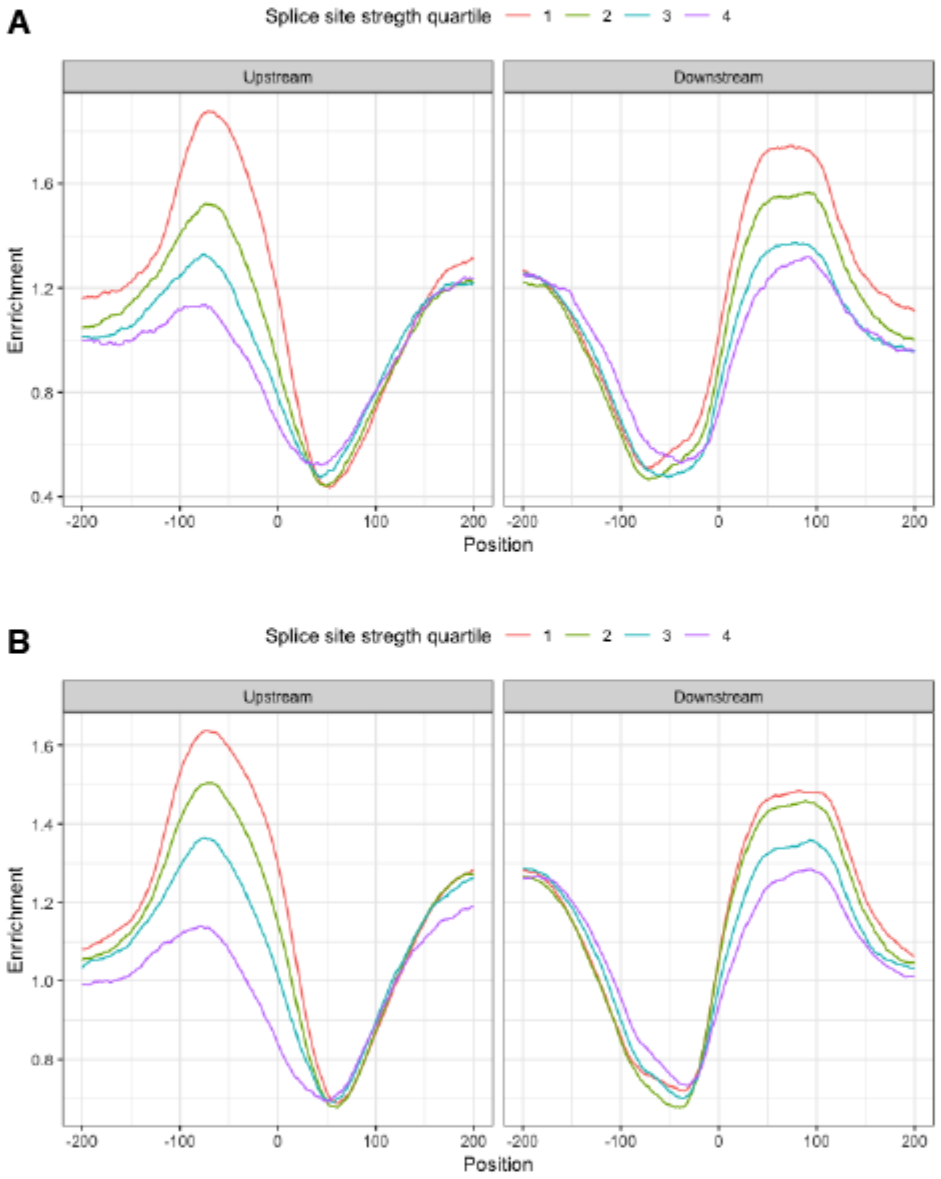

Figure 4.4: Splice site strength and distribution of G4 motifs at splicing sites. G4s display a stronger enrichment at weaker splice sites. A. Distribution of G4 peaks derived from G4-seq with K$^+$ treatment from (Marsico et al., 2019b) at splicing sites at 3′ss (upstream) and 5′ss (downstream) and the association with splicing strength. B. Distribution of G4 peaks derived from G4-seq with PDS treatment from (Marsico et al., 2019b) at 3′ss (upstream) and 5′ss (downstream) and the association with splicing strength.

4.2.2.2 G4s are preferentially found on the non-template strand

Since G4s are strand specific, we oriented each instance relative to the direction of transcription. Thus, we considered G4s found at the template (non-coding) and non-template (coding) strands separately and found them statistically enriched on the non-template strand (Binomial tests, p-value<0.001 at 3′ss and 5′ss). Moreover, G4s were enriched at both strands relative to flanking sequences (Fig 4.5). At 3′ss the enrichment was 3.01-fold and 2.78-fold enrichment scores at the non-template and template strands, respectively. At the 5′ss the difference between the strands was larger with 5.56-fold and 2.38-fold at the non-template and template strands, respectively. Therefore, there was an asymmetric enrichment between the template and non-template strands at the 5′ss, but only a weak asymmetry at the 3′ss.

We also investigated if there was a strand asymmetry when considering the splicing strength scores. Indeed, we found a bias in the splicing strength scores dependent on the strand orientation of G4s (Mann-Whitney U, 3′ss p-value<0.05, 5′ss p-value<0.001). At the 3′ss the enrichment for splice junctions with the weakest splicing strength scores at the template and non-template strand was 3.90-fold and 3.66-fold, respectively. By contrast, we observed a 6.76-fold enrichment for G4s at the 5′ss at the non-template strand, but only a 3.66-fold enrichment on the template strand at the splice junctions with the weakest splicing strength scores (Fig 4.5).

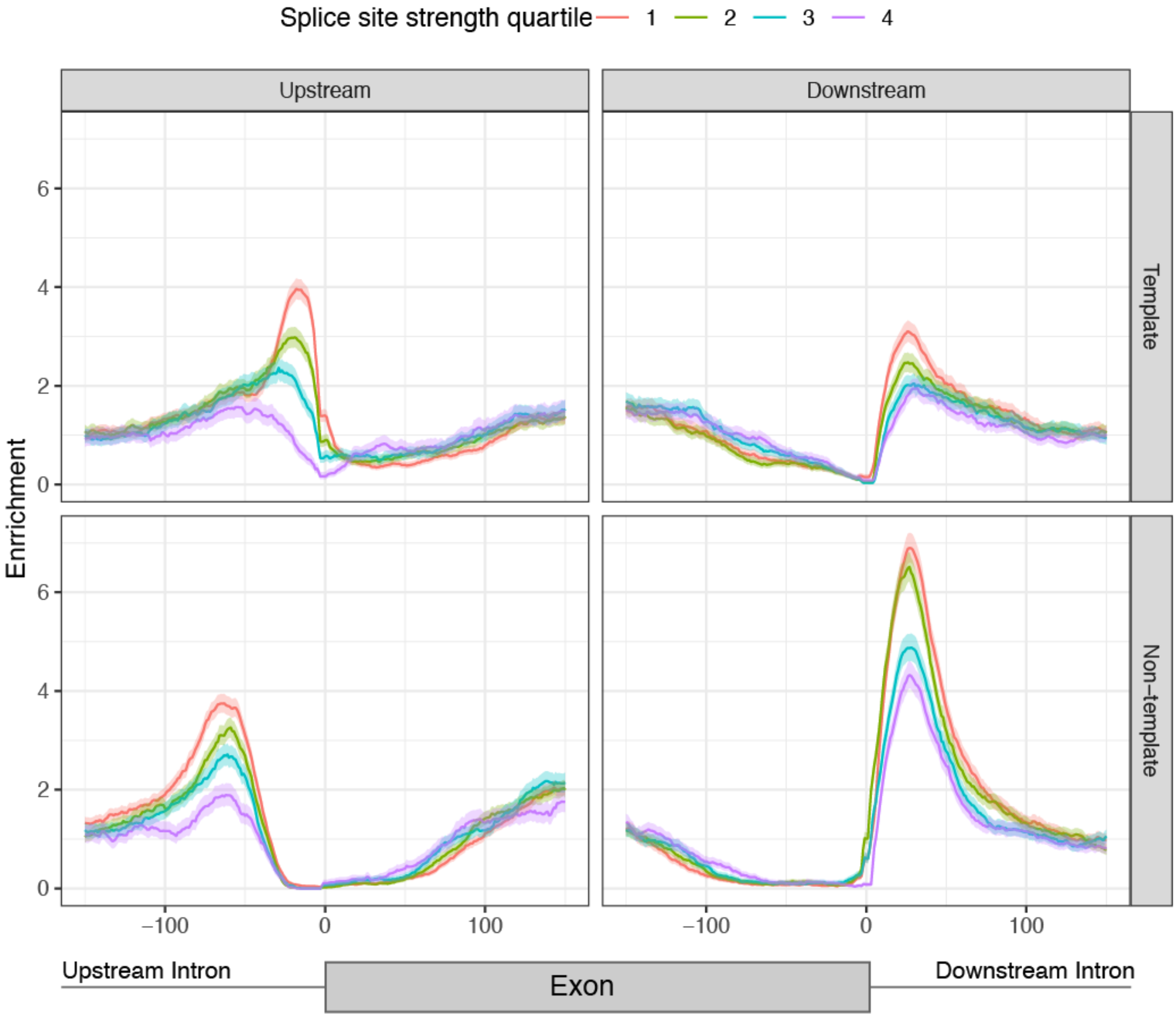

Figure 4.5: Characterisation of G4 motifs across splicing junctions. A. G4 enrichment for template and non-template strands and stratified by the splicing strength scores of the adjacent splice site. The splicing strength scores for splice junctions with a G4 are significantly lower than for splice junctions without a G4 (Mann-Whitney U, p-value<0.001). The splicing strength score bias was found to be dependent on the strand orientation of G4s for splice sites with G4s within 100 nt away (Mann-Whitney U, 3′ss p-value<0.05, 5′ss p-value<0.001).

4.2.2.2.1 Replicated effects are found in independent G4-seq experiments

In order to corroborate the observed G4 motif and G4-seq peak enrichment (Marsico et al., 2019b) patterns, we used additional G4-seq data produced independently. Chambers and collaborators performed a G4-seq experiment in primary human B lymphocytes (NA18507) under Na^{+}-K^{+} or Na^{+}-PDS conditions (Chambers et al., 2015), both of which promote G4 formation. Even though both Chambers et. al and Marsico et al. data were performed using different cell lines and different ionic solution as G4 stabilized treatments, both G4-seq data show similar enrichment patterns across splice site strength quartiles, displaying the same inverse relationship between splice strength quartile and enrichment (Fig 4.6) (Mann-Whitney U, p-values<0.001).

For both the PDS and K^{+} treatments we find that a substantial fraction of the genome is affected, with 31.72% and 10.25% of splice junctions having a G4 within 100 nt intronic window. In addition, 67% and 35% of human genes contain a G4-seq peak from PDS and K^{+} treatments within 100 nt of a splice junction, supporting our earlier observations using the consensus G4 motif. As a result of these findings, we conclude that G4s are a pervasive feature near splicing junctions.

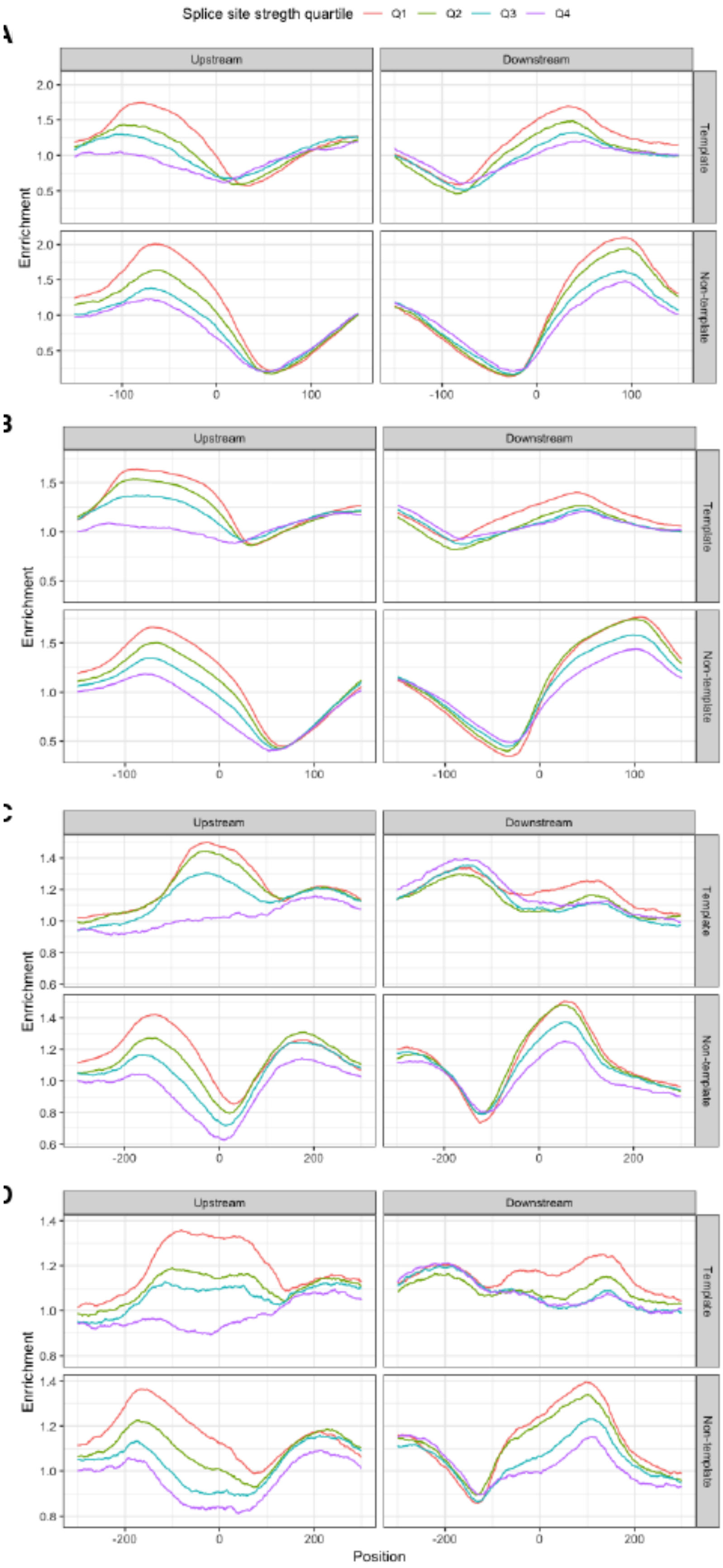

Figure 4.6: G4 enrichment patterns are consistent across different G4-seq experiments. A. Distribution of G4 peaks derived from G4-seq in presence of PDS at template and non-template strands and the association with splicing strength. **B.** Distribution of G4 peaks derived from G4-seq in presence of K$^+$ at template and non-template strands and the association with splicing strength. **C.** Distribution of G4 peaks derived from G4-seq in presence of Na$^+$-PDS at template and non-template strands and the association with splicing strength. **D.** Distribution of G4 peaks derived from G4-seq in presence of Na$^+$-K$^+$ at template and non-template strands and the association with splicing strength. **A-B.** Are based on recent G4-seq data from (Marsico et al., 2019b) with higher resolution than **C-D** (Chambers et al., 2015). But the same trend is shown, validating the observation across two independent sets of experiments.

4.2.2.2.2 Template and non-template G4 enrichment patterns across gene body

Since the enrichment of G4s around promoters have been reported to be biased towards the non-template DNA strand (Eddy and Maizels, 2008), we hypothesized that strand asymmetries could also be found for the occurrences of G4 around splice sites. In order to study the strand asymmetries patterns of G4 motifs around splice sites across the gene body, for each gene with nine or more exons, we separated the exons of its longest transcript into nine groups: the first four exons, the last four exons and the remaining middle exons. For each of the groups, we calculated G4 motif enrichment at splice sites across both template and non-template strands. We found a pervasive enrichment of G4 motifs across the gene body, however non-template G4 motifs are consistently more highly enriched at 3′ss than template strands (Fig 4.7). In the case of 5’ss, enrichment differences between template and non-template strands are smaller, but it gets bigger towards the gene ends. These findings provide evidence for widespread variation in the topography of G4s in splice junctions; these include the frequency of G4s in the exons and introns flanking the splice site, biases regarding the strand preference, the distance from the splice site and the positioning across the gene body.

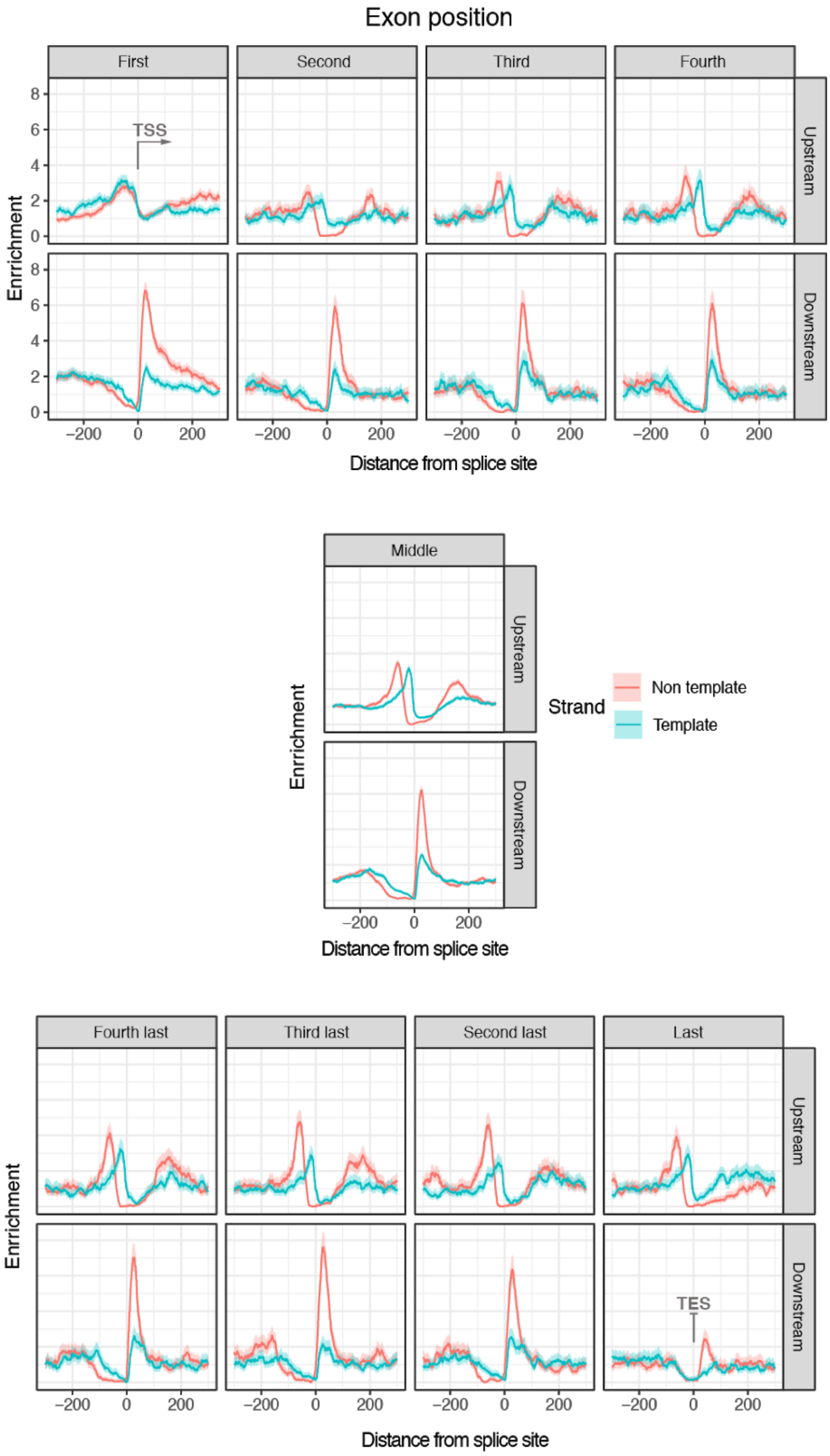

Figure 4.7: Template and non-template splice site G4 enrichment across gene body. G4 motif enrichment relative to splice sites across exons in the gene body for template and non-template strands. Exons were classified according to their gene position. Each panel shows G4 motif enchiment across their upstream or downstream regions of exon across the gene body. The start and end exonic coordinates are centered at 0 x-axis coordinate and they correspond to 3′ss or 5′ss, respectively, except for the first or last exon where transcriptional start site (TSS) and transcriptional end site (TES) are indicated.1.2.2.3 Longer G-runs exhibit higher enrichment around splice sites.

An intramolecular G4 is usually a representation of four or more consecutive G-runs. Yet, fewer consecutive G-runs can also result in G4 formation when they are complemented by additional G-runs from another DNA or RNA molecule (Bhattacharyya et al., 2016; Nasiri et al., 2016). We found minimal to no enrichment for single G-runs at both 5′ss and 3′ss (Fig 4.8). However, for two and three G-runs we observed a 1.39-fold and a 2.10-fold enrichment at the 3′ss and a 1.67-fold and a 2.47-fold enrichment at the 5′ss, which may implicate intermolecular G4s in splice sites. The highest enrichment was observed for four to six G-runs, indicating that intramolecular G4 motifs are more enriched at splice sites than their intermolecular counterparts, in accordance with our earlier findings.

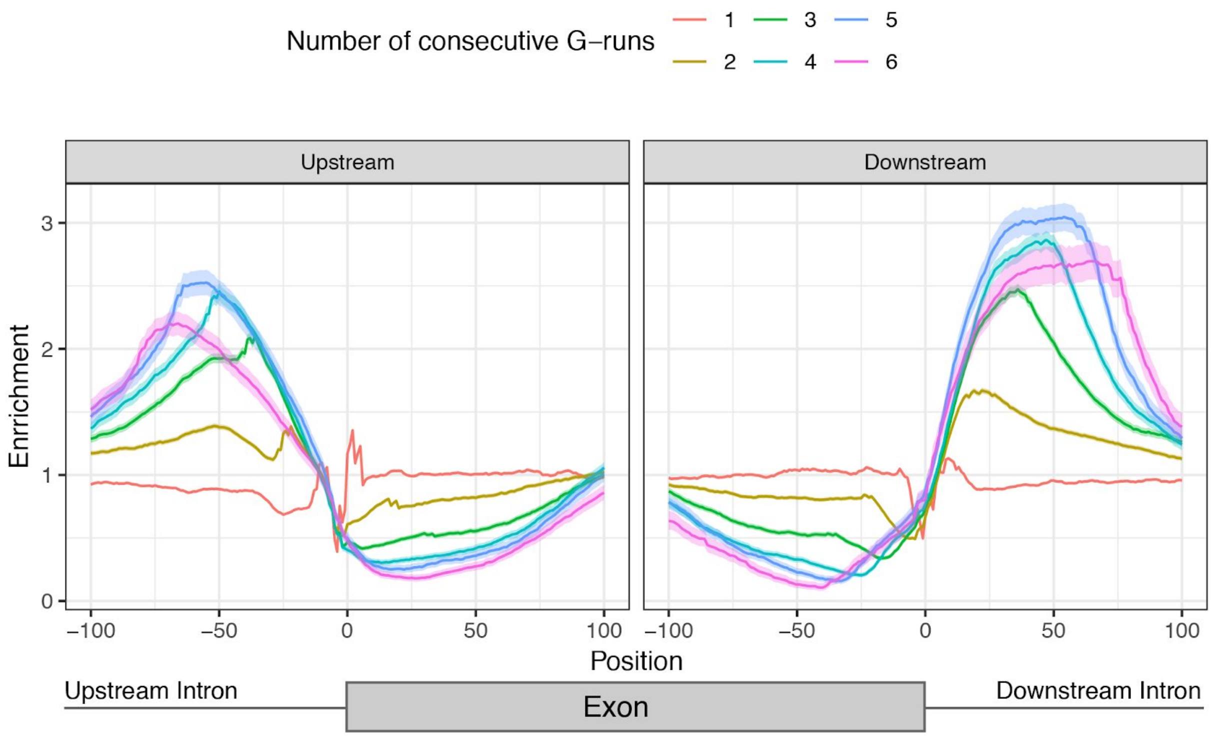

Figure 4.8: Enrichment of G4 of different G-run lengths. Number of consecutive G-runs (G stretches of at least 3 nt separated by 1-7 linker nucleotides) and relative enrichment at the splicing junction. The error bands in A-B represent 0.95 confidence intervals from the binomial error.

4.2.3 Gene architectural features associated with G4-exons

4.2.3.1 G4s are enriched for short introns

The length of introns in metazoans can vary across four orders of magnitude (Sakharkar et al., 2004). We hypothesized that the enrichment patterns of G4s at introns proximal to splicing sites would be associated with intron length. We compared the intron length of splice sites that had a G4 motif within 100 bps in the direction of the intron to the ones that did not have this motif. Consistent with our hypothesis, we found that introns with a G4 at the 3′ss had a median length of 701 nt while introns without a G4 had a median length of 1,618 nt (Figure 4.9a), (Mann Whitney U, p-value<0.001). Similarly, at the 5′ss, introns with a G4 had a median length of 379 nt, whereas introns without a G4 had a median length of 1,629 nt (Mann Whitney U, p-value<0.001). Interestingly, introns in the range of ~45-85 bps were the most enriched for G4s for both the 3′ss and the 5′ss. Moreover, the enrichment of introns in G4s declined rapidly with increased intron length, indicating that they are preferentially found in the subset of short introns (Fig 4.9b-c, Kolmogorov-Smirnov test p-value<0.001). We also investigated the association between splicing strength score and intron length at sites with G4s in the 3′ss and 5′ss and found that the highest enrichment for G4s was in short introns with weak splice site strength (Fig 4.9d-e).

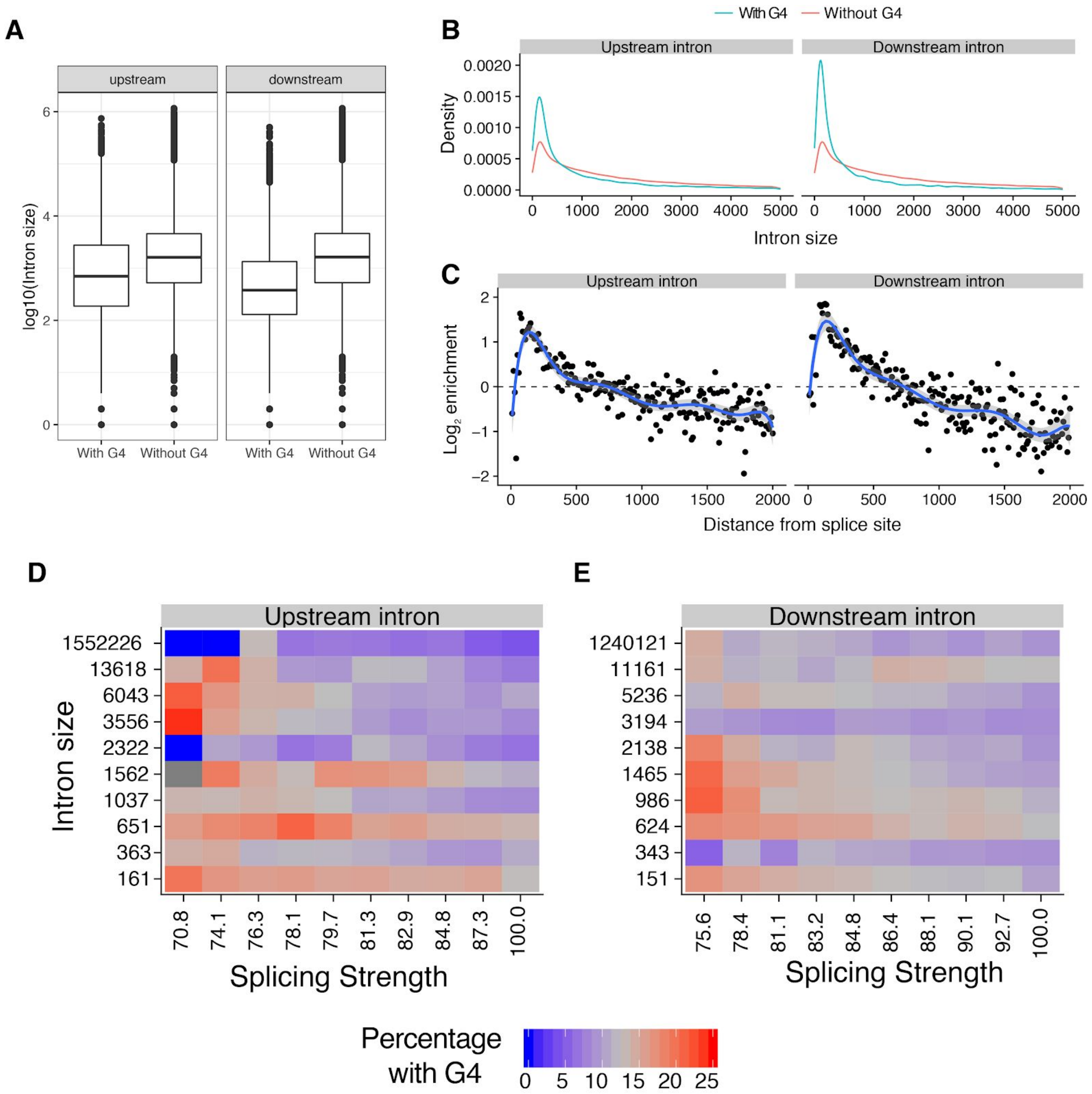

Figure 4.9: G4s are enriched in short introns A. Intron size of the upstream and downstream introns was calculated for groups with or without a G4 within 100 bps of the splice site (Mann-Whitney U p-value <0.001 for both upstream and downstream introns). **B.** Length density distribution of introns upstream and downstream from exons that are flanked by G4s or not (Kolmogorov-Smirnov test p-value<0.001). **C.** Abundance enrichment of intron sizes at upstream and downstream splice sites flanked by G4s. A bin size of 10 bps was used with the blue line representing an eighth degree polynomial model. **D-E.** Heatmap for the relationship between splicing strength score, intron length and G4 presence in a local window of 100 nt within the splice site for the upstream and downstream introns. Red color represents high proportion of splice site regions with G4s, whereas blue color represents depletion of G4s.

4.2.3.1.1 GC-content controlled associations of G4 motifs and intron size

Since short introns are more GC-rich than long introns (Lim and Burge, 2001), we wanted to compare selected groups of introns that have close GC-content distribution (Fig 4.10a). Thus, we divided the intron into two populations, short (<500 nt) and long (>500nt). Then, for different long intron size interval groups, we select groups of short and long introns that have the minimum GC-content difference. By doing this, differences in the fraction of introns containing a G4 in the splice site vicinity (within 100 nt) cannot be attributed to GC-content differences. We found a significantly higher fraction of intron splice sites with a G4 in their vicinity (within 100 nt) for long introns in comparison with short introns, suggesting an inverse enrichment direction when GC-content is controlled (Fig 4.10b).

Moreover, intron size comparisons between template and non-template strands are also not subject to GC-content effect since both strands have the same GC contribution. Thus, to further investigate the relationship regarding the intron length, we separated G4s identified using the consensus motif into non-template and template for both the 5′ss and the 3′ss. At the 3′ss introns showed small but significant differences in length if a G4 was at the non-template or the template strand with medians of 736 nt and 621 nt, respectively (Mann-Whitney U, p-value<1e-21). However, if a G4 was at the non-template strand at the 5′ss the median intron length was 267 bp, whereas if the G4 was at the template strand the median intron length was 539 nt (Mann-Whitney U, p-value<1e-16), displaying more aggravated differences in intron length. Therefore, we conclude that the highest enrichment is found for short introns, on the non-template strand, downstream of the 5′ss.

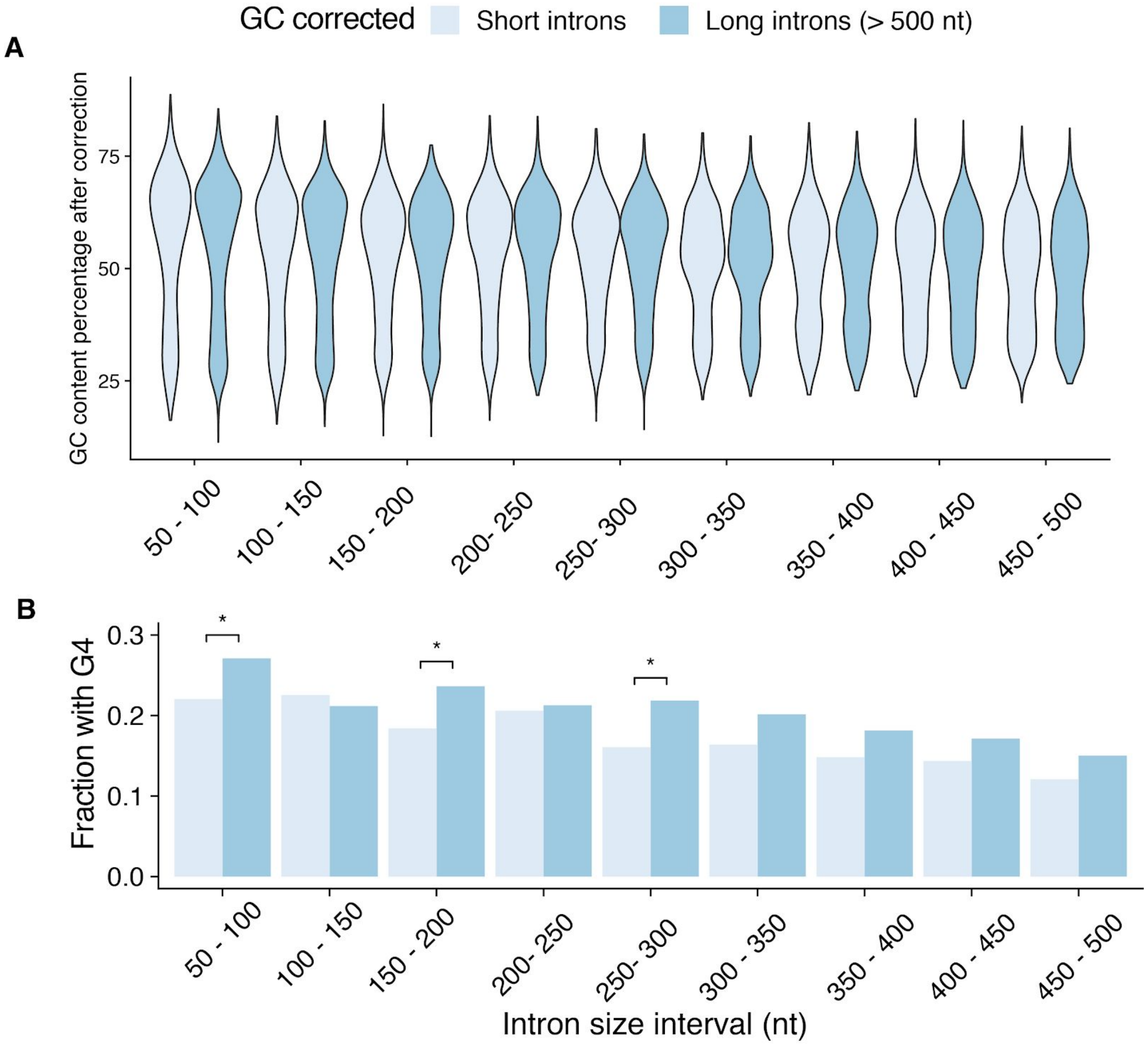

Figure 4.10: G4 enrichment at short intron splice sites is driven by GC-content **A.** GC content distribution across selected groups of short and long introns. Intron size interval refers to the size of small introns. Long introns were defined as introns > 500 bp. **B.** Fraction of splice sites with a G4, controlling for GC content between long and short introns. We use Chi-squared test to evaluate significant differences between short and long introns (* denotes p-values<0.05 after multiple testing corrections).

4.2.3.2 G4 are not enriched in microexons

We also investigated if there is an association between G4s near splice sites and exon length. We do not find a significant association between G4s and exon length at the 3′ss (median exon length without G4s: 124 bp, median exon length with G4s: 123 bp, p-value>0.05, Mann-Whitney U), but we find a significant association for smaller exons near the 5′ss, albeit with a very small magnitude (median exon length without G4s: 127 bp, median exon length with G4: 123 bp, p-value<0.001, Mann-Whitney U). Furthermore, we explored if microexons, defined as exons <30 nt long (Irimia et al., 2014), (Li et al., 2015) had an enrichment for G4s at their splice sites relative to other exons. However, we could not find a higher density of G4s at the introns flanking microexons compared to other exons.

4.2.4 Abundance of G4s at splice sites has emerged during vertebrate evolution

Alternative splicing is a pivotal step of eukaryotic mRNA processing. To understand to what extent splice site regulation by G4s is conserved we considered eleven eukaryotes: *Homo sapiens* (human), *Mus musculus* (mouse), *Sus scrofa* (pig), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Caenorhabditis elegans* (nematode) *D. melanogaster* (fruit fly), *Xenopus tropicalis* (frog), *Anolis carolinensis* (lizard), *Saccharomyces cerevisiae* (yeast) and *Arabidopsis thaliana* (flowering plant). *S. cerevisiae* was excluded from further analysis since we could not find any G4s at splice sites and G4s were rare with only 39 occurrences genome-wide. Interestingly, we found that the enrichment pattern of G4 motifs at splice sites was restricted to a subset of vertebrate species, with minimal or no enrichment in fruit fly, *Arabidopsis* and *C. elegans* (Fig 4.11). We observed strong enrichment in chicken, pig, human and mouse, while lizard displayed limited enrichment levels. Surprisingly, *X. tropicalis* and *D. rerio* displayed relative depletion. This suggests that alternative splicing regulation by G4s is found is restricted to mammals and birds, but absent in plants, other tetrapods or fish. However more comprehensive evolutionary analysis are needed to completely discard the presence of G4 enrichment in other organisms.

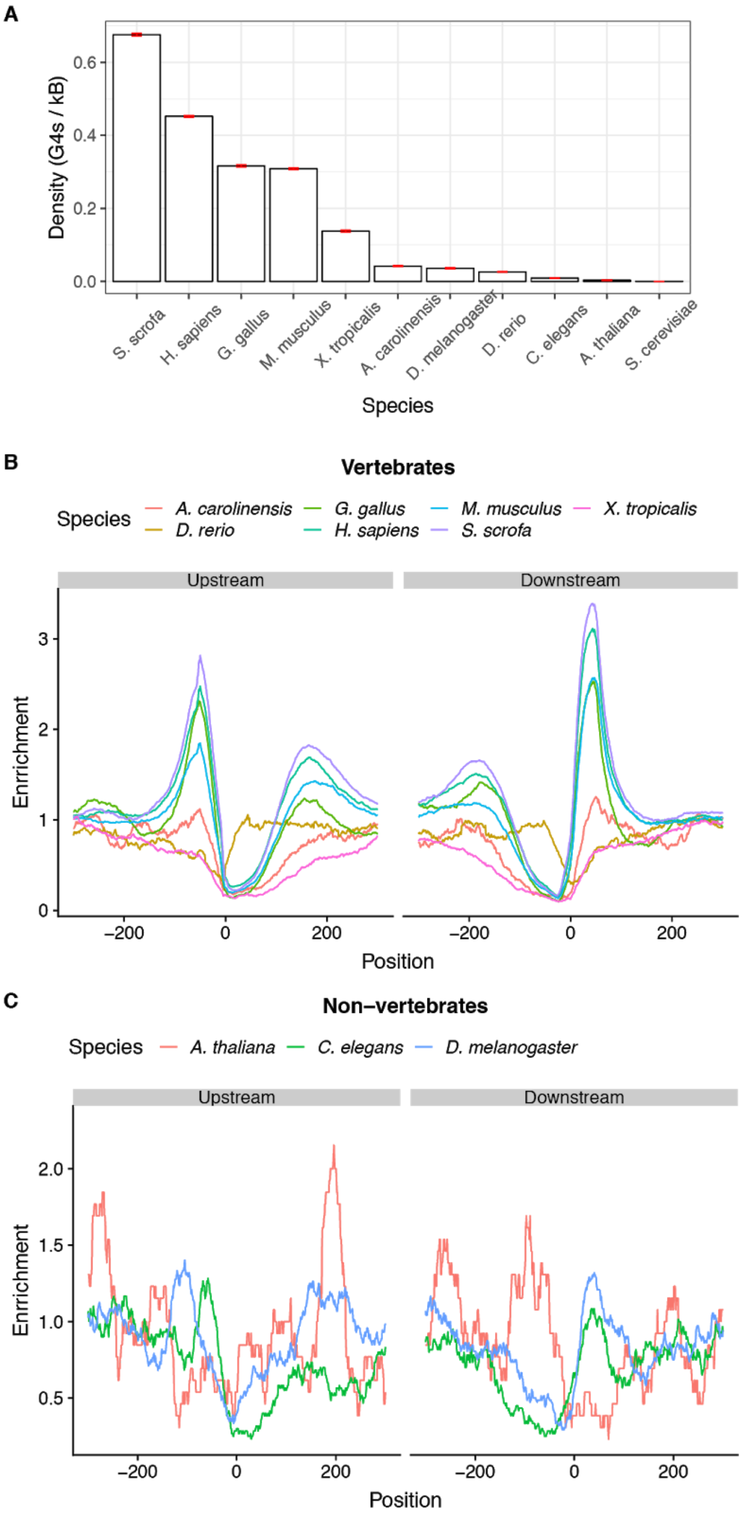

Figure 4.11: G4 motifs are enriched in a subset of vertebrates. A. Density of G4 motifs in a 100 nt window each side across all 5′ / 3′ splice sites of each species. Error bars indicate standard deviation from 1,000-fold bootstrapping with replacement. **B-C.** Enrichment of G4 motifs at splice sites for seven vertebrate (B) and three invertebrate (C) species using the consensus G4 motifs.

Additional support for this conclusion comes from our analysis of G4-seq derived G4 maps generated in PDS and K^+ conditions. These maps are available for multiple model organisms, including three vertebrates (human, mouse and zebrafish) and four non-vertebrate species (nematode, fruit fly, arabidopsis and yeast). Consistent with the analysis based on the primary sequence, we find an acute enrichment of G4s at the 5′ss and 3′ss only in humans and mouse. In particular, we could not find any G4s in the vicinity of splicing junctions for *S. cerevisiae*, there was no enrichment for *D. melanogaster* and *D. rerio*, while we observed a depletion in *A. thaliana* (Fig 4.13).

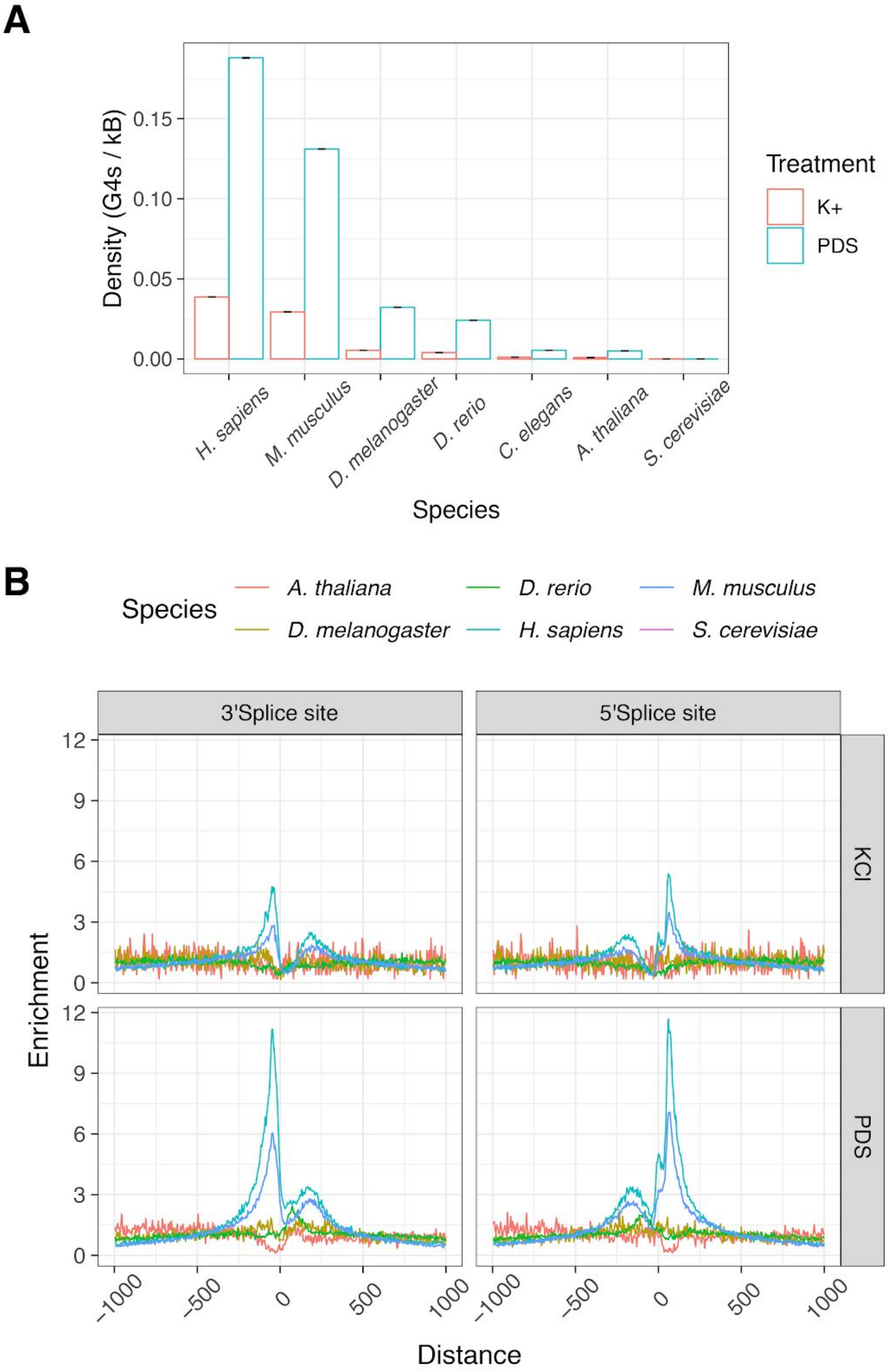

Figure 4.12: Cross specie G4-seq analyses validate findings *in-silico*. A. Enrichment of G4-seq derived G4s at splicing sites at 100 nt splicing site windows in PDS and K$^+$ treatments. Error bars indicate standard deviation from 1,000-fold bootstrapping with replacement. **B.** Enrichment of G4s at: 5′ / 3′ splice sites across six species for PDS and K$^+$ treatments.

4.3 Materials and Methods.

4.3.1 Genome and gene annotations processing

We obtained genome assemblies from the UCSC Genome Browser FTP server for eleven organisms: *Homo Sapiens* (hg19), mouse (mm10), *Saccharomyces cerevisiae* (sacCer3), chicken (galGal5), *Drosophila melanogaster* (dm6), zebrafish (danRer11), *Xenopus tropicalis* (xenTro9), *Anolis carolinensis* (anoCar2), *Arabidopsis thaliana* (Tair10) and *Caenorhabditis elegans* (ce10) reference genomes.

We downloaded the Ensembl gene annotation files for the associated genomes from UCSC Table Browser as BED files for each species (Karolchik et al., 2003). Using in-house python scripts we extracted the coordinates of internal exons flanked by canonical splice sites (GT-AG introns) for every species. To calculate the splicing strength scores, we used publicly available positional frequency matrices from the SpliceRack database (Sheth et al., 2006) and previously developed scripts used before for the same purpose (Parada et al., 2014). Splice sites were grouped into quartiles based on their splicing strength score for the downstream analyses to study the distribution of non-B DNA motifs and in particular G4 motifs. The confidence intervals were calculated using "binconf" command from "Hmisc" package in R with default parameters. Mann-Whitney U tests were performed at 100 nt each side in the upstream splice site and at the downstream splice site to compare the splicing strength scores of sites with and without G4s.

4.3.2 Genomic datasets.

4.3.2.1 Non-B DNA motifs.

Identification of each non-B DNA motif was performed using the genome-wide maps in humans and mice provided by (Cer et al., 2013) and processed as described in (Georgakopoulos-Soares et al., 2018). We focused on seven non-B DNA motifs;

inverted repeats, mirror repeats, H-DNA which forms at a subset of mirror repeats with high AG content, G4s, Z-DNA which forms at non-AT alternating purine pyrimidine stretches, short tandem repeats and direct repeats.

Regular expressions were employed to identify genome-wide consecutive G-runs across the human genome, interspersed with loops of up to 7 bps. In total, one to six consecutive G-runs were searched. For each species we generated the genome-wide G4 maps using a regular expression of the consensus G4 motif (G□3N1-7G□3N1-7G□3N1-7G□3). Orientation of G4s and G-runs was performed with respect to template and non-template strands to calculate strand asymmetries at genic regions as previously described for polyN motifs (N being Gs, Cs, Ts and As) in (Georgakopoulos-Soares et al.).

Permuted windows of 100 nt each side of each splice junction were generated using ushuffle (Jiang et al., 2008) correcting for dinucleotide content. The fold enrichment for G4s was calculated as the ratio of the number of motifs found in the real sequences and the median of 1,000 permutations of the set of all real sequences. The corrected enrichment of G4s at 3′ss and 5′ss was calculated as the ratio of the real enrichment of G4s over the background enrichment of G4s at shuffled splice site windows.

To investigate the relationship between non-B DNA motifs or G4-seq peaks and splice sites we generated local windows around the splice sites and measured the distribution of each non-B DNA motif or G4-seq dataset across the window. The enrichment was calculated as the number of occurrences at a position over the median number of occurrences across the window. Regardless of the window size shown in figures, the enrichment was calculated over a window of 1kB. The same approach was used to calculate the enrichment of G4s at splice sites across different species.

The density of G4 consensus motifs or G4-seq derived peaks at local windows was calculated as the number of occurrences of the motif or the peak over the total number of base pairs examined .

4.3.3 G4-seq data

1.4.3.1 G4-seq BedGraph data were obtained from GEO accession code GSE63874 (Chambers et al. 2015) for the human genome and analyzed with bedtools closest command to identify the closest G4 to splice sites and to calculate the distance. The analysis was performed separately for Na+-K+ and Na+-PDS conditions and it was compared to the distribution obtained from the G4 consensus motif. G4-seq BedGraph data for six species, human, mouse, D. melanogaster, C. elegans, A. thaliana and yeast, were obtained from GEO accession code GSE110582 (Marsico et al. 2019) and analyzed using the same genome annotations as those used for the generation of each G4-seq dataset.

Coordinates for internal exons flanked by canonical splice sites (GT-AG introns) were extracted for each species using the Ensembl annotation versions described in (Marsico et al., 2019b) using custom python scripts.

4.3.4 Relationship between G4s and exon / intron length

Introns and exons were grouped based on the presence or absence of G4s within 100 nt each side of the 5′ss and 3′ss and further subdivided into those containing a G4 on the template or on the non-template strand, separately for the 3′ss and the 5′ss. For each of the eight groups we calculated the median length of the intron or exon in a group and performed Mann-Whitney U tests to calculate the significance of the association between length of exons / introns and G4 presence. The R function stat_density was used to plot the length distribution of introns with and without G4s as modelled by a kernel density estimate. Abundance enrichment of intron length in 3′ / 5′ splice sites in relationship with presence of G4s was generated in R using the function geom_smooth in an eighth grade model. Correction of GC content in introns with different length was performed by grouping introns into small introns (<500nt) and large introns (>500nt). Then we calculated the GC content for both groups and for each short intron we selected a long intron with a close GC content value, in such a way that GC distribution across short and long introns groups were nearly identical.

4.3.4.1 G4s and relationship to exon number

For the longest transcript of each gene with nine or more exons we separated exons into 9 groups, the first four exons, the last four exons and the remaining middle exons. To compare the frequency of G4s in splice junctions across the gene body we calculated the distribution of G4s in each exon group relative to the 5′ / 3′ splice sites (S5c). We also calculated the distribution of G4s in each exon group relative to the 5′ / 3′ splice sites separately for the template and non-template strands.

4.3.4.2 Relationship between G4s, splicing strength score and intron length

We calculated the splicing strength score and intron length for the upstream and downstream intron of each exon. We separated introns and splicing strength scores into deciles and calculated the G4 density at each decile, from which we produced two heatmaps displaying the density of G4s as a function of splicing strength score and intron length for the upstream and downstream introns.

5 Chapter V: Dynamic non-canonical splicing responses to neuronal depolarization stimuli

Collaboration note

Most of the work described in this chapter is currently part of the same manuscript as results in Chapter IV (Georgakopoulos-Soares et al.), which is currently under revision in Nature Communications. Ilias Georgakopoulos-Soares and I had equal contributions to these results; while Ilias Georgakopoulos-Soares quantified non-B DNA motifs across the genome, I processed publicly available RNA-seq data and splice sites. We also collaborated with Hei Yuen Wong from Hong Kong University who performed the validation experiments under the supervision of Chun Kit Kwok. Ilias Georgakopoulos-Soares and I both contributed to the experimental design. I carried out the analyses to generate all the figures presented in this chapter except the ones related to the experimental validation.

5.1 Introduction

Splicing patterns undergo dramatic changes during neuronal development, not only involving conventional splicing events, but also non-canonical splicing events such as microexons. However, neuronal alternative splicing changes are not restricted to development; mature neurons also exhibit dynamic alternative splicing changes in response to neuronal activity.

The effect of neuronal activity over gene expression and alternative splicing is often studied through the induction of neuronal depolarization by introducing high extracellular potassium ion concentrations. These potassium-induced depolarization events trigger the opening of voltage-dependent calcium channels that ultimately increase the intracellular calcium concentration. Different alternative splicing events have been associated with calcium influx triggered by depolarization, among which several are present in genes that encode for ion channels, which have the potential to modulate neuronal electrochemical properties and function (Hermey et al., 2017; Sharma and Lou, 2011). For example, increase of intracellular calcium induces the skipping of exons 5 and 21 of NMDA receptor type 1 (*NMDAR1*) (An and Grabowski, 2007; Han et al., 2005; Lee et al., 2007) and the skipping of the so-called STREX exon present at the *KCNMA1* gene, which encodes for BK (Big Potassium) channel (Xie and Black, 2001). While NMDA exon 5 and STREX exon regulate some of the ion channel electrochemical properties (Rumbaugh et al., 2000; Traynelis et al., 1995, 1998; Vance et al., 2012), NMDA exon 21 encodes for a C-terminal protein domain (C1) that promotes *NMDAR1* retention at the endoplasmic reticulum, preventing it reaching the extracellular membrane (Ehlers et al., 1995; Scott et al., 2001; Standley et al., 2000).

Intracellular calcium increase is a pivotal signalling process that triggers several different cellular responses, many of which are promoted by the activation of calmodulin-dependent protein kinases (CaMKs). Depolarization experiments on excitable pituitary cell line GH3 have shown that skipping of the STREX exon involves the specific activation of CaMK IV (Xie and Black, 2001), which was later described as a mediator of several exon skipping events through the activation of CAMK IV-responsive RNA elements (CaRRE) (Lee et al., 2007; Xie et al., 2005). CaRRE are *cis*-regulatory elements that are bound by hnRNP L, which in its phosphorylated form leads to depolarization-induced splicing regulation (An and Grabowski, 2007). Additional *cis*-regulatory elements have also been found and novel ones may remain to be discovered (Sharma and Lou, 2011).

In addition, the architectural splicing code can play a role in the definition of alternative splicing events associated with changes in ionic concentrations. In the particular case of neuronal microexons, there is already evidence that indicate their high responsiveness to depolarisation stimuli; RNA-seq experiment analyses have demonstrated widespread microexon skipping events after depolarization of primary cultured hippocampal neurons (Quesnel-Vallières et al., 2016). Thus I hypothesized that other non-canonical splicing features could be involved in dynamic alternative splicing changes induced by depolarization stimuli. In this chapter, the research aims are centred on the third research aim of this thesis (see section 1.7). However, since the involvement of G4 formation in depolarization-induced alternative splicing would represent a complete novelty in the field, further analyses and experimental validations were performed to investigate associations of G4 with this alternative splicing pathway.

5.2 Results

5.2.1 Dynamic splicing responses to neuronal depolarization are associated with non-canonical features

Since non-canonical splicing events, such as microexons and circRNAs, are often differentially included in neurons, we wanted to explore how they are associated with dynamic alternative splicing responses to neuronal activity. To this end, we analysed RNA-seq experiments performed on human and mouse stem cell-derived cortical neurons (Hum-ESC^{CORT} and Mus-ESC^{CORT}), mouse developing primary cortical neurons (DIV4 Mus-PRIM^{CORT} and DIV10 Mus-PRIM^{CORT}) and aneuploid Tc1 mouse neurons (DIV10 Mus-Tc1-PRIM^{CORT}) after four hours of depolarization induced by 170mM of KCl and an L-type Ca2+ channel agonist FPL64176 (KCl/FPL) (Qiu et al., 2016). The depolarization induced by KCl/FPL treatment has been reported to lead to a strong and uniform increase in intracellular calcium concentration (>10-fold increase), which is likely to have a significant effect on the CaMK IV splicing pathway (Gaspard et al., 2008; Lee et al., 2007; Xie and Black, 2001).

5.2.1.1 Depolarization triggers genome-wide cassette exon exclusion events that are highly enriched in microexons

We used MicroExonator coupled to Whippet (Sterne-Weiler et al., 2018) to perform an integrative quantitative assessment of alternative splicing for both long exons and microexons. Splicing node quantification analyses were performed by Whippet leading to the identification of a total of 22,344 alternative splicing events, 2,633 of which corresponded to core exon differential inclusion events. Among all splicing node types from Whippet’s contiguous splice graph (CSG) model, the quantification of core exons splice (CE) nodes inclusion is a direct proxy to assess the differential inclusion of cassette exons. Further analysis of CE splicing nodes showed a bias towards cassette exon exclusion after the KCl-induced depolarization stimuli as 2,346 (89.1%) of the differentially spliced CE nodes correspond to exon exclusion events. These results are consistent with previous studies which have demonstrated exon skipping following depolarisation in individual examples (An and Grabowski, 2007; Fiszbein and Kornblihtt, 2017; Lee et al., 2007; Liu et al., 2012; Schor et al., 2009; Xie and Black, 2001), but to the best of my knowledge this provides the first genome-wide analysis demonstrating widespread exon skipping (Fig 5.1a). Interestingly, the alternative cassette exons events that were found to be triggered in response to the KCl treatment are highly enriched in microexons (Fig 5.1b), suggesting their active involvement in dynamic splicing changes in neurons.

5.2.1.2 Dynamic splicing responses to neuronal depolarization are associated with G4s proximal to splice sites

Additionally, Hum-ESC^{CORT} RNA-seq analysis showed a consistent association of G4s to depolarization-induced splicing events. We found an enrichment for differential CE splicing node inclusion[8] events when intronic G4 motifs are found within a 100 nt window from splice sites (Fig 5.1c, chi-squared test with multiple testing correction, p-value<0.001, odds-ratio=1.57). To provide further support for the findings obtained using the consensus G4 motif, we examined the distribution of G4-seq derived peaks in PDS and K$^+$ conditions around splice sites of differentially and non-differentially included exons. As expected, we found a consistent enrichment at the differentially included CE splice nodes (Fig 5.1 c-e). These observations are consistent with replicated analyses performed across the depolarization RNA-seq experiments from mouse cortical neuron samples (Fig 5.2). Moreover, the effect size was larger for the G4 motifs and the G4-seq derived G4 sites in the non-template strand at the 5′ss in comparison to those found at the template strand (chi-square test multiple testing correction, p-value<0.001 when using the consensus G4 motif and for both PDS and K$^+$ G4-seq conditions in human

[8] Herein I use the expression “differential inclusion” to refer to both inclusion and exclusion events, regardless of the bias found towards inclusion or exclusion.

neurons). Taken together, our results suggest that the presence of a G4 near the splice junction of cassette exons is associated with dynamic changes of alternative splicing in response to KCl induced depolarisation, however additional experiments will be required to prove causal effects beyond the correlations reported here.

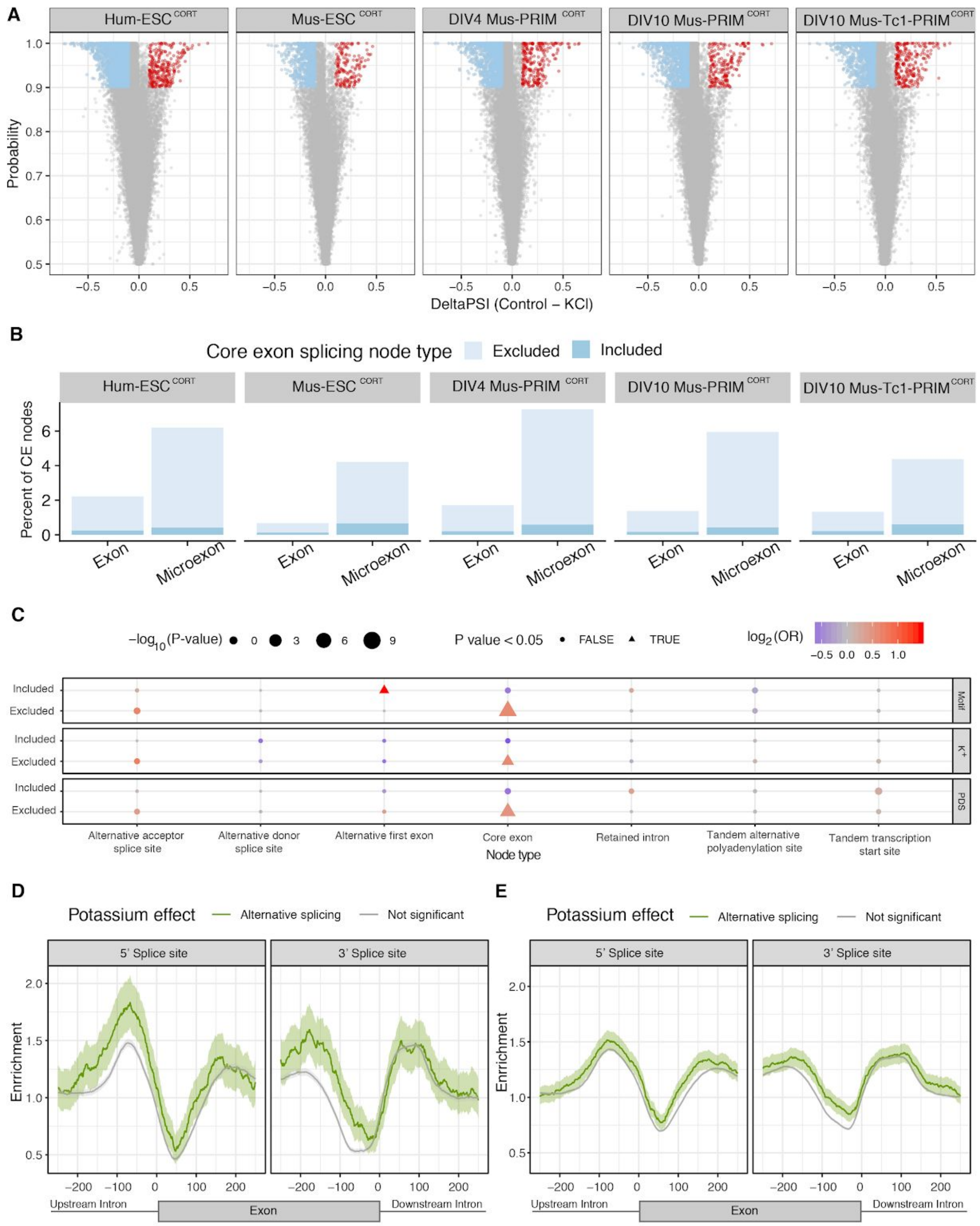

Figure 5.1: Depolarization induces genome wide exon skipping of cassette exons. Analysis of depolarization experiments across human and mouse stem cell-derived cortical neurons (Hum-ESC^{CORT} and Mus-ESC^{CORT}), mouse developing primary cortical neurons (DIV4 Mus-PRIM^{CORT} and DIV10 Mus-PRIM^{CORT}) and aneuploid Tc1 mouse neurons (DIV10 Mus-Tc1-PRIM^{CORT}) (Qiu et al., 2016). **A.** Volcano plot highlighting differentially excluded (blue dots) and included cassette

exons (red dots) after neuronal depolarization stimuli. These cassette exon inclusion quantification derives from core exon splice node quantification performed by Whippet. **B.** Differential cassette exon inclusion events after neuronal depolarization are enriched in microexons. Barplot shows the percentage of all core exons (CE) splicing nodes that correspond to cassette exon or microexon differential inclusion. Across all human and murine cortical neurons, alternative splicing changes after depolarization stimuli is significantly enriched in microexons (chi-squared tests using Yates' correction and also adjusting for multiple testing with Bonferroni multiple testing corrections, p-value < 1e-15). **C.** Differentially included splice nodes enrichment of G4 motifs in human stem cell-derived cortical neurons. Odds ratio representing the relationship between presence of G4s and alternative splicing changes. The odds-ratio significance was assessed by chi-squared tests. All p-values were calculated with chi-squared tests using Yates' correction and also adjusting for multiple testing with Bonferroni corrections. **D-E.** Enrichment of G4-seq peaks at 3' / 5' splice site vicinity for CE splicing nodes that were or were not detected by Whippet as differentially included after potassium stimulation of Hum-ESC^{CORT} (alternative spliced / not significant). G4 maps were generated after treatment with PDS (D) or KCL (E), showing consistent results. . Splice sites with G4 peaks within 100 nt were more likely to be differentially spliced following KCl treatment (chi-squared test, p-value<0.001 both at 5'ss and 3'ss). . The error bands represent 95% confidence intervals based on a binomial model.

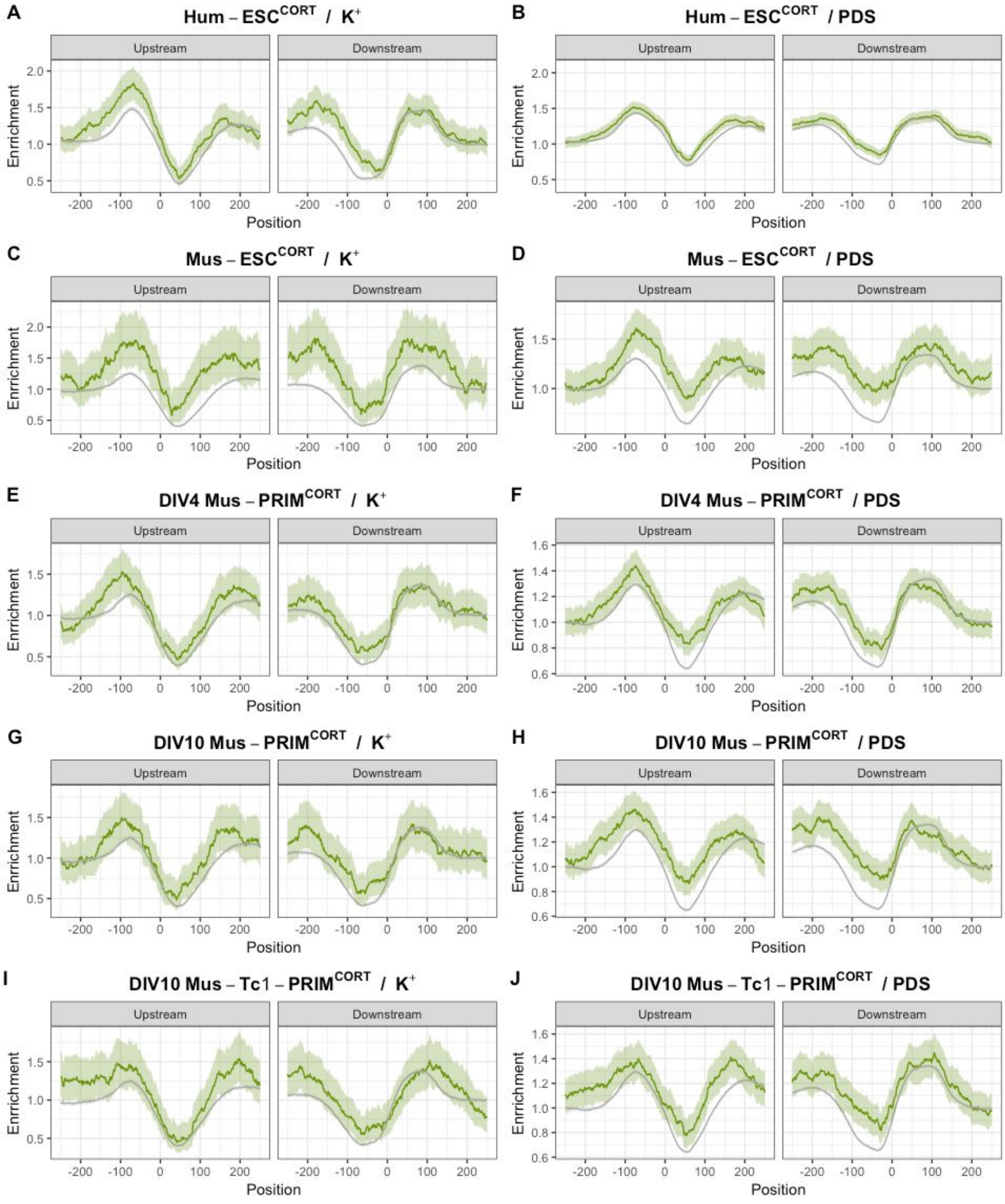

Figure 5.2: Mouse cortical neuronal depolarization experiments show G4-associated alternative splicing patterns consistent with human results. Peaks derived from G4-seq experiments under K+ and PDS are consistently enriched at the vicinity of splice sites from exons that are alternatively included after KCl-induced depolarization stimuli. The different panels correspond to different K+ and PSD G4-seq mouse experiments across different mouse cortical neurons; **A-B** Hum-ESC^{CORT} **A-B** Mus-ESC^{CORT}. **C-D** DIV4 Mus-PRIM^{CORT} **E-F** DIV10 Mus-PRIM^{CORT} **G-H** DIV10 Mus-Tc1-PRIM^{CORT}. The error bands in D-E represent 95% confidence intervals based on a binomial model.

5.2.2 Case study of G4 associated with depolarization induced exon skipping events that are evolutionarily conserved

To gain additional insights into the association between G4s and neuronal alternative splicing, we focused our attention on 3 out of 54 cassette exons that are flanked by one or more G4s and differentially included after cortical neuron depolarization in human and mouse. These exons are found at the *SLC6A17*, *UNC13A* and *NAV2* loci, and they are all flanked by one or more G4s (Fig 5.3). The criteria to select these candidates were based on the alternative splicing analysis results that we obtained from the deliparization experiments, where all three exons were differentially included in human and three or more mouse conditions. Moreover, particular focus was given to these exons since the corresponding genes have been shown to be relevant for neuronal function, which I further discuss below.

5.2.2.1 Candidate selection

SLC6A17 (NTT4/XT1) is a member of the SLC family of transporters which are involved in Na^+-dependent uptake of the majority of neurotransmitters at presynaptic neurons (Zaia and Reimer, 2009). *SLC6A17* is involved in the transport of neutral amino acids and mutations in this gene have been associated with autosomal-recessive intellectual disability (Zaia and Reimer, 2009), (Iqbal et al., 2015). Exon seven of *SLC6A17*, which is skipped after KCl treatment (Delta PSI=-0.177), has a G4 50 nt downstream of the 5′ss on the non-template strand. As the domains of *SLC6A17* include an intracellular loop, two transmembrane regions and part of extracellular domains, the KCl-induced alternative skipping of this exon may lead to functional structural changes (Fig 5.3, 5.4). Similarly, *UNC13A* encodes another presynaptic protein involved in glutamatergic transmission, and it has been associated with amyotrophic lateral sclerosis (Placek et al., 2019). We identify a G4 downstream of exon 38, which results in dramatic exon skipping (Delta PSI=-0.369), (Fig 5.3, 5.5). Finally, the third target was a G4 located downstream of exon 16 in *NAV2* (navigator protein 2), which is required for retinoic acid induced neurite outgrowth in human neuroblastoma cells (Merrill et al., 2002). Again, KCl treatment resulted in exon skipping (Delta PSI=-0.271), which affects a *NAV2* serine rich sequence region (Fig 5.3, 5.6).

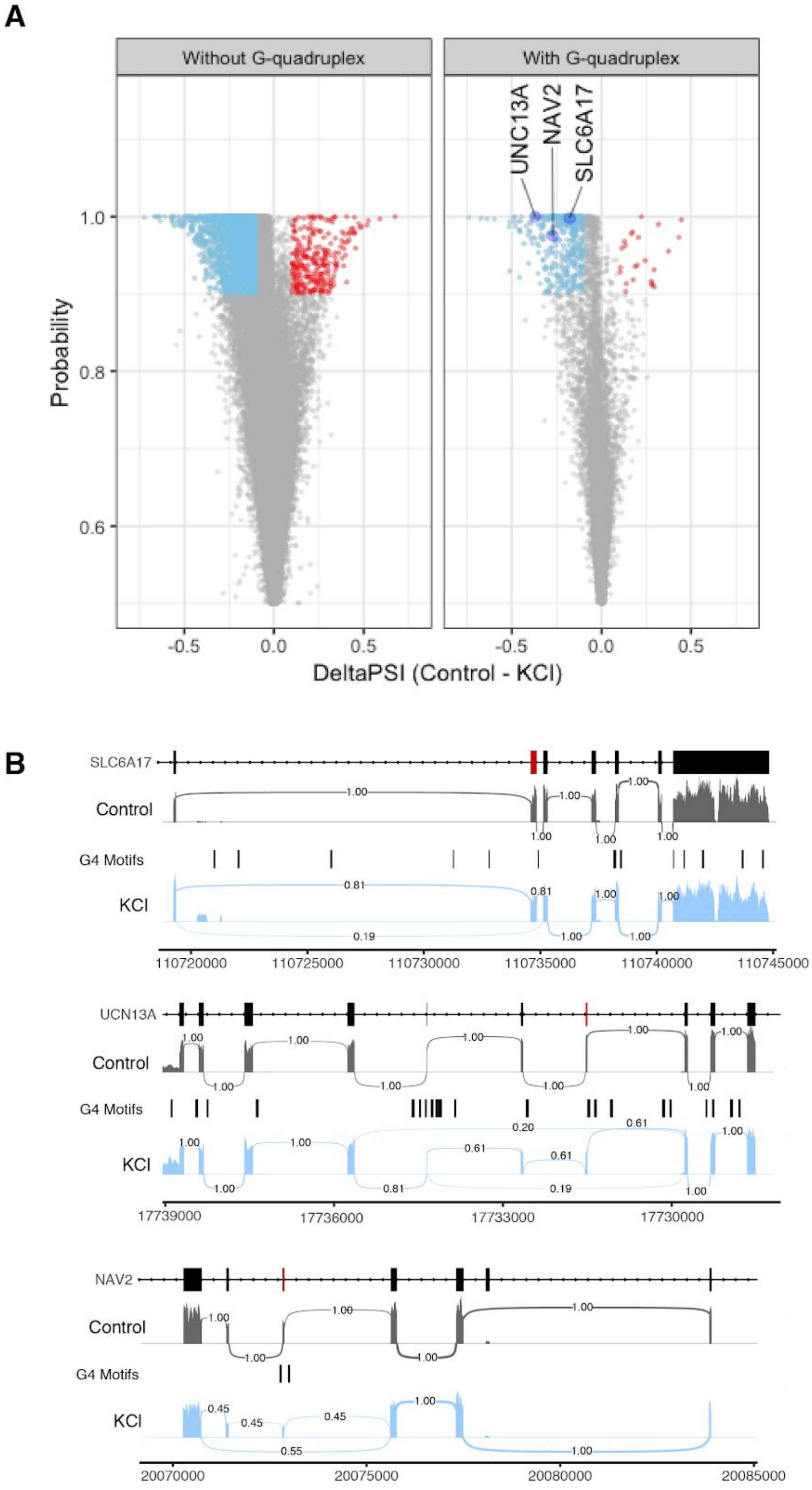

Figure 5.3: Cassette exons exhibit a strong exon exclusion pattern after KCl-induced depolarization A. Volcano plot showing differential inclusion events in presence and absence of flanking G4s and the associated probability following potassium stimulation with widespread exon skipping after depolarisation in human

neuronal cells. **B.** Sashimi plots showing alternative exon inclusion for the three candidates, namely *SLC6A17*, *UNC13A* and *NAV2* following KCl treatment. Exons flanked by a G4 that were used for validation experiments are marked in red. The numbers connecting exons represent the fraction of reads supporting each path.

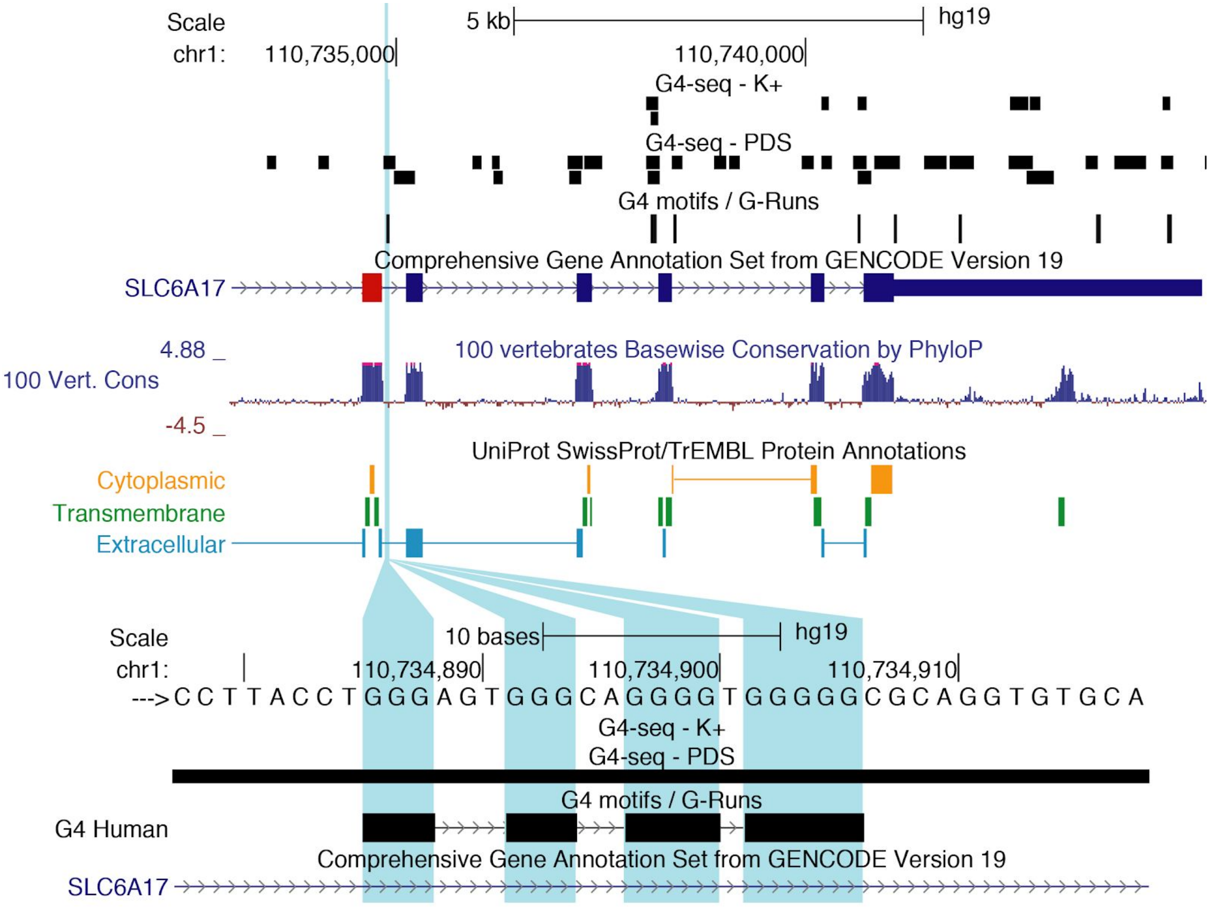
Scale
5 kb
hg19
chr1:
110,735,000
110,740,000
G4-seq - K+
G4-seq - PDS
G4 motifs / G-Runs
Comprehensive Gene Annotation Set from GENCODE Version 19
SLC6A17
4.88 _
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
-4.5 _
UniProt SwissProt/TrEMBL Protein Annotations
Cytoplasmic
Transmembrane
Extracellular
Scale
10 bases
hg19
chr1:
110,734,890
110,734,900
110,734,910
--->C C T T A C C T G G G A G T G G G C A G G G G T G G G G G C G C A G G T G T G C A
G4-seq - K+
G4-seq - PDS
G4 motifs / G-Runs
G4 Human
Comprehensive Gene Annotation Set from GENCODE Version 19
SLC6A17

Figure 5.4: Non-template G-quadruplex motif downstream from an alternatively included SLC6A17 exon. Exon highlighted in red is skipped after KCl-induced depolarization. Downstream non-template G-quadruplex motif is highlighted in blue. G4-seq and additional UCSC tracks are shown.

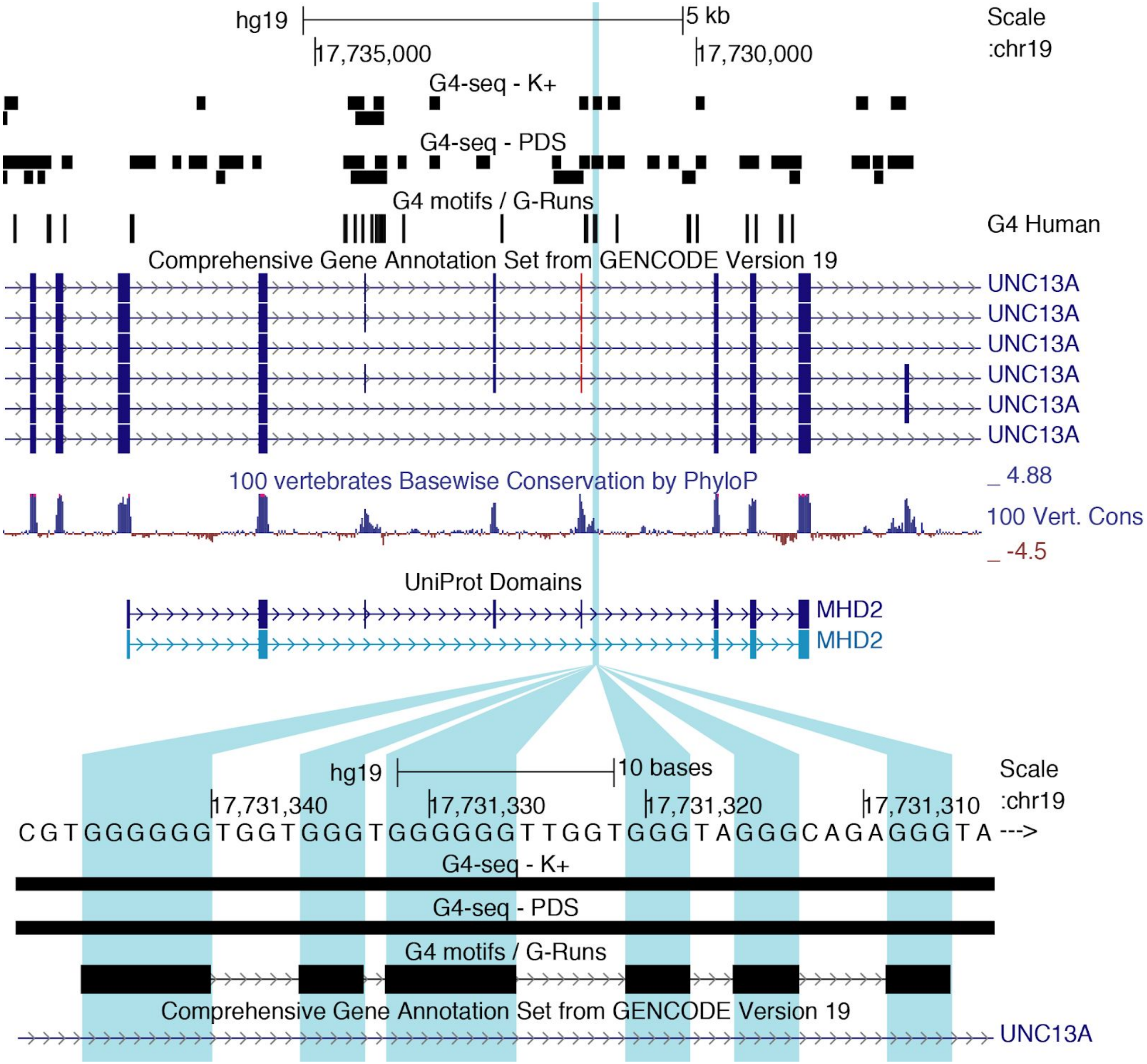

Figure 5.5: Non-template G-quadruplex motif downstream from an alternatively included *UNC13A* microexon. Microexon highlighted in red is skipped after KCl-induced depolarization. Downstream non-template G-quadruplex motif is highlighted in blue. G4-seq and additional UCSC tracks are shown.

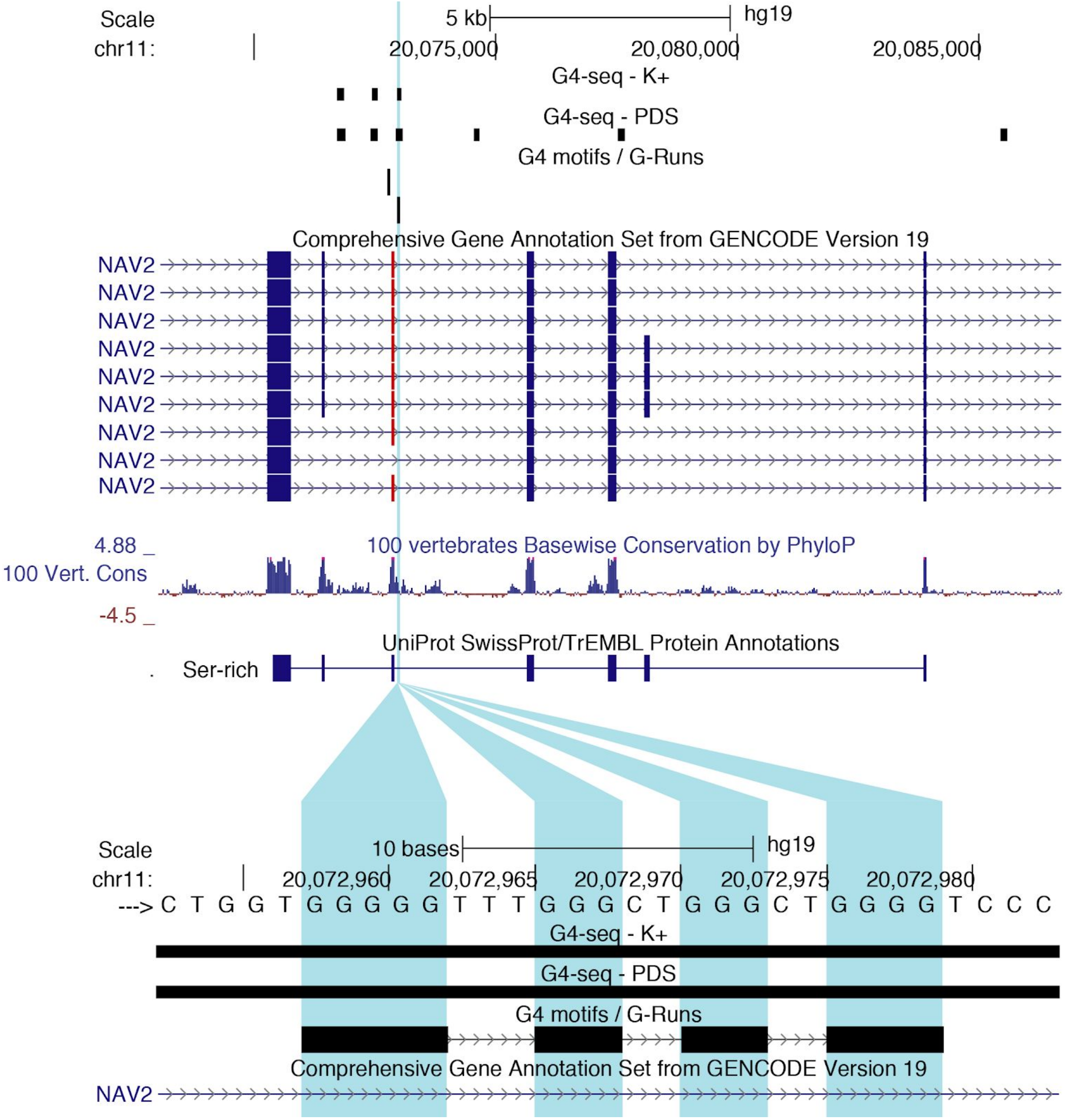

Figure 5.6: Non-template G-quadruplex motif downstream an from alternatively included *NAV2* exon. Exon highlighted in red is skipped after KCl-induced depolarization. Downstream non-template G-quadruplex motif is highlighted in blue. G4-seq and additional UCSC tracks are shown.

5.2.2.2 G4 motif sequences found at SLC6A17, UNC13A and NAV2 promote the formation of G4 structures *in vitro*

For each of the three candidates we designed RNA oligos that were used by Hei Yuen to perform multiple assays which demonstrate that these G4 not only form at the DNA level (as shown by G4-seq maps), but they are also formed at the RNA level *in vitro*. Further details the results obtained by our collaborators can be found in our preprint manuscript (Georgakopoulos-Soares et al.).

5.3 Materials and Methods.

5.3.1 Comparative analysis of RNA-seq experiment. Differential exon inclusion following depolarisation.

We analyzed available data (BioProject Accession: PRJEB19451, ENA link: ERP021488) for mouse and human ESC-derived cortical neurons, mouse primary cortical neurons from wild-type and Tc1 mice stimulated with KCl treatment and untreated followed by RNA-seq four hours post-treatment (Qiu et al., 2016). We used MicroExonator coupled to Whippet (Sterne-Weiler et al., 2018) to discover microexons and integrate them in alternative splicing analysis performed by Whippet to assess the differential inclusion of splicing nodes after KCl/FPL treatment and controls. We used absolute value of DeltaPSI greater than 0.1 and probability greater than 0.9 to define a splicing node as differentially included between treatment and controls.

We calculated the distance between the middle point of G4 motifs or G4-seq peaks from each splicing node to determine their association with G4s. Splicing nodes whose splice sites were within 100 bps of a G4 motif or 45 bps to a G4-seq peak were classified as G4 associated splicing nodes. Next, we assessed the influence of G4s to splicing changes following KCl depolarisation of human and mouse neurons by calculating the odds ratio score of each splicing node type. To determine the statistical significance of the effect we performed a chi-squared test using Yates` correction and also adjusting for multiple testing with Bonferroni corrections. The distribution of G4 motifs and G4-seq peaks was profiled around differentially included and non-differentially included core exon splicing nodes (CE). The confidence intervals were calculated using “binconf” command from “Hmisc” package in R with default parameters. Sashimi plots were generated using “ggsashimi” package (Garrido-Martín et al. 2018). Inclusion and exclusion path ratios were calculated using the total amount of spliced reads supporting each splice junction, where inclusion paths were calculated using the average read count for splice junctions flanking each exon side.

Three putative non-template G4s found in proximity to splicing junctions and which were differentially included following depolarisation in human ESC-derived neurons and in at least one condition in mice were selected for validation experiments. These were: i) a G4 downstream of exon 7 for *SLC6A17* (chr1: 110734886-110734906), ii) a G4 downstream of exon 38 in *Unc13a* (chr19:17731307-17731346) and iii) a G4 upstream of exon 16 in *Nav2* (chr11:20072958-20072979) for which RNA oligonucleotides at the G4 locations were ordered.

The RNA oligonucleotides used were (G-runs marked in bold):

1. *SLC6A17* oligonucleotide:

GGGAGT**GGG**CA**GGGG**T**GGGGG**

2. *UNC13A* oligonucleotide:

GGGGGGTGGT**GGG**T**GGGGGG**TTGGT**GGG**TA**GGG**CAGA**GGG**

3. *Nrxn2* oligonucleotide:

GGGGGTTT**GGG**CT**GGG**CT**GGGG**

6 Chapter VI - Discussion and future work

The findings that were described in previous chapters provide some novel insights about two non-canonical splicing features: extremely short exon size (microexons) and the non-canonical DNA and RNA structures associated with splice sites. In the following chapter, these findings are put into perspective and I also highlight future research directions to address yet unsolved questions in this field.

6.1 Development of computational a workflow for reproducible detection and quantification of microexons

The advent of RNA-seq technologies has provided unprecedented opportunities to explore the complexity of vertebrate transcriptomes. Numerous bioinformatics tools can be used to quantify gene expression and many alternative splicing events. However, the detection of splicing events associated with non-canonical features, such as introns with non-canonical dinucleotides, recursive splice sites, back splicing events or microexons, has required the development of specialized bioinformatics methods (Irimia et al., 2014; Li et al., 2015; Parada et al., 2014; Sibley et al., 2015; Wu et al., 2013; Zeng et al., 2017).

In chapter II, I presented MicroExonator, a complete bioinformatic workflow for reproducible discovery and quantification of microexons. Since MicroExonator was implemented using Snakemake as a workflow management system, large volumes of data can be handled using HPC systems while ensuring that the analyses are reproducible. Simulation-based benchmarking results show that MicroExonator has higher sensitivity than widely used RNA-seq alignments tools such as STAR (Dobin et al., 2013) and HISAT2 (Kim et al., 2015) (Fig 2.3, 2.4). Moreover, even though Olego (Wu et al., 2013) has a module specifically dedicated for microexon discovery, sensitivity to find short microexons was comparable to HISAT2, but still inferior to MicroExonator (Fig 2.4c).

While the benchmark results show that MicroExonator has sensitivity improvements particularly for the very short microexons (<10 nt), the evaluation of false-positive microexon rates demonstrates that MicroExonator has significantly higher specificity than START, HISAT2 and Olego at almost the full range size of microexons (Fig 2.4b). These results validate the computational strategies implemented in MicroExonator to reduce the detection of spurious microexons (Fig 2.2), which have been identified as one of the major challenges to perform microexon discovery (Wu and Watanabe, 2005).

6.2 MicroExonator enables large-scale reproducible analyses of microexon splicing

6.2.1 Microexon coordination across neuronal development

Microexon quantitative analysis revealed that the proteins containing microexons form a highly connected PPI network during mouse neuronal development. Moreover, analysis of the topology of the network suggests that the microexons for the most central nodes are included early in development. It is not yet fully understood how this coordination is achieved, but it has been proposed that microexon inclusion relies on an upstream intronic splicing enhancer which is recognized by specific neuronal splicing factors (Gonatopoulos-Pournatzis et al., 2018). However, I also identified a large group of microexons that are constitutively included across murine tissues, suggesting that their inclusion cannot be dependent on tissue-specific factors alone. Instead, our analysis points to a more straightforward explanation as the constitutive microexons have stronger splicing signals than neuronal microexons. Further analysis of neuronal microexon cis-regulatory elements is required to understand how inclusion events are coordinated and why there is a small number of microexons that are progressively excluded through brain development.

The predominant mechanism for regulating alternative splicing events during neuronal development is through RNA binding proteins (Vuong et al., 2016). In the case of microexons, *SRRM4* and *RBFOX1* have a critical role in coordinating microexon inclusion through brain development, and changes in expression of these splicing factors have been linked to misregulation of alternative splicing events in individuals with autism spectrum disorder (ASD) (Irimia et al., 2014; Li et al., 2015; Voineagu et al., 2011). In fact, alternative splicing changes associated with ASD are enriched in microexons and they are recapitulated in mutant mice haploinsufficient for SRRM4 (Irimia et al., 2014; Quesnel-Vallières et al., 2015). Moreover, a recent genome-wide CRISPR-Cas9 screen has identified two additional factors, SRSF11 and RNPS1, that contribute to SRRM4-dependent microexon regulation, and these genes have also been implicated in ASD and other neurological disorders (Gonatopoulos-Pournatzis et al., 2018). Another example of a protein where imbalances of microexon inclusion have been associated with an elevated risk of ASD is cytoplasmic polyadenylation element-binding protein 4 (CPEB4) (Parras et al., 2018). I found differential inclusion of a *CPEB4* microexon during mouse embryonic brain development, and I also found microexon changes in other protein factors that are involved in mRNA polyadenylation, such as CPEB2, CPEB3 and FIP1L1. All four members of the cytoplasmic polyadenylation element binding (CPEB) family are involved in translational control and have been found to be transcribed in the mouse transcriptome. CPEB transcriptional control has been associated with synaptic plasticity, learning and memory (Turimella et al., 2015). While the role of these microexons in neuronal function and neuropsychiatric diseases remains unexplored, CPEBs function have been associated with ALS and human episodic memory (Downie, 2017; Vogler et al., 2009).

The high degree of conservation of microexons strongly suggests that they are functionally important, however detailed mechanisms of how microexon splicing impacts neuronal function and development have not yet been carried out for most loci. A notable exception is *SRC* where microexon inclusion leads to the production of a well-characterized neuronal splice variant (n-SRC). The *SRC* microexon encodes for a positively charged residue located at an SH3 domain that has been shown to regulate Src kinase activity and specificity (Brugge et al., 1985). From the STRING analysis (Fig 3.3), I found evidence of SRC-dependent phosphorylation of

GIT1, CTNND1 and PTK2 (Chernyavsky et al., 2008; Lim et al., 2002; Wang et al., 2010a). The impact of neuronal microexon alternative splicing for these phosphorylation events remains unknown. However, recent studies show that n-SRC microexon inclusion is required for normal primary neurogenesis and the L1cam dependent neurite elongation (Keenan et al., 2017; Lewis et al., 2017b), implying a strong phenotype. Another central node in the PPI network that is known to undergo microexon alternative splicing changes that are important for axon growth is *L1CAM*, a founding member of L1 protein family. Across the L1 protein family, a sorting signal is included due to 12-nucleotide alternative microexons. In the case of *L1CAM*, the 12-nucleotide microexon mediates clathrin-mediated endocytosis by interacting with adaptor protein complex 2 (AP-2) (Kamiguchi et al., 1998). Our analysis shows that the AP-2 mu subunit (*AP2M1*) is also affected by microexon inclusion through mouse brain development.

6.2.2 Cell-type specific microexon alternative splicing across the mouse visual cortex

Single-cell RNA-seq data are providing unique opportunities to survey cell-specific expression profiles. However, with a few notable exceptions (Arzalluz-Luque and Conesa, 2018; Gokce et al., 2016; Lukacsovich et al., 2019; Zhang et al., 2016), most scRNA-seq analyses have focused on the gene rather than the transcript level. Here, I applied MicroExonator to GABA-ergic and glutamatergic cells from the visual cortex, and to increase power I developed a downstream SnakeMake workflow, snakepool. As many splicing events are undetected in single cell data due to poor coverage, a pooling strategy is necessary to increase the power to identify significant differential inclusion events.

From the analysis with snakepool, 39 microexons were detected as differently included between GABA-ergic and glutamatergic neurons. Fifteen of these cell-type specific microexons are found encoding eleven synaptic proteins. Among these, two alternatively included microexons were found for *PTPRD*, a protein known to have a key role in modulating trans synaptic interactions and having a direct impact on synapse formation (Yamagata et al., 2015a, 2015b). In addition, microexons found in

PTPRD and other proteins involved in transsynaptic protein interactions were found to have distinctive alternative inclusion profiles across GABA-ergic and glutamatergic subtypes (Fig 3.5).

The differential inclusion of microexons could have profound effects on neuronal identity, synapse formation and disease. For example, GABA-ergic neurons were found to have higher inclusion levels of an alternative microexon in GABAa receptor subunit γ (GABRG2) and this alternative splicing event may have significant repercussions for GABA-ergic neuronal function since GABRG2 microexon introduces a phosphorylation site that regulates the GABA activated current (Moss et al., 1992; Ustianenko et al., 2017; Whiting et al., 1990). Misregulation of this alternative splicing event has been associated with schizophrenia in human patients (Huntsman et al., 1998; Ustianenko et al., 2017). However, additional analyses of alternative microexon patterns across neuronal cell-types will be required to fully understand their contribution to neuronal heterogeneity and function.

6.3 Microexon alternative splicing may shape neuronal connectivity

Taken together, the results of the single cell analysis suggest that microexon alternative splicing events may have an influence over synaptic formation across different neuronal subtypes (Fig 3.5). The alternative inclusion of some of these cell-type specific microexons is also shown to be regulated through mouse embryonic development (Fig 3.3). Therefore, it is possible that microexon splicing patterns that are cell-type-specific are established during neuronal development and have a deep impact on the way the neuronal connectome develops to form a mature brain.

Only very recently have bulk RNA-seq analyses started to uncover alternative splicing differences across neuronal subtypes (Furlanis et al., 2019; Saito et al., 2019; Wamsley et al., 2018). These strategies are based on the cell-type specific mRNA isolation based on fluorescence-activated cell sorting or mRNA-pull down approaches, enabling alternative splicing analysis between cellular subtypes by standard bulk RNA-seq. Wamsley and collaborators used these approaches to study alternative splicing changes that are regulated during cortical interneuron development across specific subcellular types (SST+ and PV+ cINs), where they highlight microexon alternative splicing events in *PTPRD* and *NRXN1* as examples of developmentally regulated alternative splicing events between E18.5 and P4. Moreover, they found that Rbfox1 orchestrates a substantial part of the developmentally regulated alternative splicing events that affect synaptic proteins (Wamsley et al., 2018). Furlanis and collaborators developed a novel approach to perform neuronal cell-type-specific transcriptome profiling, by isolated ribosome-engaged transcripts of genetically defined cortical and hippocampal neuron populations. The analysis of these results showed the existence of an alternative splicing program dedicated to the control of synaptic interactions. Even though they report that only 3.8 - 5.3% of the differentially splicing events across different neuronal cell-types involve microexons, several of the synaptic proteins affected by these events are also affected by differentially included microexon events that I detected in bulk and single-cell RNA-seq data (Fig 6.1)

[redacted]

Figure 6.1: Summary of synaptic genes affected by alternative splicing events across neuronal sub-populations in mouse brain. Cell-type specific alternative splicing events found across brain regions (including forbrain, neocortex and hippocampus) affect spliceosomal proteins. Furlanis and collaborators uncovered alternative splicing programs that might control synaptic interactions and neuronal architecture (Furlanis et al., 2019). * denote genes that are also altered by differentially included microexon events (presented in Chapter III) detected by MicroExonator, either though bulk (red) or single-cell (green) RNA-seq data analysis. Schematics adapted from (Furlanis et al., 2019).

6.4 Non-neuronal microexons

Differential inclusion analyses from RNA-seq samples corresponding to different mouse tissues not only enabled the identification of neuronal microexons, but also microexons that are differentially included in SKM, heart and AG (Fig 2.2f-i). To the best of my knowledge, there are no reports of alternatively included microexons in AG. However, since microexons differentially included in AG highly overlap with microexons differentially included in the brain, they are likely to correspond to neuronal microexons that are included in neuroendocrine cells, known as chromaffin cells, which are derived from the neural crest during embryonic development (Bornstein et al., 2012; Shtukmaster et al., 2013).

Li and collaborators suggest that RBFOX proteins can regulate microexon inclusion across brain, muscle and heart (Li et al., 2015). However, the functional impact of these microexon inclusion events is largely unknown. Several of the alternative microexon events that are common between brain, SKM and heart have an impact on Mads box transcription enhancer factor 2 (MEF2), particularly on *MEF2A* and *MEF2D* subunits. MEF2 is a transcription factor involved in nervous system development, however RBOX proteins were reported to regulate *MEF2D* alternative splicing events that were required for muscle differentiation (Porter et al., 2018; Runfola et al., 2015). Even though I did not observe any obvious microexon inclusion trend in SKM and heart samples across embryonic development, extensive alternative splicing transitions have been observed during postnatal skeletal muscle development (Brinegar et al., 2017). Thus, microexon developmental changes could be potentially coordinated in later mouse developmental stages that were not included in the RNA-seq experiments that I analysed. Moreover, I found 65 microexons that were differentially included in SKM and/or heart samples (Fig 2.2i), which suggest that additional factors might regulate microexon inclusion in muscular tissues.

6.5 The G-quadruplex formation is enriched in splice sites

Even though B DNA is the most common DNA conformation, different sequence motifs are associated with the formation of non-B DNA structures (Bacolla and Wells, 2004). In Chapter IV, I presented the enrichment analyses of different non-B DNA motifs, of which G4s show the highest enrichment across splice sites. Similar enrichments have been reported by previous in-silico analyses (Maizels and Gray, 2013; Tsai et al., 2014), however I analysed recently published G4-seq data that corroborated the potential of these motifs to form G4 structures near splice sites. More in-depth characterisation of G4 enrichment indicates strong differences between template and non-template strands; while for the upstream exon intronic regions similar level of enrichment was observed for the template and non-template strand, higher non-template G4 enrichment was found for the downstream intronic regions. The high G4 enchments and the strand asymmetries that were observed suggest that G4 positioning around splice sites may be subject to purifying selection, which could be tested by analysing differences on allelic variability of template and non-template G4s that are located in the vicinity of splice sites. Moreover, recent reports showed that non-template G4 motifs can enhance promoter activity by inducing successive R-loop formation (Lee et al., 2020). This capability of non-template G4 motifs to promote R-loop formation may mediate transcriptional kinetics effects that impact mRNA splicing, however further experiments and analyses will be needed to test this hypothesis.

The evolutionary analyses showed that G4 motifs are a conserved feature in vertebrates, and it may be restricted to mammals and birds (Fig 4.11). However, a more comprehensive evolutionary analysis is needed to fully characterize G4 presence in higher eukaryotic organisms. The presence of additional regulatory mechanisms is in accordance with higher frequencies of alternative splicing events in vertebrates compared to invertebrates (Artamonova and Gelfand, 2007). Moreover, G4s display a higher likelihood of DNA mutations (Du et al., 2014) and as a result they are likely plastic in nature, enabling rapid splicing changes during evolution and the establishment of new functions through alternative splicing and the generation of isoform diversity.

There are certain alternative splicing features that are determinants of exon definition in vertebrates. One of them is splicing strength, which largely influences exon inclusion frequency across isoforms. While strong splice sites are associated with constitutive exons, weaker splice sites lead to suboptimal exon recognition (Luco et al., 2011). Thus, this enables alternative splicing events to be modulated by additional cis-regulatory elements or epigenetic factors (Ast, 2004). Here I show that there is a pronounced enrichment of G4s at weak splice sites and provide evidence for widespread contribution of G4 structures to the regulation of alternative splicing. Xiao and collaborators have also shown splice site strength-dependent association of G-runs across splice sites (Xiao et al., 2009). However, in their analysis they studied independent G-runs, that do not necessarily correspond to G4 motifs, which were found to be more enriched across splice sites with intermediate 5’ splice strength.

6.6 Mechanistic models for G4-dependent modulation

G-runs have been previously reported to be bound by hnRNP F/H, which is known to have a direct influence over splice site recognition (Caputi and Zahler, 2001; Královicová and Vorechovsky, 2006; Marcucci et al., 2007; Mauger et al., 2008; McCullough and Berget, 1997; McNally et al., 2006; Yeo et al., 2004b). Thus, part of the G4 motifs studied in Chapter IV, might overlap with G-runs that are targets for hnRNP F/H binding. However, whether the formation of G4s could enhance or undermine the binding of hnRNP F/H is still a matter of debate. Functional minigene assays suggest mutations that have deleterious effects over G4 formation inhibit exon inclusion by preventing hnRNP F binding, a positive regulator of exon definition (Huang et al., 2017). However, the interpretation of these experiments contradicts previous biophysical evidence which shows that hnRNP F binds preferentially to single-stranded G-tracts, suggesting that G4 formation could have a rather negative effect over exon inclusion (Samatanga et al., 2013). This model of G4 formation as an impediment for hnRNP F binding, is consistent with diverse evidence published by the Burge laboratory which indicates preferential binding of RBPs to unstructured RNA (Dominguez et al., 2018; Lambert et al., 2014; Taliaferro et al., 2016), however more experiments and analysis will be required resolve this conflicting evidence.

Additionally, during transcription G4 formation can be favored by unstranded DNA that is transiently generated inside the transcription fork, favoring DNA G4s at the non-template strand and RNA G4s. In fact, G4 formation has been associated with transcriptionally active promoters, which may lead to genome instability induced by double strand break generation (Hänsel-Hertsch et al., 2016; Marnef et al., 2017). G4 formation can have kinetic effects over RNA PolII by delaying RNA polymerization, in fact gene transcription relies a the co-transcriptional unwinding of G4s by dedicated helicase activity (Chakraborty and Grosse, 2011; Paeschke et al., 2011). Since transcriptional speed has an impact on alternative splicing, G4 formation may lead to alternative splicing regulation through a kinetic control of transcription (Nieto Moreno et al., 2015). Moreover, transcription can induce

RNA/DNA hybrid G-quadruplexes, which can lead to even stronger effects over RNA polymerase progression (Shrestha et al., 2014).

6.7 Depolarisation induced alternative splicing

To investigate the relationship of non-canonical splicing features with dynamic alternative splicing regulation events, we quantified the inclusion of microexons and G4-flanked exons in the context of neuronal depolarization. After KCl-induced depolarization of human and mouse ESC-derived neurons, I observed genome wide changes in alternative splicing patterns, of which around 10% corresponded to exon skipping events. These exon skipping events are much more frequently observed than exon inclusion and they are enriched in microexons and G4-flanked exons (Fig 5.1), suggesting that there is a regulated program of alternative splicing events induced by neuronal depolarization. The prominent role of G4s and microexons suggests that non-canonical splicing features are central to this process. Depolarization neurons leads to a strong increase of intracellular calcium, which is thought to mediate previously reported exon skipping events (Lee et al., 2007; Xie and Black, 2001) (Fig 6.2). However, previous studies have only focused on a handful of genes and to the best of my knowledge this is the first time that thousands of exon skipping events have been shown to be triggered after neuronal depolarization.

The molecular mechanisms by which the increase of intracellular calcium concentration leads to alternative splicing events (particularly exon skipping), are not fully understood. In the case of *NMDAR1* and *KCNMA1* exon skipping events, the increase of intracellular calcium induced by depolarization was reported to activate CaMK IV, a calcium/calmodulin-dependent kinase protein which can regulate splicing selection though specific cis-regulatory elements (CaRREs) which have been shown to be associated with hnRNP L binding sites (Ares, 2007; Lee et al., 2007; Li et al., 2009; Sharma and Lou, 2011; Xie, 2008). However, the increase of intracellular calcium can also lead to CaMK IV-independent alternative splicing changes. For example, exon 19 of RBFOX1 is skipped after depolarization, which leads to an increase of RBFOX1 nuclear localization and induction of

RBFOX1-dependent alternative splicing events (Lee et al., 2009). Another example is NCAM exon 18, which does not respond to the CaMK IV pathway, but instead its skipping is determined by an increase of H3K9ac across exon 18, that is induced after neuronal depolarization. H3K9ac may increase RNA polII elongation rates and has a kinetic effect over splicing of exon 18 of *NCAM*. The presence of G4s near splice sites may modulate the effects of intracellular calcium over alternative exon inclusion events in a CaMK IV-dependent manner. However, G4 structures could also directly respond to changes in intracellular calcium. Miyoshi and collaborators showed that structural transitions from antiparallel to parallel G4 conformations are induced by Ca2+ (Miyoshi et al., 2003). These structural changes may have a direct impact on RNA polII kinetics and splicing, but more experimental evidence will be required to prove or disprove this hypothetical mechanism.
On the other hand, microexons have already been shown to be widely skipped after depolarization of primary cultured hippocampal neurons (Quesnel-Vallières et al., 2016). Voltage-gated calcium channels are key transducers of membrane potential changes of intracellular Ca^{2+} concentration under physiological and experimentally induced conditions (Catterall, 2011). Analysing differential inclusion of microexons, I have found microexons that are differentially included in genes that encode for different subunits of voltage dependent calcium channels (*CACNB1*, *CACNB4*, *CASTSPER2* and *CASTPER2*), and Sodium/calcium exchanger (SLCA1), some of which are annotated members of the “presynaptic depolarization and calcium channel opening” Reactome pathway (Figure 6.2). Thus regulation of microexon inclusion may directly influence intracellular Ca^{2+} concentration changes after neuronal depolarisation and have the potential to regulate part of the alternative splicing events that respond to intracellular Ca^{2+}.

Finally, in several cases, both types of non-canonical splicing features herein studied there were related to the same depolarisation-induced exon skipping events. For example a 30-nt microexon at the UNC13A gene, that was skipped after KCl-induced depolarization, was found to be flanked by a downstream non-template G4 (Fig 5.5). Since UNC13A encodes for a presynaptic protein which plays a key role in glutamatergic transmission, alternative skipping of this exon may have a regulatory effect over neuronal transmission and neurological diseases (Placek et al., 2019). Another interesting example is the case of a microexon skipping event induced after neuronal depolarization at the *NRXN2* gene, which is flanked by an upstream template G4 (Fig 6.3). This microexon affect an extracellular domain of neurexin-2, implicated in trans-synaptic protein-protein interactions that regulate synaptic formation. Both of these examples are conserved between human and mouse and may be part of a fine-tuned regulatory network of alternative splicing events that coordinate synaptic formation across neuronal populations.

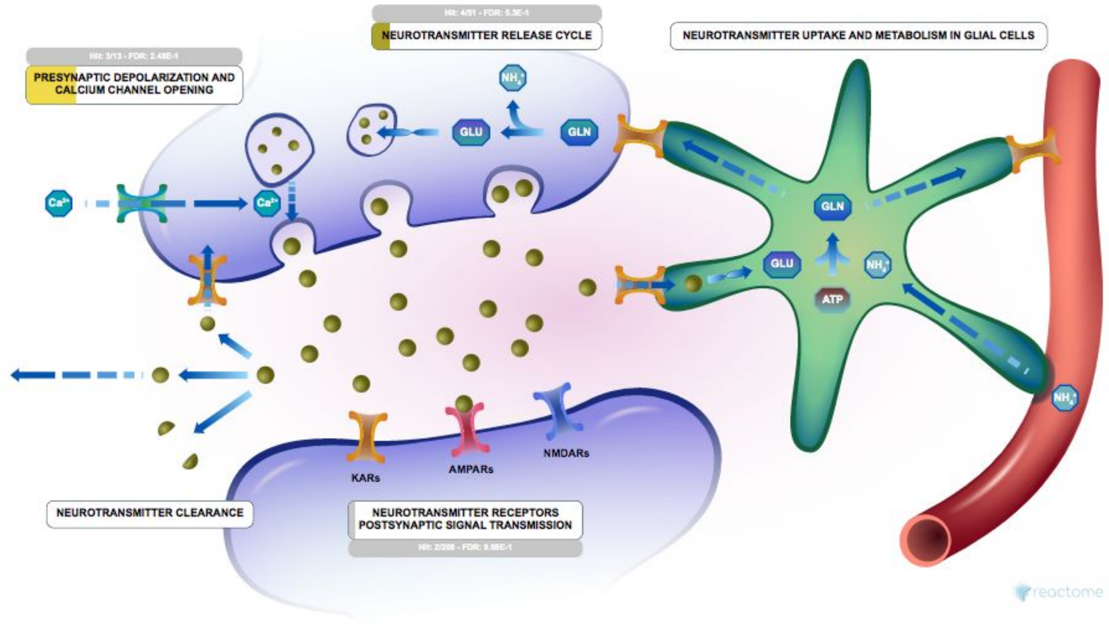
Hit: 3/13 - FDR: 2.48E-1
PRESYNAPTIC DEPOLARIZATION AND CALCIUM CHANNEL OPENING
Hit: 4/51 - FDR: 5.3E-1
NEUROTRANSMITTER RELEASE CYCLE
NEUROTRANSMITTER UPTAKE AND METABOLISM IN GLIAL CELLS
NH_4^+
GLU
GLN
Ca^{2+}
Ca^{2+}
GLN
GLU
NH_4^+
ATP
NH_4^+
KARs
AMPARs
NMDARs
NEUROTRANSMITTER CLEARANCE
NEUROTRANSMITTER RECEPTORS POSTSYNAPTIC SIGNAL TRANSMISSION
Hit: 2/208 - FDR: 9.68E-1
reactome

Figure 6.2: Developmentally regulated microexons can have an impact over transmission across chemical synapses. At least two Reactome pathways related with transmission across chemical synapses might be affected microexon inclusion, these includes "Presynaptic depolarization and calcium channel opening" and Neurotransmitter release cycle, where 3/13 and 4/51 genes involved were found to be differentially included through mouse embryonic development using MicroExonator. Bar highlighted in yellow bar indicates statistical enrichment through pathways analysis done using Reactome (Fabregat et al., 2018).

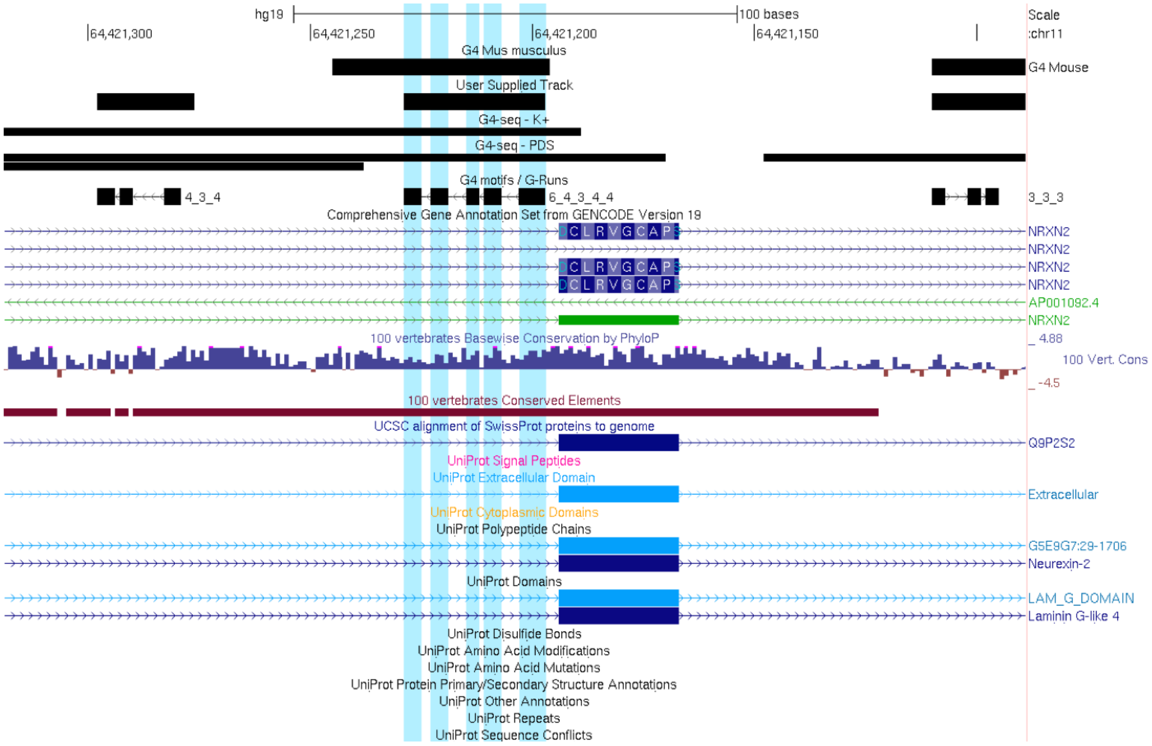
hg19
100 bases
Scale
chr11
64,421,300
64,421,250
64,421,200
64,421,150
G4 Mus musculus
G4 Mouse
User Supplied Track
G4-seq - K+
G4-seq - PDS
G4 motifs / G-Runs
4_3_4
6_4_3_4_4
3_3_3
Comprehensive Gene Annotation Set from GENCODE Version 19
C L R V G C A P
NRXN2
NRXN2
NRXN2
NRXN2
AP001092.4
NRXN2
100 vertebrates Basewise Conservation by PhyloP
4.88
100 Vert. Cons
-4.5
100 vertebrates Conserved Elements
UCSC alignment of SwissProt proteins to genome
Q9P2S2
UniProt Signal Peptides
UniProt Extracellular Domain
Extracellular
UniProt Cytoplasmic Domains
UniProt Polypeptide Chains
G5E9G7:29-1706
Neurexin-2
UniProt Domains
LAM_G_DOMAIN
Laminin G-like 4
UniProt Disulfide Bonds
UniProt Amino Acid Modifications
UniProt Amino Acid Mutations
UniProt Protein Primary/Secondary Structure Annotations
UniProt Other Annotations
UniProt Repeats
UniProt Sequence Conflicts

Figure 6.3: Template strand G-quadruplex formation upstream a depolarization-dependent microexon skipping event in neurexin 2. Template G-quadruplex motif is highlighted in blue. G4-seq and additional UCSC tracks are shown.

6.8 Concluding remarks

In this thesis, I have successfully developed novel bioinformatics methods and analysed high throughput sequencing data to characterise microexons and G4-flanked exons. These analyses not only indicate that both short exon size and presence of G4-structures are associated with alternative splicing events, but they also suggest G4s could be potential regulatory features that drive some of the dynamic alternative splicing modulations observed in neuronal cells.

Since microexons are often missing from RNA-seq analyses when standard tools are used, I developed MicroExonator, a novel computational workflow that enables reproducible discovery and quantification of microexons on RNA-seq data. MicroExonator enabled me to perform the integrated analysis of bulk and single-cell RNA-seq data to have an in-depth characterisation of microexons through mouse development and neuronal subtypes. On the other hand, I also explored the possible association of splice sites with sequence motifs that are known to promote the formation of non-canonical DNA and RNA structures. Among the analysed motifs, I found a significant and consistent enrichment of G4 motifs across splice sites of vertebrates.

Further analyses of G4-seq data and experimental validations corroborate *in vitro* formation of G4s near splice sites. Lastly, I analysed the association of microexons and G4-flanked exons to dynamic splicing changes induced by depolarisation stimuli. These analyses showed a transcriptome-wide induction of cassette exon skipping events. Moreover, these changes were enriched in microexons and G4-flanked cassette exons, suggesting the involvement of these non-canonical splicing features in the dynamic regulation of splicing changes upon neuronal depolarisation.

6.9 Future Work

6.9.1 Exploration of large scale RNA-seq sequencing experiments to study alternative splicing of microexons

MicroExonator provides new opportunities to explore large-scale RNA-seq experiments that have been made available to the public domain and this has the potential to enable a comprehensive characterization of microexon alternative splicing patterning across different biological contexts, which may lead to the identification of different cellular pathways associated with alternative splicing of microexons. MicroExonator can be configured to automatically download the data and perform the analysis. Normally, these analyses are limited by the processing power and disk storage that researchers have access to, but the versatile Snakemake workflow management system enables MicroExonator to be compatible with multiple queueing systems (such as LSF or SLURM, implemented at Sanger and Gurdon institute HPC systems respectively). Thus, during the next coming year I will explore several large collections of RNA-seq data to further elucidate the role of the microexon alternative splicing network. In this section will briefly mention the public repositories that I will use for these purposes and what potential insights they may provide.

6.9.1.1 Psychiatric diseases

The network of microexon alternative splicing events have already been shown to be associated with autism. However, a number of microexon splicing regulated events mentioned in this thesis may be involved in the pathogenesis of other neurological diseases. The PsychENCODE project is providing publicly accessible experiments generated from about 1,000 phenotypically characterized disease-affected human post-mortem brains (PsychENCODE Consortium et al., 2015). The large scale multidimensional genomic data generated has already been used to study isoform-level dysregulation associated within autism spectrum disorder, schizophrenia and bipolar disorder, where microexon alternative splicing events have been highlighted and are most likely involved in the pathogenesis of these disorders (Gandal et al., 2018). Thus, the analysis of these data has the potential to uncover the relevance of microexon alternative splicing under neurological different disease contexts.

1.9.1.2 Primate evolution

The alternative splicing of neuronal microexons is arguably the most conserved network of alternative splicing events so far described (Irimia et al., 2014; Torres-Méndez et al., 2019). Recent research conducted by Torres-Méndez and collaborators has uncovered a novel protein domain, termed as ‘enhancer of microexons’ (eMIC), which drove the evolution of the neuronal microexon splicing network (Torres-Méndez et al., 2019). While these results trace the evolutionary emergence of this network to bilaterian ancestors, the evolution through primates is currently not characterized. Evolutionary studies of microexons might highlight alternative splicing events that can act as microsurgery of proteins in regulatory regions to coordinate different developmental processes that lead to morphological and functional differences of brains across primates.

In collaboration with Ilias Georgakopoulos-Soares and Professor Nadav Ahituv[9], I am analysing RNA-seq data collected for different non-human primates to uncover the evolutionary trajectory of microexons across primate evolution. We are particularly interested in identifying microexons that are specifically gained or lost in the human lineage and we are going to focus our analysis on data published by The Non-Human Primate Reference Transcriptome Resource (NHPRTR) and genotype-tissue expression (GTEx) project (GTEx Consortium, 2013; Pipes et al., 2013).

6.9.1.2 Functional assessment of RBPs enhanced CLIP and loss-of-function experiments

Recent improvements of CLIP assays used for genome-wide identification of RBP binding sites have led to the development of an enhancer CLIP (eCLIP) protocol (Van Nostrand et al., 2016). These experimental techniques correspond to the

[9] Group leader at Department of Bioengineering and Therapeutic Sciences the University of California San Francisco.

crosslinking and immunoprecipitation of RBP RNA targets and eCLIP experiments that profiled 150 RBPs have enabled the generation of robust splicing regulatory maps (Yee et al., 2019) by the ENCODE consortium (Sloan et al., 2016). Moreover, the ENCODE consortium has recently released a large collection of RBP knockdown followed by RNA-seq experiments across two cell lines. The integration of these resources has been used to explore the role of RBPs across different RNA-processing pathways, including alternative splicing. Furthermore, independent analysis of this data has provided novel insights into the regulation of recursive splicing (Blazquez et al., 2018). Further processing of this data might identify novel regulatory elements across different exon populations, including microexons and G4-flanked exons.

6.9.1.3 Cancer

Alternative splicing defects have been recurrently associated with cancer, however the role of microexon splicing in pathogenic cancer-associated mechanisms is poorly understood. Collin and collaborators have recently identified a microexon alternative splicing event that has functional effects over Cytohesin-1 protein function (Ratcliffe et al., 2019). This corresponds to an evolutionarily conserved 3-nt microexon that induces differential affinity of Cytohesin-1 to triglycine and diglycine, being implicated in selective phosphoinositide recognition and affecting signal transduction pathways related to cancer cell migration. These results showed the potential of microexon splicing research to uncover new mechanisms to drive cancer pathogenesis.

The Catalogue Of Somatic mutations In Cancer (COSMIC) project is currently one of the most ambitious large scale projects at the Sanger Institute (Forbes et al., 2011). COSMIC is the largest and most comprehensive curated catalog of somatic mutations detected in human cancer, providing valuable resources for exploring the impact of somatic mutations in cancer. Recent releases of COSMIC have made available the exome sequencing and RNA-seq data[10] across 1020 cancer cell lines. Therefore, the processing and integration of this data may uncover the effects of

[10] Which for now is only internally available, but it will soon be published.

annotated somatic mutations that are associated with alternative splicing defects of microexons.

6.9.2 Further developments of single cell data analysis methods for microexon splicing analysis

Single cell analyses have been a revolutionary approach to catalog cellular subtypes across different model organisms. Tissues with high cellular heterogeneity, such as brain cortex, have been the target of large scale full-length scRNA experiments released by the Allen Institute for Brain Science (Tasic et al., 2016, 2018). In this thesis I have used part of this data to develop novel approaches to evaluate cell-type specific alternative splicing events (snakemake), particularly focusing on microexon splicing. However, further improvements are required to consolidate this method. Recently developed methods to perform cell-type transcriptomic profiling of neurons using bulk RNA-seq, such as RiboTRAP (Furlanis et al., 2019), could be used to benchmark different single cell approaches to study alternative splicing between two cell-types. This approach is highly attractive since the transcriptome of analogous neuronal populations have been profiled between single cell data generated by Tasic and collaborators and RiboTRAP data generated by Furlanis and collaborators. The integrative analyses of bulk and single-cell RNA-seq experiments might provide significant methodological insights as well as novel cell-type specific microexons and other splicing events.

6.9.3 Study of non-neuronal microexons

The alternative splicing of microexons has been reported to be primarily regulated in neurons, but computational analyses have shown that some of them are also included in heart, SKM and pituitary gland. Moreover quantitative analyses using MicroExonator have also led to the identification of microexons in the adrenal gland (Fig 2.2). I proposed that inclusion of microexons detected in adrenal gland can be due to the presence of chromaffin cells in the adrenal gland, which share the same primordial tissue of origin during embryonic development than other neuronal cells. This hypothesis can now be tested through RNA-seq analysis of isolated chromaffin cells, which have been recently generated (Chan et al., 2019). Furthermore, the study of other neuroendocrine glands might lead to the identification of microexon inclusion in other non-neuronal tissues, such as different components of the gastrointestinal tract, gallbladder, pancreas and thyroid.

Even though microexon inclusion has been reported across SKM and heart their functions are largely unknown. Moreover, other types of muscle, such as smooth muscle have not yet been studied. Collaborations with Professor Christopher WJ Smith will be initiated to explore microexon inclusion in smooth muscle and determined if RBPMS, a newly identified splicing factor that control smooth muscle splicing events, is implicated in microexon alternative splicing (Nakagaki-Silva et al., 2019).

6.9.4 Tissue-specific splicing of G4-flanked exons

The results that I have presented in Chapter V indicate that microexon and G4-flanked exons are enriched in alternatively included exons induced by depolarisation stimuli. In the case of microexons, alternative splicing events are strongly associated with neuronal alternative splicing programmes. However, in the case of G4-flanked exons we have not yet systematically explored their inclusion across different tissues and cell-types. Thus, similar analyses to the ones I have conducted for microexons are required to determine if G4-exons are in general associated with alternative splicing, or if they are particularly associated with tissue-specific or cell type-specific splicing events.

6.9.5 Elucidating mechanisms of G4-mediated modulation of alternative splicing

The formation of G4 structures can affect alternative splicing by altering the binding potential of RBPs to their target sites. As mentioned in section 1.6.2.1, there is still controversy about whether G4 formation would promote or block the binding of RBPs to the mRNA. Since there is an increasing amount of crosslinking and immunoprecipitation (CLIP) experiments that is being generated to determine the binding sites of multiple RBPs (Louie et al., 2018; Yee et al., 2019), I plan to do a systematic analysis of RBP binding sites across splice sites and G-quadruplexes. This analysis will provide an unbiased approach to find novel G4 interactors that may play a functional role over alternative splicing regulation.

Another non-exclusive possibility is that G4 formation leads to kinetics effects that regulate alternative splicing. G4s have already been reported to have an impact in RNA polymerase speed and kinetics, but their effects on splicing modulation remain undescribed. However, polymerase speed is already known to regulate splicing and control the recognition of weak splice sites (Luco et al., 2011). Thus the integration of genome-wide transcription pausing with G4-seq and RNA-seq data might provide new mechanistic insight about the kinetic effects of G4 formation in alternative splicing.

6.9.6 Machine learning for motif discovery

Machine learning approaches have been implemented to uncover novel *cis*-regulatory elements that control tissue-specific alternative splicing (Barash et al., 2010; Zhang et al., 2019b). Recent doctoral work of Nicholas Lee[11] has led to the development of a novel convolutional neural network approach to find regulatory motifs across different quantitative transcriptome experiments. Collaborative work between Nicholas Lee, Jacob Hepkema[12] and I have led the identification of novel microexon regulatory motifs that can quantitatively predict the inclusion patterns observed across mouse brain development. Thus, further inspection of these results can lead to the discovery of novel cis-regulatory elements involved in microexon alternative splicing.

[11] PhD candidate at the Sanger Institute and University of Cambridge, who has also been under the supervision of Martin Hemberg.

[12] Master student from Utrecht University, who have been doing analyses at Hemberg's laboratory as an internship work.

7 References

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J.M., del Pozo, A., Vázquez, J., Valencia, A., and Tress, M.L. (2015). Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS Comput. Biol. *11*, e1004325.

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., et al. (2018). How many human proteoforms are there? Nat. Chem. Biol. *14*, 206–214.

Aebi, M., Hornig, H., Padgett, R.A., Reiser, J., and Weissmann, C. (1986). Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell *47*, 555–565.

An, P., and Grabowski, P.J. (2007). Exon silencing by UAGG motifs in response to neuronal excitation. PLoS Biol. *5*, e36.

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. *19*, 1732–1741.

Ares, M., Jr (2007). Sing the genome electric: excited cells adjust their splicing. PLoS Biol. *5*, e55.

Artamonova, I.I., and Gelfand, M.S. (2007). Comparative genomics and evolution of alternative splicing: the pessimists’ science. Chem. Rev. *107*, 3407–3430.

Arzalluz-Luque, Á., and Conesa, A. (2018). Single-cell RNAseq for the study of isoforms-how is that possible? Genome Biol. *19*, 110.

Ast, G. (2004). How did alternative splicing evolve? Nat. Rev. Genet. *5*, 773–782.

Bacolla, A., and Wells, R.D. (2004). Non-B DNA conformations, genomic rearrangements, and human disease. J. Biol. Chem. *279*, 47411–47414.

Bacolla, A., Tainer, J.A., Vasquez, K.M., and Cooper, D.N. (2016). Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res. *44*, 5673–5688.

Baek, D., and Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc. Natl. Acad. Sci. U. S. A. *102*, 12813–12818.

Bang, I. (1910). Untersuchungen über die Guanylsäure. Biochem. Z. *26*, 293–311.

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

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.

Bartschat, S., and Samuelsson, T. (2010). U12 type introns were lost at multiple occasions during evolution. BMC Genomics *11*, 106.

Bartys, N., Kierzek, R., and Lisowiec-Wachnicka, J. (2019). The regulation properties of RNA secondary structure in alternative splicing. Biochim. Biophys. Acta Gene Regul. Mech. 194401.

Beachy, P.A., Helfand, S.L., and Hogness, D.S. (1985). Segmental distribution of bithorax complex proteins during Drosophila development. Nature *313*, 545–551.

Behzadnia, N., Golas, M.M., Hartmuth, K., Sander, B., Kastner, B., Deckert, J., Dube, P., Will, C.L., Urlaub, H., Stark, H., et al. (2007). Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. The EMBO Journal *26*, 1737–1748.

Benhalevy, D., Gupta, S.K., Danan, C.H., Ghosal, S., Sun, H.-W., Kazemier, H.G., Paeschke, K., Hafner, M., and Juranek, S.A. (2017). The Human CCHC-type Zinc Finger Nucleic Acid-Binding Protein Binds G-Rich Elements in Target mRNA Coding Sequences and Promotes Translation. Cell Reports *18*, 2979–2990.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank: update. Nucleic Acids Res. *32*, D23–D26.

Bevilacqua, P.C., Ritchey, L.E., Su, Z., and Assmann, S.M. (2016a). Genome-Wide Analysis of RNA Secondary Structure. Annu. Rev. Genet. *50*, 235–266.

Bevilacqua, P.C., Ritchey, L.E., Su, Z., and Assmann, S.M. (2016b). Genome-Wide Analysis of RNA Secondary Structure. Annu. Rev. Genet. *50*, 235–266.

Bhattacharyya, D., Mirihana Arachchilage, G., and Basu, S. (2016). Metal Cations in G-Quadruplex Folding and Stability. Front Chem *4*, 38.

Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nat. Chem. *5*, 182–186.

Biffi, G., Di Antonio, M., Tannahill, D., and Balasubramanian, S. (2014). Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. Nat. Chem. *6*, 75–80.

Bitton, D.A., Rallis, C., Jeffares, D.C., Smith, G.C., Chen, Y.Y.C., Codlin, S., Marguerat, S., and Bähler, J. (2014). LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. Genome Res. *24*, 1169–1179.

Black, D.L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? Genes Dev. *5*, 389–402.

Blazquez, L., Emmett, W., Faraway, R., Pineda, J.M.B., Bajew, S., Gohr, A., Haberman, N., Sibley, C.R., Bradley, R.K., Irimia, M., et al. (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. Mol. Cell *72*, 496–509.e9.

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

Bochman, M.L., Paeschke, K., and Zakian, V.A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. *13*, 770–780.

Boehm, V., Britto-Borges, T., Steckelberg, A.-L., Singh, K.K., Gerbracht, J.V., Gueney, E., Blazquez, L., Altmüller, J., Dieterich, C., and Gehring, N.H. (2018). Exon Junction Complexes Suppress Spurious Splice Sites to Safeguard Transcriptome Integrity. Mol. Cell *72*, 482–495.e7.

Bornstein, S.R., Ehrhart-Bornstein, M., Androutsellis-Theotokis, A., Eisenhofer, G., Vukicevic, V., Licinio, J., Wong, M.L., Calissano, P., Nisticò, G., Preziosi, P., et al. (2012). Chromaffin cells: the peripheral brain. Mol. Psychiatry *17*, 354–358.

Brackenridge, S. (2003). Efficient use of a “dead-end” GA 5’ splice site in the human fibroblast growth factor receptor genes. The EMBO Journal *22*, 1620–1631.

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

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., and Blencowe, B.J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. *24*, 1774–1786.

Brinegar, A.E., Xia, Z., Loehr, J.A., Li, W., Rodney, G.G., and Cooper, T.A. (2017). Extensive alternative splicing transitions during postnatal skeletal muscle development are required for calcium handling functions.

Brugge, J.S., Cotton, P.C., Queral, A.E., Barrett, J.N., Nonner, D., and Keane, R.W. (1985). Neurones express high levels of a structurally modified, activated form of pp60c-src. Nature *316*, 554–557.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell *46*, 871–883.

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

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

Burge, C.B., Tuschl, T., and Sharp, P.A. (1999). Splicing of precursors to mRNAs by the spliceosomes. Cold Spring Harbor Monogr. Ser. *37*, 525–560.

Burset, M. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Research *28*, 4364–4375.

Calarco, J.A., Superina, S., O'Hanlon, D., Gabut, M., Raj, B., Pan, Q., Skalska, U., Clarke, L., Gelinas, D., van der Kooy, D., et al. (2009). Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. Cell *138*, 898–910.

Capra, J.A., Paeschke, K., Singh, M., and Zakian, V.A. (2010). G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. PLoS Comput. Biol. *6*, e1000861.

Caputi, M., and Zahler, A.M. (2001). Determination of the RNA Binding Specificity of the Heterogeneous Nuclear Ribonucleoprotein (hnRNP) H/H′/F/2H9 Family. J. Biol. Chem. *276*, 43850–43859.

Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5' splice-site positions. RNA *10*, 828–840.

Catterall, W.A. (2011). Voltage-gated calcium channels. Cold Spring Harb. Perspect. Biol. *3*, a003947.

Cer, R.Z., Donohue, D.E., Mudunuri, U.S., Temiz, N.A., Loss, M.A., Starner, N.J., Halusa, G.N., Volfovsky, N., Yi, M., Luke, B.T., et al. (2013). Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res. *41*, D94–D100.

Chakraborty, P., and Grosse, F. (2011). Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes. DNA Repair *10*, 654–665.

Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P., and Balasubramanian, S. (2015). High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat. Biotechnol. *33*, 877–881.

Chan, W.H., Komada, M., Fukushima, T., Southard-Smith, E.M., Anderson, C.R., and Wakefield, M.J. (2019). RNA-seq of Isolated Chromaffin Cells Highlights the Role of Sex-Linked and Imprinted Genes in Adrenal Medulla Development. Sci. Rep. *9*, 3929.

Chang, H., Lim, J., Ha, M., and Kim, V.N. (2014). TAIL-seq: genome-wide determination of poly(A) tail length and 3’ end modifications. Mol. Cell *53*, 1044–1052.

Chen, L., Chen, J.-Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., Tang, P., Qian, H., Luo, D., Li, H., et al. (2017). R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. Mol. Cell *68*, 745–757.e5.

Chernyavsky, A.I., Arredondo, J., Piser, T., Karlsson, E., and Grando, S.A. (2008). Differential Coupling of M1 Muscarinic and α7 Nicotinic Receptors to Inhibition of Pemphigus Acantholysis. J. Biol. Chem. *283*, 3401–3408.

Cocquerelle, C., Mascrez, B., Hétuin, D., and Bailleul, B. (1993). Mis-splicing yields circular RNA molecules. FASEB J. *7*, 155–160.

Cocquet, J., Chong, A., Zhang, G., and Veitia, R.A. (2006). Reverse transcriptase template switching and false alternative transcripts. Genomics *88*, 127–131.

Coelho, M.B., and Smith, C.W.J. (2014). Regulation of Alternative Pre-mRNA Splicing. In Spliceosomal Pre-mRNA Splicing: Methods and Protocols, K.J. Hertel, ed. (Totowa, NJ: Humana Press), pp. 55–82.

Cogoi, S., and Xodo, L.E. (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Res. *34*, 2536–2549.

Collett, J.W., and Steele, R.E. (1992). Identification and developmental expression of Src+ mRNAs in Xenopus laevis. Dev. Biol. *152*, 194–198.

Coolidge, C.J., Seely, R.J., and Patton, J.G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. Nucleic Acids Res. *25*, 888–896.

Cooper, T.A., and Ordahl, C.P. (1985). A single cardiac troponin T gene generates embryonic and adult isoforms via developmentally regulated alternate splicing. J. Biol. Chem. *260*, 11140–11148.

Cooper, D.A., Cortés-López, M., and Miura, P. (2018). Genome-Wide circRNA Profiling from RNA-seq Data. In Circular RNAs: Methods and Protocols, C. Dieterich, and A. Papantonis, eds. (New York, NY: Springer New York), pp. 27–41.

Cox, J.S., and Walter, P. (1996). A novel mechanism for regulating activity of a transcription factor that controls the unfolded protein response. Cell *87*, 391–404.

Crossley, M.P., Bocek, M., and Cimprich, K.A. (2019). R-Loops as Cellular Regulators and Genomic Threats. Mol. Cell *73*, 398–411.

DeBoever, C., Ghia, E.M., Shepard, P.J., Rassenti, L., Barrett, C.L., Jepsen, K., Jamieson, C.H.M., Carson, D., Kipps, T.J., and Frazer, K.A. (2015). Transcriptome sequencing reveals potential mechanism of cryptic 3’ splice site selection in SF3B1-mutated cancers. PLoS Comput. Biol. *11*, e1004105.

De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA *4*, 49–60.

Desai, A., Hu, Z., French, C.E., Lloyd, J.P.B., and Brenner, S.E. (2020). Networks of Splice Factor Regulation by Unproductive Splicing Coupled With Nonsense Mediated mRNA Decay.

Dias Neto, E., Correa, R.G., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva, W., Jr, Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., et al. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc. Natl. Acad. Sci. U. S. A. *97*, 3491–3496.

Didiot, M.-C., Tian, Z., Schaeffer, C., Subramanian, M., Mandel, J.-L., and Moine, H. (2008). The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer. Nucleic Acids Res. *36*, 4902–4912.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319.

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

Dominguez, C., Fisette, J.-F., Chabot, B., and Allain, F.H.-T. (2010). Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. Nat. Struct. Mol. Biol. *17*, 853–861.

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., et al. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell *70*, 854–867.e9.

Dominski, Z., and Kole, R. (1991). Selection of splice sites in pre-mRNAs with short internal exons. Mol. Cell. Biol. *11*, 6075–6083.

Downie, J.M. (2017). Surveying the Genetic Risk Landscape Of Amyotrophic Lateral Sclerosis in The Era of Next-Generation Sequencing. The University of Utah.

Du, X., Gertz, E.M., Wojtowicz, D., Zhabinskaya, D., Levens, D., Benham, C.J., Schäffer, A.A., and Przytycka, T.M. (2014). Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. Nucleic Acids Res. *42*, 12367–12379.

Duff, M.O., Olson, S., Wei, X., Garrett, S.C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S.E., and Graveley, B.R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature *521*, 376–379.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Muñoz, M.J., Alló, M., et al. (2013). Transcriptional elongation and alternative splicing. Biochim. Biophys. Acta *1829*, 134–140.

Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., and Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev. *18*, 1618–1629.

Dye, M.J., Gromak, N., and Proudfoot, N.J. (2006). Exon tethering in transcription by RNA polymerase II. Mol. Cell *21*, 849–859.

Eddy, J., and Maizels, N. (2008). Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Res. *36*, 1321–1333.

Ehlers, M.D., Tingley, W.G., and Huganir, R.L. (1995). Regulated subcellular distribution of the NR1 subunit of the NMDA receptor. Science *269*, 1734–1737.

Ehrmann, I., Crichton, J.H., Gazzara, M.R., James, K., Liu, Y., Grellscheid, S.N., Curk, T., de Rooij, D., Steyn, J.S., Cockell, S., et al. (2019). An ancient germ cell-specific RNA-binding protein protects the germline from cryptic splice site poisoning. Elife *8*.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol. Cell *46*, 884–892.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigó, R., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods *10*, 1185–1191.

Erkelenz, S., Theiss, S., Kaisers, W., Ptok, J., Walotka, L., Müller, L., Hillebrand, F., Brillen, A.-L., Sladek, M., and Schaal, H. (2018). Ranking noncanonical 5' splice site usage by genome-wide

RNA-seq analysis and splicing reporter assays. Genome Res. *28*, 1826–1840.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Fay, M.M., Lyons, S.M., and Ivanov, P. (2017). RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms. J. Mol. Biol. *429*, 2127–2147.

Fiszbein, A., and Kornblihtt, A.R. (2017). Alternative splicing switches: Important players in cell differentiation. Bioessays *39*.

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

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.

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

Frazee, A.C., Jaffe, A.E., Langmead, B., and Leek, J.T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics *31*, 2778–2784.

Furlanis, E., Traunmüller, L., Fucile, G., and Scheiffele, P. (2019). Landscape of ribosome-engaged transcript isoforms reveals extensive neuronal-cell-class-specific alternative splicing programs. Nat. Neurosci.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of nucleosome positioning in the human genome. PLoS Genet. *8*, e1003036.

Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods *15*, 201–206.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods *8*, 469–477.

Garg, K., and Green, P. (2007). Differing patterns of selection in alternative and constitutive splice sites. Genome Res. *17*, 1015–1022.

Garrido-Martín, D., Palumbo, E., Guigó, R., and Breschi, A. (2018). ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLoS Comput. Biol. *14*, e1006360.

Gaspard, N., Bouschet, T., Hourez, R., Dimidschstein, J., Naeije, G., van den Ameele, J., Espuny-Camacho, I., Herpoel, A., Passante, L., Schiffmann, S.N., et al. (2008). An intrinsic mechanism of corticogenesis from embryonic stem cells. Nature *455*, 351–357.

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. *22*, 35–50.

Gellert, M., Lipsett, M.N., and Davies, D.R. (1962). Helix formation by guanylic acid. Proc. Natl. Acad.

Sci. U. S. A. *48*, 2013–2018.

Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M., and Nik-Zainal, S. (2018). Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res. *28*, 1264–1271.

Georgakopoulos-Soares, I., Parada, G.E., Wong, H.Y., Miska, E.A., Kwok, C.K., and Hemberg, M. Alternative splicing modulation by G-quadruplexes.

Georgakopoulos-Soares, I., Koh, G., Jiricny, J., Hemberg, M., and Nik-Zainal, S. Transcription-coupled repair and mismatch repair contribute towards preserving genome integrity at mononucleotide repeat tracts.

Ghosh, A., and Bansal, M. (2003). A glossary of DNA structures from A to Z. Acta Crystallographica Section D Biological Crystallography *59*, 620–626.

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11*, R86.

Gokce, O., Stanley, G.M., Treutlein, B., Neff, N.F., Camp, J.G., Malenka, R.C., Rothwell, P.E., Fuccillo, M.V., Südhof, T.C., and Quake, S.R. (2016). Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. Cell Rep. *16*, 1126–1137.

Gomez, D., Lemarteleur, T., Lacroix, L., Mailliet, P., Mergny, J.-L., and Riou, J.-F. (2004). Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. Nucleic Acids Res. *32*, 371–379.

Gonatopoulos-Pournatzis, T., and Blencowe, B.J. (2020). Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder. Curr. Opin. Genet. Dev. *65*, 22–33.

Gonatopoulos-Pournatzis, T., Wu, M., Braunschweig, U., Roth, J., Han, H., Best, A.J., Raj, B., Aregger, M., O'Hanlon, D., Ellis, J.D., et al. (2018). Genome-wide CRISPR-Cas9 Interrogation of Splicing Networks Reveals a Mechanism for Recognition of Autism-Misregulated Neuronal Microexons. Mol. Cell *72*, 510–524.e12.

Gonatopoulos-Pournatzis, T., Aregger, M., Brown, K.R., Farhangmehr, S., Braunschweig, U., Ward, H.N., Ha, K.C.H., Weiss, A., Billmann, M., Durbic, T., et al. (2020). Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9-Cas12a platform. Nat. Biotechnol. *38*, 638–648.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. *17*, 100–107.

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

Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., Köster, J., and Bioconda Team (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat. Methods *15*, 475–476.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Guiblet, W.M., Cremona, M.A., Cechova, M., Harris, R.S., Kejnovská, I., Kejnovsky, E., Eckert, K., Chiaromonte, F., and Makova, K.D. (2018). Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. *28*, 1767–1778.

Guo, J.U., and Bartel, D.P. (2016). RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science *353*.

Guo, M., Lo, P.C., and Mount, S.M. (1993). Species-specific signals for the splicing of a short Drosophila intron *in vitro*. Mol. Cell. Biol. *13*, 1104–1118.

Guth, S., Martínez, C., Gaur, R.K., and Valcárcel, J. (1999). Evidence for substrate-specific requirement of the splicing factor U2AF(35) and for its function after polypyrimidine tract recognition by U2AF(65). Mol. Cell. Biol. *19*, 8263–8271.

Han, K., Yeo, G., An, P., Burge, C.B., and Grabowski, P.J. (2005). A combinatorial code for splicing silencing: UAGG and GGGG motifs. PLoS Biol. *3*, e158.

Hänsel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., et al. (2016). G-quadruplex structures mark human regulatory chromatin. Nat. Genet. *48*, 1267–1272.

Hänsel-Hertsch, R., Di Antonio, M., and Balasubramanian, S. (2017). DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. Nat. Rev. Mol. Cell Biol. *18*, 279–284.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol. *7 Suppl 1*, S4.1–9.

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

Hermey, G., Blüthgen, N., and Kuhl, D. (2017). Neuronal activity-regulated alternative mRNA splicing. Int. J. Biochem. Cell Biol. *91*, 184–193.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC genome browser database: update 2006. Nucleic Acids Res. *34*, D590–D598.

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., and Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science *325*, 626–628.

Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblihtt, A.R., and Ast, G. (2016). How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? Trends Genet. *32*, 596–606.

Hood, L., and Galas, D. (2003). The digital code of DNA. Nature *421*, 444–448.

Houseley, J., and Tollervey, D. (2010). Apparent non-canonical trans-splicing is generated by reverse transcriptase *in vitro*. PLoS One *5*, e12271.

Howe, K.J., and Ares, M., Jr (1997). Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. Proc. Natl. Acad. Sci. U. S. A. *94*, 12467–12472.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinformatics *22*, 1036–1046.

Huang, H., Zhang, J., Harvey, S.E., Hu, X., and Cheng, C. (2017). RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. Genes Dev. *31*, 2296–2309.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Huntsman, M.M., Tran, B.-V., Potkin, S.G., Bunney, W.E., and Jones, E.G. (1998). Altered ratios of alternatively spliced long and short γ2 subunit mRNAs of the γ-amino butyrate type A receptor in prefrontal cortex of schizophrenics. Proc. Natl. Acad. Sci. U. S. A. *95*, 15066–15071.

Huppert, J.L., and Balasubramanian, S. (2007). G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res. *35*, 406–413.

Hurley, L.H., Von Hoff, D.D., Siddiqui-Jain, A., and Yang, D. (2006). Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. Semin. Oncol. *33*, 498–512.

Iqbal, Z., Willemsen, M.H., Papon, M.-A., Musante, L., Benevento, M., Hu, H., Venselaar, H., Wissink-Lindhout, W.M., Vulto-van Silfhout, A.T., Vissers, L.E.L.M., et al. (2015). Homozygous SLC6A17 mutations cause autosomal-recessive intellectual disability with progressive tremor, speech impairment, and behavioral problems. Am. J. Hum. Genet. *96*, 386–396.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell *159*, 1511–1523.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. *19*, 3795–3798.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA *19*, 141–157.

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics *9*, 192.

Jin, Y., Yang, Y., and Zhang, P. (2011). New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. RNA Biol. *8*, 450–457.

Kadener, S. (2001). Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. The EMBO Journal *20*, 5759–5768.

Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research Network, et al. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211–224.e6.

Kamiguchi, H., and Lemmon, V. (1998). A neuronal form of the cell adhesion molecule L1 contains a tyrosine-based signal required for sorting to the axonal growth cone. J. Neurosci. *18*, 3749–3756.

Kamiguchi, H., Long, K.E., Pendergast, M., Schaefer, A.W., Rapoport, I., Kirchhausen, T., and Lemmon, V. (1998). The neural cell adhesion molecule L1 interacts with the AP-2 adaptor and is endocytosed via the clathrin-mediated pathway. J. Neurosci. *18*, 5311–5321.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51–54.

Kaushik, M., Kaushik, S., Roy, K., Singh, A., Mahendru, S., Kumar, M., Chaudhary, S., Ahmed, S., and Kukreti, S. (2016). A bouquet of DNA structures: Emerging diversity. Biochem Biophys Rep *5*, 388–395.

Keenan, S., Wetherill, S.J., Ugbode, C.I., Chawla, S., Brackenbury, W.J., and Evans, G.J.O. (2017). Inhibition of N1-Src kinase by a specific SH3 peptide ligand reveals a role for N1-Src in neurite elongation by L1-CAM. Sci. Rep. *7*, 43106.

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

Kikin, O., D'Antonio, L., and Bagga, P.S. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. *34*, W676–W682.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kim, D., Langmead, B., and Salzberg, S. (2017). HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes.

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

Kim, E., Goren, A., and Ast, G. (2008). Alternative splicing: current perspectives. Bioessays *30*, 38–47.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-Kwon, K.-R., Benham, C.J., Casellas, R., Przytycka, T.M., et al. (2017). Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Syst *4*, 344–356.e7.

Královicová, J., and Vorechovsky, I. (2006). Position-dependent repression and promotion of DQB1 intron 3 splicing by GGGG motifs. J. Immunol. *176*, 2381–2388.

Kwok, C.K., and Merrick, C.J. (2017). G-Quadruplexes: Prediction, Characterization, and Biological Application. Trends Biotechnol. *35*, 997–1013.

Kwok, C.K., Marsico, G., Sahakyan, A.B., Chambers, V.S., and Balasubramanian, S. (2016). rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nat. Methods *13*, 841–844.

Kwok, C.K., Marsico, G., and Balasubramanian, S. (2018). Detecting RNA G-Quadruplexes (rG4s) in the Transcriptome. Cold Spring Harb. Perspect. Biol. *10*.

Lagarde, J., Uszczynska-Ratajczak, B., Santoyo-Lopez, J., Gonzalez, J.M., Tapanari, E., Mudge, J.M., Steward, C.A., Wilming, L., Tanzer, A., Howald, C., et al. (2016). Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). Nat. Commun. *7*, 12339.

Lam, E.Y.N., Beraldi, D., Tannahill, D., and Balasubramanian, S. (2013). G-quadruplex structures are stable and detectable in human genomic DNA. Nat. Commun. *4*, 1796.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–900.

Lambowitz, A.M., and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb. Perspect. Biol. *3*, a003616.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lareau, L.F., Green, R.E., Bhatnagar, R.S., and Brenner, S.E. (2004). The evolving roles of alternative splicing. Curr. Opin. Struct. Biol. *14*, 273–282.

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

Larsonneur, E., Mercier, J., Wiart, N., Floch, E.L., Delhomme, O., and Meyer, V. (2018). Evaluating Workflow Management Systems: A Bioinformatics Use Case. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2773–2775.

Lau, E., Han, Y., Williams, D.R., Thomas, C.T., Shrestha, R., Wu, J.C., and Lam, M.P.Y. (2019). Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. Cell Rep. *29*, 3751–3765.e5.

Lee, M.S., and Ji, Q.C. (2017). Protein Analysis using Mass Spectrometry: Accelerating Protein Biotherapeutics from Lab to Patient (John Wiley & Sons).

Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu. Rev. Biochem. *84*, 291–323.

Lee, C.-Y., McNerney, C., Ma, K., Zhao, W., Wang, A., and Myong, S. (2020). R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. Nat. Commun. *11*, 3392.

Lee, J.-A., Xing, Y., Nguyen, D., Xie, J., Lee, C.J., and Black, D.L. (2007). Depolarization and CaM kinase IV modulate NMDA receptor splicing through two essential RNA elements. PLoS Biol. *5*, e40.

Lee, J.-A., Tang, Z.-Z., and Black, D.L. (2009). An inducible change in Fox-1/A2BP1 splicing modulates the alternative splicing of downstream neuronal target exons. Genes Dev. *23*, 2284–2293.

Le Hir, H., Saulière, J., and Wang, Z. (2016). The exon junction complex as a node of post-transcriptional networks. Nat. Rev. Mol. Cell Biol. *17*, 41–54.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. Brief. Bioinform. *18*, 530–536.

Leung, D.W., and Amarasinghe, G.K. (2016). When your cap matters: structural insights into self vs non-self recognition of 5’ RNA by immunomodulatory host proteins. Curr. Opin. Struct. Biol. *36*, 133–141.

Leung, M.K.K., Xiong, H.Y., Lee, L.J., and Frey, B.J. (2014). Deep learning of the tissue-regulated splicing code. Bioinformatics *30*, i121–i129.

Levine, A., and Durbin, R. (2001). A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res. *29*, 4006–4013.

Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E.Y., and Ast, G. (2008). Intronic Alus influence alternative splicing. PLoS Genet. *4*, e1000204.

Levy, J.B., Dorai, T., Wang, L.H., and Brugge, J.S. (1987). The structurally distinct form of pp60c-src detected in neuronal cells is encoded by a unique c-src mRNA. Mol. Cell. Biol. *7*, 4142–4145.

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

Lewis, C.J.T., Pan, T., and Kalsotra, A. (2017a). RNA modifications and structures cooperate to guide RNA-protein interactions. Nat. Rev. Mol. Cell Biol. *18*, 202–210.

Lewis, P.A., Bradley, I.C., Pizzey, A.R., Isaacs, H.V., and Evans, G.J.O. (2017b). N1-Src Kinase Is Required for Primary Neurogenesis in Xenopus tropicalis. J. Neurosci. *37*, 8477–8485.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, X., and Manley, J.L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. Cell *122*, 365–378.

Li, H., Liu, G., Yu, J., Cao, W., Lobo, V.G., and Xie, J. (2009). In vivo selection of kinase-responsive

RNA elements controlling alternative splicing. J. Biol. Chem. *284*, 16191–16201.

Li, Y., Li, C., Li, S., Peng, Q., An, N.A., He, A., and Li, C.-Y. (2018). Human exonization through differential nucleosome occupancy. Proc. Natl. Acad. Sci. U. S. A. *115*, 8817–8822.

Li, Y.I., Sanchez-Pulido, L., Haerty, W., and Ponting, C.P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. *25*, 1–13.

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

Lim, Y., Han, I., Kwon, H.J., and Oh, E.-S. (2002). Trichostatin A-induced detransformation correlates with decreased focal adhesion kinase phosphorylation at tyrosine 861 in ras-transformed fibroblasts. J. Biol. Chem. *277*, 12735–12740.

Lin, C.-F., Mount, S.M., Jarmołowski, A., and Makałowski, W. (2010). Evolutionary dynamics of U12-type spliceosomal introns. BMC Evol. Biol. *10*, 47.

Lin, C.-L., Taggart, A.J., Lim, K.H., Cygan, K.J., Ferraris, L., Creton, R., Huang, Y.-T., and Fairbrother, W.G. (2016). RNA structure replaces the need for U2AF2 in splicing. Genome Res. *26*, 12–23.

Liu, G., Razanau, A., Hai, Y., Yu, J., Sohail, M., Lobo, V.G., Chu, J., Kung, S.K.P., and Xie, J. (2012). A conserved serine of heterogeneous nuclear ribonucleoprotein L (hnRNP L) mediates depolarization-regulated alternative splicing of potassium channels. J. Biol. Chem. *287*, 22709–22716.

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

Louie, A.L., Aigner, S., Bergalet, J., Zhou, B., and Su, A. (2018). A large-scale binding and functional map of human RNA binding proteins. bioRxiv.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat. Struct. Mol. Biol. *20*, 1434–1442.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell *144*, 16–26.

Lukacsovich, D., Winterer, J., Que, L., Luo, W., Lukacsovich, T., and Földy, C. (2019). Single-Cell RNA-Seq Reveals Developmental Origins and Ontogenetic Stability of Neurexin Alternative Splicing Profiles. Cell Rep. *27*, 3752–3759.e4.

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.

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

Mader, R.M., Schmidt, W.M., Sedivy, R., Rizovski, B., Braun, J., Kalipciyan, M., Exner, M., Steger, G.G., and Mueller, M.W. (2001). Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. J. Lab. Clin. Med. *137*, 422–428.

Maizels, N., and Gray, L.T. (2013). The G4 genome. PLoS Genet. *9*, e1003468.

Marcel, V., Tran, P.L.T., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M.-P., Hall, J., Mergny, J.-L., Hainaut, P., and Van Dyck, E. (2011). G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis *32*, 271–278.

Marcucci, R., Baralle, F.E., and Romano, M. (2007). Complex splicing control of the human

Thrombopoietin gene by intronic G runs. Nucleic Acids Res. *35*, 132–142.

Marnef, A., Cohen, S., and Legube, G. (2017). Transcription-Coupled DNA Double-Strand Break Repair: Active Genes Need Special Care. J. Mol. Biol. *429*, 1277–1288.

Marsico, G., Chambers, V.S., Sahakyan, A.B., McCauley, P., Boutell, J.M., Di Antonio, M., and Balasubramanian, S. (2019a). Whole genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic Acids Research *47*, 3862–3874.

Marsico, G., Chambers, V.S., Sahakyan, A.B., McCauley, P., Boutell, J.M., Di Antonio, M., and Balasubramanian, S. (2019b). Whole genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic Acids Res.

Martinez, R., Mathey-Prevot, B., Bernards, A., and Baltimore, D. (1987). Neuronal pp60c-src contains a six-amino acid insertion relative to its non-neuronal counterpart. Science *237*, 411–415.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. Mol. Cell *12*, 525–532.

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

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

Mauger, D.M., Lin, C., and Garcia-Blanco, M.A. (2008). hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc. Mol. Cell. Biol. *28*, 5403–5419.

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

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

McGlincy, N.J., and Smith, C.W.J. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends Biochem. Sci. *33*, 385–393.

McManus, C.J., and Graveley, B.R. (2011). RNA structure and the mechanisms of alternative splicing. Curr. Opin. Genet. Dev. *21*, 373–379.

McNally, L.M., Yee, L., and McNally, M.T. (2006). Heterogeneous nuclear ribonucleoprotein H is required for optimal U11 small nuclear ribonucleoprotein binding to a retroviral RNA-processing control element: implications for U12-dependent RNA splicing. J. Biol. Chem. *281*, 2478–2488.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.

Merrill, R.A., Plum, L.A., Kaiser, M.E., and Clagett-Dame, M. (2002). A mammalian homolog of unc-53 is regulated by all-trans retinoic acid in neuroblastoma cells and embryos. Proc. Natl. Acad. Sci. U. S. A. *99*, 3422–3427.

Metzakopian, E., Bouhali, K., Alvarez-Saavedra, M., Whitsett, J.A., Picketts, D.J., and Ang, S.-L. (2015). Genome-wide characterisation of Foxa1 binding sites reveals several mechanisms for regulating neuronal differentiation in midbrain dopamine cells. Development *142*, 1315–1324.

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

Miyoshi, D., Nakao, A., and Sugimoto, N. (2003). Structural transition from antiparallel to parallel

G-quadruplex of d(G4T4G4) induced by Ca2+. Nucleic Acids Res. *31*, 1156–1163.

Montell, C., Fisher, E.F., Caruthers, M.H., and Berk, A.J. (1982). Resolving the functions of overlapping viral genes by site-specific mutagenesis at a mRNA splice site. Nature *295*, 380–384.

Morgan, M., Much, C., DiGiacomo, M., Azzi, C., Ivanova, I., Vitsios, D.M., Pistolic, J., Collier, P., Moreira, P.N., Benes, V., et al. (2017). mRNA 3' uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. Nature *548*, 347–351.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Moss, S.J., Doherty, C.A., and Huganir, R.L. (1992). Identification of the cAMP-dependent protein kinase and protein kinase C phosphorylation sites within the major intracellular domains of the beta 1, gamma 2S, and …. J. Biol. Chem.

Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., and Fields, C. (1992). Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res. *20*, 4255–4262.

Müllner, D., and Others (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. *53*, 1–18.

Murtagh, F., and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J. Classification *31*, 274–295.

Nahkuri, S., Taft, R.J., and Mattick, J.S. (2009). Nucleosomes are preferentially positioned at exons in somatic and sperm cells. Cell Cycle *8*, 3420–3424.

Nakagaki-Silva, E.E., Gooding, C., Llorian, M., Jacob, A.G., Richards, F., Buckroyd, A., Sinha, S., and Smith, C.W. (2019). Identification of RBPMS as a mammalian smooth muscle master splicing regulator via proximity of its gene with super-enhancers. Elife *8*.

Nasiri, A.H., Wurm, J.P., Immer, C., and Weickhmann, A.K. (2016). An intermolecular G-quadruplex as the basis for GTP recognition in the class V–GTP aptamer. RNA.

Nguyen, H.D., Zou, L., and Graubert, T.A. (2019). Targeting R-loop-associated ATR response in myelodysplastic syndrome. Oncotarget *10*, 2581–2582.

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

Nickless, A., Bailis, J.M., and You, Z. (2017). Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci. *7*, 26.

Nieto Moreno, N., Giono, L.E., Cambindo Botto, A.E., Muñoz, M.J., and Kornblihtt, A.R. (2015). Chromatin, DNA structure and alternative splicing. FEBS Lett. *589*, 3370–3378.

Nogués, G., Kadener, S., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2002). Transcriptional Activators Differ in Their Abilities to Control Alternative Splicing. J. Biol. Chem. *277*, 43110–43114.

Novikova, O., and Belfort, M. (2017). Mobile Group II Introns as Ancestral Eukaryotic Elements. Trends Genet. *33*, 773–783.

Ohnishi, T., Shirane, M., and Nakayama, K.I. (2017). SRRM4-dependent neuron-specific alternative splicing of protrudin transcripts regulates neurite outgrowth. Sci. Rep. *7*, 41130.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of

60,770 full-length cDNAs. Nature *420*, 563–573.

Padgett, R.A. (2012). New connections between splicing and human disease. Trends Genet. *28*, 147–154.

Paeschke, K., Capra, J.A., and Zakian, V.A. (2011). DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. Cell *145*, 678–691.

Pai, A.A., Paggi, J.M., Yan, P., Adelman, K., and Burge, C.B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. PLoS Genet. *14*, e1007588.

Pan, Q. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes & Development *20*, 153–158.

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

Pannunzio, N.R., and Lieber, M.R. (2018). Concept of DNA Lesion Longevity and Chromosomal Translocations. Trends Biochem. Sci. *43*, 490–498.

Paoletti, P., Bellone, C., and Zhou, Q. (2013). NMDA receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. Nat. Rev. Neurosci. *14*, 383–400.

Papasaikas, P., and Valcárcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends Biochem. Sci. *41*, 33–45.

Parada, G.E., Munita, R., Cerda, C.A., and Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. Nucleic Acids Res. *42*, 10564–10578.

Park, J., and Belden, W.J. (2018). Long non-coding RNAs have age-dependent diurnal expression that coincides with age-related changes in genome-wide facultative heterochromatin. BMC Genomics *19*, 777.

Parras, A., Anta, H., Santos-Galindo, M., Swarup, V., Elorza, A., Nieto-González, J.L., Picó, S., Hernández, I.H., Díaz-Hernández, J.I., Belloc, E., et al. (2018). Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. Nature *560*, 441–446.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K., Pandey, A., and Salzberg, S.L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. *19*, 208.

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

Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. *32*, 577–591.

Pipes, L., Li, S., Bozinoski, M., Palermo, R., Peng, X., Blood, P., Kelly, S., Weiss, J.M., Thierry-Mieg, J., Thierry-Mieg, D., et al. (2013). The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. Nucleic Acids Res. *41*, D906–D914.

Placek, K., Baer, G.M., Elman, L., McCluskey, L., Hennessy, L., Ferraro, P.M., Lee, E.B., Lee, V.M.Y., Trojanowski, J.Q., Van Deerlin, V.M., et al. (2019). UNC13A polymorphism contributes to frontotemporal disease in sporadic amyotrophic lateral sclerosis. Neurobiol. Aging *73*, 190–199.

Popp, M.W.-L., and Maquat, L.E. (2013). Organizing principles of mammalian nonsense-mediated mRNA decay. Annu. Rev. Genet. *47*, 139–165.

Porter, R.S., Jaamour, F., and Iwase, S. (2018). Neuron-specific alternative splicing of transcriptional machineries: Implications for neurodevelopmental disorders. Mol. Cell. Neurosci. *87*, 35–45.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–D763.

PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., et al. (2015). The PsychENCODE project. Nat. Neurosci. *18*, 1707–1712.

Pucker, B., and Brockington, S.F. (2018). Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. BMC Genomics *19*, 980.

Pulido, R., Krueger, N.X., Serra-Pagès, C., Saito, H., and Streuli, M. (1995a). Molecular Characterization of the Human Transmembrane Protein-tyrosine Phosphatase δ: EVIDENCE FOR TISSUE-SPECIFIC EXPRESSION OF ALTERNATIVE HUMAN TRANSMEMBRANE PROTEIN-TYROSINE PHOSPHATASE δ ISOFORMS. J. Biol. Chem. *270*, 6722–6728.

Pulido, R., Serra-Pagès, C., Tang, M., and Streuli, M. (1995b). The LAR/PTP delta/PTP sigma subfamily of transmembrane protein-tyrosine-phosphatases: multiple human LAR, PTP delta, and PTP sigma isoforms are expressed in a tissue-specific manner and associate with the LAR-interacting protein LIP.1. Proc. Natl. Acad. Sci. U. S. A. *92*, 11686–11690.

Qiu, J., McQueen, J., Bilican, B., Dando, O., Magnani, D., Punovuori, K., Selvaraj, B.T., Livesey, M., Haghi, G., Heron, S., et al. (2016). Evidence for evolutionary divergence of activity-dependent gene expression in developing neurons. Elife *5*.

Quesnel-Vallières, M., Irimia, M., Cordes, S.P., and Blencowe, B.J. (2015). Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. Genes Dev. *29*, 746–759.

Quesnel-Vallières, M., Dargaei, Z., Irimia, M., Gonatopoulos-Pournatzis, T., Ip, J.Y., Wu, M., Sterne-Weiler, T., Nakagawa, S., Woodin, M.A., Blencowe, B.J., et al. (2016). Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. Mol. Cell *64*, 1023–1034.

Raj, B., O'Hanlon, D., Vessey, J.P., Pan, Q., Ray, D., Buckley, N.J., Miller, F.D., and Blencowe, B.J. (2011). Cross-regulation between an alternative splicing activator and a transcription repressor controls neurogenesis. Mol. Cell *43*, 843–850.

Raj, B., Irimia, M., Braunschweig, U., Sterne-Weiler, T., O'Hanlon, D., Lin, Z.-Y., Chen, G.I., Easton, L.E., Ule, J., Gingras, A.-C., et al. (2014). A global regulatory mechanism for activating an exon network required for neurogenesis. Mol. Cell *56*, 90–103.

Ratcliffe, C.D.H., Siddiqui, N., Coelho, P.P., Laterreur, N., Cookey, T.N., Sonenberg, N., and Park, M. (2019). HGF-induced migration depends on the PI(3,4,5)P3-binding microexon-spliced variant of the Arf6 exchange factor cytohesin-1. J. Cell Biol. *218*, 285–298.

Ribeiro, M.M., Teixeira, G.S., Martins, L., Marques, M.R., de Souza, A.P., and Line, S.R.P. (2015). G-quadruplex formation enhances splicing efficiency of PAX9 intron 1. Hum. Genet. *134*, 37–44.

Rissland, O.S., and Norbury, C.J. (2009). Decapping is preceded by 3' uridylation in a novel pathway of bulk mRNA turnover. Nat. Struct. Mol. Biol. *16*, 616–623.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. *10*, 84–94.

Roweis, S.T. (1998). EM Algorithms for PCA and SPCA. In Advances in Neural Information

Processing Systems 10, M.I. Jordan, M.J. Kearns, and S.A. Solla, eds. (MIT Press), pp. 626–632.

Roy, S.W., and Irimia, M. (2008). When good transcripts go bad: artifactual RT-PCR “splicing” and genome analysis. Bioessays *30*, 601–605.

Rumbaugh, G., Prybylowski, K., Wang, J.F., and Vicini, S. (2000). Exon 5 and spermine regulate deactivation of NMDA receptor subtypes. J. Neurophysiol. *83*, 1300–1306.

Runfola, V., Sebastian, S., Dilworth, F.J., and Gabellini, D. (2015). Rbfox proteins regulate tissue-specific alternative splicing of Mef2D required for muscle differentiation. J. Cell Sci. *128*, 631–637.

Ruskin, B., Zamore, P.D., and Green, M.R. (1988). A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. Cell *52*, 207–219.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Mol. Cell *58*, 870–885.

Saito, Y., Yuan, Y., Zucker-Scharff, I., Fak, J.J., Jereb, S., Tajima, Y., Licatalosi, D.D., and Darnell, R.B. (2019). Differential NOVA2-Mediated Splicing in Excitatory and Inhibitory Neurons Regulates Cortical Development and Cerebellar Function. Neuron *101*, 707–720.e5.

Sakharkar, M.K., Chow, V.T.K., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Biol. *4*, 387–393.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell. Biol. *28*, 4320–4330.

Samatanga, B., Dominguez, C., Jelesarov, I., and Allain, F.H.-T. (2013). The high kinetic stability of a G-quadruplex limits hnRNP F qRRM3 binding to G-tract RNA. Nucleic Acids Res. *41*, 2505–2516.

Santoni, M.J., Barthels, D., Vopper, G., Boned, A., Goridis, C., and Wille, W. (1989). Differential exon usage involving an unusual splicing mechanism generates at least eight types of NCAM cDNA in mouse brain. EMBO J. *8*, 385–392.

Santos-Pereira, J.M., and Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. Nat. Rev. Genet. *16*, 583–597.

Sasaki-Haraguchi, N., Shimada, M.K., Taniguchi, I., Ohno, M., and Mayeda, A. (2012). Mechanistic insights into human pre-mRNA splicing of human ultra-short introns: potential unusual mechanism identifies G-rich introns. Biochem. Biophys. Res. Commun. *423*, 289–294.

Schellenberg, M.J., Ritchie, D.B., and MacMillan, A.M. (2008). Pre-mRNA splicing: a complex picture in higher definition. Trends Biochem. Sci. *33*, 243–246.

Schor, I.E., Rascovan, N., Pelisch, F., Alló, M., and Kornblihtt, A.R. (2009). Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc. Natl. Acad. Sci. U. S. A. *106*, 4325–4330.

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

Scott, D.B., Blanpied, T.A., Swanson, G.T., Zhang, C., and Ehlers, M.D. (2001). An NMDA receptor ER retention signal regulated by phosphorylation and alternative splicing. J. Neurosci. *21*, 3063–3072.

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

Semlow, D.R., and Staley, J.P. (2012). Staying on message: ensuring fidelity in pre-mRNA splicing.

Trends Biochem. Sci. *37*, 263–273.

SEQC/MAQC-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. *32*, 903–914.

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

Sharma, A., and Lou, H. (2011). Depolarization-mediated regulation of alternative splicing. Front. Neurosci. *5*, 141.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

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

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. *34*, 3955–3967.

Shimada, M.K., Sasaki-Haraguchi, N., and Mayeda, A. (2015). Identification and Validation of Evolutionarily Conserved Unusually Short Pre-mRNA Introns in the Human Genome. Int. J. Mol. Sci. *16*, 10376–10388.

Shrestha, P., Xiao, S., Dhakal, S., Tan, Z., and Mao, H. (2014). Nascent RNA transcripts facilitate the formation of G-quadruplexes. Nucleic Acids Res. *42*, 7236–7246.

Shtukmaster, S., Schier, M.C., Huber, K., Krispin, S., Kalcheim, C., and Unsicker, K. (2013). Sympathetic neurons and chromaffin cells share a common progenitor in the neural crest in vivo. Neural Dev. *8*, 12.

Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J., et al. (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371–375.

Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. Nat. Rev. Genet. *17*, 407–421.

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., and Hurley, L.H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc. Natl. Acad. Sci. U. S. A. *99*, 11593–11598.

Singh, R.K., and Cooper, T.A. (2012). Pre-mRNA splicing in disease and therapeutics. Trends Mol. Med. *18*, 472–482.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.

Small, S.J., Haines, S.L., and Akeson, R.A. (1988). Polypeptide variation in an N-CAM extracellular immunoglobulin-like fold is developmentally regulated through alternative splicing. Neuron *1*, 1007–1017.

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

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell *36*, 245–254.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164–1167.

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

Standley, S., Roche, K.W., McCallum, J., Sans, N., and Wenthold, R.J. (2000). PDZ domain suppression of an ER retention signal in NMDA receptor NR1 splice variants. Neuron *28*, 887–898.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet.

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

Sterner, D.A., Carlo, T., and Berget, S.M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci. U. S. A. *93*, 15081–15085.

Sterne-Weiler, T., Weatheritt, R.J., Best, A.J., Ha, K.C.H., and Blencowe, B.J. (2018). Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. Mol. Cell *72*, 187–200.e6.

Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc. Natl. Acad. Sci. U. S. A. *99*, 16899–16903.

Strobel, E.J., Yu, A.M., and Lucks, J.B. (2018). High-throughput determination of RNA structures. Nat. Rev. Genet. *19*, 615–634.

Su, C.-H., D, D., and Tarn, W.-Y. (2018). Alternative Splicing in Neurogenesis and Brain Development. Front Mol Biosci *5*, 12.

Südhof, T.C. (2017). Synaptic Neurexin Complexes: A Molecular Code for the Logic of Neural Circuits. Cell *171*, 745–769.

Szafranski, K., Schindler, S., Taudien, S., Hiller, M., Huse, K., Jahn, N., Schreiber, S., Backofen, R., and Platzer, M. (2007). Violating the splicing rules: TG dinucleotides function as alternative 3’ splice sites in U2-dependent introns. Genome Biol. *8*, R154.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368.

Takahashi, H., and Craig, A.M. (2013). Protein tyrosine phosphatases PTPδ, PTPσ, and LAR: presynaptic hubs for synapse organization. Trends Neurosci. *36*, 522–534.

Talerico, M., and Berget, S.M. (1994). Intron definition in splicing of small Drosophila introns. Mol. Cell. Biol. *14*, 3434–3445.

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., and Burge, C.B. (2016). RNA Sequence Context Effects Measured *In Vitro* Predict In Vivo Protein Binding and Regulation. Mol. Cell *64*, 294–306.

Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768.

Tarn, W.Y., and Steitz, J.A. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron *in vitro*. Cell *84*, 801–811.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346.

Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature *563*, 72–78.

Thakurela, S., Garding, A., Jung, J., Schübeler, D., Burger, L., and Tiwari, V.K. (2013). Gene regulation and priming by topoisomerase IIα in embryonic stem cells. Nat. Commun. *4*, 2478.

Thanaraj, T.A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res. *29*, 2581–2593.

Thomas, J.D., Polaski, J.T., Feng, Q., De Neef, E.J., Hoppe, E.R., McSharry, M.V., Pangallo, J., Gabel, A.M., Belleville, A.E., Watson, J., et al. (2020). RNA isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. Nat. Genet. *52*, 84–94.

Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996–1001.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Tipping, M.E., and Bishop, C.M. (1999). Probabilistic Principal Component Analysis. J. R. Stat. Soc. Series B Stat. Methodol. *61*, 611–622.

Torres-Méndez, A., Bonnal, S., Marquez, Y., Roth, J., Iglesias, M., Permanyer, J., Almudí, I., O'Hanlon, D., Guitart, T., Soller, M., et al. (2019). A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons. Nat Ecol Evol *3*, 691–701.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Traynelis, S.F., Hartley, M., and Heinemann, S.F. (1995). Control of proton sensitivity of the NMDA receptor by RNA splicing and polyamines. Science *268*, 873–876.

Traynelis, S.F., Burgess, M.F., Zheng, F., Lyuboslavsky, P., and Powers, J.L. (1998). Control of voltage-independent zinc inhibition of NMDA receptors by the NR1 subunit. J. Neurosci. *18*, 6163–6175.

Tress, M.L., Abascal, F., and Valencia, A. (2017a). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends in Biochemical Sciences *42*, 98–110.

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

Trotman, J.B., and Schoenberg, D.R. (2019). A recap of RNA recapping. Wiley Interdiscip. Rev. RNA *10*, e1504.

Tsai, Z.T.-Y., Chu, W.-Y., Cheng, J.-H., and Tsai, H.-K. (2014). Associations between intronic non-B DNA structures and exon skipping. Nucleic Acids Res. *42*, 739–747.

Turimella, S.L., Bedner, P., Skubal, M., Vangoor, V.R., Kaczmarczyk, L., Karl, K., Zoidl, G., Gieselmann, V., Seifert, G., Steinhäuser, C., et al. (2015). Characterization of cytoplasmic polyadenylation element binding 2 protein expression and its RNA binding activity. Hippocampus *25*, 630–642.

Ustianenko, D., Weyn-Vanhentenryck, S.M., and Zhang, C. (2017). Microexons: discovery, regulation, and function. Wiley Interdiscip. Rev. RNA *8*.

Uzilov, A.V., and Underwood, J.G. (2016). High-Throughput Nuclease Probing of RNA Structures Using FragSeq. RNA Structure Determination 105–134.

Vance, K.M., Hansen, K.B., and Traynelis, S.F. (2012). GluN1 splice variant control of GluN1/GluN2D NMDA receptors. J. Physiol. *590*, 3857–3875.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.

Varizhuk, A.M., Sekridova, A.V., Tankevich, M.V., Podgorsky, V.S., Smirnov, I.P., and Pozmogova, G.E. (2017). Conformational polymorphysm of G-rich fragments of DNA Alu-repeats. II. The putative role of G-quadruplex structures in genomic rearrangements. Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry *11*, 146–153.

Verma, B., Akinyi, M.V., Norppa, A.J., and Frilander, M.J. (2018). Minor spliceosome and disease. Semin. Cell Dev. Biol. *79*, 103–112.

Vogler, C., Spalek, K., Aerni, A., Demougin, P., Müller, A., Huynh, K.-D., Papassotiropoulos, A., and de Quervain, D.J.-F. (2009). CPEB3 is associated with human episodic memory. Front. Behav. Neurosci. *3*, 4.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380–384.

Volfovsky, N., Haas, B.J., and Salzberg, S.L. (2003). Computational discovery of internal micro-exons. Genome Res. *13*, 1216–1221.

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

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

Wamsley, B., Jaglin, X.H., Favuzzi, E., Quattrocolo, G., Nigro, M.J., Yusuf, N., Khodadadi-Jamayran, A., Rudy, B., and Fishell, G. (2018). Rbfox1 Mediates Cell-type-Specific Splicing in Cortical Interneurons. Neuron *100*, 846–859.e7.

Wang, G., and Peng, B. (2019). Script of Scripts: A pragmatic workflow system for daily computational research. PLoS Comput. Biol. *15*, e1006843.

Wang, G., and Vasquez, K.M. (2017). Effects of Replication and Transcription on DNA Structure-Related Genetic Instability. Genes *8*.

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

Wang, J., Yin, G., Menon, P., Pang, J., Smolock, E.M., Yan, C., and Berk, B.C. (2010a). Phosphorylation of G protein-coupled receptor kinase 2-interacting protein 1 tyrosine 392 is required for phospholipase C-gamma activation and podosome formation in vascular smooth muscle cells. Arterioscler. Thromb. Vasc. Biol. *30*, 1976–1982.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010b). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, X., Codreanu, S.G., Wen, B., Li, K., Chambers, M.C., Liebler, D.C., and Zhang, B. (2018).

Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. Mol. Cell. Proteomics *17*, 422–430.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. *35*, 169–178.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Wei, C.M., Gershowitz, A., and Moss, B. (1975). Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. Cell *4*, 379–386.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J.S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B.T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biol. *13*, R35.

Weitzmann, M.N., Woodford, K.J., and Usdin, K. (1996). The development and use of a DNA polymerase arrest assay for the evaluation of parameters affecting intrastrand tetraplex formation. J. Biol. Chem. *271*, 20958–20964.

Weldon, C., Dacanay, J.G., Gokhale, V., Boddupally, P.V.L., Behm-Ansmant, I., Burley, G.A., Branlant, C., Hurley, L.H., Dominguez, C., and Eperon, I.C. (2018). Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X. Nucleic Acids Res. *46*, 886–896.

Wells, S.E., Hillner, P.E., Vale, R.D., and Sachs, A.B. (1998). Circularization of mRNA by eukaryotic translation initiation factors. Mol. Cell *2*, 135–140.

Weyn-Vanhentenryck, S.M., Feng, H., Ustianenko, D., Duffié, R., Yan, Q., Jacko, M., Martinez, J.C., Goodwin, M., Zhang, X., Hengst, U., et al. (2018). Precise temporal regulation of alternative splicing during neural development. Nat. Commun. *9*, 2189.

Whiting, P., McKernan, R.M., and Iversen, L.L. (1990). Another mechanism for creating diversity in gamma-aminobutyrate type A receptors: RNA splicing directs expression of two forms of gamma 2 phosphorylation site. Proc. Natl. Acad. Sci. U. S. A. *87*, 9966–9970.

Wiestler, O.D., and Walter, G. (1988). Developmental expression of two forms of pp60c-src in mouse brain. Mol. Cell. Biol. *8*, 502–504.

Wilusz, C.J., Wormington, M., and Peltz, S.W. (2001). The cap-to-tail guide to mRNA turnover. Nat. Rev. Mol. Cell Biol. *2*, 237–246.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., et al. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. *41*, W557–W561.

Won, S.Y., and Kim, H.M. (2018). Structural Basis for LAR-RPTP-Mediated Synaptogenesis. Mol. Cells *41*, 622–630.

Wright, J.C., Mudge, J., Weisser, H., Barzine, M.P., Gonzalez, J.M., Brazma, A., Choudhary, J.S., and Harrow, J. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat. Commun. *7*, 11778.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

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

Wu, J., Anczuków, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res. *41*, 5149–5163.

Xiao, X., Wang, Z., Jang, M., Nutiu, R., Wang, E.T., and Burge, C.B. (2009). Splice site strength-dependent activity and genetic buffering by poly-G runs. Nat. Struct. Mol. Biol. *16*, 1094–1100.

Xie, J. (2008). Control of alternative pre-mRNA splicing by Ca(++) signals. Biochim. Biophys. Acta *1779*, 438–452.

Xie, J., and Black, D.L. (2001). A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. Nature *410*, 936–939.

Xie, J., Jan, C., Stoilov, P., Park, J., and Black, D.L. (2005). A consensus CaMK IV-responsive RNA sequence mediates regulation of alternative exons in neurons. RNA *11*, 1825–1834.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yamagata, A., Yoshida, T., Sato, Y., Goto-Ito, S., Uemura, T., Maeda, A., Shiroshima, T., Iwasawa-Okamoto, S., Mori, H., Mishina, M., et al. (2015a). Mechanisms of splicing-dependent trans-synaptic adhesion by PTPδ-IL1RAPL1/IL-1RAcP for synaptic differentiation. Nat. Commun. *6*, 6926.

Yamagata, A., Sato, Y., Goto-Ito, S., Uemura, T., Maeda, A., Shiroshima, T., Yoshida, T., and Fukai, S. (2015b). Structure of Slitrk2-PTPδ complex reveals mechanisms for splicing-dependent trans-synaptic adhesion. Sci. Rep. *5*, 9686.

Yang, D., and Hurley, L.H. (2006). Structure of the biologically relevant G-quadruplex in the c-MYC promoter. Nucleosides Nucleotides Nucleic Acids *25*, 951–968.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805–817.

Yang, Y., Zhan, L., Zhang, W., Sun, F., Wang, W., Tian, N., Bi, J., Wang, H., Shi, D., Jiang, Y., et al. (2011). RNA secondary structure in mutually exclusive splicing. Nat. Struct. Mol. Biol. *18*, 159–168.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. Supplement to the Journal of the Royal Statistical Society *1*, 217–235.

Yee, B.A., Pratt, G.A., Graveley, B.R., Van Nostrand, E.L., and Yeo, G.W. (2019). RBP-Maps enables robust generation of splicing regulatory maps. RNA *25*, 193–204.

Yeo, G., Holste, D., Kreiman, G., and Burge, C.B. (2004a). Variation in alternative splicing across human tissues. Genome Biol. *5*, R74.

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

You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., Akbalik, G., Wang, M., Glock, C., Quedenau, C., et al. (2015). Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. Nat. Neurosci. *18*, 603–610.

Zaia, K.A., and Reimer, R.J. (2009). Synaptic Vesicle Protein NTT4/XT1 (SLC6A17) Catalyzes Na -coupled Neutral Amino Acid Transport. Journal of Biological Chemistry *284*, 8439–8448.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453–466.

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput. Biol. *13*, e1005420.

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

Zhang, D.-H., Fujimoto, T., Saxena, S., Yu, H.-Q., Miyoshi, D., and Sugimoto, N. (2010). Monomorphic RNA G-quadruplex and polymorphic DNA G-quadruplex structures responding to cellular environmental factors. Biochemistry *49*, 4554–4563.

Zhang, J., Harvey, S.E., and Cheng, C. (2019a). A high-throughput screen identifies small molecule modulators of alternative splicing by targeting RNA G-quadruplexes. Nucleic Acids Res. *47*, 3667–3679.

Zhang, X., Chen, M.H., Wu, X., Kodani, A., Fan, J., Doan, R., Ozawa, M., Ma, J., Yoshida, N., Reiter, J.F., et al. (2016). Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell *166*, 1147–1162.e15.

Zhang, Z., Pan, Z., Ying, Y., Xie, Z., Adhikari, S., Phillips, J., Carstens, R.P., Black, D.L., Wu, Y., and Xing, Y. (2019b). Deep-learning augmented RNA-seq analysis of transcript splicing. Nat. Methods *16*, 307–310.

Zhao, J., Bacolla, A., Wang, G., and Vasquez, K.M. (2010). Non-B DNA structure-induced genetic instability and evolution. Cell. Mol. Life Sci. *67*, 43–62.

Zheng, S. (2016). Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression. International Journal of Developmental Neuroscience *55*, 102–108.