# Functional impact of alternative splicing in vertebrate proteomes across tissues and cell types

Javier Tapial Rodríguez

TESI DOCTORAL UPF / 2018

THESIS DIRECTOR:

Dr. Manuel Irimia Martínez

Centre for Genomic Regulation

Department of Systems Biology

CRG
Centre
de Regulació
Genòmica

upf.
Universitat
Pompeu Fabra
Barcelona

*“There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.”*

J. R. R. Tolkien, “The Hobbit”

# AGRADECIMIENTOS / ACKNOWLEDGEMENTS

Para mí, el doctorado siempre ha sido un viaje, uno de los más importantes de mi vida. Un viaje hacia fuera, en busca siempre de llenar la mochila con las cosas aprendidas y tratar (¡con suerte!) de adelantar siquiera un paso mínimo la frontera de lo conocido. Pero desde el principio, el viaje más importante ha sido siempre el otro, el viaje hacia dentro. En buena medida, ambos viajes han sido posibles gracias al apoyo de un gran número de personas con quienes he tenido el privilegio de compartir camino, así que no quiero desperdiciar la ocasión para mencionar a varios de ellos. Inevitablemente, olvidaré muchos nombres, así que espero que me perdonéis.

Primero de todo, a Manu. No tengo palabras para agradecerte la confianza inquebrantable que siempre has tenido en mí, en las luces y en las sombras. Gracias de corazón por tu franqueza y tu paciencia, por apoyar mis ideas, por valorar mis opiniones y por intentar sacar siempre lo mejor de mí. Gracias por respetar tanto nuestras diferencias y por las palabras de ánimo. Gracias por la disponibilidad casi sobrehumana, los intercambios febriles de correos de madrugada, y los goles por el palo corto.

Gracias también a Nuno, por la oportunidad de pasar unos meses inolvidables aprendiendo todo lo que me falta por aprender. Por ayudarme a cambiar el punto de vista, por reconciliarme con este oficio, que es una manera de vivir, y por enseñarme a ser mejor científico.

A todos los que estuvisteis conmigo en la trinchera: a Vicky, por ser ejemplo de superación y perseverancia, mirando siempre adelante sin perder la sonrisa, y por la ayuda en el empujón final; a Laura, por devolverme en el microscopio el asombro que a veces se me pierde entre las carpetas del ordenador y por todas las charlas de café; a

Chris, con quien nunca imaginé que acabaría compartiendo tanto el día que me lo crucé saliendo del despacho de Manu tras las entrevistas; a Antonio, que un día no muy lejano triunfará en la ciencia; a Patryk, por el desparpajo y el buen humor; y a Demi, por señalar el camino para los que venimos detrás sin perder el cachondeo.

A los miembros de los dos labs en los que he tenido la inmensa suerte de estar: a Yamile. por las noches de lluvia en Cracovia, por tolerar mi pésimo acento mexicano, y por todo lo que me has enseñado; a Lucía, por los mates y las risas. A Jon, por ser la voz de la experiencia. A André, por preocuparte de la salud de mi pobre teclado. A Beth, Bárbara, Thom, Marta, Cris, María y Quirze, por todas vuestras aportaciones, consejos y ánimos. A Nuno Agostinho, Marie, Mariana, Bernardo y Nora, por hacerme uno de vosotros desde el primer momento. Espero que volvamos a vernos pronto.

A toda la gente del CRG con la que he tenido la oportunidad de interaccionar, intercambiar ideas y buenos ratos. En especial, a Imma Falero, por su extraordinaria dedicación a todos y cada uno de los doctorandos del CRG. Muchas gracias por aquel mensaje, tras cuatro años creo que puedo confirmar que el clima de Barcelona efectivamente supera con mucho al de Oxford.

A Pauli, Pol y Hikaru, por todos los cafés, las charlas, las cenas, los juegos de mesa, las salidas al monte, y por recordarme que hay vida fuera del laboratorio. A todos los miembros de la Asociación Catalana de Esgrima Antigua (especialmente a Oriol, Elisabeth y José Luis), por ayudarme a recuperar una de mis grandes pasiones. A todos los futboleros del PRBB, por ayudarme a descubrir una nueva. A Marc, Edu, Toni y Jordi, por las horas de rol. Entre todos me habéis mantenido cuerdo.

Cómo no, a Álvaro, porque pasara lo que pasara siempre has tenido ánimos para dar y un rato para un café, una salida al monte, unos espadazos o alguna idea peregrina para intentar distraerme cuando las cosas no salían. Gracias por todo.

A los amigos de Toledo de los que no consigo librarme por lejos que me exilie, y por más que pasen los años: David, Anto, Guillermo, Dani, Carlos, Corbacho, Raúl, Miguel, Germán… Especialmente, a Mario, por cuidarme tanto y enseñarme a no rendirme; y a Juan, por haber traído un trocito de Toledo a Barcelona cuando más lo necesitaba. También a Carlos, por Veintiuno y por las mil preguntas respondidas sobre la próxima etapa. Y a todos los demás que esperáis en casa con los brazos abiertos.

A mi familia. Sobre todo, a mis padres, Ángel Luis y Pilar: gracias por ser ejemplo de vida, por hacerme posible perseguir mis sueños y por vuestro apoyo incondicional a cada momento, sin esperar nada a cambio. A Irene, por resistir sin cansarte mis intentos de encandilarte con la biología. A mis abuelos, Teresa, Eusebio y Sagrario, por vuestra paciencia y por recordarme lo que es verdaderamente importante. A mi abuelo Amalio, allí donde estés.

Y, por supuesto a Alicia, por estar a mi lado a cada paso de la Ruta sin desfallecer nunca. País a país, ciudad a ciudad, en cada obstáculo y en cada conquista, has sido la calma en medio de la tormenta. Desde el fondo de mi corazón, mil gracias. Al final, lo hemos conseguido.

# ABSTRACT

Alternative splicing (AS) is a post-transcriptional mechanism for gene expression regulation affecting almost every multi-exonic gene in mammals. However, the impact of AS at the proteomic level remains under debate. In this thesis, we explored the landscape of AS across a variety of tissues, cell types and developmental stages in human, mouse and chicken, through the analysis of a panel of more than 300 publicly available RNA-seq samples. We confirm the presence of highly specific AS programmes in neural, muscle, testis and pluripotent cells in these species, and identify additional conserved modules of co-regulated AS events in kidney, liver, adipose tissue and immune cells by the application of a network analysis approach. In addition, we describe a subset of exons undergoing AS in virtually all analysed tissues and cell types. These exons are enriched in genes of pathways related to regulation of gene expression, where they likely operate through their translation into alternative protein isoforms coexisting at the single cell level.

# RESUMEN

El corte y empalme alternativo del ARN (CEAA) es un mecanismo post-transcripcional de regulación de la expresión génica que afecta a un alto porcentaje de genes en mamíferos. Sin embargo, la magnitud del efecto del CEAA a nivel proteómico permanece todavía en debate. En esta tesis se explora el efecto del CEAA en un panel de más de 300 muestras públicas de experimentos de secuenciación masiva de ARN, correspondientes a diversos tejidos, tipos celulares y etapas del desarrollo de humano, ratón y pollo. Nuestros resultados confirman la presencia de importantes programas de CEAA en tejidos neurales, musculares, en testículo y en células pluripotentes. Además, hemos identificado módulos adicionales de exones co-regulados en riñón, hígado, tejido adiposo y células del sistema inmune, mediante la aplicación de métodos de análisis de grafos. Por otra parte, describimos un conjunto de exones regulados por CEAA en la práctica totalidad de las muestras analizadas. Estos exones se localizan preferentemente en genes relacionados con procesos de regulación de la expresión génica, donde podrían actuar mediante su traducción a isoformas proteicas coexistentes a nivel celular.

# PREFACE

The application of next-generation sequencing technologies to the study of gene expression has sparked the interest in alternative splicing (AS) and its role in the establishment of tissue and cell identity.

Numerous studies highlight that, far from being an isolated biological peculiarity, AS affects virtually every multi-exonic gene in mammals, and alternative RNA isoforms are known to be often differentially expressed at the tissue and developmental level. The global potential of AS to confer tissue-specific plasticity to vertebrate proteomes without the need for a drastic increase in gene number or genome size is enormous. However, experimental detection rates of alternative protein isoforms do not match the estimates obtained from RNA-based techniques, and therefore the actual impact of AS in the phenotypic variability between cell types is under question. The present thesis is a humble attempt to contribute to this debate.

The first two chapters act as an introduction to the field. Chapter 1 is a general introduction to the diverse mechanisms regulating gene expression in eukaryotes, and Chapter 2 provides an overview of the current knowledge about AS, its regulation, biological functions and its role in three of the contexts where it has been more thoroughly studied: neurons, muscular tissues and pluripotent cells.

The central part of the thesis is devoted to the results obtained during this work, summarised by means of the three publications where they were released. The two first chapters in this section correspond to two collaborations providing insights into specific AS programmes in two different tissues. Chapter 3 is focused on the splicing landscape of neurons and neuronal differentiation, with a particular emphasis on microexons (a set of short exons with remarkable biological features and highly specific inclusion profiles in neurons). Chapter 4 is dedicated to the relationship between AS and pluripotency in planarians. The regenerative abilities of planarians and the high evolutionary distance between this organism and the vertebrate lineage allow for an assessment of the conservation of global mechanisms governing AS programmes in the process of self renewal and pluripotency. The first-author publication in Chapter 5 constitutes a more comprehensive survey of AS in vertebrates, by a systematic analysis across more than 300 RNA-seq samples from tissues, cell types and developmental stages in human, mouse and chicken. Finally, in Chapter 6 we discuss our view about the contribution of AS to tissue identity and proteomic complexity in the light of our results.

# OBJECTIVES

- To comprehensively characterize novel tissue-regulated AS programmes in vertebrate species, with a focus on human, mouse and chicken.

- To study regulatory patterns and mechanisms of those AS programmes.

- To characterize the effect of alternative exons at the protein level.

- To establish a centralized repository of AS exons, including their inclusion patterns and their effect on their protein counterparts, where they can be accessed by the research community in order to develop new testable hypotheses about AS function.

# TABLE OF CONTENTS

# 1. REGULATION OF GENE EXPRESSION IN EUKARYOTES

## 1.1 Cell theory and biochemical bases of cellular composition

Of all the features of life on Earth that have gathered the interest of mankind, its diversity is probably among the most important ones. Throughout millions of years of natural history, a variety of organisms have roamed our planet. The number of living species is believed to be in the range of 1,000 million (Larsen et al., 2017), and new species are described year after year.

Despite this diversity, it is among the main aims of biology to discover basic principles applicable to all living organisms. The application of the light microscope to the observation of biological tissue samples paved the way to the discovery of one of such landmarks of biology: the development of the cell theory by the botanist Matthias Schleiden and the zoologist Theodor Schwann. In its classical form, the cell theory is typically summarized in three short statements: ‘All organisms are composed by cells’, ‘The cell is the basic unit of life’, and ‘All cells arise from pre-existing cells’ (Schwann and Schleiden, 1839). Altogether, this theory provided a unified view of the composition of animal and vegetal living matter, which, until then, had been considered intrinsically different in their internal organization. While many of the details that Schleiden and Schwann predicted about cells have been proved wrong in the 150 years since their theory was published, these three main conclusions remain at the foundations of cell biology as a field.

The fact that cells compose all living matter does not mean that all cells of the same species, or even organism, are equal to each other. Although the life cycle of unicellular organisms can also go through several stages, it is in multicellular organisms where this is most clearly observed. Cells have very specialised morphologies, adapted to the functions that they perform in their respective tissues. However, under their apparent differences, all cells are made of the same basic molecule types. On average, up to 75% of a cell is just a solution of water, ions, sugars, lipids and other small molecules. The second most abundant fraction —about 20%— corresponds to proteins. Nucleic acids (DNA and RNA), and polysaccharides make up the rest of the composition of a cell.

Proteins are the most functionally diverse kind of biomacromolecule and, as such, they are the most numerous and varied effectors in a cell. In the first place, proteins serve as structural scaffolds for many intracellular structures. In addition, most biochemical reactions that living organisms need to stay alive can happen at physiological conditions only because they are catalysed by proteins, which maintain the substrates in conformations that lower the activation energy threshold needed for those reactions to happen. Finally, proteins can serve as signalling molecules used by the cell to sense its environment and to transmit this information to other cells.

The reason why a protein can perform a function is because it has the precise three-dimensional shape needed for it. Intermolecular interactions in a living organism only happen because the molecules taking part in them show structural complementarity, in such a way that the interaction is energetically favourable. A protein is a linear polymer of amino acids, joined together by peptide bonds. But three main factors act together to force this chain into a particular three-dimensional structure. In order of importance, these factors are: the *hydrophobic effect* (the tendency to minimise the global entropy of the system, by minimising the size of the solvation sphere of the protein), physicochemical interactions between protein residues (such as disulphide bonds between cysteine residues, salt bridges, hydrogen bonds and Van der

Waals interactions), and interactions between the protein and its —usually aqueous— environment.

Given a relatively constant set of conditions, the effect of these factors depends primarily on the amino acid sequence of the protein itself. As there are 21 different amino acids forming part of proteins sequences (including selenocysteine), and only four different nucleotides in either DNA or RNA polymers, the potential for structural diversity in proteins is much higher than that of nucleic acids. The transition from an RNA-based world (where biochemical reactions were catalysed mainly by RNA molecules) to a protein-based metabolism is considered a major event in the history of evolution, as it allowed for a vast increase in metabolic complexity of cells.

Although the cytoplasm is mainly composed of water, there is a relatively high concentration of proteins and other solutes —it is estimated that a prototypic mammalian cell can have as many as 2.7 million protein molecules per cubic micrometre (Milo, 2013)—, as well as high numbers of organelles and vesicles. All of these molecules and structures are surrounded by their own solvation spheres, whose existence constrains the mobility of around 25% of water molecules in the cell (reviewed in Ball, 2017). It is estimated that the average distance between macromolecules in the cytoplasm is about 1 nm, which would correspond to three to four layers of water molecules (Ball, 2017). In such a crowded environment, with intermolecular contacts happening at a high frequency, the structural complementarity between proteins and their interacting partners is the main mechanism to ensure selectivity in biological reactions.

Biochemical processes in the cell are, however, tightly regulated. Comprehensive elucidation of how this regulation is achieved is, still today, one of the most intensely pursued questions in biological research. One of the foundational steps in this pursuit were the experiments of Beadle and Tatum in 1941, which proved that the repertoire of enzymatic activities of a cell and its descendants could be altered by affecting their DNA (Beadle and Tatum, 1941). Only three years later, DNA was discovered as the molecule storing inheritable information about an organism (Avery et al., 1944). Together, these two experiments brought together the worlds of genetics and biochemistry, which used to be considered as separate disciplines.

A turning point in the history of biology was the postulation of the structural model of double-stranded DNA by Watson and Crick (Watson and Crick, 1953), heavily influenced by the experimental work made by Rosalind Franklin and Maurice Wilkins (Franklin and Gosling, 1953b, 1953a; Wilkins et al., 1953). This model led Crick to the idea of the DNA sequence acting as a code for protein sequence (Crick, 1958), and later to the so-called *central dogma* of molecular biology (Figure 1), where RNA was postulated as an intermediary in the information flow from DNA to protein (Crick, 1970). The discovery of the genetic code —the universal correspondence between triplets of RNA nucleotides and protein amino acid residues— finally provided a mechanistic explanation of how protein sequences can be encoded in a DNA-based genome (Lengyel et al., 1961; Nirenberg and Matthaei, 1961; Nirenberg et al., 1965; Nishimura et al., 1964).

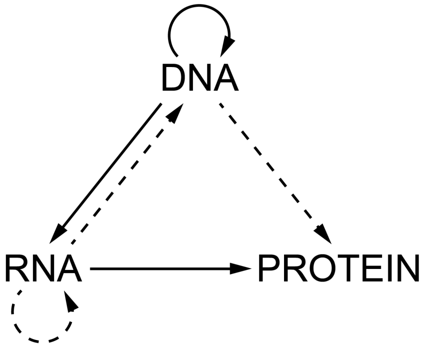


**Figure 1: Information transfers in biological organisms according to the 'central dogma' on its 1970 version.** Solid arrows show general transfers; dotted arrows show special transfers. The absent arrows are undetected transfers (Crick, 1970).

Although the genome is the container of all the information that a cell needs to survive and perform its functions, the genomic sequence of a cell does not suffer big changes throughout its lifespan. In contrast, biochemical processes taking place inside cells change significantly during the same period. By modulating the relative rates of the different metabolic pathways taking place inside a cell, it can adapt to changes in its environment, respond to external stimuli, and reproduce successfully. In many multicellular organisms, an the existing diversity between cell types —each of them with specialised functions— is reminiscent of the diversity of species on Earth. In those organisms, cell growth and division rates must be regulated in a way that ensures renewal of damaged or old tissues, while avoiding uncontrolled proliferation that would lead to development of a tumour. Furthermore, some processes that are essential for the survival of an organism, such as development, can only happen through coordinated replication, migration, cross-talk and death of many cells from different tissues.

The reason behind this apparent paradox —the existence of many different cell types, and the ability of each cell type to adapt its behaviour to its environment, despite having an essentially identical genome in all cells and situations throughout the lifespan of an organism— is that not every gene is equally expressed in every cell type and condition. In eukaryotes, regulation of gene expression is a complex process, with multiple layers of regulation operating simultaneously on each gene. Typically, these regulation mechanisms are classified as transcriptional or post-transcriptional, depending on the stage of the expression process in which they take place.

## 1.2 Regulation of gene expression by transcriptional mechanisms

Although the first studies of transcriptional regulation were performed in prokaryotes (Jacob and Monod, 1961), several elements are shared by eukaryotes. In both domains, transcriptional regulation follows the *cis-trans* model: genes are activated or repressed through certain regions in their same chromosome (*cis*-acting elements), which can be bound by other molecules (*trans*-acting factors, usually proteins). The key event for transcription initiation is the recruitment of an RNA polymerase to the gene promoter. This is triggered by the binding of several *trans*-acting factors (called *basal transcription factors*), to a set of *cis-acting* elements located in the 5' start of the gene, in a region named *promoter* (reviewed in Sainsbury et al., 2015).

Nevertheless, eukaryotic transcriptional regulation presents several particularities. As a general rule, eukaryotic genes are repressed by default, and their transcription can only be initiated by means of several additional *trans-activators* (Struhl, 1999). These transcription factors specifically recognise and bind additional regulatory sequences (*enhancers*), usually located in genomic regions distant to the core promoter (reviewed in Cho, 2012). Transcription factors bound to enhancers are recognised by an additional type of regulators: *Mediator* proteins (Conaway and Conaway, 2011). These proteins lack DNA binding activity, but they have affinity for different *trans*-activators and for the RNA polymerase II. Together with basal transcription factors, the Mediator complex is also essential for recruitment of RNA polymerase II to promoters in order to induce transcriptional initiation. Although different enhancers acting on the same gene can be distant from each other and from the gene promoter in the DNA sequence, loops in the DNA structure stabilised by cohesin complexes can bring these elements together in the three-dimensional space, where their respective *trans*-activator complexes can interact with Mediator proteins (Kagey et al., 2010; Li et al., 2012). Ultimately, these large multi-protein complexes trigger transcription initiation in most eukaryotic genes, inducing phosphorylation of specific serine residues in the carboxy terminal domain (CTD) of RNA polymerase II by TFIIH (Plaschka et al., 2015). Genes transcribed by RNA polymerases I and III, such as ribosomal genes, transference RNA genes, and other small functional RNAs, follow different variations of this mechanism (reviewed in Vannini and Cramer, 2012).

The specific requirement for activation of eukaryotic genes avoids the need to actively repress every gene that does not need to be transcribed in a particular situation, which would require the production of the corresponding repressor proteins. Since eukaryotes tend to have larger genomes than prokaryotes, both in size and in number of genes, this mechanism is more efficient for the needs of a eukaryotic organism. Several studies in human and mouse suggest that the percentage of these genomes covered by enhancers is significantly higher than their transcribed fraction, highlighting the level of regulatory complexity in these organisms (Cho, 2012; Dunham et al., 2012; Shen et al., 2012). Usually, transcriptional activation of each gene is mediated by several enhancers, presumably as a mechanism of combinatorial control to minimise off-target activation. Although pervasive transcription still takes place in eukaryotes, including human (Djebali et al., 2012), many mechanisms, such as DNA methylation, operate to minimise its effects (Neri et al., 2017, and see below).

Another source of differences between prokaryotes and eukaryotes comes from the high-order structural organisation of eukaryotic genomes. In eukaryotes, most of the genome is assembled into nucleosomes. Moreover, large portions of the genome are condensed in heterochromatic regions. These two factors greatly influence the accessibility of the transcriptional machinery to the corresponding *cis*-acting sequences, and therefore, they have a significant impact in transcription rates (reviewed in Lee and Young, 2013). While local differences in DNA sequence affect nucleosome occupancy (because of the differences in bendability introduced by changes in sequence composition), it is well established that the position of some nucleosomes is heavily regulated by additional factors (Gross and Garrard, 1988). Therefore, nucleosome occupancy is difficult to predict from sequence alone in a genome-wide fashion (van der Heijden et al., 2012).

In a genome, some regions consistently show depletion in nucleosome occupancy. Classically, this has been quantified by nuclease digestion assays, and therefore, nucleosome-depleted regions are also called *nuclease hypersensitive sites* (NHS). These sites are usually flanked by nucleosomes with very strict positioning, and they tend to overlap with *cis*-acting elements of genes (Crawford et al., 2004; Wu, 1980). Transcriptional activation of a gene is associated with changes in its nucleosome position pattern, with nucleosomes being released from NHS, or mobilized to adjacent regions. Reduced nucleosome occupancy of NHS can be regulated by chromatin remodelling agents —such as the SWI/SNF complex, that destabilizes histone-DNA interactions and mobilizes nucleosomes—, post-translational modification of histones (like acetylation of H3K27 (H3K27ac), trimethylation of H3K4 (H3K4me3), or ubiquitylation—, usage of histone variants —such as H3.3 and H2A.Z—, or directly by *trans*-acting factor binding, which hinders packing of new nucleosomes around the NHS (reviewed in Bell et al., 2011).

Epigenetic regulation of gene transcription is deeply intertwined with regulation by transcription factors. Even genes in heterochromatic regions usually have at least one NHS accessible. Accessible sites can be bound by a subset of transcription factors, called *pioneer factors* (Lai et al., 2018; Zaret and Mango, 2016; Zaret et al., 2016), which induce recruitment of chromatin remodellers to the region. The subsequent increase in chromatin accessibility enables access of additional *trans*-activators, and, ultimately, the beginning of transcription. Histone variants and post-translational modifications of histones also contribute to the process, as they serve as recognition sites for binding of certain transcription factors (Ahsendorf et al., 2017; Xin and Rohs, 2018).

Some genomic *cis*-acting regions act as transcriptional repressors, although they tend to be less abundant than enhancers. Very frequently, repression is mediated by proteins with co-repressor activity, that sequester other positive transcription regulators and impede their binding to their corresponding *cis*-acting regions (reviewed in Payankaulam et al., 2010). Similarly, epigenetic modifications can also contribute to maintain genes in a repressed state. Full transcriptional silencing is associated to post-translational modifications of histones such as H3K9 trimethylation (H3K9me3) and H4K20 trimethylation (H4K20me3) (Feng et al., 2010; Lejeune and Allshire, 2011; Schotta et al., 2004). Several chromatin remodellers, such as the Polycomb group proteins, introduce other histone modifications like trimethylation of H3K27 (H3K27me3) at genes that must be repressed at certain developmental stages (Cao et al., 2002; Müller et al., 2002). In general, all these mechanisms trigger chromatin condensation as a way to ensure long-term efficient repression.

DNA methylation is another epigenetic mechanism related to transcriptional repression of gene expression. There is a great deal of heterogeneity among methylation patterns observed in eukaryotic genomes, both in terms of methylation frequency and in the variability of sequences that get methylated. Within metazoans, methylation usually takes place in position 5 of cytidines, if they are followed by guanosine (that is, in CpG dinucleotides), although there are reports of CpT methylation in the *D. melanogaster* embryo (Lyko et al., 2000), and low levels of CpA and CpT methylation in mammalian embryonic stem cells (ESCs) —but not in differentiated tissues— have also been detected (Laurent et al., 2010; Lister et al., 2009; Ramsahoye et al., 2000).

The clearest distinction in metazoan methylation occurs between vertebrates and invertebrates (Hendrich and Tweedie, 2003; Tweedie et al., 1997). Some invertebrate genomes seem to lack methylation completely —including several model organisms like the nematode *C. elegans* and the yeast species *S. cerevisiae* and *S. pombe* (Antequera et al., 1984; Proffitt et al., 1984; Tweedie et al., 1997)—, while others show intermediate degrees of methylation. In contrast, the vast majority of CpG dinucleotides in vertebrate genomes are methylated, including introns, exons, intergenic regions, repetitive sequences and many regulatory elements. The exceptions are some unmethylated CpG-rich regions of about 1 kb in length, often mapping to promoters or distal regulatory regions, named *CpG islands* (Bird, 1986). Methylation of CpG islands —especially at vertebrate promoters— has been extensively linked to transcriptional repression, since methylated CpG islands can be recognised by proteins of the methyl-CpG binding containing (MBD) family, which in turn recruit epigenetic modifiers responsible for chromatin condensation (reviewed in Bogdanović and Veenstra, 2009). According to this model, genome-wide methylation of enhancers and promoters acts as a safe-lock mechanism to avoid pervasive transcription in vertebrates (Hargan-Calvopina et al., 2016).

In the last years, nuclease-based methods such as DNase and MNase sensitivity assays are often replaced by other approaches such as chromatin immunoprecipitation coupled with sequencing (ChIP-seq) or assays for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013), based on chromatin immunoprecipitation and transposase accessibility, respectively. Recently, ATAC-seq has been applied in combination with the use of microfluidic devices to analyse the chromatin accessibility landscape of human lymphoblastoid cells at single-cell resolution, revealing two subsets of *trans*-acting factors associated to increased and decreased cell-to-cell variability in chromatin accessibility (Buenrostro et al., 2015).

## 1.3 Regulation of gene expression through post-transcriptional mechanisms

Eukaryotic pre-mRNAs undergo a complex sequence of chemical modifications between their transcription and their translation. These *RNA maturation* reactions take place mainly in the cell nucleus, and their effect, together with other processes —such as RNA export, RNA stability and subcellular localisation, translational regulation and post-translational protein modifications— form an additional regulatory layer that, in combination with transcriptional control mechanisms, shapes cellular behaviour in every set of conditions. A summary of the levels of regulation operating in gene expression is shown in Figure 2.

For simplicity, we will only describe a few of these mechanisms in this thesis: pre-mRNA capping and polyadenylation, RNA editing and, especially, pre-mRNA splicing, whose patterns and regulation are the core of this work (see section 1.4). Although we generally refer to these as post-transcriptional regulatory mechanisms of gene expression, in the case of splicing it is worth noting that most introns are spliced co-transcriptionally (Tilgner et al., 2012), even though it is true that some introns, especially alternative introns, may be retained in the nucleus for a long time after transcription of their pre-mRNA has finished. In fact, splice site choices are far from independent from transcriptional parameters such as RNA polymerase elongation speed, as lower elongation rates tend to favour usage of weaker splice sites (SS) instead of downstream stronger SS that are transcribed later because of their position in the gene (reviewed in Bentley, 2014). A withstanding open question is whether transcriptional elongation is regulated to affect AS.

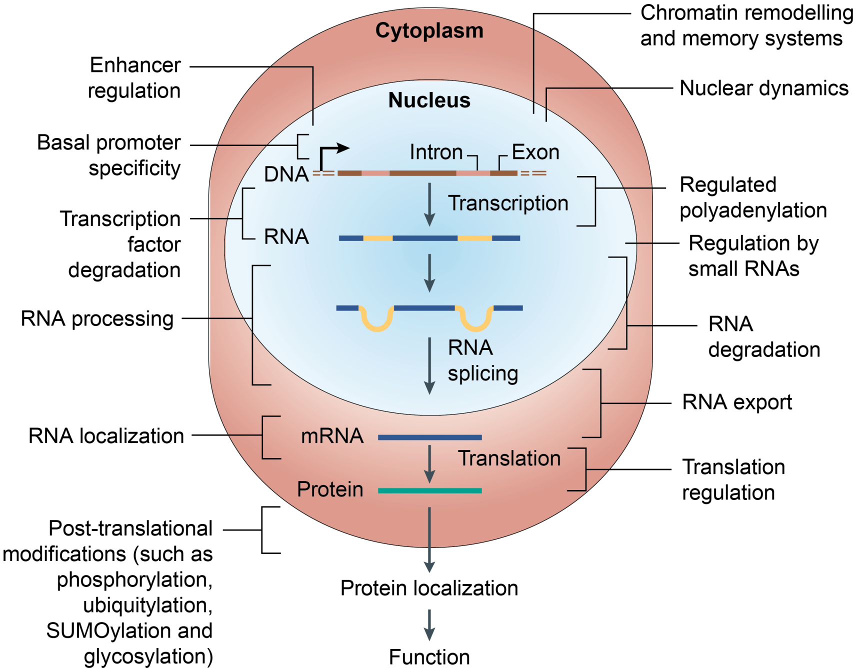


**Figure 2: Overview of the levels of regulation in gene expression.** Adapted from (Alonso and Wilkins, 2005).

### 1.3.1 Pre-mRNA capping and polyadenylation:

RNA polymerase II transcripts are ‘capped’ with a 7-methylguanosine group bound at the 5’ end by a 5’,5’-triphosphate bond. The process of capping happens co-transcriptionally, after the first 20-30 nucleotides of the pre-mRNA has been transcribed. This cap protects the transcript from the action of exonucleases, serves as a recognition site for protein complexes responsible for further RNA processing, and improves translation efficiency (Shatkin, 1976; reviewed in Furuichi, 2015 and Shuman, 2015).

Most of the transcripts produced by RNA polymerase II are also cleaved at the 3’ end after transcription termination, and then elongated with a poly(A) tail. The length of this tail is highly variable between species and genes, with an average length of 250-300 adenosines in human (Elkon et al., 2013). The poly(A) tail is essential in the maturation process, as it has an important function in regulation of gene expression, and in signalling the mRNA for nuclear export. Poly(A) tails with reduced length cause transcript degradation by mechanisms such as nonsense-mediated decay (NMD), or their storage in a dormant state, where they are not translated (D’Ambrogio et al., 2013; Guhaniyogi and Brewer, 2001; reviewed in Elkon et al., 2013).

3’ modifications of RNA depend on the recognition of several *cis*-acting elements in the 3’ region of the pre-mRNA by certain protein complexes, such as CPSF, CSTF, CFIm/IIm and a poly(A) RNA polymerase (PAP) (Elkon et al., 2013), as well as several single proteins such as the Symplekin scaffolding protein. These factors perform their function by binding to several *cis*-regulatory elements in the pre-mRNA: a poly-A signal (PAS) with the consensus sequence AAUAAA located 15-30 nucleotides upstream of the cleavage site, and several upstream and downstream sequence elements (USE and DSE, respectively), usually GU-rich. While most of these factors are conserved, the intervening *cis*-elements are highly variable in eukaryotes. An overview of the machinery involved in pre-mRNA cleavage and polyadenylation is shown on Figure 3.

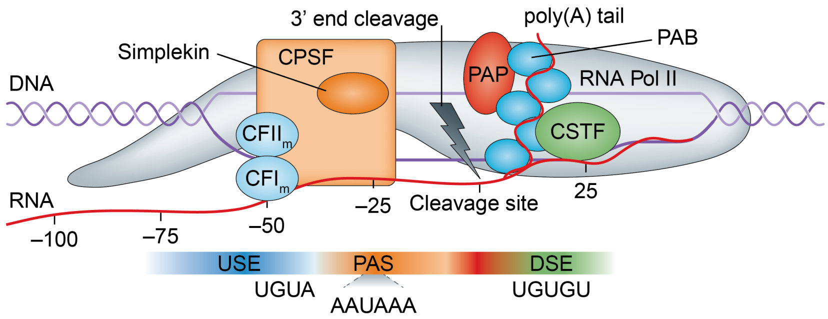


**Figure 3: Main factors involved in transcript cleavage and polyadenylation.** PAS: polyadenylation signal. CPSF: cleavage and polyadenylation specificity factor. CSTF: cleavage stimulating factor. PAP: poly(A) polymerase. CFIm: cleavage factor Im. CFIIm: cleavage factor IIm. Adapted from (Elkon et al., 2013).

## 1.3.2 RNA editing:

In some cases, RNA sequences are altered after they are transcribed. This process is known as *RNA editing*, and it has important effects in the expression of several eukaryotic and prokaryotic genes. While more than a hundred of these modifications have been described in eukaryotes (summarised in Table 1), the most common editing process is the hydrolytic deamination of adenosines in their sixth carbon of the purine ring, to yield inosine.

While adenosine can form Watson-Crick base pairs with thymidine or uridine, inosine pairs with cytidine instead (Figure 4). This causes inosine to be effectively recognized as guanosine by the cellular machinery responsible for RNA splicing and translation, which, in turn, can generate non-synonymous substitutions at the protein sequence level (known as *recoding*), alter the splicing pattern of the transcript, or affect miRNA binding sites.

In mammals, this process is catalyzed by the adenosine deaminases ADAR1 and ADAR2. Both are ubiquitously expressed, although ADAR1 has higher expression. There are two isoforms of ADAR: ADAR1p110 (constitutive, and localized in the nucleus), and ADARp150 (cytoplasmic and induced by interferon). While ADAR1 edits most of the sites in mammalian genomes, the function is to avoid triggering of an immune response in the cytosol by the interaction of MDA5 with double-stranded RNA (dsRNA), that can be sensed as viral RNA. ADAR2 is responsible for most of the recoding events, and a third member of the family, ADAR3, is only expressed in the brain and lacks catalytic activity.

| **Editing type** | **Mediated by** | **Number of sites** | **Targeted transcripts** | **Organisms** |
|---|---|---|---|---|
| Insertion and deletion of U | RECC | 1,000s | Mitochondrial mRNAs; recoding | Kinetoplastids |
| C-to-U | APOBECs | 100s | Mostly non-coding; in a few tissues | Mammals |
| C-to-U | PPR proteins | 100s | Chloroplast and mitochondrial mRNA; mostly recoding | Plants |
| A-to-I | ADAR proteins | 1,000,000s | Mostly non-coding | Metazoa |

**Table 1: Description of the four main known types of mRNA editing.** Adapted from (Eisenberg and Levanon, 2018).

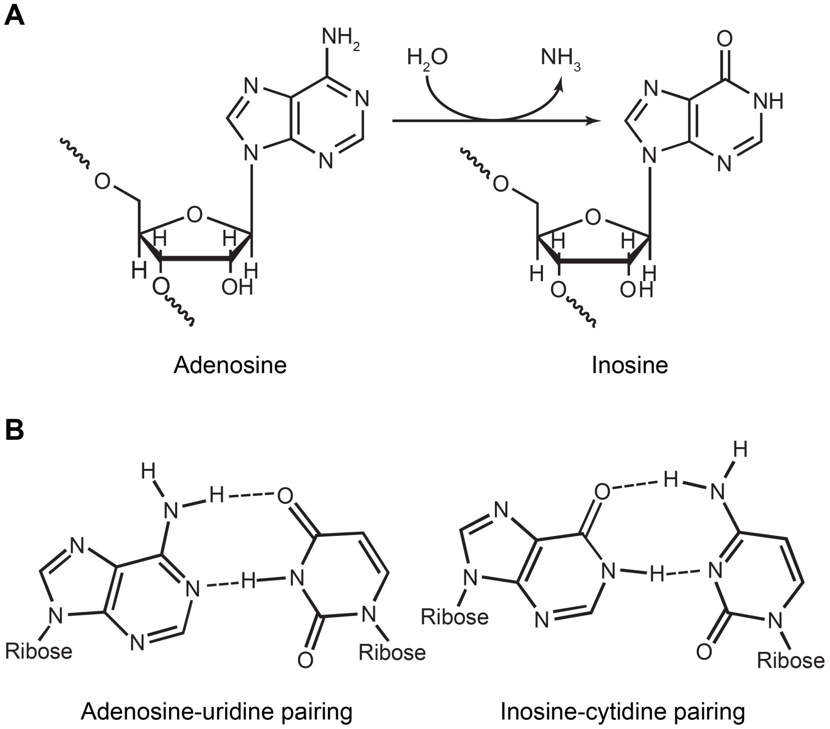


**Figure 4: A) Deamination reaction of adenosine, catalysed by ADAR proteins. B) Effect on deamination on Watson.Crick base pairing.** Adapted from (Zinshteyn and Nishikura, 2009).

The human genome is estimated to contain more than 100 million editing sites, although most of these sites are edited by ADAR1, and they are edited at low levels, below 1% (Bazak et al., 2014). The majority of human editing events map to intronic positions or untranslated regions (UTRs) of their pre-mRNAs, especially in the 3' end (Chen, 2013).

On the other hand, only around 1,000 recoding sites have been mapped in human. Regarding tissues, the human brain transcriptome is particularly enriched in editing sites (Nishikura, 2010), and recoding sites are enriched in genes with neural functions, both in human and in invertebrate species (reviewed in Rosenthal & Seeburg, 2012). An example is the glutamate receptor GRIA2. The editing site in this gene is conserved in mouse, and it is the only tested ADAR2 editing site that is essential for survival in mice embryos (Higuchi et al., 2000).

Most editing sites discovered in human are either species-specific, or conserved only in primates (Eisenberg and Levanon, 2018). The dsRNA structural motif recognized by ADAR proteins can be generated by neighbouring inverted pairs of repetitive elements. This explains the high proportion of primate-specific editing sites, as more than 99% of editing sites in the human genome overlap with repetitive elements of the Alu family, which are found only in primates (Bazak et al., 2014). Interestingly, recruitment of ADAR1 to one of these sites can enhance editing in additional sites of the same pre-mRNA, at distances of several hundreds of nucleotides, which would not be edited otherwise (Daniel et al., 2014).

The mammalian editing landscape is substantially different from other clades. Besides ADAR1, cephalopods have two isoforms of ADAR2. While ADAR2b is similar to the mammalian ADAR2, ADAR2b has an additional RNA binding domain, more editing activity in vitro, and increased resistance to saline environments such as in squid neurons, which have higher salt concentrations than those of other species (Palavicini et al., 2012). More than 10,000 conserved recoding events have been characterized in this lineage, suggesting that this regulatory mechanism plays a key role in proteomic diversification of cephalopods (Alon et al., 2015; Liscovitch-Brauer et al., 2017)

Misregulation of RNA editing has been linked to several diseases. Mutations in ADAR1 cause Aicardi-Goutières syndrome, an autoimmune disease affecting the brain and the skin, characterized by an increased production of interferon α (Rice et al., 2012). Significant differences in editing levels have also been found in amyotrophic lateral sclerosis (ALS) (reviewed in Kwak & Kawahara, 2005) and some neuropsychiatric disorders (Schmauss, 2005). Altered editing is also associated to tumour malignancy (Paz and Levanon, 2007), and there is a negative correlation between global editing levels and patient survival (Paz-Yaacov et al., 2015).

## 1.4 Pre-mRNA splicing

### 1.4.1 'Why genes in pieces?' and the discovery of pre-mRNA splicing

In 1977, the detection of single-stranded DNA (ssDNA) loops in electron micrographs of dsDNA restriction fragments from adenovirus incubated with their corresponding mRNA surprised two teams of molecular biologists at MIT and Cold Spring Harbor Laboratory (Berget et al., 1977; Chow et al., 1977). The micrographs indicated that adjacent segments in the viral mRNA formed stable hybrids with regions that were distant to each other in the viral genomic sequence (Figure 5). They explained the result by postulating the existence of a previously unknown step in viral RNA processing: the formation of mRNA by "intramolecular joining" of segments separated by intervening sequences in the original pre-mRNA (Berget et al., 1977).

Soon afterwards, R-loops were also observed in eukaryotic pre-mRNAs (Konkel et al., 1978; Leder et al., 1978; Tilghman et al., 1978). The advent of the first DNA sequencing methods (Maxam and Gilbert, 1977; Sanger et al., 1977) allowed the elucidation of the first complete sequences of eukaryotic genes, such as the mouse β-globin gene (Konkel et al., 1978). These studies established the consensus sequences of *cis*-elements operating in several mRNA maturation processes, including the presence of conserved CU and AG dinucleotides at the boundaries of the intervening sequences of genes, even in different species (Table 2, Figure 5).

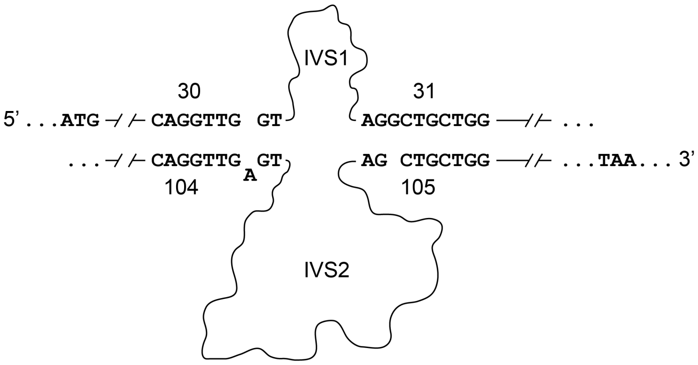


**Figure 5: Sequence comparison between the splice junctions of the first two introns described in the mouse $\beta^{maj}$ globin gene first described in mouse.** Numbers in the figure represent amino acid codon positions. IVS: intervening sequence (intron). Adapted from (Konkel et al., 1978).

A seminal 'News and Views' piece written by Walter Gilbert in 1978 analysed the impact of this new paradigm of gene expression on the existing models of genomic evolution, with three main ideas (Gilbert, 1978). First, the existence of intervening sequences (introns) implied that point mutations could have a greater effect at the protein level than previously expected, by creating or disrupting splice sites and altering the expressed sequences (exons) of a gene. Second, Gilbert also predicted the possibility of alternative splicing patterns in a single gene, and with it, the end of the 'one gene – one polypeptide' axiom. Finally, he highlighted introns as hotspots of future evolution (due to their non-coding nature), as well as their role in facilitating evolution of new protein products after recombination, by exon shuffling.

| Gene | 5' boundary | Intron | 3' boundary |
|---|---|---|---|
| Mouse β$^{maj}$ | GCAGGTTG | IVS1 | TTAGGCTGCTG |
| Mouse β$^{maj}$ | TCAGGGTG | IVS2 | ACAG-CTCCTG |
| Mouse β$^{min}$ | TCAGGGTG | IVS2 | ? |
| Mouse α | TCAGGTAT | IVS2 | ? |
| Mouse λ$_1$ | TCAGGTCA | IVS1 | GCAGGGGCCA |
| Mouse λ$_1$ | CTAGGTGA | IVS2 | CCTGCGGCCA |
| Mouse λ$_2$ | TCAGGTCA | IVS1 | GCAGGAGCCA |
| **Prevalent sequence** | **TCAGGT** | | **CAGG** |

**Table 2: One of the first comparisons of splice junction sequences across different globin and immunoglobulin mouse genes. IVS: intervening sequence (intron). Adapted from (Konkel et al., 1978).**

## 1.4.2 Types of introns and their evolution

The current picture of RNA splicing distinguishes four classes of introns (reviewed in (Irimia and Roy, 2014)): spliceosomal introns, group II introns, group I introns and tRNA introns.

Spliceosomal introns take their name from the need of a complex assembly of snRNAs and auxiliary proteins (collectively called spliceosome) for their excision. They are divided in two groups. The most abundant subgroup of spliceosomal introns in eukaryotes are U2 introns, as they are present in all nuclear eukaryotic genomes studied to date. As shown in Figure 6, in most species U2 introns are characterized by their short 5' SS and minimal 3' SS, with the sequences GT and AG almost always present at the intron borders (exceptionally, GC-AG, see Thanaraj, 2001). Additionally, they include a fixed adenine nucleotide (called *branch point* (BP)) with catalytic function near the 3' SS, and a polypyrimidine tract (PPT) located between the branch point and the 3' SS. These introns are excised by the so-called *major* spliceosome and released as a lariat structure (see section 1.4.2). The second subgroup of spliceosomal introns are U12 introns, present in many distant eukaryotic clades (including vertebrates). Their consensus sequence on the 5'

SS is different from that of U2 introns (GT or AT) (Figure 6), and the distance between the BP and the 3' SS usually ranges between 10 and 15 nucleotides. These introns lack a PPT, and are excised by a spliceosome with a different set of snRNAs and auxiliary proteins than U2 introns, also called *minor* spliceosome, although most components are common between both spliceosome types.

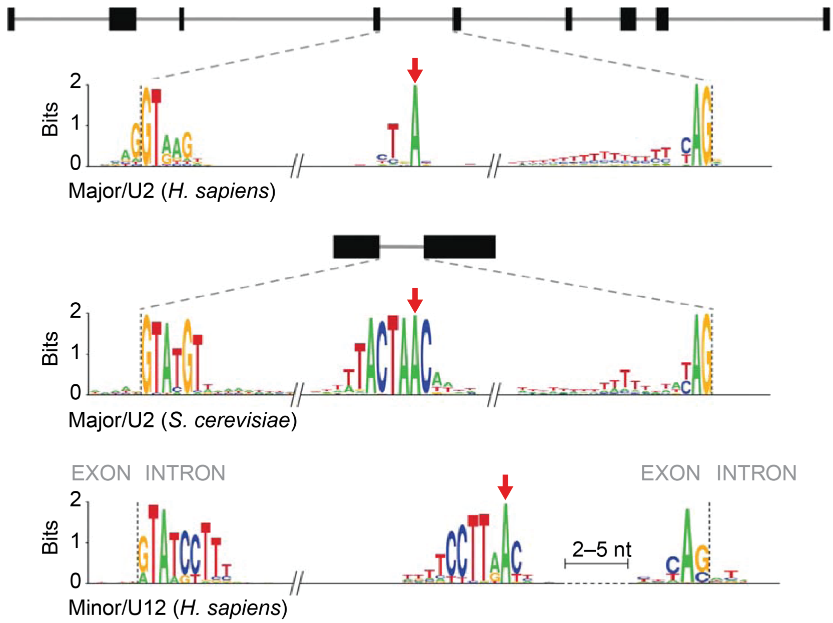


**Figure 6: Consensus sequences for spliceosomal intron core splicing signals for human U2 introns, yeast U2 introns and human U12 introns.** The branc point adenosine is indicated by a red arrow. From (Irimia and Roy, 2014).

distribution. They are present in around 25% of the sequenced bacterial genomes, as well as some archaea and the mitochondrial and chloroplast genomes of plants, fungi and protists (Lambowitz and Zimmerly, 2004; Robart and Zimmerly, 2005). Unlike spliceosomal introns, group II introns are able to catalyse their own splicing, although *in vivo* they make use of auxiliary proteins. Often, these proteins are encoded by the intron itself, and then they receive the generic denomination of intron-encoded proteins (IEPs). The fact that IEPs sometimes have reverse transcriptase activity able to produce a DNA form of the intron —either in the same polypeptide or in a separate protein from the one involved in splicing—, makes these introns more similar to transposable elements. Because of their splicing mechanism, the sequence of these introns is evolutionary constrained throughout the intron length, instead of showing strong conservation in isolated locations of the intron.

Group I introns are present in some bacteria and viruses, as well as nuclear rRNAs or organellar genomes of distant eukaryotic clades. These introns are self-spliced, although they use a free guanosine nucleotide as a phosphate donor instead of an endogenous branch point. As a result, these introns are not excised in a lariat configuration.

A fourth group of introns, usually called tRNA introns, are found in tRNAs of eukaryotes and archaea, and also in several archaeal coding genes. As opposed to the three types described above, these introns are spliced by protein enzymes, and not by RNA.

The evolutionary history of introns, especially spliceosomal introns, was the subject of an intense debate in the research community for approximately 30 years after their discovery (reviewed in (Irimia and Roy, 2014; Koonin, 2006)), given that introns are almost absent from bacterial genomes, and widespread in many eukaryotic genomes. For a long time, experts were divided between two main hypotheses. The *introns-late* hypothesis postulated that introns are an evolutionary novelty in eukaryotes —which could have gained them progressively after their divergence from prokaryotes, perhaps by accumulation of transposable elements—. This theory was supported by the absence of introns in prokaryotes, and the detection of transposable elements inside introns (Doolittle and Stoltzfus, 1993). On the other hand, the *introns-early* hypothesis defended that introns were already present in the common ancestor of prokaryotes and eukaryotes --where they could play a decisive role in the assembly of functional mature transcripts, as replication and transcription machineries were likely error-prone in this organism-- and some clades, including bacteria, underwent a process of intron loss in order to streamline their gene expression and accelerate their genome replication. Intron loss would confer these organisms a selective advantage by decreasing their duplication rate, at the cost of the genome plasticity conferred by introns (Darnell, 1978; Doolittle, 1978).

The current consensus combines elements from both theories (Figure 7). The discovery of group II introns in bacteria and eukaryotic organellar genomes, and their similarity to spliceosomal introns, led to the hypothesis that type II introns were carried by the genome of the endosymbiotic prokaryote (probably an α-proteobacteria) that originated mitochondria (Martin et al., 2015; Mereschkowsky, 1905; Sagan, 1967). These introns would have invaded the host genome, where they proliferated and eventually evolved into spliceosomal introns (Cech, 1986; Koonin, 2006). By the time of the last eukaryotic common ancestor (LECA), both types of spliceosomal introns would already exist (Irimia and Roy, 2014). The very presence of a nucleus could have evolved as a defence mechanism to prevent unspliced or mis-spliced transcripts to be translated into proteins (Martin and Koonin, 2006)

During the divergence of eukaryotic lineages from their common ancestor, several clades —such as yeast— lost most of their introns. Massive intron loss is associated to loss of heterogeneity in intronic *cis*-elements involved in splicing (Irimia and Roy, 2008; Irimia et al., 2007; Schwartz et al., 2008). In extreme cases, U12 introns were completely lost, as it happens in *S. cerevisiae*. However, many eukaryotic lineages (including most vertebrates) retained a high amount of ancestral introns, and even acquired many novel ones (Irimia and Roy, 2014).

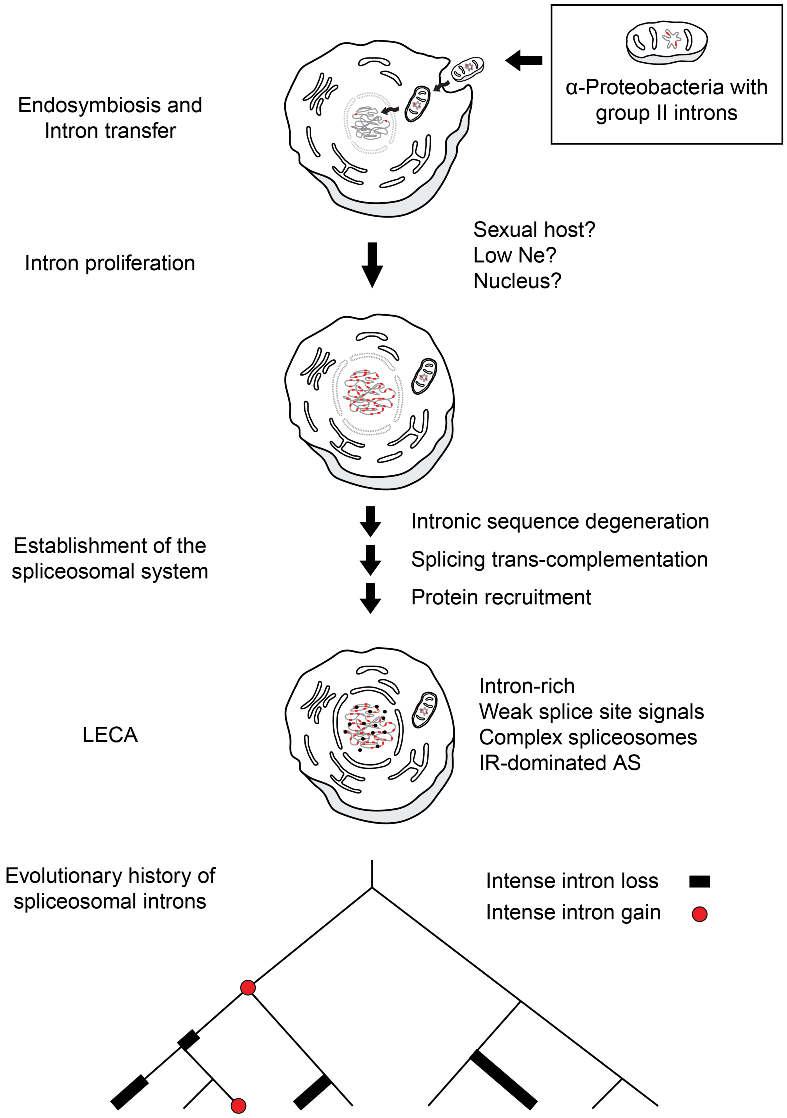

α-Proteobacteria with group II introns
Endosymbiosis and Intron transfer
Intron proliferation
Sexual host?
Low Ne?
Nucleus?
Establishment of the spliceosomal system
Intronic sequence degeneration
Splicing trans-complementation
Protein recruitment
LECA
Intron-rich
Weak splice site signals
Complex spliceosomes
IR-dominated AS
Evolutionary history of spliceosomal introns
Intense intron loss ▬
Intense intron gain ●

**Figure 7: A reconstruction of the process of establishment of the spliceosomal system in LECA during eukaryogenesis, and a representation of the diverse evolution of intron number across eukaryotic clades.** Adapted from (Irimia and Roy, 2014).

### 1.4.3 Excision mechanisms of spliceosomal introns

Spliceosomal introns are removed through two bimolecular transesterification reactions (SN2). In the first step, the 2’-OH group of the BP adenosine carries out a nucleophilic attack on the phosphodiester bond at the exon-intron boundary on the 5’ SS of the intron, leaving a free 3’-OH group at the end of the 5’ exon. In the second step, this group acts as a nucleophile on the phosphodiester bond at the exon-intron boundary on the 3’ SS. The final products of the reaction are a shortened pre-mRNA with both exons joined together, and the intron in a lariat configuration, with its 5’ SS bound to the BP adenosine. Interestingly, both reactions are reversible.

The spliceosomal machinery consists of several RNA-protein complexes (U1, U2, U4/U6, U5, U6atac/U4atac, U11 and U12), collectively called small nucleolar ribonucleoproteins (snRNPs), and assembled in a step-wise manner during the splicing reaction. Each snRNP is made of a single snRNA (two, in the case of U4/U6 and U4atac/U6atac), a common set of seven Sm proteins, and a variable set of snRNP-specific proteins. Additionally, many proteins not associated with any snRNP take part in the process. Altogether, proteins constitute two thirds of the total mass of the spliceosome when it is assembled on short introns. Unlike other RNP molecular machines, such as the ribosome, snRNPs do not have pre-assembled catalytic sites, and the splicing reaction depends on a sequence of complex structural rearrangements in the spliceosomal subunits happening while the spliceosome is being assembled upon the mRNA.

#### 1.4.3.1 Splicing of U2 introns

U2 spliceosomal introns are spliced by a set of snRNPs denominated *major spliceosome*, made of U1, U2, U4/U6 and U5. The canonical splicing mechanism for these introns (summarised in Figure 8) involves the recognition of intron-exon borders by spliceosomal factors, and the assembly of the spliceosome across the intron.

In a first step, the 5’ region of the U1 snRNA binds to the 5’ SS of the intron, and the binding is stabilised by the snRNP-associated protein U1C. It is known that the U1 snRNP can interact with the CTD domain of RNA polymerase II, which could have a role in co-transcriptional positioning of U1 at this stage. In parallel, the SF1 protein recognises and binds the BP sequence. In metazoans, additional interactions are established between the pre-mRNA and U2 auxiliary factors: U2AF65 binds to the polypyrimidine tract, and U2AF35 to the 3’ SS. These interactions are individually weak, but stabilised because of their cooperative character, as the proteins involved establish contacts with each other. For instance, it is known that U2AF65 interacts with the C-terminal RNA recognition motif (RRM) of SF1. At this point, binding of spliceosomal components to the pre-mRNA do not require ATP hydrolysis. In metazoans, SR proteins and the cap-binding complex also contribute to the stabilization of this complex, known as complex E.

From this point, the splicing reaction requires ATP hydrolysis in vivo, in order to catalyse conformational changes of the spliceosome. The internal region of the U2 snRNA pairs the BP and the PPT, displacing SF1, and forcing the intron into a conformation where the BP adenosine bulges out from the structure, exposing the 2’-OH group that will act as nucleophile in the first splicing reaction. The pairing is stabilised by the U2 proteins SF3a and SF3b, as well as the SR domain of U2AF65. At this point, the spliceosome-pre-mRNA complex is denominated A complex.

The next step involves the recruitment of U4/U6 and U5, which enter the spliceosome as a protein-RNA complex called tri-snRNP. In the tri-snRNP, U4 and U6 are bound together by base pairing, while U5 is attached to the complex through protein-RNA interactions. In this conformation, U4 masks the 5’ region of U6 —which will form part of the catalytic core of the spliceosome at a later stage—, avoiding premature cleavage of the pre-mRNA. With the incorporation of the tri-snRNP, the spliceosome (now called B complex) has incorporated all the snRNPs, although in a catalytically inactive conformation.

Then, the U4/U6 hybrid is dissociated by Prp8, and U4 and U1 are released from the spliceosome. This exposes the 5’ end of U6, yielding a complex known as the activated B complex. Subsequent rearrangements of the spliceosome by Prp2 make the 5’ end of U6 pair with the 5’ SS, using an ACAGA motif located in U6. At the same time, the region of U6 downstream of this motif pairs with U2, and the central region of U6 assembles into an internal stem-loop called U6-ISL. This loop is responsible for positioning a pair of divalent ions (usually $Mg^{2+}$), which also take part in the splicing reaction. At this step (B* complex), catalytic activation of the spliceosome by the Prp2 helicase initiates the first transesterification between the BP 2’-OH bound to U2 and the 5’ SS bound to U6.

After the first transesterification, the spliceosome is at the C complex stage. In this conformation, the stem-loop in U5 contacts nucleotides from both exons in regions adjacent to the intron-exon boundaries, bringing them together for the second transesterification. In a final stage, the processed RNA products —the pre-mRNA with both exons bound together, and the intron in a lariat form— are released, and the snRNPs are separated and recycled for subsequent splicing rounds.

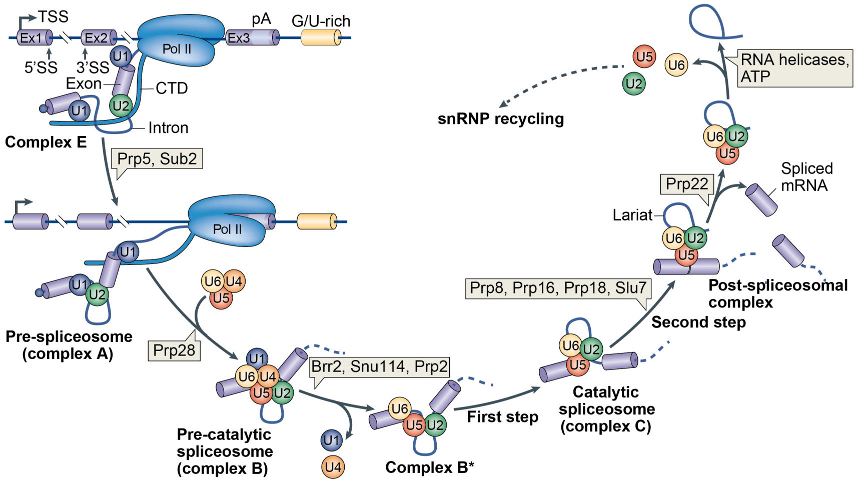


**Figure 8: Step-wise assembly mechanism of the U2 spliceosome.** Boxes represent auxiliary proteins intervening in each step of the process. Adapted from (Matera and Wang, 2014).

The relative contribution of snRNAs and proteins in the catalysis of the intron excision reaction is a subject of intense debate, although it is known that snRNAs are essential in this process, as shown above. In fact, splicing of a U2 intron has been achieved in an *in vitro* protein-free system with only U2 and U6 snRNAs, although the reaction ocurred at a very slow rate, highlighting the role of peptidic splicing factors in increasing splicing efficiency and speed. In metazoans, the spliceosome is believed to include more than 170 proteins, although this number is reduced to 80 proteins in yeast. This can be attributed to the higher homogeneity of *cis*-acting sequences in yeast, in comparison with metazoan transcriptomes.

At any individual stage of splicing —except for the E complex— the spliceosome contains about 125 protein subunits. Each spliceosomal snRNP has approximately 45 distinct snRNP-associated proteins. These proteins belong mainly to two classes: DExD/H-box helicases and peptidyl-prolyl isomerases (PPIases). DExD/H-box helicases use ATP for structural rearrangements of the spliceosome throughout the splicing reaction. In the formation of the A complex, Sub2 and Prp5 facilitate the exchange of SF1 for the U2 snRNA at the BP. In the B complex, the helicase Brr2 unwinds the U4/U6 duplex, and Prp28 contributes to the transfer of the 5' SS from U1 to U6, yielding the B* complex. These two activities need to be highly coordinated in time, and Prp8 — which occupies a central position at the core of the catalytic site— mediates this coordination. Prp2 is also required after U4/U6 dissociation, but before the first catalytic step. In the second splicing reaction, there is an essential contribution from Prp16 and Prp22. Finally, spliceosome disassembly requires the activity of the helicase Prp43. In general, DExD/H-box helicases also have a key role in increasing splicing fidelity, as they act as kinetic proofreaders of the process, giving time for release of undesired intermediates. On the other hand, PPIases are responsible for conformational changes in proline residues of other protein factors, accelerating spliceosomal rearrangements. At least six PPIases take part in the splicing of U2 introns.

The splicing process is characterized by the high degree of protein exchange between stages. In yeast, the transition from the A complex to the B complex involves the recruitment of about 25 proteins as part of the tri-snRNP, and 35 additional non-snRNP proteins. In human, approximately 65 proteins are recruited and 10 are lost. Many of these proteins belong to the Prp19 complex (the homologue of the yeast NTC complex). In yeast, this complex has 7 subunits with 4 copies of Prp19 each. In the transition from the B complex to the C complex, about 35 proteins are recruited to the spliceosome, and a similar number of proteins are lost. Most of the released proteins belong to the U1, U2 and U4 snRNP.

Many of these protein exchanges are mediated by conformational changes induced by post-translational modifications (PTMs), especially phosphorylations and dephosphorylations (reviewed in Wahl et al., 2009). It is known that SR proteins need to be phosphorylated to function, and Prp8 must be phosphorylated before tri-snRNP addition to the spliceosome. On the other hand, U1-70K needs to be dephosphorylated before the first splicing reaction occurs. PTMs in some proteins are applied and reversed at different stages. For instance, the U2 protein Sf3b155 is hyperphosphorylated before the first catalytic step, but it needs to be dephosphorylated for the second splicing reaction to take place. Besides phosphorylations and dephosphorylations, other PTMs, such as protein ubiquitylation and acetylation, also happen during the splicing process. Interestingly, binding interfaces of several spliceosomal proteins map to intrinsically disordered regions (IDRs). For example, SF3b155 has a disordered N-terminal region of nearly 450 amino acids, while it undergoes a disorder-to-order transition upon binding to SF3b14a. It is thought that the higher radius of gyration of the disordered state allows SF3b155 to sample a larger fraction of its vicinity, thus increasing its effective interaction radius.

Alternative assembly pathways have also been proposed for the U2 spliceosome. In yeast, a penta-snRNP complex without pre-mRNA has been detected in vitro. This complex can then transition to an active spliceosome upon addition of a pre-mRNA and other splicing factors, without an intermediate step of dissociation (Stevens et al., 2002).

When introns are much longer than their neighbouring exons, the interaction between U1 and U2 can happen across an exon, instead of across the intron, in a process called exon definition (Berget, 1995). In vitro assays show that splicing in Drosophila pre-mRNAs occurs through formation of an exon definition complex when introns have lengths above 250 nt (Fox-Walsh et al., 2005). This means that splicing of many metazoan introns requires formation of an exon definition complex instead of the canonical cross-intron assembly pathway shown above, as metazoan exons are usually small (50-250 bp) compared to their flanking introns (Robberson et al., 1990).

During the splicing reaction, the exon definition complex is equivalent to complex A in the canonical spliceosome assembly pathway. Eventually, this complex transitions to an intron-defined complex equivalent to complex E (De Conti et al., 2013; Reed, 2000; Smith and Valcárcel, 2000). The precise mechanisms for this transition are not completely understood, although it is believed that many regulatory mechanisms for splice site selection operate at this stage (Bonnal et al., 2008; House and Lynch, 2006; Sharma et al., 2008). In fact, it is known that, because of the heterogeneity of their *cis*-acting elements with respect to consensus sequences, most of the introns in metazoans cannot be constitutively spliced by the spliceosome, and additional splicing factors are required for intron removal (Lim and Burge, 2001). One of these factor families are SR proteins, which bind to *cis* elements inside exons. There, they can stabilise the exon definition complex by bridging the structural gap between U2 and U1, located upstream and downstream of the exon, respectively (Hoffman and Grabowski, 1992; Reed, 2000). In turn, these proteins can be antagonized by additional splicing factors, which greatly expands the regulatory landscape of splice site selection in metazoans (see section 2.1).

#### 1.4.3.2 Splicing of U12 introns

The U12 spliceosome, also called minor spliceosome consists of five snRNPs: U11, U12, U4atac, U5atac and U5 (Tarn and Steitz, 1996). U11 and U12 perform similar functions to U1 and U2, respectively, while the U4atac/U6atac di-snRNP replaces U4/U6. U5 is present both in U2 and U12 spliceosomes. Despite their functional analogy, the percentage of sequence identity between each U2 spliceosome snRNAs and their U12 counterparts is surprisingly low. As an example, human and yeast U6 snRNAs have more sequence identity —about 60% (Brow and Guthrie, 1988)— than human U6 and human U6atac.

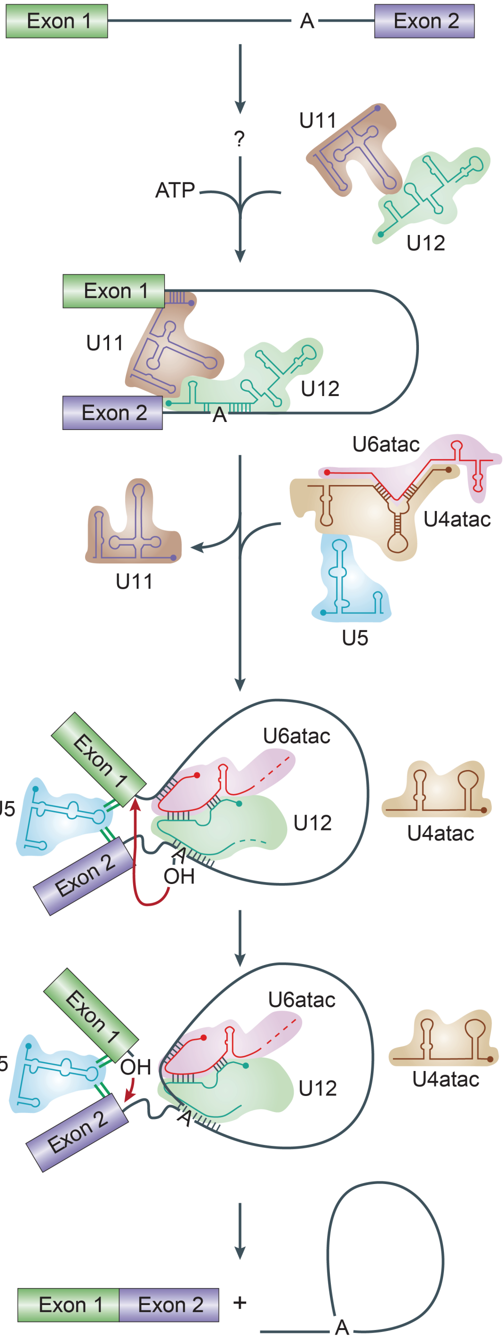
Important similarities can be found between the assembly pathways of both spliceosomes (Figure 8, Figure 9). First, U11 and U12 bind to the 5' and 3' SS of the intron and bring them together, although they do so as a di-snRNP (Frilander and Steitz, 1999; Wassarman and Steitz, 1992). In the next stage, a tri-snRNP formed by U4atac, U6atac and U5 is incorporated to the complex, and U11 and U4atac are released, allowing U6atac to interact with both U12 and the 5'SS, which triggers both catalytic steps. Interestingly, while in U2 introns the interaction between U6atac and the 5' SS must happen before U4/U6 dissociation and U6/U2 pairing, in U12 introns the interactions between U4atac and the 5' SS, and between U4atac and U12, can happen in any order (Frilander and Steitz, 2001).

Besides the U5 snRNA, a large set of proteins are shared between both spliceosomes, including key splicing factors as the Sm proteins, SF3b, the protein fraction of the tri-snRNP, and Prp8 (Luo et al., 1999; Schneider et al., 2002; Will et al., 1999)

An exon definition mechanism can also operate in excision of U12 introns, even if they are flanked by U2 introns —which is usually the case—. Splicing rates of U12 introns are usually increased upon removal of neighbouring U2 introns (Wu and Krainer, 1996), and purine-rich splicing enhancers bound by SR proteins have been also been detected in exons flanked by U12 introns (Dietrich et al., 2001; Hastings et al., 2001; Wu and Krainer, 1996), which suggests that these elements can have a similar role than in U2 introns.

U12 introns are present in plants, as well as many metazoan taxa (including vertebrates and insects), and they have not been detected in several model organisms such as *S. cerevisiae*, *S. pombe* and *C. elegans*. Despite the scarcity of U12 introns in the organisms where they have been found —there are only 404 characterised U12 introns in the human genome (Levine and Durbin, 2001)— they are very conserved, and the U12 spliceosome has been found to be essential for development in *Drosophila* (Otake et al., 2002), which reveals their functional importance. Since U12 introns are removed more slowly than U2 introns in vivo, and the exchange of some U12 introns for U2 analogues has been found to increase the concentration of the corresponding mRNA and protein up to six-fold, it has been suggested that the reason of this conservation could be their role as rate-limiting steps in splicing of their pre-mRNAs (Patel et al., 2002).

*(Next page)* **Figure 9: Assembly of the U12 spliceosome.** Transesterification reactions are indicated by red arrows. Small ribonucleolar particles (snRNPs) are drawn as a small nucleolar RNA (snRNA) with the 5' terminus marked as a dot), surrounded by a shaded area representing associated proteins. Adapted from (Patel and Steitz, 2003).


Exon 1
A
Exon 2
?
ATP
U11
U12
Exon 1
U11
U12
Exon 2
A
U6atac
U4atac
U11
U5
U6atac
Exon 1
U5
U12
Exon 2
A
OH
U4atac
U6atac
Exon 1
U5
OH
U12
Exon 2
A
U4atac
Exon 1
Exon 2
+
A

# 2. ALTERNATIVE SPLICING

The discovery of splicing in adenovirus included a surprising additional finding: the same viral transcript was processed into different mRNAs, which in turn generated more than ten different proteins (Chow et al., 1977). This differential processing of transcripts through usage of multiple sets of splice junctions to originate multiple mRNA isoforms receives the name of alternative splicing (AS).

AS can involve several types of molecular events, as shown in Figure 10A. Complete exons (called cassette exons, hereafter AltEx events) may be included or skipped from their transcripts, or they can have alternative SS at their 5’ or 3’ end (Alt3 and Alt5 events). In other cases, polyadenylated mRNAs can also show intron retention (IR) events, when an intron is not excised before the transcript exits the nucleus. Moreover, many of these events —of the same or different categories— can coexist in the same transcript, giving rise to more complex AS event combinations like mutually exclusive exons. Some other sources of transcriptomic variation, such as the existence of alternative promoters or alternative polyadenylation sites in a gene, do not involve AS, although they can also introduce changes in the sequence of a mature mRNA.

From an evolutionary point of view, the question of which genomic features underpin the biological complexity of organisms with a higher number of distinct cell types has been extensively studied since the publication of the first draft sequence of the human genome (Lander et al., 2001; Venter et al., 2001). The discovery that the number of human genes is not substantially different from those of other model organisms, such as the nematode *C. elegans* or the thale cress *A. thaliana* (Betrán and Long, 2002), known as the *G-value paradox* (Betrán and Long, 2002; Hahn and Wray, 2002), sparked a search for alternative explanations to the

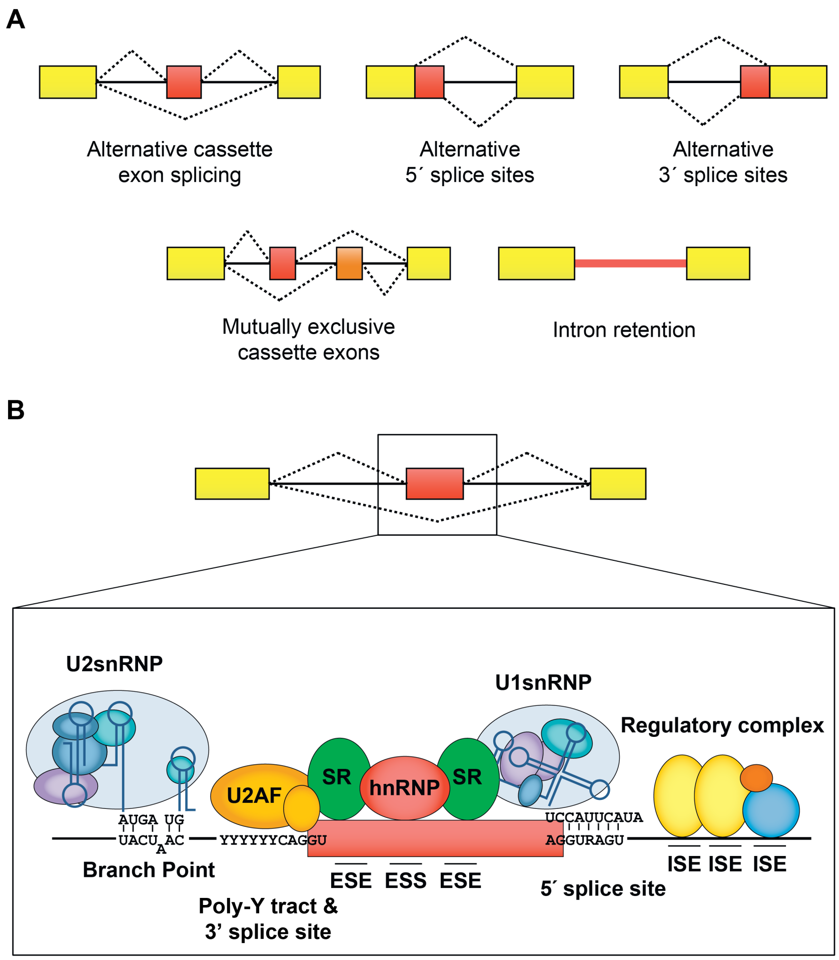


**Figure 10: A) Main types of AS events. B) Representation of some of the main *cis*- and *trans*-acting regulators of most splicing reactions in vertebrates.** SR: Ser/Arg-repeat containing protein. hnRNP: heteogeneous ribonucleoprotein. U2AF: U2 snRNP auxiliary factor. ISE: intronic splicing enhancer. ESE: exonic splicing enhancer. ESS: exonic splicing silencer. Adapted from (Irimia and Blencowe, 2012).

higher diversity of human cell types (which, under this model, is assumed as equivalent to increased biological complexity).

Nowadays, the hypothesis that AS could be one of the main explanations for this missing genomic complexity is widely supported, especially in the metazoan lineage (Barbosa-Morais et al., 2012; Blencowe, 2006; Chen et al., 2014; Kim et al., 2007). An important reason for this are several studies published in the last decade that highlight the widespread occurrence of AS in vertebrates. According to these studies, about 95% of human multi-exonic genes undergo AS (Pan et al., 2008; Wang et al., 2008).

## 2.1 Regulation of AS

Computational analyses in eukaryotic model organisms reveals that, although the information content of 5’ and 3’ SS, PPTs and BP regions may be enough to specify exon-intron boundaries in *S. cerevisiae*, additional recognition mechanisms are required in other organisms such as human and *A. thaliana* (Lim and Burge, 2001). This is related with the degeneration of splicing-related sequences in these organisms, as shown on section 1.4.3 (Figure 6). In many cases, intervention of these additional factors makes splice site selection highly variable between tissues, cell types, developmental stages and biological conditions. Consequently, a plethora of *cis*- and *trans*-acting factors —in addition to the core spliceosomal components and the essential regulatory sequences described in section 1.4— are involved in AS regulation, in order to ensure both robustness and flexibility of the system. An overview of some of these factors is shown in Figure 10B.

*Cis*-acting sequences involved in AS are typically catalogued depending on their intronic or exonic location, and the effect they induce on splicing of their nearest SS. Therefore, we can distinguish four types of such sequences: intronic splicing silencers (ISS), intronic splicing enhancers (ISE), exonic splicing silencers (ESS) and exonic splicing enhancers (ESE). *Trans*-acting factors involved in splicing regulation are typically RNA-binding proteins, regulating SS choice through very diverse molecular mechanisms of action. Further elucidation of these mechanisms is a field of active research, as many of them are still poorly understood (reviewed in Nilsen & Graveley, 2010). Several splicing enhancers can actively recruit basal spliceosomal components to near SS, but there are also reports of enhancement of distal SS by alternative mechanisms involving communication between splice sites in a transcript. Splicing inhibitors, on the other hand, usually act by direct steric hindrance of the inhibited SS, but their incorporation to the pre-mRNA can also lead to the formation of *dead-end* complexes —assemblies of RBPs that prevent the formation of a full spliceosome for a pair of SS—.

In terms of *trans*-acting factors, many splicing regulators belong to the SR and hnRNP families. An early classification established that SR proteins —which receive their name from characteristic domains with high serine and arginine content— recognize ESE elements both in constitutive and alternative exons, from where they recruit components of the core spliceosomal machinery. On the other hand, hnRNP proteins —an acronym for *heterogeneous nuclear ribonucleoproteins*— usually act as repressors, by binding both ISS and ESS elements and hindering the assembly of the spliceosome (Mayeda and Krainer, 1992). However, the current interpretation of AS regulation points to a much more complex model. Two main principles seem evident from the great deal of studies about AS regulation performed after the advent of high-throughput technologies such as splicing arrays and RNA-seq.

First, AS regulation is fine-tuned by the collective effect of many regulatory proteins, through multiple cooperative low-affinity interactions that exert combinatorial control over the spliceosome (Smith and Valcárcel, 2000). In many cases, cross-regulatory relationships exist between these proteins, meaning that expression levels of a single splicing factor are often a poor predictor of inclusion profiles in an AS programme. One of the causes for this is the existence of abundant cross-regulatory relationships between splicing factors, in which some of these proteins mediate AS events in transcripts of genes also involved in splicing regulation, two splicing factors act in complex with each other, or their joint action at different points of the pre-mRNA is required for an effect on splice site selection.

Second, identification of protein-RNA interactions in transcriptome-wide studies by means of RNA maps (Witten and Ule, 2011) has revealed that each splicing regulator can have antagonistic functions depending on its binding site with respect to the excised intron (reviewed in Fu and Ares, 2014). This is the case for many proteins both in the hnRNP and SR families.

A particular set of proteins stands out among the repertoire of trans-acting factors involved in AS regulation. Some splicing factors are only expressed in very restricted sets of tissues, where they act in combination with the rest of regulators and their *cis*-acting elements to activate and repress entire AS programmes. Because of their dominant effect on the AS landscape in the tissues where they are expressed, these factors act as *master regulators* of tissue-specific AS. Abundant examples of these regulators, the cross-regulatory relationships established between them and their context-dependent impact on the splicing patterns of several tissues are described in sections 2.2.1 to 2.2.3.

## 2.2 Functions of AS

The potential of AS to produce different transcript isoforms of the same gene can have very significant effects at the proteomic level, by generating diverse protein isoforms upon translation (see Discussion). Because these protein isoforms differ in their peptidic sequences, they can acquire different features in their three-dimensional structures, and therefore, they can present differences in their function, subcellular localization, activity, binding affinity for interaction partners, or ligand specificity.

The degree of isoform diversification introduced by AS can vary drastically between genes. While genes with a single AS event have the potential to express only two different transcript isoforms, several examples of extreme isoform diversification have been described. For example, the *Dscam* gene of *D. melanogaster*, encoding a cell adhesion molecule, contains 95 AltEx exons, arranged in four modules so that exons in each module are mutually exclusive. The *Dscam* gene has been reported to originate 38,016 distinct isoforms, which means that this gene alone generates a number of isoforms exceeding the number of genes in the whole *D. melanogaster* genome (Schmucker et al., 2000).

Importantly, mRNA isoforms originated by AS can show differences in features not directly related to their coding sequences with other effects in quantitative regulation of gene expression. Alterations in the 5’ and, especially, 3’ untranslated regions (UTRs) of a transcript can affect miRNA binding sites, which in turn can lead to variations in transcript stability. Introduction of an in-frame termination codon through AS can influence transcript stability through mechanisms such as the nonsense-mediated decay (NMD) pathway of mRNA degradation (reviewed in Chang et al., 2007). Several studies suggest that the triggering of the NMD pathway by a particular mRNA can be predicted according to the relative position of its first in-frame termination codon and its last exon-exon junction (EEJ). If the termination codon lies more than 50 nt upstream of an EEJ, the mRNA is likely to undergo NMD (reviewed in Kurosaki & Maquat, 2016; Popp & Maquat, 2016), contributing to an effective decrease in gene expression levels.

Although every kind of AS event can potentially introduce premature termination codons (PTCs), this is especially frequent in the case of IR events, since PTCs located in intron sequences are not subject to negative selection, as opposed to their counterparts in exons. In fact, short eukaryotic introns are under selective pressure to cause NMD through PTC introduction (Jaillon et al., 2008). While the majority of tissue-regulated IR events down-regulate the expression levels of their transcripts through NMD, others trigger nuclear sequestration of their polyadenylated transcripts, that are eventually degraded in the nucleus by alternative pathways (Braunschweig et al., 2014). Interestingly, some studies in human and mouse cells report the detection of a conserved program of retained introns causing both nuclear retention and a reduction of nuclear degradation rates of their transcripts (Boutz et al., 2015). These introns are known as detained introns, and their excision can be triggered by inhibition of Clk proteins —a family of kinases responsible for SR protein phosphorylation in the nucleus, whose transcripts also have detained introns— or DNA damage. Detained introns are enriched in genes belonging to RNA processing pathways, including splicing, and therefore, their regulation can indirectly affect AS patterns of many other genes upon these environmental cues.

In quantitative terms, high-throughput mass spectrometry (MS) studies report the detection of multiple protein isoforms for approximately 37% of human protein-coding genes (Kim et al., 2014). The contribution of AS to tissue- and developmentally-regulated transcriptomic diversification, and its potential to introduce alterations both in translated mRNA regions and UTRs —as previously highlighted— suggest that AS-induced transcript variability between tissues could in turn have important effects at the proteomic level. However, the extent of this effect is currently under debate (see Discussion).

Many biological processes at very different pathways are known to be regulated by certain AS events (reviewed in Kelemen et al., 2013), ranging from sexual differentiation and dosage compensation in Drosophila (Valcárcel, Singh, Zamore, & Green, 1993; reviewed in Lucchesi & Kuroda, 2015), to regulation of apoptosis (Izquierdo et al., 2005; Li et al., 2007), modulation of ion channel activity (Adams et al.; Dabertrand et al., 2006; Schlesinger et al., 2005), and hormone secretion (Fernández et al., 2009). With the revelation of the widespread occurrence of AS in vertebrates —and especially, in mammals—a more systematic view has arisen, as some of the studies highlighting the prevalence of AS in human and other metazoans also remark the existence of co-regulated AS programmes in these organisms, with differential splicing signatures characteristic of certain tissues, cell types and developmental stages (Baralle and Giudice, 2017; Wang et al., 2008). After these discoveries, the importance of AS as a genome-wide regulatory layer of gene expression has become more evident, and the hypothesis that proteome diversification through AS is a major molecular source of organismal complexity in metazoans has gained strength.

## 2.3 AS programmes

The landscape of AS in a particular biological context —such as a certain cell type in physiological conditions— can be represented as a network with two different types of nodes, corresponding to splicing regulators (RBPs) and their target exons and introns, each of them connected to the RBPs influencing their splicing pattern. In these representations, master regulators have a prominent role, since they act as hubs orchestrating the behaviour of the whole network, both directly and through cross-regulatory relationships with other splicing factors. The topology of the network will diverge between biological conditions because of variations in the expression levels of regulatory RBPs and their target genes.

It cannot be forgotten that, after accounting for variations in gene expression, AS is only the first among many layers of molecular diversification across cell types. However, transcriptomic complexity induced by AS can induce a cascade effect when it is considered in conjunction with downstream regulatory mechanisms such as translation rates, mRNA and protein stability, and post-translational protein modifications, among others. The compound effect of variations in all these regulatory layers across the landscape of internal and external conditions can induce a dramatic amount of molecular variability, which ultimately constitutes the identity of each cell type in an organism (Alonso and Wilkins, 2005; Jangi and Sharp, 2014; Niklas et al., 2015).

AS programmes have been described in many different vertebrate tissues and cell types. The following sections of this thesis contain a brief description of some of these programmes, as well as their corresponding regulators. For simplicity, we will focus on the landscape of tissue-specific AS in neural tissues, muscular tissues, and ESCs. For a more detailed overview of AS in other tissues, the reader is referred to many excellent reviews previously published on the same topic (Baralle and Giudice, 2017; Gallego-Paez et al., 2017; Jangi and Sharp, 2014).

The following sections of this chapter contain a brief description of some of the tissue-specific AS programmes characterized to date, as well as their corresponding regulators. For a more detailed overview, the reader is referred to many other articles previously published on the same topic (Baralle and Giudice, 2017; Gallego-Paez et al., 2017; Jangi and Sharp, 2014).

### 2.3.1 AS in vertebrate neural tissues and neuronal differentiation

Of all vertebrate tissues and cell types, neurons are arguably the one with the most characteristic, evolutionary conserved and best characterized specific AS programmes (Baralle and Giudice, 2017; Barbosa-Morais et al., 2012; Licatalosi and Darnell, 2006; Melé et al., 2015). As previously mentioned, the main driving force for the establishment of these programmes is the combined effect of the tissue-restricted expression of several master splicing regulators, and a variety of cross-regulatory interactions between splicing factors causing some of these proteins to undergo neural-specific AS. In this section, we will focus on the role of splicing factors from the NOVA, RBFOX and PTBP families, as well as the SRRM4 protein.

AS events up-regulated in neurons and during neurogenesis are often related to functions characteristic to this cell type. As an example, a comparison between the transcriptomes of embryonic and adult cerebral cortices in mouse revealed a programme of more than 400 AS events with differential inclusion patterns between embryonic and adult samples. Events selectively included in adult samples showed enrichment in genes associated with ion transport and homeostasis and transmission of nerve impulse, while up-regulated events in embryonic samples were enriched in genes related to cell cycle regulation and mitosis. The fact that 31% of these AS changes were not complemented by changes in global expression levels of their genes is a strong indicative of the importance of post-transcriptional regulation, and specifically AS, in neuronal development (Dillman et al., 2013).

RBPs from the NOVA family were among the first characterized neural-specific splicing factors (Ule et al., 2005). Both mammalian paralogues —NOVA1 and NOVA2— are highly enriched in neurons and NPCs, where they regulate AS programmes essential for neuronal survival, inhibitory synaptic transmission, and axon guidance (Ule et al., 2005). Events regulated by Nova proteins include an exon in Dab1 critical for postnatal migration of cortical neurons (Yano et al., 2010), and an exon in the Agrin gene critical for neuromuscular junction formation (Ruggiu et al., 2009). The search for splicing targets and mechanisms of action of Nova proteins led to the development of several novel approaches now well established in the field of AS. Two examples of this are the detection of *cis*-regulatory RNA sequences responsible for binding of particular RBPs by coupling cross-linking and immunoprecipitation with RNA sequencing (CLIP-seq) (Licatalosi et al., 2008) —of which many variations would be developed later (reviewed in Li, Song, & Yi, 2014), including single-nucleotide resolution CLIP (iCLIP) (Huppertz et al., 2014)— and the description of transcriptome-wide binding preferences of RBPs with RNA maps (Ule et al., 2006; Witten and Ule, 2011). These efforts revealed the strong sequence specificity that mediates the binding of NOVA proteins to their target alternative exons and introns, strongly enriched in YCAY motifs.

The RBFOX protein family has three paralogues in mammalian species —RBFOX1, RBFOX2 and RBFOX3 in human—. All of them have tissue-regulated expression and are expressed in neurons, although only RBFOX3 is neural-specific. As an example of cross-regulation between splicing factors involved in tissue-specific AS programmes, RBFOX1 and RBFOX2 are translated into a rich variety of tissue-specific protein isoforms (Kuroyanagi, 2009). Skipping of exon 19 in RBFOX1 —mediated by the Rbfox1 protein itself— leads to the production of a nuclear protein product, while the inclusion isoform can be translated into a cytoplasmic protein (Lee et al., 2009). Up-regulation of RBFOX1 nuclear isoforms in an in vitro neuron differentiation assay triggered an AS programme of approximately 900 exons, many of them with canonical RBFOX1 binding sites in the corresponding downstream introns. This programme includes RNA processing factors such as HNRNPH1, HNRNPA1 and HNRNPD, as well as many genes involved in neuronal development and cytoskeletal organization (Fogel et al., 2012; Lee et al., 2016). In turn, nuclear RBFOX1 regulates the stability of mRNAs related to synaptic transmission, including candidate genes for autism spectrum disorder (ASD) susceptibility (Lee et al., 2016; Voineagu et al., 2011).

An antagonistic temporal variation in the expression levels of PTBP1 and SRRM4 exists during mammalian neurogenesis. The high expression of PTBP1 in neural progenitors mediates skipping of a 34 nt exon in the PTBP2 mRNA, which is subjected to NMD because of the introduction of an in-frame PTC. However, in later stages, a drastic reduction of PTBP1 levels and up-regulation of SRRM4 induces inclusion of this exon. Successful translation of PTBP2 then initiates a splicing programme responsible for neural differentiation (Boutz et al., 2007; Makeyev et al., 2007; Spellman et al., 2007).

Therefore, PTBP1 and PTBP2 have antagonistic functions in differentiating neurons (reviewed in Vuong et al., 2016). Transcriptome-wide RNA-binding profiles of PTBP1 and PTBP2 largely overlap and, interestingly, depletion of PTBP1 induces trans-differentiation from fibroblasts into neurons (Xue et al., 2013), but its germline knockout in mice increases embryonic mortality (Shibayama et al., 2009; Suckale et al., 2011). A combination of the higher affinity of PTBP1 for its targets and its repressive action on PTBP2 itself keeps PTBP2-regulated exons skipped during early stages of neuronal differentiation. In later stages, PTBP2 is down-regulated again, which induces a transition towards an AS landscape characteristic of mature neurons (Linares et al., 2015). This illustrates the complexity of the AS landscape in the context of neuronal differentiation.

The SR protein Srrm4 is among the few splicing factors with truly neural-specific expression. As previously mentioned, it also shows cross-regulatory interactions with other regulators of neural splicing, as it has been shown to down-regulate up-regulate inclusion of exon 10 of PTBP2, contributing to decrease its protein levels as neuronal differentiation progresses (Calarco et al., 2009). SRRM4 regulates an extensive group of neural exons, which often show sequence motifs for additional factors like CELF, RBFOX and MBNL proteins. In most cases, Srrm4 binds to UGC motifs located directly upstream of its regulated exons. Among the AS programme regulated by SRRM4 in mammals there is a set of very short exons (*microexons*), between 3 and 27 nt in length, that show extremely high conservation in vertebrates, neural-specific inclusion, and are strongly enriched for genes related to neuronal differentiation, vesicular transport, cytoskeletal reorganization and synaptic transmission (see Chapter 3). Srrm4-depleted mice show a tremor phenotype, and they die soon after birth because of defective diaphragm innervation and subsequent asphyxia. At the tissue level, they also show important defects in cortical layering, neurite outgrowth and brain midline crossing, supporting the essential role of SRRM4 for neuronal differentiation (Quesnel-Vallières et al., 2015). As other SRRM-regulated exons, microexons are also subject to regulation by RBFOX and PTBP1 proteins (Li et al., 2014b). Interestingly, many SRRM4 target exons lie in ASD candidate genes, and down-regulated SRRM4 expression and decreased microexon inclusion has been found in ASD patients (see Chapter 3).

### 2.3.2 AS in vertebrate muscular tissues

Together with neurons, striated muscle also shows a wide repertoire of specific AS events. As an example, a microarray-based study in 48 human tissues and cell types found more than 1,000 AS events with differential inclusion in skeletal muscle, and more than 500 events in heart (Castle et al., 2008). In particular, extensive splicing regulation has been observed during heart development, involving splicing factors such as RBM24, as well as proteins from the CELF, MBNL and RBFOX families (reviewed in Giudice & Cooper, 2014)

One of the best-known regulation mechanisms of muscle- and heart-specific AS programmes is the expression switch between CELF and MBNL proteins. As an example, 206 alternative exons have been found to respond similarly to Mbnl1 depletion or CELF1 induction in murine heart and skeletal muscle samples. In this study, both MBNL1 depletion and CELF1 induction in differentiated cardiomyocytes caused splicing transitions in opposite directions of those found in normal development, suggesting that replacement of CELF1 by MBNL1 throughout muscular and cardiomyocyte differentiation is a key driver of AS transitions in this developmental context (Wang et al., 2015). This is in agreement with expression patterns of these proteins, as CELF is down-regulated and MBNL is up-regulated as postnatal heart development progresses (Kalsotra et al., 2008).

The fact that exons up-regulated in the differentiated state are also enriched in GO terms related to cell-cell junctions, cell differentiation and microtubule cytoskeleton supports their role in establishing tissue identity of differentiated muscular tissue. Moreover, more than half of these AS events preserved the reading frame of their transcripts, suggesting that the effect of this AS programme is largely mediated by generation of alternative protein isoforms in the proliferative and differentiated states (Wang et al., 2015). Other studies report even larger AS transitions during normal postnatal heart development, involving changes of more than 500 AltEx events with inclusion changes of more than 20% in their percentage of inclusion, and enrichments in functions related to cardiac muscle contraction and vesicular trafficking in up-regulated exons, and DNA replication and cell cycle in down-regulated exons (Giudice et al., 2014). Validation of four of these events in skeletal muscle in vivo revealed that disruption of their splicing patterns induces defects in muscular function, including misregulation of calcium signalling and muscle force generation (Giudice et al., 2016).

Both CELF and MBNL proteins can regulate splicing of particular exons in a positive and negative manner, depending on the positioning of their *cis*-acting sequences with respect to the regulated exon. In general, exons up-regulated by MBNL1 and MBNL2 show MBNL binding sites in exonic or upstream exonic sequences, while down-regulated exons have MBNL binding sites in their downstream intronic region (Charizanis et al., 2012; Du et al., 2010; Wang et al., 2012). Interestingly, AS events regulated by CELF1 tend to show an opposite pattern, since events with CELF1 binding sites in the exonic region are generally repressed by this protein (Wang et al., 2015). As previously discussed, this context-dependent effect is also observed in many tissue-specific splicing factors, such as NOVA proteins (Licatalosi et al., 2008).

Besides their effect on splicing, there is also evidence for antagonistic effects of CELF and MBNL proteins in mRNA stability during heart development. Several mRNAs of genes related to acting cytoskeleton organization and mesoderm morphogenesis show CELF and MBNL binding sites in their 3’ UTRs, and the effect of these factors in mRNA stability seems dependent on the ratio of CELF to MBNL binding sites (Wang et al., 2015). This study suggests a model in which CELF-mediated degradation of mRNAs with higher expression in differentiated muscle can be avoided by the binding of MBNL proteins, which leads to the establishment of a muscular gene expression programme during postnatal heart development.

A particularly interesting case of heart-specific AS is the titin (TTN) gene. This gene has the highest number of exons of all known human genes, and many of them are regulated during postnatal heart development. In this context, AS of TTN induces an isoform switch regulated by RBM20, from a short neonatal isoform (N2B) to a long adult isoform (N2A) (Bang et al., 2001; Labeit and Kolmerer, 1995; Li et al., 2013). Because the alternatively spliced region encodes for a region (I-band) influencing the elasticity of the titin molecule, the ratio between N2A and N2B isoforms affects myocardium wall stiffness and sarcomere length. Consequently, mutations in RBM20 have been linked to several cardiomyopaties (Beqqali et al., 2016; Brauch et al., 2009; Guo et al., 2012). Other splicing events of special interest are located in additional genes related to cytoskeletal regulation during cardiomyocyte contraction —for example, cardiac troponin T (Goo and Cooper, 2009; Warf et al., 2009)—, and genes involved in calcium signalling, such as the Ca2+ ATPase transporter SERCA2, and the ryanodine receptor RYR2 (George et al., 2007; Ver Heyen et al., 2001).

Splicing networks related to skeletal muscle differentiation have also been detected, often with the use of myoblast cell lines such as C1C12 murine cells. The most important AS programmes found in differentiation of these cells are regulated by RBFOX2, QK and PTBP proteins (Hall et al., 2013; Runfola et al., 2015; Singh et al., 2014). Among the set of regulated events, there is a cassette exon in the MEF2D transcription factor regulating susceptibility to phosphorylation of its protein and, as a result, its intracellular localisation. While the ubiquitous MEF2D isoform is phosphorylated and bound to co-repressors, the muscular isoform is able to recruit ASH2L as a co-activator and enter the nucleus, where it triggers transcription of muscle-specific genes (Sebastian et al., 2013).

### 2.3.3 AS and pluripotency in vertebrate embryonic stem cells, reprogramming and trans-differentiation

The existence of AS transitions between ESCs and several differentiated cell types has been well documented (Cloonan et al., 2008; Han et al., 2013; Venables et al., 2013; Wu et al., 2010). Moreover, global regulation of AS landscapes have been also described to play an important role during differentiation of multi- and unipotent stem cells, as is the case during hematopoiesis (Clien et al., 2014), neurogenesis (Vuong et al., 2016a) and myogenesis (Bland et al., 2010). In line with their role as positive AS regulators of exons specific to other differentiated tissues and cell types, the RBPs MBNL1 and MBNL2 have been shown to have a supplementary function in differentiated cells, acting as splicing repressors of a set of 56 out of 119 exons differentially up-regulated in human ESCs (where these regulators are only expressed at very low levels) (Han et al., 2013). Events in this AS programme and the role of MBNL proteins in their regulation are largely conserved in mammals, and are enriched in genes associated to the cytoskeleton, kinase activity, and plasma membrane. Among the genes regulated by this AS programme there are also several transcription factors and chromatin regulators, which suggests important contributions of AS regulation to the transcriptional switch between the pluripotent and differentiated states.

Among these events, mutually exclusive inclusion of exons 18/18b (16/16b in mouse) of the transcription factor FOXP1 is especially relevant. Although several splicing variants of FOXP1 have been reported (Brown et al., 2008), this AS event regulates the production of two alternative protein isoforms where the alternative region encodes for part of the DNA-binding domain of FOXP1. Each of these two FOXP1 isoforms has different sequence specificity and binding affinity for its DNA targets. Inclusion of the 18b/16b exon (which is up-regulated in pluripotent cells) affects the transcription of genes related to early development and maintenance of pluripotency. While the pluripotent-specific FOXP1 isoform acts mainly as a transcriptional repressor, it also induces transcription of several genes with documented roles in the establishment and maintenance of pluripotency, such as OCT4 and NANOG (Gabut et al., 2011).

For more than a decade, it has been known that these two factors form part of a minimal set of four genes whose induction is sufficient to trigger cellular reprogramming —the reversal of the differentiation process through transcriptional, post-transcriptional and epigenetic changes, in order to confer pluripotency to cells from a differentiated lineage (Takahashi and Yamanaka, 2006)—. Consequently with the differences in AS profiles found during the process of differentiation, major AS transitions conserved in vertebrates and regulated by a reduction in the levels of MBNL1 and RBFOX2 have been characterized in the process of cellular reprogramming of fibroblasts to induced pluripotent stem cells (iPSCs) (Venables et al., 2013). This programme affects proteins involved in cell adhesion, migration and polarity, and can be reversed upon expression of MBNL1 and RBFOX2.

The initial stage in the process of cell reprogramming from fibroblasts to iPSCs also includes important AS transitions mediated by splicing factors of the ESRP family. Consequently, ectopic expression of ESRP has been linked to increased reprogramming efficiency (Cieply et al., 2016). ESRP1 and ESRP2 act as regulators of an AS programme characteristic of epithelial cells (Bebee et al., 2015). This programme is instrumental in epithelial-to-mesenchymal transitions (EMTs), a process that mediates morphogenesis of several embryonic structures during development (reviewed in Lamouille, Xu, & Derynck, 2014), as demonstrated by the cleft lip phenotype of mice embryos lacking ESRP1, and the generalized developmental aberrations found in double ESRP knockouts (Bebee et al., 2015). Co-option of ESRP-mediated AS for different morphogenetic programmes has been shown in many deuterostome species, although the number of overlapping ESRP targets between human and other species decreases rapidly outside of the mammalian lineage (Burguera et al., 2017).

One of the most interesting ESRP targets is an AS event conserved in chordates, which affects two mutually exclusive exons encoding for the IgIII domain in several paralogues of the FGFR family of receptors. This event alters the ligand specificity of the receptor, as each isoform is sensitive to epithelial or mesenchymal signals (reviewed in Turner & Grose, 2010). In accordance with the function of these two factors during development, the AS programme regulated by ESRP1 and ESRP2 during reprogramming also includes an AltEx event in exon 5 of the GRHL1 gene. Skipping of this exon causes a frameshift leading to the appearance of a PTC just upstream of the DNA-binding domain, and produces a truncated version of the GRHL1 protein, with reduced efficiency as a transcription factor. ESRP-mediated inclusion of this exon is associated to increased reprogramming efficiency, presumably because of the role of GRHL1 in inducing transcription of more genes related to pluripotency (Cieply et al., 2016).

Overall, this and other high-throughput analyses of transcriptome changes during reprogramming have shown multiphasic reversion of AS landscapes from the differentiated-state towards a pluripotent state (Cieply et al., 2016; Ohta et al., 2013; Tanaka et al., 2015) indicating that a widespread AS reprogramming is required for the acquisition of pluripotency. Interestingly, a genome-wide search for critical regulators of early stages of reprogramming identified the splicing factor SFRS11 as a crucial player (Toh et al., 2016). SFRS11 acts as a repressor of reprogramming partially through the splicing of ZNF207 isoforms. Moreover, this study showed over 400 genes that were alternatively spliced upon SFRS11 depletion. Amongst them were MBNL1, SFRS3 and U2AF1, strongly suggesting that SFRS11 could be regulating reprogramming via the splicing of downstream splicing factors.

# 3. FIRST ARTICLE: “A HIGHLY CONSERVED PROGRAM OF NEURONAL MICROEXONS IS MISREGULATED IN AUTISTIC BRAINS”

In this chapter, we make a first description of vast-tools, the RNA-seq analysis pipeline used for AS quantification throughout the present thesis, and apply this pipeline to a set set of more than 50 tissues and cell types in human and mouse, with the aim of making a comprehensive characterization of the AS landscape in neuronal tissues.

Among neural-regulated alternative exons, we found a strong negative association between exon size and neural specificity, which led to the discovery of a set of highly neural-specific microexons, strongly regulated by the SR protein SRRM4. These exons show strikingly high evolutionary conservation in vertebrates, and they are enriched in pathways related to neuronal biology. Consequently with this, they show sharp switch-like behaviour in their inclusion profiles during neuronal differentiation, and most of them peak during the postmitotic stage. Interestingly, we observed down-regulation of SRRM4 and microexons in brain samples from autistic patients, suggesting a decisive functional role for these events during neural development and for establishment of healthy neuronal function.

With respect to their protein properties, microexons show a striking contrast with other alternative exons, as they have an increased trend to map in or near ordered protein domains. However, our structural data shows that microexon inclusion usually does not significantly remodel the fold of the nearby domains. At the same time, microexon-containing proteins tend to occupy central positions in protein-protein interaction (PPI) networks. With these results, we hypothesize that microexons act in neurons by fine-tuning the function of their nearby domains.

My personal contribution to this work was the description of microexons at the structural and domain level, by mapping microexons to available PDB structures where possible, and obtaining structural models using Phyre2 for the rest of cases. This structural dataset allowed for visualisation of protein features associated with microexons that could not be predicted from sequence-based analyses alone. In addition, the code developed for this publication served as a foundation for a more comprehensive structure mapping described in Chapter 5.

Reference for this article and its supplemental materials:

Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell. 2014 Dec 18;159(7):1511–23. DOI: 10.1016/j.cell.2014.11.035

# 4. SECOND ARTICLE: “CONSERVED FUNCTIONAL ANTAGONISM OF CELF AND MBNL PROTEINS CONTROLS STEM CELL-SPECIFIC ALTERNATIVE SPLICING IN PLANARIANS.”

In this article, we apply vast-tools to neoblasts —pluripotent cells involved in regeneration— and differentiated cells from the planarian species *Schmidtea mediterranea*, in order to compare the landscape of neoblast-regulated AS with the splicing programmes observed in mammalian ESCs and their differentiation process. We detected several hundreds of conserved neoblast-regulated AS events, including a programme of exons in genes related to the extraordinary regenerative abilities of this organism. We show that many of these exons are regulated by the CELF and MBNL splicing factors. Our results point to the existence of a conserved programme of AS contributing to regulate pluripotency and differentiation in mammalians and planarians, despite more than 500 million years of independent evolution of these two lineages.

My personal contribution to this article was the analysis of the impact of some of these conserved neoblast-regulated exons at the protein level, as well as their relationship with the domain architecture of their proteins.

Reference for this article, including supplemental material and additional figures:

Solana J, Irimia M, Ayoub S, Orejuela MR, Zywitza V, Jens M, et al. Conserved functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative splicing in planarians. Elife. 2016 Aug 9;5. DOI: 10.7554/eLife.16797

# 5. THIRD ARTICLE: "AN ATLAS OF ALTERNATIVE SPLICING PROFILES AND FUNCTIONAL ASSOCIATIONS REVEALS NEW REGULATORY PROGRAMS AND GENES THAT SIMULTANEOUSLY EXPRESS MULTIPLE MAJOR ISOFORMS".

In this article, we thoroughly describe an updated version of the vast-tools pipeline, including a benchmark against several standard AS quantification tools widely used in the field. We also extend our previous analyses to a set of more than 300 RNA-seq samples from a diverse repertoire of tissues, cell types and developmental stages from human, mouse and chicken, in order to identify previously uncharacterized splicing programmes in these tissues.

Our results confirm the presence of highly tissue-specific AS signatures in brain, muscle, testis and pluripotent cells, and describe additional conserved modules of co-regulated exons in other tissues, such as liver, adipose tissue, colon, kidney and immune cells.

In addition, we describe a set of exons alternatively spliced in virtually all analysed tissues and cell types, which we named PanAS exons. These exons show high evolutionary conservation and a trend to preserve the open reading frame of their transcripts. Our analyses suggest that the alternative character of PanAS exons is maintained at the single cell level and in ribosome-engaged RNA, suggesting that these exons are able to originate multiple coexisting protein isoforms at the cellular level. In stark contrast with switch-like exons, PanAS exons are enriched in genes involved in processes related to transcriptional regulation of gene expression, suggesting distinct functional roles for both types of AS programmes.

Finally, this work also describes VastDB, our centralized repository of AS events across vertebrate species, tissues, cell types and developmental stages. VastDB integrates AS profiles and gene expression levels with a variety of biological features that contextualize these events, including the impact of AS events at the protein domain and structure level, evolutionary conservation of individual events, and their respective genomic context.

My personal contribution to this work included the description and characterization of PanAS events, the application of network analysis to reveal co-regulated splicing programmes in tissues distinct from brain and muscle, the assessment of evolutionary conservation of these programmes between our three species of study, and extensive involvement in the development and maintenance of VastDB.

Reference for this article, including supplementary material and additional figures:

Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017 Oct;27(10):1759–68. DOI: 10.1101/gr.220962.117

# 6. DISCUSSION

The pervasiveness of AS in vertebrate transcriptomes has been repeatedly confirmed. An indicative of this is the amount of human multi-exonic genes estimated to undergo AS, which has suffered a steady increase during the last twenty years, reaching 95% of multiexonic genes by using RNA-seq datasets (Pan et al., 2008; Smith and Valcárcel, 2000; Wang et al., 2008). The fact that virtually all mammalian genes are transcribed and spliced into more than one RNA isoform seems now evident from the growing number of transcriptome-wide studies performed in different tissues and organisms during the last decade. In this thesis, we contributed to this body of knowledge, applying vast-tools to study two AS programmes in neurons and ESCs, two of the tissues with the most distinctive transcriptomic profiles.

Of the many biological contexts where AS has been analysed, neuronal differentiation is one of the processes where the role of AS has been most frequently documented. In Chapter 3, we confirmed the existence of more than 2,500 AS events differentially included in neurons from human and mouse. More than 50% of these events are predicted to preserve the open reading frame of their transcripts, which constitutes an important enrichment in neural-specific events generating alternative protein isoforms in comparison with non-specific events. In addition, these frame-preserving events are enriched in genes related with biological functions especially relevant for neuronal biology, such as axonogenesis, synaptic transmission, vesicle trafficking, ion channels, and cytoskeletal organisation. Genes regulated at the gene expression and AS levels constitute two largely independent sets, with only 8.5% of genes regulated by both mechanisms. Overall, this suggests an important role of AS in regulating neuronal-specific processes through the production of alternative protein isoforms fine-tuning biological pathways that require special regulation in neurons in comparison to other tissues.

Among neuronal-regulated exons, we unveiled the existence of 225 microexons conserved between human and mouse. These exons, between 3 and 27 nt in length, show a set of features that make them particularly interesting. Their levels of conservation and neural specificity are significantly higher than those of longer exons, and they show a stronger trend to generate alternative protein isoforms and higher enrichment in neuronal pathways. These features situate microexons at the core of the neuronal-specific AS programme. According to our sequence analyses and structural predictions, and in contrast to longer alternative exons —which tend to encode for disordered regions—, microexons show a trend to map near structured protein domains, presumably fine-tuning the behaviour of their proteins to adapt them to the specific requirements of neuronal biology.

In Chapter 4, we gained insights on the role of AS in ESC biology by applying vast-tools to characterize AS programmes in two different cell populations on S. mediterranea, an invertebrate evolutionary distant from human and mouse showing great regenerative abilities mediated by neoblasts, a population of cells with similar pluripotency and replicative characteristics to vertebrate ESCs. We showed that there is a group of approximately 500 AS events with differential inclusion between neoblasts and differentiated cells, regulated by a conserved mechanism involving Bruli and MBNL-1, which are planarian homologues of the CELF and MBNL families of splicing factors in vertebrates. 68% of these exons potentially generate alternative protein isoforms, affecting to disordered protein regions of proteins related to cytoskeletal regulation and cellular signalling.

Interestingly, this neoblast-regulated AS programme is unexpectedly conserved in D. japonica, a planarian species with an phylogenetic distance to S. mediterranea similar to that between human and mouse. 15 of the genes regulated by this AS programme in planarian also show orthologous gene groups in human and, of those, 10 have AS-regulated exons with differential splicing in human ESCs. Our analysis also shows conservation at the exon sequence level in four of these events, highlighting the importance of some of these exons in the shared pluripotency and proliferative character found in planarian neoblasts and human ESCs.

While these levels of conservation do not seem very high at first sight, we must take into account the high evolutionary distance between these two organisms. More importantly, the regulatory role of MBNL proteins in an AS programme related to ESC pluripotency and differentiation, and the importance of CELF in the splicing landscape of several precursors, are well documented in vertebrates (see sections 2.3.1 and 2.3.2). Altogether, these results point to the existence of a conserved regulatory mechanism of an AS programme key for these biological functions, both in planarians and vertebrates, with only partial conservation in the regulated exons. In planarians, CELF proteins act as inducers of splicing profiles contributing to the maintenance of the pluripotent state, while MBNL proteins act in both clades repressing these exons as cells commit to differentiated lineages, and promoting inclusion of other exons characteristic of differentiated states. Future studies may confirm a role for CELF proteins in vertebrate ESCs, confirming the similarity of AS regulation in both organisms in this biological context.

In Chapter 5, we thoroughly benchmarked vast-tools against several other tools for AS quantification. Our pipeline performs at least as well as every other tool in the comparison in terms of computing time, memory usage, and event discovery rate, —especially in the detection of microexons— both in real and simulated datasets. We then applied vast-tools to detect AS events in a diverse set of tissues, cell types and developmental stages in human, mouse and chicken (with 108, 139 and 61 RNA-seq samples, respectively). Our results confirm the high degree of tissue specificity of neural, muscular, testis and embryonic AS programmes, and we have used a network analysis approach to identify additional conserved modules of co-regulated exons with differential inclusion in tissues with previous preliminary evidence of specific AS regulation, such as liver, adipose tissue and colon (Barash et al., 2010).

In the same chapter, we also describe PanAS exons, a set of exons alternatively included across most of the tissues and cell types analysed. This pattern of AS is radically different from switch-like profiles observed in many other alternative exons, including tissue-specific ones. Interestingly, PanAS exons are enriched in genes with different biological functions as switch-like exons, including many genes regulating the process of gene expression at the transcriptional level, in contrast with the effector functions usually related to other alternative exons (see sections 2.3.1 and 2.3.2, and Chapters 3 and 4). Similarly to switch-like exons, PanAS exons show relatively high evolutionary conservation between mammals, which suggests that they have an important biological role. The fact that PanAS are also enriched in frame-preserving exons, and their intermediate inclusion frequencies in ribosome-engaged RNA and single-cell RNA-seq data suggests that their effect is probably mediated by the coexistence at the cellular level of a regulated balance of multiple protein isoforms of the same gene. Taken together, our results suggest that these two types of AS programmes —PanAS exons and switch-like exons— evolved to regulate cellular processes at different levels, in many cases through production of alternative protein isoforms.

In spite of these inferences, the effective contribution of AS to proteomic diversity between tissues and cell types remains a subject of intense debate in the field. Despite the attempts to systematically detect alternative protein isoforms using mass spectrometry (MS), only a minor percentage of AS-regulated transcripts have been proved to exist at the protein level (Abascal et al., 2015; Tress et al., 2017). However, other studies show that a high fraction of AS-regulated exons can be found in ribosome-engaged RNA, especially those in genes with medium and high expression levels (Weatheritt et al., 2016). Individual AS events have been often shown to play a role in a variety of biological processes, either through translation of transcript isoforms with differential inclusion of AS-regulated exons into alternative protein products, or through isoform-specific changes in mRNA stability introduced by inclusion or skipping of these exons (see section 2.2). Besides these effects of individual exons, more comprehensive analyses showed that genes with tissue-specific AS exons have a strong general trend to occupy central positions in protein-protein interaction (PPI) networks (Buljan et al., 2012; Ellis et al., 2012), and a high-throughput co-immunoprecipitation analysis in mouse indicates that the introduction of neural-specific exons has an important effect on the interactomic profiles of the protein products of their genes. Specifically, one-third of the tested exons affected a PPI by at least 2-fold, either disrupting or enhancing it (Ellis et al., 2012). Other large-scale interactomic studies also show important differences in the repertoire of interacting partners of alternative protein isoforms in PPI networks (Corominas et al., 2014; Yang et al., 2016). Collectively, these results indicate a strong effect of AS-regulated exons in rewiring PPI networks in a tissue- and disease-specific manner.

A major hurdle to elucidate this controversy is, to our knowledge, the lack of isoform-specific information in PPI databases, despite the high number of resources listing experimentally determined PPIs for any given gene (Chatr-Aryamontri et al., 2017; Franceschini et al., 2013; Mosca et al., 2013, 2014; Orchard et al., 2014). An accessible resource including systematic annotation of the protein isoforms studied in PPI screens would constitute a major contribution to the field.

Similarly, structural databases such as the Protein Data Bank (PDB) (Mir et al., 2018) rarely reflect isoform diversity introduced by AS, as the number of genes with structural data for multiple isoforms is usually very reduced. In the case of human microexons, we could retrieve PDB protein structures describing inclusion isoforms for 7 out of 301 events, probably due to the highly neural-specific character of these isoforms (see Chapter 3). For a more general set of nearly 30,000 alternative human exons, our pipeline only retrieved reliable PDB structures of inclusion isoforms containing the exon of interest and both flanking constitutive exons for 2,465 events (see Chapter 5). Even in cases where structural characterization of these alternative isoforms has been attempted and structures of inclusion isoforms are deposited in the PDB, AS-regulated exons often map to disordered regions, which means that spatial coordinates for the protein residues encoded by these exons are missing from the structure. As we show in the two aforementioned studies, structural modelling tools can be helpful in addressing these situations. However, care must be taken when interpreting these structures, as the disordered nature of the protein regions of interest means that their structure in solution is inherently variable, and models will only capture a static conformation by definition.

This trend of alternative exons to encode for disordered regions of their protein is well known (Buljan et al., 2012; Ellis et al., 2012; Romero et al., 2006). By no means it must be concluded that the lack of defined structure in these AS-encoded regions implies a lack of function, since disordered regions have a variety of mechanisms to influence biological processes by affecting PPIs (reviewed in Buljan et al., 2013 and Latysheva et al., 2015). Precisely thanks to their increased conformational variability, disordered regions confer greater interaction promiscuity to their proteins than ordered domains (Schreiber and Keating, 2011). PPIs mediated by disordered regions are usually more dynamic and transient than those between structured domains, because of their smaller interface area and reduced affinity (Tompa et al., 2014). Consequently, proteins with disordered regions tend to display two main characteristics. First, they act as hubs in PPI networks, connecting modules of the interactome that would otherwise be very distant (Cumberworth et al., 2013). Second, they are enriched in genes with biological functions where transient interactions are particularly advantageous, such as genes belonging to signalling pathways (Wright and Dyson, 2015). Interestingly, these two features are shared by tissue-specific exons, suggesting that, rather than corresponding to transcriptional noise, these exons may have a very relevant function regulating biological processes mediated by disordered regions in their respective proteins, as previously remarked.

Our own studies indicate an additional effect of AS at the proteomic level through the action of PanAS exons. Although we did not try to systematically detect PanAS-regulated isoforms at the protein level, we did detect intermediate inclusion levels of these exons both in ribosome-engaged RNA and in single-cell RNA-seq samples. Our results suggest that PanAS exons, especially those preserving the reading frame of their transcripts, are very likely to give rise to multiple protein isoforms of the same gene coexisting in individual cells. Given that PanAS exons make up for more than 18% of the protein-coding genes in the human genome, our data contradicts —at least to some extent— conclusions from other researchers suggesting that most genes may express a single protein isoform (Tress et al., 2017), rather indicating that AS has a systemic effect on proteomic diversity.

Our results indicate that genes with tissue-regulated exons preserving the reading frame of their transcripts are often not differentially expressed in the tissues where these exons show differential inclusion. This suggests that AS contributes to cellular and tissue identity as a complementary but largely distinct regulatory layer with respect to transcriptional mechanisms. While AS shows poor evolutionary conservation between species in comparison with transcriptional regulatory mechanisms (Barbosa-Morais et al., 2012; Merkin et al., 2012) several studies show that some tissue-specific AS programmes have higher conservation levels, which supports the importance of their biological role (Barbosa-Morais et al., 2012; Ellis et al., 2012; Han et al., 2013). Our results confirm the high conservation in vertebrates of tissue-specific exons in certain tissues such as neurons, muscle, testis, and pluripotent cells. Moreover, in the latter case, we showed that fundamental regulatory mechanisms of tissue-specific AS programmes can also be conserved in very distant organisms, such as planarians, even if individual regulated events show weaker conservation (see Chapter 4).

Elucidation of the functional impact of microexons at the proteomic level is still far from complete. While global phenotypes for microexon down-regulation are well characterized in a SRRM4-depleted mouse line (Quesnel-Vallières et al., 2015), and reduced levels of SRRM4 expression and microexon inclusion have been detected in ASD patients, only a few microexons have described biological functions (Laurent et al., 2015; Ohnishi et al., 2017; Parras et al., 2018; Quesnel-Vallières et al., 2015; Toffolo et al., 2014; Zibetti et al., 2010).

We showed that microexons display different patterns of inclusion during neuronal differentiation, with most of them peaking in maturing postmitotic neurons (see Chapter 3). However, Inclusion levels of microexons can vary throughout the lifespan of a mature neuron. An AS transition including about 100 microexons have been detected upon neuronal membrane depolarization in culture, as well as during neuronal activation, with decreased inclusion in almost all cases. Down-regulation of these exons is likely mediated by a decrease of SRRM4 observed under the same conditions. This AS programme regulated by neuronal activity is enriched in genes related to synaptic function, especially genes related with cytoskeletal organisation, vesicular trafficking, and membrane reorganisation (Quesnel-Vallières et al., 2016). Therefore, these microexons could potentially contribute to long-term homeostasis of the synaptic response in response to changes in membrane depolarization patterns. Other microexons have been reported to contribute directly to the process of synapse formation. As an example, several microexons in presynaptic proteins of the LAR-RPTP family is determinant in the binding affinity of these proteins for different post-synaptic ligands. Inclusion of microexon B allows binding of LAR-RPTP proteins to SALM3 and ILRAPL1, while only the skipping LAR-RPTP isoforms can bind TrkC (Li et al., 2015; Yoshida et al., 2011; reviewed in Furlanis and Scheiffele, 2018).

SRRM4 is not the only splicing factor involved in microexon regulation, although we show it is probably the most important regulator of these exons in neurons. At the same time, several studies report microexon inclusion in non-neural tissues. Our results show inclusion of some microexons in muscular samples (see Chapter 3). In agreement with this, 159 and 113 microexons show regulatory motifs conserved between human and mouse, corresponding to up-regulation by RBFOX proteins and down-regulation by PTBP1, respectively. The tissues where some of these microexons are included correspond to those where expression of RBFOX proteins has been detected, that is, neurons, heart and muscle, suggesting a function for RBFOX1 in promoting microexon inclusion in muscle (Li et al., 2014b). In addition, up-regulation of SRRM4 has been described in treatment-induced neuroendocrine prostate cancer (t-NEPC), an aggressive and usually metastatic type of tumour sometimes found in prostate adenocarcinoma patients after therapy. Interestingly, an isoform switch of the *SH3GLB1* gene is frequently observed in t-NEPC patients. This switch, also observed in neurons, includes up-regulation of a 24 nt microexon and a 39 nt exon in the t-NEPC-specific isoforms. Inclusion of these two exons changes the function of the protein from pro-apoptotic to anti-apoptotic, suggesting a role for these exons in the poor prognosis of t-NEPC compared to prostate adenocarcinoma (Gan et al., 2018).

Our data suggests the existence of conserved programmes of co-regulated alternative exons in several tissues where the contribution of AS to cellular identity, its effects at the proteomic level, and its role in disease are only starting to be unveiled (see Chapter 5). As an example, inclusion levels of a mammalian-specific exon of 36 nt in the insulin receptor gene (INSR) have been reported to decrease in adipose tissue of obese patients in response to weight loss, without significant variations in global INSR expression levels. Inclusion of this exon is also correlated with basal insulin levels in blood (Kaminska et al., 2014).

The INSR isoform including this exon (INSR-A) predominates in fetal tissues, where it promotes cellular growth by binding to IGFII and to proinsulin (Frasca et al., 1999). In contrast, the skipping isoform (INSR-B), 12 aa shorter, is more abundant in adult differentiated tissues including liver, muscle and adipose tissue (Benecke et al., 1992), where it acts as a receptor specific for insulin with much lower affinity for IGFII. A correlation between the INSR-A:INSR-B ratio and the expression levels of several splicing factors, such as HNRNPA1, SF3A1 and SFRS7 has also been observed in these patients, suggesting a regulatory mechanism for this event (Kaminska et al., 2014). Promotion of exon inclusion by SRSF3, SRSF1 and HNRNPF and skipping induced by CELF1 and HNRNPA1 have also been detected in culture (Sen et al., 2009; Talukdar et al., 2011). Interestingly, insulin levels can regulate phosphorylation and expression levels of several splicing factors affecting AS events in genes of signalling pathways related to of lipid metabolism in liver and muscle (Jiang et al., 2009; Pihlajamäki et al., 2011). This suggests that inclusion levels of this INSR exon might also be regulated by insulin in adipose tissue through changes in expression levels and PTMs of splicing factors triggered by insulin-dependent signalling pathways.

The complexity of the landscape of cross-regulatory relationships between splicing factors, the context-dependent action of many of this proteins and the extensive cross-talk between AS and other mechanisms regulating gene expression in eukaryotes are three of the main factors hindering the elucidation of a universal ‘splicing code’ for each organism (a reliable prediction of the splicing profiles of all AS events in a cell as a function of their intronic and exonic sequences, the expression levels of the different splicing factors, the chromatin state of the cell, and other parameters) (Baralle and Baralle, 2018). While attempts to predict the impact of genetic variants at the AS level have been made for human (Xiong et al., 2015), our knowledge of splicing networks at the tissue and cell type level is still incomplete. Novel AS events, novel regulators, or novel functions for known splicing factors still appear today. Even factors initially expected to be mere effectors of the splicing reaction, such as core spliceosomal snRNAs, have been shown to have a function as AS regulators across different tissues and cell types (Dvinge, 2018; Dvinge et al., 2018). Surely, more research is still needed to discover the importance of AS in specifying cellular identity throughout vertebrate organisms.

With this in mind, we have developed a public web resource with AS event inclusion and gene expression levels derived from more than 300 RNA-seq samples of human, mouse and chicken tissues and cell types, plus additional datasets such as RNA-seq samples from the Genotype-Tissue Expression Project (GTEx) (Melé et al., 2015) (see Chapter 5). These values are shown together with other layers of information, such as the impact of events in protein domains and structure, their evolutionary conservation at the exon level, and the genomic context surrounding each event. Currently, VastDB is visited about 500 times a month by researchers in the splicing field from dozens of countries. In the following years, we will maintain and extend this resource, increasing the number of available species and the amount of information supplied about each gene and AS event. As an example, we are gathering a catalogue of biological functions for AS events supported by published experiments. With collaboration from the VastDB user community, we have included on our website a total of 621 such annotations, a number that we expect to raise in subsequent releases. Our aim is to make VastDB a reference resource in the splicing community, in order to help other researchers to devise new hypotheses and strategies that will set avenues for further studies.

# 7. CONCLUSIONS

1) Neuronal, muscular, testis and pluripotent samples exhibit the strongest AS signatures in human, mouse and chicken, confirming results from previous literature.
2) Among neuronal-specific exons, a programme of more than 200 microexons between 3 and 27 nt in length show the highest conservation in vertebrates, as well as the highest levels of neural specificity.
3) These microexons are enriched in frame-preserving events, and show an increased trend to map inside or near protein domains (without significantly disrupting the fold of their proteins) in comparison with longer neural-specific exons.
4) Our network analysis reveals the existence of a second layer of conserved co-regulated AS programmes in other tissue types, including kidney, liver, adipose tissue and immune cells.
5) Based on their regulatory pattern, we define a third type of alternative exons, the PanAS exons, which are alternatively spliced in more than 80% of our analysed tissues and cell types, presumably contributing to proteomic complexity in processes related with regulation of gene expression.
6) We have developed VastDB, a publicly accessible web repository of AS events integrating their inclusion levels across a catalogue of tissues, cell types and developmental stages in human, mouse and chicken, as well as other features.

# BIBLIOGRAPHY

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.

Adams, P.J., Garcia, E., David, L.S., Mulatz, K.J., Spacey, S.D., and Snutch, T.P. Ca(V)2.1 P/Q-type calcium channel alternative splicing affects the functional impact of familial hemiplegic migraine mutations: implications for calcium channelopathies. Channels (Austin). *3*, 110–121.

Ahsendorf, T., Müller, F.J., Topkar, V., Gunawardena, J., and Eils, R. (2017). Transcription factors, coregulators, and epigenetic marks are linearly correlated and highly redundant. PLoS One *12*, 1–25.

Alon, S., Garrett, S.C., Levanon, E.Y., Olson, S., Graveley, B.R., Rosenthal, J.J.C., and Eisenberg, E. (2015). The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. Elife *2015*, 1–17.

Alonso, C.R., and Wilkins, A.S. (2005). The molecular elements that underlie developmental evolution. Nat. Rev. Genet. *6*, 709–715.

Antequera, F., Tamame, M., Villanueva, J.R., and Santos, T. (1984). DNA methylation in the fungi. J. Biol. Chem. *259*, 8033–8036.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J. Exp. Med. *79*, 137–158.

Ball, P. (2017). Water is an active matrix of life for cell and molecular biology. Proc. Natl. Acad. Sci. *2017*, 201703781.

Bang, M.L., Centner, T., Fornoff, F., Geach, A.J., Gotthardt, M., McNabb, M., Witt, C.C., Labeit, D., Gregorio, C.C., Granzier, H., et al. (2001). The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. Circ. Res. *89*, 1065–1072.

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

Baralle, M., and Baralle, F.E. (2018). The splicing code. BioSystems *164*, 39–48.

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., Çolak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science (80-. ). *338*, 1587–1593.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F.J., Rechavi, G., Li, J.B., Eisenberg, E., et al. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. *24*, 365–376.

Beadle, G.W., and Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. Proc. Natl. Acad. Sci. U. S. A. *27*, 499–506.

Bell, O., Tiwari, V.K., Thomä, N.H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. Nat. Rev. Genet. *12*, 554–564.

Benecke, H., Flier, J.S., and Moller, D.E. (1992). Alternatively spliced variants of the insulin receptor protein. Expression in normal and diabetic human tissues. J. Clin. Invest. *89*, 2066–2070.

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

Beqqali, A., Bollen, I.A.E., Rasmussen, T.B., Van Den Hoogenhof, M.M., Van Deutekom, H.W.M., Schafer, S., Haas, J., Meder, B., Sørensen, K.E., Van Oort, R.J., et al. (2016). A mutation in the glutamate-rich region of RNA-binding motif protein 20 causes dilated cardiomyopathy through missplicing of titin and impaired Frank-Starling mechanism. Cardiovasc. Res. *112*, 452–463.

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

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

Betrán, E., and Long, M. (2002). Expansion of genome coding regions by acquisition of new genes. Genetica *115*, 65–80.

Bird, A.P. (1986). CpG-Rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bland, C.S., Wang, E.T., Vu, A., David, M.P., Castle, J.C., Johnson, J.M., Burge, C.B., and Cooper, T.A. (2010). Global regulation of alternative splicing during myogenic differentiation. Nucleic Acids Res. *38*, 7651–7664.

Blencowe, B.J. (2006). Alternative Splicing: New Insights from Global Analyses. Cell *126*, 37–47.

Bogdanović, O., and Veenstra, G.J.C. (2009). DNA methylation and methyl-CpG binding proteins: Developmental requirements and function. Chromosoma *118*, 549–565.

Bonnal, S., Martínez, C., Förch, P., Bachi, A., Wilm, M., and Valcárcel, J. (2008). RBM5/Luca-15/H37 Regulates Fas Alternative Splice Site Pairing after Exon Definition. Mol. Cell *32*, 81–95.

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

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

Brauch, K.M., Karst, M.L., Herron, K.J., de Andrade, M., Pellikka, P.A., Rodeheffer, R.J., Michels, V. V., and Olson, T.M. (2009). Mutations in Ribonucleic Acid Binding Protein Gene Cause Familial Dilated Cardiomyopathy. J. Am. Coll. Cardiol. *54*, 930–941.

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.

Brow, D.A., and Guthrie, C. (1988). Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. Nature *334*, 213–218.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

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.

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

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.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R., and Zhang (2002). Role of Histone H3 Lysine 27 Methylation in X Inactivation. Science (80-. ). *298*, 1039–1043.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A. V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. *40*, 1416–1425.

Cech, T.R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell *44*, 207–210.

Chang, Y.-F., Imam, J.S., and Wilkinson, M.F. (2007). The Nonsense-Mediated Decay RNA Surveillance Pathway. Annu. Rev. Biochem. *76*, 51–74.

Charizanis, K., Lee, K.Y., Batra, R., Goodwin, M., Zhang, C., Yuan, Y., Shiue, L., Cline, M., Scotti, M.M., Xia, G., et al. (2012). Muscleblind-like 2-Mediated Alternative Splicing in the Developing Brain and Dysregulation in Myotonic Dystrophy. Neuron *75*, 437–450.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. *45*, D369–D379.

Chen, L. (2013). Characterization and comparison of human nuclear and cytosolic editomes. Proc. Natl. Acad. Sci. *110*, E2741–E2747.

Chen, L., Bush, S.J., Tovar-Corona, J.M., Castillo-Morales, A., and Urrutia, A.O. (2014). Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol. Biol. Evol. *31*, 1402–1413.

Cho, K.W.Y. (2012). Enhancers. Wiley Interdiscip. Rev. Dev. Biol. *1*, 469–478.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell *12*, 1–8.

Cieply, B., Park, J.W., Nakauka-Ddamba, A., Bebee, T.W., Guo, Y., Shang, X., Lengner, C.J., Xing, Y., and Carstens, R.P. (2016). Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins. Cell Rep. *15*, 247–255.

Clien, L., Kostadraia, M., Martens, J.H.A., Canu, G., Garcia, S.P., Torro, E., Downes, K., Macaolay, L.C., Bielczyk-Maezynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science (80-. ). *345*.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods *5*, 613–619.

Conaway, R.C., and Conaway, J.W. (2011). Origins and activity of the Mediator complex. Semin. Cell Dev. Biol. *22*, 729–734.

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

Corominas, R., Yang, X., Lin, G.N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S.A., et al. (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat. Commun. *5*, 3650.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., National Institutes of Health, Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., et al. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. *101*, 992–997.

Crick, F.H. (1958). On Protein Synthesis. Symp. Soc. Exp. Biol. *12*, 138–166.

Crick, F.H. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

Cumberworth, A., Lamour, G., Babu, M.M., and Gsponer, J. (2013). Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochem. J. *454*, 361–369.

D’Ambrogio, A., Nagaoka, K., and Richter, J.D. (2013). Translational control of cell growth and malignancy by the CPEBs. Nat. Rev. Cancer *13*, 283–290.

Dabertrand, F., Morel, J.L., Sorrentino, V., Mironneau, J., Mironneau, C., and Macrez, N. (2006). Modulation of calcium signalling by dominant negative splice variant of ryanodine receptor subtype 3 in native smooth muscle cells. Cell Calcium *40*, 11–21.

Daniel, C., Silberberg, G., Behm, M., and Öhman, M. (2014). Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biol. *15*, 1–17.

Darnell, J.E. (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science *202*, 1257–1260.

Dietrich, R., Shukla, G., Fuller, J., and Padgett, R. (2001). Alternative splicing of U12-dependent introns in vivo responds to purine-rich enhancers. RNA *7*, 1378–1388.

Dillman, A.A., Hauser, D.N., Gibbs, J.R., Nalls, M.A., McCoy, M.K., Rudenko, I.N., Galter, D., and Cookson, M.R. (2013). MRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. Nat. Neurosci. *16*, 499–506.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Doolittle, W.F. (1978). Genes in pieces: were they ever together? Nature *272*, 581–582.

Doolittle, W.F., and Stoltzfus, A. (1993). Genes-in-pieces revisited. Nature *361*, 403.

Du, H., Cline, M.S., Osborne, R.J., Tuttle, D.L., Clark, T.A., Donohue, J.P., Hall, M.P., Shiue, L., Swanson, M.S., Thornton, C.A., et al. (2010). Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. Nat. Struct. Mol. Biol. *17*, 187–193.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Dvinge, H. (2018). Regulation of alternative mRNA splicing: old players and new perspectives. FEBS Lett. *592*, 2987–3006.

Dvinge, H., Guenthoer, J., Porter, P.L., and Bradley, R.K. (2018). RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. BioRxiv 1–24.

Eisenberg, E., and Levanon, E.Y. (2018). A-to-I RNA editing - Immune protector and transcriptome diversifier. Nat. Rev.

Genet. *19*, 473–490.

Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: Extent, regulation and function. Nat. Rev. Genet. *14*, 496–506.

Ellis, J.D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T.H., 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.

Feng, S., Jacobsen, S.E., and Reik, W. (2010). Epigenetic reprogramming in plant and animal development. Science (80-. ). *330*, 622–627.

Fernández, C.M., Moltó, E., Gallardo, N., del Arco, A., Martínez, C., Andrés, A., Ros, M., Carrascosa, J.M., and Arribas, C. (2009). The expression of rat resistin isoforms is differentially regulated in visceral adipose tissues: effects of aging and food restriction. Metabolism. *58*, 204–211.

Fogel, B.L., Wexler, E., Wahnich, A., Friedrich, T., Vijayendran, C., Gao, F., Parikshak, N., Konopka, G., and Geschwind, D.H. (2012). RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. Hum. Mol. Genet. *21*, 4171–4186.

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. *102*, 16176–16181.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C., et al. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*, D808-15.

Franklin, R.E., and Gosling, R.G. (1953b). Evidence for 2-chain Helix in crystalline structure of sodium deoxyribonucleate. Nature *172*, 156–157.

Franklin, R.E., and Gosling, R.G. (1953a). Molecular configuration in sodium thymonucleate. Nature *171*, 740–741.

Frasca, F., Pandini, G., Scalia, P., Sciacca, L., Mineo, R., Costantino, A., Goldfine, I.D., Belfiore, A., and Vigneri, R.

(1999). Insulin Receptor Isoform A, a Newly Recognized, High-Affinity Insulin-Like Growth Factor II Receptor in Fetal and Cancer Cells. Mol. Cell. Biol. *19*, 3278–3288.

Frilander, M.J., and Steitz, J.A. (1999). Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. Genes Dev. *13*, 851–863.

Frilander, M.J., and Steitz, J.A. (2001). Dynamic exchanges of RNA interactions leading to catalytic core formation in the U12-dependent spliceosome. Mol. Cell *7*, 217–226.

Fu, X.D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Furlanis, E., and Scheiffele, P. (2018). Regulation of Neuronal Differentiation, Function, and Plasticity by Alternative Splicing. Annu. Rev. Cell Dev. Biol. 1–19.

Furuichi, Y. (2015). Discovery of m 7 G-cap in eukaryotic mRNAs. Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci. *91*, 394–409.

Gallego-Paez, L.M., Bordone, M.C., Leote, A.C., Saraiva-Agostinho, N., Ascensão-Ferreira, M., and Barbosa-Morais, N.L. (2017). Alternative splicing : the pledge, the turn, and the prestige The key role of alternative splicing in human biological systems. Hum. Genet.

Gan, Y., Li, Y., Long, Z., Lee, A.R., Xie, N., Lovnicki, J.M., Tang, Y., Chen, X., Huang, J., and Dong, X. (2018). Roles of Alternative RNA Splicing of the Bif-1 Gene by SRRM4 During the Development of Treatment-induced Neuroendocrine Prostate Cancer. EBioMedicine *31*, 267–275.

George, C.H., Rogers, S.A., Bertrand, B.M.A., Tunwell, R.E.A., Thomas, N.L., Steele, D.S., Cox, E. V., Pepper, C., Hazeel, C.J., Claycomb, W.C., et al. (2007). Alternative splicing of ryanodine receptors modulates cardiomyocyte Ca2+ signaling and susceptibility to apoptosis. Circ. Res. *100*, 874–883.

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

Giudice, J., and Cooper, T.A. (2014). RNA-Binding Proteins in Heart Development. In Advances in Experimental Medicine and Biology, pp. 389–429.

Giudice, J., Xia, Z., Wang, E.T., Scavuzzo, M.A., Ward, A.J., Kalsotra, A., Wang, W., Wehrens, X.H.T., Burge, C.B., Li, W., et al. (2014). Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. Nat. Commun. *5*, 3603.

Giudice, J., Loehr, J.A.A., Rodney, G.G.G., and Cooper, T.A.A. (2016). Alternative Splicing of Four Trafficking Genes Regulates Myofiber Structure and Skeletal Muscle Physiology. Cell Rep. *17*, 1923–1933.

Goo, Y.H., and Cooper, T.A. (2009). CUGBP2 directly interacts with U2 17S snRNP components and promotes U2 snRNA binding to cardiac troponin T pre-mRNA. Nucleic Acids Res. *37*, 4275–7286.

Gross, D.S., and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. *57*, 159–197.

Guhaniyogi, J., and Brewer, G. (2001). Regulation of mRNA stability in mammalian cells. Gene *265*, 11–23.

Guo, W., Schafer, S., Greaser, M.L., Radke, M.H., Liss, M., Govindarajan, T., Maatz, H., Schulz, H., Li, S., Parrish, A.M., et al. (2012). RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. Nat. Med. *18*, 766–773.

Hahn, M.W., and Wray, G.A. (2002). The g-value paradox. Evol. Dev. *4*, 73–75.

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

Han, H., Irimia, M., Ross, P.J., Sung, H.-K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., et al. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. Nature *498*, 241–245.

Hargan-Calvopina, J., Taylor, S., Cook, H., Hu, Z., Lee, S.A., Yen, M.R., Chiang, Y.S., Chen, P.Y., and Clark, A.T. (2016). Stage-Specific Demethylation in Primordial Germ Cells Safeguards against Precocious Differentiation. Dev.

Cell *39*, 75–86.

Hastings, M.L., Wilson, C.M., and Munroe, S.H. (2001). A purine-rich intronic element enhances alternative splicing of thyroid hormone receptor mRNA. RNA *7*, 859–874.

van der Heijden, T., van Vugt, J.J.F.A., Logie, C., and van Noort, J. (2012). Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. Proc. Natl. Acad. Sci. U. S. A. *109*, E2514-22.

Hendrich, B., and Tweedie, S. (2003). The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet. *19*, 269–277.

Ver Heyen, M., Heymans, S., Antoons, G., Reed, T., Periasamy, M., Awede, B., Lebacq, J., Vangheluwe, P., Dewerchin, M., Collen, D., et al. (2001). Replacement of the muscle-specific sarcoplasmic reticulum Ca2+-ATPase isoform SERCA2a by the nonmuscle SERCA2b homologue causes mild concentric hypertrophy and impairs contraction-relaxation of the heart. Circ. Res. *89*, 838–846.

Higuchi, M., Maas, S., Single, F.N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P.H. (2000). Point mutation in an AMPA receptor gene recues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature *406*, 1998–2001.

Hoffman, B.E., and Grabowski, P.J. (1992). U1 snRNP targets an essential splicing factor, U2AF65 to the 3’ splice sit by a network of interactions spanning the exon. Genes Dev. *6*, 2554–2568.

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

Huppertz, I., Attig, J., D’Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: Protein-RNA interactions at nucleotide resolution. Methods *65*, 274–287.

Irimia, M., and Roy, S.W. (2008). Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet. *4*.

Irimia, M., and Roy, S.W. (2014). Origin of spliceosomal introns and alternative splicing. Cold Spring Harb. Perspect. Biol. *6*.

Irimia, M., Penny, D., and Roy, S.W. (2007). Coevolution of genomic intron number and splice sites. Trends Genet. *23*, 321–325.

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

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jaillon, O., Bouhouche, K., Gout, J.F., Aury, J.M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B.M., Ségurens, B., et al. (2008). Translational control of intron splicing in eukaryotes. Nature *451*, 359–362.

Jangi, M., and Sharp, P.A. (2014). Building robust transcriptomes with master splicing factors. Cell *159*, 487–498.

Jiang, K., Patel, N.A., Watson, J.E., Apostolatos, H., Kleiman, E., Hanson, O., Hagiwara, M., and Cooper, D.R. (2009). Akt2 regulation of Cdc2-like kinases (Clk/Sty), serine/arginine-rich (SR) protein phosphorylation, and insulin-induced alternative splicing of PKCβJII messenger ribonucleic acid. Endocrinology *150*, 2087–2097.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., Van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kalsotra, A., Xiao, X., Ward, A.J., Castle, J.C., Johnson, J.M., Burge, C.B., and Cooper, T.A. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. Proc. Natl. Acad. Sci. *105*, 20333–20338.

Kaminska, D., Hämäläinen, M., Cederberg, H., Käkelä, P., Venesmaa, S., Miettinen, P., Ilves, I., Herzig, K.H., Kolehmainen, M., Karhunen, L., et al. (2014). Adipose tissue INSR splicing in humans associates with fasting insulin level and is regulated by weight loss. Diabetologia *57*, 347–351.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. Gene *514*, 1–30.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kim, N., Alekseyenko, A. V., Roy, M., and Lee, C. (2007). The ASAP II database: Analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res. *35*, 93–98.

Konkel, D.A., Tilghman, S.M., and Leder, P. (1978). The sequence of the chromosomal mouse β-globin major gene: Homologies in capping, splicing and poly(A) sites. Cell *15*, 1125–1132.

Koonin, E. V. (2006). The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate? Biol. Direct *1*, 1–23.

Kurosaki, T., and Maquat, L.E. (2016). Nonsense-mediated mRNA decay in humans at a glance. J. Cell Sci. *129*, 461–467.

Kuroyanagi, H. (2009). Fox-1 family of RNA-binding proteins. Cell. Mol. Life Sci. *66*, 3895–3907.

Kwak, S., and Kawahara, Y. (2005). Deficient RNA editing of GluR2 and neuronal death in amyotropic lateral sclerosis. J. Mol. Med. *83*, 110–120.

Labeit, S., and Kolmerer, B. (1995). Titins: Giant proteins in charge of muscle ultrastructure and elasticity. Science (80- . ). *270*, 293–296.

Lai, X., Verhage, L., Hugouvieux, V., and Zubieta, C. (2018). Pioneer Factors in Animals and Plants—Colonizing Chromatin for Gene Regulation. Molecules *23*, 1914.

Lambowitz, A.M., and Zimmerly, S. (2004). Mobile Group II Introns. Annu. Rev. Genet. *38*, 1–35.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Larsen, B.B., Miller, E.C., Rhodes, M.K., and Wiens, J.J. (2017). Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life. Q. Rev. Biol. *92*, 229–265.

Latysheva, N.S., Flock, T., Weatheritt, R.J., Chavali, S., and Babu, M.M. (2015). How do disordered regions achieve comparable functions to structured domains? Protein Sci. *24*, 909–922.

Laurent, B., Ruitu, L., Murn, J., Hempel, K., Ferrao, R., Xiang, Y., Liu, S., Garcia, B.A., Wu, H., Wu, F., et al. (2015). A Specific LSD1/KDM1A Isoform Regulates Neuronal Differentiation through H3K9 Demethylation. Mol. Cell 1–14.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Leder, A., Miller, H.I., Hamer, D.H., Seidman, J.G., Norman, B., Sullivan, M., and Leder, P. (1978). Comparison of cloned mouse alpha-and beta-globin genes: conservation of intervening sequence locations and extragenic homology. Proc. … *75*, 6187–6191.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

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.

Lee, J.A., Damianov, A., Lin, C.H., Fontes, M., Parikshak, N.N., Anderson, E.S., Geschwind, D.H., Black, D.L., and Martin, K.C. (2016). Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. Neuron *89*, 113–128.

Lejeune, E., and Allshire, R.C. (2011). Common ground: Small RNA programming and chromatin modifications.

Curr. Opin. Cell Biol. *23*, 258–265.

Lengyel, P., Speyer, J.F., and Ochoa, S. (1961). Synthetic polynucleotides and the amino acid code. Proc. Natl. Acad. Sci. U. S. A. *47*, 1936–1942.

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

Li, D., Jin, C., Yin, C., Zhang, Y., Pang, B., Tian, L., Han, W., Ma, D., and Wang, Y. (2007). An alternative splice form of CMTM8 induces apoptosis. Int. J. Biochem. Cell Biol. *39*, 2107–2119.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Li, S., Guo, W., Dewey, C.N., and Greaser, M.L. (2013). Rbm20 regulates titin alternative splicing as a splicing repressor. Nucleic Acids Res. *41*, 2659–2672.

Li, X., Song, J., and Yi, C. (2014a). Genome-wide Mapping of Cellular Protein-RNA Interactions Enabled by Chemical Crosslinking. Genomics, Proteomics Bioinforma. *12*, 72–78.

Li, Y., Zhang, P., Choi, T.Y., Park, S.K., Park, H., Lee, E.J., Lee, D., Roh, J.D., Mah, W., Kim, R., et al. (2015). Splicing-Dependent Trans-synaptic SALM3-LAR-RPTP Interactions Regulate Excitatory Synapse Development and Locomotion. Cell Rep. *12*, 1618–1630.

Li, Y.I., Sanchez-Pulido, L., Haerty, W., and Ponting, C.P. (2014b). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. *25*, 1–13.

Licatalosi, D.D., and Darnell, R.B. (2006). Splicing Regulation in Neurologic Disease. Neuron *52*, 93–101.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

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

Linares, A.J., Lin, C.-H., Damianov, A., Adams, K.L., Novitch, B.G., Black, D.L., Adams, K., Rousso, D., Umbach, J., Novitch, B., et al. (2015). The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. Elife *4*, e09268.

Liscovitch-Brauer, N., Alon, S., Porath, H.T., Elstein, B., Unger, R., Ziv, T., Admon, A., Levanon, E.Y., Rosenthal, J.J.C., and Eisenberg, E. (2017). Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods. Cell *169*, 191–202.e11.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lucchesi, J.C., and Kuroda, M.I. (2015). Dosage compensation in drosophila. Cold Spring Harb. Perspect. Biol. *7*, 1–21.

Luo, H.R., Moreau, G.A., Levin, N., and Moore, M.J. (1999). The human Prp8 protein is a component of both U2- and U12-dependent spliceosomes. Rna *5*, 893–908.

Lyko, F., Ramsahoye, B.H., and Jaenisch, R. (2000). DNA methylation in Drosophila melanogaster. Nature *408*, 538–540.

Makeyev, E. V., Zhang, J., Carrasco, M.A., and Maniatis, T. (2007). The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. Mol. Cell *27*, 435–448.

Martin, W., and Koonin, E. V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. Nature *440*, 41–45.

Martin, W.F., Garg, S., and Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. Philos. Trans. R. Soc. B Biol. Sci. *370*.

Maxam, a M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A. *74*, 560–564.

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

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). The human transcriptome across tissues and individuals. Science (80-. ). *348*, 660–665.

Mereschkowsky, C. (1905). Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol. Zent. Bl. *25*, 593–604.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science (80-. ). *338*, 1593–1599.

Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. BioEssays *35*, 1050–1055.

Mir, S., Alhroub, Y., Anyango, S., Armstrong, D.R., Berrisford, J.M., Clark, A.R., Conroy, M.J., Dana, J.M., Deshpande, M., Gupta, D., et al. (2018). PDBe: Towards reusable data delivery infrastructure at protein data bank in Europe. Nucleic Acids Res. *46*, D486–D492.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. Nat. Methods *10*, 47–53.

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: A catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. *42*, D374-9.

Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. (2002). Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell *111*, 197–208.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72–77.

Niklas, K.J., Bondos, S.E., Dunker, A.K., Newman, S.A., Niklas, K.J., Bondos, S.E., Dunker, A.K., and Newman, S.A. (2015). Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. Front. Cell Dev. Biol. *3*, 1–13.

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

Nirenberg, M.W., and Matthaei, J.H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. U. S. A. *47*, 1588–1602.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., and O'Neal, C. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc. Natl. Acad. Sci. *53*, 1161–1168.

Nishikura, K. (2010). Functions and Regulation of RNA Editing by ADAR Deaminases. Annu. Rev. Biochem. *79*, 321–349.

Nishimura, S., Jacob, T.M., and Khorana, H.G. (1964). Synthetic Deoxyribopolynucleotides As Templates for Ribonucleic Acid Polymerase: the Formation and Characterization of a Ribopolynucleotide With a Repeating Trinucleotide Sequence. Proc. Natl. Acad. Sci. U. S. A. *52*, 1494–1501.

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.

Ohta, S., Nishida, E., Yamanaka, S., and Yamamoto, T. (2013). Global Splicing Pattern Reversion during Somatic Cell Reprogramming. Cell Rep. *5*, 357–366.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N., et al. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358-63.

Otake, L.R., Scamborova, P., Hashimoto, C., and Steitz, J.A. (2002). The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in

Drosophila. Mol. Cell *9*, 439–446.

Palavicini, J.P., Correa-Rojas, R.A., and Rosenthal, J.J.C. (2012). Extra double-stranded RNA binding domain (dsRBD) in a squid RNA editing enzyme confers resistance to high salt environment. J. Biol. Chem. *287*, 17754–17764.

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.

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., McCarthy, M., and Steitz, J.A. (2002). The splicing of U12-type introns can be a rate-limiting step in gene expression. EMBO J. *21*, 3804–3815.

Payankaulam, S., Li, L.M., and Arnosti, D.N. (2010). Transcriptional repression: Conserved and evolved features. Curr. Biol. *20*, 764–771.

Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, H.T., Danan-Gotthold, M., Knisbacher, B.A., Eisenberg, E., and Levanon, E.Y. (2015). Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. Cell Rep. *13*, 267–276.

Paz, N., and Levanon, E. (2007). Altered adenosine-to-inosine RNA editing in human cancer. Genome … 1586–1595.

Pihlajamäki, J., Lerin, C., Itkonen, P., Boes, T., Floss, T., Schroeder, J., Dearie, F., Crunkhorn, S., Burak, F., Jimenez-Chillaron, J.C., et al. (2011). Expression of the splicing factor gene SFRS10 is reduced in human obesity and contributes to enhanced lipogenesis. Cell Metab. *14*, 208–218.

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E. V., Borchers, C.H., Baumeister, W., Herzog, F., et al. (2015). Architecture of the RNA polymerase II-Mediator core initiation complex.

Nature *518*, 376–380.

Popp, M.W., and Maquat, L.E. (2016). Leveraging rules of nonsense-mediated mRNA decay for genome engineering and personalized medicine. Cell *165*, 1319–1332.

Proffitt, J.H., Davie, J.R., Swinton, D., and Hattman, S. (1984). 5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA. Mol. Cell. Biol. *4*, 985–988.

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.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, a P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. U. S. A. *97*, 5237–5242.

Reed, R. (2000). Mechanisms of fidelity in pre-mRNA splicing. Curr. Opin. Cell Biol. *12*, 340–345.

Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., et al. (2012). Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type i interferon signature. Nat. Genet. *44*, 1243–1248.

Robart, A.R., and Zimmerly, S. (2005). Group II intron retroelements: Function and diversity. Cytogenet. Genome Res. *110*, 589–597.

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.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., et al. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc. Natl. Acad. Sci. *103*, 8390–8395.

Rosenthal, J.J.C., and Seeburg, P.H. (2012). A-to-I RNA Editing: Effects on Proteins Key to Neural Excitability. Neuron *74*, 432–439.

Ruggiu, M., Herbst, R., Kim, N., Jevsek, M., Fak, J.J., Mann, M.A., Fischbach, G., Burden, S.J., and Darnell, R.B. (2009). Rescuing Z+ agrin splicing in Nova null mice restores synapse formation and unmasks a physiologic defect in motor neuron firing. Proc. Natl. Acad. Sci. U. S. A. *106*, 3513–3518.

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.

Sagan, L. (1967). On the origin of mitosing cells. J. Theor. Biol. *14*, 255–274.

Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 129–143.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. *74*, 5463–5467.

Schlesinger, F., Tammena, D., Krampfl, K., and Bufler, J. (2005). Desensitization and resensitization are independently regulated in human recombinant GluR subunit coassemblies. Synapse *55*, 176–182.

Schmauss, C. (2005). Regulation of Serotonin 2C Receptor Pre-mRNA Editing by Serotonin. Int. Rev. Neurobiol. *63*, 83–100.

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

Schneider, C., Will, C.L., Makarova, O. V, Makarov, E.M., and Luhrmann, R. (2002). Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. Mol Cell Biol *22*, 3219–3229.

Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D., and Jenuwein, T. (2004). A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev. *18*, 1251–1262.

Schreiber, G., and Keating, A.E. (2011). Protein binding specificity versus promiscuity. Curr. Opin. Struct. Biol. *21*, 50–61.

Schwann, S., and Schleiden, H. (1839). Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachstum der Pflanzen und Thiere.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. *18*, 88–103.

Sebastian, S., Faralli, H., Yao, Z., Rakopoulos, P., Palii, C., Cao, Y., Singh, K., Liu, Q.C., Chu, A., Aziz, A., et al. (2013). Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. Genes Dev. *27*, 1247–1259.

Sen, S., Talukdar, I., and Webster, N.J.G. (2009). SRp20 and CUG-BP1 Modulate Insulin Receptor Exon 11 Alternative Splicing. Mol. Cell. Biol. *29*, 871–880.

Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C., and Black, D.L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat. Struct. Mol. Biol. *15*, 183–191.

Shatkin, A.J. (1976). Capping of eucaryotic mRNAs. Cell *9*, 645–653.

Shen, Y., Yue, F., Mc Cleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Ren, B., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Shibayama, M., Ohno, S., Osaka, T., Sakamoto, R., Tokunaga, A., Nakatake, Y., Sato, M., and Yoshida, N. (2009). Polypyrimidine tract-binding protein is essential for early mouse development and embryonic stem cell proliferation. FEBS J. *276*, 6658–6668.

Shuman, S. (2015). RNA capping: Progress and prospects. Rna *21*, 735–737.

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

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

Spellman, R., Llorian, M., and Smith, C.W.J. (2007). Crossregulation and Functional Redundancy between the Splicing Regulator PTB and Its Paralogs nPTB and ROD1. Mol. Cell *27*, 420–434.

Stevens, S.W., Ryan, D.E., Ge, H.Y., Moore, R.E., Young, M.K., Lee, T.D., and Abelson, J. (2002). Composition and functional characterization of the yeast spliceosomal penta-snRNP. Mol. Cell *9*, 31–44.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell *98*, 1–4.

Suckale, J., Wendling, O., Masjkur, J., Jäger, M., Münster, C., Anastassiadis, K., Stewart, A.F., and Solimena, M. (2011). PTBP1 is required for embryonic development before gastrulation. PLoS One *6*.

Talukdar, I., Sen, S., Urbano, R., Thompson, J., Yates, J.R., and Webster, N.J.G. (2011). HnRNP A1 and hnRNP F modulate the alternative splicing of exon 11 of the insulin receptor gene. PLoS One *6*.

Tanaka, Y., Hysolli, E., Su, J., Xiang, Y., Kim, K.Y., Zhong, M., Li, Y., Heydari, K., Euskirchen, G., Snyder, M.P., et al. (2015). Transcriptome Signature and Regulation in Human Somatic Cell Reprogramming. Stem Cell Reports *4*, 1125–1139.

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.

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.

Tilghman, S.M., Tiemeier, D.C., Seidman, J.G., Peterlin, B.M., and Sullivan, M. (1978). Intervening sequence of DNA identified in the structural portion of a mouse B-globin gene. Proc. Natl. Acad. Sci. U. S. A. *75*, 725–729.

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

Toffolo, E., Rusconi, F., Paganini, L., Tortorici, M., Pilotto, S., Heise, C., Verpelli, C., Tedeschi, G., Maffioli, E., Sala, C., et al. (2014). Phosphorylation of neuronal Lysine-Specific Demethylase 1LSD1/KDM1A impairs transcriptional repression by regulating interaction with CoREST and histone deacetylases HDAC1/2. J. Neurochem. *128*, 603–616.

Toh, C.X.D., Chan, J.W., Chong, Z.S., Wang, H.F., Guo, H.C., Satapathy, S., Ma, D., Goh, G.Y.L., Khattar, E., Yang, L., et al. (2016). RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. Cell Rep. *15*, 2597–2607.

Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A Million peptide motifs for the molecular biologist. Mol. Cell *55*, 161–169.

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

Tweedie, S., Charlton, J., Clark, V., and Bird, A. (1997). Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol. Cell. Biol. *17*, 1469–1475.

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

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

Valcárcel, J., Singh, R., Zamore, P.D., and Green, M.R. (1993). The protein Sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA. Nature *362*, 171–175.

Vannini, A., and Cramer, P. (2012). Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries. Mol. Cell *45*, 439–446.

Venables, J.P., Lapasset, L., Gadea, G., Fort, P., Klinck, R., Irimia, M., Vignal, E., Thibault, P., Prinos, P., Chabot, B., et al. (2013). MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. Nat. Commun. *4*, 1–10.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science (80-. ). *291*, 1304–1351.

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*, 380–386.

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

Vuong, J.K., Lin, C.-H., Zhang, M., Chen, L., Black, D.L., and Zheng, S. (2016b). PTBP1 and PTBP2 Serve Both Specific and Redundant Functions in Neuronal Pre-mRNA Splicing. Cell Rep. *17*, 2766–2775.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718.

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, E.T., Cody, N.A.L., Jog, S., Biancolella, M., Wang, T.T., Treacy, D.J., Luo, S., Schroth, G.P., Housman, D.E., Reddy, S., et al. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell *150*, 710–724.

Wang, E.T., Ward, A.J., Cherone, J.M., Giudice, J., Wang, T.T., Treacy, D.J., Lambert, N.J., Freese, P., Saxena, T., Cooper, T.A., et al. (2015). Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. Genome Res. *25*, 858–871.

Warf, M.B., Diegel, J. V., von Hippel, P.H., and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proc. Natl. Acad. Sci. *106*, 9203–9208.

Wassarman, K.M., and Steitz, J. a (1992). The low-abundance U11 and U12 small nuclear ribonucleoproteins (snRNPs) interact to form a two-snRNP complex. Mol. Cell. Biol. *12*, 1276–1285.

Watson, J., and Crick, F. (1953). Molecular structure of nucleic acids. Nature *171*, 737–738.

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

Wilkins, M.H.F., Stokes, A.R., and Wilson, H.R. (1953). Molecular structure of deoxypentose nucleic acids. Nature *171*, 738–740.

Will, C.L., Schneider, C., Reed, R., and Lührmann, R. (1999). Identification of both shared and distinct proteins in the major and minor spliceosomes. Science (80-. ). *284*, 2003–2005.

Witten, J.T., and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. Trends Genet. *27*, 89–97.

Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. *16*, 18–29.

Wu, C. (1980). The 5’ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature *286*, 854–860.

Wu, Q., and Krainer, A.R. (1996). U1-mediated exon definition interactions between AT-AC and GT-AG introns. Science (80-. ). *274*, 1005–1008.

Wu, J.Q., Habegger, L., Noisa, P., Szekely, A., Qiu, C., Hutchison, S., Raha, D., Egholm, M., Lin, H., Weissman, S., et al. (2010). Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. Proc. Natl. Acad. Sci. *107*, 5254–5259.

Xin, B., and Rohs, R. (2018). Relationship between histone modifications and transcription factor binding is protein family specific. Cold Spring Harb. Lab. Press January *16*, 1–13.

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). The human splicing code reveals new insights into the genetic determinants of disease. Science (80-. ). *347*, 1254806.

Xue, Y., Ouyang, K., Huang, J., Zhou, Y., Ouyang, H., Li, H., Wang, G., Wu, Q., Wei, C., Bi, Y., et al. (2013). Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated MicroRNA circuits. Cell *152*, 82–96.

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.

Yano, M., Hayakawa-Yano, Y., Mele, A., and Darnell, R.B. (2010). Nova2 Regulates Neuronal Migration through an RNA Switch in Disabled-1 Signaling. Neuron *66*, 848–858.

Yoshida, T., Yasumura, M., Uemura, T., Lee, S.-J., Ra, M., Taguchi, R., Iwakura, Y., and Mishina, M. (2011). IL-1 Receptor Accessory Protein-Like 1 Associated with Mental Retardation and Autism Mediates Synapse Formation by Trans-Synaptic Interaction with Protein Tyrosine Phosphatase . J. Neurosci. *31*, 13485–13499.

Zaret, K.S., and Mango, S.E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev. *37*, 76–81.

Zaret, K.S., Lerner, J., and Iwafuchi-Doi, M. (2016). Chromatin Scanning by Dynamic Binding of Pioneer Factors. Mol. Cell *62*, 665–667.

Zibetti, C., Adamo, A., Binda, C., Forneris, F., Toffolo, E., Verpelli, C., Ginelli, E., Mattevi, A., Sala, C., and Battaglioli, E. (2010). Alternative Splicing of the Histone Demethylase LSD1/KDM1 Contributes to the Modulation of Neurite Morphogenesis in the Mammalian Nervous System. J. Neurosci. *30*, 2521–2532.