Alternative splicing during neuronal development

Mariana Ascensão Ferreira

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors:

Dr. Nuno Luís Barbosa Morais
Prof. Gabriel António Amaro Monteiro

Examination Committee:

Chairperson: Prof. João Pedro Estrela Rodrigues Conde
Supervisor: Dr. Nuno Luís Barbosa Morais
Member of the Committee: Prof. Maria Margarida Fonseca Rodrigues Diogo

October 2016
Acknowledgments

First, I would like to express gratitude to my supervisor, Dr. Nuno Barbosa Morais, for his expertise, enthusiasm and acuity during all the discussions that shaped this work. My stay in the Lab, under his supervision, has largely enriched my academic experience and strongly reinforced my ambition to become a scientist in this field. Also, I am thankful to all the Computational Biology Lab, at IMM - FMUL, for providing me the great opportunity to learn through a summer project, in which my ever growing enthusiasm for science became a solid motivation for the future. I would also want to thank Professor Gabriel Monteiro, my supervisor from IST, for his very valuable comments on this thesis and for the assistance related to bureaucracy issues.

I would also like to thank the experts from whom I had teeming insights that were fundamental for the ongoing of the project, namely Dr. Manuel Irimia, during his visit to the Lab, and Dr. Domingos Henrique, from IMM. Also, I want to thank Dr. Ana Pombo and Dr. Sérgio Almeida for sharing some of the resources used in this work.

I am deeply thankful to all my colleagues from NMorais Lab at IMM: Bárbara Caravela, Bernardo Almeida, Carolina Leote, Juan Carlos Verjan, Lina Paez, Marie Bordone, Nuno Agostinho and Teresa Maia for the helpful and cheerful contribute to all my doubts, since the beginning of my stay in the Lab. Also, I very much appreciate your patience to explain important biological details that bypass the technical-formatting reasoning I brought from my previous experiences. You are all part of my growing enthusiasm about alternative splicing and the development of nervous systems.

I want to thank my fellow classmates of Biomedical Engineering at IST, for going along with me during this wonderful yet winding path that was my stay at IST, in which we unfailing managed to encourage and cheer up each other.

I am thankful to all my friends and family for the support during this journey. I want to express my deepest gratitude to Bernardo Calado, who has been an inexhaustible source of self-confidence, motivation and belief, whenever my mood did not match my passion for this work. Last but not least, I want to thank my parents, for transferring to me the love for learning and for doing their best to help me grow up matching academics, sports, books and everything I have been passionate about. Finally, I want to thank my sister, for her insanely faith in overestimating my skills, which I frequently find myself disbelieving.
Abstract

Alternative splicing (AS) is one of the mechanisms expanding transcriptome diversity when comparing organisms with similar number of genes and is known to occur in 95% of human multi-exon genes. Remarkably, AS is more frequent and more conserved, across vertebrate species, in nervous tissues, and there are relatively more neuronal-specific RNA-binding proteins, suggesting that neuronal-specific regulated alternative splicing may be one major contributor to functional complexity of neurodevelopmental processes. Importantly, links between alternative splicing misregulations and neurodevelopmental disorders have already been established.

Widespread usage of high-throughput sequencing techniques allowed the current availability of transcriptomic data from previous studies, which combined with a bioinformatics approach supported by statistics potentiates the accurate quantification of gene expression and exon inclusion levels. Through gene expression and alternative splicing quantification, the patterns of exon inclusion during neuronal differentiation enabled the identification of AS events changing as cells commit to the neuronal fate, as well as characterization of a neural progenitor cell line widely used to model neurogenesis in vitro (NPC 46C).

The NOVA1 RNA-binding protein (RBP) is known to promote exon skipping upon binding to the last nucleotides of the upstream intron. Interestingly, neuronal function AS events with switch-like changes during differentiation showed inclusion levels significantly anti-correlated with the expression of the Nova1 gene. Moreover, similar association between nSR100 RBP and neuronal-required exons is observed. These results provide a relevant contribution for profiling the role of alternative splicing in fine-tuning neuronal function.

Keywords: Alternative splicing, neuronal development, transcriptomics, neuroprogenitor cells
Resumo

O splicing alternativo (SA) é um dos mecanismos responsáveis pelo aumento no número de transcritos em organismos com número semelhante de genes, ocorrendo em 95% dos genes humanos multi-exónicos. Em vertebrados, o SA é mais frequente e conservado em tecidos neuronais, além de existirem mais proteínas de ligação ao RNA específicas nestes tecidos, sugerindo que é um dos mecanismos que contribui para a maior complexidade funcional do tecido nervoso. Além disso, já foram estabelecidas associações entre desregulações no SA e doenças de neuro-desenvolvimento.

A sequenciação high-throughput tem vindo a contribuir para a disponibilização pública de dados transcriptómicos, que, através de análises bioinformáticas assentes em princípios de estatística, potenciam a quantificação de níveis de expressão gênica e de inclusão de exões. Usando este tipo de quantificações, a evolução dos níveis de inclusão de exões durante a diferenciação neuronal permitiu identificar eventos que transitam à medida que as células entram no fenótipo neuronal, facilitando ainda a caracterização de uma linha celular de células neuro-progenitoras, comum em estudos de neurogénese in vitro.

A proteína NOVA1 promove a exclusão de exões através da sua ligação ao intrão a montante e a sua expressão revelou-se significativamente anti-correlacionada com os níveis de inclusão progressivamente maiores de exões com funções neuronais. O mesmo tipo de associação foi identificado para uma outra proteína de ligação ao RNA, nSR100. Estes resultados permitem ajudar a elucidar as funções do splicing na função neuronal.

Palavras-chave: Splicing alternativo, desenvolvimento neuronal, transcriptómica, células neuroprogenitoras
Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>Resumo</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>List of Acronyms and Symbols</td>
<td>xiv</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Molecular biology of splicing</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Genomics, transcriptomics and beyond</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Alternative splicing and species’ complexity</td>
<td>9</td>
</tr>
<tr>
<td>1.1.3 Alternative splicing in the brain</td>
<td>11</td>
</tr>
<tr>
<td>1.2 Neuronal development</td>
<td>15</td>
</tr>
<tr>
<td>1.2.1 Neural differentiation in vitro</td>
<td>17</td>
</tr>
<tr>
<td>1.2.2 Alternative splicing misregulations in neurodevelopmental disorders</td>
<td>18</td>
</tr>
<tr>
<td>1.3 Motivation and goals</td>
<td>20</td>
</tr>
<tr>
<td>1.4 Document structure</td>
<td>21</td>
</tr>
<tr>
<td>2 Materials and Methods</td>
<td>23</td>
</tr>
<tr>
<td>2.1 RNA-Sequencing data</td>
<td>23</td>
</tr>
<tr>
<td>2.1.1 Gene expression quantification</td>
<td>27</td>
</tr>
<tr>
<td>2.1.2 Splicing quantification</td>
<td>28</td>
</tr>
<tr>
<td>2.2 Datasets</td>
<td>29</td>
</tr>
<tr>
<td>2.2.1 Human datasets</td>
<td>30</td>
</tr>
<tr>
<td>2.2.2 Mouse datasets</td>
<td>30</td>
</tr>
<tr>
<td>2.2.3 MultiTissue dataset</td>
<td>31</td>
</tr>
<tr>
<td>2.3 Statistical analyses</td>
<td>31</td>
</tr>
<tr>
<td>2.3.1 Principal component analysis</td>
<td>32</td>
</tr>
<tr>
<td>2.3.2 Filtering</td>
<td>33</td>
</tr>
<tr>
<td>2.3.3 Statistical tests</td>
<td>34</td>
</tr>
<tr>
<td>2.3.4 Beta distribution</td>
<td>36</td>
</tr>
<tr>
<td>2.3.5 K-means clustering</td>
<td>38</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Loss of function experiments in mice RNA-binding proteins with neural-specific functions and neuronal function consequences .................................................. 12

2.1 Publicly available datasets analysed ...................................................... 30
2.2 Tissues used to study ortholog AS events between human and mouse (MultiTissue dataset) ................................................. 31
2.3 Summary of R functions and packages used for analysis .......................... 40

3.1 Illustrative contingency table used in statistical test for the hypothesis of a regulatory role of nSR100 in target exons .................................................. 54
List of Figures

1.1 Overview of the general mechanisms involved in gene expression in eukaryotes .......... 4
1.2 Reactive regions in intron definition ................................................. 5
1.3 Intron removal process ................................................................. 6
1.4 Regulation of an exon inclusion ....................................................... 7
1.5 Alternative splicing event types ..................................................... 8
1.6 Comparative genomics of splicing levels in metazoans ......................... 10
1.7 Examples of neuronal-specific alternative splicing events ...................... 13
1.8 Position-dependent activity of RNA-binding proteins regulating splicing .......... 14
1.9 Neural stem cells differentiation into neurons and glial cells .................. 16

2.1 General overview of an RNA sequencing experiment .......................... 24
2.2 Bridge amplification ........................................................................ 25
2.3 Illumina® sequencing procedure ....................................................... 26
2.4 RNA-Seq reads used in quantification of exon inclusion ....................... 29
2.5 Principal component analysis example ............................................. 33
2.6 Shape parameters effect in Beta distribution ..................................... 36
2.7 K-means clustering algorithm ......................................................... 38

3.1 Beta distribution shape differences in a model for PSI quantification .......... 43
3.2 Comparison between Beta distribution and VAST-TOOLS PSI values ........ 44
3.3 Beta distribution modelling of exon inclusion levels ............................ 45
3.4 Example of tissue-specific conservation of an orthologous human/mouse alternative splicing event ............................................................... 46
3.5 Principal component analysis of gene expression data in Hubbard’s dataset .... 47
3.6 Time-correlated alternative splicing events ........................................ 49
3.7 Exon switches anti-correlated with Nova1 expression during neuronal differentiation ... 51
3.8 Exon switches correlated with nSR100 expression during neuronal differentiation .... 52
3.9 Mapping RNA-binding proteins motifs in the vicinity of an exon skipping event ........ 53
3.10 K-means clustering by correlation of PSI profile along neuronal differentiation .... 55
3.11 Expression patterns of lineage markers in NPC 46C RNA-Seq data ............ 56
3.12 Examples of exons progressively more included/skipped in NPC 46C ............ 57
3.13 Principal component analysis of all gene expression data in neurogenesis/nervous systems samples ............................................................... 59
3.14 Principal component analysis of all exon skipping alternative splicing events data in neurogenesis/nervous systems samples ........................................... 60
3.15 Principal component analysis for microexon skipping alternative splicing events data in neurogenesis/nervous systems samples .................................................. 61

A.1 Variance and PSI range analysis for removal of low varying AS events .................. A-1
B.1 Principal component analysis of centred but unscaled gene expression values ........ A-3
B.2 Principal component analysis of standardized (mean of zero and variance of one) gene expression values ............................................................... A-4
B.3 Principal component analysis of centred and unscaled log-transformed gene expression values A-4
C.1 General scheme illustrating pipeline for analysis of PSI data ................................. A-5
# List of Acronyms and Symbols

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'ss</td>
<td>3' splice site</td>
</tr>
<tr>
<td>5'ss</td>
<td>5' splice site</td>
</tr>
<tr>
<td>AS</td>
<td>Alternative splicing</td>
</tr>
<tr>
<td>BPS</td>
<td>Branch point sequence</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementar DNA</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>ESC</td>
<td>Embryonic stem cell</td>
</tr>
<tr>
<td>ESE</td>
<td>Exonic splicing enhancer</td>
</tr>
<tr>
<td>ESS</td>
<td>Exonic splicing silencer</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>ISE</td>
<td>Intronic splicing enhancer</td>
</tr>
<tr>
<td>ISS</td>
<td>Intronic splicing silencer</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>NA</td>
<td>Not available</td>
</tr>
<tr>
<td>NMD</td>
<td>Nonsense mediated decay</td>
</tr>
<tr>
<td>NPC</td>
<td>Neural progenitor cell</td>
</tr>
<tr>
<td>NPC 46C</td>
<td>Sox1-reporter cell line</td>
</tr>
<tr>
<td>NSC</td>
<td>Neural stem cells</td>
</tr>
<tr>
<td>P-value</td>
<td>Probability value</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PPT</td>
<td>Polypyrimidine tract</td>
</tr>
<tr>
<td>pre-mRNA</td>
<td>Precursor messenger ribonucleic acid</td>
</tr>
<tr>
<td>PTB</td>
<td>Polypyrimidine tract binding</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNP</td>
<td>Ribonucleoprotein</td>
</tr>
<tr>
<td>snRNA</td>
<td>Small nuclear ribonucleic acid</td>
</tr>
<tr>
<td>snRNP</td>
<td>Small nuclear ribonucleoprotein</td>
</tr>
<tr>
<td>SSO</td>
<td>Splice-switching antisense oligonucleotide</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>U2AF</td>
<td>U2 RNP auxiliary factor</td>
</tr>
<tr>
<td>Greek symbols</td>
<td></td>
</tr>
<tr>
<td>$\alpha$, $\beta$</td>
<td>Beta distribution shape parameters</td>
</tr>
<tr>
<td>Subscripts</td>
<td></td>
</tr>
<tr>
<td>fit</td>
<td>relative to the fitting process</td>
</tr>
<tr>
<td>Superscripts</td>
<td></td>
</tr>
<tr>
<td>$\sim$</td>
<td>Homozygous knock-out</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

“For the real amazement, if you wish to be amazed, is this process. You start out as a single cell derived from the coupling of a sperm and an egg; this divides in two, then four, then eight, and so on, and at a certain stage there emerges a single cell which has as all its progeny the human brain. The mere existence of such a cell should be one of the great astonishments of the earth. People ought to be walking around all day, all through their walking hours calling to each other in endless wonderment, talking of nothing except that cell.”

Lewis Thomas (1979)

1.1 Molecular biology of splicing

By the late 1960s, Francis Crick and James Watson expanded crystallography work by Rosalind Franklin to propose the deoxyribonucleic acid (DNA) double-helical structure, contributing to a major keystone in biology. This proposal, which was part of the work that was awarded a Nobel Prize of Medicine in 1962, led to the establishment of what is commonly termed the Central Dogma of Molecular Biology, stating that genes’ sequences encode all information needed to produce a protein and they transfer that information through transcription into messenger ribonucleic acid (mRNA), which, after leaving the cell’s nucleus, is translated to synthesize proteins, in a unidirectional process of reading genes’ instructions to generate proteins [1,2].

More than fifty years after, the proposal of the DNA double helix not only provided plenty of validation to both DNA structure and transcription and translation mechanisms but also enabled much more insight into the complex molecular interactions encompassing protein synthesis, genomic replication and other cellular phenomena. Interestingly, RNA, which in the light of the simplicity of Crick’s model served only as a template carrier for information, has been in the centre of the steady criticisms to the naivety of the Central Dogma.

Considering molecular functions of the cell, a number of important discoveries have proven the Central Dogma to be incompatible with our current perspective on the cell’s complex genetic mechanisms [3]:

• Reverse transcription: the process by which RNA serves as template to generate a new, single-stranded DNA molecule;

1 in Scott F. Gilbert, Developmental Biology
• **DNA proofreading and repair**: stability of the genome is maintained through protein-based corrections, if DNA is subjected to chemical or physical damage;

• **Post-translational protein modifications**: to achieve its functional structure, most proteins undergo covalent alterations, such as adenylation or phosphorylation, posterior to ribosomal translation;

• **Post-transcriptional RNA processing**: RNA is modified after copied from DNA and messenger RNAs encoding proteins are the result of the internal cleavage of some regions.

These critical points, amongst many others currently well established in biology, show unequivocally that the genetic information flow is not unidirectional from DNA to proteins and that there are interactions between each of the three major elements on the Central Dogma: DNA, RNA and proteins. For the purpose of the current work, one important consequence of this more accurate perspective on the molecular mechanisms of DNA replication and protein synthesis arises: during the process of reading DNA to produce a protein, a number of events take place to tune gene expression in a much more complex way than what was first described by Crick. The following sections aim to explore the role of precursor messenger RNA (pre-mRNA) splicing in the process of reading information out of the genome and, consistently with cell's environmental stimuli, produce the adequate amount of the correct gene product to achieve a specific functional role.

### 1.1.1 Genomics, transcriptomics and beyond

**Decoding genomic information**

The hereditary molecular instructions leading to cell’s biological functions are encoded as the sequence of nucleotides in specific regions – genes – of DNA molecules inside the nuclei of cells of all living organisms. These instructions further lead to the synthesis of products which exert a specific activity in the cell’s metabolism. Human DNA molecules contain approximately 3 300 million base pairs, distributed along 23 pairs of chromosomes in tightly packed, two-meter-long chromatin fibre fitting into small, micrometric nuclei. The human genome, with roughly 25 000 genes, is distributed throughout chromatin and encodes the information to synthesize the biological products, such as proteins or RNA molecules, required to achieve all the functions inside a living cell [4, 5].

The genome is kept relatively constant during the cell’s life cycle but information contained in genes’ sequences must be constantly decoded to synthesize its products accordingly to the environment. This decoding starts by the actual copy of the gene sequence into a single stranded molecule of ribonucleic acid (RNA) by base-pair complementarity between DNA and RNA nucleotides. Since this copy is achieved using the same type of chemical language, i.e. in the form of nucleotides, it is called transcription [4].

Transcription starts when RNA polymerase, the enzyme catalysing transcription, recognizes the transcription start site (TSS) of a gene, after the assembly with accessory proteins called transcription factors (TFs), and recognizes specific sequences within the DNA sequence. These enzymes slide through the DNA molecule and detect promoter sequences, regions upstream from the point where transcription should start. It is important to note that RNA polymerases are only able to synthesize RNA molecules in the 5' → 3' direction. The polarity of the promoter sequence ensures the binding of the RNA polymerase in the correct strand and consequent transcription of the adequate template sequence [4].
RNA polymerases promote the unwinding of the DNA double strand, rendering it accessible for complementary ribonucleotides to bind. Available ribonucleoside triphosphates in the medium stochastically approach the open DNA molecule and every time a complementary ribonucleotide reaches its position, RNA polymerase catalyses the formation of phosphodiester bonds between the 3’ hydroxyl and 5’ phosphate of consecutive ribonucleotides, providing the backbone for the RNA molecule, called RNA transcript [6]. Contrarily to what happens in DNA replication, RNA molecules do not stay bound to the DNA template strand but they detach, enabling the DNA molecule to fold back into double-helical structure. Transcription ends when RNA polymerase reaches the terminator region, where specific nucleotide sequences are recognized by the enzyme [4].

Figure 1.1 depicts a schematic general overview of the mechanisms involved in regulation of gene expression. Transmembranar receptors allow the detection of signals from the environment, and a sequence of steps leads to the activation of transcription factors that act in promoting transcription of target genes. Then, decoding of genetic information takes place, followed by a series of steps that finally enable the formation of a gene product (a protein, in the case illustrated). Next sections focus on the important steps of precursor messenger RNA processing.

The splicing reaction

After being synthesized in the nucleus, the pre-mRNA molecule must undergo processes to ensure the resulting transcript is stable enough and mature to travel into the cytoplasm of the cell, where protein synthesis takes place. Precursor messenger RNA transcripts are modified in the 5’ end, with the inclusion of a methylated guanine nucleotide – RNA capping. Moreover, the 3’ end of the RNA transcript is modified with the addition of a sequence of a few hundred adenine nucleotides (polyadenylation), forming the poly-A tail [4]. The polyadenylation process is important for the subsequent steps of pre-mRNA maturation, not only because it promotes the stabilization of the RNA transcript but also because the poly-A tail will be recognized as a marker for the end of the coding region, which is upstream from the region where the poly-A tail can be found. The pre-mRNA contains the protein coding regions and two terminal untranslated regions [6].

One of the most surprising discoveries in genetics was the fact that genes are distributed along chromatin in a discontinuous manner, meaning that coding regions – exons – are intercalated with longer non-coding interrupting regions – introns. The average human gene has 9 exons and 8 introns and, while exons have usually less than 200 nucleotides, introns can be up to 10 000 nucleotides long [4–6]. This implies that, in order to further process a pre-mRNA into an mRNA containing only coding, “meaningful” sequences and fully translatable, introns have to be removed and exons joined together, in a process called splicing. Similarly, pre-mRNAs that do not encode for proteins do also need to be spliced to originate long non-coding RNAs. In most cases, the splicing process results in a product mRNA having 10% of the pre-mRNA’s length [8].

Since the splicing reaction aims to join consecutive exons and remove the introns in between, the first fundamental step is to distinguish between exonic and intronic regions and, more specifically, to pinpoint the exon/intron junctions. Within the region comprised between two exons there are important reactive “special” sequences of nucleotides, signalling exon/intron junctions and other points important for splicing. These regions, depicted in Figure 1.2(a) are the 5’ splice site (5’ss), defining the junction between the first exon and
Figure 1.1: Overview of the general mechanisms involved in gene expression in eukaryotes

External factors are detected by the cell through their binding to transmembrane receptors, leading to the activation of specific transcription factors that are able to enter the nucleus and exert its role in promoting transcription of target genes. Decoding genetic information involves transcription from DNA to a transcript of precursor messenger RNA (pre-mRNA). This molecule is then subjected to different processing mechanisms, removing non-coding regions (introns) ligating coding regions (exons). Polyadenylation and 5' capping ensure stabilization of the transcript, enabling its export to the cytoplasm, where translation and protein synthesis takes place. With post-translational modifications, proteins achieve their functional conformations. Adapted from [4,7].

The spliceosome, the largest and one of the most sophisticated ribonucleoprotein complexes in the nucleus of eukaryotic cells, is composed of five uridine-rich small nuclear ribonucleoproteins (snRNPs) subunits – U1, U2, U4, U5 and U6 – assembled to hundreds of proteins, and catalyses two cleavage...
and one ligation reactions needed for splicing, based on consecutive interactions with the specific reactive regions in Figure 1.2(a). snRNPs in the spliceosome contain small nuclear RNAs (snRNAs) that may bind to pre-mRNA through base pair complementarity, enabling splice site recognition, as shown in Figure 1.2(b). The splicing process starts when the U1 snRNP binds to the 5' splice site through base pair complementarity of its U1 snRNA with six nucleotides of the intron and up to three nucleotides in the exon, followed by the recognition of the branch point adenosine by the U2 snRNA of U2 snRNP combined with the recognition of the polypyrimidine tract and 3'ss by the U2 snRNP auxiliary factor (U2AF). By rearrangement of both the U1 and U2 snRNPs and binding of the other spliceosomal snRNPs, the 3'ss, 5'ss and branch point are brought in close proximity [8–11].

The splicing reaction is depicted in Figure 1.3. The 2' hydroxyl group of the adenosine within the BPS makes a nucleophilic attack to the phosphate of the 5'ss, cleaving the phosphodiester bond in the exon/intron junction and releasing exon 1 with a hydroxyl group at the 3’ end and simultaneously forming a new phosphodiester bond between the adenosine of the BPS and the 5’ end of the intron. At this point, the intron is still bound to the exon 2, forming a lariat structure. Afterwards, the 3' hydroxyl group of exon 1 makes a nucleophilic attack at the phosphate of the 3’ splice site, enabling the ligation between both exons and the release of the intron lariat. After intron removal, the U snRNPs and proteins involved in the spliceosome assembly are recycled to catalyze other reactions [8].
The extra complexity layer of alternative splicing

In addition to the fact that RNA transcripts need to be spliced in order to remove the non-coding intronic regions, not all exons are always joined together. During the process of removing introns and joining exons together, regulatory regions within neighbouring introns or the exon itself promote or inhibit the recognition of splice sites for a given exon, controlling the chances of the exon to be detected and included into the final transcript. As a consequence, the same pre-mRNA can generate a number of different messenger RNAs through combining different exons together. Consistent with this, some exons are more often included because the splice site chemical recognition is stronger – constitutive exons – while others may be skipped in some isoforms and included in others. Moreover, alternative splice sites may also be recognized by the spliceosomal components and lead to alternative definition of the exonic region, creating different length alternative exons. This splicing feature greatly expands the number of combinations within the coding regions of a pre-mRNA and consequently the number of possible outcomes of the transcript. An alternative splicing (AS) event is dependent on the existence of multiple competitor splice sites and the influencing behaviour of regulatory elements in the vicinity [9].

The regulatory sequences which lay within the pre-mRNA sequence itself are called cis-acting factors and have affinity with external factors – trans-acting factors. Depending on their positioning either within the intronic or exonic region and on the effect they have in promoting or inhibiting the splice site selection and the inclusion of the exon, those cis-regulatory sequences can be intronic splicing enhancers (ISE), intronic splicing silencers (ISS), exonic splicing enhancers (ESE) or exonic splicing silencers (ESS). Trans-acting RNA-binding proteins (RBPs) recognize these sequences and bind to either enhance or repress the binding of the spliceosomal components involved in exon definition.

RNA-binding proteins that have a role in splicing regulation are usually divided into three classes: canonical heterogeneous nuclear ribonucleoproteins (hnRNPs), serine/arginine repeat (SR) proteins and tissue-specific RNA-binding proteins [6]. In most cases, hnRNPs act as repressors for the neighbour splice site selection and SR proteins show an enhancement effect, promoting splice site selection. However, opposite effects have also been identified and therefore the current belief is that, depending on the binding position (intron or exon, for instance), RBPs may have antagonistic effects. The same has been reported
for a number of tissue-specific RNA-binding proteins, which may show both repressor or promoter effect, by binding to different regions of the pre-mRNA [5]. Since the splicing reaction is dependent on the interaction of a number of different elements, each step of the mechanism is subjected to regulation and a weak, not completely assembled spliceosome may be disassembled such that other stronger splice sites may be selected [10]. Figure 1.4 shows the mechanism of regulation controlling the inclusion of an alternative exon, with both cis and trans-regulatory elements depicted.

Figure 1.4: Regulation of an exon inclusion by binding of trans-acting factors to cis-acting elements in the proximity of the splice sites
C1 and C2 are constitutive exons and Alt is the alternative exon. Reddish colours indicate silencer effect and blueish colours indicate enhancer effect. ISS (intronic splicing silencer), ISE (intronic splicing enhancer), ESE (exonic splicing enhancer) and ESS (exonic splicing silencer), U1/U2 (spliceosomal small nuclear RNP's), U2AF (U2 auxiliary factor) are spliceosomal components [5, 9].

The interactions influencing splice site choice, which strongly suggest a highly regulated process instead of a stochastic one, result in more complex forms of alternative splicing than the inclusion or not of an exon [12]. The different types of alternative splicing event are depicted in Figure 1.5. In the top left panel, Figure 1.5(a) shows the most commonly known, canonical alternative splicing event, which is the skipping of the whole exonic sequence, termed exon skipping, skipped exon or cassette exon. At the top right panel, Figure 1.5(b) shows a similar case but using two different consecutive mutually exclusive exons, meaning that mature transcripts may only have one exon or the other. Figure 1.5(c) and Figure 1.5(d) both show the case where competitive, alternative splice sites are enhanced to be recognized by the splicing machinery, instead of the constitutive splice sites for the same exon, and alternative exons end up having different lengths. Moreover, first and last exons may also be influenced by alternative splice site selection and Figure 1.5(e) depicts alternative promoter regions, resulting in a different 5' or start exon, while Figure 1.5(f) shows an alternative last exon, with the poly-A tail indicating the end of transcription. Another possible AS event is the retention of an intron, shown in Figure 1.5(g), sometimes referred to as "lack of splicing". Finally, in figure Figure 1.5(h)
a situation of alternative polyadenylation is depicted, with 3’ untranslated regions with alternative, varying lengths [9, 13].

Figure 1.5: Alternative splicing event types
Light blue boxes correspond to constitutive exons, green/blue boxes represent alternative exons, gray segments indicate introns, gray boxes indicate poly-A tails and coloured lines connecting boxes correspond to the specific junctions in each type of event [9, 13].

Interestingly, the splicing reaction in vivo is reported to happen too fast to be supported only by one spliceosome per intron, and the vast majority of genes are multiexonic. Following this suggestion, recent studies showed that four spliceosomes and the pre-mRNA being spliced form a macromolecular structure, the "supraspliceosome", which is theoretically enough for splicing every transcript, irrespectively of length and number of introns [8].

Functional consequences of alternative splicing

The first evident consequence of having differently spliced mRNA transcripts being produced in a highly regulated manner is the impact on producing proteins with modified functions. Some studies on the functional impact of alternative splicing report that isoform change in proteins may alter the protein’s catalytic activity, molecular interactions, structure, localization in the cell or stability, which has a impact in a great number of cellular processes [13]. Alternative splicing has been associated with regulatory functions in sex determination, diversity in neuronal wiring or determination of functions in membrane receptors of the mammalian nervous system [5].

Remarkably, AS has also an important role in controlling gene expression, namely by the introduction of a premature termination codon, that triggers the nonsense mediated decay (NMD) pathway, leading transcripts
to be degraded and therefore avoiding protein synthesis. This coupling of AS to NMD seems to occur to a great percentage of transcripts in human and mouse. Interestingly, transcripts that proceed to NMD are often regulated by the protein they produce, in a feedback loop controlling gene expression and maintaining homeostasis [12]. Furthermore, the inclusion of a single exon can work as a biological switch-like mechanism to achieve differential regulation in different types of cells, such as cells in different developmental stages [13]. Studies on alternative splicing events in mammals concluded that tissue-specific regulated intron retention events were responsible for the reduction of transcripts that do not have an important physiological role in the specified tissue [14].

Furthermore, alternative splicing has already been associated with a number of human diseases, such as cancer, muscular dystrophies and both developmental and degenerative diseases. One of the alterations linked to diseases is the occurrence of mutations in splice junctions or within cis-factors, that alter regulatory mechanisms and therefore shift isoform ratios. In some other cases, a more global set of splicing events is deregulated and therefore abnormalities are likely related to trans-acting factors. One important question involved in constant debate is whether splicing alterations associated to diseases are causes or consequences of other deregulated mechanisms [5,13].

Alternative splicing is one of the main mechanisms expanding the transcriptomic capacity of genes with more than one exon and contributes to the explanation of why evolutionary distant species may sometimes have similar number of genes but yet show huge phenotypic complexity differences. Although genetic information is maintained relatively constant within a cell, excluding the sporadic occurrence of mutations, the vast majority of genes is not being expressed at all times. The dynamic behaviour of the cell, by which it is able to react to external environment stimuli and produce different proteins in different contexts, is based on the ability to make decisions about the genes to be expressed at a given moment which are consistent with the environment [4]. By coupling the regulated transcription of the genes in the genome with alternative splicing, a processing mechanism that largely expands the number of possible RNA transcripts, the cell is empowered with a biologically efficient and elegant mechanism for fine-tuning gene expression without being fully limited by the number of genes. Bringing back the Central Dogma into this discussion, it is now clear that RNA complexity, which is dependent on the number of unique different forms of RNA inside a cell, is strongly responsible for biological complexity [12].

1.1.2 Alternative splicing and species’ complexity

The advent of genome sequencing techniques and genomic analysis that followed has provided data for a number of breakthroughs in biology. Since information from genomic annotations became available, one major conclusion was made clear: the number of genes of a species does not correlate with organismal complexity. This surprising observation led to questioning what other biological mechanisms may be explaining differences across species and through evolutionary distance. Figure 1.6 shows the number of total, multi-exon and protein-coding genes, as well as the number of different isoforms for four species of well-studied metazoans [5]. One striking observation is the growing number of unique isoforms with species complexity without a relevant increase in the number of genes. Remarkably, while the number of multi-exon genes is similar when comparing human and worm, the number of different isoforms reflects, among other
factors, the alternative splicing role in expanding the human proteome when compared to the worm’s.

Figure 1.6: Comparative genomics of splicing levels in metazoans
Comparison between the total numbers of genes and isoforms, showing transcriptomic amplification power in species with prevalence of alternative splicing [5].

More than 95% of human multi-exon genes are known to undergo alternative splicing [15]. Human multi-exon genes have on average seven isoforms, while mouse’s have slightly over three and fly’s and worm’s have less than two [5]. Alternative splicing complexity was compared between different tissues from ten vertebrate species separated by 350 million years of evolutionary distance [16] and results show that tissue-specific AS patterns cluster species together instead of tissues. Moreover, this species-specific AS signature is more likely cis-regulated than dependent on trans-regulatory factors. The tissue-specificity of gene expression profiles is indeed conserved across species. The AS species-specific signature suggests that, in vertebrates, evolutionary pressure may have forced alternative splicing mechanisms to evolve rapidly, resulting in a more relevant contribution to species phenotypic diversity than gene expression evolutionary changes. However, AS events with important functional consequences for a tissue are relatively conserved. In conclusion, AS seems to be a mechanism through which species greatly expand the transcriptome, although tight regulation seems to be coherent with events that have a strong and conserved physiological impact [16].

Examples have been reported of single alternative splicing events having as consequence the misregulation of a great number of other splicing events. Namely, an exon skipping event affecting PTBP1, an RNA-binding protein with functions in controlling brain-specific AS patterns, has the consequence of affecting the downstream regulation of a number of alternative splicing events, reinforcing the idea that it is a rapid and efficient mechanism for evolutionary variation between species [17].
Almost half of AS isoforms show differential expression between tissues. Importantly, a number of tissue specific AS events and regulatory molecules influencing tissue-dependent ratios between isoforms have already been suggested. Tissue-specific regulators include MTMR1 in muscle, CD44 in cancer and the AS pattern of TCF7L2. Therefore, evidence suggests that alternative splicing is frequently regulated in a tissue-specific manner [18].

### 1.1.3 Alternative splicing in the brain

In vertebrates, alternative splicing is more prevalent and more conserved in brain and cerebellum [16]. In addition to this, there are a number of different neural-specific splicing *trans*-regulators, which does not seem to be the case for the generality of other tissues [19]. Taken together, these observations strongly point out that brain-specific alternative splicing is one of the main players contributing to functional complexity of the nervous system.

The human brain is estimated to express half of all human genes [18]. This high pattern of gene expression shows different distributions both in space and time [18]. Alternative splicing may be the most important mechanism capable of providing the biological diversity needed for the functioning of nervous systems. More specifically, it has been suggested that AS would be the most versatile mechanism to account for the amplification on the number of cell adhesion molecules which are fundamental for the establishment of synapses [18].

#### Splicing factors with brain-specific functions

NOVA proteins (NOVA1 and NOVA2) were the first splicing regulators to be described as having a neural-specific mode of action. These proteins were identified in patients with a motor nervous system disorder being targeted by autoimmune responses. Studies on the impact of loss of function in Nova proteins have provided validation for its role in controlling neuronal alternative splicing programmes. Table 1.1 sums up the evidence for the neurologic impact of lost of function of each or both Nova proteins in mice. Bioinformatic approaches using probabilistic models led to the establishment of the *Nova* regulatory network of alternative exons in genes with functions in synapse formation and axon guidance [19]. The mode of action of Nova proteins in brain-specific regulation of splicing outcomes suggests that these splicing factors directly regulate 700 splicing events in the mouse brain [20].

RNA-binding protein fox homologue (RBFOX) is a family of three paralogs which are also neural-enriched splicing factors and their interactions with RNA have been mapped in the mouse brain [19]. Moreover, AS events with evidence to be regulated by RBFOX proteins exhibit a changing pattern during brain development [19]. Interestingly, the evidence for a differential distribution of RBFOX proteins in human and mouse brains suggests that RBFOX1 and RBFOX2 have related but distinct roles in AS regulation in neuronal tissues. Table 1.1 shows that experiments with *Rbfox1/2* deletions in mice central nervous system had caused excitability and structural alterations [19].

PTBP1 and PTBP2 are polypyrimidine tract binding (PTB) proteins known to be regulators of brain-specific alternative splicing, acting in a mutually exclusive manner. During *in vitro* neuronal differentiation from
Table 1.1: Loss of function experiments in mice RNA-binding proteins with neural-specific functions and neuronal function consequences

<table>
<thead>
<tr>
<th>RBP family</th>
<th>Genomic alteration</th>
<th>Functional consequences in neuronal development</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTBP</td>
<td>Ptbp1-/-</td>
<td>Precocious neurogenesis with depletion of neural stem cell pool</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>Ptbp2-/-</td>
<td>Early embryonic mortality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ptbp1 RNAi knock-down</td>
<td>104 target splicing events altered in embryonic stem cells and neural progenitor cells (70 repressed, 34 induced)</td>
<td>[21]</td>
</tr>
<tr>
<td></td>
<td>Knock-out in developing cortex</td>
<td>Inhibits differentiation of neurons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nSR100/SRRM4</td>
<td>Accumulation of Pax6+ progenitor cells in ventricular zone</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>Germline deletion</td>
<td>Fewer Pax6+ progenitor cells in ventricular zone</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Srrm4-/-</td>
<td>Fewer late-born upper-layer neurons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nova1 knock-out</td>
<td>Apoptosis in motor neurons leading to motor failure and eventual death</td>
<td>[9]</td>
</tr>
<tr>
<td>NOVA</td>
<td>Nova2-/-</td>
<td>Misregulation of activity-dependent long-term potentiation on slow inhibitory postsynaptic current</td>
<td>[19]</td>
</tr>
<tr>
<td></td>
<td>Nova double knock-out</td>
<td>Mislocalization of cortical layers</td>
<td></td>
</tr>
<tr>
<td>RBFOX</td>
<td>Rbox1 deletion in CNS</td>
<td>Susceptibility to seizures</td>
<td>[19]</td>
</tr>
<tr>
<td></td>
<td>Rbox2 deletion in CNS</td>
<td>Deficiencies in migration and dendrites of Purkinje cells</td>
<td></td>
</tr>
</tbody>
</table>

embryonic stem cells to motor neurons, expression levels of Ptbp1 decrease while Ptbp2 increase [21]. Exon 10 of the Ptbp2 transcript is neural-specific, meaning that it is usually included in a neural context but not in other types of tissues. Interestingly, this AS event is regulated by its paralog, Ptbp1, which in non-neuronal cells induces the skipping of Ptbp2 exon 10. The Ptbp2 isoform lacking exon 10 contains a premature termination codon which results in its targeting by NMD, causing a decrease on the levels of the protein in non-neuronal cells. Ptbp2 is expressed in neural precursor cells but in developing neurons the micro-RNA miR-124 acts as a silencer for Ptbp1. This results in the inclusion of exon 10 and higher levels of Ptbp2 expression in later neural development. This RBP acts in the regulation of the inclusion of a number of embryonic stage exons. In summary, mismatched but consistent levels of expression of Ptbp1 and Ptbp2 in developing neurons ensure the maintenance of the correct neuronal development AS programme [19,21].

The neural-specific SR-related protein of 100 kDa, nSR100 or SRRM4, is conserved in vertebrates but not present in invertebrates, which corroborates the hypothesis of being a player in the mechanisms contributing for neuronal functional complexity. Although nSR100 lacks a canonical RNA-binding domain, it recognizes UGC-rich sequences located between the polypyrimidine tract and the 3' splice site of the exons it targets. Moreover, typical motifs for PTBP1 binding are enriched in regions nearby the splice sites of target exons for nSR100 and the events regulated by both significantly overlap, with PTBP1 and nSR100 showing opposing effects in the exons they regulate [9].

nSR100 regulates a network of neuronal-specific AS events which are involved in remodelling protein-protein interactions in neurogenesis. One of the roles of nSR100 in neuronal development, illustrated in the rightmost panel of Figure 1.7, is related to its impact on the control of the inclusion of one exon in the repressor element 1 silencing transcription factor (REST), which acts as a silencer in neurogenesis. In non-neural cells, the exon is skipped and a full length, functional isoform is generated to repress neurogenesis-related genes. However, in neural cells, the presence of nSR100 induces the inclusion of the exon, generating the truncated isoform REST4, which lacks domains required for DNA-binding. As a consequence, its repressive activity in pro-neurogenesis genes is compromised. More studies focused on the combinatorial effects of trans-acting neuronal-specific AS factors provided evidence that the nSR100 function in neurogenesis is to overcome the repression regulated by PTBP1 and ensure correct exon inclusion [19,22].
RNA splicing maps

Cross-linking immunoprecipitation sequencing (CLIP-Seq) enables the sequencing of RNA regions which are bound by specific RNA-binding proteins. Recent findings using both RNA and CLIP sequencing have provided more insight into the dependence between RBP binding positions and splicing outcomes. CLIP-Seq enables the identification of the RNA motif sequences to which an RBP binds. Taking this information along with RNA-Seq data, from which inclusion levels of exons can be quantified and correlated with expression levels of the RBP, it is possible to associate the preferred binding position of the RBP, i.e. the region within the vicinity of the exon with a significant enrichment for its binding motifs, to the increase or decrease in that exon inclusion levels. These associations, that have been called RNA splicing maps, [19] are represented in a simple version in Figure 1.8. NOVA, RBFOX and PTBP1 were already predicted to promote, in general, the inclusion of their target exons upon binding in the proximal region of the downstream exon, while they promote skipping when they bind in upstream intron or within the first nucleotides of the exon itself [19, 23]. This type of insight is extremely valuable for further understanding the regulatory mechanisms by which splicing factors control exon inclusion.

Alternative splicing roles in neuronal development processes

Experiments using small interfering RNA (siRNA) transfections to knock down the Ptbp1 gene, combined with CLIP-Seq data, revealed a number of alternative splicing events that are targets for PTBP1 and significantly change between embryonic stem cells and neural progenitor cells. PTBP1 regulates a particular
CLIP-Seq data enables the identification of regions of RNA with enrichment for RBP binding motifs. Nova, Rbfox and Ptbp1 promote target exon inclusion upon binding in downstream intron, while the same RBPs and Ptbp2 promote the skipping of the target exon upon binding in the upstream intron or the first nucleotides of the exon; on the contrary, nSR100 promotes the inclusion of the target exon when bound to upstream intronic regions [19].

Figure 1.8: Position-dependent activity of RNA-binding proteins regulating splicing

Alternative splicing event in Pbx1 (homeodomain transcription factor) with fundamental impact in neuronal function. This regulatory effect is illustrated in Figure 1.7. The Pbx1a isoform is conserved across mammals and contains exon 7, while Pbx1b lacks exon 7 and has an alternative termination codon in exon 8, generating a shorter isoform. Interestingly, early embryonic tissues express more of the Pbx1b isoform while more mature neural tissues express the Pbx1a isoform, which is known to activate transcription of neuronal genes. As Ptbp1 expression decreases as cells acquire the neural phenotype, inclusion of exon 7 is derepressed and the Pbx1a exon 7-containing isoform levels increase [21].

Different types of alternative splicing events have specific functional roles. Microexons (3 to 27 nucleotides in length) are shorter than regular cassette exons and they are more included in neural tissues than in any other type of tissue. Furthermore, microexons, which also happen to be more conserved than other types of AS events, show an inverse relationship between their length and their percentage of inclusion. Microexons are more included in the late stages of neuronal development and roles for transcripts containing these AS events have been associated with synaptic function and neurite outgrowth. nSR100 has been reported as regulating the inclusion of many microexons. Individuals with autism spectrum disorder have misregulations in neural microexons, coherent with reduced levels of nSR100 [15,19].

Neurons have a specialized system for transporting mRNA to synapses. Interestingly, unspliced introns have been observed in RNA extracted from dendrites and these intronic regions contain sequences which serve as cis-regulatory elements for controlling the localization of fully spliced transcripts to the correct localization in the synapse, by a mechanism of “hitchhiking” of fully spliced mRNA in the same mRNA granules [24]. Intron retention events have been suggested as an efficient mechanism for targeting for degradation transcripts that are not needed for that tissue, therefore contributing to tuning tissue
specificity [14].

**Neuronal developmental-specific alternative splicing events**

Neurexins are proteins involved in cell adhesion during synaptogenesis and interact with neuroligins to establish unambiguous affinity bindings that define synapse identity. There are only three neurexin genes in the human genome, each containing several cassette exons. Alternative splicing of neurexins generates thousands of different isoforms, with different affinities to post-synaptic receptors, generating unique responses to neural signals [18].

A recently published work [25] explored the role of histone and DNA modifications in the regulation of transcriptional programs and maintenance of phenotypes. G9a or EHMT2 methyltransferase (euchromatic histone lysine N-methyltransferase 2) is the enzyme responsible for H3 lysine 9 dimethylation (H3K9me2) in mammalian euchromatin, marks that usually cause transcriptional silencing, and has been associated with neuronal differentiation, namely in repression of non-neuronal genes. G9a methyltransferase has two isoforms differing in the inclusion of exon 10. This inclusion increases and is required during neuronal development. Interestingly, an increase in the levels of exon 10 inclusion does not alter the catalytic activity of the enzyme but still stimulates H3K9me2 levels by promoting G9a nuclear localization. G9a exon 10-containing isoform regulates its own alternative splicing, promoting neuron differentiation in a positive feedback loop in which neuron differentiation enhances the production of the isoform containing exon 10, which in turn improves more inclusion of the exon, promoting the nuclear localization of the enzyme [25].

RNA-binding motif protein 4 (RBM4) is a known regulator of alternative splicing with evidence for roles in pancreatic cell differentiation. RBM4 was found in mouse brains during gestation and its levels are reduced in Down's syndrome fetal brains. Moreover, it is known to regulate the neuronal differentiation alternative splicing of Numb’s exons 3 and 9. Mammalian NUMB is involved in the maintenance of neural progenitor cells and promotion of neuronal differentiation. Differential inclusion of exons 3 and 9 generates isoforms that differ in their phosphotyrosine-binding domain lengths and either promote proliferation or differentiation. RBM4 modulates exon selection of the NUMB isoforms to promote the expression of pro-neuronal genes, such as MASH1 [26].

### 1.2 Neuronal development

The vertebrate embryo dorsal ectoderm contains a region of columnar cells, the neural plate, which gives rise to the nervous system. A structural folding process – neurulation – leads to the formation of the neural tube, a hollow tubular structure that narrows and lengthens along the anterior-posterior axis of the embryo. The neural tube faces processes of folding, bending and closure to give rise to central nervous system structures, such as the brain and spinal cord. Cell differentiation from neural tube multipotent cells occurs afterwards to originate the neural progenitor cells, which are able to further differentiate into the three neural lineages: neurons and glial cells (astrocytes and oligodendrocytes). At the cellular level, neural tube cells are influenced by neighbour cells’ signalling in order to maintain the neural progenitor proliferation and undergo timely differentiation into neurons and glial cells [27,28].
Neural lineages, neurons and glial cells, depicted in Figure 1.9, are derived from the embryonic neuroectoderm in a process that starts with neural induction. Following, neural cells expand and spatially and temporally coordinated cell-fate decisions further result in regional specialization of neuronal types [29]. Neural induction is the first step for cells to acquire the neural fate and happens through alterations in the molecular environment surrounding ectodermal cells. Some molecules, such as noggin, follistatin and chordin show neural activity in binding to bone morphogenic protein (BMP), antagonizing BMP signalling. Interestingly, a number of studies have proposed that ectodermal cells have an intrinsic tendency to follow the neural fate unless this is inhibited by BMP signalling, while differentiation of ectodermal cells into non-neural tissues, such as epidermis, is induced by BMP. This predisposition of ectodermal cells to become neural progenitors, unless they experience an antagonistic environment, is known as "the default model" of neural induction [29,30].

Neural lineages, neurons and glial cells, depicted in Figure 1.9, are derived from the embryonic neuroectoderm in a process that starts with neural induction. Following, neural cells expand and spatially and temporally coordinated cell-fate decisions further result in regional specialization of neuronal types [29]. Neural induction is the first step for cells to acquire the neural fate and happens through alterations in the molecular environment surrounding ectodermal cells. Some molecules, such as noggin, follistatin and chordin show neural activity in binding to bone morphogenic protein (BMP), antagonizing BMP signalling. Interestingly, a number of studies have proposed that ectodermal cells have an intrinsic tendency to follow the neural fate unless this is inhibited by BMP signalling, while differentiation of ectodermal cells into non-neural tissues, such as epidermis, is induced by BMP. This predisposition of ectodermal cells to become neural progenitors, unless they experience an antagonistic environment, is known as "the default model" of neural induction [29,30].

Notch signalling pathway regulates timely production of neurons and glia, either promoting or inhibiting differentiation. Genes known as specific lineage markers are highlighted. Adapted from [31].

Neural specification of pluripotent stem cells relies on sequential biological processes related to specificity of the neural phenotype the cells acquire. Pluripotent cells become neural competent when they are exposed to a combination of transcription factors that targets neural differentiation. Cells become specified when they were already exposed to such a specific molecular environment but differentiation can still be stopped by inhibitory factors. Afterwards, cells commit to the neural fate and enter the differentiation process into neuronal and glial cells, even if there are inhibitory signals [27]. Moreover, in the beginning of neural development, neural stem cells undergo symmetrical divisions in order to expand the number of neural stem cells but, at a given point in gestation – the neurogenic phase –, these cells start its asymmetrical cell divisions, originating one neural stem/progenitor cell and one neuron. Further on in development, neural stem cells divide asymmetrically, producing one neural stem/progenitor cell and one glial cell – the gliogenic phase [32]. This timely separation of neurogenesis and gliogenesis is depicted in Figure 1.9, also highlighting genes.
known as specific lineage markers.

Neuronal differentiation during the neurogenic phase is dependent on the action of transcription factors with roles also in the regulation of differentiation/proliferation of neural stem cells [29]. PAX6, amongst others, induces expression of other transcription factors, such as proneural basic helix-loop-helix (bHLH), whose increase in expression promotes neuronal differentiation and inhibits astroglial differentiation [29]. Other transcription factors are required for context-specific neuronal differentiation, such as Neurogenin1/2 and ASCL1/MASH1 [29].

1.2.1 Neural differentiation \textit{in vitro}

Neural stem cells (NSCs) and neural progenitor cells (NPCs) are self-renewable, undifferentiated and multipotent cells present in the embryonic and adult brain. The cellular and molecular characterization of neural commitment to differentiation relies on \textit{in vitro} experiments where pluripotent stem cells undergo neural induction and are monitored in culture for a period of time, for instance in terms of gene expression. Moreover, already well-established lineage-specific markers enable the quantification of heterogeneity of cell types in cultures of differentiating cells. Upon neural induction \textit{in vitro}, embryonic stem cells (ESCs) are first converted to show a primitive ectodermal phenotype. Afterwards, cells proceed to become neuroectoderm, followed by the establishment of a pool of neural progenitor cells. These, in turn, will gradually give rise to neurons, first, and glial cells, in a posterior phase [28].

Transcriptomics of neuronal development

Gene expression profiles and developmental changes in embryonic structures were mimicked \textit{in vitro} during neural induction of a mouse embryonic stem cell line reporting the expression of the neuronal marker Sox1 (NPC 46C cell line) [28]. This cell line is a knock-in of Sox1-GFP (green fluorescent protein), therefore emitting green fluorescence when Sox1 is expressed, which marks the entering of cells into the neuronal phenotype.

When cultured in serum-free medium, NPC 46C cells express pluripotency markers in the first days in colony, such as Fbxo15, and embryonic stem cell (ESC) markers, such as Nanog, Pou5f1/Oct4, and Sox2, which is one of the earliest markers of neural stem cell identity. However, pluripotency genes expression is reduced as soon as after one day in culture, upon neural induction. Also, reduction in the expression levels of markers for endoderm (such as Gata1) is consistent with the increased levels of markers for ectoderm phenotype (such as Calcr, Fgf5 and Tnnc2), confirming commitment to an ectodermal phenotype after neural induction and loss of endodermal potential. After, from day six in culture on, the expression of neural precursor markers, such as Pax6, Sox3 and Chrdl1, is stabilized when cells enter the neural fate, although these genes are lowly expressed at neural induction. Expression levels for proneural genes increase with days in culture, consistently with the commitment and differentiation of the cells to acquire the neural phenotype [28].
Spacio-temporal regulation of neuronal development

Nervous system neurons have different functions which depend on their positioning with respect to the anterior-posterior and dorsal-ventral neural axis of the embryo. Patterning of this axis is shaped through position-dependent gradients of signalling molecules when cells are differentiating. The neural tube anterior-posterior axis is divided into the prosencephalon, mesencephalon, rhombencephalon and the spinal cord. Caudalizing molecules such as retinoic acid, Wnt and FGF with a spatial distribution of concentrations are responsible for the specification of these structures. After neural tube closure, dorsal-ventral axis follows anterior-posterior patterning. Dorsalization is promoted in a similar manner throughout all the neural tube and is promoted by BMP and Wnt, while ventralization is promoted by Shh (Sonic hedgehog). Temporal information is encoded in progenitor cells, which are able to generate central and peripheral nervous system neurons before and glial cells afterwards. Signalling molecules as fibroblast growth factor, bone morphogenic protein and Noggin are part of the regulatory mechanisms, which combined with intrinsic timing information, induce a temporal specification in stem cells [30,33].

One important regulator of neural development is the Notch signalling pathway, which controls the rate of neural progenitor commitment to differentiation, ensuring that neuron formation is timely. The equilibrium between proliferating and differentiating neural stem cells is also regulated by the Notch signalling pathway. In order to maintain NPC pluripotency, Notch activates downstream transcription factors, such as HES1 and HES5, which act as repressors of proneural genes and therefore help maintaining neural stem cells in a undifferentiated state [29]. Figure 1.9 shows the effect of Notch signalling at different point of neural differentiation. Notch acts positively to induce differentiaion from glial progenitors into astrocytes, but inhibits the differentiation of both neurons and oligodendrocytes [31].

Neural rosettes of cultured differentiating cells, which resemble the neural tube, also present Notch receptors and ligands [28]. This observation is consistent with what happens in the in vivo embryo, where Notch signalling is responsible for inducing timely production of neurons and glia. Notch signalling inhibition is known to result in premature production of neurons. Moreover, NPC 46C cells in culture show a switch in potential from early increased neurogenic potential to later gliogenic potential, also consistent with what happens in the embryo. [28]

1.2.2 Alternative splicing misregulations in neurodevelopmental disorders

As previous sections elucidated, differentiation into neurons and glial cells is strongly subjected to regulation by external factors. The connective network of roughly $10^{11}$ neurons eventually interacting with the $10^{12}$ glial cells [27] in the human brain is in part dependent on the correct differentiation, patterning and maturing of neural stem cells that takes place in the embryo and afterwards. However, during the process of neuronal development, defects in synaptic maturation or connectivity may lead to neurodevelopmental disorders, such as autism spectrum disorders or Rett syndrome, which drastically impair cognitive functions, locomotion and communication [34,35].

Alternative splicing, which is more prevalent in the brain than in any other tissue [16], has already been associated to a number of alterations that may compromise diverse aspects of neuronal development.
There is indeed growing evidence that alternative splicing misregulations play a role in neurodevelopmental disorders. Neurexin alternative splicing regulated by the RNA-binding protein SLM2 is misregulated in Slm2 knock-out mice, causing behavioural alterations consistent with loss of functional synapse specification. Genetic correction of the knock-out successfully rescues synaptic plasticity and behavioural alterations [36]. Irimia and colleagues have shown that the splicing pattern of a set of neuronal-specific, highly conserved microexons is disregulated in autism spectrum disorder [15]. Another study on the targets of the RBFOX family of RNA-binding proteins established a network of RBFOX-regulated AS events showing dynamic changes during neuronal development, that are misregulated in autistic brains [37]. Moreover, a mouse model of Rett syndrome revealed the misregulation of hundreds of alternative splicing events caused by the loss of function of MeCP2 protein, which acts as a transcription regulator and interacts with splicing modulators [38]. Taken together, these results reinforce the relevance alternative splicing has in shaping the function of the brain and suggest that novel evidence on the molecular mechanisms of neuronal differentiation may be further explored as potential therapeutic targets in neurodevelopmental disorders.

**Therapeutics of splicing**

Considering the effect that misregulations in splicing patterns may have in normal and pathological mechanisms in a number of diseases, splicing has been recently addressed as a potential therapeutic target. Splice-switching oligonucleotides (SSOs) are short synthetic molecules comprised of nucleotides that bind to a target, complementary sequence. By binding to target cis-regulatory regions controlling splice site selection, these 15-30 nucleotide long oligonucleotides are able to block the binding of other regulatory proteins and therefore inhibit their effect in splicing regulation. Chemical modifications make SSOs hardly degradable and its short length increases their affinity to the target sequence in pre-mRNA, by reducing the likelihood of unwanted binding. Although *in vivo* delivery of SSOs can be achieved using regular methods of administration such as intravenous or intramuscular, central nervous system applications are more demanding because these molecules do not cross the blood brain barrier. Fortunately, direct delivery to cerebrospinal fluid has provided good results, at least for more superficial regions of the brain. Therapies using SSOs in mice have shown alterations in splicing and disease symptoms for up to one year after administration [39].

After administration, SSOs can bind to circulating proteins and enter the cell by binding to receptors complementary to those proteins, targeting compartmentalization by endosomes or lysosomes. Afterwards, vesicle released SSOs are able to target pre-mRNA transcripts inside the nuclei of cells. SSOs can be used both to enhance or to inhibit splice site selection, depending on the therapeutic purpose, by having their sequences designed to bind to either splice sites or *cis* enhancer regions or *cis* silencer regions. By targeting splice sites or *cis*-regulatory regions, protein or snRNA binding is inhibited and exon definition is not processed. By binding to silencer regions, RNA-binding proteins are unable to bind to this regions and cannot exert their repressive activity [39].

Another trend in the field of therapeutically modulating splicing is to directly target spliceosome components. A number of molecules interfering with the spliceosomal assembly have been identified and are generically called *splicing inhibitors* because they block the splicing machinery. The interference may occur through inhibition of the enzymatic activity of ATPases that promote spliceosomal structural
rearrangements, blocking interactions between components of the spliceosome or enhancing stabilization of the same components in a way that suspends splicing. One promising technique is to use inhibitors for SF3B1, a spliceosomal protein first known as having a role in recruitment and stabilization of U2 snRNP at the branch point sequence but recently suggested to also have other roles along the splicing process. Three compounds have been identified as targeting and inhibiting SF3B1: pladienolide B, spliceostatin A and herboxidiene. In the presence of splicing inhibitors, selection of splice sites is affected, resulting in changes in exon inclusion levels and increased intron retention. These molecules are being used to further investigate biological functions of each component of the spliceosome and are also promising new drugs to target undesirable alternative splicing [40].

1.3 Motivation and goals

Evidence shows that alternative splicing works as an evolutionary advantageous strategy that higher organisms make use of for expanding the transcriptomic potential without being limited by gene number or genome size. Alternative splicing highly regulating brain specificity further corroborates this observation, since it is one of the main contributors for the functional complexity of the human nervous system. Previous studies focusing on alternative splicing roles during neuronal development provided more insight into the important activity of neuronal-specific RBPs in the regulation of the inclusion of exons that are important for neural functions. Also, different types of alternative splicing events seem to have specific functions in shaping the specificity of neuronal tissues, as it is evident by the inverse relation between the length of microexons and their tendency to be included in the brain or intron retention events removing non-essential transcripts from the cell.

However, prediction of functional consequences for alternative splicing events that change through development is required and will help understanding AS dysregulations in neurodevelopmental disorders. Also, although a number of RNA-binding proteins regulate AS in a brain-specific manner, global networks of co-regulated events are not established. Therefore, the identification of global AS alterations during neuronal development, both in terms of differences in inclusion levels of the exons and the regulatory mechanisms influencing them, is valuable and may provide therapeutic targets.

The goal of the current work is to use bioinformatics tools to analyse human and mouse available transcriptomic data in order to unveil the pattern of alternative splicing during neuronal differentiation and explore its regulatory mechanisms, by correlating alternative splicing quantifications with the expression levels of genes that encode RNA-binding proteins whose binding to RNA is predicted to be regulating them. Other than that, a method for quantification of exon inclusion levels taking into account differences in gene expression between biological replicates is proposed. Furthermore, characterization of a neural progenitor cell line which is widely used to study neural induction and neural differentiation from pluripotent stem cells in vitro is performed, contributing to a better understanding of gene expression patterns at each stage of neuronal differentiation in culture.
1.4 Document structure

This document is organized in four main chapters: Introduction, where the biological context of alternative splicing in neuronal development is explained, as well as the state of the art in the field of neuronal-specific alternative splicing regulation; Materials and Methods, where the analysed datasets are introduced and technical details about the methods used are explained; Results and Discussion, where the procedures that were followed to answer the proposed questions are detailed, with produced results explained and discussed; and Conclusions, where the main findings are summarized and additional questions are proposed for future work.
Chapter 2

Materials and Methods

The current work made use of publicly available transcriptomic datasets from human and mouse neuronal developmental cells, which were processed in order to quantify alternative splicing and gene expression. Splicing levels and gene expression quantifications were afterwards analysed and statistical hypotheses were tested. The following sections focus on the details of the tools and methodologies applied to do so. This work was developed at Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, in the NMorais Lab (Computational Biology) between February and September, 2016.

2.1 RNA-Sequencing data

In a given moment and for a certain condition, the cell’s transcriptome comprises all the different RNA transcripts in a cell, such as messenger, ribosomal, transfer or non-coding RNAs. Transcriptome characterization and quantification first appeared in 1991 as a database of approximately 600 randomly selected brain cDNA clones produced with the then recently developed Sanger sequencing technology, in a publication already predicting the future generalization of the approach to discover new coding regions and analyse comparative genomics faster and at a lower cost [41, 42].

The major interests in transcriptomics studies are the characterization of all types of transcripts within a cell, the accurate determination of genes’ coding region structures, such as their 5’ and 3’ ends or their exon/intron junctions, and the quantifications of gene expression when comparing different conditions [43]. Hybridization-based technologies have been used for the characterization of transcriptomes, where the identification of a transcript occurs by complementary binding of labelled complementar DNA (cDNA) to previously known, immobilized oligonucleotides in microarrays. Transcripts labelled with fluorophores, for instance, will emit fluorescence when bound to the oligonucleotides in the array and the intensity is measured and compared between conditions [43].

However, hybridization technologies have a number of limitations. The most striking one is the fact that the identification of a transcript in a cell is dependent on the previous knowledge of the genome sequence, which is specially inadequate in splicing analysis, where the discovery of new, not annotated isoforms is most valuable. Moreover, transcriptomics analyses based on hybridization are usually noisy due to the background
signal. The quantification of gene expression levels with microarrays is of low sensitivity, because detection of a transcript is limited to a range of expression, masking low and very high expression genes [43,44].

On the other hand, sequence-based methods directly identify the nucleotide sequence of cDNA. RNA sequencing (RNA-Seq) methodology uses high-throughput sequencing techniques to sequence cDNA complementary to RNA fragments of the transcriptome of interest [44]. In 2006, the first publication using RNA-Seq came out, describing a set of 200 000 reads, 110 base pair long, produced using the 454/Roche technology that characterized the transcriptome of a prostate cancer cell line [45]. RNA-Seq is currently the dominant technology for studying splicing, since it has a number of advantages over microarrays. Besides the fact that the identification of new sequences is possible in parallel with transcript quantification, because there is no need for a reference genome, sequencing methods are not as noisy as microarrays. Moreover, RNA-Seq is much more sensitive, provides single-base resolution and can detect a larger range of gene expression quantification (9 000-fold for RNA-Seq when compared to 100-fold for microarrays) [43,44,46].

![Diagram of RNA sequencing experiment](image)

**Figure 2.1: General overview of an RNA sequencing experiment**

RNA-Seq involves the extraction of the interest population of RNA from the samples, followed by fragmentation of the transcripts. Afterwards, RNA fragments are converted to complementary DNA (cDNA) and adaptors are bound to both ends of each fragment. High throughput sequencing is then performed, and resulting reads are aligned to a reference. In the example, coloured boxes depict exons. The reference genome contains one exon (light blue) which was not included in the transcript depicted, illustrating the definition and quantification of the skipping of an alternative exon through RNA-Seq [43].

Although RNA-Seq technologies may vary in the technical details, the common pipeline is diagrammed in Figure 2.1. The RNA population of interest is extracted from the cell and converted to complementary DNA (cDNA), which is more stable than RNA molecules. Fragmentation of longer cDNA is required for the creation of a cDNA library. Small adapter sequences are bound to each of the cDNA fragments of the library, either only in one end (single-end) or to both ends of the fragment (paired-end). Each fragment from the library, with or without amplification, is then sequenced using one of the high-throughput sequencing methods. The sequencing process produces reads, resulting in genome-wide transcription mapping, from which gene expression levels may be quantified and other questions can be addressed, such as the characterization of
splicing patterns [43].

RNA-Seq results in reads with a known length, typically 30 to 400 base pairs, although it depends on the technology [44]. Read length influences the sequencing and following analysis of data, because longer reads are more likely to be uniquely mappable to a given annotated region in the reference [46]. Another limiting aspect is the sequencing depth or the number of reads (library size) for a given sample. A higher number of reads, or a deeper sequencing, means more transcripts will be detected and therefore quantified. Recommended library sizes range from five million (for gene expression analysis in high or medium-expressed genes) to one hundred million reads (for more accurate quantification also for low expression genes) [46]. Moreover, ideal library size depends on the target transcriptome complexity because at a certain point an increase in sequencing depth does not bring more information about the sequences in the library. Therefore, decisions on the sequencing depth to use depend on the targeted transcriptome, the consequent expected library complexity and budget restrictions [46].

**Illumina HiSeq® 2000 System**

One of the common used next-generation sequencing technologies in RNA-Seq, used in the datasets analysed in this work, is the Illumina HiSeq® 2000 System. In this technology, cDNA fragments with paired-end adaptors undergo amplification before the sequencing process. Both the adaptors contain universal priming sites, which are complementary to primers (oligonucleotides) clustered and immobilized in solid surfaces (glass slides). Each slide contains a high number of both direct and reverse primers, complementary to the adaptors, which are responsible for the high rate of amplification [47].

---

**Figure 2.2: Bridge amplification**

Bridge amplification enables the generation of clones for each of the fragments in the library: high density forward and reverse primers are attached to the slide, promoting the bending of the free-end of the fragment to hybridize to the complementary primer on the slide. Polymerization in a "bridge" occurs and cycles of changing temperatures promote the denaturation of double strands and removal of reverse strands. Each fragment is amplified in clusters. Adapted from [47].
For the amplification reaction, simplified in Figure 2.2, one fragment binds to one of the spatially distributed, immobilized primers and, in the presence of nucleotides and polymerases, the reverse strand of the fragment is synthesized. The double-stranded molecule is denatured and the original template is washed away. Then, reverse strand bends over in a way that the free adapter binds one complementary primer attached to the slide. By bridge amplification polymerases use free nucleotides to synthesize a complementary strand that is a clone of the original sequence. This double-stranded molecule is denatured, resulting in two clones. This process is repeated in cycles and occurs simultaneously in many clusters, originating clones of all the fragments in the library, which remain tethered to the glass slide in the place of origin [47, 48].

In the sequencing phase, depicted in Figure 2.3, complementary primers bind to the adaptors in each single-stranded clone of the library. At each cycle, four labelled nucleotides (A, T, C and G), each given a specific fluorescent dye, compete for the binding to the first accessible nucleotide after the primer. The complementary nucleotide for each position will be more likely incorporated than the other three nucleotides, although errors can occur. Afterwards, four-colour imaging enables the identification of the dye, and associated nucleotide, that bound to a specific position. The ligation between the nucleotide and the dye is then cleaved and unbound nucleotides are washed away. The process, called cyclic reversible termination [47], is repeated until all the nucleotides in the sequence are identified. The sequencer software ensures correct association between dye colours and the fragment cluster at each particular position, since it is a paralleled process happening in all clusters at the same time [47].

**Figure 2.3: Illumina® sequencing procedure**

In the Illumina® sequencing procedure, modified nucleotides with fluorescence dyes are incorporated by polymerases at each cycle. Four-colour imaging after unincorporated nucleotides being washed away enables the identification of the nucleotide. The process is repeated in cycles until the whole sequence is determined. Adapted from [47].

**Alignment to a reference genome/transcriptome**

Resulting reads may be aligned to a reference transcriptome or genome [46]. Although aligning to a reference transcriptome is usually faster, it does not enable the discovery of new transcripts, which is relevant in the case of studying splicing [44, 46]. Therefore, transcripts containing exon/exon junctions that are not annotated in the reference transcriptome will not be mapped. One common strategy used by alignment softwares is to map reads first to a reference transcriptome and then map the unmapped reads to a reference genome, allowing the identification of novel splice sites [44, 46].
The assembly of reads may be quite challenging because the algorithm needs to deal with indels and mismatches in the sequence, as well as possible sequencing errors. The reference is usually the genome/transcriptome of one individual, meaning that it may differ from the studied sample and, consequently, the assembly task is to map reads to the best region of the reference, even if there is not a perfect match [49].

The usage of the paired-end method provides more accuracy to the assembly process because the distance between both adaptors is known, therefore increasing probability of finding the correct position of the read compared to the reference.

### 2.1.1 Gene expression quantification

Raw RNA-Seq data are, in general terms, a list of millions of short read sequences, to whom a quality score is associated [49]. The quality score is a function (-log₁₀) of the probability of error occurrences by the sequencing method, p, which should not exceed 10% [44]. FASTQ files coming from each short read of high throughput sequencing contain plain text with a sequence identifier, an optional description, the sequence itself (where N denotes an unknown nucleotide) and the quality scores [50]. Gene expression is associated with the counting of reads mapping to a given gene. However, for comparative analysis of gene expression, both between samples and within the same sample, normalization needs to be performed to reduce the intrinsic bias of read counts.

Considering the RNA fragmentation needed to gain coverage during the sequencing, the final number of reads attributed to a single gene is proportional to the number of transcripts produced by that gene (the gene expression layer), multiplied by the gene length (since longer genes will have more reads mapping to them) [51]. This means that reads mapping to a gene need to be normalized considering the gene length from which they were originated. This is a matter of within-sample normalization, since it affects comparisons between different genes coming from the same sample [49].

Moreover, raw read counts are biased for the sequencing depth. Although the expected scenario would be that genes accumulate reads proportionally to the library sizes, the fact that a low number of genes whose expression is several orders of magnitude above the mean levels of expression accumulate most reads compromises precision in measuring expression for lowly expression genes [49]. This effect must be considered such that between-sample normalization corrects this bias to a faithful comparison between samples sequenced with different depths [44,49].

Reads per kilobase of exon model per million mapped reads (RPKMs) are one widespread measure for normalized read counts. RPKMs are obtained, as explained in Equation 2.1, by taking the total number of reads assigned to a given gene and divide them by the length of the gene (in kilobases) times the total number of mapped reads (in millions). This normalized read count ensures faithful analysis of gene expression, without the bias for longer genes or samples sequenced at a deeper level [46,52].

\[
\text{RPKM}_{\text{gene } A} = \frac{\text{number of reads for gene } A \times 1 \text{ kb}}{\text{length of gene } A \times 1 \text{ M}} \times \frac{1}{\text{total number of mapped reads}} \quad (2.1)
\]

---

However, RPKMs do not take into account the ambiguous mappability of reads to the reference. Hence, some authors [53] proposed a correction method for multiple mapping that disregards reads mapped to a gene that are not uniquely mappable. Correction for mappability is obtained by only considering the portion of the gene sequence with uniquely mappable positions - the "effective length" [53]. These normalized gene counts (Equation 2.2) are then comparable between different genes within the same sample and between distinct samples, enabling a more accurate analysis of gene expression.

\[
c_{\text{RPKM}}_{\text{gene } A} = \frac{\text{number of reads for gene } A \times 1 \text{ kb uniquely mappable positions in gene } A}{\text{total number of mapped reads}} \times 1 \text{ M} \tag{2.2}
\]

### 2.1.2 Splicing quantification

Splicing quantification depends on the profiling of RNA-Seq reads aligning to splice junctions. In this work, splicing was analysed based on the exon-centered approach, where each potential inclusion of an exon in a final transcript is called an event and the exon is treated as an alternative exon.

Each event is defined by the genomic coordinates of its defining splice sites. Importantly, this definition is used for quantification of most of the types of alternative splicing events in Figure 1.5. Exceptions are intron retention events, whose quantification requires extra information associated with the intron body and more complex AS events, where some adjustments are needed to cope with combinations of different splice site possibilities.

For each event, the percent-spliced-in (PSI) is obtained as the ratio (Equation 2.3) between the reads supporting the inclusion of the exon and the total number of reads attributed to that event, that is the sum of reads supporting inclusion and reads supporting exclusion, as explained in Figure 2.4 [15, 16].

\[
\text{PSI} = \frac{\text{inclusion reads}}{\text{exclusion reads} + \text{inclusion reads}} \times 100 \tag{2.3}
\]

Since the number of reads mapping to both ends of the exon is eventually different, inclusion reads are taken as the mean between both sets of reads supporting inclusion of the alternative exon (both C1-Alt and Alt-C2 as explained in Figure 2.4). Therefore, quantification of the percentage of exon inclusion (PSI) is done as in Equation 2.4. Constitutive exons tend to have many more reads supporting inclusion and consequently its PSI values will tend to 100, while alternative exons will have more variable PSI values, near zero if seldom included [15, 16].

\[
\text{PSI} = \frac{\text{Mean}(C1-\text{Alt};\text{Alt-C2})}{(C1-C2) + \text{Mean}(C1-\text{Alt};\text{Alt-C2})} \times 100 \tag{2.4}
\]

### Vertebrate Alternative Splicing and Transcription Tools (VAST-TOOLS)

Vertebrate Alternative Splicing and Transcription Tools (VAST-TOOLS)\(^2\) is a software package for characterization and comparison of alternative splicing from RNA-Seq data [15]. VAST-TOOLS is able to align reads and provide both cRPKM values for gene expression and PSI values. The authors recommend

\(^2\)VAST-TOOLS: https://github.com/vastgroup/vast-tools
Figure 2.4: **RNA-Seq reads used in quantification of exon inclusion**
Percent-spliced-in (PSI) quantification (in Equation 2.4) is obtained by comparing the number of reads that align to each exon/exon junction defined for an alternative splicing event. Reads supporting inclusion of the alternative exon are reads spanning the constitutive (C1 or C2) and the alternative exons (Alt), while reads supporting exclusion are reads aligned to junctions from both C1 and C2 [15, 16].

RNA-Seq experiments using a minimum of 70 million reads to ensure alternative splicing analysis are not biased to highly expressed genes [15]. In the present work VAST-TOOLS was used to obtain cRPKMs and PSI values for the public RNA-Seq datasets analysed.

**Mixture of isoforms (MISO)**

Mixture of isoforms (MISO)[3] uses a probabilistic model to estimate isoform abundances from RNA-Seq data. MISO models the probability of reads being originated by a specific isoform and infers the isoform ratios based on this [54]. Percentages of exon inclusion calculated by the MISO pipeline were also addressed in this work.

### 2.2 Datasets

This work made use of publicly available RNA-Seq datasets of neuronal differentiation both in human and mouse. Moreover, another dataset (MultiTissue), part of the lab’s resources and provided by collaborator Manuel Irimia, has been used to identify the ortholog alternative splicing events between human and mouse.

---

### Table 2.1: Publicly available datasets analysed

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Species</th>
<th>Samples</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPC 46C</td>
<td>Mouse</td>
<td>5 NPC samples from 5 timepoints</td>
<td>Sox1-GFP reporter cell line of neural progenitors (46C)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Differentiation days: 0, 1, 3, 16 and 30</td>
<td></td>
</tr>
<tr>
<td>Hubbard [55]</td>
<td>Mouse</td>
<td>31 samples from 8 timepoints</td>
<td>ESCs differentiation into glutamatergic cortical neurons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Days in vitro: -8, -4, 0, 1, 7, 16, 21, 28</td>
<td>ESCs, NeSCs, Radial glia, developing neurons</td>
</tr>
<tr>
<td>Linares [21]</td>
<td>Mouse</td>
<td>30 samples CRISPR Cas9 deletions and siRNAs</td>
<td>Neuronal differentiation of ESCs into motor neurons</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ESCs, NPCs and motor neurons</td>
</tr>
<tr>
<td>Goff [56]</td>
<td>Human</td>
<td>20 samples</td>
<td>NPC differentiation timecourse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Days: 0, 1, 2, 4, 5, 11 and 18</td>
<td></td>
</tr>
<tr>
<td>MultiTissue</td>
<td>Mouse</td>
<td>65 samples from 38 tissues or cell lines</td>
<td>Resource from NMorais lab</td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>52 samples from 30 tissues or cell lines</td>
<td></td>
</tr>
</tbody>
</table>

### 2.2.1 Human datasets

Sauvageau, Goff and colleagues\(^4\) have used high throughput sequencing to study human neuronal precursor differentiation in order to address the expression profile of long non-coding RNAs and test their impact during brain development [56]. Human neural stem cells were cultured and neuronal differentiation was induced. RNA extraction followed by high throughput sequencing in different timepoints led to RNA-Seq data that therefore reflect a timecourse of neural progenitor cells’ differentiation \textit{in vitro}. Details about this dataset are summarized in Table 2.1.

### 2.2.2 Mouse datasets

A study from Hubbard and colleagues [55] provided RNA sequencing data profiling deep transcriptional changes during differentiation of mouse embryonic stem cells into a population of glutamatergic cortical neurons (see Table 2.1 for more details). Cells were cultured from eight days before induction of neuronal differentiation until 28 days after and neurons were harvested at seven different timepoints. For each timepoint, high throughput sequencing was performed using Illumina HiSeq\(^\circledR\) 2000 with paired-end library preparation. This procedure led to a set of samples that ranges from embryonic stem cells to different stages of developing neurons [55]. Most of the biological questions addressed in this work used or at least started being tested in this dataset, because the wide range of developmental stages available, the number of replicates in each stage (at least three) and the number of different timepoints (eight) all contributed to a better interpretation of results, after which decisions could be made in relation to the analysis of other datasets.

Linares and colleagues [21] explored the role of the splicing factor Ptbp1 in controlling alternative splicing during neuronal development. ESCs were differentiated into NPCs. RNA was extracted from ESCs, NPCs and motor neurons and paired-end library were constructed followed by high throughput sequencing using the Illumina HiSeq\(^\circledR\) 2000. This dataset also contains ESC, NPC and motor neuron samples corresponding to small interfering RNA transfections targeting Ptbp1 and/or Ptbp2, interesting to address the regulatory role of this splicing factor during mouse neuronal differentiation.

Also, a mouse neuronal differentiation dataset which is part of the lab’s resources was addressed.

---

\(^4\)Gene Expression Omnibus (GSE49581)
Although these cells are embryonic stem cells that undergo neural differentiation, the denomination of neural progenitors (NPC 46C) will be used to refer to this cell line on the following sections of this document, consistently to their differentiation into the neural phenotype. These data respects to the Sox1-GFP reporter cell line NPC 46C, which is commonly used as a model for studying mouse neuronal differentiation in vitro. NPC 46C cells report the entering of cells into the neuronal commitment phase by fluorescence. These data was available through a collaboration with Dr. Ana Pombo, from the Max Delbrück Center for Molecular Medicine, Berlin. Details about this dataset are in Table 2.1.

2.2.3 MultiTissue dataset

The MultiTissue dataset is part of the lab' resources and consists of a set of RNA-Seq experiments analysed using the VAST-TOOLS pipeline (section 2.1.2), therefore resulting in gene expression data quantified as cRPKMs and alternative splicing events as PSI values. These samples form a collection of RNA-Seq experiments obtained after extraction of RNA followed by high-throughput sequencing. These data was widely used for comparison of theoretical ortholog alternative splicing events between human and mouse and for testing a PSI quantification method taking into account the different number of reads supporting inclusion/exclusion of alternative splicing events. Although this dataset contains 38 different human tissues/cell lines and 30 mouse tissues/cell types, only tissues present in both datasets were used, as pairwise comparisons. The tissues used in this work are summarized in Table 2.2

Table 2.2: Tissues used to study ortholog AS events between human and mouse (MultiTissue dataset)

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Mouse</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrenal</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Brain</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Colon</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ESC</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Heart</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>iPS</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Kidney</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Liver</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Lung</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Muscle</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ovary</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Placenta</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>PREC</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Testis</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

2.3 Statistical analyses

As focused on previous sections, although RNA-Seq provides a more accurate way for transcriptome profiling, it is intrinsically biased due to technical issues related to the way the sequencing is performed but also to gene expression patterns themselves. Moreover, both gene expression and alternative splicing quantifications may sometimes result in "not available" (NA) values, in the cases where specific transcripts
were not covered by the sequencing step, because their expression levels were null or too low for detection, under the sequencing depth used.

Given that transcriptomic data, after alignment and normalization, consist in huge tables that are not humanly possible to interpret feature by feature (e.g.: gene by gene), statistical methods are required for extracting interesting information from them. Importantly, the application of statistical tests and subsequent analysis of results is to be critically interpreted by the investigator, under the light of the biological question that is being addressed.

2.3.1 Principal component analysis

Principal component analysis (PCA) is considered to be the oldest multivariate technique [57] and is used to extract the most of variance out of data when multidimensionality hampers interpretation of the relation between variables. By taking into consideration all the variables and correlations between them, principal components are obtained as linear combinations of the original variables [57, 58], reflecting an orthogonal projection of data into a reduced dimensional space, in which the variance of projected data is maximized [59]. PCA is intimately related to the linear algebra problem of single value decomposition, since evaluating the mean and covariance of variables is dependent on the finding of eigenvalues and eigenvectors [59]. Principal components accommodate the variance of data in a descending order, meaning that principal component one is the direction along which the data shows the largest variance, followed by the second component, and from that on.

Figure 2.5 depicts an example of a PCA using only gene expression data for two genes: Nanog, a pluripotency marker, and Ascl1, a marker for the neuronal lineage. Left panel of Figure 2.5 shows the plot of the expression of both genes in a mouse dataset of neuronal differentiation, each plot representing one sample. Unsurprisingly, these two genes allow a separation between pluripotent cells and neurons, with pluripotent/multipotent cells (ESCs and NeSCs) expressing higher levels of Nanog and lower levels of Ascl1, while radial glial and neurons in the first developmental stages (DS12) express high levels of Ascl1 but low levels of Nanog. Moreover, principal components (PCs) axis depicted show the directions accumulating most variance.

Right panel in Figure 2.5 illustrates the utility of PCA to address variance of data, since most of the separation which is clearly dependent on the different expression of both genes in the previous plot is projected onto the first PC, while the second PC separates high and low expression genes. This example is also useful to understand PCA in geometrical terms. Since both Ascl1 and Nanog are related with the separation of neuronal differentiation samples, PC1 obtained is a linear combination with contributions from both genes’ expression, accumulating most of the variance in the data points. In addition, in the particular example of Figure 2.5, dimensional reduction is not achieved, as there are still two informative dimensions, although the orthogonal transformation into two new uncorrelated directions (PCs) enables better extraction of information.

PCA is used in different fields as an exploratory tool to reduce the dimensions in data analysis. In the context of this work, PCA was used to explore the gene expression and AS patterns and extract from those variables the ones that better separate samples, helping to highlight the genes/AS events that better explain
Principal component analysis

Example of a PCA (principal component analysis) applied to gene expression levels of two genes – Nanog and Ascl1 – in a mouse dataset of neuronal differentiation. Each point in left panel is the correspondence between the expression levels (cRPKMs) of both genes. Right panel shows principal components for the same data, projecting the differential gene expression separation onto the first principal component axis, while the second principal component reflects high versus low expressed genes. Legend: DS - Developmental stage

timepoint differences during neuronal development. For performing a PCA, tables containing samples as rows and variables to be considered as columns were first centred around zero using the stdize function. This standardization function also allows scaling of each variable, a transformation that results in all variables having unit variance. When variables are scaled, high and low variances across samples will have similar weights in the PCA, while unscaled variables will account for differences in variables’ ranges. The PCA function from the FactoMineR package was used to obtain principal components and the variables needed for their plotting.

2.3.2 Filtering

Missing data

One of the most common problems when dealing with gene expression or alternative splicing quantifications is the presence of missing data (NA values), for some genes or AS events. Although one possible way to deal with NA values is to simply remove the genes/events containing at least one, in some cases this may be counter-productive, since usually a high proportion of the data contains at least one NA, and the analysis will end up with much fewer data than expected and relevant information eliminated. Because of this, in each analysis performed in this work, NA occurrences were counted and its proportions were calculated in comparison to the whole universe of that variable. After this, a threshold was defined as the maximum proportion of NAs allowed for each variable, depending on the analysis performed. In any case, the maximum allowed proportion of NAs was 0.2.
Variance

The variance expresses the mean of the squared deviations to the mean of a distribution of values, providing a measure for variability [60]. The alternative splicing events with a PSI close to the upper limit – 100 – or to the lower limit – 0 – across all samples, correspond to exons that are either frequently or rarely included. Moreover, an AS event that does not vary much across samples, even if the PSIs are in the medium levels (say around 50), would not be interesting since it would represent a similar pattern of AS across all samples. Given that this work’s interest is in differences between samples, filtering for general low varying events was also performed before any analysis related to alternative splicing quantifications.

Furthermore, when the question driving this work was related to the progressive inclusion or exclusion of exons along neuronal differentiation, filtering was performed to ensure the difference between the maximum and minimum PSI value for each event, across samples – PSI range – was above a certain threshold. The threshold considered in this particular analysis was 20, meaning that for this case, only PSIs varying more than 20% will be considered. This value was selected after exploratory analysis, namely by plotting, for each AS event, the PSI range against the logarithm of variance, in order to select events that vary substantially across samples.

Log-transformation

Gene expression values, after normalization for gene length, library size and correction for mappability, cover several orders of magnitude and, since many statistical tests assume normality of data, this should be taken into consideration. Accordingly, the variation in gene expression comes from multiplicative effects, contrarily to what happens in other types of variables, in which variation is explained by additive factors [61,62]. Logarithms are therefore more adequate to deal with this type of data, as they transform products into sums, converting the multiplicative scatter of log-normal into additive scatter (normal distribution) [61, 63]. Hence, log-transformation of cRPKM values was often applied when performing correlations or when plotting gene expression values versus some other normal-distributed variables, since log-transformed values are normally distributed [61–63].

2.3.3 Statistical tests

Scientific reasoning should be supported on the conceiving of a specific question and further design of a context that enables to statistically test the underlying hypothesis. Thus, application of statistical tests and interpretation of their results helps the decision-making and judging process of either considering the hypothesis to be true or decide that there is no evidence to say so [61]. Indeed, statistical tests are based on the rigorous definition of a null hypothesis, usually representing the case where nothing is happening. For instance, when testing the correlation between two variables, the null hypothesis should be that they are uncorrelated. In parallel, an alternative hypothesis should also be defined, representing the hypothesis under test, that is, in the correlation example, that the variables are actually correlated. However, careful and critical interpretation of statistical tests results is required.
Significance

The goal of a statistical test is to determine how different a dataset is from what should be expected under the null hypothesis. Statistical tests have an associated metric for the relation between variables that is being addressed and also a probability value (P-value). Evaluation of the outcome of a test should be done taking into consideration the P-value, which reflects the probability of having the same or a more extreme result, by random sampling, if the null hypothesis is true. Indeed, the P-value is associated to the likelihood of getting the same results just by chance. By defining a threshold of significance, that is the probability from which you decide to risk that correspondent results are very unlikely to have been generated by chance, and therefore evidence supports the rejection of the null hypothesis. The significance level used is dependent on the type of error believed to be more misleading to the analysis. For instance, a false positive occurs when results lead to the rejection of the null hypothesis when it is true, while a false negative occurs when there is really a deviation from what would happen at random, but decision is to not reject the null hypothesis. In the current work, false positives were to be avoided and significance levels of 0.05 were used, meaning the probability of rejecting the null hypothesis while actually being true is merely 5%, although frequently much lower P-values were taken into consideration [63].

Correction for multiple testing

Moreover, when a statistical test is performed a number of times, there is a probability of getting significant results just by chance, because if all null hypothesis under testing were true, P-values were expected to follow a distribution of random points between zero and one [61]. If the interest is to find within the data a bulk of variables for which there is significant evidence to reject the null hypothesis, P-values need to be corrected for the number of comparisons, as there will be occurrences of low P-values by chance [63]. In the current work, the false discovery rate (FDR) was used for the purpose [61].

Correlation analysis

The main hypothesis test used in this work was a correlation test, used to search for the strength of the relationship between variables such as PSI values, gene expression levels or time (days after induction of differentiation). The results of a correlation test are expressed as correlation coefficients and associated P-values, which, as previously referred before, should guide the decision to either reject the null hypothesis (that there is no relationship between variables) or not. Correlation tests calculate a correlation coefficient, dependent on the covariance and standard deviation of two variables, which follows a distribution in the interval $[-1, 1]$. Negative values for correlation coefficients correspond to inverse relations (sometimes called anti-correlations), while positive coefficients reflect positively related variables (the increase of the first variable is accompanied by the increase of the other) [60, 63].

Two different correlation methods were used in this analysis. Pearson's correlation coefficient reflects monotonic, linear relations between variables, meaning that it is suitable for testing associations between similar-distributed variables. On the other hand, Spearman's correlation coefficient is processed by considering the variables as ranks, and therefore good correlations correspond to coherent variations in
the values, without reflecting a linear regression. Spearman’s correlation was used whenever the two variables into consideration did not follow similar distributions, as when comparing PSI values with time or with cRPKMs [63]. Importantly, correlation tests should be performed in order to search for evidence of relationship between values, but good correlation coefficients (either positive or negative) and agreeing significant P-values should not claim per se a cause-effect relation between variables, which should be explored in mechanistic terms, considering the question under analysis [61,63].

2.3.4 Beta distribution

The Beta distribution, defined by two shape parameters, is used to model phenomena with values that are constrained to the interval [0, 1], being widely applicable in modelling probabilities or proportions. As depicted in Figure 2.6, the shape of the Beta distribution is dependent on the shape parameters, $\alpha$ and $\beta$, with both $\alpha, \beta > 0$. These are called hyperparameters [59] because they control the mean of the distribution.

![Figure 2.6: Shape parameters effect in Beta distribution](image)

Left panel shows density functions for different Beta distributions, while right panel shows a set of 500 randomly generated points following each of the same distributions. As the difference between the parameters increases, so does the distribution’s skewness. The $\alpha$ and $\beta$ parameters can be seen as proportional to the evidence (i.e. read counts) for inclusive and exclusive isoforms respectively, making the Beta distribution adequate for modelling PSI values.

If $\alpha$ and $\beta$ are the same and $\alpha, \beta > 1$, the Beta distribution is unimodal and symmetric, with the specific case of $\alpha = \beta = 1$ making the distribution become Uniform(0,1). The ratio between $\alpha$ and $\alpha + \beta$ defines the mean value of the distributions, while increase in the values of the parameters turns the distribution more peaked. That is, for the same ratio between parameters, the distribution gets more peaked when the magnitude of $\alpha$ and $\beta$ increase. Moreover, the ratio between parameters shifts the mean to either one of the sides. However, if $\alpha$ and $\beta$ are similar but $\alpha, \beta < 1$ (seagreen distribution in Figure 2.6), the function is "U-shaped". The Beta distribution will be left-skewed (the expected value shifted to the right) if $\alpha > \beta$, and it will be right-skewed when $\alpha < \beta$.

In general, higher values for the shape parameters lead to more peaked distributions and the distribution’s mean value is more sensitive to alterations in the parameters when these are higher. The shape of the Beta
distribution is therefore determined by the compromise between the magnitude of the values and the ratio between $\alpha$ and $\alpha + \beta$.

Interestingly, this density function is appropriate for modelling PSI values when studying alternative splicing, because its mean is the ratio between $\alpha$ and $\alpha + \beta$, which is directly comparable to the PSI quantification, obtained as the ratio between inclusive reads over all reads (see Equation 2.3). Moreover, when quantifying PSI values, the more inclusive/exclusive reads there are, the higher is the confidence for the PSI estimation, what is consistent with a narrower Beta distribution. This coherence makes the Beta distribution a natural choice for fitting sets of inclusive and exclusive reads, enabling alternative splicing quantification, as the obtained results demonstrate.

The Beta was used, in the context of this work, to extrapolate unique PSI values when several biological replicates show distinct PSI values and different evidence for inclusive and exclusive read counts. Taking into consideration the practical equivalence between $\alpha$ and $\beta$ parameters and inclusion/exclusion read counts, it is clear that the number of reads will influence PSI the same way the shape parameters influence the shape of Beta distribution. However, fitting distributions to read count data requires robust optimization methods. In this work, using `fitdistr` function as the algorithm for finding the shape parameters did sometimes resulted in undefined parameters, since the optimization procedure was not able to converge in a given number of iterations, when initial shape parameters were $\alpha = \beta = 1$. Likewise, the method of moments was used to provide initialization parameters that will favour convergence of the optimization procedure.

**Method of moments**

The method of moments was first proposed by Pearson in the late nineteenth century [63] and establishes that the estimation of parameters with respect to a distribution can be obtained by defining equations relating the moments of the distribution to the mean and variance of the sample of points to fit [63,64]. The method works by constructing as many equations as the number of parameters to be estimated. Hence, in the specific case of estimating $\alpha$ and $\beta$, only two equations are required. Considering the definition of uncorrected moments, as the $i^{th}$ uncorrected moment of a distribution being $\mu'_i = E[X^i]$ [63], the first moment of the Beta distribution is the mean, $E(X)$, and the second is $E(X^2)$, as defined in Equation 2.5 and Equation 2.6, respectively [63,64].

$$\mu'_1 = E[X] = \frac{\alpha}{\alpha + \beta} \tag{2.5}$$

$$\mu'_2 = E[X^2] = \frac{(\alpha + 1)\alpha}{(\alpha + \beta + 1)(\alpha + \beta)} \tag{2.6}$$

Given that the definition of variance is dependent on both moments $\text{Var}(X) = E(X^2) - (E(X))^2$, it can also be defined in terms of $\alpha$ and $\beta$. Likewise, by rearranging the two defined equations, the parameters can be found by setting the sample mean and variance equal to the distribution’s mean and variance. Two equations are then obtained, which written in order to the shape parameters enable the definition of initialization constants for the optimization of the fitting to more likely converge. The final equations used to do so are Equation 2.7 and Equation 2.8, where $\overline{X}$ is the mean value and $S^2$ the variance of the sample.
\[ \alpha = X \left( \frac{X(1 - X)}{S^2} - 1 \right) \] (2.7)

\[ \beta = (1 - X) \left( \frac{X(1 - X)}{S^2} - 1 \right) \] (2.8)

### 2.3.5 K-means clustering

Clustering algorithms facilitate the unsupervised grouping of closely related variables, considering a given metric for distance. K-means is an algorithm for finding K groups of variables which are similar within each group when compared to the other groups, using a given coefficient for measuring distances. Given a group of data points distributed in a multidimensional space, our goal is to partition data into a number, K, of clusters, in such a way that data points belonging to each cluster show inter-point distances smaller than the distances to points in outside clusters [59]. Importantly, the distances calculated may not be actually euclidean distances but other similarity metrics, such as correlation coefficients. For instance, in the context of this work, K-means algorithm was used to cluster PSI event that evolved with a similar trend along neuronal differentiation. For such particular case, Spearman’s correlation was the more adequate metric to use, since highly correlated events would correspond to similar pattern of evolution along development.

![Figure 2.7: K-means clustering algorithm](image)

Figure 2.7: K-means clustering algorithm

K-means algorithm starts by randomly attributing K centres for datapoints (1). Afterwards, the chosen metric for distance is calculated between each point and each centre (2). Each point is then attributed to the closest centre (3) and new centres are re-calculated (4). The process is repeated until convergence, that is, until new calculation of distances and consequent assignment of new centres does not change the previously defined ones. Adapted from [59].

K-means algorithm works as explained in Figure 2.7, after a given metric chosen to compare variables. In the case of the current work, the metric was Spearman’s correlation coefficient. K-means algorithm is
processed in the following way:

1. Define number of clusters (K) and randomly set K cluster centres (“means”).
2. For each point, calculate the distance (based on the selected metric) to each centre: associate each point with the closest centre.
3. Recalculate cluster centre as the mean of points assigned to that cluster.
4. Repeat until convergence, i.e., until there is no change in cluster centres.

**Finding the best number of clusters**

One common problem when using unsupervised clustering algorithms is the fact that no relation is defined *a priori* between variables, and therefore it is not known how many clusters there should be in the data. Moreover, as explained before, K-means and other clustering algorithms rely on knowing how many clusters should the data be arranged in. Therefore, methods for choosing the best number of clusters, based on the data points, have been proposed and associate a cost function that evaluates the assignment of each data point to the cluster assigned. The silhouette width method was used in this work to find the best number of clusters. This method, as others used to find an optimal number of clusters, iterate the number of clusters and compute K-means for each defined K. Afterwards, the cost function is computed and it is possible to evaluate which K provided the best values for the cost function [65]. Silhouette width measures how well a data point fits into its cluster by calculating the within-cluster dissimilarity, $a_i$, and the minimum of average between cluster dissimilarity, $b_i$. The silhouette width is consequently defined as in Equation 2.9.

$$\text{sil}_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (2.9)$$

The silhouette width method works the following way:

1. Calculate average within-cluster dissimilarity ($a_i$).
2. Calculate minimum of average between-cluster dissimilarity ($b_i$).
3. Calculate silhouette width, $\text{sil}_i$.
4. Choose K in order to maximize the average of $\text{sil}_i$.

**2.4 R statistical software**

All the analysis were performed using R software environment for statistical computing ⁵, that provides a wide range of statistical and graphical techniques, while being easily extensible. R was used on most steps of data analysis in this project, from filtering of tables coming from aligners until the production of final plots showed in this work. Moreover, RStudio, an integrated development environment for R language ⁶ is a user-friendly open-source software, including tools as console, syntax editor supporting direct code execution and a number of other applications, including plot, help and debug tools. A summary of functions and packages used in the analysis is presented in Table 2.3.

---

⁵The R Project for Statistical Computing: https://www.r-project.org/
⁶RStudio: https://www.rstudio.com/products/RStudio/
<table>
<thead>
<tr>
<th>Function</th>
<th>Package</th>
<th>Utilization</th>
<th>Purpose in this work</th>
</tr>
</thead>
<tbody>
<tr>
<td>stdize</td>
<td>R</td>
<td>Standardization (centering and/or scaling) of data matrix</td>
<td>Centering of data matrices, prior to PCA</td>
</tr>
<tr>
<td>PCA</td>
<td>FactoMineR</td>
<td>Performs principal component analysis</td>
<td>Perform PCA</td>
</tr>
<tr>
<td>p.adjust</td>
<td>stats</td>
<td>Adjusts P-values in multiple comparisons</td>
<td>P-value correction (FDR method)</td>
</tr>
<tr>
<td>cor.test</td>
<td>stats</td>
<td>Performs a correlation test</td>
<td>Correlation test (both Pearson’s and Spearman’s)</td>
</tr>
<tr>
<td>fitdistr</td>
<td>MASS</td>
<td>Fitting of univariate distributions</td>
<td>Fit set of points into distribution (Beta distribution)</td>
</tr>
<tr>
<td>silhouette</td>
<td>cluster</td>
<td>Computes silhouette in k clusters</td>
<td>Finding the best number of clusters</td>
</tr>
<tr>
<td>Kmeans</td>
<td>amap</td>
<td>Performs k-means clustering in a data matrix</td>
<td>K-means clustering (by correlation)</td>
</tr>
<tr>
<td>Dist</td>
<td>amap</td>
<td>Computes distance matrix using a specified distance measure</td>
<td>Compute distance matrix based of Spearman’s correlation</td>
</tr>
<tr>
<td>colorRampPalette</td>
<td>RColorBrewer</td>
<td>Provides colour schemes for graphics</td>
<td>Customization of colour sets</td>
</tr>
<tr>
<td>ggplot</td>
<td>ggplot, ggplot2</td>
<td>Graphics for data analysis</td>
<td>Graphical customization</td>
</tr>
<tr>
<td>geom_label_repel</td>
<td>ggrepel</td>
<td>Text and labels customization using ggplot</td>
<td>Graphical customization</td>
</tr>
<tr>
<td>try</td>
<td>base</td>
<td>Wrapper to run an expression that might fail</td>
<td>Prevent optimization convergence problem to abort the whole code run</td>
</tr>
</tbody>
</table>
Chapter 3

Results and Discussion

The following sections aim to trace the sequence of biological questions that were posed and further explored, based on the data available for the deployment of the present work. The project started after a revision of the literature supporting that alternative splicing plays a major role in fine-tuning the biology of the nervous system and its patterns seem to distinguish organisms with higher functional complexity from others. Following this, human and mouse alternative splicing and gene expression data, in the context of neuronal differentiation, was analysed, with the motivation of unveiling new AS switching events during formation of nervous systems. Interestingly, some questions opened opportunities for exploration of unanticipated issues, which globally lead to the opening of new paths in the conceptual framework of the project. Some questions remain to be fully addressed and serve as a strong motivation for future work.

3.1 Alternative splicing quantification

This project made use of both human and mouse neuronal development transcriptomics data. Since alternative splicing events that have physiological relevance are in general more conserved across species, comparison between human and mouse alternative splicing orthologous events may provide important evolutionary information. Consistent with this, in the current project, analysis of alternative splicing events using mouse datasets were to be somewhat extrapolated to the human biology. For that reason, there was the need to directly compare gene expression patterns and, more relevantly, PSI values between human and mouse. NMorais Lab’s resources include a dataset of gene expression and splicing quantifications for several human and mouse tissues (subsection 2.2.3), including the inter-species correspondence of annotated orthologous alternative splicing events. Moreover, the Ensembl Project [66], which proposes to maintain automatic annotation of eukaryotic genomes, contains an online tool, Ensembl Biomart\(^1\), that was used to download tables containing the orthologous genes between human and mouse. Furthermore, resources from the Lab include the same type of correspondence produced for each of the annotated AS events that are part of VastDB\(^2\), a database associated with VAST-TOOLS (see section 2.1.2). Using orthology annotations, genes and AS events could be compared between human and mouse. Importantly, all the following analyses

\(^1\)Ensembl Biomart: http://www.ensembl.org/biomart/
\(^2\)VastDB: vastdb.crg.eu
that refer to specific annotated exons were accompanied by browsing the exons’ coordinates in the UCSC Genome Browser [67], a tool that enables dynamic visualization of each required region of the genome of a species, strongly clarifying interpretation of alternative splicing events, amongst other types of information.

It is consensual that tissue-specific gene expression patterns are conserved amongst vertebrates, i.e., more similar when comparing the same tissue in different species than when comparing different tissues from the same species [16]. Considering that less tissue-specific conservation is reported for alternative splicing, one of the first tasks in this work was to correlate PSI values between human and mouse orthologous AS events to determine those more likely to have homologous functional behaviour.

Likewise, by using the MultiTissue dataset (subsection 2.2.3) and matching tissues between human and mouse, correlations could be obtained for PSI values on the set of tissues under analysis. However, although the MultiTissue dataset contained some homologous and therefore comparable tissues/organisms, there were different numbers of samples between the two species for the same tissue. Considering this, there was the need to find a suitable method for pairwise comparison of a set of PSI values correspondent to each tissue, for each event.

3.1.1 Beta distribution in multiple sample splicing quantification

The Beta distribution empirically arose as an adequate function for modelling PSI values because it shows three main properties matching PSI quantification: a) it is restricted to the interval [0, 1], such as PSI values are, the same way as every proportion measure; b) the ratio between shape parameters defines the expected value of the distribution, the same way the ratio between inclusive and exclusive reads influences PSI values and c) the distribution becomes more peaked, well-defined, when shape parameters increase, as the probability of having a specific PSI value should increase as the evidence for inclusion and exclusion of the exon is stronger. Consequently, PSIs may be modelled by a beta-shaped distribution whose parameters, \( \alpha \) and \( \beta \), relate to the inclusive and exclusive read counts, respectively. The application of the Beta distribution to the quantification of PSI values from read counts was inspired by a developing component of the VAST-TOOLS pipeline (section 2.1.2) that provides a means to statistically quantify differences of PSI values taking into consideration different sample replicates and using the same principle to compare PSI values between two interest groups.

In order to further understand the consistency between PSI values and its modelling through this function, Figure 3.1 illustrates the effect of read counts in the definition of inclusion levels, as modelled by a Beta distribution. Both green and blue representative alternative splicing events have a percent-spliced-in of 80, resulting from the calculation of a proportion of reads supporting inclusion of 0.8. However, by generating Beta-shaped distributions (bottom panels) taking into consideration the differences in magnitude of the number of reads, very different distributions are obtained. The distribution associated to the green event is much more skewed than the blue one, as there is less evidence to support the \( \text{PSI} = 80 \) value. On the contrary, the blue event distribution is much sharper, since the higher values of read counts (and therefore, \( \alpha \) and \( \beta \) parameters) turn the distribution much less skewed. This reinforces that, when comparing biological replicates (in this case, samples considered to be from the same tissue), correction for non consistent

---

3https://genome-euro.ucsc.edu/
Figure 3.1: The Beta distribution shape differences in a model for PSI quantification

Beta distribution may be used to model PSI values because it empirically matches the characteristics of PSI quantification. Inclusive and exclusive read counts affect the skewness of the distribution, which is consistent with the PSI quantification being affected by less evidence on the number of reads. Both green and blue representative events show read counts that lead to a PSI of 80. However, an event with more reads supporting PSI is translated into a much sharper Beta distribution.

This principle was used for quantification of PSI values in tissues having more than one sample per tissue, in order to determine one global representative PSI value. For this, read counts supporting inclusion or exclusion for each event, which are part of the VAST-TOOLS splicing quantification files, were used for generating a conjugate Beta distribution that best fits the individual contributing distributions supported by each samples’s reads. Likewise, in the case of having \( n \) samples for a tissue, the procedure was the following:

1. For each of the \( n \) samples, get number of reads supporting inclusion and exclusion:
   
   (a) Generate a Beta distribution with a set of 500 random points (with the \( \texttt{rbeta} \) function in R) using number of inclusion/exclusion reads + 1 as parameters – there is the need to avoid zeros, since \( \alpha, \beta \neq 0 \) by definition of the Beta distribution function.

   \[
   \alpha = \# \text{ inclusion reads} + 1 \quad \beta = \# \text{ exclusion reads} + 1
   \]

   (b) Concatenate the \( n \) sets of points, therefore keeping and joining together less skewed distributions when read counts are high and more skewed when it is the other way around.

2. Calculate mean and variance for the concatenation of the sets of points.
3. Calculate equation for shape parameters of the whole distribution, using the method of moments estimators for parameters (section 2.3.4).

4. Try to fit a new Beta distribution, using \texttt{fitdistr} and providing estimators as initialization values.

5. If convergence is not achieved and consequently parameters of fitting are not provided, try different combinations of initialization values, obtained empirically by try and error experiences. Dividing both $\alpha$ and $\beta$ by $\frac{\alpha + \beta}{10}$ is tried first, then division by $\frac{\max(\alpha, \beta)}{200}$ is tried. Finally, $\alpha = \beta = 1$ is tested. Only events for which convergence is thereby achieved are considered.

6. Use shape parameters of the overall joint distribution to calculate the PSI value, as the following ratio:

$$\text{Global PSI} = \frac{\alpha_{\text{fit}}}{\alpha_{\text{fit}} + \beta_{\text{fit}}} \times 100$$

Figure 3.2: \textit{Comparison between Beta distribution and VAST-TOOLS PSI values}

Each point represents the correspondence of the mean PSI value from the samples as calculated using the general PSI quantification method and the Beta distribution method to account for differences in the number of reads. Points reflect 10000 AS events and are coloured by the variance within VAST-TOOLS PSI values, with lighter colours corresponding to higher variances.

The application of the Beta distribution to MultiTissue dataset resulted in new PSI quantifications for each of a pair of alternative splicing events annotated as orthologs between human and mouse, in each tissue containing more than one sample (considered biological replicates). Therefore, new tables were generated for both species, containing a unique PSI value for each event, per tissue. This method for determining a PSI value that takes into consideration a number of different samples, possibly with large differences in the number of reads accounting for inclusion and exclusion, should improve, at least in theory, the accuracy of the global PSI value since samples where the conventional PSI is supported by a low number of reads should influence the global PSI less than samples having much more evidence for the inclusion level obtained.

Although mean or median PSI values may be convenient for dealing with several PSI replicates, they do
not take into consideration the influence of the total number of reads supporting a PSI value. For visualization purposes, for each case in which the Beta distribution was applied, comparison between the global PSI and the mean of PSI values corresponding to the regular PSI calculation was performed. The resulting Figure 3.2 shows a general consistency in PSI values obtained with the Beta method and VAST-TOOLS as reflected by the linear trend, although there is a considerable number of differences suggested by the points apart from the diagonal. When PSI values are close to the extremes, zero and one, the Beta distribution will show more different results, since the expected value of the distribution would not turn out to be neither zero nor one, since it is incompatible with restrictions for the shape parameters. Accumulation of points in vertical lines likely reflects the cases where the number of reads is low for some samples and those are the situations where the Beta method is more interesting.

The result of applying the Beta distribution method described before into the quantification of PSI is illustrated for some events in Figure 3.3. Top panels correspond to three alternative splicing events in human brain samples, while bottom panels show similar examples in two human muscle samples. It is clear that variation between the individual values of each replicate influences the global PSI calculation but the main advantage of using the Beta distribution is the stronger impact of estimates supported by higher read evidence on the global PSI.

Figure 3.3: Beta distribution modelling of exon inclusion levels

Beta distribution as a model for PSI quantification aims to correct for differences in the number of reads supporting either inclusion or skipping of exons when taking several replicates. Top panels correspond to inclusion level quantification for three different exon skipping events in five human brain samples, whereas bottom panels illustrate three events in two muscle samples. Colored values indicate the parameters of each individual beta distribution generated, which are coherent with the number of inclusion and exclusion reads supporting the PSI value.
3.1.2 Functional orthologous alternative splicing events between human and mouse

This exercise was extremely valuable for better understanding splicing quantification and also served as a methodology pipeline to address the issue of summarizing PSIs from biological replicates with very different read coverage. Moreover, it was used to directly compare and correlate splicing levels between orthologous events in human and mouse, in order to infer about the conservation of a given interest event. For instance, a side project in the lab focused on the regulatory effect of PTBP1 in the inclusion levels of an exon involved in the inflammation processes related to ageing. Since the human event considered had an annotated mouse ortholog event and there was available data for knock-out of PTBP1 and splicing quantifications on those samples, the previously described procedure enabled direct comparison of inclusion of the exons between both species. The linear relation between PSI values for those orthologous events across comparable tissues is shown in Figure 3.4.

![Figure 3.4: Example of conserved ortholog human/mouse alternative splicing events across tissues](image)

Figure 3.4: Example of conserved ortholog human/mouse alternative splicing events across tissues
Skipping of an exon in the EXOC7 human/mouse gene is an example of an alternative splicing event conserved between both species. Points represent a pair of PSI values for each tissue/cell line.

3.2 Alternative splicing patterns during neuronal development and their regulatory mechanisms

The main focus of this project was to study alternative splicing events varying during neuronal development and explore their regulatory mechanisms and biological consequences. For addressing this question, available neurogenesis datasets were examined in terms of number of replicates and differentiation
Figure 3.5: Principal component analysis of gene expression data in Hubbard’s dataset
Gene expression data clearly differentiate neurogenesis samples within clusters of samples in the same developmental stage.

timepoints. For most of the questions proposed in this work, Hubbard’s dataset was used in the first place, because it contained the highest number of replicates and different timepoints, favouring statistical evaluation of results and therefore biological interpretation, although procedures were expanded to some of the other datasets whenever justified.

Exploratory principal component analysis of gene expression within the Hubbard dataset was performed and results are in Figure 3.5. As expected, samples are very similar within each developmental stage and the genes that contribute the most to the variance between groups certainly distinguish pluripotent cells from more mature neural progenitors, what is clear by the separation achieved through the first principal component, accumulating roughly 43% of the variance within the gene expression quantifications. The second component, which also accumulates a great percentage of data variance, is distinguishing genes related to the early neuronal samples from others, suggesting this may accommodate genes that are responsible for the entering of cell into the neural fate.
3.2.1 Splicing events with a specific neuronal-development pattern evolution

As mentioned before, VAST-TOOLS output files contain quantification of gene expression values (cRPKMs) and exon inclusion levels (PSI). These tables contain a high number of events/genes, which can only be interpreted by the human reasoning through statistics. With the motivation of exploring AS changes during neuronal development, PSI values corresponding to the Hubbard datasets were analysed in terms of the variance and PSI range for each event across tissues. This procedure aimed to determine a threshold for both minimum variance and PSI range to consider only events that vary enough for inferring a biological consequence in neuronal development. Filtering after this analysis resulted in the consideration only of events whose PSI value varies more than 20% across all samples, removing low variance events and all events for which there was missing data in at least one sample.

After the filtering process, data were prepared to explore if there were events having their PSI value significantly increasing or decreasing along time of differentiation, without assuming a linear relation between PSI values and time in days after differentiation. Hence, Spearman’s correlation was tested for all events matching the minimal variance and PSI range criteria described before. Under the null hypothesis that there was no correlation of the PSI values with time, the alternative hypothesis of correlated variables was tested for all events. Since this was a case of multiple testing, obtained P-values were then adjusted using the false discovery rate correction method.

In this dataset, a great number of alternative splicing events were significantly correlated or anti-correlated with the number of days in culture under the differentiation protocol. Figure 3.6 shows four examples of events with the PSI value significantly (FDR < 0.01) correlated or anti-correlated with days in culture. The top event, corresponding to the skipping of exon 4 of *Tmem63b* gene, encoding for a calcium channel, transmembrane protein 4, shows PSI values close to 100 in the undifferentiated stages, but they drop down around day zero, the timepoint corresponding to the neural induction. Afterwards, PSI values approach zero in more mature stages of differentiation. This means that exon 4 is always included in pluripotent stem cells (PSI ≃ 100) but always skipped in more mature neurons. Since calcium channels are important for synaptic functions of neurons, this is a potentially interesting neuronal-specific AS event.

*Cacna1a*, for instance, encodes for calcium voltage-dependent channels that trigger the neurotransmitter release in synapses. Interestingly, *Cacna* gene family is known to generate tens or thousands of different proteins through alternative splicing of their multiple exons, resulting in functionally tuned calcium channels efficiently performing depolarization of the membrane. Moreover, pools of *Cacna1a* transcripts differ in different developmental stages and distinct neuronal types, reinforcing the role of this expanded AS mechanism in neuronal function [68]. In the example depicted in Figure 3.6, exon 44 of the *Cacna1a* gene is frequently skipped in pluripotent cells (PSI ≃ 0) but included in neuronal ones, where PSI values are increased up to ≃ 90.

The other alternative splicing events serving as examples for significant correlations with time have similar interpretations. *Dnm1l* exon 3 is progressively more included along differentiation, probably meaning it also has a role in neuronal functions, while *Ube2j1* is more often skipped in neuronal cells (PSI ≃ 40) than in ESC, with PSI values close to 100.

\[\text{http://www.uniprot.org/uniprot/} \text{ (Accessed in September, 2016)}\]
Figure 3.6: **Time-correlated alternative splicing events**
Examples of exons progressively more included or excluded along differentiation, selected after testing for Spearman’s correlation across several AS events in Hubbard’s dataset.

This type of dynamic changes across neuronal differentiation are motivating for the profile of alternative splicing patterns during differentiation, but lack a causal link that can be coupled to the evidence for correlation between variables. Taking this into consideration, the next natural step was to search for a link between the time-correlated AS events and the action of splicing regulatory proteins, explored in the next section.
3.2.2 Exploratory analysis of the regulatory mechanisms of neuronal development-specific splicing events

In order to further explore the regulatory mechanisms underlying the progressively included or skipped exons during neuronal development, correlation was tested between PSI values of time-correlated exons and gene expression levels of RNA-binding proteins that have already been proposed to be regulating AS events in the neuronal tissues.

As explained in section 1.1.3, the binding position of splicing regulators in the RNA with respect to the alternative exon vicinity influences the event outcome. Recent efforts in the functional characterization of RNA-binding proteins used an RNA compete assay to infer RNA-binding motifs for a number of RBPs. With this procedure, incubation of each RBP with all possible 9-nucleotide sequences of RNA, competing for binding preference, and further isolation of the binding complex provided a compendium of binding motifs, a valuable resource for studying RNA-binding regulated processes [69].

Considering this, the mapping of known RNA-binding motifs to AS events may be used to infer the putative regulatory role of the respective RBPs on time-correlated AS [70]. As a preliminary analysis, Pearson’s correlations between expression levels (cRPKMs) of RBPs with known RNA-binding preferred motifs and PSI values for the significantly time-correlated alternative splicing events were calculated.

This procedure led to the identification of a set of events, depicted in Figure 3.7, that show evidence to be relevant for neuronal differentiation and whose PSI evolution is significantly anti-correlated with the expression of Nova1. An alternative splicing event in the Axin1 gene is depicted in the top panel. Axin1 is a scaffold protein that interacts with GSK3β and β-catenin proteins promoting the activating of neuronal genes, and its subcellular localization controls the balance between neural progenitor proliferation and neuronal production. Moreover, Axin1 is essential for correct formation of axon and dendrite polarity and, consequently, neural connectivity. Axin1 deregulation during neuronal development is associated to brain malformations in mice, such as micro or macrocephaly [71]. Importantly, a mouse model with high levels of Axin1 resulted in autism-like behaviors. Furthermore, Axin1 interaction with β-catenin favour degradation of the former, acting as a tumour suppressor. In colorectal cancer samples, alternative splicing of exon 9 in Axin1 gene has been detected as alternatively spliced when comparing tumour samples of colorectal cancer [72].

Figure 3.7 shows that exon 9 of Axin1 is included in between a quarter to half of the isoforms in embryonic stem cells but its inclusion becomes even rarer when cells proceed to the neuronal lineage. This result may suggest that the skipping of exon 9 is associated with the functional alterations and subcellular localization that is needed for the timely promotion of transcription of neurogenesis genes. However, functional impact in the activity of the scaffold protein of the skipping of exon 9 is required for exploration of this hypothesis.

Although the reduction in the inclusion levels of exon 5 of Dip2a gene is modest (see Figure 3.7), this seems to be specific in timing terms, since alterations are notorious at day 8 in culture. This is still of relevant importance, since this gene encodes for the disconnected disco-interacting protein 2 homolog A, which is expressed in ectodermal-derived tissues during embryonic development. In mouse nervous systems, DIP2A acts as a receptor for signals involved in axon path finding and patterning. Moreover, this protein, broadly expressed in mouse brains, has been proposed to have a role in specification of synapses through
Figure 3.7: **Exon switches anti-correlated with Nova1 expression during neuronal differentiation**

Top panels show alternative splicing events that are progressively more skipped and whose PSI are anticorrelated with Nova1 expression levels (bottom panel) during neuronal differentiation.

Furthermore, 

Furthermore, *Mprip* exon 9 is progressively less included as cells differentiate into neurons, as shown by the decreasing of the PSI values in the top right panel of Figure 3.7. In neuronal cells, this protein is involved in neurite outgrowth and suggested to target myosin phosphatase to the actin cytoskeleton. Interestingly, the same alternative splicing event in the *Mprip* gene shows differential splicing patterns in post-mortem brains of myotonic distrophy type I, a multisystemic genetic disorder that in the central nervous system shows deficits in visual spatial and attention deficits or avoidant behavior, amongst other symptoms [74]. Although knock-out mice for *Mbnl1* (a splicing regulator protein) exhibit reduced inclusion levels for this particular event, the pattern is not fully recapitulated, which may suggest that other proteins may be involved in the AS event regulation, consistent to the potential role of NOVA1.
Figure 3.8: Exon switches correlated with \textit{nSR100} expression during neuronal differentiation

Top panels show alternative splicing events that are progressively more included and whose PSIs are correlated with \textit{nSR100} expression levels (bottom panel) during neuronal differentiation.

A similar analysis was performed for \textit{nSR100}, an RNA-binding protein with reported functions in promoting neural-specific exon inclusion in neuronal cells. Alternative splicing events that significantly correlated with time were tested for correlation with the expression values of \textit{nSR100}.

Figure 3.8 depicts three examples of time-correlated AS events in genes that have a function in shaping the nervous system, also showing significant correlation with \textit{nSR100} expression levels. The \textit{Camkk2} gene is expressed in the hippocampus, neocortex and cerebellum and is associated with hippocampal learning and
memory. Deletion of this gene in mice, causing the loss of the calcium/calmodulin-dependent protein kinase kinase 2 resulted in no apparent morphological alterations, but spatial long term memory was significantly delayed, when comparing to wild type littermates [75]. Interestingly, some isoforms of calcium/calmodulin-dependent protein kinase kinase 2, but not all, were needed for formation of long-term memory. The AS event is eventually affecting the ratios of these isoforms, and therefore playing a role in the function of long-term memory.

Also, exon 9 of Clip2 gene is progressively more included as cells differentiate, as it is clear in the middle top panel of Figure 3.8. The encoded protein, CAP-Gly domain containing linker protein 2, is part of the cytoplasmatic linker protein family, which mediates the interaction between membranous organelles and microtubules, with a role in the formation of dendrites 5.

Finally, the progressive inclusion of exon 17 in R3hdm1 is also potentially interesting, since this gene is known to encode a protein that binds to poly-A RNA, thus possibly having a role in transcription regulation. These examples represent only some of the exons for which there is a correlation with the expression of a splicing regulator and these linear relations motivate the test for evidence in causality between expression of a splicing regulator and the outcome in alternative splicing. Further analysis taking into consideration the mapping of RNA-binding motifs in alternative splicing events may provide more potentially relevant regulated AS neuronal-specific events, for which in vitro validation may then be addressed.

![Figure 3.9: Mapping RNA-binding proteins motifs in the vicinity of an exon skipping event](image)

Nomenclature used for the definition of regions where motifs for RNA-binding proteins are thought to be regulating the exon skipping event, defined considering constitutive (C1, C2) and alternative (Alt) exons.

Indeed, by using the information about the preferred RNA motifs for the binding of a given RBP together with the annotation of alternative splicing events, the lab produced resource tables where, for each RBP binding motif, occurrences of those motifs are mapped to the vicinity of alternative splicing exon skipping events. This mapping makes use of a code of eight different regions defined based on the AS events’ coordinates, as indicated in Figure 3.9. Moreover, as referred in chapter 1, the association between binding position and the splicing consequences is known for some RBPs, such as the case of NOVAs, PTBP1 and nSR100.

In order to explore the evidence for a mechanism of regulation explaining correlations between splicing regulators and exon inclusion levels described before, search for enrichment in preferred RNA-motifs for each RBP can be applied in the future. This method, which is already been applied in the lab for inference AS regulatory mechanisms, is based on statistical comparison between the number of motifs found in the

---

regulatory regions for each event correlated with the expression of the putative regulatory RBP.

For instance, NOVA1 protein is known to promote exon skipping when bound to the last nucleotides of the upstream intron or the first nucleotides in the exon itself (regions I12 and A1 in the representation of Figure 3.9). To search for evidence of a regulatory role, it would be necessary to compare, from the time-correlated AS events that are anti-correlated with NOVA1 expression (namely the ones in Figure 3.7) how many of them do have an enrichment in the NOVA1-associated RNA motif(s) in the expected region (true positives), in comparison to the ones without a significant enrichment in motifs but that are also significantly correlated (false positives). The events that are enriched in motifs for NOVA1 but do not show significant correlations should also be taken into consideration. The null hypothesis that nothing is happening, that is, in this case, that there is no evidence for a NOVA1-associated regulatory mechanism, should only be rejected if the set of events containing the NOVA1-associated motif in the expected region that are anti-correlated with the expression of NOVA1 is statistically significant when compared to the whole universe of time-correlated AS events.

nSR100 RNA-binding protein is known to promote the exon inclusion when binding in the last nucleotides of the upstream intron (region I12 in Figure 3.9) relative to its target exons. Hence, by using the results for the correlations between AS events PSI evolution along differentiation and cRPKMs of the expression of the RBP, inference about potentially nSR100-regulated events would be possible, if a statistical procedure similar to what described before is applied. This may be achieved using a Fisher’s test. Since the underlying hypothesis of regulation by NOVA1 or nSR100 requires that exons that are targets for a splicing regulator contain the proteins’ preferred binding motif in specific regions, true positives would be events enriched in motifs at that specific region and that are correlated with the RBP’s expression, while true negatives will be events that lack an enrichment in the motif but do not show significant correlation. On the other hand, false positives will be correlated events that lack the motif enrichment and false negatives would be events not correlated or lacking the motif. An example of contingency table for the application of this methodology in statistical inference of a regulatory role for nSR100 would be as in Table 3.1.

Another effect that should be taken into consideration in the exploration of a regulatory mechanism is the fact that time-correlated AS events that show a significant correlation with one of these RBPs, for instance, may also be correlated with other splicing regulators. In that case, the motif enrichment for both RBPs should be performed and decisions on either to reject the null hypothesis or not would be related to the strength of evidence in both cases. This type of procedures will be applied to search for evidence in regulatory mechanisms during the future ongoing of this project.
3.2.3 Alternative splicing events with similar pattern of evolution during neuronal development

Another question addressed during the project was whether splicing regulators controlling the inclusion rates of specific exons would have a specific mode of action in all their target exons. In other words, the hypothesis was that RNA-binding proteins regulating neuronal-specific AS events would influence the PSI in a similar pattern. In order to explore this hypothesis, K-means clustering algorithm was used to group together events with a similar profile of PSI values across neuronal differentiation.

The silhouette width metric was used to test what would be the best number of clusters, that is the number of clusters that maximizes the wellness of attributing each event to a given cluster. Importantly, the distance metric used was the Spearman’s correlation and therefore clusters of events accumulate the events whose PSI have a more similar pattern evolution along differentiation. Two clusters would provide the best value for the silhouette width, but would not be relevant if we consider the biological question, since events would be separated into increasing and decreasing PSI values. Following, by increasing the number of clusters, five clusters provided an increase in the silhouette width and would therefore be the best number of clusters, if \( K = 2 \) is excluded.

![Figure 3.10: K-means clustering by correlation of PSI profile along neuronal differentiation](image)

The result for the five clusters is depicted in Figure 3.10, where different PSI evolution profiles are clear. All clusters show a shift around day 1 from differentiation, suggesting that a great number of PSI values are changed after the start of neural induction. Moreover, by comparing, for instance, the orange and green clusters, it is clear that the first group of exons increase their inclusion after differentiation, while the latter group’s PSI decrease. Pink and blue clusters may also be interesting, since they show a more oscillatory profile. However, a deeper analysis is required for characterization of closely related profiles in the PSI values.
during differentiation and inference of the relationship each cluster has with splicing regulatory proteins.

### 3.3 Neuroprogenitor cell line 46C characterization

The neuroprogenitor 46C cell line (NPC 46C) is used to study neuronal differentiation in vitro, since these cells’ differentiation in feeder-free adherent monocultures recapitulate neuronal commitment under treatment with retinoic acid [28]. This cell line was obtained through a gene targeting technique in which a green fluorescent protein (GFP) was inserted in the reading frame of the Sox1 gene in a mouse embryonic stem cell line [76]. Sox1 is the earliest specific neural marker and therefore these cells are convenient for monitoring, quantifying or purifying of neural committed cells. RNA-Seq data from NPC 46C were available in the Lab, following a collaboration with Dr. Ana Pombo, from the Max Delbrück Center for Molecular Medicine in Berlin, and Dr. Sérgio Almeida, from the IMM.

![Image of lineage markers expression](image)

**Figure 3.11: Expression patterns of lineage markers in NPC 46C RNA-Seq data**

Expression levels of pluripotency and pro-neuronal markers reinforces the importance of the NPC 46C cell line in modelling neuronal differentiation in vitro. Pluripotency and ESCs markers’ decrease along time, while proneural genes, such as Ascl1, Neurog1 and Neurog2, increase consistently during differentiation. Ectodermal and neural progenitor markers’ expression is somewhat maintained during differentiation while the endodermal marker Gata1, serving here as a control, decreases in expression levels during differentiation. Lineage markers selection according to [28].

Gene expression quantifications and inclusion levels for alternative splicing events were obtained for the five timepoint samples of cultured NPC 46C. Although these cells model neuronal differentiation in vitro, it was not clear what was the specific transcriptomic profile of each sample, as compared to "anchors" of neuronal differentiation, such as embryonic stem cells or neurons. Likewise, the expression of a set of known lineage-specific markers was profiled, in order to compare the transcriptomic patterns of these samples with those from other datasets.

As depicted in Figure 3.11, NPC 46C express pluripotency markers (Fbxo15) and ESC markers
(Nanog, Sox2 and Oct4) in day zero of culture, when neural induction treatment is started. Further on, neural progenitor markers, such as Chrdl1, Sox3 and Pax6, show an increase in expression levels, while pluripotency/ESC markers levels start to decrease. Proneural genes, as Neurog1/2 or Ascl1, consistently increase in expression levels, as cells acquire the neural phenotype. The decrease in expression of ectoderm-specific genes suggests that cells lose the capacity to differentiate into other types of cells from the ectoderm, such as skin cells.

![Exons progressively more included/skipped in NPC 46C](image)

**Figure 3.12: Examples of exons progressively more included/skipped in NPC 46C**

A number of alternative splicing events were also found to be increasing/decreasing in PSI values along neuronal differentiation.

Analysis of correlation with time was also performed in this dataset. However, as the number of points is very low, Spearman’s good correlation are very likely to occur by chance, making correlation tests not significant after correction for multiple testing. Some examples of events with increasing levels of inclusion and one example of decreasing values are depicted in Figure 3.12.

### 3.3.1 Gene expression

One of the most relevant questions to be addressed through the analysis of this dataset was where exactly would these cells fall in terms of timing during neuronal differentiation. To answer this question, the proposed idea was to perform a principal component analysis of expression data from all genes in all mouse neurogenesis or nervous system tissue samples for which there was quantification of cRPKMs. By taking into consideration all genes and samples, the dimension reduction enabled by PCA would make it possible to:

(a) Visualize clusters of samples with similar transcriptomic behaviour, most likely samples labelled as belonging to the same developmental stage, but eventually also samples labelled differently but showing in fact a similar gene expression profiles;

(b) Identify genes contributing the most to differences in expression patterns along differentiation;
(c) Allow the identification of sample outliers that may correspond to a different biological profile or lead to the identification of technical biases.

Principal component analysis of all mouse neurogenesis gene expression data is depicted in Figure 3.13, where the three first components, accumulating approximately 60% of the variance, seem to distinctly separate groups of developmentally-related samples. Although this was expected, since cells belong to different timepoints from differentiation and therefore are subjected to different external stimuli, it made clear that different protocol conditions do not change greatly the expression patterns of differentiating cells. Moreover, the almost perfect timewise distribution of samples across the three dimensional plot, reinforced by the colored arrows, enabled indirect characterization of the transcriptomic profile of the NPC 46C cell line. These plots incorporate samples from MultiTissue, Hubbard and NPC 46C datasets (see section 2.2).

Principal component 1 clusters together all embryonic stem cell samples, even those coming from different datasets, and induced pluripotent stem cells. Then, along the PC1 axis, neuroepithelial stem cells and NPC 46C are distributed coherently with their associated timepoints of days in culture, after neural induction. Along the PC1, the final developmental stages for neurons (DS45) show similar patterns as those of developed nervous tissue samples. Therefore, principal component 1 is likely dominated by genes specifically expressed in pluripotent cells and genes specifically expressed in neuronal tissues.

Furthermore, principal component 2 separates differentiated tissues from developmental samples, suggesting that genes associated to mature tissues dominate PC2. Principal component 3 is more intriguing, since it seems to be a representation of a set of genes associated to intermediate stages of development, as it distinguishes radial glia and neurons in the first stages of development from other samples. This may be indicative of a defined set of genes having roles in the initial definition of neuronal commitment. Interestingly, the rightmost part of the PC1 versus PC3 shows a very well defined axis of differentiation incorporating timewise distributed samples ranging from embryonic, pluripotent stem cells to the first stages of neuronal development, reinforcing the accuracy of NPC 46C cells in modelling the early stages of neuronal differentiation.

### 3.3.2 Alternative splicing

Interestingly, performing PCA of exon inclusion levels for all the exon skipping events, after removal of events containing missing data, recapitulates the timewise distribution of samples accordingly to their developmental stage, as Figure 3.14 shows. Principal component 1 clearly accumulates a strong contribution from alternative splicing events distinguishing pluripotent cells from developed tissues, while principal component 2 shows a more diluted effect, greatly affected by one ESC sample outlier. In this particular case, AS events contributing for PC1 likely occur in similar ways in the later developmental stages of neurons and in mature tissues, although the sequential disposition of earlier developmental stages (from ESCs to neurons in developmental stage I/II), similar to what as described for Figure 3.13, is clear. This means that alternative splicing events also clearly distinguish cells by their developmental stage. However, in order to correctly interpret this results, alternative splicing events should be decoupled from the gene expression of their cognate genes, if possible. This would be relevant because alternative splicing quantifications generally
Figure 3.13: Principal component analysis of all genes expression data of neurogenesis/nervous systems samples

Principal components 1 to 3 accumulate most of the variance within the dataset (roughly 60%) and clearly distinguish neurodevelopmental stages. Moreover, three major tendencies show up: PC1 accumulates genes that show varying levels of expression between pluripotent stages and neuronal stages, while PC2 seems to distinguish fully differentiated tissues from developing ones. PC3 separates samples labelled as radial glia and neurons at medium developmental stages (DS12 labelling corresponds to developmental stages I/II) from other groups.
Figure 3.14: **Principal component analysis of all exon skipping alternative splicing events data in neurogenesis/nervous systems samples**

Principal component 1 clearly distinguishes neurodevelopmental stages. Separation between pluripotent samples and mature tissues is clear.

correlate with cognate gene expression, as transcription and splicing are known to often influence each other.

One striking observation is the deviation of the NPC 46C sample correspondent to day 1 from the other samples of this cell line and also from other pluripotent stem cells. This needs to be further explored but is suggesting a switch in exon inclusion patterns that follows neuronal induction *in vitro* and is not prominent for gene expression.

Furthermore, the role of microexons in this separation of different stage cells was inspected, because microexons a) are known to be more conserved when compared to other types of events; b) show a switch-like regulation and c) are known to be involved in neurogenesis [15]. When only microexons were kept into the principal component analysis, resulting in Figure 3.15, the timewise distribution of samples is still clear. In this particular case, PC1, accumulating much of the variance within microexon PSI values, separates samples in three distinct groups: the first contains pluripotent ESC, NPC 46C, NeSC and radial glia; the second group is composed by the samples considered by Hubbard and colleagues to be neurons in the first stages of development; and developing neurons and brain tissues, that are apparently similar in terms of their microexon PSI values. PC2 here separates, once again, developed tissues from *in vitro* samples. Nevertheless, if we disregard the more mature tissues (Brain, cerebellum and neurons) an almost linear axis of differentiation becomes clearer, starting with the embryonic stem cell cluster and finishing in developed neurons.
Figure 3.15: Principal component analysis for microexon skipping alternative splicing events data in neurogenesis/nervous systems samples

Microexons PSI values preserve the distinct timewise separation, although principal component 1 is separating three groups: ESCs, NeSCs, radial glia and NPC 46C together, DS12 neurons and mature tissues.
Chapter 4

Conclusions

The previous sections focused on the description of the work course followed in the present project, with the purpose of discovering switch-like alternative splicing events associated to specific neuronal development stages and exploration of their regulatory mechanisms. The Beta distribution was applied to the quantification of alternative splicing events, since it matches the characteristics of the percent-spliced-in metric and adds the dependence on the number of reads supporting an alternative splicing event. Therefore, it enables correction for biases caused by low number of reads when the interest is to have a PSI value globally representing a set of samples, namely biological replicates. Importantly, the described Beta distribution procedure is also suitable for AS quantification in different contexts, such as is the case of comparing two different conditions, like normal samples versus disease samples, if a suitable statistical test for the differences between two or more Beta distributions is implemented.

Moreover, comparison of interesting alternative splicing events between species was enabled, and therefore biological mechanisms which are comparable between human and mouse, in this case, could be addressed. This is important when studying brain and nervous systems, because availability of human samples is limited by ethical reasons. Furthermore, since alternative splicing is believed to be one of the mechanisms that play a role in the functional complexity of higher animals, orthology comparisons are very insightful.

Analysis of alternative splicing events shifting during differentiation provided promising results, which serve as strong motivation for future questions. Moreover, when matching events significantly correlated with timecourse with those correlated to the expression of RBPs that could be acting as their regulators, some alternative splicing events came out. NOVA1 expression is anti-correlated with the PSI of skipping exons in genes with important neuronal roles in axon guidance, synapse formation and other neuronal critic events, such as *Axin1*, *Dip2a* or *Mprip*, which would be further tested for statistical significance of NOVA1 RNA-binding motifs enrichment. Also, nSR100 expression levels are correlated with PSIs of exons that seem to be neuronal-specific in *Camkk2* and *Clip2*, for instance, genes known to have roles in the correct brain function.

Another layer of this project led to the transcriptomic profiling of NPC 46C, a widely used resource to model neuronal differentiation *in vitro*, and comparison with other neuronal development datasets. Preliminary characterization of alternative splicing patterns in those samples was also determined.
4.1 Future work

One of the resources required to apply the Beta distribution in the quantification of PSI values is a robust statistical method to compare two different distributions representing the two conditions under analysis. Although the conceptual idea of a possible solution was conceived, the implementation of such a method will be explored in the future. In the context of the Beta distribution applied to PSI quantifications, this could be used for comparison between two sets if we consider both density functions and for each value of delta PSI we compute the probability of the difference in values in group A and in group B to be higher than the value of delta PSI (or, in the discrete case, we count the number of values that are greater in group A when compared to B). After setting up a probability threshold for decision, for instance, 0.95 of probability of the difference to be higher or equal than that value, selection of the delta PSI would be the value of difference accumulating that probability.

Moreover, considering the RNA-binding proteins’ regulatory role in alternative splicing, the analysis will be expanded and processed with more accurate statistical procedures. For inference of the effect upon binding of an RBP to the regions in the vicinity of splice sites defining an alternative splicing event, RNA binding statistical maps will be produced. For this, for each RBP, its preferred RNA-binding motifs need to be mapped across all AS events, with respect to the regions defined in Figure 3.9. Simultaneously, expression levels of the RBP are to be tested for correlation with all events PSIs. Afterwards, in order to statistically extract binding regions for which there is motif enrichment, a sliding window encompassing a given number of nucleotides, within the same interest regions, is iterated across the sequence and a Fisher's test is performed comparing AS events with the binding motif that are correlated, anti-correlated or not correlated. For this particular case, true positives will be regions with the RNA-binding motif and correlated (or anti-correlated), while false positives will be regions that are correlated without having enrichment in the motif. This procedure leads to the determination of regions that are statistically enriched in binding motifs associated to the predicted RBP activity. A number of compromises has to be made in applying this method such that decisions lead to robust answers. Namely, a score associated to the strength of evidence for a motif to be considered as a target of each RBP, and its associated P-value, as well as the threshold to define correlated/not correlated AS PSIs. This method is already being applied in the Lab, and statistical RNA maps for some RBPs are already available.

The application of statistical RNA maps to the analysis of AS events switching during neuronal development will hopefully provide more evidence for some relations inferred using the methods previously described. One of the possible contributions would be to couple this information with the clustering of closely evolving events, eventually leading to the identification of clusters of events with similar patterns along differentiation that are regulated by the same RBPs. In a scenario where a number of AS targets is predicted to be regulated by a certain protein, enrichment in those events in predefined clusters may provide this type of associations. Interestingly, there is evidence for RBPs regulating a number of AS events, providing an amplification mechanism to the switch effects described for specific exons that has been associated to the evolutionary diversity of splicing events [17].

Furthermore, principal component analysis of the different-origin datasets of neuronal differentiation
enabled the identification of a pattern showing that both in gene expression and alternative splicing there were two distinct groups of timewise evolution in samples: first, embryonic stem cells, neural progenitors from the Hubbard dataset and the NPC 46C cell line; then, the developing neurons, mature neurons and the brain/nervous system tissues. For that reason, and with the purpose of providing more resolution to the study of alternative splicing events in neuronal development, the well defined axis of differentiation encompassing the early progenitors will be addressed. NPC 46C cells, besides being widely used, are described as recapitulating the in vivo neuronal differentiation very accurately, and therefore characterization of their profile is important.

Also, neurodevelopmental disorders have been a focus of major interest from the scientific community but the inaccessibility of living human neuronal tissues have limited the investigation on the underlying mechanisms of these diseases. Recent studies in animal models provided high-throughput sequencing data which are publicly available [38] and may be used to explore the pattern of neuronal development associated alternative splicing events. Moreover, human datasets are also available, as well as the orthology annotations required to compare human and mouse genes, meaning that these data could also be explored.

4.2 Concluding remarks

Fortunately, bioinformatics has the potential of addressing a great number of questions, which in this particular case opened many new questions to be explored. Analysis of transcriptomics and genomics involves "big data" and is consequently demanding. Each step in data mining, filtering, plotting or interpreting results is the product of a series of issues that arise from the biological questions, for which a solution is constructed and the result serves as a baseline for the next question. This "step-by-step" procedure, that enables to solve greater problems, is very motivating and serves as a continuous source of defiance. However, high dimensionality of the results is of difficult interpretation, since there are many more potentially interesting events than the number of variables to possibly consider and evaluate by human reasoning.

This project served as a fundamental part of my academic career and enabled the discovery of a science field in which I got really interested. The biological consequences of alternative splicing are a continuous source of wonderment in which I am very motivated to continue working. The human brain is made of approximately 85 billion neurons forming a functional network of signals which, in normal conditions, is able to shape human cognitive, communication and emotional skills. However, during the process of neurodevelopment, defects in synaptic maturation or connectivity may lead to neurodevelopmental disorders, impairing cognitive function, locomotion and communication. After the learning process that was part of this project, I will continue searching for the links between alternative splicing and the shaping of functional brains.
References


Appendix A

A.1 Exploratory analysis for filtering low varying AS events

Figure A.1: Variance and PSI range analysis for removal of low varying AS events
Points represent AS events, plotted based on their PSI range and log-transformed variance across samples in the Hubbard’s dataset. Events in green, with a PSI range $\geq 20$, were considered for correlation analyses.
Appendix B

B.1 Study on the impact of scaling gene expression values

The impact of scaling gene expression values in the principal component analysis was studied. Since gene expression levels are distributed across several orders of magnitude, highly expressed genes tend to dominate the first principal components when values are unscaled, given that proportionally small variations in their expression across samples will still be larger, in absolute value, than those of lowly expressed genes. Figure B.1, depicting a PCA obtained with centred but unscaled gene expression values, shows clustering of samples according to their developmental group, although PC1 does not enable a clear separation in intermediate groups, such as NPC 46C and early developing neurons. Also, mature brain samples are placed between pluripotent stages’ samples and later developing neurons. Two possible interpretations of this may be considered: a) genes that dominate PC1 are genes that distinguish ESC from later stages of developing neurons and are not genes distinguishing pluripotent cells and functional neurons (that will be expected in samples from mature brains) or b) later stages of developing neurons in Hubbard’s dataset do not recapitulate the transcriptomic profile of functional neurons.

Log-transformation or scaling of the data both enable medium expression genes to weight more on the principal component analysis and the dominance of highly expressed genes is dissipated. These transformations enabled the observation of a neuronal differentiation timewise distribution of samples with respect to the first principal components, which made it easier to interpret the specific transcriptomic profile
of NPC 46C samples and better understand the timing of these cells’ differentiation into neurons.

<table>
<thead>
<tr>
<th>ESC_day−8</th>
<th>NeSC_day−4</th>
<th>RG_day0</th>
<th>DS34_day7</th>
<th>DS45_day16</th>
<th>DS45_day21</th>
<th>DS45_day28</th>
<th>DS12_day1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurons_M</td>
<td>Brain_FrontalLobe</td>
<td>Cerebellum</td>
<td>Neurons_DS34</td>
<td>Neurons_DS45</td>
<td>Neurons_DS45</td>
<td>Neurons_DS45</td>
<td>Neurons_DS45</td>
</tr>
</tbody>
</table>

Figure B.2: Principal component analysis of standardized (mean of zero and variance of one) gene expression values

<table>
<thead>
<tr>
<th>ESC</th>
<th>iPS</th>
<th>NPC 46C</th>
<th>NeSC</th>
<th>Radial Glia</th>
<th>Neurons DS12</th>
<th>Neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain_B12</td>
<td>Brain_BNP</td>
<td>Brain_Cortex_b</td>
<td>Brain_FrontalLobe</td>
<td>Cerebellum</td>
<td>Neurons_DS34</td>
<td>Neurons_DS45</td>
</tr>
</tbody>
</table>

Figure B.3: Principal component analysis of centred and unscaled log-transformed gene expression values
Appendix C

C.1 General procedure for processing AS and gene expression data

1. Resources tables
   - Upload of all resources tables
     - RBP, Ortholog genes list, all MultiTissue data, (..)

2. Neuro PSI Data
   - Upload gene expression data (Hubbard, Linares, Goff, NPC)
   - Add samples table with descriptions

3. General Hubbard
   - Variance analysis
   - Choose only skipped exons
   - Filter NA, variance, ΔPSI
   - Sort columns as samples
   - Result: only SE, Var > min, ΔPSI, NA

3.1 Hubbard SF Expression
   - Sort column names
   - Subset to what’s in Mouse orthologs SF
   - Remove only NA rows
   - Result: MouseH_SF_cRPKM

3.2 Time Correlation
   - x = Timepoints
   - y = MouseH_SE_events
   - Keep only correlation coefficient > 0.7
   - Get annotation of events
   - Result: Correlation_MouseH_SE_events

3.3 K-means clustering
   - Run silhouette with Spearman’s correlation as distance between 2 and 50 K
   - Run Kmeans for optimal K

4. SF PSI Correlation Hubbard
   - Only corrected P-value < 0.05
     - MouseH_signif_SE
   - Result: Correlation_SF_PSIs

Figure C.1: General scheme illustrating pipeline for analysis of PSI data