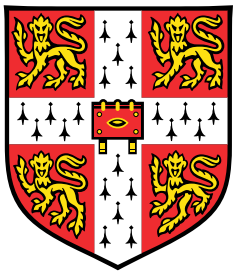
Assessing the usage of alternative transcripts in human tissues from RNA-seq and proteomics data



Sérgio Miguel Cardoso Marcelino dos Santos

EMBL - European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Wolfson College

May 2019

Dedicated to my parents.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Sérgio Miguel Cardoso Marcelino dos Santos
May 2019

Acknowledgements

I would like to thank my supervisor Alvis Brazma for the support and guidance during my studies. I am extremely grateful for being given the opportunity to do a PhD in his group. I would also like to thank all the members of my Thesis Advisory Committee: Pedro Beltrao, Juan Antonio Vizcaino, Anne-Claude Gavin, Jyoti Choudhary, and John Marioni. Their feedback and valuable suggestions were very important, as well as their genuine interest in my work. In particular, I thank Juan for the support at the beginning of my PhD and feedback on my thesis. Regarding the work I have done, I would also like to thank Vihandha Wickramasinghe, whom I collaborated with and Nuno Fonseca for the technical bioinformatics advice, for feedback on my work and thesis. Lastly, I would like to thank Mar Gonzàlez-Porta for starting the studies that led to most of my PhD work, giving me a direction to works towards from the beginning. I am also grateful for the people that gave me feedback every time I publicly presented my work.

I would like to thank all the people I contacted daily, such as the current and former members of the group: Mitra, Liliana, Natalja, Claudia, Zichao and Manik. They made my everyday life pleasant and enjoyable. In particular, I thank Mitra Barzine for making me feel at home from the beginning, for helping me, for the valuable tips and suggestions.

I am grateful for the EMBL predoc community for making me feel welcome and especially facilitating the start of the PhD period. I feel very fortunate for the nice people I have met on a regular basis at EBI. I would also like to thank Lynn French, Tracey Andrew and Lorraine McAlister for taking care of bureaucracies, making my life easier and with fewer worries.

I would like to thank all my friends, in particular, Ricardo Adaixo and António Ribeiro for giving me the extra push I needed to start a PhD and I am very grateful for all friends who have supported me in my personal life, my family, my friends, the salsa dancing community and the natural lifestyles community. I thank my parents for listening to me, for making the best out of their lives in my absence and my family to always making me feel at home everytime I go back to Portugal.

Lastly, I thank Maria for the support through this time, for being present and supportive when life was not the easiest, for the work-related suggestions and for overall contributing to making my life more stable and much better.

Abstract

Alternative splicing is an important step in gene expression regulation in eukaryotes, through which a single gene can express different transcript isoforms. We can now use RNA sequencing (RNA-seq) data to identify which isoforms of each gene are being expressed in a specific condition, by quantifying the expression level of each isoform of a gene, even though quantification of isoforms remains a difficult task. In this way, we can better understand how prevalent this process is and how often a gene expresses different isoforms. It can also be evaluated if all isoforms of a gene are about equally expressed or if there is one dominant isoform that is significantly more expressed than the others. Moreover, by applying this analysis to different tissues, it can be assessed if there are changes in splicing between different conditions and if such a change has a biological role.

A dataset of 32 normal human tissues was used in this study. The results show that, although alternative splicing can lead to the expression of different transcripts of a gene, many genes have an n-fold dominant transcript – a transcript that is expressed at n times higher level than the second most expressed one [1]. On average, 68% of protein-coding genes expressed in a given tissue have a 2-fold dominant transcript and 47% have a 5-fold dominant transcript.

It was observed that the dominant transcript of a gene tends to be the same across tissues, but there are cases where the dominant isoform switches between tissues, these cases are designated switch events. For a given pair of tissues, there are on average around thirty 2-fold switch events and just below four 5-fold switch events. The switching exons often significantly overlap and the most common types of alternative splicing are alternative 3' selection (24% of the cases) and alternative 5' selection (21%).

To evaluate the conservation of the transcripts, the dominant transcripts were compared to APPRIS principal isoforms. These isoforms are annotated based on their function, protein structure, and cross-species conservation [2]. 69.2% of the 2-fold dominant transcripts and 81.1% of the 5-fold dominant transcripts are APPRIS principal isoforms. It was also observed that in 80% of the switches there are no protein domain changes.

Similar results were obtained when the same analysis was done using the GTEx dataset [3], which has a much higher number of samples, containing data from 54 conditions. In

this case, on average 59% of expressed genes in a given condition had a 2-fold dominant transcript and 31% had a 5-fold dominant transcript. The number of switch events was again low, given the number of dominant transcripts, indicating that dominant transcripts tend to be conserved across normal tissues.

A comparative analysis of matching tissues common to the two mentioned datasets was also performed and, although the datasets are different, there were switch events in common between both of them. 5 examples of 5-fold switches involving domain swaps were analysed in detail and it was revealed that the type of genes affected by switches can be quite distinct and the protein domains that change between isoforms can vary both in number and function. The tissues found to be particularly more represented on these switch events were skeletal muscle, testis and cerebral cortex.

These results show that in most cases, changes in alternative splicing do not change transcripts significantly, and respectively, the changes at the protein level are minor. This and similar observations [4, 5] indicate that alternative splicing may not be the main process responsible for generating protein diversity.

In the last study presented in this thesis, it is analysed how RNA-seq data can be integrated with a data-independent acquisition (DIA) mass spectrometry method, SWATH-MS (sequential window acquisition of all theoretical spectra-mass spectrometry), to study the impact of depleting PRPF8, a core spliceosomal component, on the proteome. The results show that intron retention events lead to decreased protein abundance. It is also shown that differential transcript usage and gene expression have effects on protein abundance, altering it proportionally to transcript levels. Overall, some links between transcript and protein level are revealed and it is demonstrated how perturbed systems can be used in the study of alternative splicing [6].

Table of contents

List of figures

List of tables

Chapter 1

Biology of gene expression

At the center of molecular biology is deoxyribonucleic acid (DNA). DNA was first discovered in 1869 by Friedrich Miescher, who called it "nuclein" and, far from knowing what he had actually discovered, he simply described it as a phosphorus-containing substance. Many studies followed its discovery but ones of significant importance were done by Erwin Chargaff in 1950, where he found that DNA nucleotide composition varies among species and that the amount of adenine (A) was similar to the amount of thymine (T) and the amount of cytosine (C) was similar to the one of guanine (G). These discoveries were essential for the later construction of the model structure of DNA (Figure 1.1) proposed in 1953 by James Watson and Francis Crick and presented as the molecule that carries the genetic information from one generation to the other [28]. Previously in 1866, Gregor Mendel had published his work on how certain traits, like shape and color, were passed on peas from one generation to the next. He also recognized that certain traits had different characteristics that could be dominant or recessive. Without knowing, he was studying genes, DNA regions that code for functional RNA molecules, the functional unities of genomes, which contain the complete set of genetic information in an organism [28].

The understanding of molecular biology has evolved much since these discoveries but one of the most remarkable landmarks was sequencing the first complete human genome, a joint effort of a group of publicly funded researchers, the International Human Genome Sequencing Consortium. First, the draft genome was published in 2001 [29, 30] and two years later the final version was made available [31].

It was understood that there was potential in the discovery of the source code for human life and that it could help to find the answers to many biological problems, although the complexity of this discovery was greatly underestimated. It was soon evident that extracting useful information from a sequence of DNA was a far harder task than it was anticipated.

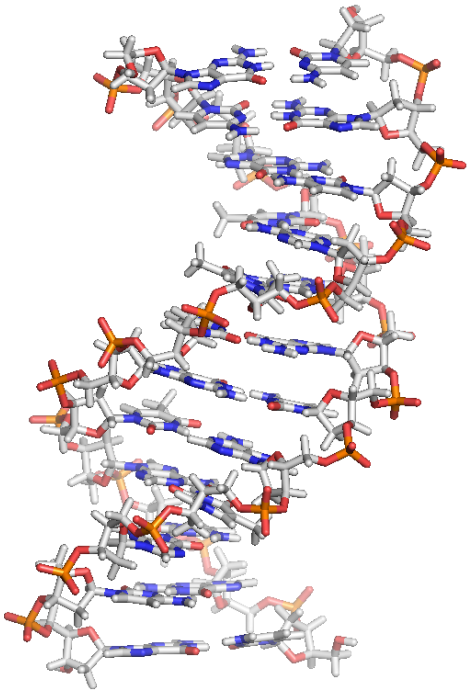


Figure 1.1 Structure of the double helix of DNA.

Since the human genome was first sequenced, studies such as the ENCODE project analysed sequencing data in a large scale manner to extract biologically relevant information. In this study, regions of transcription were identified and mapped, as well as, regions of transcription factor association, regions involved in establishing chromatin structure and histone modification. In the ENCODE project, biochemical functions were assigned to 80% of the genome and relationships between regulatory elements and gene expression were discovered, helping to understand mechanisms of gene regulation [32].

With the advances in sequencing technologies, more genomes are being sequenced at cheaper prices. Nevertheless, it is still not a straightforward process to extract comprehensive information to answer complex biological questions. Also, many genomes are sequenced to study very specific topics but there are more questions that could be potentially addressed with that same data. Therefore, there is a big potential in the data to explore and there is a need to implement automatic methods that can extract information from all the data that is being continuously generated and many times made publicly available.

1.1 Central dogma of molecular biology - RNA

The central dogma of molecular biology was first stated by Francis Crick in 1958 [33] as a hypothesis for how pathways of information can be transferred between biological molecules. It stated that the information can be transferred from nucleic acid to nucleic acid or from nucleic acid to protein but not between protein to protein or protein to nucleic acid. By information, was meant the precise determination of sequence, either nucleic acid bases or amino acid residues. As a general statement this is right and not only allows the transfer of information from DNA to ribonucleic acid (RNA) but also from RNA to DNA, DNA to DNA and RNA to RNA. This is a more complete version of the dogma than the perhaps more popular one which was proposed by James Watson in 1965 which only allows information to flow from DNA to RNA and from RNA to protein [34].

The dogma describes the common flow of information in a biological system. It describes the most obvious processes of DNA replication, RNA transcription, protein translation, as well as, RNA replication and reverse transcription, that are common in viruses. The way it is stated is in fact so generic that it includes all major biological processes. However, it would also allow processes that are not encountered in nature, like proteins being translated directly from DNA, without using messenger RNA (mRNA). Despite this and other exceptions, as a generalization the central dogma is true and its main message is that information cannot flow from protein to nucleic acid [35].

The DNA sequence is however only one of the components in the system, and to fully understand gene expression, one must look at transcription and translation, and analyse the expression levels of the gene products (Figure 1.2).

1.1.1 Transcription

Transcription is the process of transferring information from a sequence of DNA to a molecule of RNA. In eukaryotes, it occurs in the nucleus where DNA is located and is mediated by a complex of multiple enzymes. The main enzyme involved is RNA polymerase, which matches the RNA bases to the DNA sequence, assembling the new RNA molecule. The activity of RNA polymerase is mediated by transcription factors, proteins that bind to specific DNA sequences, promoters or enhancers, to initiate transcription. The complex of RNA polymerase and transcription factors is called the transcription initiation complex and is responsible for initiation, elongation, and termination of transcription (Figure 1.3).

Initiation is the first step of transcription and, in the case of mRNA, starts with RNA polymerase II (RNAPII) binding to the $5'$ end of the gene to a sequence called promoter, which affects the transcription rate depending on its affinity for RNA polymerase, and

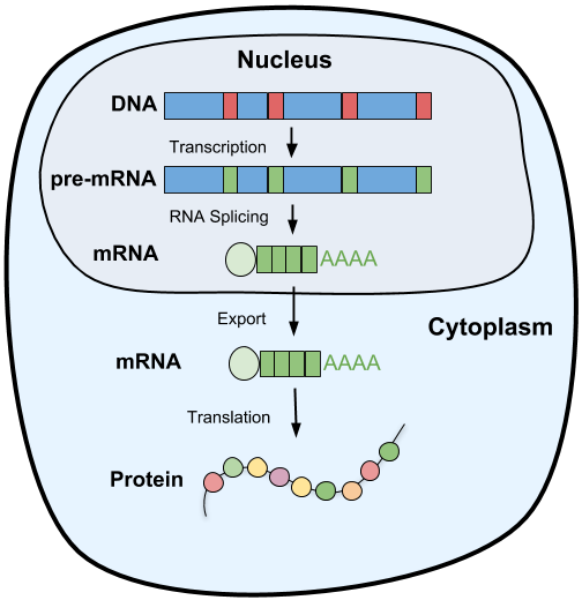


Figure 1.2 Basic steps of gene expression in eukaryotes. The information stored in DNA is transcribed to RNA in the nucleus. The RNA is processed and it can undergo alternative splicing, originating one or multiple mRNA molecules. The mRNA is then exported to the cytoplasm where it is translated into a protein (adapted from [7]).

determines the localization of the transcription start site (TSS). The promoter also determines the direction of transcription and the DNA strand that is transcribed. It is an important region in the regulation of gene expression and it is composed of a core promoter and regulatory domains [36].

Other regulatory elements, such as enhancer sequences, can be thousands of bases far from the transcription site but are brought close by the looping of DNA, which is a result of the interaction of proteins bound to the enhancer and others bound to the promoter. These proteins are either activators or repressors, depending on the effect they have on transcription. DNA changes its 3-D structure to facilitate the activity of RNA polymerase. A number of specialized proteins is needed to stabilize DNA which is tightly packaged as chromatin [37].

After initiation, the elongation process begins and the new RNA sequence is extended from the start site to the termination site. After transcription initiation, the RNA polymerase unwinds the DNA double helix, reads the DNA template and adds nucleotides to the $3'$ end of the growing RNA sequence [37].

The final phase of transcription is termination. The termination process in eukaryotes occurs differently depending on the polymerase that is transcribing. In the case of RNAPII, terminator sequences are at the end of noncoding sequences and are recognised by protein factors. In the case of RNA polymerase I and III, a termination factor stops transcription

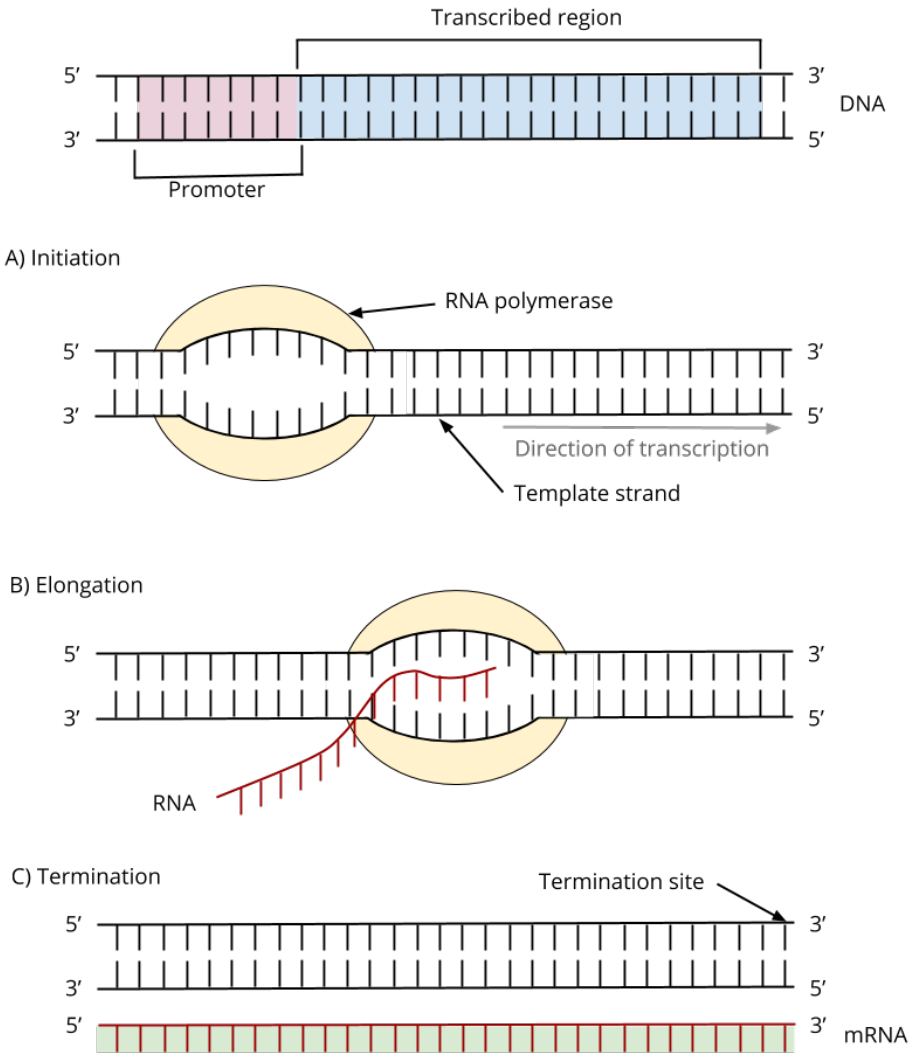


Figure 1.3 The three basic steps of DNA transcription. (A) Transcription is initiated when the enzyme RNA polymerase binds to a promoter sequence on the template DNA strand. (B) The elongation process starts with the unwinding of the DNA double helix. The enzyme RNA polymerase continues reading the template DNA strand and adding nucleotides to the $3'$ end of a growing RNA molecule. (C) The transcription is terminated when the RNA polymerase reaches a termination sequence. At this point both the mRNA transcript and RNA polymerase are released from the complex.

only after transcribing a polyuracil stretch. The transcription can continue for hundreds or thousands of nucleotides after the end of a noncoding sequence until a cleavage and polyadenylation specific factor, and a cleavage stimulation factor, finally cleave the mRNA. Cleavage occurs at a consensus sequence and it is coupled with termination. Then the mature mRNA is polyadenylated at the $3'$ end, resulting in a poly(A) tail. This process is also coordinated with termination, but how these processes relate to each other still remains unclear [37, 38].

1.1.2 Translation

In eukaryotes, translation and transcription occur in different compartments of the cell. Transcription occurs in the nucleus and translation in the cytoplasm. The mRNA is transported through the nuclear pore complex channel, bound to ribonucleoproteins, and the shuttling mRNA binding proteins are removed in the cytosol before translation [39].

Translation starts with the small subunit of the ribosome and an initiator transfer RNA (tRNA) molecule binding the mRNA transcript. tRNAs are adaptor molecules that carry amino acids for the emerging peptide chain and can read the triplet code in the mRNA through complementary base-pairing. In eukaryotes, the complex of initiator tRNA and the small ribosomal subunit is firstly formed, and it then binds the mRNA transcript, which means there is a simultaneous binding of the tRNA and the small ribosomal subunit to the mRNA (Figure 1.4).

Ribosomes sequentially read groups of three bases of mRNA (codons), that correspond to a particular amino acid, and translation starts at the AUG codon, which is translated to methionine. A complex of initiation and elongation factors brings aminoacylated tRNAs to the ribosome-mRNA complex and matches the mRNA codon to the tRNA anticodon, which is complementary to the mRNA triplet. Each tRNA has the appropriate amino acid coupled to it, so as the process repeats itself the elongation of the amino acid chain occurs, and the chain simultaneously folds into conformation. The end of translation occurs with a stop codon that can be one of three options: UAA, UGA or UAG. The polypeptide chain may require further processing to become fully formed or active, so it might undergo processing by chaperones to get into the right conformation. Some sections of the peptide chain might also be cleaved from the protein and then discarded, those are called inteins. Other proteins might also be split into different sections, undergo cross-linking or cofactors binding and other post-translational modifications [40].

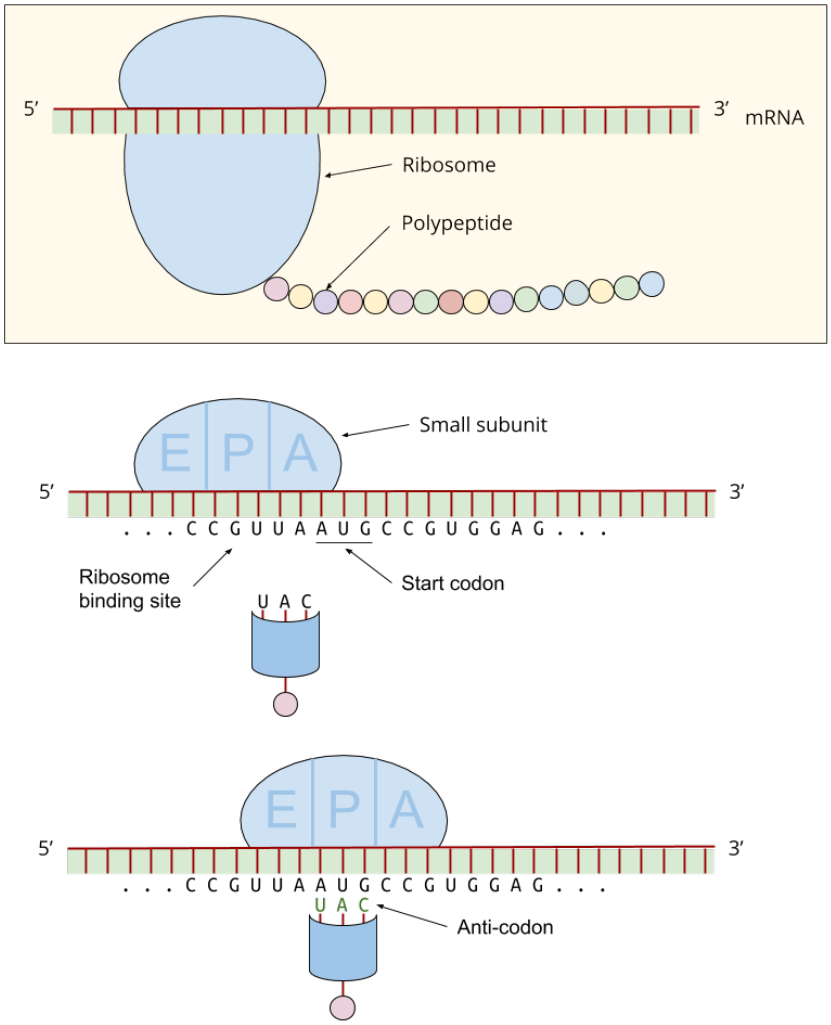


Figure 1.4 The process of RNA translation. There are three binding sites in the small subunit of the ribosome: an amino acid site (A), a polypeptide site (P), and an exit site (E). The process starts with the initiator tRNA molecule, which carries the amino acid methionine, binding to the AUG start codon of the mRNA at the P site of the ribosome where it will become the first amino acid of the polypeptide chain (adapted from [8]).

1.2 Alternative splicing

Before the human genome project was concluded in 2001, the predictions of the number of protein-coding genes were between 30,000 and ~140,000 [30]. However, it was then realised that the number of protein-coding genes was much lower, around 20,000, which was quite surprising considering the complexity of human cells when compared to simpler organisms, like *Arabidopsis thaliana* with ~25,000 or *Caenorhabditis elegans*, a nematode 1 mm in length and ~20,000 protein-coding genes. So it was concluded that the number of genes might not be a good indicator of organism complexity, especially in eukaryotic genomes [41]. Besides correcting the number of genes over time, the definition of gene has also changed due to the realization that the most abundant products of transcription are not mRNA molecules but rather tRNA, rRNA and many others involved in gene expression regulation [42].

Alternative splicing is a mechanism by which a pre-mRNA can be processed into different mRNA molecules. As a direct consequence, a single gene can code for multiple proteins or other gene products, depending on how exons are spliced together. However, the downstream effects of splicing vary both among genes and species [41].

The complexity of an organism does not correlate well with the size of the genome and the number of protein-coding genes. With that said, alternative splicing is only one of the processes that enable the creation of complexity without increasing genome size. This process combined with the action of regulatory elements, such as promoters and enhancers, will lead to an end result that is, not only difficult to predict but also to understand without knowing all the components that intervene in gene expression and regulation [41].

There are other mechanisms that contribute to eukaryotic genome capacity for complexity. These include RNA editing, trans-splicing, and tandem chimerism [43–45]. RNA editing is the process by which RNA can be modified after transcription. For example, a modification of an RNA base in a mRNA, which can affect the protein that is translated [43]. Trans-splicing, in opposition to alternative splicing, is the splicing together of separate pre-mRNAs to form a new mRNA [44]. Tandem chimerism is the transcription of two adjacent transcription genes into a single “chimeric” mRNA, which can then be translated into a fused protein, having parts of both original proteins [45].

1.2.1 Discovery of alternative splicing

RNA splicing was discovered in 1977 by Richard J. Roberts and Phillip A. Sharp [46, 47]. Before that time, a gene was thought to be a long uninterrupted sequence of DNA, however, they discovered that was not always the case. They found that genes could be discontinuous, meaning that multiple segments of DNA could be separated but end up together at the RNA

level. Both Roberts and Sharp were studying the localization of genes on the genome of adenovirus and found out that the mRNA did not behave as expected. Shortly after, it was shown that split genes could be found in higher organisms and that exons could be differently spliced in different transcripts through a process named alternative splicing. These findings drastically changed the view, not only on gene expression but also on gene evolution [48].

They used electron microscopy to determine where the segments of DNA were located in the genome. They discovered that the mRNA was derived from four different segments of DNA located separately in the genome, concluding that certain genes could have genetic information discontinuously organized in the genome [46]. After this initial discovery, it was shown that this split gene structure was common and also the most frequently found in higher organisms. In terms of evolution, this discovery implied that genes not only evolve by accumulating minor mutations, resulting in a gradual change but could also suffer meaningful rearrangements that result in larger nucleotide and functional changes. An exon could actually correspond to an entire protein domain, adding or removing an entire functional unit to the protein, which had considerable effects during evolution [48].

1.2.2 The process of alternative splicing

In the process of alternative splicing, once the pre-mRNA is processed into different mRNA molecules, information is lost and it is no longer possible to reconstruct the original DNA sequence from the final mRNA. When pre-mRNA is transcribed, it is composed of introns and exons but only exons are typically part of the mature mRNA, which may include protein-coding sequences and untranslated regions at either end of the transcript. Introns are portions of the pre-mRNA that are removed during splicing. Splicing controls which exons end up in the mature RNA and these may vary according to the tissue, cell type or condition, which creates a possibility of a gene expressing different transcripts in different conditions through alternative splicing [49].

The process of alternative splicing has some variations depending on how the exons are split. There are four main types of alternative splicing, exon skipping (Figure 4.16 - a) is the most common in higher eukaryotes, accounting for about 40% of splicing events, but is quite rare in lower eukaryotes. The second and third most common types in higher eukaryotes are alternative $3'$ and $5'$ splice site selection (Figure 4.16 - b and c), accounting for 18.4% and 7.9% of the events, respectively [9]. These events occur when two or more splice sites are recognized at one of the ends of an exon. The fourth most common type is intron retention (Figure 4.16 - d) and it occurs when the intron is kept in the mature mRNA transcript. It is the rarer type of event in vertebrates and invertebrates, only 5% of the cases, but is the most common type in plants, fungi, and protozoa. There are also other types: mutually

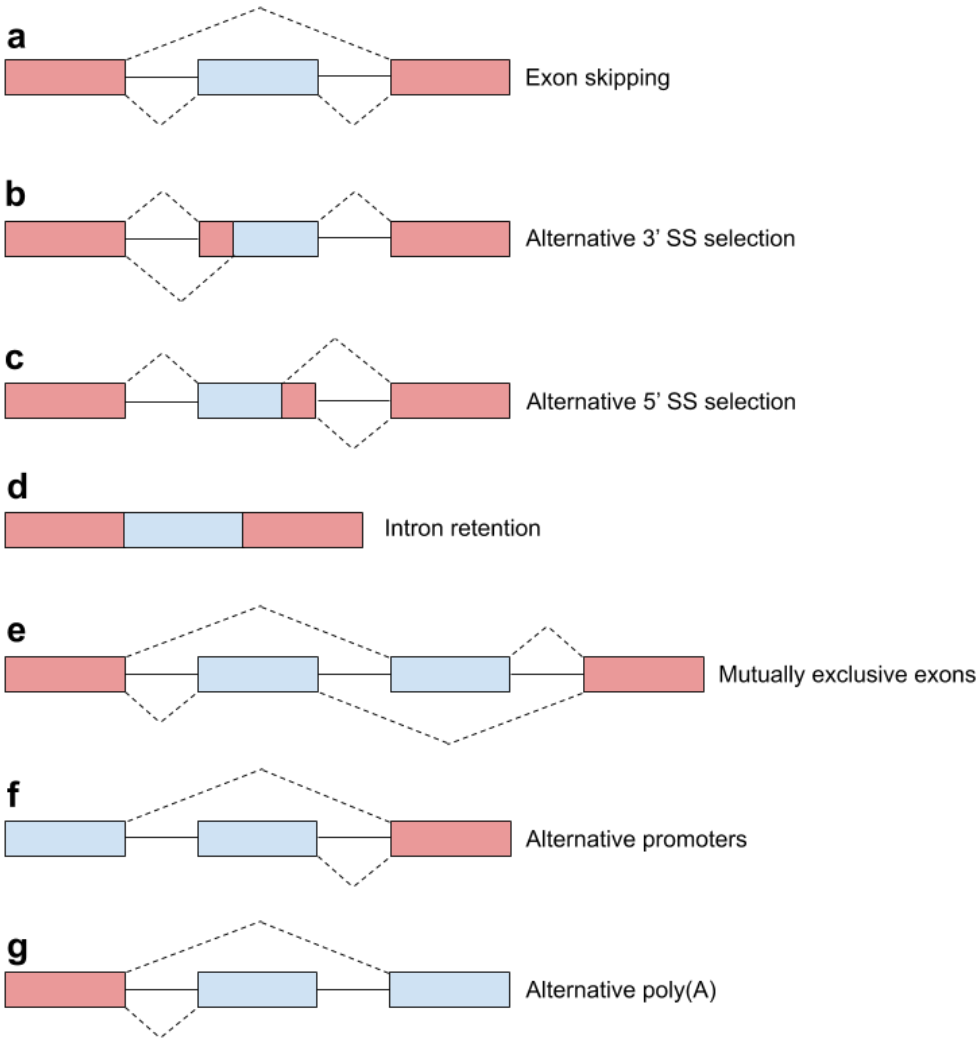


Figure 1.5 Different types of alternative splicing (adapted from [9]).

exclusive exons (Figure 4.16 - e), alternative promoter usage (Figure 4.16 - f) and alternative polyadenylation (Figure 4.16 - g) [9].

The splicing process uses specific sequences at the splice junctions of the pre-mRNA that identify where exons and introns are located. Therefore, mutations in these regions can lead to the production of unstable or out of frame mRNAs. As a consequence, protein expression can also be affected [50]. The percentage of genetic diseases involving splicing mutations that could be direct mutations of splice sites or disruption of other splicing components can be as high as 60% [51]. This is one of the reasons why understanding alternative splicing is important.

Functional alternative splicing has been acquired throughout vertebrate evolution and it is a key mechanism of gene evolution alongside amino acid change [52]. This is one of the

processes through which ~20,000 human genes are able to express approximately ~160,000 transcripts, according to Ensembl annotation [53]. RNA-seq has allowed the identification of a large number of transcripts [54], resulting in an increased number of annotated transcripts in the databases as the technology advances.

Studying alternative splicing at the RNA level has certain limitations because RNA expression level does not always correlate with the protein expression level [55]. On the other hand, some splicing events are difficult to detect with mass spectrometry (MS) experiments because an isoform might only be identified by one peptide only, which may not be detected [56]. Therefore, it is possible to detect more splicing events using RNA-seq data.

It has been suggested that 92-94% of human genes undergo alternative splicing and that it varies 2- to 3-fold less between individuals than between tissues. Interestingly, switch-like regulation has been associated with sequence conservation in regulatory regions and with the generation of full-length open reading frames. There is also a correlation between patterns of alternative splicing, alternative cleavage, and polyadenylation, suggesting that these have coordinated regulation and that there is high sequence conservation of regulatory motifs in alternative introns and 3′ UTRs [57].

Although gene expression has strong tissue dependence both across individuals [58] and different species [59], the same is not observed for alternative splicing patterns. Most of the variance found between human tissues is due to differential gene expression, with alternative splicing playing a small role in it. However, there is more variance of alternative splicing between individuals, which might be seen as an indication that this process is more stochastic [60]. There is also more variance between species than between tissues [61]. Given these general observations, it has to be noticed that there are exceptions and in fact, some exons are, not only tissue specific but also conserved across species [62].

1.2.3 Function of alternative splicing

Alternative splicing can generate RNA diversity from the genome [63] but the idea that alternative splicing is also the primary source of protein diversity has been questioned based on the results of some large-scale proteomics. Although most genes have the potential to express multiple isoforms, most genes express a single protein isoform. It has also been shown that most alternative exons are under neutral selection, which imposes the question of what is the role of alternative splicing and how prevalent it really is [4]. There are well-known examples of alternative splicing functions such as regulating transcription factors, localization of proteins, enzymatic activity and protein interactions, but in many cases, it is difficult to detect the changes brought about by it [63]. So despite some of its functions being known,

there are still mechanisms that are not yet fully understood. Especially in large-scale analysis, it is difficult to make an assessment for each specific case.

One of the evolutionary advantages of alternative splicing is that new genes can be created through nondisruptive recombination at introns, by a process called exon shuffling [64]. Another advantage is to expand the coding capacity of the genome, without the need to create totally new genes. Both vertebrates and invertebrates have a similar number of genes, around 20,000, however, the number of genes that can undergo alternative splicing is much higher in vertebrates [65], suggesting that this process is closely related to increased transcriptome complexity. In fact, alternative splicing is approximately twice as frequent in organs from primates, as in the equivalent organs from mouse and other species [66]. The difference is even higher between unicellular and multicellular organisms, the latter having longer and more numerous introns [67]. It has also been proposed that alternative splicing could be the main driver of evolution of phenotypic complexity in mammals, especially in primates which have the highest alternative splicing complexity [66].

Alternative exons predominantly encode regions located on the outside of the protein. As a result, the general structure of the protein does not change, there is only an effect on the protein surface [68]. In this way, alternative exons can introduce new domains without disrupting the overall protein structure [69]. This suggests that alternative splicing can introduce protein changes without radically altering the function of a protein.

1.2.4 Alternative splicing regulation

Alternative splicing increases the coding capacity of genomes and its regulation is essential for determining cell- and tissue-specific features. Its regulation is mediated by chromatin structure, DNA methylation, histone marks, and nucleosome positioning, which provide a dynamic scaffold for interactions between the spliceosome and transcription complex [70].

Splicing is performed by the spliceosome, a ribonucleoprotein (RNP) mega particle that binds splice sites in each intron. These splice sites consist of consensus sequences that are recognized by the spliceosome and have different binding affinities according to their similarity to each consensus sequence. Splice sites compete for splicing and their comparative binding affinity determines how they are spliced [71]. It should be noticed that, although transcription and splicing are often depicted as a two-step independent process, splicing has actually a cotranscriptional nature, so a full-length primary transcript might not exist, meaning that in very long genes, some sites might be spliced before the end of transcription [72].

The regulation of tissue-specific alternative splicing is usually done by a combination of tissue-specific and ubiquitously expressed RNA-binding factors that interact with cis-acting

RNA elements, controlling spliceosome assembly near splice sites [73]. Some factors can activate or repress splicing depending on the context, that is largely influenced by the location of binding relative to specific core spliceosomal components [74].

Splicing is regulated by some cis-regulatory sequences such as exonic splicing enhancers, exonic splicing silencers, intronic splicing enhancers, and intronic splicing silencers, that affect how the splice site is used according to their location. There are also *Trans*-acting factors that bind splicing enhancers and silencers, which include serine–arginine (SR)-rich and heterogeneous nuclear ribonucleoprotein (hnRNP) families of proteins, as well as tissue-specific factors, such as polypyrimidine tract–binding protein (PTB) [75], NOVA [76], and FOX [77]. Regulation is also dependent on the position of the binding on the pre-mRNA [78]. The interaction between RNA-binding factors can coordinate regulation of polyadenylation and splicing, making sure that both UTR regulatory sequences and coding regions are expressed in conjunction to obtain the appropriate tissue-specific isoforms.

One of the models of alternative splicing suggests that there is an interaction between splicing, transcription, and chromatin organization, that controls the process temporally and spatially [71].

Alternative splicing is regulated by splicing factors, both by their abundance and their post-translational modifications. It is also dependent on the functional and physical coupling that is established between the transcription and splicing machinery. Coupling is an RNAPII-dedicated mechanism and it can occur by two mechanisms: recruitment coupling and kinetic coupling. Recruitment coupling requires splicing factor recruitment to transcription sites by the transcription machinery. On the other hand, in kinetic coupling, it is the speed of RNAPII elongation that controls alternative splicing by affecting the rate at which splice sites and regulatory sequences emerge in the pre-mRNA that is being formed during transcription [71].

1.2.5 The spliceosome

The spliceosome machinery is very accurate and flexible, due to the fact that its conformation and composition is extremely dynamic. In most eukaryotes, there are two unique spliceosomes: the U2-dependent spliceosome, which removes U2-type introns, and the U12-type spliceosome, which is less abundant and not present in all eukaryotes [10].

On the pre-mRNA, the information that defines an intron is limited to conserved sequences at the branch site, 3′ splice site, and 5′ splice site. The branch site is located 18-40 nucleotides (nt) upstream of the 3′ splice site and is followed by a polypyrimidine tract in higher eukaryotes (Figure 1.7), a region rich in pyrimidine nucleotides, especially uracil, that is usually 15-20 nt long. U2- and U12-type introns have different branch sites. There are several *cis*-acting elements that can be splicing enhancers or silencers, both intronic and

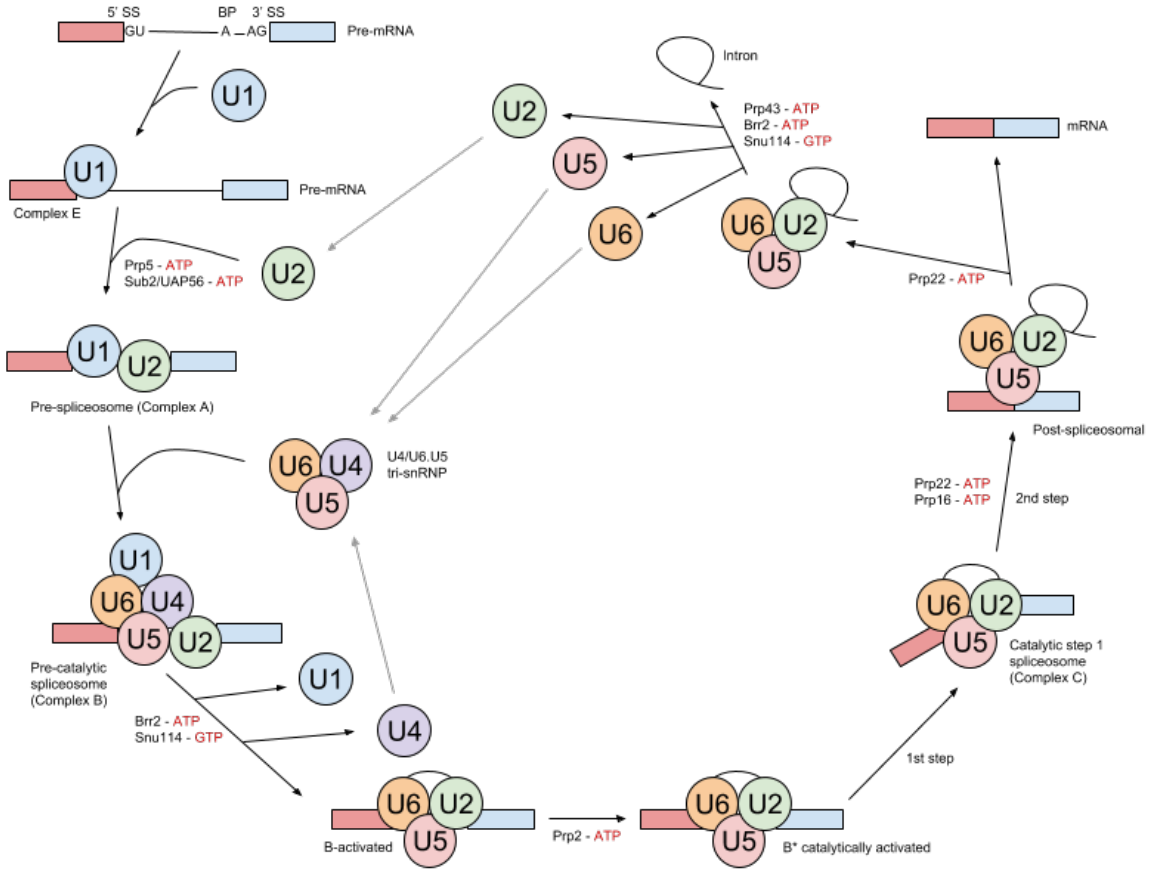


Figure 1.6 The U2-dependent spliceosome pathway. All the snRNPs are represented by circles and non-snRNP proteins are just represented by their names. In the transcripts, exons are represented by blue boxes and introns are represented by lines (adapted from [10]).

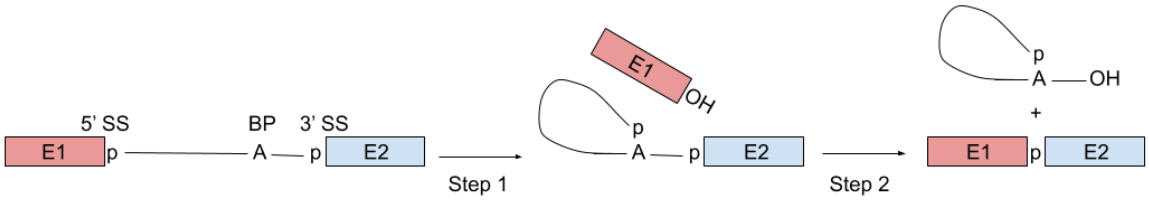


Figure 1.7 The splicing mechanism. Exons are represented in boxes (E1 and E2) and introns in solid lines. The letter 'A' represents the branch site adenosine and p represents the phosphate groups at the $5'$ and $3'$ splice sites (adapted from [10]).

exonic. They have short and diverse sequences to which regulatory proteins bind, modulating in this way both constitutive and alternative splicing, stimulating or repressing the assembly of spliceosome complexes [10].

The information contained in the splicing substrate is limited, therefore the regulation of alternative splicing is also controlled by a large number of *trans*-acting factors that together with the pre-mRNA form the spliceosome. The components of the U2-dependent spliceosome are U1, U2, U5, and U4/U6 small nuclear ribonucleoproteins (snRNPs), as well as numerous non-snRNP proteins. It has to be noticed that the U12-dependent spliceosome contains different snRNP components. During splicing, snRNAs undergo structural rearrangements and none of them possess an active site, unlike ribosomal subunits [10].

The assembly of the spliceosome happens in a predefined order of interactions between snRNPs and the other splicing factors. First, U1 is recruited to the $5'$ splice site and a set of non snRNP factors interact with the branch site and the polypyrimidine tract. Then, U2 binds the branch site forming the A complex (the prespliceosome). A pre-assembled complex of U5 and U4/U6 snRNPs is recruited by the A complex, forming the B complex. As the RNA and protein interactions go through rearrangements, U1 and U4 are destabilized and leave the spliceosome complex, forming the B activated complex, which suffers catalytic activation, subsequently catalyzing the two first steps of splicing. This reaction gives rise to the C complex, which catalyzes the next step of splicing, causing the spliceosome to dissociate, and the snRNPs can then take part in additional rounds of splicing (Figure 1.6).

Exons in mammalian pre-mRNAs have a rather constrained length of ~120 nt on average. On the other hand, introns have varying sizes that can go from hundreds to more than 1 000 000 nt, being on average ~5000 nt. When intron lengths exceed 200-250 nt, splicing is undergone through another pathway, called exon definition [79]. This process is quite prominent in mammals and it starts with U1 binding the $5'$ splice site of an exon, facilitating the association of U2AF with the polypyrimidine tract upstream of it. As a consequence, U2 is recruited to the branch site upstream of the exon, splicing enhancers recruit SR family

proteins and a network of protein-protein interactions is established across the exon around which the exon-defined complex has been formed. After exon definition, the $3'$ splice site pairs with an upstream $5'$ splice site across the adjacent intron, and there is a switch from exon-defined to intron-defined splicing complex, where the cross-exon complex is disrupted and converted into a cross-intron A complex, forming a molecular bridge between U2 and U1 bound to a $5'$ split site. It is through this process that it is determined which exons are ultimately spliced together. Exon-defined complexes contain other snRNPs and can be converted directly into cross-intron B complexes. There are also multiple pathways for an exon-defined complex to be converted into an intron-defined spliceosome, but the pairing of splice sites usually occurs during the A complex formation, with very few exceptions when it might actually occur in later stages [10].

Chapter 2

Uncovering the human transcriptome with RNA-seq

The transcriptome is the set of all transcripts in a cell or condition. By studying the transcriptome, one can better understand which regions of the genome are being expressed, as well as estimate the relative expression values of genes and individual transcripts. Comparative transcriptome analysis can help to understand the differences between conditions, being cell types, tissues, development stages, or between normal and disease states. Transcriptomics studies might include transcripts other than mRNAs, such as non-coding RNAs, helping to uncover the role of different components of the transcriptome and the genome. In this way, the structure of genes can also be better understood, the start and stop sites of exons can be mapped, and alternative splicing patterns can be analysed [54].

2.1 Functions of RNA

RNA has diverse functions that go beyond the ones of mRNA and tRNA. Of these two, mRNA is a copy of a DNA sequence, the protein blueprint, and tRNA carries amino acids to assemble each new polypeptide chain. Besides tRNA, another RNA type that participates in translation is ribosomal RNA (rRNA), which is functionally the most important component of the ribosome. These are the three types of RNAs that have a role in protein synthesis but there are others that have different functions [80].

The diversity of RNAs and their functions gave rise to the hypothesis that RNA preceded the evolution of DNA and proteins. The RNAs that do not encode proteins are generally called non-coding and its most common types are the above-mentioned rRNA and tRNA. However, there are also long non-coding RNAs (lncRNAs) and others that are generally described as

small regulatory RNAs (sRNAs), which can exert their activity through a combination of complementary base pairing, through complexing with proteins or even by having their own enzymatic activities [80].

One of the subcategories of sRNAs is small nuclear RNA (snRNA), which play an important role on alternative splicing. There are also microRNAs (miRNAs), which regulate gene expression by binding to mRNAs and repressing their translation, generally by imperfectly pairing, although there are exceptions. A number of these miRNAs appear to be linked to cancer and other diseases [80, 81]. Small interfering RNAs (siRNAs) are another type of sRNA that also inhibits gene expression. They can be single or double-stranded and can be incorporated in a complex called RISC (RNA-induced silencing complex), which can bind to a sequence of mRNA with the complementary sequence, inhibiting transcription. siRNAs might have evolved as a defense mechanism against double-stranded viruses. They are generated in a similar process to miRNAs, but they are derived from a longer RNA molecule and processed by the Dicer enzyme. Their mechanisms of inhibition of gene expression are different in most cases but there are cases in which both act in similar ways [82]. Another type of RNA is small nucleolar RNA (snoRNA), which is located inside the nucleus in a structure called nucleolus, where rRNA processing and ribosomal assembly occurs. snoRNAs take part in the processing of rRNA, namely with methylation and pseudouridylation of specific nucleosides [83].

There are also less common RNAs, such as riboswitches which modulate gene expression by detecting environmental and metabolic cues and affecting expression in accordance with those. Ribozymes are another example, in this case, they have catalytic functions similar to enzymes, intervening in replication, mRNA processing, and splicing.

Finally, there are lncRNA that interact with DNA, RNA and transcription factors, participating in processes such as DNA methylation, histone modification, and chromatin remodeling. They operate as signals, decoys, guides, or scaffolds to regulate gene expression [84].

There is a big variety of noncoding RNAs, and maybe some will still be discovered. Understanding the biology of RNA is essential to have a full picture of the molecular biology of the cell [80, 81].

2.2 Methods to study RNA

Although the variety of RNA types is vast, alternative splicing is commonly studied by analysing mRNA expression data. All products of gene expression from the same locus are called isoforms and they might vary in transcription start sites (TSSs), untranslated regions

(UTRs) and protein-coding DNA sequences (CDSs). It has to be taken into account that assessing the extent of the differences in mRNA isoform expression between conditions has presented substantial technical challenges in the past and still presents some nowadays [85].

High-throughput techniques to study RNA started being used more than 20 years ago. These studies were first done using expressed sequence tags (ESTs), which yielded relatively low estimates of tissue specificity. The difficulty of this method on detecting differences in isoform expression levels is related to its limited statistical power [86]. EST methods enable the analysis of gene expression by partially sequencing complementary DNA (cDNA), obtaining in this way the sequence, as well as the abundance of RNAs. Although this method has been of importance in the past, the high sequencing costs and the relatively low throughput limited its application in expression analysis. Additionally, the data obtained is semi-quantitative [87].

Tag-based methods tried to overcome some of the mentioned limitations. Serial Analysis of Gene Expression (SAGE) [88] significantly reduced the cost of expression analysis per gene, because it only relied on sequencing a short tag region per cDNA, 15 to 21 base pairs. Other methods such as Cap Analysis of Gene Expression (CAGE) [89], and Massively Parallel Signature Sequencing (MPSS) [90] have higher throughput and provide gene expression levels. However, they all have serious limitations related to their dependency on Sanger sequencing technology and to the fact that many of the short tags used cannot be uniquely mapped to the reference genome. It has also to be noticed that only a portion of the transcripts is analysed and isoforms might not be distinguishable from each other [54].

These methods were largely replaced by DNA microarrays when they appeared, mostly because of the lower costs for large-scale studies and the consistently higher coverage achieved across tissues [91]. DNA microarrays are based on hybridization of transcript derived targets fluorescently labeled to probes attached to a solid surface. The method requires the sequence information to be previously known or the reference genome to be available. This technology has low specificity and low sensitivity for some genes, due to the background signals and the occurrence of cross-hybridization. Because of these issues, this technique also has a limited dynamic range [54, 87].

High-throughput Next Generation Sequencing (NGS) overcame some of the mentioned pitfalls. RNA-seq has less background noise and greater dynamic range than microarrays. It also generates high coverage of mRNAs and by sequencing the transcripts, reveals sequence identity directly, instead of relying on hybridization techniques to distinguish and quantify them [87, 92]. It allows not only the identification of transcripts in a given condition but also the quantification of these same transcripts at a low cost, that keeps getting lower with the technology advances [93].

2.3 Experimental workflow of RNA-seq

Although direct sequencing of RNA is possible, most RNA-seq experiments are done through sequencing complementary DNA (cDNA), which is DNA synthesized from mRNA by the enzyme reverse transcriptase. This means that a cDNA library has to be prepared. There is a wide variety of library preparation protocols but in all of them, DNA or RNA fragments are fused with adapters that contain the necessary elements for immobilization on a solid surface. During this step, it has to be considered how the RNAs of interest are going to be captured, how the RNA is going to be reverse transcribed into double-stranded DNA and how the adapters are going to be placed on the cDNA ends for amplification and sequencing.

There are some size selection steps that are also performed, as well as PCR amplification. The library preparation should ensure a good molecular recovery of the original fragments so that the genomic coverage is high even with little sequencing. Some library preparation protocols might introduce biases in sample composition, which poses technical challenges and may lead to biased data.

In general, RNA-seq protocols are more challenging than DNA-seq protocols and can also produce more bias. Low complexity is one of the types of bias and is the result of the production of many reads with the same starting point. Another bias is the uneven coverage across transcripts which can produce antisense artifacts when using standard libraries [87].

2.3.1 Library preparation

RNA-seq procedures start with rRNA depletion or mRNA enrichment. In the case of eukaryotic transcriptomes, after the RNA is extracted, polyadenylated mRNAs are extracted with oligo-dT beads. Alternatively, rRNAs can be selectively depleted using ribonucleases, which has the advantage of not restricting the analysis to polyadenylated RNAs. These latter protocols are usually referred to as total RNA protocols, in contrast with polyA-selected protocols. Oligo-dT beads protocols are cheaper and are the most popular also because the most commonly sequenced RNAs are polyadenylated RNAs, either mRNA or lncRNA. Selection can also be done with oligo-dT priming for reverse transcription but priming based methods can exhibit $3'$ bias, resulting in reads enriched in the $3'$ portion of the transcripts. Therefore, poly(A) purification is the preferred method for selecting RNAs [94].

This technology has some constraints regarding the size of reads that can be sequenced, therefore the extracted transcripts need to be fragmented in reads with the adequate size. Additionally, the reads also need to cover the whole length of each transcript. This step is also necessary because of size limitations of most current sequencing platforms. There are many techniques for RNA fragmentation, RNase III digestion and chemical zinc-induced hydrolysis

being two of them. RNase III cleaves RNA in a sequence and structure-specific manner, unlike zinc-mediated cleavage which has no such specificity but still is not completely random [94]. In this process, uneven fragmentation can be a source of bias, because it can lead to differential representation of specific RNA regions [87].

After fragmentation, random hexamer primers hybridize to the RNAs and retrotranscription is initiated, generating double-stranded cDNA. The following step is ligation of adapter sequences to both ends of each cDNA before amplification and sequencing (Figure 2.1). The adapters have the purpose of enabling hybridisation of the cDNA fragments into the flow cell and also work as primers for the sequencing reaction. The resulting cDNA fragments are then size-selected using gel electrophoresis to fit the adequate size for the sequencing machine and then the cDNA library is amplified by PCR [94].

2.3.2 Sequencing

After library preparation, the samples are loaded into a flow cell for sequencing. The adapters in the flow cell are complementary to the ones ligated to both ends of the cDNA fragments to allow hybridisation. A second amplification step follows and bridge amplification is used in order to increase the signal for sequencing. This consists of the synthesis of cDNA complementary fragments, which then hybridise with adjacent adapters, allowing subsequent rounds of synthesis. The result is clusters of identical sequences that are then sequenced and read by synthesis. This procedure uses modified versions of the four bases, that incorporate a reversible terminator and a fluorescent dye. During each sequencing cycle, the new reagents are added, the new bases are incorporated and then elongation is blocked. The incorporated bases are identified by measuring the fluorescent signal. This process is repeated for the whole sequence and at the end, a set of images enables the identification of each base by using base calling software, which converts the images to sequences or reads. These reads are a copy of the expressed RNA in the initial samples and their length is the same as the number of cycles performed during sequencing. At the end of the process, both the sequence and the probability of a base being wrong are often saved in a FASTQ file [94].

There are some biases that are introduced during library preparation. One of them is that random hexamer priming is not totally random; there are preferences for certain sequences and as a consequence, there are fragments that undergo preferred conversion to cDNA [95]. PCR amplification also has the same type of bias, leading to differential amplification of fragments dependent on their GC content [96]. Additionally, there can be wrong base calls if the elongation reaction is not blocked or the fluorescent dye is not removed at the right time [97]. Technology advances have reduced the biases over time by introducing new protocols and algorithms to take them into account. Cufflinks [14] algorithm is an example of such a

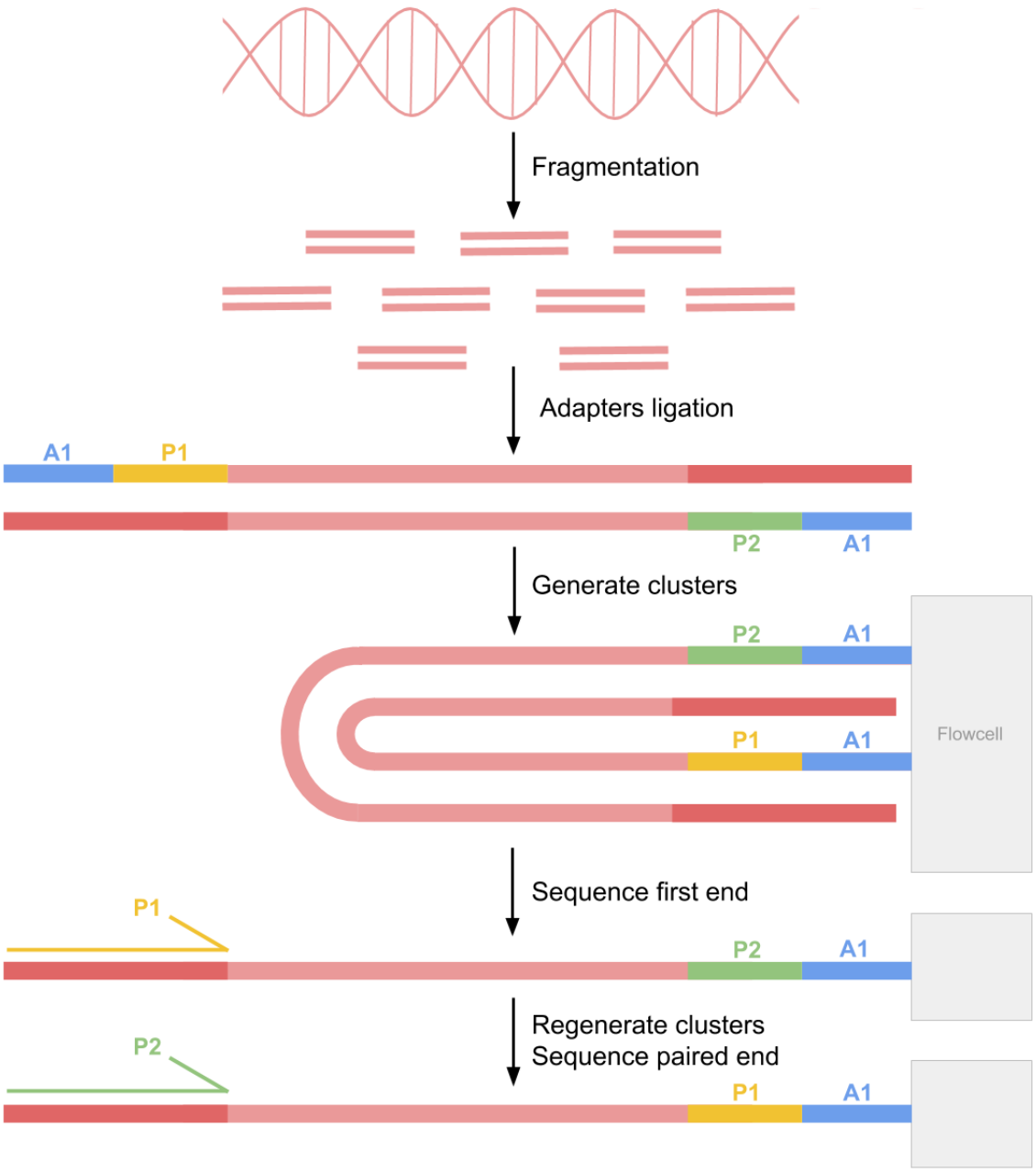


Figure 2.1 Library preparation and sequencing of an Illumina platform. This is the paired-end workflow, which includes the ligation of adaptors at each end of the cDNA molecule. Both ends of the cDNA fragment are independently sequenced. Letter 'A' designate adapters and 'P' designate primers (adapted from [11]).

case. Regarding the PCR protocols, the use of molecular identifiers, such as random barcodes also reduced the mentioned bias [98].

2.4 Read mapping and transcript quantification

One of the challenges of RNA-seq is to reconstruct the full set of transcripts that were present in the original samples. In addition to identifying the transcripts, estimating the expression levels also presents its own challenges.

There are two main strategies to map sequencing reads. One consists of aligning the reads to a reference genome or transcriptome, and the other is *de novo* assembly of the reads into contigs, without using a reference sequence. The first method is used in most situations, and the second one is used in particular cases, such as identifying new transcripts. Read mapping is a time consuming and computationally expensive task due to its complexity. It is also a task that may introduce some errors because of the decreasing quality of the reads at the $3'$ end, specifically in Illumina platforms [99]. Therefore, quality control checks are usually performed on the reads prior to mapping and trimming of the $3'$ end of reads is a common practice to avoid low-quality nucleotides that might lead to less accurate mapping. Reads with overall low quality are also removed for the same reason and also to decrease the complexity of the time-consuming mapping procedure.

When a reference genome is available, reads can be aligned directly to that sequence, and if transcriptome annotation is available, reads can be aligned to the transcriptome. On the one hand, aligning directly to the transcriptome, simplifies the alignment process, drastically reducing the complexity of the task. On the other hand, this approach does not allow the discovery of new transcripts or any other analysis that requires detecting expression in intronic or unannotated regions [98].

Mapping of RNA-seq reads is more complex than DNA-seq reads because the mRNA does not contain introns, so if the aligner is using a genome reference, it must handle gapped reads and these gaps might be 10 to 100,000 bases long. Another issue that the aligner needs to deal with is processed pseudogenes, which may cause the incorrect mapping of many exon-spanning reads [98].

2.4.1 TopHat2 aligner

The alignment tool TopHat2 [98] implements a compromise strategy, combining the ability to identify novel splice sites with mapping of known transcripts. TopHat2 aligns reads of various lengths to the reference, allowing indels of multiple dimensions. It also has the option

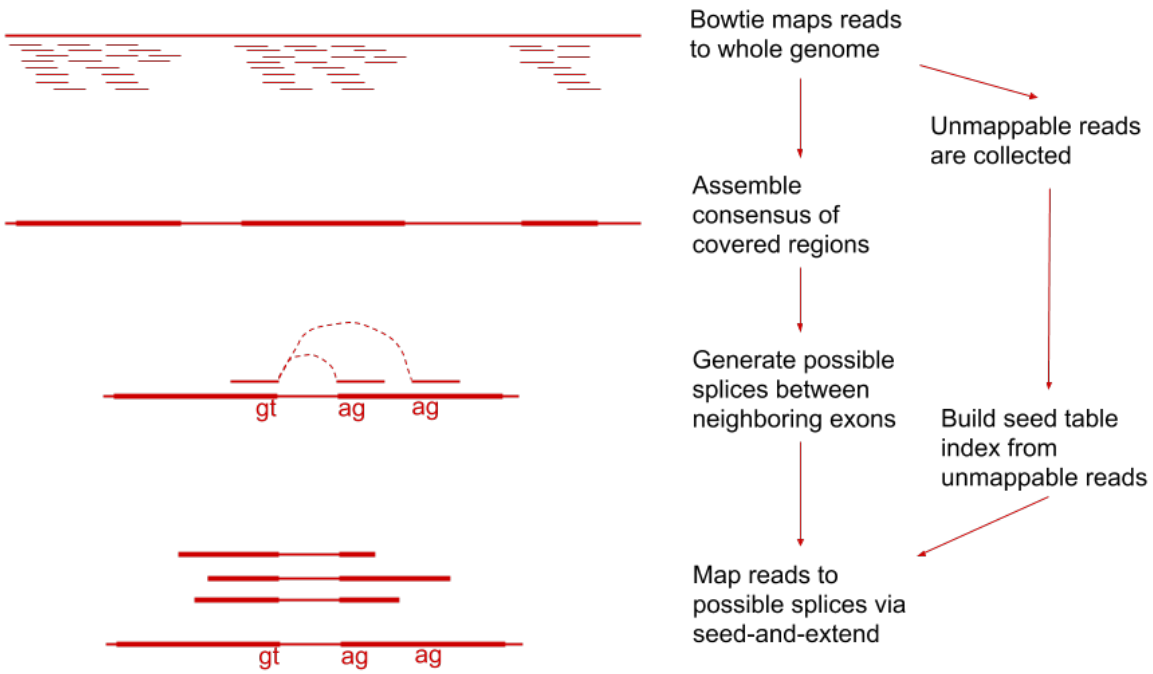


Figure 2.2 TopHat pipeline. First, all reads are mapped against the reference genome and the ones that are not able to be mapped, are set aside. The first consensus of mapped reads is computed, followed by the determination of potential splice junctions, using sequences potentially surrounding splice sites. Then the initially unmapped reads are indexed and TopHat tries to align them to previously determined splice junctions (adapted from [12]).

of allowing read mapping across fusion breaks, corresponding to genomic translocations. It combines the identification of novel splice sites with mapping to known annotated transcripts, producing accurate alignments even in highly repetitive genome regions and pseudogenes.

TopHat2 implements a two-step approach: first potential splice sites for introns are detected; then the candidate splice sites are aligned to multiexon-spanning reads (Figure 2.2).

TopHat2 starts by finding exact matches between k-mers of reads and the transcriptome. Those that are unable to be mapped, can then be mapped to the genome. Optionally, reads can be directly mapped to the genome in the first place. After this step, the mapper tries to identify splicing junctions to correctly align multiexon-spanning reads. These splice junctions and their flanking sequences are concatenated to form the full transcripts. Reads that are not aligned in the initial phase, as well as low-quality ones, are then split into smaller fragments and realigned. Paired-end reads are mapped separately and only in the final phase is the additional information, such as fragment length and read orientation, taken into account. The output of mapping is recorded on SAM/BAM files [100].

In the first step of the alignment, TopHat2 can use full-length transcripts defined by the annotation, improving sensitivity and accuracy. With that said, some of the transcripts from

the target genome might differ from the reference genome due to insertions, deletions or bigger structural variations, which can cause problems in the alignment of reads to these regions. To overcome these problems, TopHat2 implements some procedures to ensure the correct mapping is done. To do so, it uses Bowtie2 [101], that runs under TopHat2 and is quite efficient at detecting short indels. Optionally it may also use the TopHat-Fusion algorithm to detect large indels, inversions, and translocations involving different chromosomes [98].

One of the problems in the alignment of RNA-seq data is the presence of processed pseudogenes in the reference genome. Reads that span across multiple exons can also map to a pseudogene version of a functional gene, which can lead to inaccurate mapping. Pseudogenes often have no introns but some can be very similar to functional intron-containing genes, making it hard for the aligner to determine from which region the reads are actually from. Some reads might even align perfectly to both gene and pseudogene. To avoid this problem, TopHat2 uses these reads in the splice alignment phase, so even if these reads align perfectly to a potential pseudogene region, they are still tested to see if can be split and aligned to a functional gene [98].

2.4.2 *De novo* assembly

There are some specific cases when *de novo* assembly has to be used. Situations where a quality reference genome annotation is not available or in cases where a given sample is expected to be considerably different from the reference, like in the case of cancer samples. The assembly of contigs from short reads is, however, a complex task, even with paired-end reads many regions are still difficult to assemble. Since these methods largely rely on matching overlapping reads, the read size greatly affects the efficiency of the task. With technology advances, read lengths are increasing, which facilitates this procedure, so it is expected for this methods to be more frequently used.

2.4.3 Transcript quantification

After aligning the sequencing reads, the transcripts can be quantified. Typically the expression values are estimated for genes and transcripts. This process can make use of an annotation or not, if *de novo* transcript identification is being performed. One of its main difficulties is that transcript isoforms from the same gene can be highly similar in sequence and share most of the same exons, difficulting the assignment of reads to the transcript of origin.

The simplest approach for estimating expression levels relies on counting the number of reads that align to a certain locus. HTSeq [102] is a program commonly used to count reads. In the process of counting reads, there must be taken into special consideration reads that

map to multiple locations and reads mapping to highly repetitive regions, because both can lead to overestimation of counts. HTSeq opts for excluding these reads from the analysis in order to avoid this problem. Another strategy is to distribute uniformly the reads across all mapping regions, such as TopHat2 does [98]. In the case of overlapping reads, HTSeq offers different options to deal with them, as well as several execution modes for using information provided by the annotation (Figure 2.3) [102].

Besides this more simplistic and straightforward quantification strategy, there are several quantification packages that estimate normalised quantification scores. One of these is RSEM [103], which implements iterations of EM (Expectation-Maximization) algorithms to assign reads to isoforms. A more recent tool that uses a similar approach is eXpress [104]. There are different strategies like the Bayesian inference method used by TIGAR2 [105], which is optimized to work better with longer reads, and Cufflinks [14], which can be used for *de novo* transcript discovery and quantification.

All methods have to deal with common challenges. Some reads overlap with exons that are shared across multiple isoforms of the same gene, which hinders the assignment of reads to specific transcripts. In these cases, inference approaches, such as the ones implemented in Cufflinks [14], RSEM [100] or Salmon [106] can be used. Due to this challenge, one of the strategies often used to estimate expression values relies on the use of reads that uniquely map to a specific annotated transcript of a gene. Another important source of information are split reads, which are particularly informative to detect splice junctions. Paired-end reads are also extremely useful in these cases. By sequencing two ends of the initial cDNA, a bigger genomic region is covered, facilitating the mapping and consequently the estimation of transcript abundance. Another useful complementary information is the distribution of fragments length, which can be used to deconvolute ambiguous mapping, allowing to assign lower likelihoods to specific mappings that would involve very long distances between paired reads.

2.4.4 Cufflinks - transcriptome assembly and quantification

The alignments produced by TopHat2 can be used by Cufflinks to identify an expressed locus in a given sample. In the case of paired-end reads, Cufflinks aligns each pair of reads as a unit. Fragment assembly (Figure 2.4 - b) starts with the identification of pairs of incompatible fragments which belong to distinct mRNA isoforms. The fragments are connected in an overlap graph if they are compatible and the alignments overlap in the genome. There are one node and one edge in the graph for each fragment, and each of these is placed between a pair of other compatible fragments (three examples are represented in yellow, blue and red in Figure 2.4 - c) that have been originated from distinct isoforms. However the other

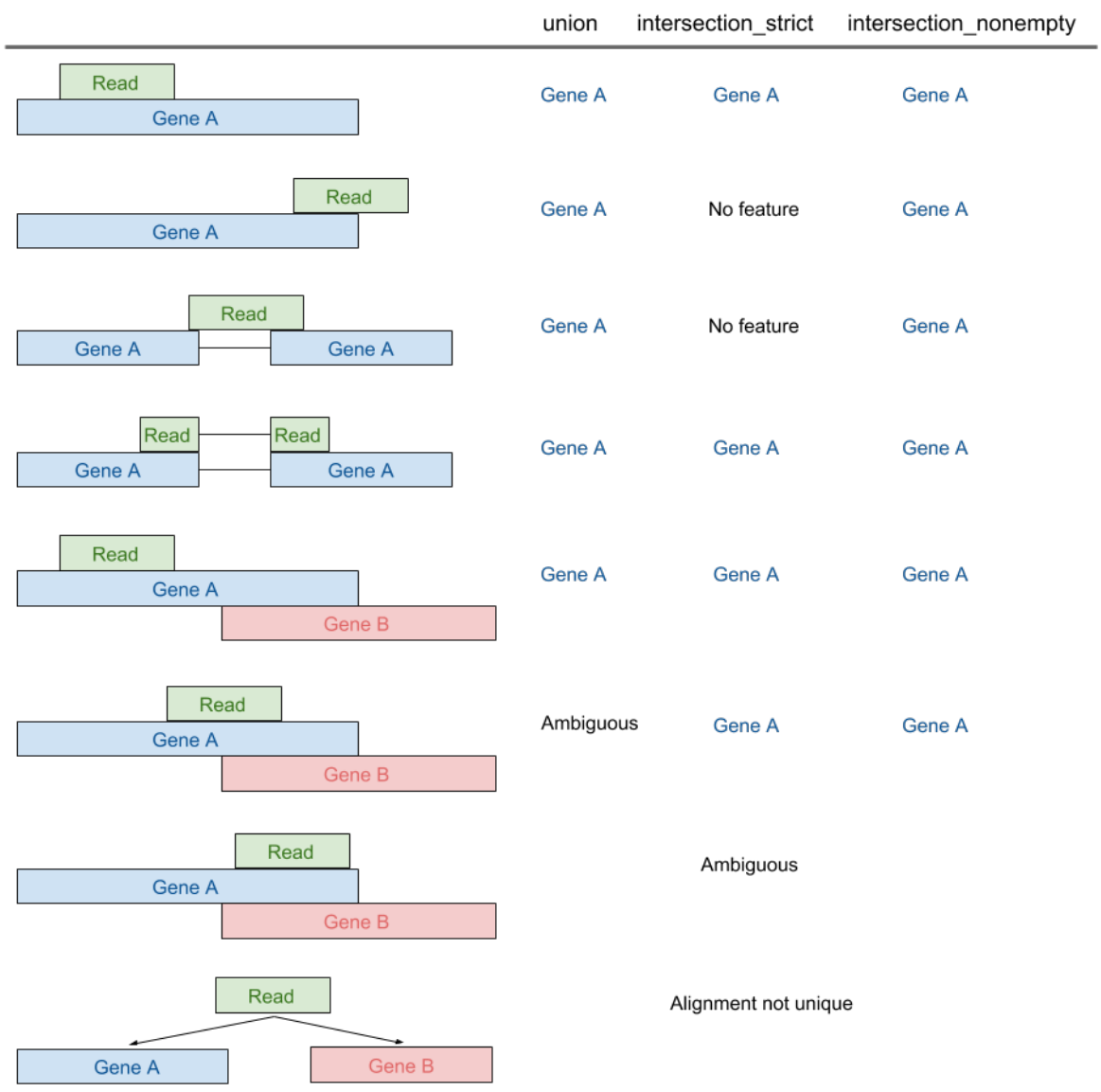


Figure 2.3 The three different modes of running htseq-count. The different modes provide flexibility to choose how reads that do not totally overlap with transcripts are treated (adapted from [13]).

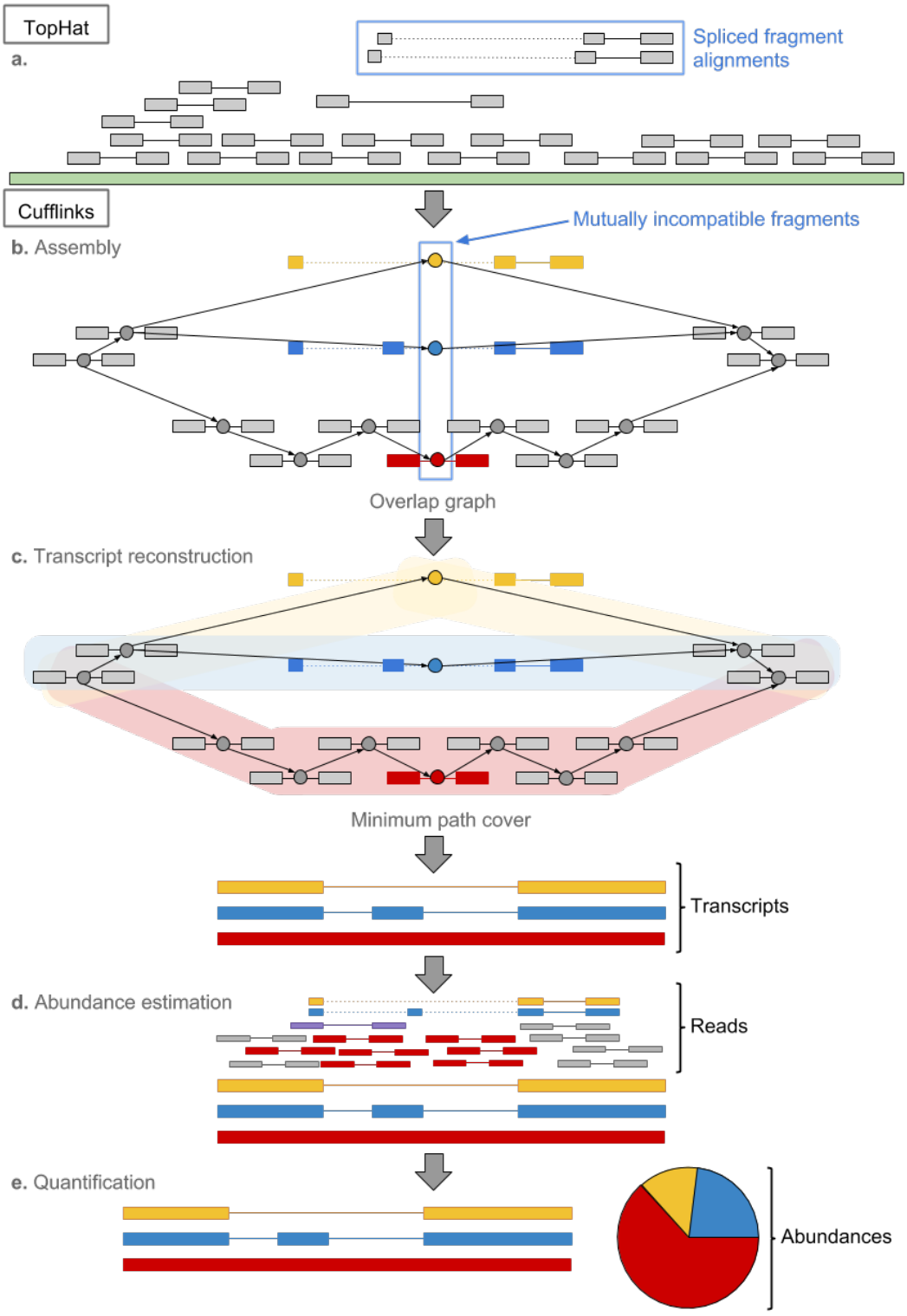


Figure 2.4 Cufflinks pipeline. The input of Cufflinks is reads (cDNA fragment sequences) that have been aligned by TopHat (a) or a compatible program. In the case of paired-end reads, each pair of fragment reads is treated as a single alignment. The overlapping fragment alignments are assembled separately (b-c), reducing CPU and memory usage. Then the abundances of the assembled transcripts are estimated (d-e) (adapted from [14]).

fragments can belong to any of the transcripts that the three colored ones belong to. The isoforms are assembled from the overlap graph. The graph paths represent sets of mutually compatible fragments which can be assembled into the full isoforms. In (Figure 2.4 - c) there are three different minimal paths, each representing a different isoform. According to Dilworth's Theorem, the number of mutually incompatible reads equals the minimum number of isoforms that are required to explain all the fragments. The implementation of a proof of Dilworth's Theorem in Cufflinks finds the largest set of reads assuring that no two could have been originated from the same transcript. It does so by producing a minimal set of paths which cover all fragments. The estimation of transcript abundance (Figure 2.4 - d) is done by assigning fragments to the isoforms they could have been originated from. In Figure 2.4 - d, both the blue and red isoforms could have been originated from the violet fragment. In the Cufflinks statistical model, the probability of observing each fragment is a linear function of the abundance of transcripts from which it could have been originated. The program can also use the distribution of fragment lengths to assist the assignment of fragments to isoforms. Finally, Cufflinks numerically maximizes the function which calculates a likelihood for all possible sets of relative abundances of all isoforms ($\gamma 1$, $\gamma 2$, $\gamma 3$ in Figure 2.4 - e), generating the abundance distribution that best describe the observed fragments, as can be seen in the pie charts.

One of the reasons why cufflinks is still often used is the fact that it can be used for *de novo* transcript identification. Cufflinks can identify reads that do not belong to already annotated transcripts. *De novo* assembly is still a complex task and read length is certainly an important limiting factor. Regions with very low expression also complicate this process, making it hard for the algorithm to find the best solution for the constructed graph. Due to the way the algorithm works, trying to extend to the maximum the graph paths, alternative start and end sites become also difficult to identify.

2.4.5 MMSEQ

MMSEQ [107] is another assembly and quantification tool that uses a strategy similar to Cufflinks. The main difference between these two tools is the implementation of the inference method and the type of input they use. Cufflinks uses reads mapped to a reference genome, relying on a frequentist inference approach to determine the expression levels that best explain the data. As a consequence, it does not quantify the uncertainty of the determined expression values. MMSEQ, on the other hand, uses a Bayesian model and takes reads mapped to a previously known transcriptome as input, which limits the type of analyses that can be performed downstream. Both methods try to correct sequence biases in their models and have strategies to use reads that map to multiple sites. To correct sequence-dependent

bias, these tools attribute a weight to each position in the expressed loci by taking into account the sequence context. These weights are used during the abundance inference step to model the non-uniform location of reads across the isoforms [108, 109].

2.4.6 Kallisto - pseudoaligments and quantification

More recently, transcript quantification methods that do not rely on a full alignment to a reference genome or transcriptome have been developed. These methods are generally faster and rely on alignment-free or pseudoalignment strategies. Sailfish [110] was one of the first programs and it used a k-mer approach. Later it was changed to incorporate the same mapper used in Salmon [106], which uses a "quasi-mapping" approach. It implements a two-phase procedure with both online and offline iterations of EM, as well as, two different modes of quantification. Salmon can use its own mapper, RapMap [111], or take BAM files as input.

One other program that is now becoming popular is Kallisto [15], which makes use of a *de Brujin* graph to achieve efficient pseudoalignments. Its pseudoalignment algorithm produces a list of compatible transcripts for each read, avoiding individual bases alignment (Figure 2.5). It uses *de Brujin* graphs built from k-mers of the transcripts in the annotation, instead of k-mers from reads, and the paths in the graph correspond to the transcripts. The quantification made from pseudoalignments is calculated using a likelihood function that takes into account the set of mapped fragments and the set of transcripts [15].

2.4.7 Transcript expression levels

Estimating transcript expression levels is a more complex task than gene expression levels because transcripts of the same gene can be very similar to each other and share multiple exons, which makes it hard for the computational method to determine to which transcript each read belongs to. To some extent, transcript identification can rely on an annotation and paired-end reads. Reads that map uniquely to one of the annotated transcripts are valuable for this task and split reads that extend across multiple exons are also informative. Paired-end reads also facilitate the assignment to the correct transcript, not only because this type of sequencing covers larger genomic regions, but also because it restricts the possible locus that the reads can map to, due to restrictions on the distance between the pair of reads.

The quantification method produces read counts for features of interest. These counts are proportional to the expression levels of this feature but other factors will affect them. The length of the feature, the sequencing depth of the experiment, as well as other experiment biases, can influence read counts. Therefore, a normalisation is required to try to minimize these effects and make read counts comparable both between features and samples. FPKMs

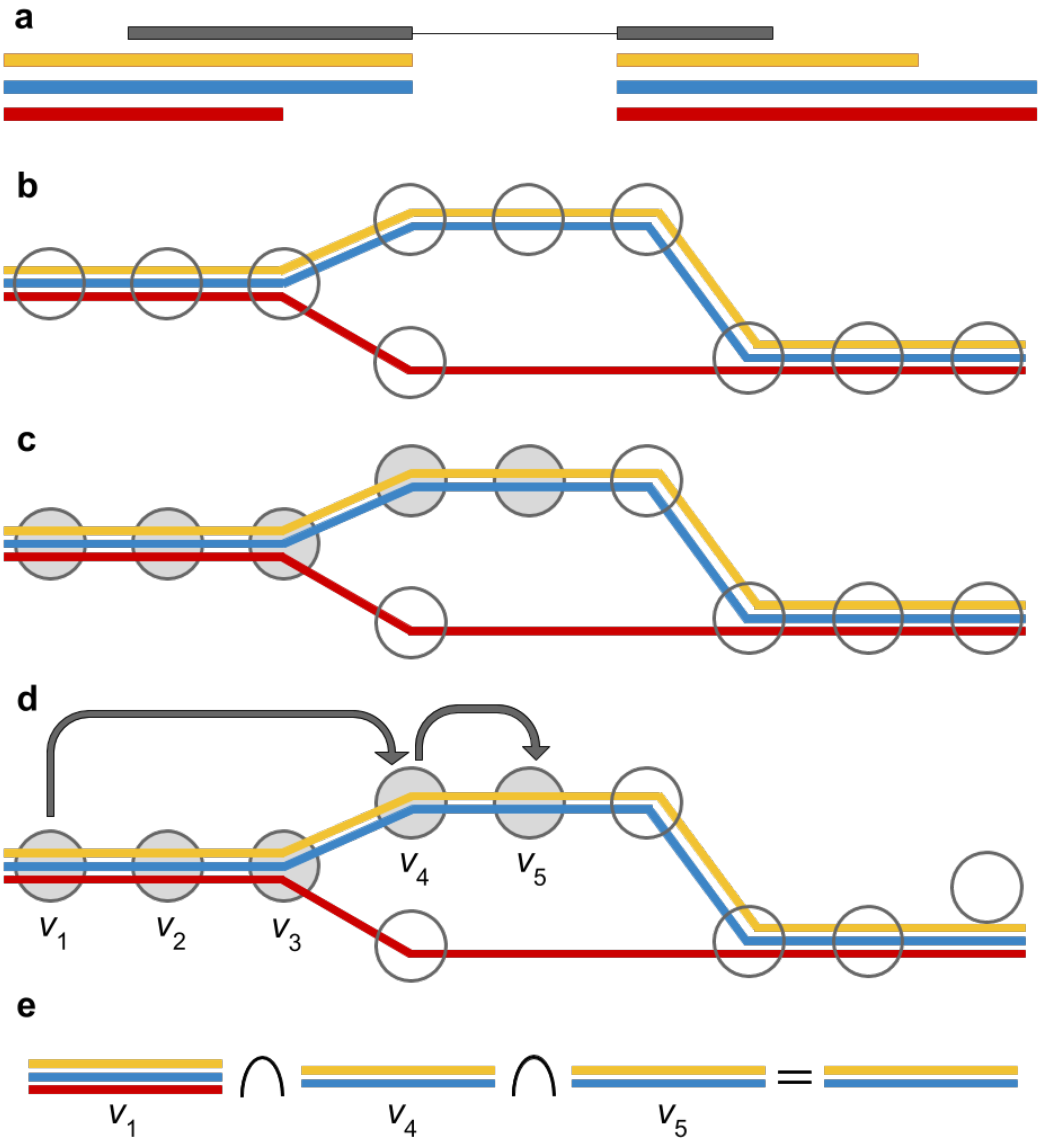


Figure 2.5 Kallisto pseudoalignment strategy. The input of Kallisto is a reference transcriptome and RNA-seq reads. (a) There is a read represented in black and three overlapping transcripts with different colors are shown with two exonic regions each. (b) As the *de Brujin* graph of the transcriptome (T-DBG) is created, an index is constructed. In the graph, the nodes (v1, v2, v3, ...) are k-mers and each transcript is a colored path. The path cover creates a k-mer compatibility class for each k-mer. (c) The black dots are the k-mers of a read that are hashed to identify the k-compatibility class of the read. (d) The black dashed lines represent a skipping method that uses the T-DBG information to skip redundant k-mers and accelerate the process. (e) The k-compatibility class can be determined by intersecting the k-compatibility classes of the k-mers that constitute them (adapted from [15]).

- Fragments per Kilobase per Million mapped reads - is a commonly used normalised expression measure, which is used by Cufflinks [112] but there are others that are similar, such as RPKMs and TPMs, used by other programs. This measure accounts for the length of the feature and for the total number of mapped reads in the dataset. It assumes that the expression levels across samples are identical and each condition has the same amount of mRNA [113]. Due to this assumption, this type of normalisation is not ideal for cases where libraries have significantly different expression levels [114]. Even cases where samples have identical expression levels for most genes but there are some differential expressed genes can be a challenge to interpret using these methods and comparison between libraries in such situations can be misleading.

2.4.8 Differential gene expression

Identifying differentially expressed genes is useful in studies comparing different conditions. In such studies, RNA-seq data is commonly used and several algorithms, such as the one implemented in DESeq2 R package [115], can be used to determine which genes are significantly differentially expressed.

In comparative transcriptomics analysis, it is common to test the null hypothesis that the logarithmic fold change (LFC) of the expression of a gene is zero between two different conditions. The goal of a differential analysis is often to produce a list of genes ranked by p-value. This approach has some limitations because small changes in expression might not be biologically important even if they are statistically significant. Additionally, LFC estimates for genes with low counts can be noisy and the number of differentially expressed genes depends, not only on biological factors but also on the sample size and experimental design. DESeq2 implements a statistical framework that produces a stable estimation of effect sizes, facilitating gene ranking and visualization. It also tests differential expression taking into account biological relevant thresholds defined by the user [115].

DESeq2 is based on DESeq [116], the previous version of the program, which detects very low expression estimates and corrects them using a model of the dependence of dispersion on the average expression levels of all samples. DESeq2 implements additional features that improve gene ranking, hypothesis tests above and below a threshold logarithm transformation for quality assessment and clustering of overdispersed count data. This methodology has high precision and sensitivity while controlling the false positive rate [115].

The DESeq2 method takes a count matrix as input and the reads are modeled using a negative binomial model. The matrix entries indicate the number of reads that mapped unambiguously to a specific gene in a sample and a generalized linear model is fitted to each gene. First the counts of reads are normalised across libraries, then the variability of

each gene is estimated based on all biological replicates data, and finally, each gene is tested to see if it is significantly differentially expressed. It uses Negative Binomial Generalised Linear Models to evaluate the significance of the detected changes. The null hypothesis is that the observed counts across the two conditions being compared are similar enough to be derived from the same distribution. The alternative hypothesis is that two separated distributions would explain those better. The negative binomial distribution is used to avoid over-dispersion produced by previous models [117].

One important step in the workflow of the method is the estimation of the amount of variation across biological replicates because the evaluation of significance depends on it. The low number of replicates in RNA-seq experiments does not allow for the calculation of variance, therefore it has to be estimated based on the data. DESeq2 assumes that genes with similar expression level also have similar variance, so it estimates the variance using the observed dispersion for a given gene and also of all other genes. It fits a regression curve to the data and uses it to modify the observed dispersion values, and decompose the mean into a function of independent variables, taking all sources of variation into account [115].

There are other tools to assess differential expression in genes, such as edgeR [118] and baySeq [119]. There are also tools integrated into the framework of transcript abundance estimation of other programs, with the advantage that these tools might take into account the uncertainty in the process of assigning reads to isoforms. Cuffdiff2 [120] is one of these tools, which is integrated within Cufflinks framework and can be used following the estimation of transcript expression levels. Another example is MMDIFF [121], which is integrated into MMSEQ.

2.4.9 Studying alternative splicing

The detection of differences in alternative splicing can be done by studying differential exon usage (DEU) or differential transcript usage (DTU). The advantage of the first approach is that it does not require transcript reconstruction, which means it does not have all the limitations associated with transcript quantification. Another advantage is that, even if it relies on existing annotation, it makes it possible to indirectly detect new isoforms. However, this comes with the cost of the results being difficult to interpret sometimes, while transcript-based analyses are more straightforward.

The Bioconductor R package DEXSeq [122] uses an exon-based approach and it accounts for biological variation, just like DESeq2 does for differential transcript expression. DEXSeq identifies significant differences in the proportion of reads that overlap each exon, relative to the total number of reads that overlap the full gene. In contrast to DEXSeq, MMDIFF [121] implements a DTU method that uses Bayesian mixed models, integrating the uncertainty in

transcript expression estimates in the regression model that it uses for testing, improving in this way the power to detect alternative splicing events.

Chapter 3

Uncovering the proteome with mass spectrometry

The proteome is the entire set of proteins that is expressed by an organism, tissue or particular cell type. Proteomics is the integrative study of the proteome and its goal is to obtain a complete and quantitative analysis of the expressed proteins. Proteomics includes not only the identification and quantification of proteins but also protein interaction networks, protein complexes, cellular localization and post-translational modifications [123].

Mass spectrometry (MS) is the method of choice to identify and quantify proteins. This technique has advanced considerably over the years with the development of new instrumentation, alternative fragmentation technologies, and advanced data acquisition strategies. As a consequence, there have been improvements in the throughput and the depth of the proteomics analysis. The range of applications has increased as well and it is now possible to do a global analysis of post-translational modifications, reconstruction of protein interaction networks on a large scale, and quantitative proteome profiling of organisms. MS can also be used in systems biology in parallel with other technologies, such as transcriptomics and metabolomics, and both the repositories and proteomics related resources have been increasing as well [124].

Proteomics datasets may contain a significant number of false positives. Specifically, datasets that describe novel genomics events, such as the ones of the human proteome draft paper [125, 126], should use restrictive quality criteria [127], otherwise, false-positive identifications accumulate due to underestimations of the false-discovery rate (FDR). This issue can be avoided by applying robust statistical and computational methods [128] and this well-known pitfall has been addressed more effectively over time. There are several commercial and open source pipelines that allow efficient and transparent analysis of proteomics data. However, the complexity, size, and diversity of proteomics datasets has increased. Therefore

there is always a need for the development of new methods to process high-throughput proteomics data [123].

3.1 Discovery proteomics: shotgun mass spectrometry

Shotgun MS is a data-dependent acquisition (DDA) proteomics method for identifying proteins in a high-throughput context. Its main advantage is the ability to identify and quantify a large number of proteins in a single analysis. However, shotgun proteomics has relatively low reproducibility due to random sampling and low sensitivity when compared to other methods.

Typically in this procedure, the proteins in the sample are firstly digested to obtain peptides. The resulting mixture of peptides can be processed to capture specific classes of peptides, and then the peptides are size selected using liquid chromatography (LC) coupled online to the MS instrument. In a second phase, the ionized peptides are subjected to tandem mass spectrometry (MS/MS), being further fragmented, and the MS/MS spectra are acquired. Finally, the MS/MS spectra are assigned to the peptides.

3.1.1 Protein digestion and separation

The digestion of proteins into peptides is a key step in shotgun proteomics and trypsin is a commonly used enzyme, although alternatively multiple enzymes can also be used. In the case of trypsin, the cleavage occurs after arginine and lysine residues, except if followed by proline, so most peptides are expected to match these cleavage rules and there should be few or no missed cleavages cases.

Protein digestion methods have advantages over intact protein MS/MS sequencing protocols but also have drawbacks. The peptide mixture can be quite complex because each protein produces on average around 50 peptides. That is why protein separation, such as 1-D SDS-PAGE or organellar based separation, should be done prior to digestion. This allows for the total protein content to be separated into sub-fractions, reducing the complexity of the sample and facilitating the identification of proteins. Protein digestion can be followed by peptide enrichment (or depletion) in order to capture peptides with specific properties, such as phosphorylated peptides. Reverse phase chromatography is then used to further separate the resulting peptide fractions, and this procedure is coupled online to the mass spectrometer [123].

3.1.2 Tandem mass spectrometry

The time that the peptides take to elute from the reverse phase column is called the retention time. As they elute, they are ionized, transferred into the gas phase, and then subjected to MS/MS fragmentation for the fragment ion spectra to be obtained. To acquire the data to produce the spectra, the instrument first scans all peptide ions at each time point, recording the MS^1 spectrum, which consists of mass-to-charge ratios (m/z values) and intensities for all peptides. Then selected peptide ions ('precursor' or 'parent' ions), are broken down into fragment ions in the collision cell, predominantly producing b- or y-type ions (N- or C-terminal charged fragments, respectively), and MS/MS, or MS^2, spectra are acquired. They consist of a list of m/z values and intensities for each fragment ion. The amino acid sequence of the peptide is determined using the fragmentation pattern encoded in the MS/MS spectrum [123]. The quantification of each peptide is typically done at the MS^1 spectrum level, although there are exceptions, and identification is performed at the MS^2 level. It is also possible to identify post-translational modifications by including amino acid modifications in the database search (Figure 3.1).

Collision-induced dissociation (CID) is the type of fragmentation most often used but there are others used in particular cases, these include electron transfer dissociation (ETD) and Higher energy Collision dissociation (HCD). Additionally, some instruments are also able to operate in a multi-stage mode that includes automated data-dependent triggering of MS^3 acquisition or Multistage Activation (MSA) [123].

The spectrum information content largely depends on the mass accuracy and resolution of the MS analyzer, so these will affect the subsequent peptide identification. It also should be noted that protein abundances vary drastically between the cells and tissues, and the dynamic range of an MS instrument varies from several parts per million (ppm), in high accuracy instruments, to more than 500 ppm. Even with high accuracy instruments, it is required to fine tune the instrument, control the room temperature, and use both internal or external (computational) calibration. The mass resolution of the instrument affects the accuracy to determine the charge state of the peptide ion [123].

3.1.3 Assignment of peptide sequences to spectra

Following the acquisition of experimental data is the computational analysis to identify the peptides that compose the obtained spectra. Database searching is the main method for assigning a peptide sequence to MS/MS spectra. The computational tools that do this task take MS/MS spectra as input and compare them against theoretical fragmentation spectra which are generated from the proteins in the database. Not all peptides in the database are

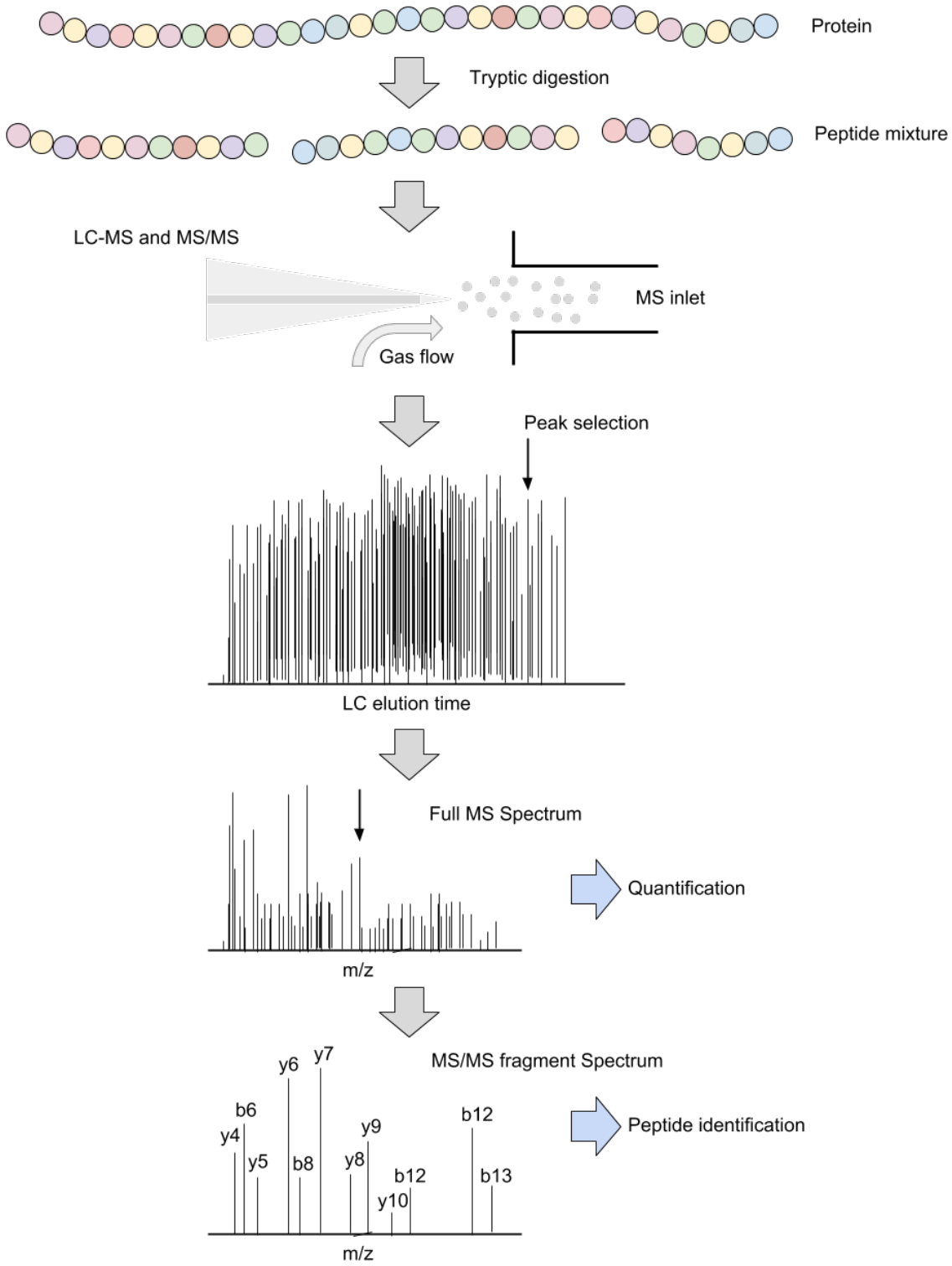


Figure 3.1 An overview of the workflow for shotgun proteomics (adapted from [16]).

used; a much smaller list of candidates is selected using in silico database digestion. There are several criteria that have to be met during the selection phase, such as parent ion mass tolerance, enzyme digestion constraints, and post-translational or chemical modifications. A scoring function is used to evaluate the peptide to spectra assignment and this process also requires some search parameters to be adjusted. These might include the expected type of ions in the spectrum, and the fragment ion mass tolerance. The final output of the program includes a list of peptides for each spectrum, that are ranked according to the search score. Most frequently, only the top best scoring peptide is used as the potential peptide which is then statistically validated [123].

3.2 Targeted proteomics: selected reaction monitoring

Selected reaction monitoring (SRM) is an MS technique applicable in cases where there are specific, predetermined analytes with known fragmentation properties in complex backgrounds. It is useful in such situations because it has higher sensitivity and specificity than traditional shotgun workflows. SRM is commonly used in liquid chromatography-coupled mass spectrometry (LC-MS), which has a capillary chromatography column connected inline to the ionization source of the instrument. SRM makes use of the advantages of triple quadrupole (QQQ) mass spectrometers, that act as mass filters and can selectively monitor a specific analyte molecular ion, as well as, one or more analyte fragment ion generated by collisional dissociation. The number of fragment ions reaching the detector is counted over time and the chromatographic trace registers the retention time and signal intensity. The precursor-fragment ion pairs are called SRM transitions and are measured sequentially and repeatedly. Since the periodicity of the measurement is fast compared to the analyte chromatographic elution, the chromatographic peaks allow for the concurrent quantification of multiple analytes [17, 129].

In a proteomics context, molecular ions with a mass in the range of the target peptide are selected during Q1, the first mass analyzer. In Q2, the molecular ions are fragmented by collision-activated dissociation and in Q3, which is the second analyzer, the fragment ions derived from the targeted peptide are measured (Figure 3.2). The quantification of peptides is supported by the integration of chromatographic peaks for each transition and the quantification might be relative or absolute if heavy isotope-labeled reference standards are used. The identification and quantification of a protein are therefore done by inference using a set of chosen SRM transitions of specific target peptides [17].

The application of SRM to proteomics has certain challenges due to the big dimensions of proteins. It is also difficult to determine the ideal number of peptides and which of them

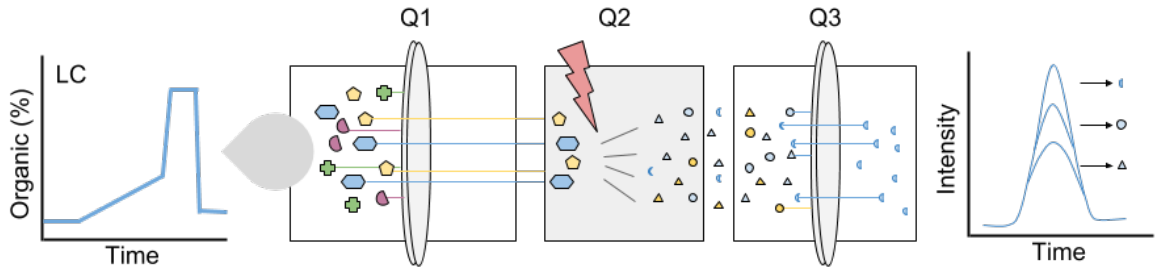


Figure 3.2 Overview of the SRM workflow. The procedure starts with electrospray ionization (ESI). In Q1 molecular ions of an analyte are selected and in Q2 they are fragmented. A specific fragment ion of the target analyte is then selected in Q3 and directed to the detector. Over time the number of target fragment ions is measured, generating the SRM trace. In this figure are shown three srm traces of three transitions that correspond to three different analytes (adapted from [17]).

should be used as input in SRM assays. Even when the peptides are correctly selected, they still yield fragment-ion patterns that are more complex than the ones of metabolites or drugs, which makes it difficult to choose the appropriate SRM transitions. The complexity and dynamic range of the proteome also create challenges that compromise the specificity of the SRM assay and the detection of low-abundance species. All challenges aside, SRM is extremely valuable in the field of proteomics and its range of applications has increased over time [124].

3.3 SWATH-MS

The most widely used methods for the identification and quantification of proteins are the two already mentioned: the main one is shotgun or discovery proteomics; and the second one is targeted proteomics. In both, proteins are first converted into peptides by proteolysis, and then the peptides are separated using liquid chromatography.

There have been developed some alternative strategies to overcome limitations of these methods, not having to rely on the detection or knowledge of the precursor ions to acquire the fragment ion spectra and having increased the reproducibility when compared to traditional DDA workflows. Such methods are designated data-independent acquisition (DIA) methods and they rely on cycle recording throughout the LC time range, executing multiple survey scans and spectra for all precursors present in a predefined isolation window. DIA methods can use isolation windows of various widths and during the scans can lose the link between the fragment ions and the precursors, which hinders the analysis of the acquired datasets. Additionally, when large width windows are used, the number of concurrently fragmented precursors increases, increasing the complexity of the composite fragment ion spectra [130].

As an alternative approach for proteome quantification, SWATH-MS is a technique that combines high specificity of DIA with targeted data analysis building on the high-throughput SRM scoring. SWATH-MS uses a sequential isolation window acquisition principle used in some DIA studies with high-resolution MS instruments. So this is a time- and mass-segmented method which generates fragment ion spectra of all precursor ions in a single injection and records the ensemble of these spectra as complex fragment ion maps. The resulting maps have a very high fragment ion specificity when compared to other techniques. This strategy is designated SWATH-MS in reference to the series of isolation windows acquired for a mass range of a given precursor. The combination of fragment ion maps of high specificity and targeted data analysis allowed by the use of information from spectral libraries confers useful advantages for the qualitative and quantitative analysis of proteomes [131].

Chapter 4

Studying alternative splicing at the RNA level

In this chapter I:

(i) analyse how many of the annotated genes and transcripts are actually expressed at significant levels in normal human tissues;

(ii) analyse the relative expression level of transcripts, in particular investigating dominant transcripts of genes;

(iii) explore the role of alternative splicing via a comparative analysis of dominant transcripts across tissues to identify switch events – changes in the dominant transcript of the same gene – and report the transcript changes controlled by alternative splicing;

(iv) infer what are the potential consequences of switch events at protein level.

The computational analyses herein described were performed by myself under the supervision of Dr. Alvis Brazma. This study is a continuation of the previous work started by Mar Gonzàlez-Porta.

4.1 Introduction

The role of alternative splicing has been extensively debated because of the evidence that has accumulated from RNA-seq transcriptomics and MS proteomics studies [4, 132]. Closely related to the function of alternative splicing is the question of how protein diversity is created.

There are around 22,000 protein-coding genes annotated by Ensembl version 79, and over 160,000 transcripts, of which 86,430 have protein-coding potential [53]. RNA-seq, in particular, has enabled the identification of a large number of transcripts and has shown that over 95% of multi-exon genes have multiple alternative splice isoforms [133]. Almost 19,000 of protein-coding genes have multiple annotated transcript isoforms [134]. However, it has been suggested in transcriptomics studies that the majority of genes have a dominant transcript, a transcript that is significantly more expressed than the others [1]. These results suggest that the role of alternative splicing should be further explored and that this process, along with differential gene expression, might only be part of the explanation for cell complexity and diversity.

It has been shown that although many multi-exon protein-coding genes express multiple transcripts, 79% of the protein-coding genes have a 2-fold dominant transcript, a transcript that is expressed at least twice as much as the second most expressed transcript [1]. This result was in accordance with the prediction made in a study based on EST data [135], in which the authors concluded that 80% of the genes have a dominant isoform. It was also shown that the dominant transcript tends to be recurrent across tissues. This suggests that protein-coding genes have a main transcript, and therefore a single main function. On the proteomics side, it has been shown that most genes have unequivocal peptide evidence for only one of the protein products [4]. This suggests that, although there is the potential for a gene to express different isoforms, for many genes there is one transcript and one protein. However, there are exceptions. Not only do some genes express multiple isoforms, having no dominant transcript, but also some express different dominant transcripts in different tissues. The latter cases are called *switch events* and are relatively rare but not less important. When a gene expresses different dominant transcripts in different conditions, it is meaningful to understand how significant are the changes and what is the impact on the function of the transcript and protein.

Alternative splicing is an important biological process, however, the prevalence of this process, as well as its biological impact must be studied in more detail, and the understanding of its function must be readdressed. Here I expand the previous study done in the group [136], using a larger dataset with more biological replicates, allowing to get better estimates of dominant transcripts. Also, I further explore the differences of alternative splicing at the RNA level across tissues, characterizing the changes controlled by it and inferring the possible consequences of the switch events at the protein level.

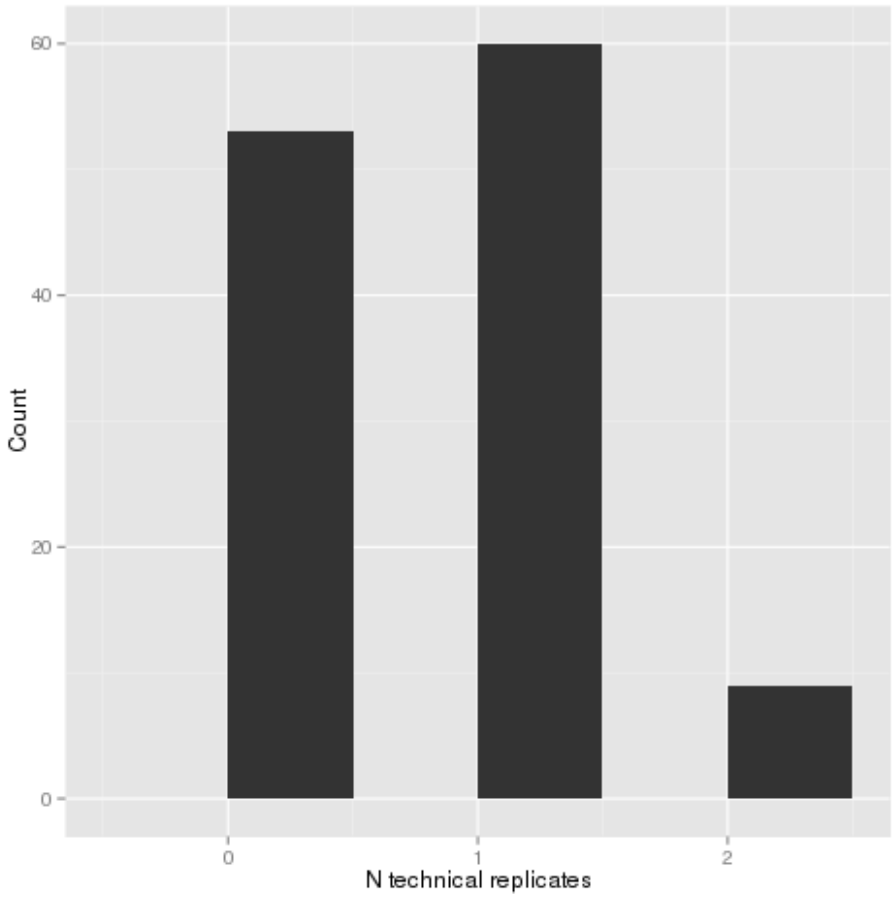


Figure 4.1 Distribution of the number of technical replicates per biological sample in the dataset.

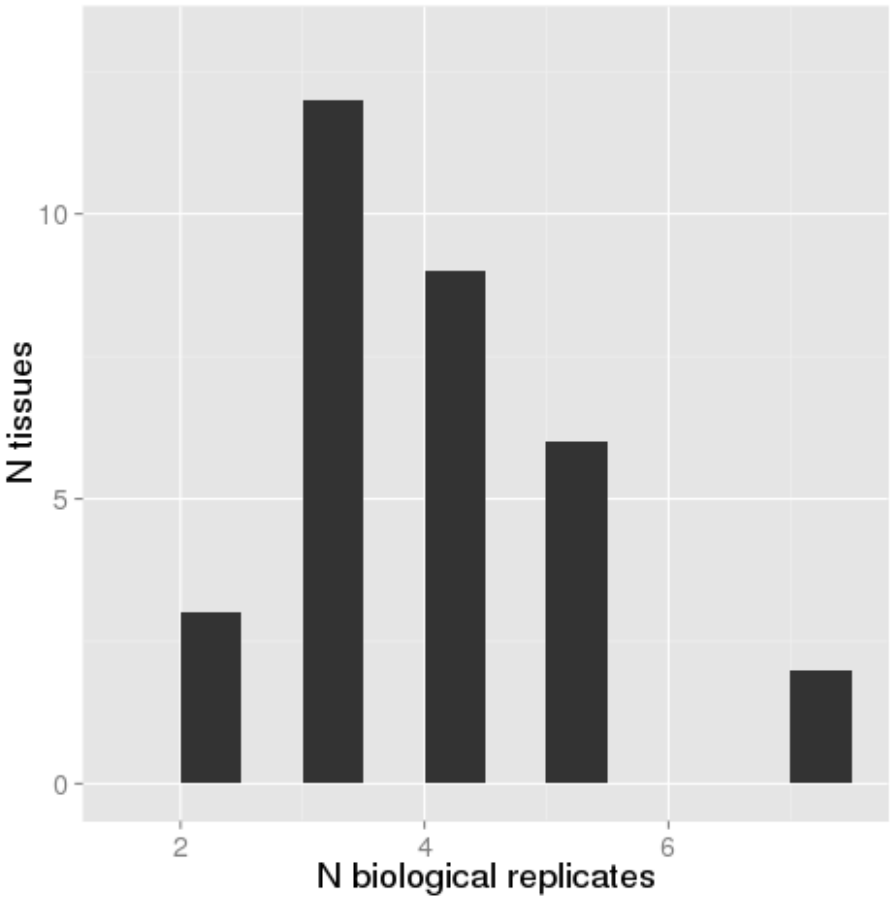


Figure 4.2 Distribution of the number of biological samples per tissue in the dataset.

4.2 Results

The dataset used in this study [137] contains a total of 200 samples of RNA-seq data of coding RNA from tissue samples of 122 human individuals representing 32 different tissues (ArrayExpress accession id: E-MTAB-2836). It is an extended version of the dataset used previously [138]. The number of technical replicates per biological sample varies between 0 and 2 (Figure 4.1) and the number of biological samples per tissue varies between 2 and 7 (Figure 4.2).

The goal of this study was to make comparisons between tissues, therefore the expression levels from individual samples had to be integrated to obtain an expression level for each transcript at the tissue level. To do so, the expression levels of each transcript was averaged across technical replicates, generating an expression level value for each transcript of each biological sample. Then the expression levels were averaged in the same manner across biological samples of the same tissue, generating the expression value per transcript per tissue. The dominant transcripts were then calculated for each gene of each tissue to allow the comparison between tissues and determination of switch events.

4.2.1 Transcript dominance analysis

For genes with more than one transcript, the dominant transcript was determined by calculating the ratio between the most expressed transcript and the second most expressed transcript of the gene (dominance ratio). The transcripts were classified in 2- and 5-fold dominant transcripts, if the ratio was higher than 2 or 5, respectively. Two additional conditions had to be met for a transcript to be considered dominant: it had to be expressed in all samples of the tissue; and it had to be the most expressed isoform in all biological replicates. In the case of genes with only one transcript, this transcript was considered dominant. The dominant transcript analysis was made using quantification scores calculated with both Cufflinks [14] and Kallisto [15].

It was observed that the average number of protein-coding genes expressed in a tissue was 10,138, of which 6942 have a 2-fold dominant transcript and 4721 have a 5-fold dominant transcript (Table 4.1).

The tissue with the highest number of 2-fold dominant transcripts was bladder with 8172 and the lowest was skeletal muscle with 4447 (Table 4.1). In the case of 5-fold dominant transcripts, the maximum was gallbladder with 5404 and the minimum was bone marrow with 3009. This means that on average around 68% (60.3% to 73.6%, SD = 2.71) of protein-coding genes have a 2-fold dominant transcript and 47% (37.8% to 49.3%, SD = 2.18) have a 5-fold dominant transcript. It should be noticed that the number of 2-fold dominant

transcripts is lower than the ones determined in the previous study done in the group [1], where it was found that 79% of protein-coding genes have a 2-fold dominant transcript. This is a consequence of more stringent criteria, requiring that all replicates behave consistently. The change in methodology was aimed at later selecting switch events with extra confidence that are real and not artifacts produced by either the method or the data.

A similar analysis was also done across all tissues. It was observed that 16,811 genes were expressed across all tissues, of which only 1484 had only one annotated isoform (Table 4.2 - *exp genes*). Therefore the vast majority of expressed genes have multiple isoforms. Of the genes being expressed 1587 and 890 had a 2- and 5-fold dominant transcript, respectively, across all tissues in the dataset (Table 4.2 - *2fold-intersect*, *5fold-intersect*). So there are a relatively small number of genes that express a dominant transcript in all tissues. However, it was found that 13,726 and 11,768 genes have a 2- and 5-fold dominant transcript, respectively, in at least one of the tissues (Table 4.2 - *2fold-union*, *5fold-union*), which corresponds to 81% of the expressed genes having a dominant transcript in at least one tissue.

The dominant transcripts determined using Cufflinks were also compared with the ones determined using Kallisto transcript expression scores (Table 4.3). Not only was the number of dominant transcripts similar using both quantification methods but also, the intersection between the sets of dominant transcripts was high: on average 89% (83.4% to 90.7%, SD = 1.36) and 82% (74.5% to 86.1%, SD = 2.48), for 2- and 5-fold dominant respectively. This high overlap provides extra confidence in the relative expression of transcripts and on the determination of transcript dominance.

Only Cufflinks quantification values were used in the determination of switch events, due to the high similarity between the results obtained with both methods.

4.2.2 Effect of varying number of biological samples

The dataset used contains a varying number of samples for each tissue. This can affect the number of dominant transcripts obtained, especially because of the criteria used to select them. In order for a transcript be considered dominant, it has to be expressed in all samples of a tissue and it has to be the most expressed transcript in all samples. There are tissues with 2 up to 7 biological samples (Figure 4.2), so it was predictable that the larger the number of samples, the more difficult would be to meet the previously mentioned criteria.

To evaluate the effect of the number of samples on the determination of dominant transcripts, the samples of a tissue were grouped in all possible combinations of sets with varying number of elements, from 1 to the maximum number of samples. The dominant transcripts were determined for each set of samples and the obtained number was plotted. The results for colon can be seen in Figure 4.4. Colon is one of the tissues with the most

	n exp	n 2-fold	ratio 2-fold	n 5-fold	ratio 5-fold
adipose_tissue	9106	5729	0.63	4123	0.45
adrenal_gland	10833	7495	0.69	5054	0.47
animal_ovary	10403	7156	0.69	4779	0.46
appendix	11040	7458	0.68	4819	0.44
bladder	11453	8172	0.71	5353	0.47
bone_marrow	7954	4793	0.60	3009	0.38
cerebral_cortex	10673	7281	0.68	4892	0.46
colon	9474	6195	0.65	4529	0.48
duodenum	11265	7857	0.70	5085	0.45
endometrium	9883	6523	0.66	4514	0.46
esophagus	10845	7783	0.72	5344	0.49
fallopian_tube	10974	7612	0.69	5279	0.48
gall_bladder	11429	8027	0.70	5404	0.47
heart	9225	6561	0.71	4494	0.49
kidney	10330	6900	0.67	4852	0.47
liver	8371	5870	0.70	4064	0.49
lung	10550	6977	0.66	4883	0.46
lymph_node	9872	6610	0.67	4503	0.46
pancreas	7568	5569	0.74	3633	0.48
placenta	10439	7258	0.70	4938	0.47
prostate	10685	7383	0.69	5063	0.47
rectum	10849	7581	0.70	5184	0.48
salivary_gland	9479	6742	0.71	4567	0.48
skeletal_muscle	6661	4447	0.67	3206	0.48
skin	10384	7203	0.69	5044	0.49
small_intestine	10720	7447	0.69	4993	0.47
smooth_muscle	10318	7192	0.70	4866	0.47
spleen	10817	7258	0.67	4857	0.45
stomach	10242	6867	0.67	4599	0.45
testis	11433	7383	0.65	4947	0.43
thyroid	10249	7167	0.70	4904	0.48
tonsil	10886	7634	0.70	5278	0.48
Average	10137.8	6941.6	0.68	4720.6	0.47

Table 4.1 Analysis of gene expression and transcript dominance per tissue (Cufflinks quantification scores). The columns designate the following categories:
n exp - number of genes expressed per tissue;
n 2-fold and *n 5-fold* - number of genes with 2- and 5-fold dominant transcripts;
ratio 2-fold and *ratio 5-fold* - ratio between the number of genes with dominant transcripts and the number of genes expressed.

	exp genes	2fold-intersect	5fold-intersect	2fold-union	5fold-union
all	16811	1587	890	13726	11768
one iso	1484	171	171	1484	1484

Table 4.2 Analysis of gene expression and transcript dominance across tissues. The rows contain the number of genes of the following categories:
all - set of all protein-coding genes in the annotation;
one isoform - set of protein-coding genes with only one annotated transcript.
The columns designate the number of genes of the following categories:
exp genes - expressed genes;
2fold-intersect - genes with a 2-fold dominant transcript across all tissues;
5fold-intersect - genes with a 5-fold dominant transcript across all tissues;
2fold-union - genes with a 2-fold dominant transcript in at least one of the tissues;
5fold-union - genes with a 5-fold dominant transcript in at least one of the tissues.

samples, containing a total of 7. As predicted, the number of samples affects the number of dominant transcripts obtained and, the highest the number of samples, the lowest the number of dominant transcripts. The biggest drop on the average number of dominant transcripts (represented by the red line in Figure 4.4) occurred between single samples (~6600 transcripts) and sets of 2 samples (~5700 transcripts), which is the minimum number of samples per tissue.

The effect of the number of samples was noticed and the method to determine dominant transcripts was maintained because it is a more conservative approach than trying to correct this effect. It was also considered that any attempts to correct this effect could artificially introduce dominant transcripts, possibly creating bias in the switch event determination step.

4.2.3 APPRIS analysis

The APPRIS database was developed within the GENCODE consortium [139] with the aim of annotating alternative protein isoforms with functional information. APPRIS has a series of modules that evaluate each isoform making use of functionally important protein residues, 3D structure information, signal peptides, Pfam domains [140], and attributes a score for each isoform model according to its cross-species conservation.

APPRIS uses reliable annotations for function, protein structure, and cross-species conservation. With this information, a reference CDS is selected as the 'principal' isoform. This specific isoform is the one containing the most conserved features. Some of the other isoforms might be designated 'alternative', in case they contain unusual, missing or non-conserved features [2]. Besides the isoforms being classified in principal or alternative, they

	2-fold_c	2-fold_k	2-fold_i_r	5-fold_c	5-fold_k	5-fold_i_r
adipose_tissue	5729	6492	0.91	4123	4383	0.85
adrenal_gland	7495	8266	0.89	5054	5168	0.82
animal_ovary	7156	8046	0.89	4779	4993	0.82
appendix	7458	8182	0.89	4819	4688	0.80
bladder	8172	9180	0.88	5353	6086	0.85
bone_marrow	4793	5437	0.83	3009	2921	0.74
cerebral_cortex	7281	8088	0.90	4892	4981	0.82
colon	6195	7019	0.91	4529	4795	0.85
duodenum	7857	8663	0.89	5085	5151	0.79
endometrium	6523	7353	0.89	4514	4669	0.82
esophagus	7783	8715	0.88	5344	5914	0.85
fallopian_tube	7612	8748	0.89	5279	5764	0.84
gall_bladder	8027	8991	0.88	5404	5807	0.82
heart	6561	7337	0.89	4494	5005	0.86
kidney	6900	7743	0.91	4852	4940	0.83
liver	5870	6516	0.89	4064	4395	0.86
lung	6977	7815	0.90	4883	5062	0.84
lymph_node	6610	7293	0.88	4503	4687	0.82
pancreas	5569	6062	0.89	3633	3952	0.84
placenta	7258	8084	0.90	4938	5172	0.84
prostate	7383	8100	0.89	5063	5096	0.82
rectum	7581	8454	0.88	5184	5323	0.80
salivary_gland	6742	7484	0.90	4567	4793	0.83
skeletal_muscle	4447	5029	0.90	3206	3214	0.81
skin	7203	7978	0.87	5044	4981	0.79
small_intestine	7447	8240	0.90	4993	4968	0.81
smooth_muscle	7192	8064	0.88	4866	5284	0.84
spleen	7258	7987	0.88	4857	4732	0.81
stomach	6867	7657	0.90	4599	4586	0.81
testis	7383	8392	0.88	4947	5006	0.78
thyroid	7167	8127	0.88	4904	5349	0.83
tonsil	7634	8646	0.89	5278	5633	0.83
Average	6941.6	7755.9	0.89	4720.6	4921.8	0.82

Table 4.3 Comparison between the set of dominant transcripts found using quantification scores from Tophat2+Cufflinks and Kallisto. The column names with the suffix '_k' correspond to Kallisto and the '_c' to Cufflinks. The columns with the suffix '_i_r' contain the ratios calculated by dividing the intersection by the union of the two sets.

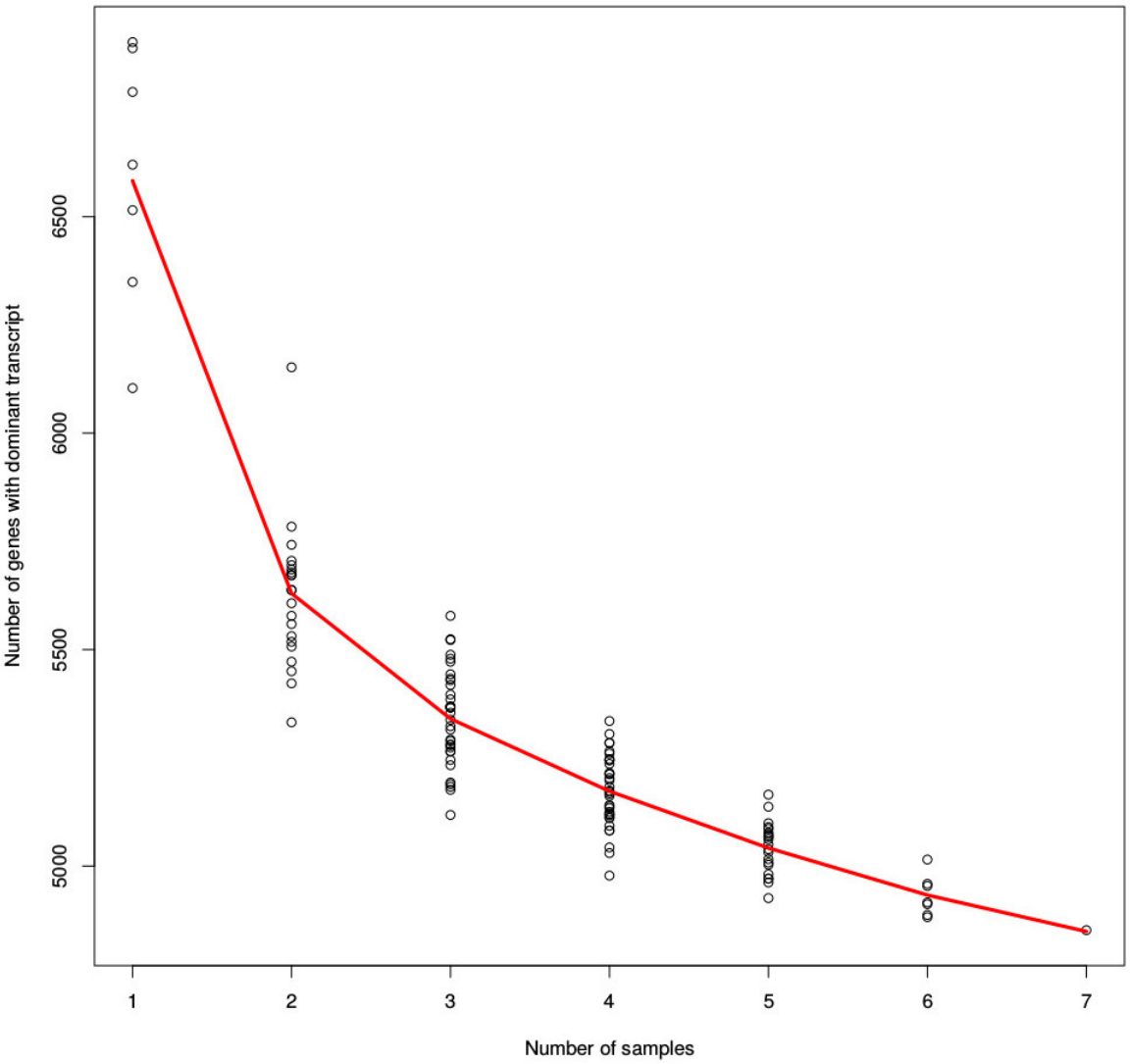


Figure 4.3 Number of genes with 5-fold dominant transcripts for all possible sets of biological samples for colon tissue. On the x-axis is the number of samples per set and each dot represents a set of samples. The red line unites the average number of dominant transcripts for each specific set size.

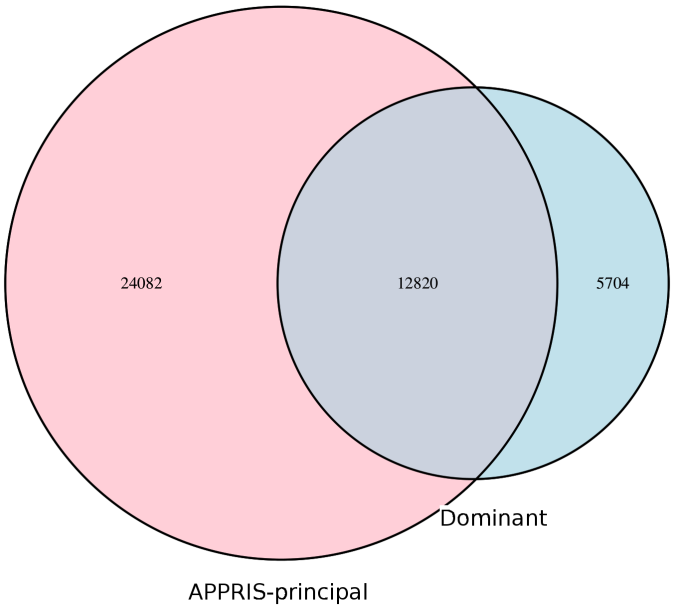


Figure 4.4 Overlap between 2-fold switch events and APPRIS principal isoforms. Transcript quantification values determined with Cufflinks.

are also given numbers from 1-5 or 1-2, respectively, according to the confidence in the classification and the criteria they met.

APPRIS was developed by Michael Tress' group, which has been publishing studies [141, 142] supporting the idea that in most cases genes have a single dominant protein isoform and that the others might be alternative isoforms expressed less frequently in specific tissues, development stages, or they might simply have a short half-life.

There are commonly used strategies to determine dominant isoforms, such as selecting the longest. This is a simple criterion that often fails and between 20-25% of the isoforms selected in this way are likely not to be the main protein product of the gene [141]. On the other hand, APPRIS principal isoforms are often the main isoforms detected in proteomics studies, this is true for around 98% of comparable genes. APPRIS effectiveness relies on its ability to identify regions of conserved structure or function, which are often lost in alternative isoforms, as well as non-conserved exons that can also be found inserted in conserved regions [2].

The list of dominant transcripts was further compared with the principal isoforms for the same gene in the APPRIS database. The aim was to determine if the main transcript and protein isoforms were the same, and infer if there is a relationship between transcript dominance and protein conservation.

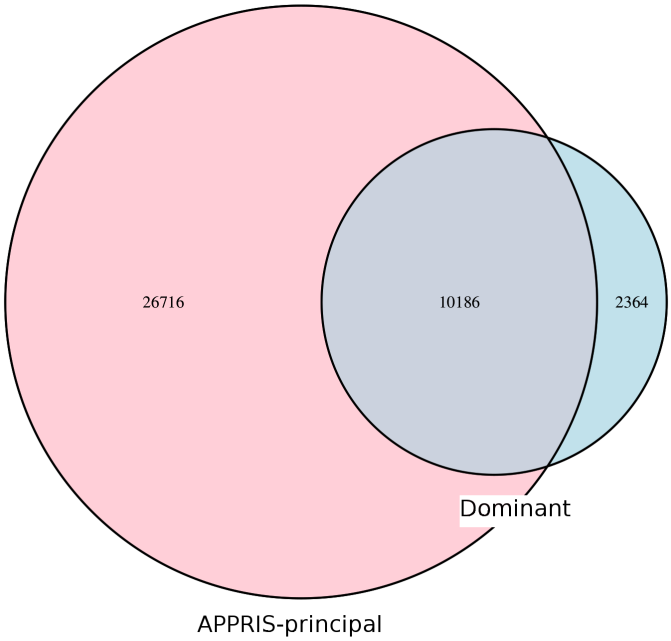


Figure 4.5 Overlap between 5-fold switch events and APPRIS principal isoforms. Transcript quantification values determined with Cufflinks.

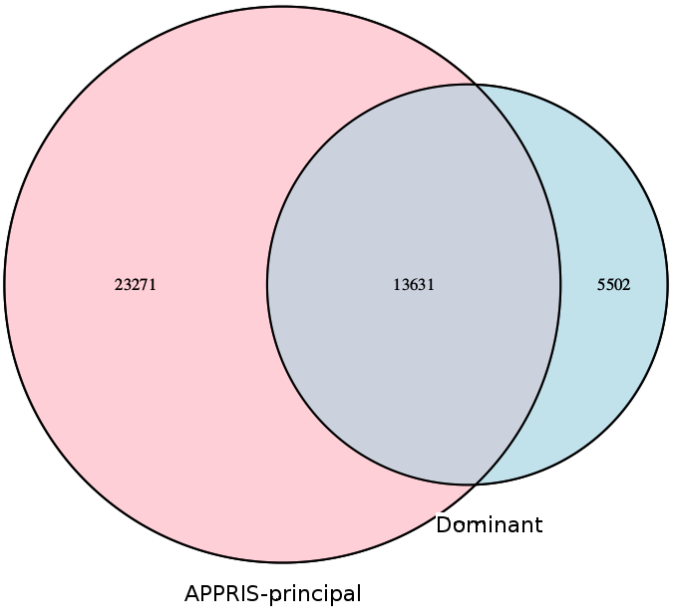


Figure 4.6 Overlap between 2-fold switch events and APPRIS principal isoforms. Transcript quantification values determined with Kallisto.

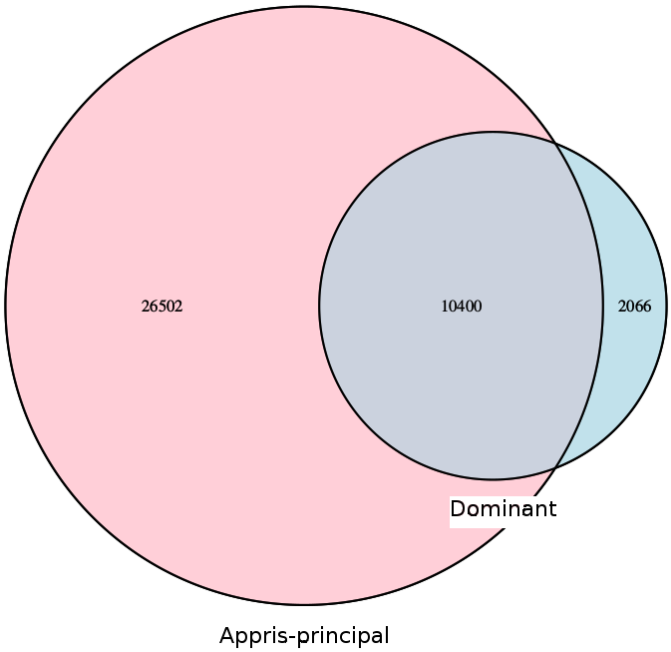


Figure 4.7 Overlap between 5-fold switch events and APPRIS principal isoforms. Transcript quantification values determined with Kallisto.

It was observed that 68% of the 2-fold dominant transcripts and 81% of the 5-fold dominant transcripts are also APPRIS principal transcripts (Figure 4.4, Figure 4.5). This corresponds to 12,820 and 10,186 genes for 2- and 5-fold dominant transcripts, respectively. The same analysis was done with the lists of dominant transcripts determined with Kallisto expression values and a similar result was obtained: 71% of the 2-fold dominant transcripts are APPRIS principal isoforms (Figure 4.6); 83% of the 5-fold dominant transcripts are principal isoforms (Figure 4.7). This corresponds to 13,631 and 10,400 genes for 2- and 5-fold dominant transcripts, respectively. Both Cufflinks and Kallisto expression values yielded similar results, which is related to the fact that dominant transcripts calculated with both strategies also tend to be the same.

The overlap of the 5-fold dominant transcripts with APPRIS principal isoforms was higher than the 2-fold dominant transcripts. This suggests that there is a relationship between the relative expression level of a transcript and its likelihood to be functional.

4.2.4 Switch events

The previous results suggest that most genes express a single dominant transcript in a specific condition. Nevertheless, it is relevant to know if the dominant transcript of a gene is the

same across conditions or if it changes. The cases where a gene expresses different dominant isoforms in different tissues are designated switch events. These events can help to understand the function of alternative splicing, how it operates and how conserved it is across normal human tissues.

Formally a switch event can be defined in the following way (definition adapted from the one used in *Switch Seq* [136]). Given a gene G, a pair of transcripts I_k and I_l, and two tissues T_i and T_j, we say that gene G undergoes an x-fold switch between transcripts I_k and I_l in tissues T_i and T_j, if G is expressed in both T_i and T_j and the ratio of the expression of I_k to I_l is at least x in T_i and no more than $1/x$ in T_j. As an example, gene SEMA4D (ENSG00000187764) has ENST00000356444 as 2-fold dominant transcript in pancreas and ENST00000455551 as 2-fold dominant transcript in heart: this is a 2-fold switch event.

The software used to determine switch events was developed using Python programming language and all the switch events with differentially expressed genes were filtered out from this analysis to assure that the differences in expression were caused by isoform switches and not gene differential expression. The differentially expressed genes were identified using DESeq2 (p-value < 0.01) [115].

To determine switch events, the 32 tissues in the dataset were compared in a pairwise manner. For each gene, the 2- and 5-fold dominant transcripts were compared across conditions. A change of a 2-fold dominant transcript to another in a different condition is designated 2-fold switch event (Figure 4.8) and a 5-fold dominant transcript changing to another is a 5-fold switch event (Figure 4.10). The heatmaps show the number of switch events that occur between a given pair of tissues. As can be seen, the number of switch events is relatively low, considering that the number of genes expressed in a given tissue is around 10,000 and 68% of them have at least a 2-fold dominant transcript (Table 4.1). On average, for a given pair of tissues, there are 30.3 2-fold switch events (Figure 4.13) and there are 1968 genes involved across all tissues. There are also 3.7 5-fold switch events (Figure 4.14) involving 367 genes across all tissues. So even if there are some cases of dominant transcripts switching, most are conserved across tissues.

In the case of 2-fold switch events, the maximum number of switches was 93 and it was found between bladder and cerebral cortex. The tissue with the highest median value of switches was bone marrow with 52. The minimum number of switches was found between endometrium and adipose tissue, prostate, and lung, as well as endometrium and lung. The tissue with the lowest median of switches was lung with 10.

In the case of 5-fold switch events, the maximum number of switches was 15 and it was found between kidney and smooth muscle. The tissue with the highest median value of switches was skeletal muscle with 7. The minimum number of switches found between a pair

of tissues was 0 and it was found in a large number of cases. Thus supporting the hypothesis that most genes have a single dominant transcript.

To determine the level of similarity between tissues, the matrixes of the heatmaps were used to do multidimensional scaling (MDS) analyses using the MDS function in python's scikit-learn library with the default parameters [143]. MDS was used instead of principal component analysis (PCA) because the matrix contains the number of differences in dominant transcripts which can be interpreted as a measure of distance between tissues. In both 2- and 5-fold switch events, it is not clear if there is a biological reason for the way the tissues cluster but there are definitely some tissues that are isolated and more distant from the main cluster of tissues in the MDS plots, implying that these cases are less similar to any of the others. In the case of the 2-fold switch events, it can be observed that bone marrow, testis, liver, cerebral cortex, skeletal muscle, and kidney are tissues that seem to be more distinct from the others (Figure 4.9). In the case of 5-fold switch events, the tissues that stand out are duodenum, kidney, testis, and skeletal muscle (Figure 4.11).

During the study, it was hypothesized that genes involved in switch events could tend to have a higher number of isoforms because can potentially undergo more alternative splicing events. This was however proven not to be the case. Four different sets of transcripts were compared (Figure 4.12). All four sets of genes compared seem to have a similar distribution of the number of transcripts, which indicates that the number of isoforms does not have a strong effect on the determination of dominant isoforms.

Although the number of switch events in normal tissues is low, it does not mean that these cases are not meaningful. The following steps of the study examine these events in more detail.

4.2.5 Effects of the dominant transcript definition on switch events

As mentioned before, the number of switch events detected was relatively low. To evaluate what is the impact of the method of selecting dominant transcripts in the number of switch events, the selection criteria were changed. One of the conditions to consider a transcript as dominant was that the transcript had to be the most expressed in all biological samples of the tissue. This last condition will be designated *support*, as in all samples supporting the specific transcript as major transcript. This condition was removed to test the effect that it has on switch events and to have an indication of the variability of dominant transcripts between biological samples. It can be seen in the heatmaps of 2- and 5-fold switch events (Figure 4.13, Figure 4.14) that removing this criterion considerably increases the number of switch events. In the case of 2-fold switch events, it might increase the number of events more than 4-fold and in some 5-fold switches there was an increase of more than double in

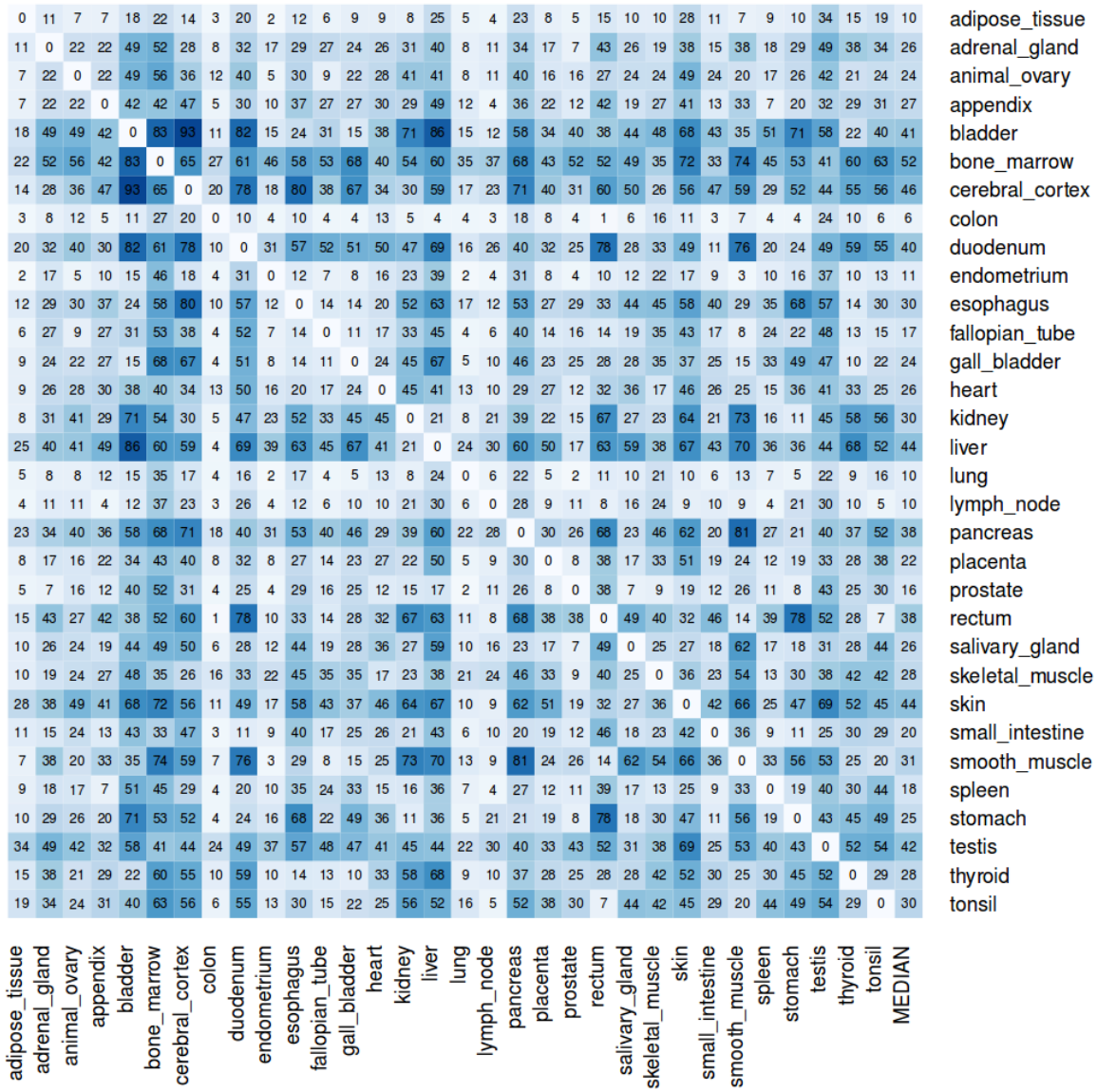


Figure 4.8 Number of 2-fold switch events for all pairs of tissues in the dataset. The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

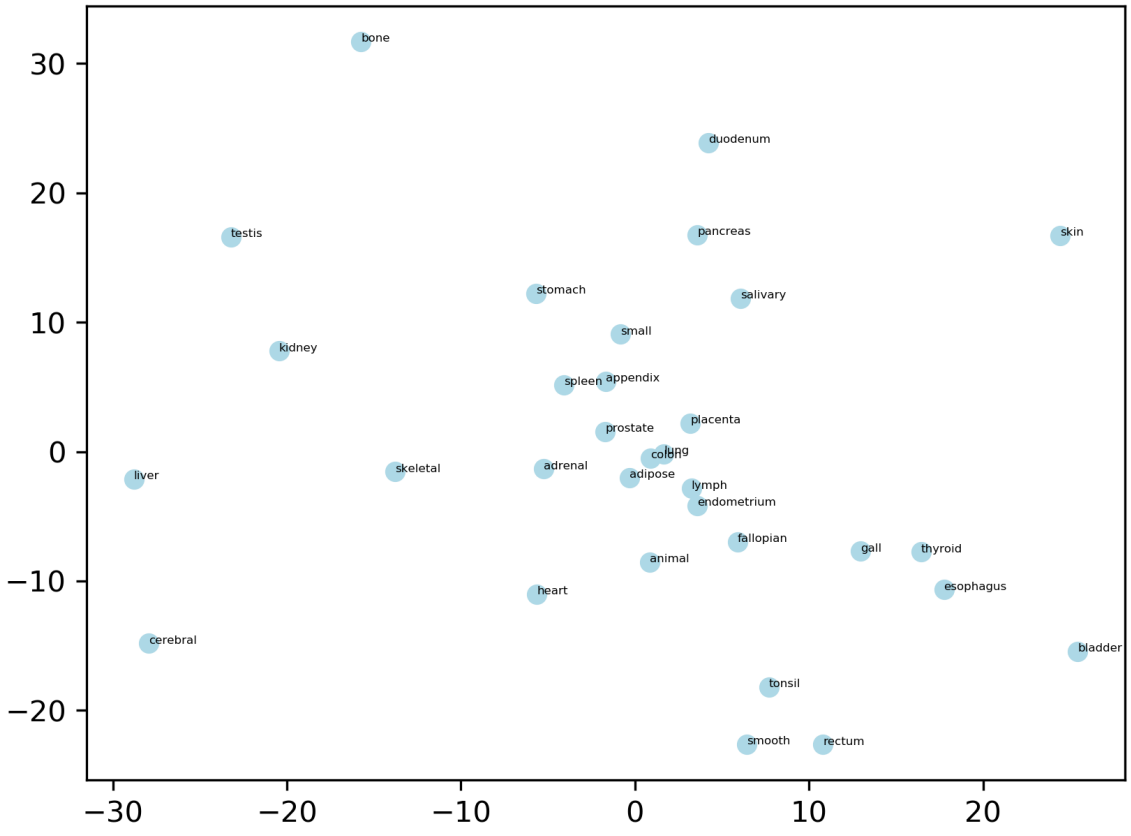


Figure 4.9 Multidimensional scaling applied to 2-fold switch events. To facilitate the visualization, only the prefix of the tissue names are shown.

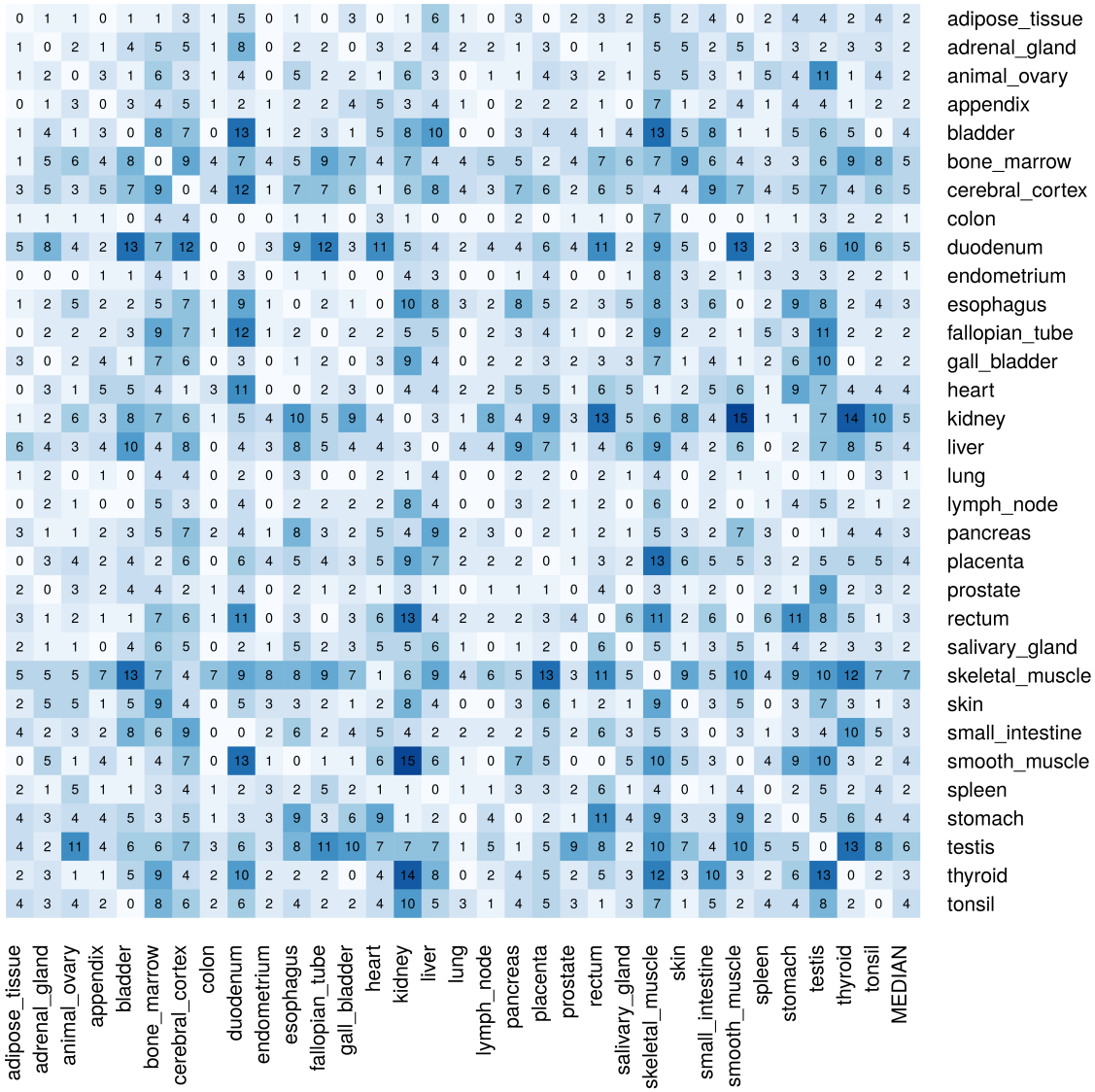


Figure 4.10 Number of 5-fold switch events for all pairs of tissues in the dataset. The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

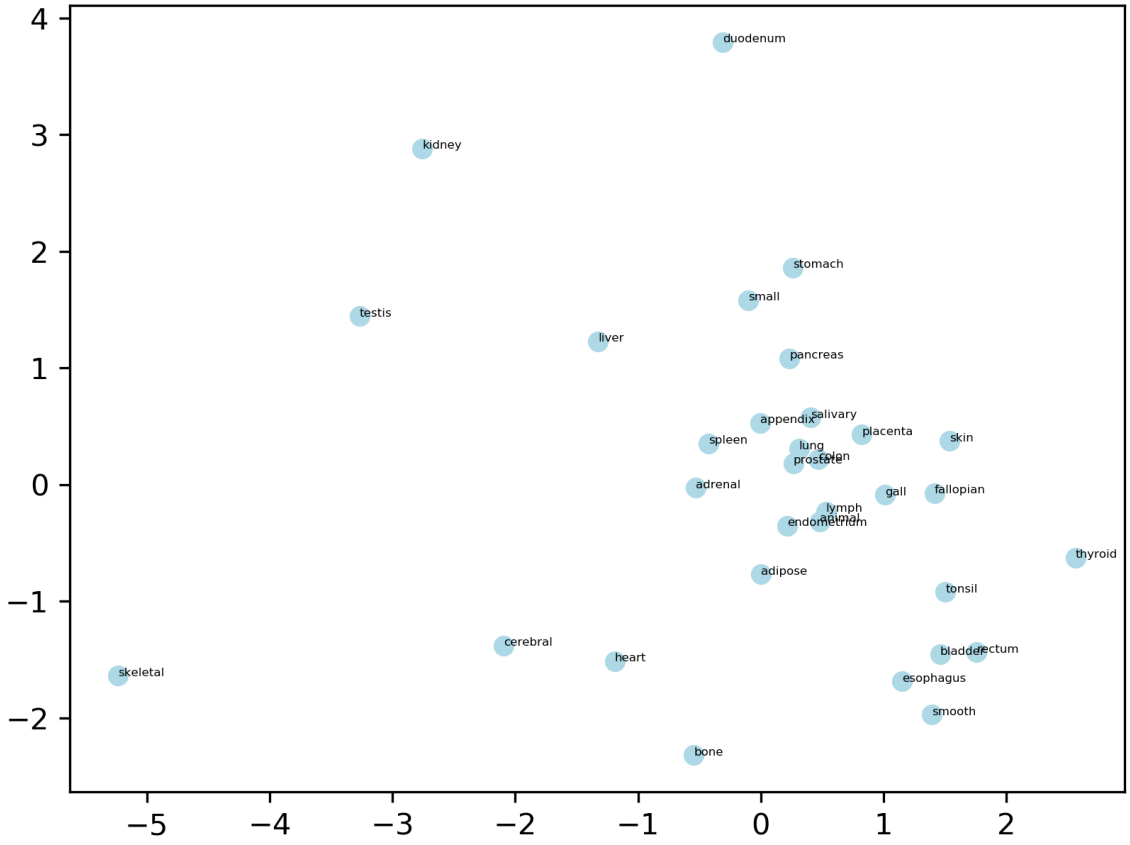


Figure 4.11 Multidimensional scaling applied to 5-fold switch events. To facilitate the visualization, only the prefix of the tissue names are shown.

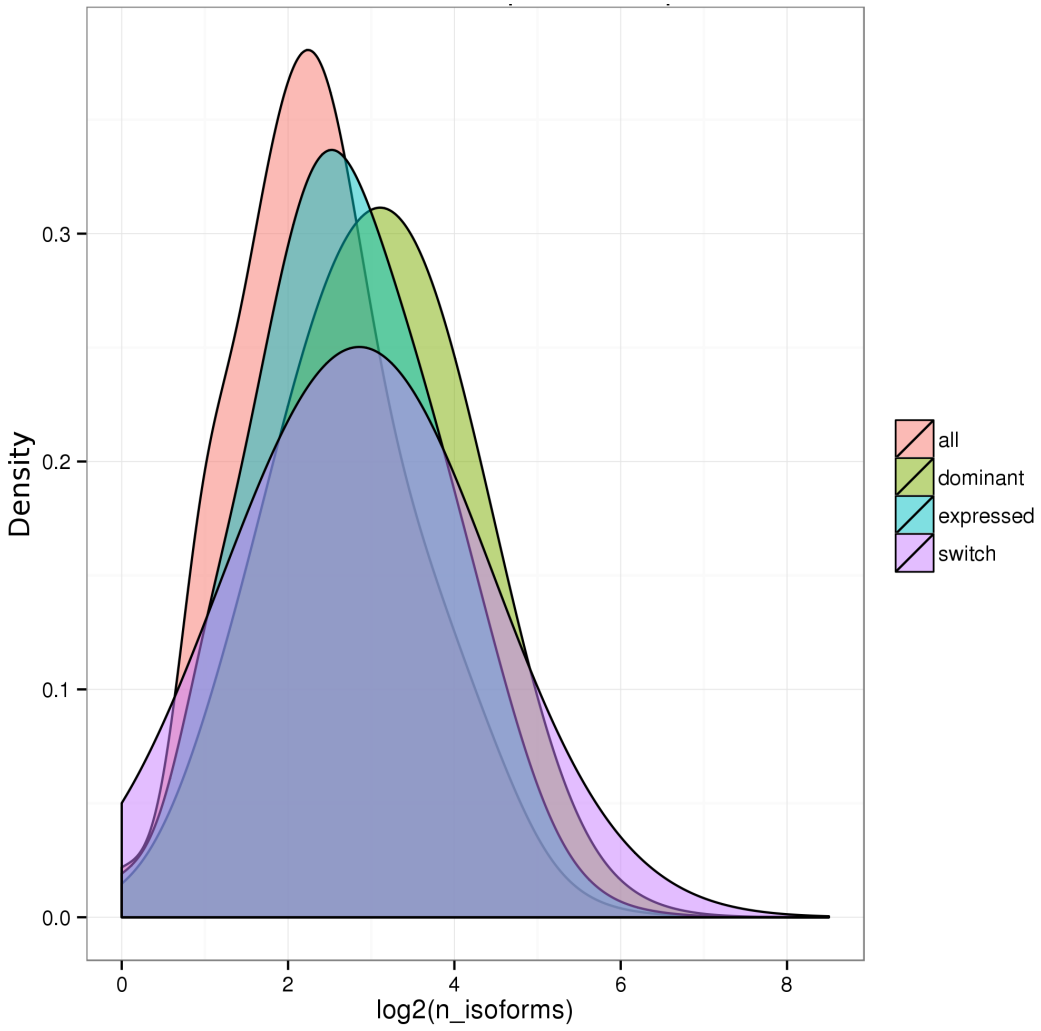


Figure 4.12 Distribution of the number of isoforms (*n_isoforms*) per gene for four categories of genes: all annotated protein-coding genes; genes with dominant transcripts; genes with expressed transcripts; and genes whose transcripts switch.

some cases. This means that for some genes, there is variability between biological samples of a tissue and some transcripts might be dominant in only some of the samples.

The previously defined criteria were maintained because the aim was to find alternative splicing events conserved across samples with a strong signal, which could also confer extra confidence to the analysis. With that said, given these results, it is clear that stringent criteria come with the cost of considerably reducing the number of switch events. It also does not mean these additional switch events are not real because alternative splicing can be variable across individuals.

4.2.6 Comparison of exons in switch transcripts

To evaluate the impact of switch events on the transcriptome, the pairs of transcripts that switch were compared. First, a comparison was made between the exons that constitute both isoforms by comparing the annotated exon identifiers for each transcript. It can be seen that most switch transcripts differ in a relatively small number of exons, being 4 the most common number of different exons between two switching transcripts (Figure 4.15). It should be noticed that there are transcripts differing by a very high number of exons.

Although the number of exon changes seems to be small, even a change in a single exon can add or remove a protein domain, which might add or remove a specific function to a protein. Therefore further investigation is needed to understand the changes induced by switch events.

4.2.7 Alternative splicing types

To further explore the changes driven by alternative splicing, the types of alternative splicing that occur between pairs of transcripts in switch transcripts were analysed. The two most common types of alternative splicing found were alternative $3'$ and alternative $5'$ splice site selection, corresponding to 23.8% and 21.3% respectively. These were followed by alternative polyadenylation and alternative promoter, 17.6% and 16.3% respectively (Figure 4.16). The first two cases account for 45% of the cases and both represent relatively small changes in the transcripts because both events are changes in a splice site of an exon (changes of less than one exon between the two transcripts). Some of these cases were checked individually and it was observed that some exons differed in only a few bases even though they were annotated as different exons. So although the previous analysis showed that the number of exons that differ between isoforms was between 1 and 78 with 4 being the most common number (Figure 4.15), it does not mean that all exon changes have a big impact or cause a large number of bases to differ, in fact, the opposite seems to be more frequent.

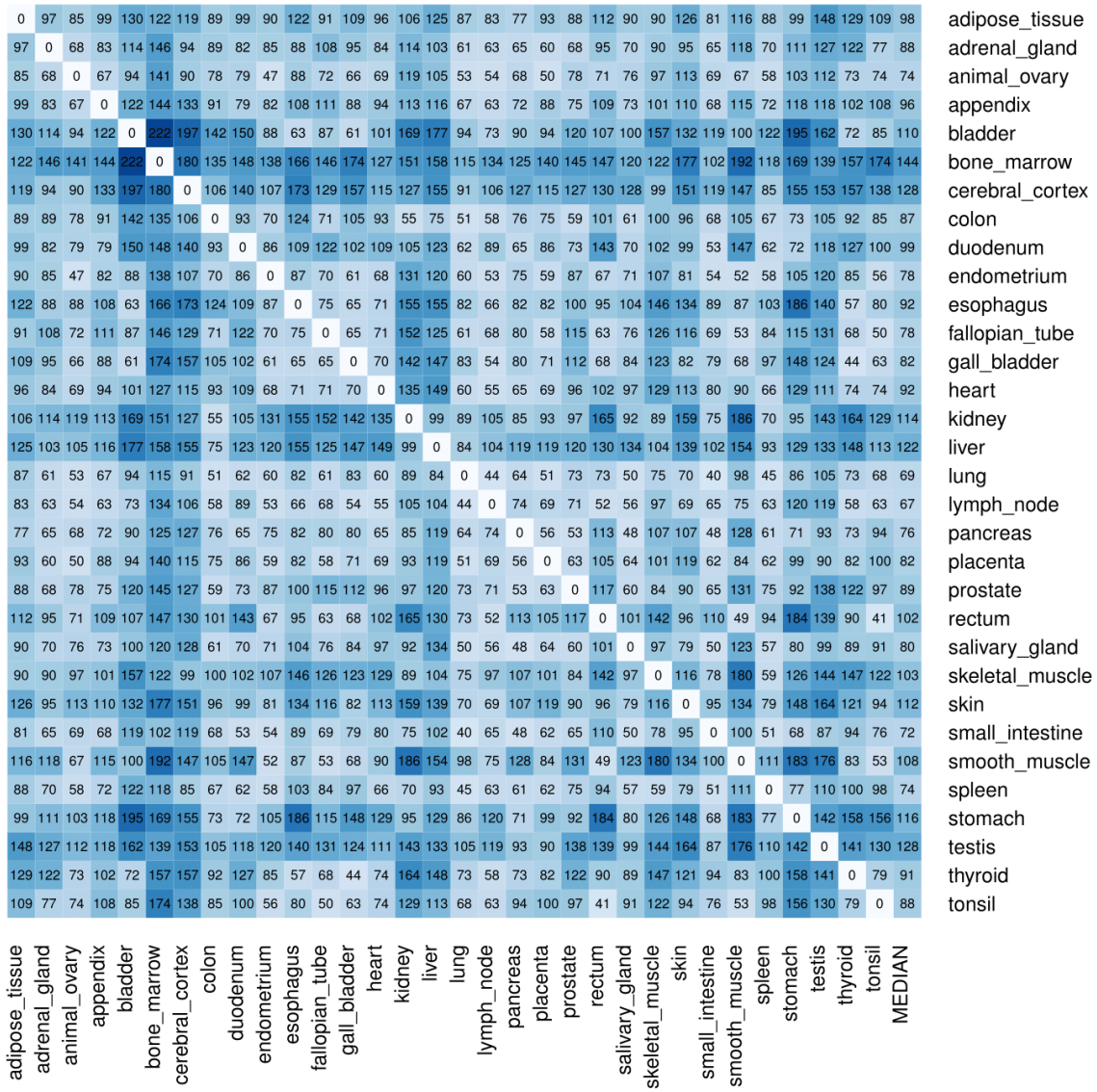


Figure 4.13 Number of 2-fold switch events for all pairs of tissues in the dataset. Results obtained with relaxed criteria for determining transcript dominance.

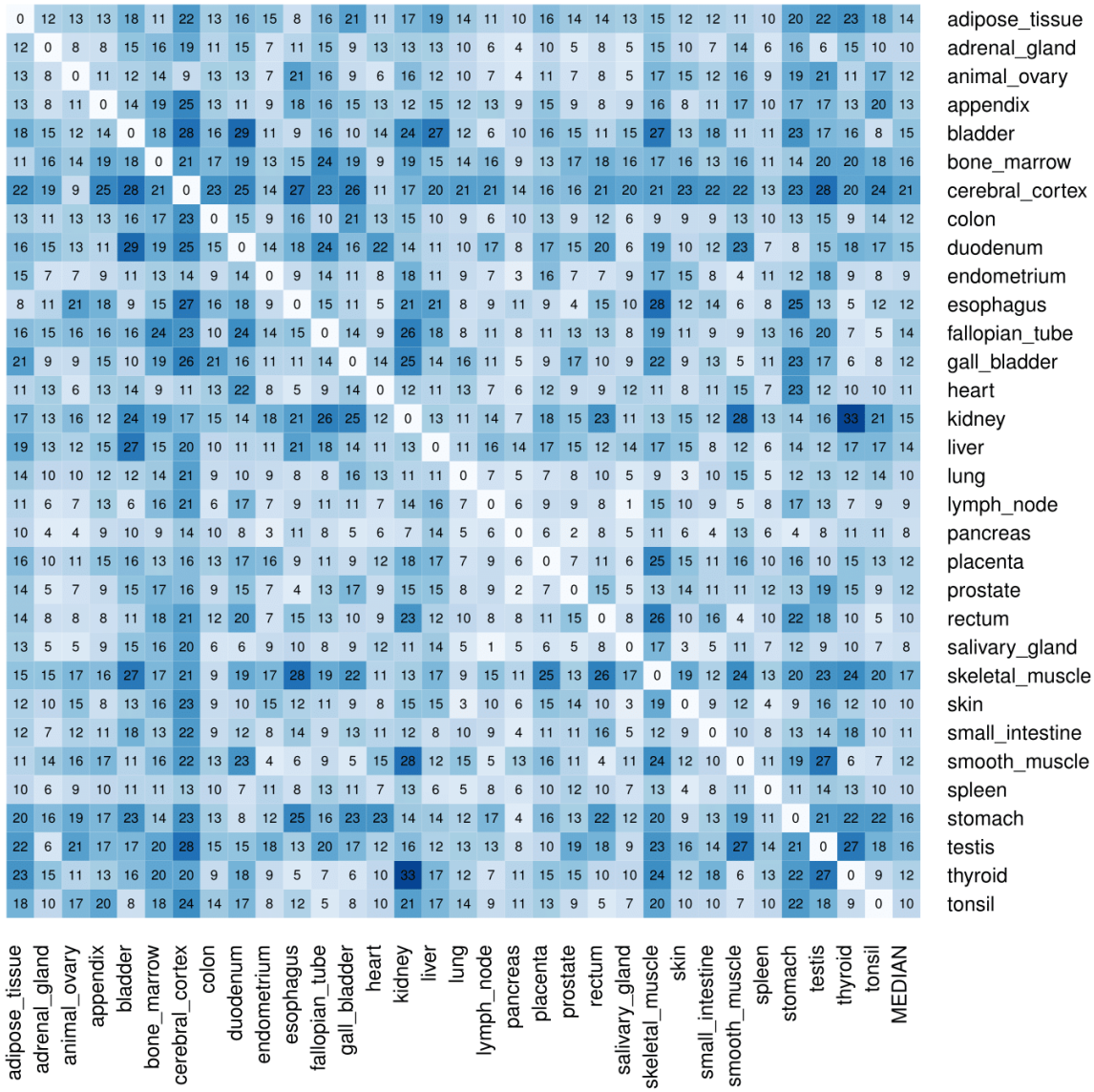


Figure 4.14 Number of 5-fold switch events for all pairs of tissues in the dataset. Results obtained with relaxed criteria for determining transcript dominance.

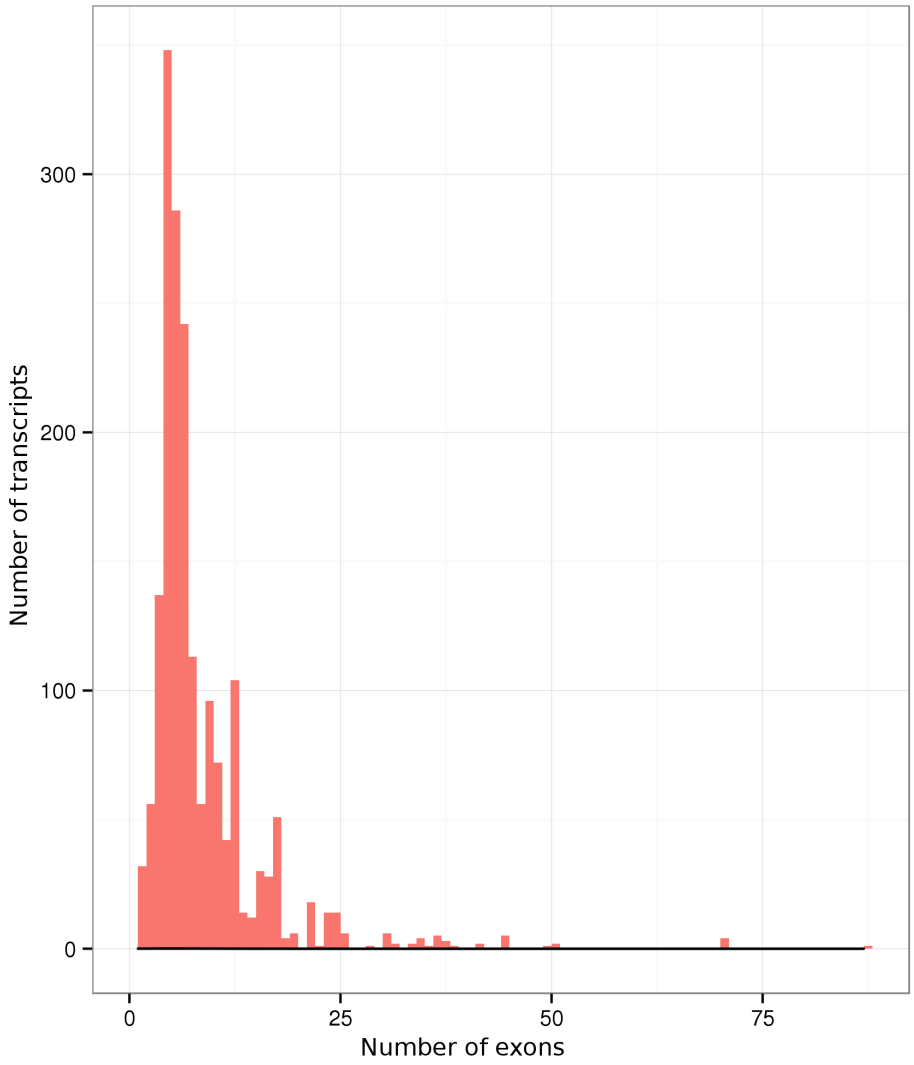


Figure 4.15 Distribution of the number of exons that differ between pairs of transcripts in a 5-fold switch event. On the x-axis is represented the number of exons that are different and on the y-axis is the number of pairs of transcripts (number of switches).

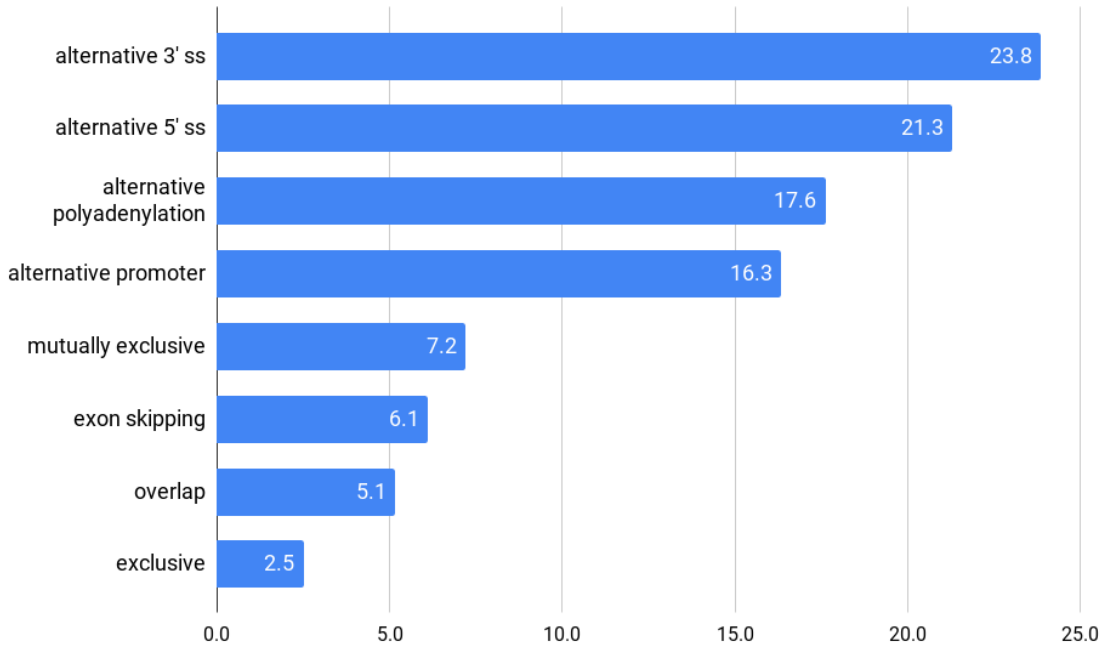


Figure 4.16 Percentage of alternative splicing types occurring between transcripts in switch events. The types of alternative splicing are alternative 3′ splice site selection; alternative 5′ splice site selection; alternative polyadenylation; alternative promoter, mutually exclusive exons; exon skipping; overlapping exons; and exclusive exon. Besides the most common types of alternative splicing, two other categories were added: "overlap", for cases of overlapping exons that do not fit any of the other splicing categories; and "exclusive", for cases of exons that are exclusive to one of the transcript and do not fit any other splicing category.

4.2.8 Sequence identity

In order to understand the impact of the exon changes in the sequence of the transcripts, the nucleotide sequence identity of switch event isoforms was calculated using BLASTN [18], a commonly used local sequence aligner, and using an implementation of the Needleman-Wunsch algorithm [144, 145], a global alignment algorithm.

Sequence identity can be used as an indicator of how similar the protein structure and function are, although it has limitations and some indels and non-homologous substitutions produce proteins with completely different folds. Therefore, not only some proteins have relatively high sequence identity and have different functions, but also there are cases of proteins with as little as 10% identity which conserve the same overall structure [56].

When the sequence identity of pairs of switch transcript sequences was determined, it was observed that there is a tendency for the transcripts to have high sequence identity (Figure 4.17, Figure 4.18). Although there are definitely cases of pairs of transcripts with low sequence identity, this suggests that transcripts often switch to another similar one. Therefore, it is important to investigate if the function of the potentially translated proteins is affected or not.

4.2.9 Exon overlapping analysis

As mentioned before, there is a considerable number of exons that change between switch transcripts. On the other hand, these same transcripts have a high sequence identity. To understand why exon differences do not translate into more significant changes in sequence identity, the exons were compared. In particular, it was checked if the exons that differ between switch transcripts overlapped with other exons in the annotation. This was done by comparing the annotated exon coordinates between overlapping exons. First, it was determined how many differently annotated transcripts overlap in a switch event (Figure 4.19) and in most switches, the number of exons that overlap is less than 5. Then, the percentage of overlap between the exons was calculated and it was observed that there is a considerable number of exons that overlap more than 90% (Figure 4.20). This explains in part why the sequence identity between switch transcripts tends to be high, even if there are exon changes.

4.2.10 Transcript biotypes

To better understand how alternative splicing operates through switch events, the biotypes of the transcripts were analysed (Figure 4.21, Figure 4.22). Biotypes are indicators of biological significance and in this case, the transcripts were classified into four categories:

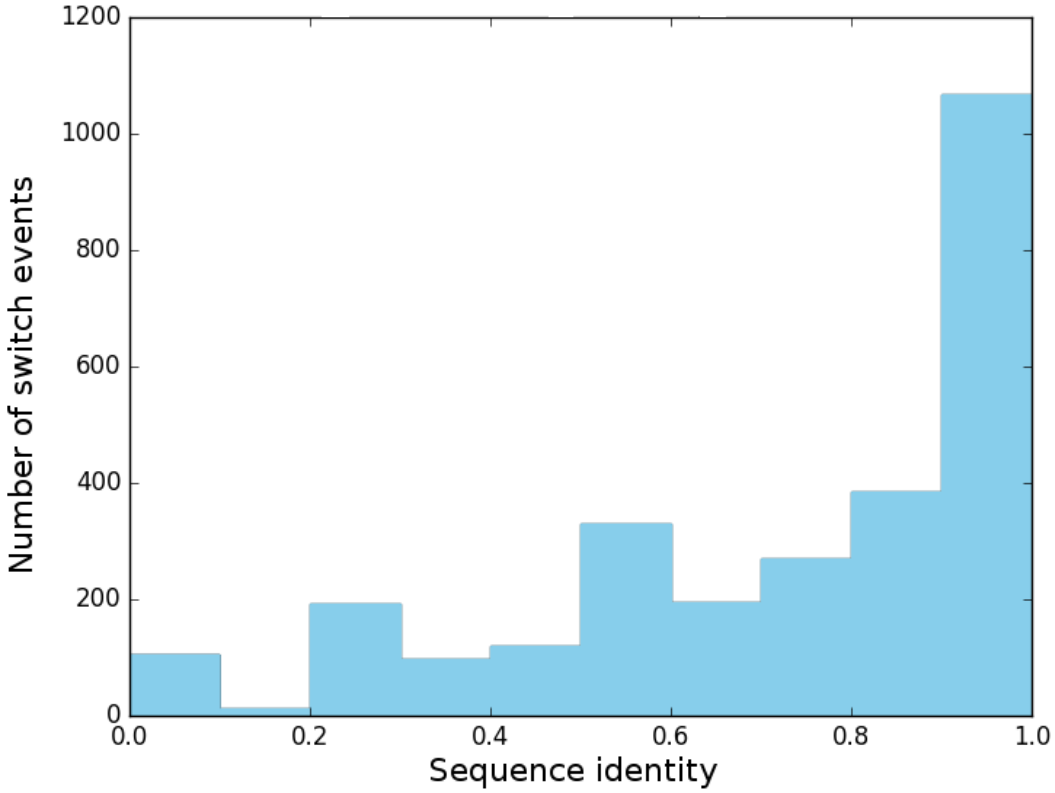


Figure 4.17 Distribution of the DNA sequence identity between pairs of switch transcripts calculated with BLASTN [18] for 5-fold switch events.

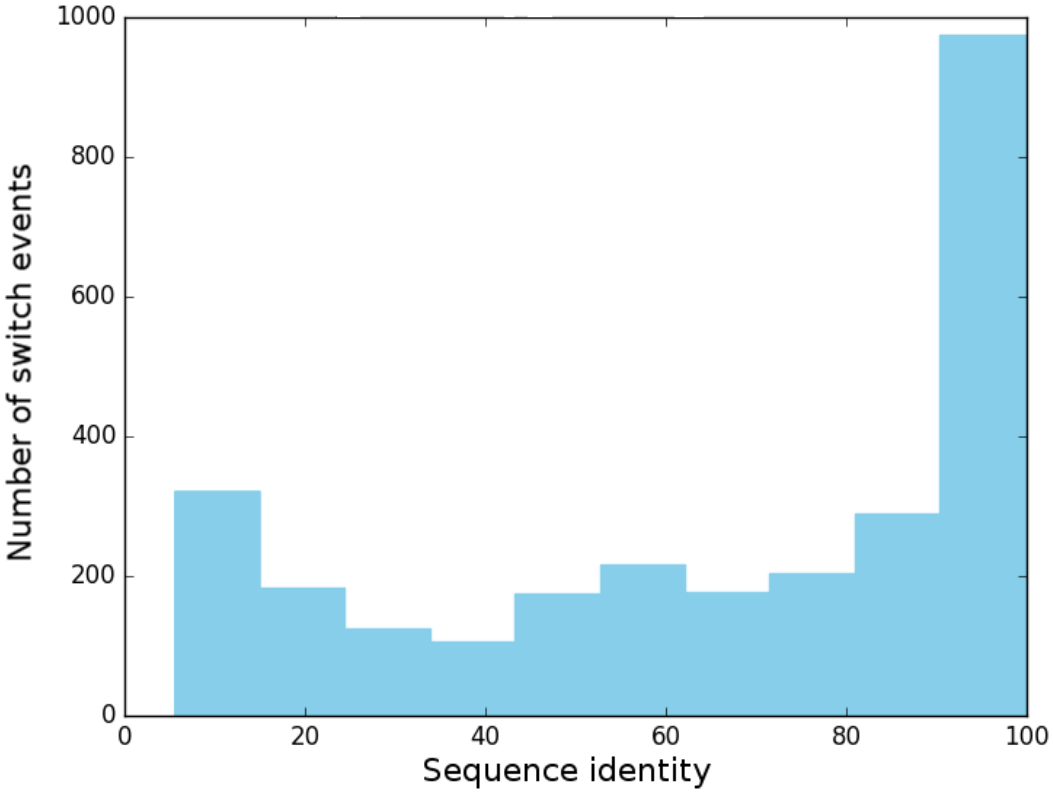


Figure 4.18 Distribution of the DNA sequence identity between pairs of switch transcripts calculated with a global aligner (needle) for 5-fold switch events.

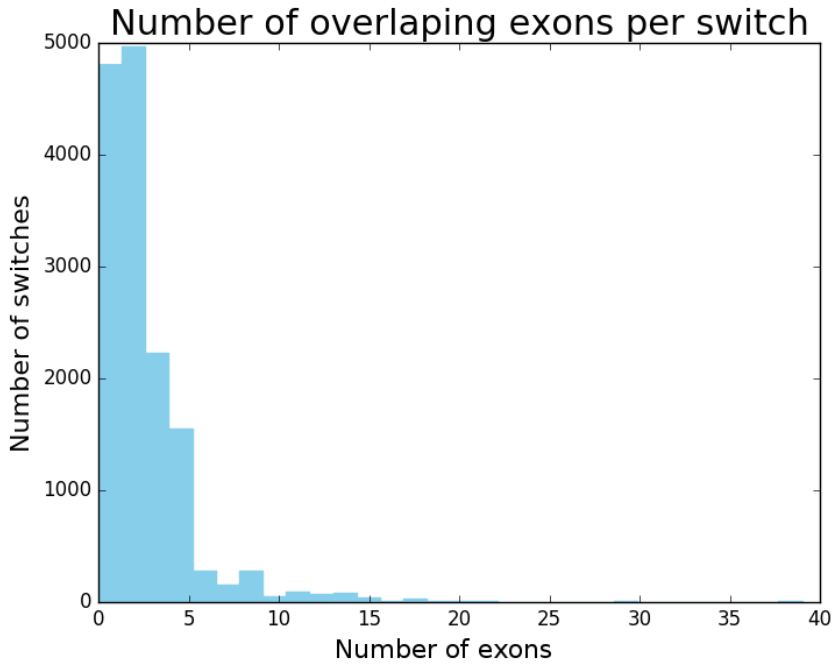


Figure 4.19 Distribution of the number of exons that overlap per switch event. These data refers to 5-fold switch events only.

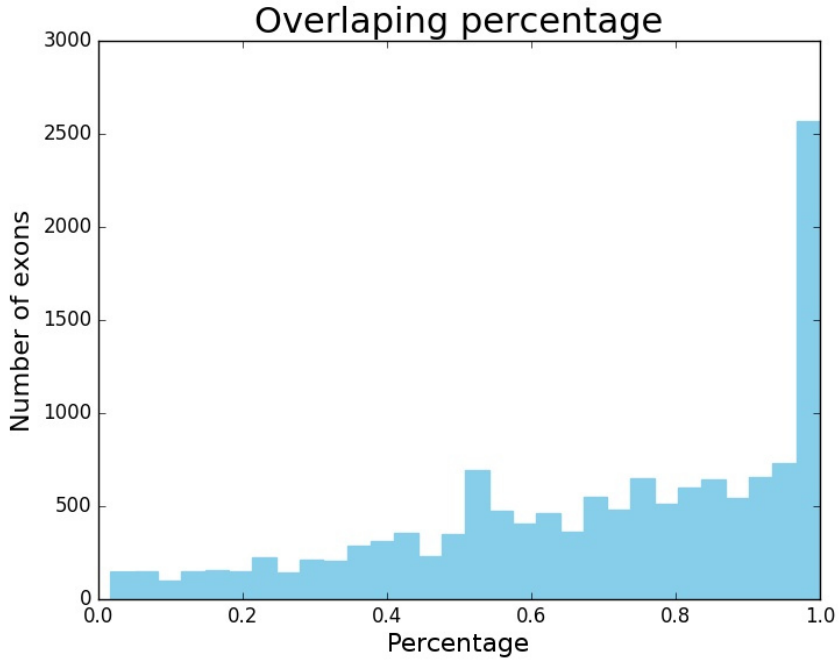


Figure 4.20 Distribution of the overlap percentage between exons in the annotation.

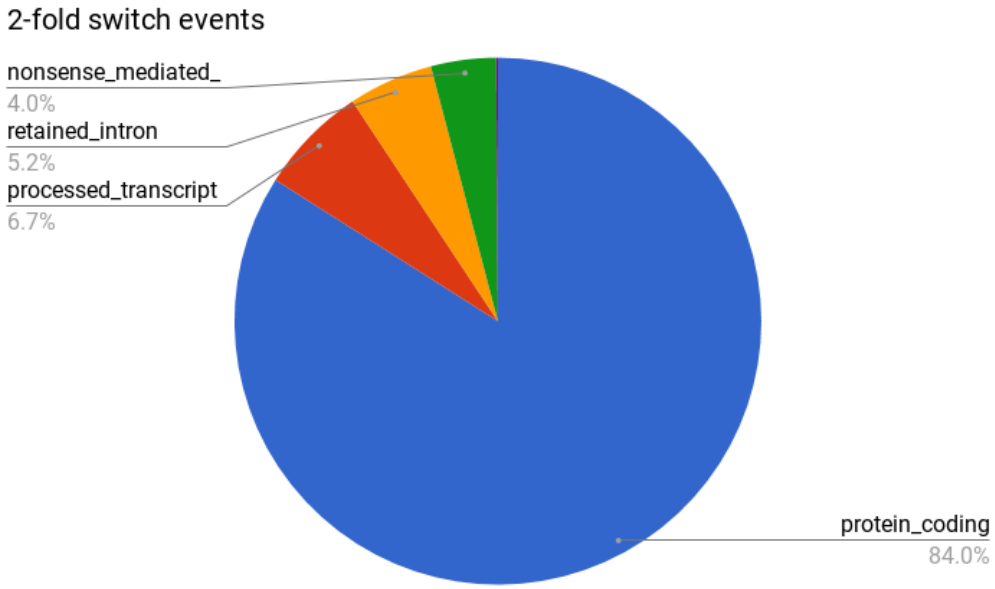


Figure 4.21 Percentage of transcript biotypes in 2-fold switch events. List of biotypes: protein-coding; processed transcripts; retained intron; and nonsense-mediated decay.

protein-coding; processed transcripts; retained intron; and nonsense-mediated decay. Both 2- and 5-fold switches have almost identical percentages of the transcript biotypes. Around 84% of the transcripts in the switch events are protein-coding transcripts and the others are distributed among 3 categories of non-coding transcripts, processed transcripts being the most common ones, followed by retained intron transcripts and finally nonsense-mediated decay transcripts. This indicates that the majority of switches has at least the potential to have an effect at the protein level. It has also to be noticed that 70% of the 5-fold switches occur between two protein-coding transcripts, so in these cases, alternative splicing can potentially change the protein being expressed, independently of changing the function or not. In 27% of the cases, there is a change between coding and non-coding transcripts, indicating that alternative splicing might be used to turn off genes or regulate gene expression. The last 3% of the switches occurs between two non-coding transcripts, these cases might be related to the regulation of other genes, but these must be judged on a case by case basis and further investigation is needed to know what is exactly happening.

4.2.11 Protein domain analysis

The exon changes mostly result in small changes in the transcript sequence but that does not give information about affecting the function of the isoform or not. In the case of protein-

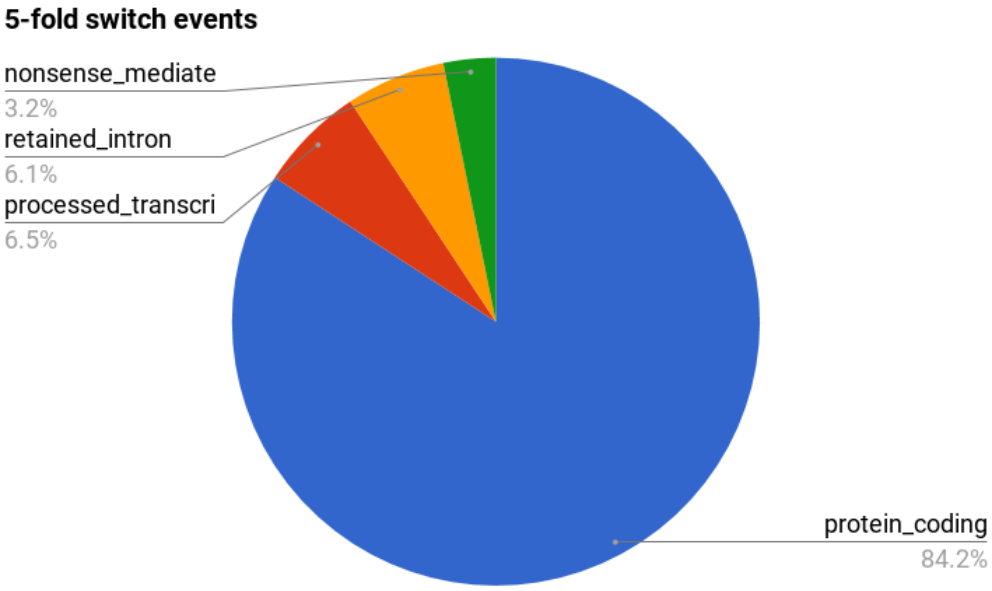


Figure 4.22 Percentage of transcript biotypes in 5-fold switch events. List of biotypes: protein-coding; processed transcripts; retained intron; and nonsense-mediated decay.

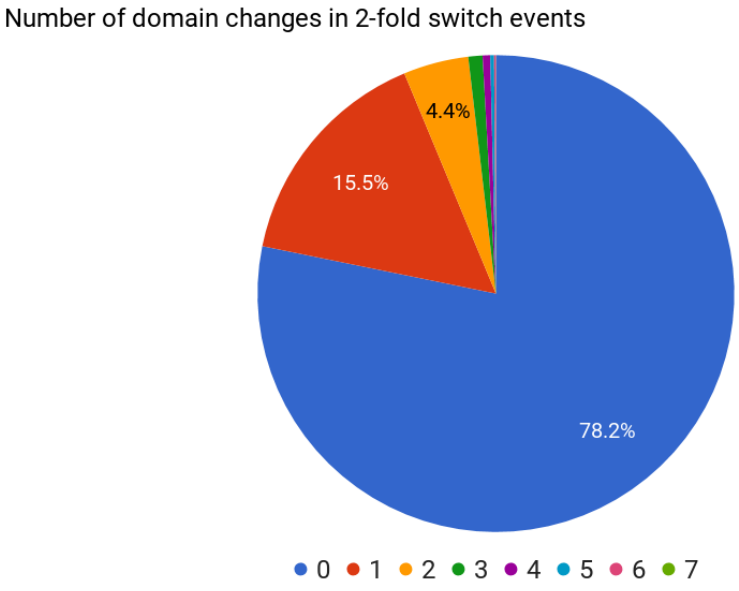


Figure 4.23 Percentage of 2-fold switch events with domain changes. Each color represents a different number of domain changes.

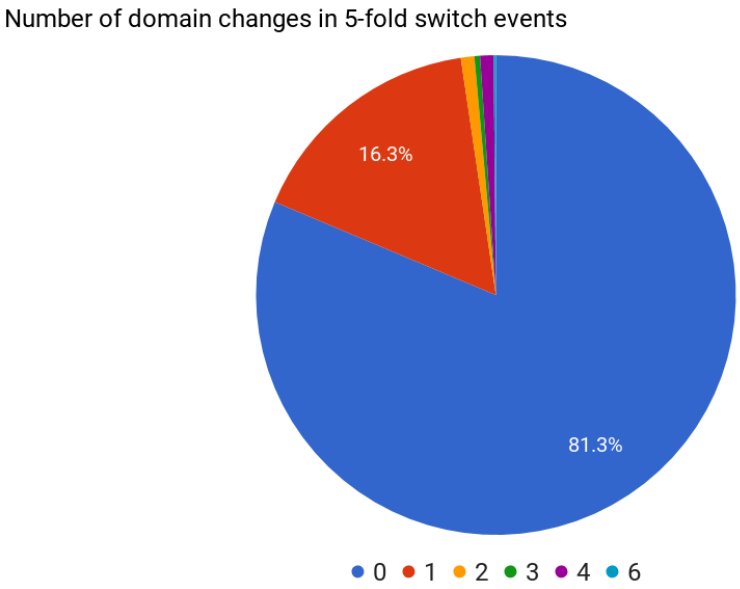


Figure 4.24 Percentage of 5-fold switch events with domain changes. Each color represents a different number of domain changes

coding transcripts, it is relevant to know if there are changes in the function of the expressed protein. To determine how switch events affect protein function, the Pfam domains [140] annotated for each transcript in a switch event were compared. Pfam is a database with a large collection of protein families and each is represented by multiple sequence alignments and a profile Hidden Markov model (HMM). Protein domains are functional regions that can be combined in a protein to confer its specific activity.

It was observed that in 80% of both the 2- and 5-fold switches, there were no protein domain changes (Figure 4.24, Figure 4.21), indicating that in these cases the proteins are not likely to change their function. Still, there are domain changes in 20% of the cases, indicating potential changes in protein function. These results are similar to what was observed in this proteomics study [56] where the authors concluded that 84.4% of splice events leave Pfam domains untouched.

These results do not imply that these switches have no biological function. It has been previously reported that some splicing events influence DNA or protein binding in the transcription factors complex and this regulation is typically cell-type or tissue-specific [146]. There are also cases where splicing events control protein localization in the cell by modifying localization signals, sequences for post-translational modification or interaction sites with other proteins, enabling tissue-specific protein interaction networks [147, 148]. Alternative splicing also controls mechanisms that regulate enzyme activity, protein secretion,

and control of substrate binding. Finally, some splicing events might occur in non-coding regions of mRNA, affecting miRNA binding sites and therefore the regulation of mRNA [63].

4.3 Case studies

Addressing the role of alternative splicing can be challenging, therefore it is important to find switch events where there is confidence they are real and not artifacts produced by the methodology or dependency on the annotation. Finding evidence for particular cases of these switch events from other methods provides extra support to the obtained results.

As mentioned before, Michael Tress' group works on alternative splicing, specifically trying to address and identify dominant protein isoforms. Here, five genes selected by Michael as potentially having multiple principal protein isoforms are analysed. Additionally, the isoforms of each gene were identified as having domain swaps, which indicates that not only the isoforms change but also there is a change in their function.

I looked in detail to see if there was a change in the relative isoform expression levels between different conditions and checked if there were any switch event cases among these specific genes. An analysis of the information generated for these genes through the switch events protocol is presented here.

The following analysis was done using the quantification scores generated by both Cufflinks and Kallisto.

4.3.1 MOCS2 - ENSG00000164172

This is the molybdopterin synthase catalytic subunit gene (molybdenum cofactor synthesis 2). It has a total of 10 annotated transcript isoforms, 8 of which are protein-coding (Figure 4.25). The expression analysis revealed that this gene is expressed in all 32 tissues of the dataset and both Cufflinks and Kallisto showed identical results. There are two isoforms (ENST00000396954 and ENST00000450852) in the list of protein domain swaps (Table 4.4) but only one of the isoforms (ENST00000450852) is expressed in all tissues of the dataset. This isoform is also the dominant isoform in all tissues and in most of them is also the only isoform of this gene that is expressed, with all the other annotated transcripts being either not expressed or expressed at a very low level. Therefore, there were no switch events between transcripts of this gene. An example of the expression profile can be seen in Figure 4.31a, all other tissues displayed similar transcript expression profiles.

Gene name	Gene id	Transcripts id
CUX1	ENSG00000257923	ENST00000292535 ENST00000360264 ENST00000292538
NEBL	ENSG00000078114	ENST00000377122 ENST00000417816
DST	ENSG00000151914	ENST00000361203 ENST00000244364 ENST00000370765
MOCS2	ENSG00000164172	ENST00000396954 ENST00000450852
ZNF451	ENSG00000112200	ENST00000370706 ENST00000370708

Table 4.4 List of genes and respective transcripts found to have evidence of domain swaps at the protein level.

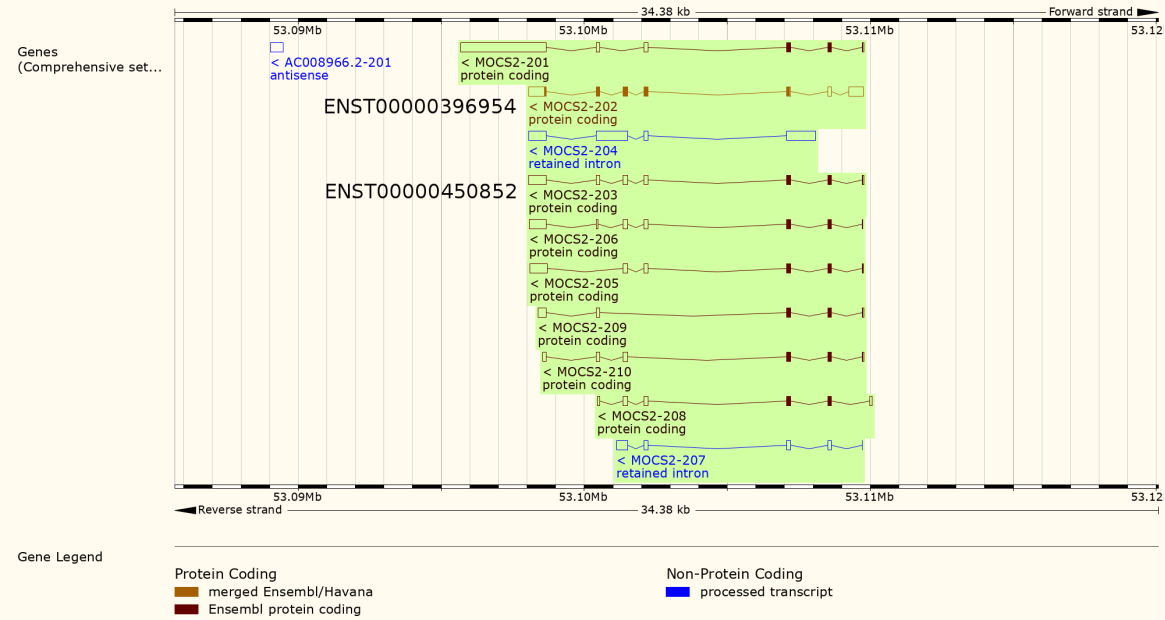


Figure 4.25 View of the transcript isoforms of MOCS2 gene in Ensembl genome browser [19]. The isoforms with domain swaps are identified with their ensembl IDs (ENST00000396954 and ENST00000450852).

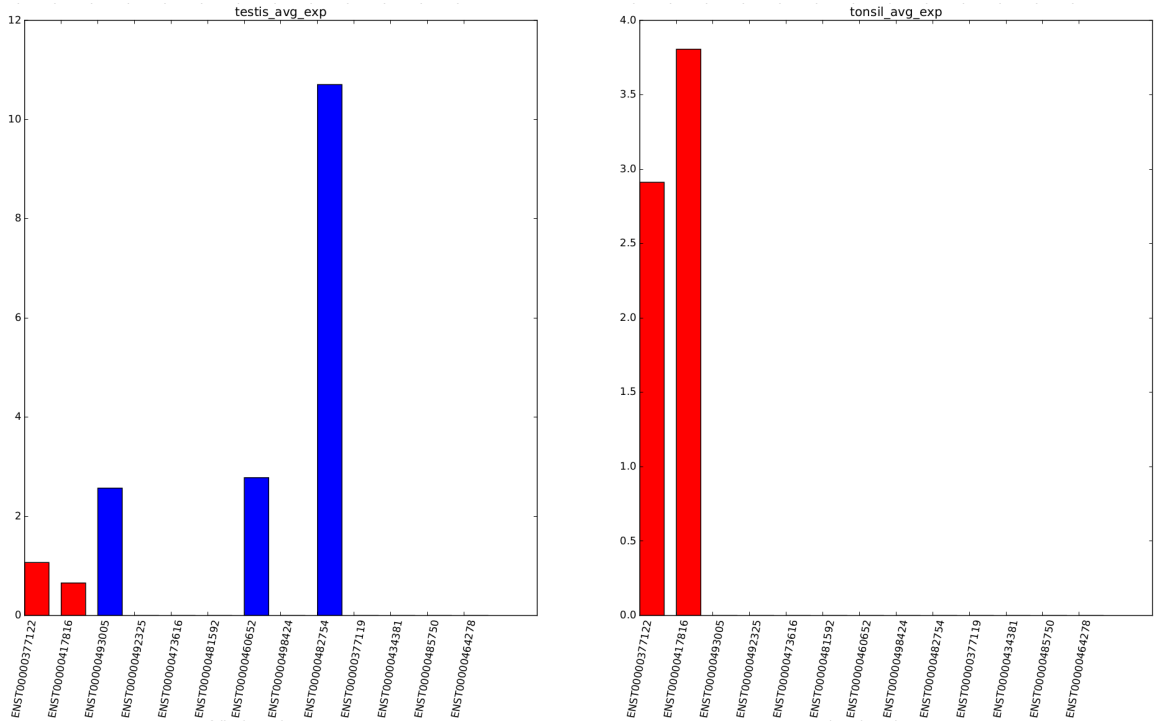


Figure 4.26 Expression profiles for the NEBL gene: testis (left) and tonsil (right). The expression values were determined with Cufflinks. The columns in red correspond to the transcripts in the list of interest, and the others are displayed in blue.

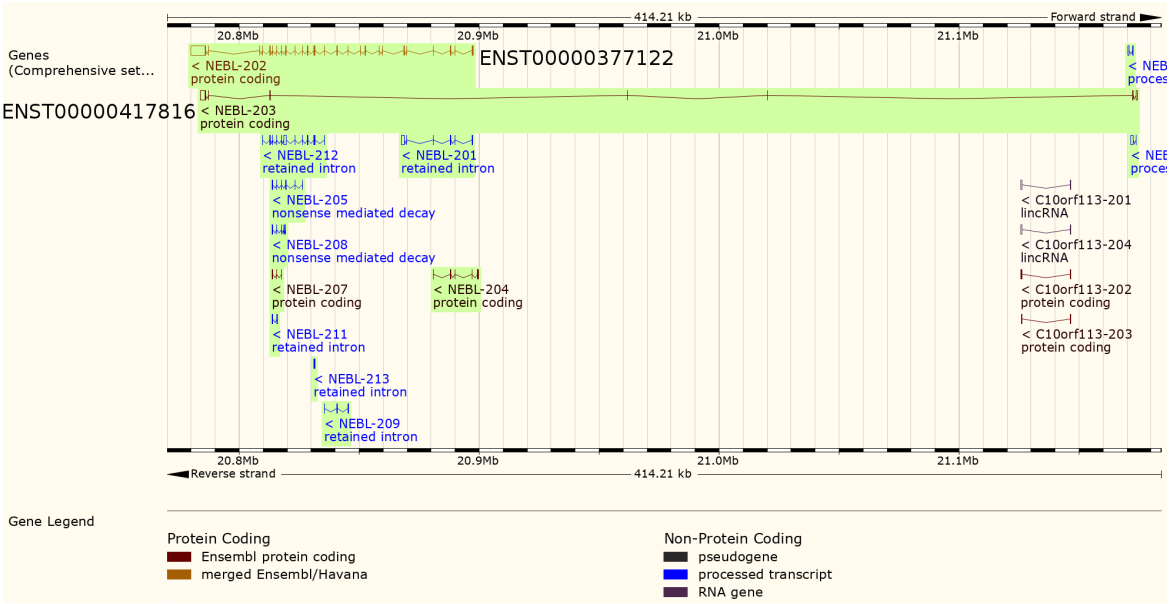


Figure 4.27 View of the transcript isoforms of NEBL gene in Ensembl genome browser [19]. The isoforms with domain swaps are identified with their ensembl IDs (ENST00000377122 and ENST00000417816).

4.3.2 NEBL - ENSG00000078114

This is the nebulette gene, a cardiac-specific protein belonging to the nebulin family of proteins. It has a total of 13 annotated transcript isoforms, 4 of which are protein-coding transcripts (Figure 4.27). Both Cufflinks and Kallisto results show that this gene is not expressed in 11 of the 32 tissues in the dataset: endometrium, fallopian tube, skeletal muscle, spleen, liver, lymph node, adipose tissue, small intestine, appendix, duodenum, and bone marrow. On all the other tissues, with the exception of testis, both isoforms on the list (ENST00000377122 and ENST00000417816; Table 4.4) are expressed approximately at the same level. This means there are no dominant transcripts and consequently no switch events.

The transcripts expressed in testis are not protein-coding. In this tissue, the dominant transcript is ENST00000482754 (*NEBL-010* in Figure 4.30) and there are two other isoforms being expressed, ENST00000493005 and ENST00000460652. Testis and tonsil transcript expression profiles can be seen in Figure 4.26. All tissues where the gene is expressed have transcript expression profiles similar to tonsil.

4.3.3 ZNF451 - ENSG00000112200

This is the zinc finger protein 451 gene. It has a total of 18 annotated transcript isoforms, 9 of which are protein-coding. This gene is expressed in all tissues of the dataset and the

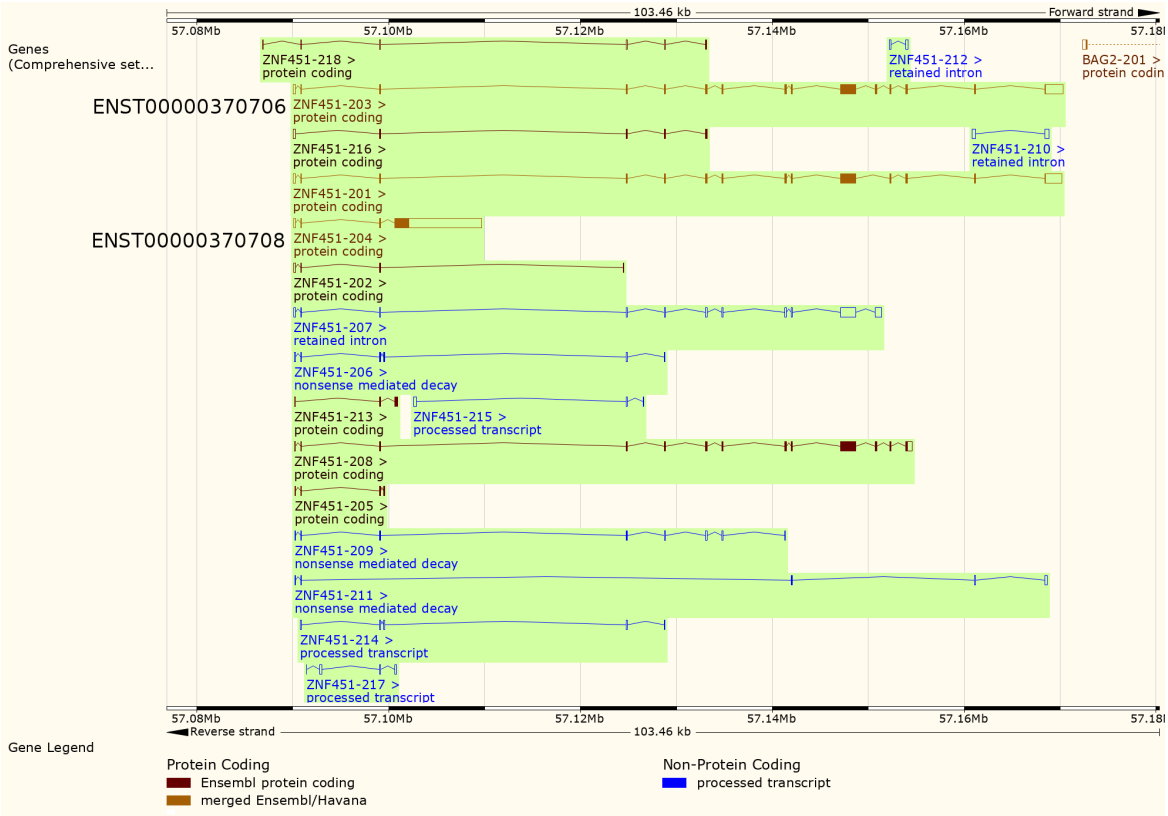


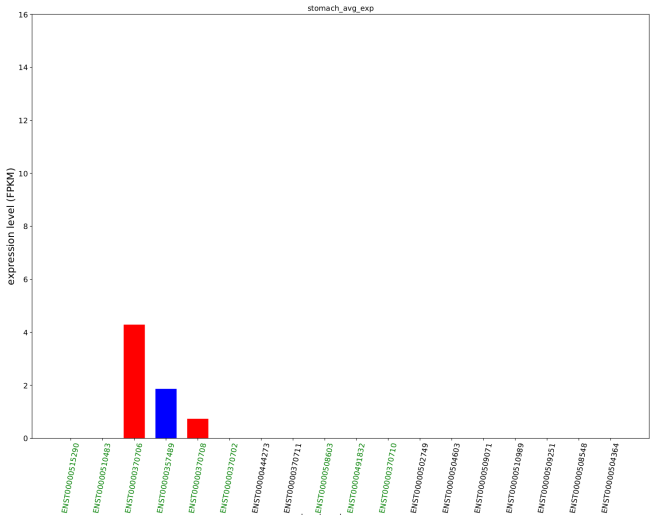
Figure 4.28 View of the transcript isoforms of ZNF451 gene in Ensembl genome browser [19]. The isoforms with domain swaps are identified with their ensembl IDs (ENST00000370706 and ENST00000370708).

results obtained with both Cufflinks and Kallisto were similar. In most tissues, the dominant isoform is one of the transcripts in the list (Table 4.4): ENST00000370706. Bone marrow and testis are two exceptions. In bone marrow, the dominant transcript is ENST00000370708, which is also one of the transcripts in the list. This means there are switch events between these two transcripts, both RNA-seq data and domain swaps information are in agreement (Figure 4.29a and Figure 4.29c).

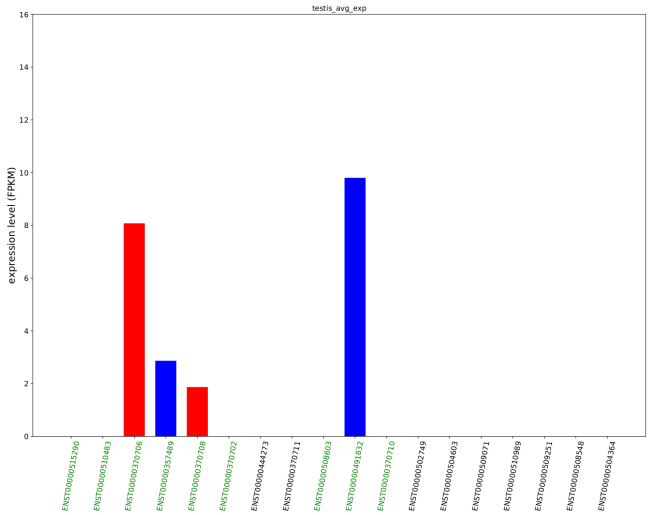
As in the previous case, testis has an expression profile different from all other tissues and its most expressed isoform is ENST00000491832 (*ZNF451-008* in Figure 4.28; Figure 4.29b). This is a protein-coding transcript but is not one of the transcripts in the list of isoforms with protein domain swaps.

4.3.4 CUX1 - ENSG00000257923

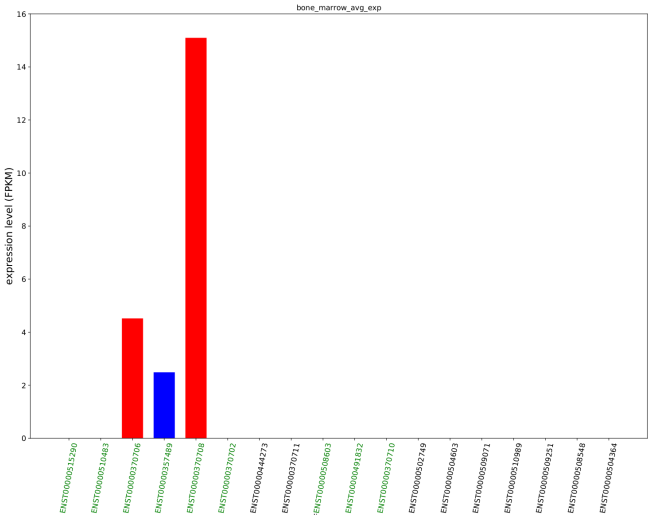
This is the cut like homeobox 1 gene that codes for golgi integral membrane protein 6. It has a total of 21 annotated transcript isoforms, 12 of which are protein-coding transcripts. Cufflinks



(a) ZNF451; stomach; Cufflinks.



(b) ZNF451; testis; Cufflinks.



(c) ZNF451; bone marrow; Cufflinks.

Figure 4.29 Expression profiles for the ZNF451 gene. The gene, tissue and software used for quantification are indicated in each label. The identifiers of protein-coding transcripts are displayed in green and the identifiers of non-coding transcripts are in black. The columns in red correspond to the transcripts in the list of interest, and the others are displayed in blue.

and Kallisto show considerably different expression profiles for this gene, an example can be seen in Figure 4.31b and Figure 4.31c. Kallisto shows that one of the transcripts in the list (ENST00000292538) is the major transcript in all tissues, with exception of testis, being dominant in almost all. Cufflinks shows that, besides the ENST00000292538 isoform, ENST00000558469 (*CUX1-021* in Figure 4.30) is also highly expressed in some tissues and, not only is this transcript dominant in some tissues, but also the overall expression of the gene is dispersed among more isoforms. Since there is no agreement between Cufflinks and Kallisto results, there is not as much confidence to make any observations regarding transcript dominance and switch events for this gene.

4.3.5 DST - ENSG00000151914

This is the dystonin gene, that codes for bullous pemphigoid antigen 1. It has a total of 35 annotated transcript isoforms, 17 of which are protein-coding. ENST00000370765 is the only transcript in the list (Table 4.4) that is expressed (>1 fpkm) and it is a dominant transcript in skin, tonsil, and esophagus (Figure 4.32a). There are two other protein-coding transcripts that are dominant in other tissues (ENST00000340834 and ENST00000523292, corresponding to Figure 4.32b and Figure 4.32c, respectively). ENST00000523292 (*DST-017* in Figure 4.33) is dominant in adrenal gland, duodenum, lung, prostate, small intestine, and spleen. ENST00000340834 (*DST-018* in Figure 4.33) is dominant in ovary. There are switch events between the dominant transcripts mentioned, however, there is no switch between the two isoforms in the list of isoforms with protein domain swaps.

4.4 Discussion

In this study, RNA-seq data from normal human tissues was used to assess the impact of switch events on the transcriptome and on protein function. It was shown that in a particular condition, there are around 10,000 genes being expressed and most of them have a dominant isoform in most tissues. More specifically, on average, around two-thirds of the genes being expressed have a 2-fold dominant transcript and slightly more than one third have a 5-fold dominant transcript. These were confirmed using both Cufflinks and Kallisto quantification scores, even if these programs rely on quite different strategies for transcript quantification. The fact that the intersection between sets of dominant transcripts determined using Cufflinks and Kallisto data was on average 89% and 82%, for 2- and 5-fold dominant transcripts respectively, shows that for the majority of cases there is agreement between the two methods and a strong signal for transcript dominance. Nevertheless, it should be

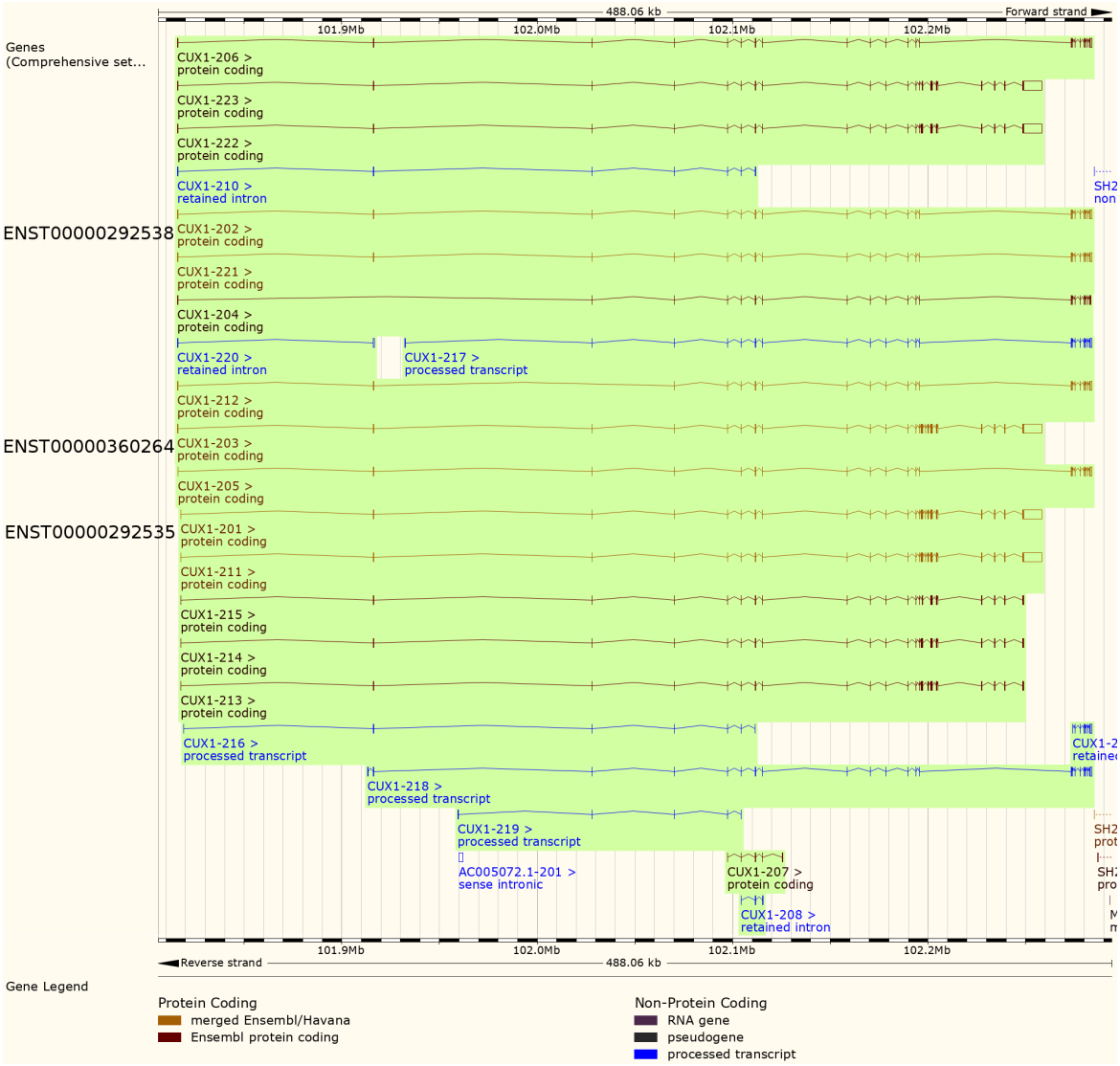
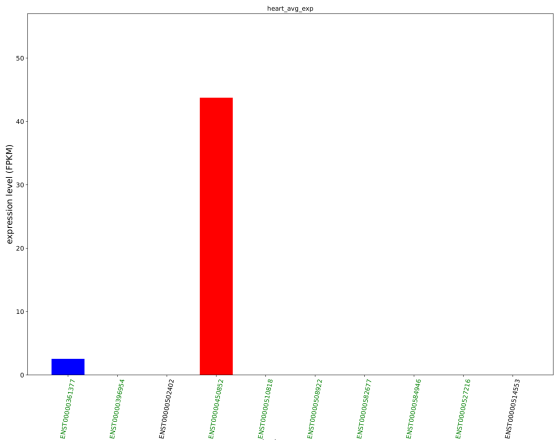
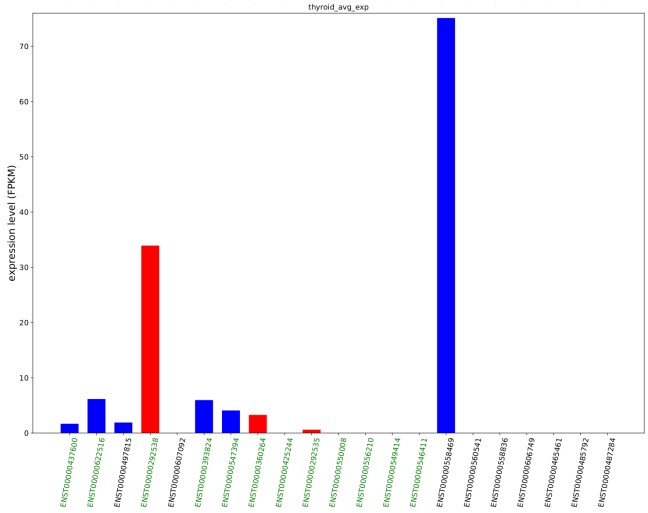


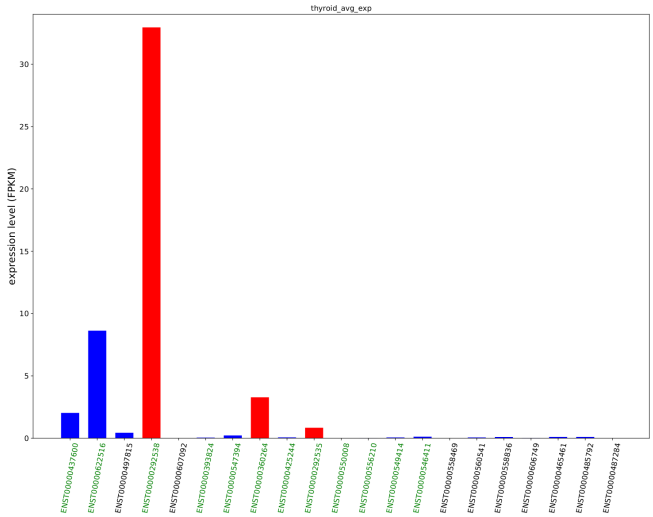
Figure 4.30 View of the transcript isoforms of CUX1 gene in Ensembl genome browser [19]. The isoforms with domain swaps are identified with their ensembl IDs (ENST00000292538, ENST00000360264 and ENST00000292535).



(a) MOCS2; heart; Cufflinks.

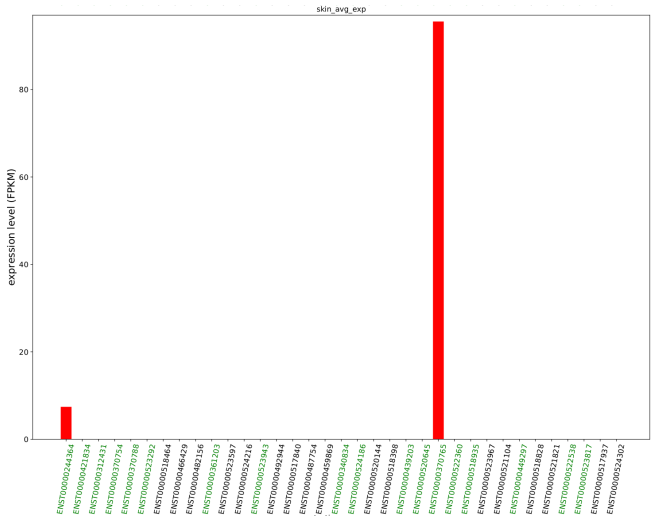


(b) CUX1; thyroid; Cufflinks.

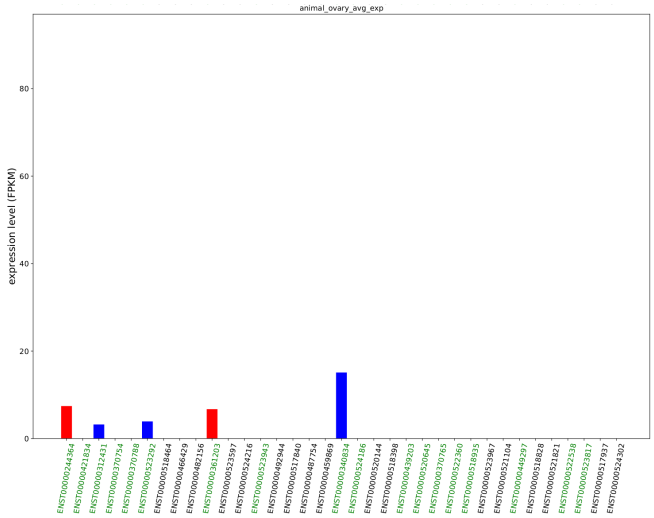


(c) CUX1; thyroid; Kallisto.

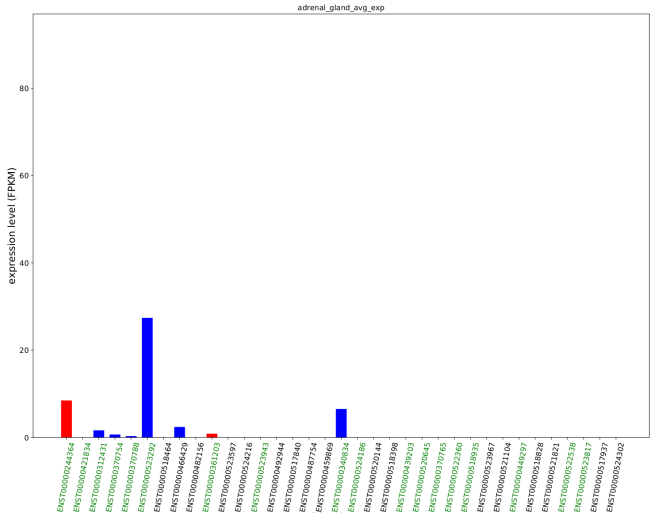
Figure 4.31 Expression profiles for the CUX1 and DST gene. The gene, tissue and software used for quantification are indicated in each label. The identifiers of protein-coding transcripts are displayed in green and the identifiers of non-coding transcripts are in black. The columns in red correspond to the transcripts in the list of interest, and the others are displayed in blue.



(a) DST; skin; Cufflinks.



(b) DST; ovary; Cufflinks.



(c) DST; adrenal gland; Cufflinks.

Figure 4.32 Expression profiles for the genes of interest. The gene, tissue and software used for quantification are indicated in each label. The identifiers of protein-coding transcripts are displayed in green (if different colors are used) and the identifiers of non-coding transcripts are in black. The columns in red correspond to the transcripts in the list of interest, and the others are displayed in blue.

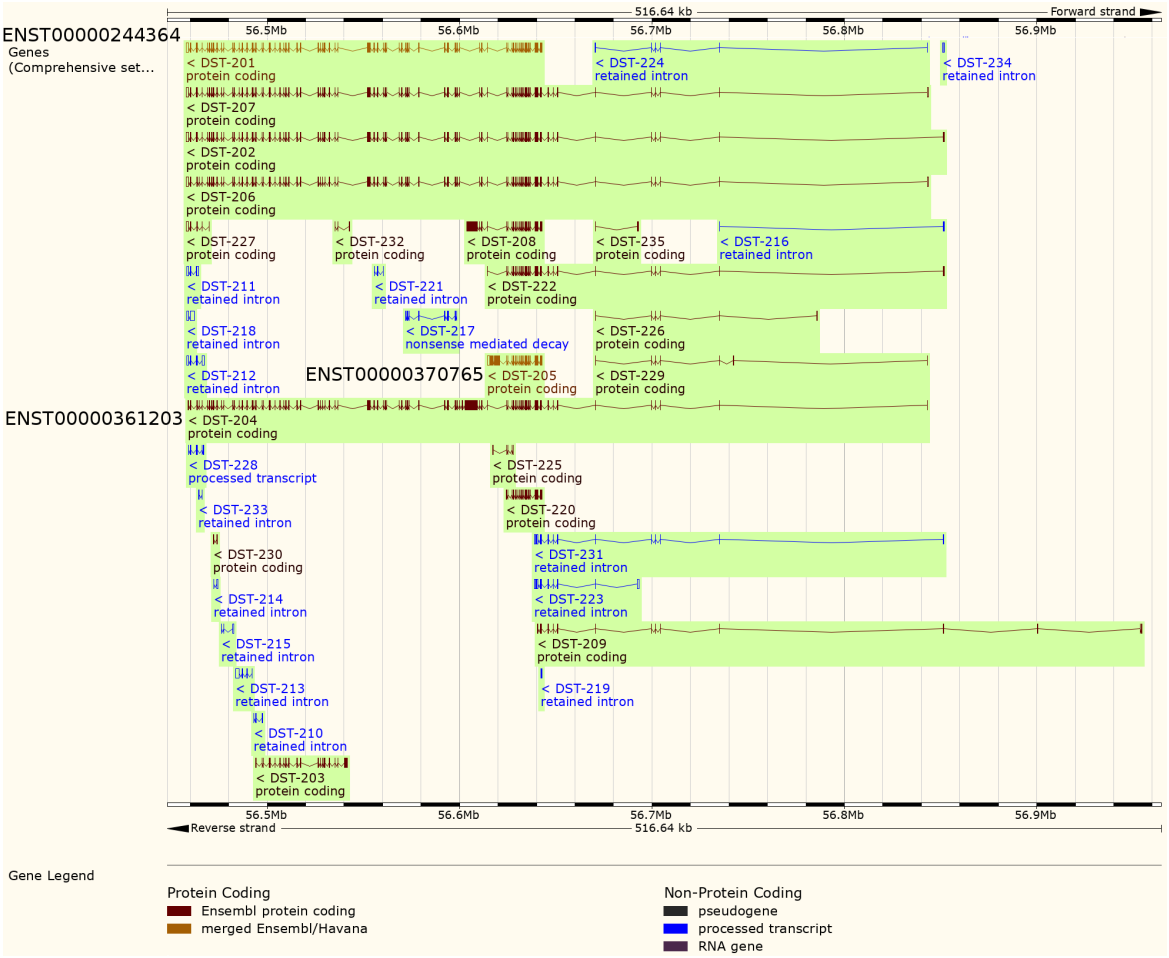


Figure 4.33 View of the transcript isoforms of DST gene in Ensembl genome browser [19]. The isoforms with domain swaps are identified with their ensembl IDs (ENST00000361203, ENST00000244364 and ENST00000370765).

mentioned that transcript quantification from small reads is still not a straightforward task and the dependence on gene and transcript annotation comes with both advantages and compromises. Some of the possible problems with the annotation were partially exposed by the comparison of exons annotation where it was evident that, relying only on exon identifiers can be misleading, but also it was unclear why some exons with identical genome coordinates were annotated as distinct exons.

The transcript dominance analysis is in agreement with a previous study done in our group [1], where it was found that for normal human tissues, 79% of protein-coding genes have a 2-fold dominant transcript and 56% have a 5-fold dominant transcript. Still, the current analysis resulted in slightly lower estimations of dominant transcripts due to the use of more conservative criteria.

Although there is a large number of transcripts annotated, only a modest part of them (16,811 – *exp genes* in Table 4.2) was found to be expressed in normal human tissues, possibly indicating that most isoform variants are unlikely to have a functional role in the cell, as it has been suggested based on proteomics studies [4]. Nevertheless, it is important to consider that the isoforms that were not found to be expressed, might only be expressed in other cell types or under specific conditions. For instance, alternative splicing plays a role in specific cellular processes such as apoptosis, where it can act as an on/off switch for several genes that code for pro-apoptotic and anti-apoptotic isoforms [63]. This might be one of the reasons why certain transcript isoforms are not detected in the dataset used in this study.

APPRIS principal isoforms have been shown to be good predictors of dominant protein isoforms, overlapping 97.6% of comparable genes [141]. The overlap between the determined dominant transcripts and APPRIS principal isoforms was relatively high, indicating that there is a relationship between the relative expression of transcripts of a gene and their cross-species conservation. The results obtained are similar to the ones observed in proteomics studies that show there is a tendency for the most functional and structural conserved isoforms to be expressed [141]. It should also be noticed that the overlap was higher for 5-fold than 2-fold dominant transcripts (83% vs. 71%), suggesting that the higher the dominance, the more likely is for the transcript to be conserved.

The analysis of switch events revealed that when a gene has a dominant transcript, this tends to be the same across tissues. However, there are exceptions and some dominant isoforms switch across conditions. On average there are 30.2 2-fold switch events between a pair of tissues and 3.7 5-fold switch events. This relatively low number of switch events is in accordance with what has been suggested before by the GTEx consortium, that gene expression levels might be the drivers of tissue specificity, and that 84% of the variance observed between human tissues is due to gene expression [60]. So this can be part of

the explanation why the number of switch events is small. Also, across the 32 tissues in the dataset, there are only 1968 genes involved in switch events, which is less than 9% of protein-coding genes.

At the protein level, there is also evidence for very few splicing events. Of 64% of annotated human protein-coding genes, there is evidence of splice events for only 282 of them, suggesting that at the protein level there is a main isoform for most genes [56]. In contrast with the dataset used in this chapter's study, the one used in the mentioned proteomics study covered a wide range of cell types, tissues, and development stages. However, it should be mentioned that MS experiments have relatively low coverage of the proteome and low sensitivity, being difficult to detect very lowly expressed peptides.

Regarding the criteria that define a dominant transcript, when these criteria were relaxed, the number of switch events considerably increased for certain pairs of tissues, but if we consider the total of annotated transcripts and genes with dominant isoforms, the number of switches was still low. On one hand, this reinforces the idea that the dominant transcripts are conserved across tissues. On the other hand, it shows that there is variability between samples of a tissue, in this case, between individuals because each sample of this dataset belongs to a different individual.

The most used methods for protein function annotation typically use sequence or structure homology strategies, even if there are other strategies that explore different types of data [149]. Although these are the most commonly used criteria, it has been observed that in some cases proteins might share high sequence or structural similarity but still evolve distinct functions or sub-functions. In these cases, similar proteins perform different functions, while sharing a general functional feature, which is quite common in enzymes that have a common step between their reaction pathways. These groups of homologous proteins that share the same function are typically designated functional families, which can be used to annotate uncharacterized sequences [150]. The higher the sequence divergence, the higher the likelihood of protein function divergence. However there are exceptions and proteins can diverge to a related function or, through the process of recruitment, proteins with similar sequence can perform very distinct functions. It should also be noticed that below the threshold of 50% sequence similarity, the functional divergence is enhanced [151]. In the case of switch events, it was observed that, in general, transcripts have high sequence similarity, which suggests that their function is not affected in a significant way.

To investigate how alternative splicing is operating in the switch event cases, the exons of the switch transcripts were compared and it was observed that there is a large number of exons that overlap close to 100% (e.g. 20,391 exons have a percentage of overlap higher than 95%; Figure 4.20). Although it was also observed that some exons have the same coordinates,

despite being annotated as different exons with different identifiers (e.g. ENSE00003570356 and ENSE00003479361 exons, from ENST00000452550 and ENST00000380631 transcripts, respectively). This might be related to what is observed on the proteomics side, where more than 20% of the splicing events identified correspond to a substitution of one exon by a homologous one. These homologous substitutions are extremely conserved [56].

The annotation of transcript biotypes contains a subset of all categories of non-coding transcripts and, although intron retention and nonsense-mediated decay are two different categories, in some situations these processes are linked [152]. Nevertheless, the transcript biotypes annotation allowed a simple straightforward analysis that revealed that 84% of both transcripts in switch events are protein-coding. So in these cases, there is the potential for the changes that occur at the transcript level to be reflected at the protein level. The other 16% of the cases might actually be more difficult to interpret because the downstream consequences of some events might differ and intron retention in particular is, in fact, the less well understood type of alternative splicing. Intron retention might lead to the expression of alternative protein products but it most often leads to a wide reduction of the expression level of transcripts that are not physiologically required in a cell or tissue type [153]. It acts through the mechanism of nonsense-mediated decay, through nuclear sequestration and also through fast turnover of intron retention transcripts. Intron retention is associated with a global checkpoint-type mechanism of localized stalling of RNA polymerase II and reduction of the availability of spliceosomal components, that suppresses inappropriately expressed transcripts [154].

Similarly to retained intron cases, processed transcripts may also lead to a diverse range of downstream effects. This category of transcripts includes long non-coding RNA, ncRNA and unclassified processed transcripts that do not fit into any of the previous categories. This is a very heterogeneous set of RNA molecules with a broad range of biological functions. Nevertheless, it can be said that alternative splicing has an impact on the transcriptome which can affect the proteome downstream but can also regulate gene expression levels upstream.

84% of switch transcripts are protein-coding but in 80% of switch events, there are no protein domain changes, indicating that the isoforms that switch are similar and possibly have a similar function. The other 20% of the switch events have domain changes, which suggests that in some cases evolution might have selected certain isoforms to fulfill specific functions in different conditions or there might also be cases of co-regulated exons without any established function [56]. Given these observations, the main role of alternative splicing does not seem to be generating drastic changes and produce high protein diversity, but rather regulate small transcript and protein changes.

The small number of protein domain changes is most likely related to the function of alternative splicing. Some studies suggest that alternative splicing main role might be to regulate tissue-specific protein-protein interaction networks. This is supported by the fact that constitutive exons tend to map to protein domains more than other exons but tissue-specific exons map to protein regions with no well defined three-dimensional structure [147]. Alternative splicing also plays a significant role in the localization of proteins and it can act by altering localization signals, post-translational modification motifs or by altering protein interaction sites [63]. In cases like these, there might not be needed the insertion or exclusion of new domains.

In some cases, alternative splicing affects untranslated regions of mRNA, thus protein domains are not affected. However, it does not mean that there are no effects at the protein level. In these cases, alternative splicing can regulate the stability of mRNA, through affecting the miRNAs binding sites frequently located in UTRs [63].

Similarly to the study presented in this chapter, most splice events identified at the protein level correspond to small changes, resulting in a few functional domains disruptions. The majority of the alternative splicing events identified in proteomics do not have a significant effect on the structure or function of the protein isoform. Most identified indels are short or located in unstructured protein regions. As a consequence, these events rarely affect Pfam functional domains. One of the theories for explaining the non-disruption of domains is that isoforms with disrupted domains are more likely to affect cellular processes and their expression must be regulated by cellular quality control pathways [56].

Alternative splicing is also known for regulating enzymatic properties. Certain kinases are frequently inactivated by deletion or inclusion of protein regions in their active center. This enzyme region is extremely sensitive to changes so, even if small, they can considerably affect enzyme activity. In some cases, enzyme activity can even be totally abolished. These particular isoforms are typically generated by creating a premature stop codon, which is the third most common type of alternative splicing event observed in switch events, accounting for 17.6% of the cases. Alternative splicing can also control enzyme secretion and even promote the formation of heterodimers, having a net negative effect on the active isoform activity [63].

As mentioned before, alternative exons often correspond to amino acids located on the protein surface, so they can regulate protein binding. There are cases of exons that regulate complete interaction domains or part of binding domains, in most cases however binding is not abolished is simply modulated, implying again small changes. Besides modulating protein binding, alternative splicing also has a role in binding to DNA or smaller ligands, such as hormones. One of the well-known examples is the case of the insulin receptor, where

alternative splicing modulates its affinity to IGF-II (insulin-like growth factor II). It acts like a fine tuner rather than changing it completely. The effects alternative splicing have on channel proteins are in most cases also small. It is known to minutely control iron channels by regulating every aspect of it: gating times, gating voltages, ion sensitivity and inhibition by other molecules [63].

Alternative splicing can also modulate transcription by affecting protein interaction affinities between transcription factor components. Transcription factors might even lose their ability to bind to promoters. In this way, alternative splicing of transcription factors can control gene expression of downstream genes. Another known mechanism to regulate gene expression is by regulating transcription factors intracellular or intranuclear localization. ncRNAs can also be used to indirectly regulate gene expression. Localization of proteins can be changed by alternative splicing and sometimes these changes can be of the "all or nothing" type, for instance, an exon can code for a nuclear localization site. Most often the changes brought upon by alternative splicing just gradually change the relative localization of protein isoforms across cellular compartments, where they can acquire different functions [63].

Because alternative splicing has such a wide spectrum of activities, it can make it hard to pinpoint or detect its effects just by looking at mRNA expression values and isoform switches.

Although there are parallels between the cases of switch events and proteomics studies, it must be clarified that at the transcript level one-third of protein-coding genes has no dominant transcript at the tissue level, meaning that they are expressing more than one isoform. This is not what is observed at the protein level, where there is generally no evidence for more than one isoform being expressed [56]. This suggests that, although a third of the genes express multiple transcripts, they might not all be translated or the proteins might have a short half-life, preventing its detection by MS. It is also possible that the expressed RNA might have other functions or the expression levels are too low to be detected using current MS techniques.

With the analysis made in this chapter, it is difficult to predict what are the real effects of switch events at the protein level. The comparison of Pfam domains gives an indication of the possible consequences of an isoform switch but just based on RNA-seq data, it is not even certain that the protein is expressed. Also, it was observed in the case studies that even if there is evidence for protein domain swaps between isoforms, these are not necessarily expressed. In fact, there were switch events between protein domain swaps isoforms in only one of the five genes analysed (ZNF451). It is also relevant that in the case of the NEBL gene, two of the isoforms are consistently being expressed in a large number of tissues, which

shows that alternative splicing enables the simultaneous expression of multiple isoforms with different functions. Finally, there is a relatively modest correlation between transcript and protein expression levels, around 56%-64% [155–157], so even if there are protein domain changes, there is not enough information on RNA-seq data to understand the full impact of such changes. Ideally, a dataset of paired transcriptomics and proteomics data with high isoform coverage will give answers to more questions and allow to better understand the role of alternative splicing in switch event cases.

4.5 Methods

4.5.1 Dataset

The data was generated using Illumina Hiseq2000 and Hiseq2500, using the standard Illumina RNA-seq protocol with a read length of 2 × 100 bases [138].

4.5.2 Gene and transcript quantification

The raw reads were filtered before mapping by trimming the last five nucleotides. The pre-processed reads were mapped, using TopHat2 [98], to gene and transcript annotations from GENCODE genome assembly version GRCh38 and Ensembl release version 79 of the human reference genome [19]. The GENCODE annotation is the result of merging the Havana manual gene annotation with Ensembl automated gene annotation.

The study presented in this chapter was based on protein-coding genes only. These are genes that contain at least one protein-coding transcript. The annotation for these genes and respective transcripts was selected and retrieved using Biomart web-based tool [158].

The reads pre-aligned with TopHat2 were used as input to Cufflinks. The quantification scores for all genes and transcripts were calculated using Cufflinks [14] and Kallisto [15]. The expression value threshold to consider a transcript as expressed was 1 FPKM – fragments per kilobase of exon model per million of mapped reads. This is a commonly used threshold and it has been suggested that 1 FPKM is the minimum expression necessary for protein detection [159], which is relevant because of the comparisons made between dominant and APPRIS principal transcripts.

4.5.3 APPRIS isoform annotation

APPRIS selects ‘PRINCIPAL’ or ‘ALTERNATIVE’ CDS variants for each gene based on a range of protein features. Principal isoforms are classified from 1 to 5 (1 being the most

reliable) and alternative isoforms are classified from 1 to 2 [160]. In the analysis presented in this chapter, only principal isoforms (1 to 5) were compared to the dominant transcript isoforms.

4.5.4 Differential gene expression

The R package DESeq2 (version 1.8.1) [115] was used to identify differentially expressed genes across all pairs of tissues in the dataset. The input of DESeq2 is a matrix of gene counts, therefore the python package HTSeq [102] was used to obtain those counts. The switch events involving differentially expressed genes were filtered out from the analysis.

4.5.5 Determining differences in exons between transcripts

The exon identifiers were retrieved from the above-mentioned annotation. The exon identifiers of the transcripts of a switch event were compared and the differences were determined by comparing the two sets of exons. Each exon that exclusively belonged to one transcript counted as one difference. Given two sets of exons A and B of two different transcripts, the number of differences is defined by:

$$|A \cup B| - |A \cap B| \tag{4.1}$$

4.5.6 Determining isoform sequence identity

The DNA sequence identity of switch event isoforms was calculated using BLASTN [18] and biopython [161] was used to parse the output files and directly obtain the sequence identity value. The BLASTN parameters used were the default ones: E value = 10; word size = 11; gap opening penalty = 5; gap extension penalty = 5.

The Needleman-Wunsch algorithm [145], a global alignment algorithm, was also used with the default parameters: match award = 10; mismatch penalty = -5; gap penalty = -5.

4.5.7 Percentage of overlap between exons

The percentage of overlap between two exons was determined by comparing the exonic coordinates of each exon of the pair of transcripts in a switch event. The coordinates were retrieved from the annotation and the percentage of overlap was calculated by dividing the number of overlapping nucleotides by the length of the largest exon. In this way, the maximum nucleotides overlap (corresponding to 100%) would be the length of the largest exon.

4.5.8 Determining alternative splicing types

The types of alternative splicing were determined by comparing the exon coordinates retrieved from the annotation. The two transcripts in a switch event were compared exon by exon and the ones that differed were classified according to the type of alternative splicing found (Figure 4.16, [9]). Two additional categories were added to describe cases that did not fit any of the most common alternative splicing types, these were:

Overlap Cases of overlapping exons that do not fit any of the other splicing category.

Exclusive Cases of exons that are exclusive to one of the transcript and do not fit any other splicing category.

4.5.9 Biotype definitions

The transcript biotypes used in this analysis were obtained from Ensembl using Biomart web-based tool [158]. There are four categories of transcript biotypes:

Protein-coding Contains an open reading frame (ORF).

Retained intron Has an alternatively spliced transcript believed to contain intronic sequence relative to other, coding, variants.

Processed transcripts A noncoding transcript that does not contain an open reading frame.

Nonsense-mediated decay A transcript with a coding sequence that finishes >50bp from a downstream splice site or if the variant does not cover the full reference coding sequence.

Nonsense-mediated decay is a process that detects nonsense mutations and prevents the expression of truncated or erroneous proteins.

4.5.10 Pfam domain analysis

The Pfam database contains a large collection of protein families, as well as protein domain annotation for a large number of protein isoforms. The annotation for these isoforms was retrieved using Biomart web-based tool [158]. The protein domains corresponding to each protein-coding transcript isoform in a switch event were compared and the number of different domains was reported.

Chapter 5

Integrating transcriptomics data from two datasets

The computational analyses herein described were performed by myself under the supervision of Dr. Alvis Brazma.

5.1 Introduction

The study presented in Chapter 4 describes a method for studying alternative splicing by exploring dominant transcripts of genes and investigating dominant transcript switches across tissues. To confirm the previous findings, I additionally investigated an independent RNA-seq dataset, the Genotype-Tissue Expression (GTEx) consortium dataset, which contains a much higher number of samples, as well as more tissues represented [3].

The GTEx project established a public resource database and associated tissue bank for studying tissue-specific gene expression and regulation in human tissues [3]. It contains samples from 54 conditions across almost 1000 individuals, for molecular assays such as whole genome sequencing and RNA-Seq. All GTEx tissue samples were examined histologically and were only included in the project if the tissue was both non-diseased and in the normal range considering the age of the donor. The RNA was extracted from postmortem samples as donors were enrolled in the study [162].

Here the GTEx dataset is used to investigate transcript dominance and switch events across conditions with the same approach used in Chapter 4. The results obtained with GTEx were compared with the ones obtained for the matching tissues of the dataset used in Chapter 4, which is called *Uhlen* dataset in the present chapter. In particular, the study here presented analyses if the same switch events are observed between pairs of the matching tissues across

the two datasets and if alternative splicing patterns are shared. This also allows to evaluate if the conclusions presented in Chapter 4 are robust.

The quantification scores for all genes and transcripts were calculated using Kallisto [15] for both datasets. The computation of these scores was done by Nuno A. Fonseca using Kallisto version 0.42.4.

5.2 Results

The dataset used in this study contains 19,972 samples of RNA-seq of coding RNA from samples representing 50 solid tissues, whole blood and 3 cell lines [162]. The number of biological replicates per tissue varies between 5 (Cervix - Endocervix) and 564 (Muscle - Skeletal), being on average 221 samples (Figure 5.1).

Similarly to the study described in the previous chapter, the goal was to compare isoform expression in the tissues for which I averaged the expression level of each transcript across the biological samples of each tissue. The exact same procedure was used in the current study to calculate transcript dominance, followed by the determination of dominant transcripts for each gene in each tissue. Lastly, a pairwise comparison of the tissues of the dataset allowed the determination of switch events.

5.2.1 Transcript dominance analysis

As before, for a transcript to be considered dominant, its average expression value has to be at least 2 times higher than the average expression value of the second most expressed transcript across the samples of a tissue. As defined in the previous chapter, two additional conditions had to be met: the gene had to be expressed in all samples ($\geq$ 1fpkm); and the transcript had to be the most expressed transcript in all samples of the tissue. This last condition was called support and gave extra confidence in the results, especially because the number of biological replicates per tissue was relatively low. The dataset used in the present chapter contains a considerably higher number of biological replicates for most tissues, so the effect of the support criterion was analysed by comparing the number of dominant transcripts using and excluding this condition.

In the GTEx dataset, the average number of genes expressed in a tissue was 9655 (Table 5.1), while in Uhlen dataset was 10,137 (Table 4.1). The values are similar, which is quite remarkable given that for a gene to be considered expressed, it has to be expressed in all samples of that tissue and, while some tissues in GTEx have hundreds of biological replicates (Figure 5.1), in Uhlen dataset the maximum number of samples is only 7. Of the

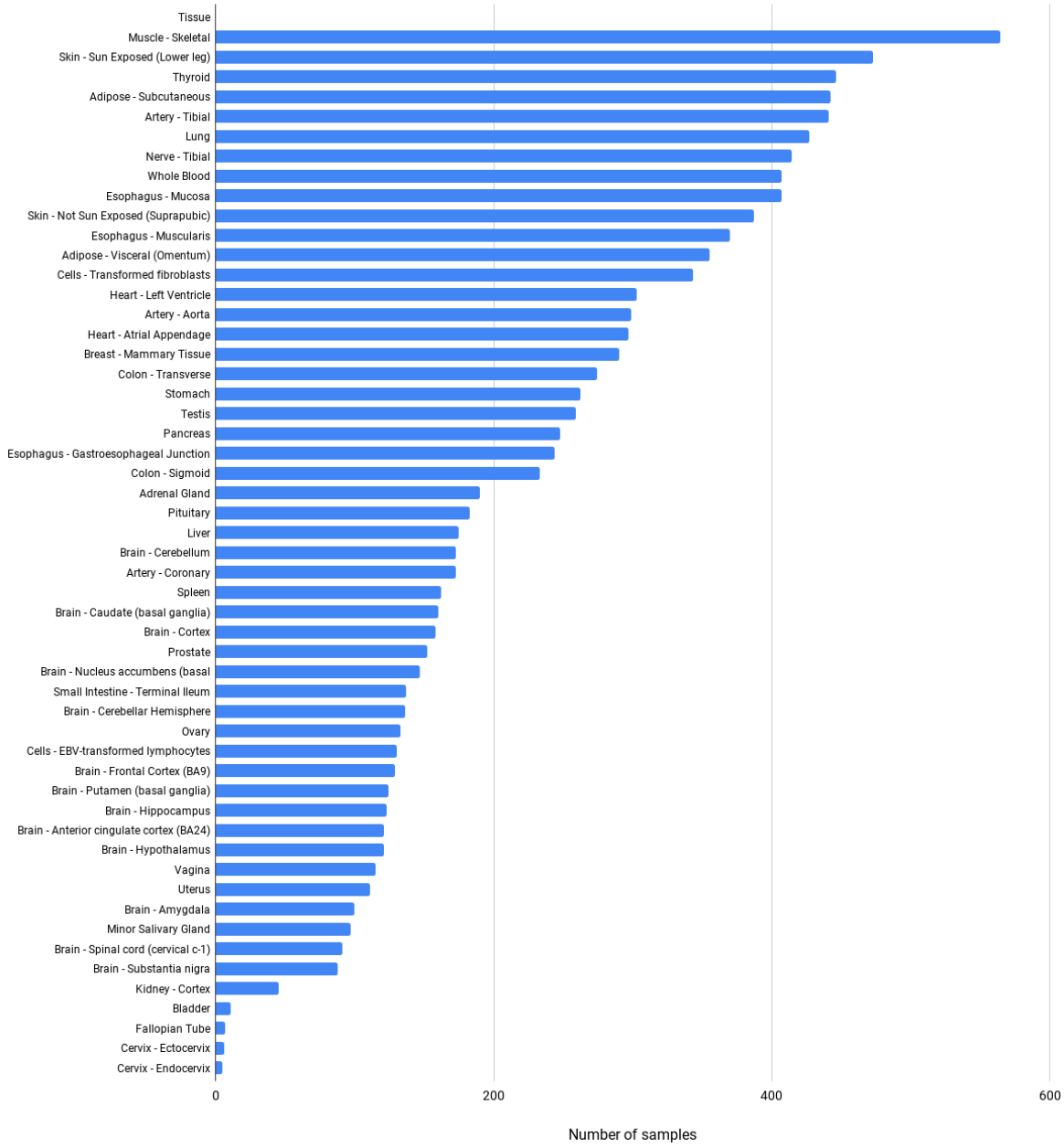


Figure 5.1 Number of samples of each tissue represented on GTEx dataset [20].

Tissue	n_exp	n_2-f		n_5-f		n_2-f_s		n_5-f_s	
Adipose_-_Subcutaneous	10135	6098	0.60	3214	0.32	3148	0.31	2542	0.25
Adipose_-_Visceral_(Omentum)	9763	5915	0.61	3129	0.32	3669	0.38	2778	0.28
Adrenal_Gland	9890	6020	0.61	3172	0.32	3513	0.36	2648	0.27
Artery_-_Aorta	9875	6053	0.61	3224	0.33	3047	0.31	2406	0.24
Artery_-_Coronary	10052	6160	0.61	3267	0.33	3821	0.38	2881	0.29
Artery_-_Tibial	9332	5900	0.63	3230	0.35	3072	0.33	2505	0.27
Bladder	11826	7141	0.60	3767	0.32	6444	0.54	3709	0.31
Brain_-_Amygdala	7463	4544	0.61	2568	0.34	2458	0.33	2013	0.27
Brain_-_Anterior_cingulate_cortex_(BA24)	9880	6128	0.62	3530	0.36	3761	0.38	3040	0.31
Brain_-_Caudate_(basal_ganglia)	8593	5257	0.61	2937	0.34	3275	0.38	2578	0.30
Brain_-_Cerebellar_Hemisphere	10402	6018	0.58	3040	0.29	3576	0.34	2662	0.26
Brain_-_Cerebellum	10812	6051	0.56	2925	0.27	2546	0.24	2021	0.19
Brain_-_Cortex	10140	6094	0.60	3324	0.33	2742	0.27	2329	0.23
Brain_-_Frontal_Cortex_(BA9)	9437	5858	0.62	3321	0.35	3335	0.35	2759	0.29
Brain_-_Hippocampus	9168	5615	0.61	3175	0.35	3325	0.36	2730	0.30
Brain_-_Hypothalamus	9041	5496	0.61	3109	0.34	3342	0.37	2700	0.30
Brain_-_Nucleus_accumbens_(basal_ganglia)	8059	4902	0.61	2723	0.34	2809	0.35	2272	0.28
Brain_-_Putamen_(basal_ganglia)	6776	4210	0.62	2449	0.36	2674	0.39	2145	0.32
Brain_-_Spinal_cord_(cervical_c-1)	8818	5342	0.61	2936	0.33	3407	0.39	2608	0.30
Brain_-_Substantia_nigra	9609	5909	0.61	3280	0.34	3803	0.40	2931	0.31
Breast_-_Mammary_Tissue	10219	6053	0.59	3116	0.30	3816	0.37	2785	0.27
Cells_-_EBV-transformed_lymphocytes	9478	5715	0.60	2948	0.31	3682	0.39	2462	0.26
Cells_-_Leukemia_cell_line_(CML)	10586	6448	0.61	3485	0.33	3883	0.37	2748	0.26
Cells_-_Transformed_fibroblasts	9304	5860	0.63	3229	0.35	2742	0.29	2112	0.23
Cervix_-_Ectocervix	11832	6910	0.58	3522	0.30	6150	0.52	3471	0.29
Cervix_-_Endocervix	12677	7359	0.58	3779	0.30	6837	0.54	3735	0.29
Colon_-_Sigmoid	10084	6060	0.60	3181	0.32	3619	0.36	2813	0.28
Colon_-_Transverse	9645	5771	0.60	2998	0.31	3513	0.36	2613	0.27
Esophagus_-_Gastroesophageal_Junction	9980	6011	0.60	3182	0.32	3720	0.37	2822	0.28
Esophagus_-_Mucosa	9081	5332	0.59	2713	0.30	2770	0.31	2192	0.24
Esophagus_-_Muscularis	9776	5957	0.61	3128	0.32	595	0.06	495	0.05
Fallopian_Tube	12480	7291	0.58	3751	0.30	6629	0.53	3690	0.30
Heart_-_Atrial_Appendage	8815	5290	0.60	2687	0.30	2588	0.29	2074	0.24
Heart_-_Left_Ventricle	5974	3603	0.60	1840	0.31	1591	0.27	1319	0.22
Kidney_-_Cortex	6802	3998	0.59	2091	0.31	2807	0.41	1918	0.28
Liver	8436	4911	0.58	2531	0.30	2713	0.32	2151	0.25
Lung	10077	5733	0.57	2631	0.26	406	0.04	362	0.04
Minor_Salivary_Gland	10462	6091	0.58	3097	0.30	4395	0.42	2887	0.28
Muscle_-_Skeletal	7112	4582	0.64	2545	0.36	1814	0.26	1603	0.23
Nerve_-_Tibial	10859	6281	0.58	3147	0.29	3489	0.32	2627	0.24
Ovary	10679	6050	0.57	2902	0.27	4003	0.37	2655	0.25
Pancreas	8925	5237	0.59	2637	0.30	3178	0.36	2313	0.26
Pituitary	11038	6321	0.57	3113	0.28	4110	0.37	2841	0.26
Prostate	10806	6152	0.57	3068	0.28	4051	0.37	2833	0.26
Skin_-_Not_Sun_Exposed_(Suprapubic)	9804	5754	0.59	2917	0.30	3317	0.34	2517	0.26
Skin_-_Sun_Exposed_(Lower_leg)	9087	5399	0.59	2765	0.30	525	0.06	453	0.05
Small_Intestine_-_Terminal_Ileum	10579	6093	0.58	3048	0.29	3925	0.37	2748	0.26
Spleen	11625	6671	0.57	3403	0.29	4540	0.39	3138	0.27
Stomach	9064	5455	0.60	2851	0.31	3316	0.37	2507	0.28
Testis	11738	6214	0.53	2814	0.24	3736	0.32	2498	0.21
Thyroid	10901	6138	0.56	2949	0.27	3425	0.31	2547	0.23
Uterus	11100	6439	0.58	3273	0.29	4238	0.38	2976	0.27
Vagina	10360	6009	0.58	2909	0.28	3861	0.37	2646	0.26
Whole_Blood	2926	1497	0.51	571	0.20	61	0.02	56	0.02
AVERAGE	9655.2	5729.6	0.59	2984.1	0.31	3366.9	0.35	2441.9	0.25

Table 5.1 Analysis of gene expression and transcript dominance per tissue in GTEx dataset (Kallisto quantification scores). The columns designate the following categories:
n_exp - number of genes expressed in the tissue;
n_2-f and *n_5-f* - number of genes with 2- and 5-fold dominant transcripts not using support criterion (defined on page 57);
n_2-f_s and *n_5-f_s* - number of genes with 2- and 5-fold dominant transcripts using support criterion. These last four columns contain two values per row, being the second the ratio between the first value and the number of expressed genes.

genes that are expressed, 59% and 31% have a 2- and 5-fold dominant transcript, respectively, but only if the support criterion is not used. If support is used to select dominant transcripts, only 35% and 25% of genes have 2- and 5-fold dominant transcript, respectively. Therefore, the use of support considerably reduces the number of genes with dominant transcripts. This is a result of requiring that, for a specific gene, all samples of a tissue to have as major transcript the dominant transcript determined for that gene and tissue. This means that, even if just a single sample does not match this condition, the transcript no longer is considered dominant. This condition is, of course, more difficult to meet in larger datasets, leading to the already mentioned effect. This effect can be observed even just by comparing GTEx sample tissues. The four tissues with the lowest number of samples (Cervix - Endocervix, Cervix - Ectocervix, Fallopian Tube and Bladder) show a less pronounced reduction of dominant transcripts when the support criterion is used, compared to the other tissues. This effect is particularly noticed between 2-fold dominant transcripts including and excluding the support condition (Table 5.1).

The number of dominant transcripts found in Uhlen dataset was higher than the ones found in GTEx, on average 68% versus 59% of expressed genes have 2-fold dominant transcripts. As mentioned before, the difference in number of samples influences the number of dominant transcripts but even the GTEx tissues with number of samples similar to Uhlen ("Cervix - Endocervix", "Cervix - Ectocervix" and "Fallopian Tube") have lower percentage of dominant transcripts (58% with no support and 52-54% with support).

5.2.2 Switch events

To evaluate if the dominant transcript of a gene changes across conditions, the number of switch events was determined for all tissues of GTEx in a pairwise manner and the level of similarity between tissues was determined using MDS as described in Chapter 4 (subsection 4.2.4).

It was observed that the number of switch events was low, considering the number of genes expressed in a given condition and the number of genes with dominant transcripts. The conditions with the highest number of 2-fold switch events were the 3 cell lines, testis and two brain regions (cerebellar hemisphere and cerebellum). Of the normal tissues, skeletal muscle, testis and all the brain regions had a high number of switches when compared to the others, with the median number of switches being close to 40 or more. With that said, the leukemia cell line was the condition with the highest median number of switches, a total of 140 (Figure 5.2) and the lowest median number of switches was 9 for stomach.

The MDS analysis revealed a clear separation of the brain regions, as well as a separation of the multiple cell lines (dark green circles and orange circles, respectively, in Figure 5.3 -

bottom). Similarly to what was observed in Uhlen dataset in the previous chapter (Figure 4.9), skeletal muscle, testis, and liver appeared isolated from the rest of the tissues. However, kidney cortex does not particularly stand out as in the Uhlen dataset, appearing now as part of the main cluster of tissues. It should also be noticed that, although tissues from the same regions (all the ones represented in circles) tend to be close to each other, the zoom plot (Figure 5.3 - top) shows that there are exceptions such as the two esophagus tissues.

In the case of 5-fold switch events, the cerebellar hemisphere (brain) was the condition with the highest median number of switches, a total of 7, and the lowest median number of switches was 0 for 22 tissues in the dataset (Figure 5.4). It was again observed that in the case of normal tissues, skeletal muscle, testis and all the brain regions had a higher number of switches when compared to the others, although even in these cases the median never exceeded 7. Regarding normal tissues, 14 was the maximum number of 5-fold switch events obtained and was observed between 3 pairs of tissues: "Brain - Cerebellar Hemisphere"/Testis; "Brain - Cerebellar Hemisphere"/"Muscle - Skeletal"; and "Brain - Cerebellum"/Testis.

The MDS analysis showed a clear separation of the brain regions and cell lines (Figure 5.5). Of the other tissues, skeletal muscle, testis, and liver are also isolated as before and similarly to what was observed in Uhlen dataset (Figure 4.11). With that said, pancreas is a tissue that appears more isolated than in the case of the 2-fold switch events and that can be more evidently seen in the zoom plot (Figure 5.5 - top). Once again, kidney cortex does not stand out and is part of the main cluster, unlike what was observed for Uhlen dataset. Lastly, it should be mentioned that, although the results are similar to what was observed for the 2-fold switch events, there is less separation between tissue types, which is quite evident in the zoom plot (Figure 5.5 - top).

The tissues with minimum and maximum median values differ from the ones determined for Uhlen dataset switch events but these datasets differ in several aspects that influence these results. GTEx contains a significantly higher number of tissues and number of biological replicates per tissue, contains 3 cell lines and 13 brain regions, which, although they are similar between each other, differ substantially from the other tissues, as can be seen from the number of switch events. Therefore, it is not straightforward to compare both datasets. On the one hand, the range of values for 2-fold switch events is similar, which can be observed in both heatmaps where the support criterion was not used (Figure 4.13, Figure 5.2). On the other hand, the number of 5-fold switch events in GTEx is consistently lower than the ones determined for Uhlen dataset (Figure 4.14, Figure 5.4) and the median number of switches for 24 tissues in GTEx is zero. This particular result might be related with the fact that GTEx has more samples and the transcript dominance calculation is based on a ratio of averages. The Uhlen dataset has a low number of samples, which means that a single abnormal value

in one can significantly shift the mean expression of a transcript, leading to a determination of a different dominant transcript and, consequently, causing switch events across all tissues that express another transcript. In datasets with hundreds of samples per tissue, such as GTEx, this effect is very unlikely to happen. Therefore, the GTEx results, specifically the more extreme cases of 5-fold dominant transcripts, strongly support the existence of a single dominant transcript dominance that is conserved across tissues. With all that being said, there are some tissues that standout as being more involved in switch events, these are skeletal muscle, testis and the brain regions.

5.2.3 GTEx/Uhlen datasets comparison

Following the determination of switch events for the tissues in GTEx dataset, a comparison was made between the switches for these and the matching conditions in Uhlen dataset. The tissues were matched according to their names and when a tissue of a dataset matched multiple tissues of the other, the first one was matched to all the other tissues, like in this example: "adipose tissue" in Uhlen dataset was matched to "Adipose - Subcutaneous" and "Adipose - Visceral (Omentum)" in GTEx. A total of 29 GTEx tissues were matched to tissues in Uhlen dataset. 2- and 5-fold switches were compared both using and excluding the criterion of support (page 57).

There were 9969 2-fold switch events found in common between the datasets, with testis being again the tissue with the highest median number of cases (55 switches) and the pair of tissues with the highest number of switch events was testis with cerebral cortex (78 switches) (Figure 5.6). As for the lowest number of switch events, lung was the lowest with 11, and there were several pairs of tissues with zero events. All the pairs of tissues with no switch events were tissues from the same regions (e.g. both cerebral cortex regions, both skin regions, etc.). This makes sense because these are cases of highly similar tissue regions.

The number of 2-fold switch events considerably decreased when the support criterion was used to define dominant transcripts. To illustrate this effect, it can be highlighted that testis had only 1 as the median number of 2-fold switches and 18 out of 29 tissues had a median of zero (Figure 5.7). The maximum number of switches was 11 and was observed between bladder with spleen and bladder with cerebral cortex. Overall, this again shows that using the support criterion in datasets with a large number of samples is a very stringent condition but even using this criterion, there were 450 2-fold switch events in common between datasets.

The tissue with the highest median number of 5-fold switches was skeletal muscle (2 switches) and the maximum number was 6, which was found for two pairs of tissues: skeletal muscle with bladder and skeletal muscle with fallopian tube (Figure 5.8). There were a high

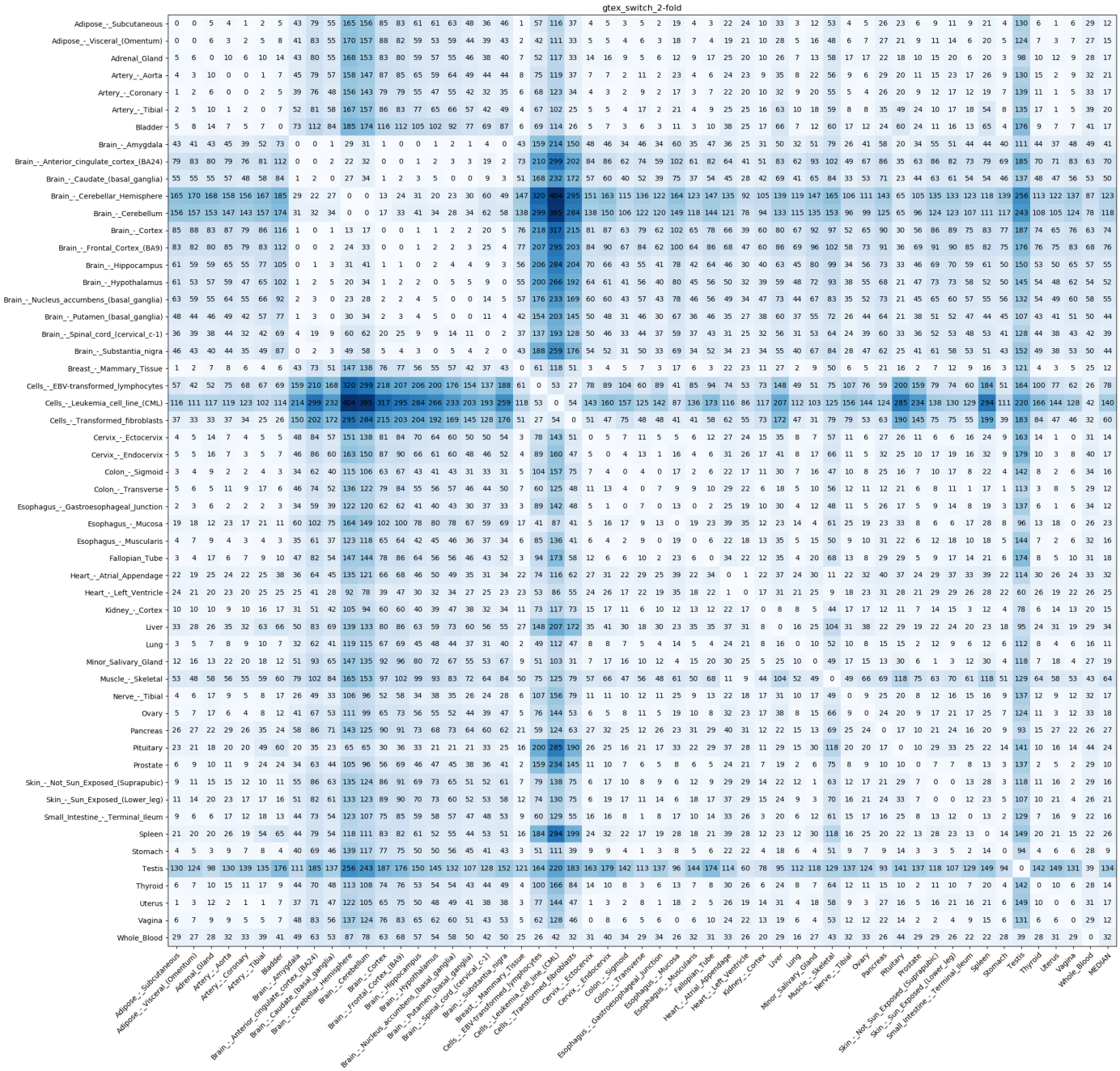


Figure 5.2 Number of 2-fold switch events for all pairs of tissues in GTEx dataset. Support criterion not used. The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

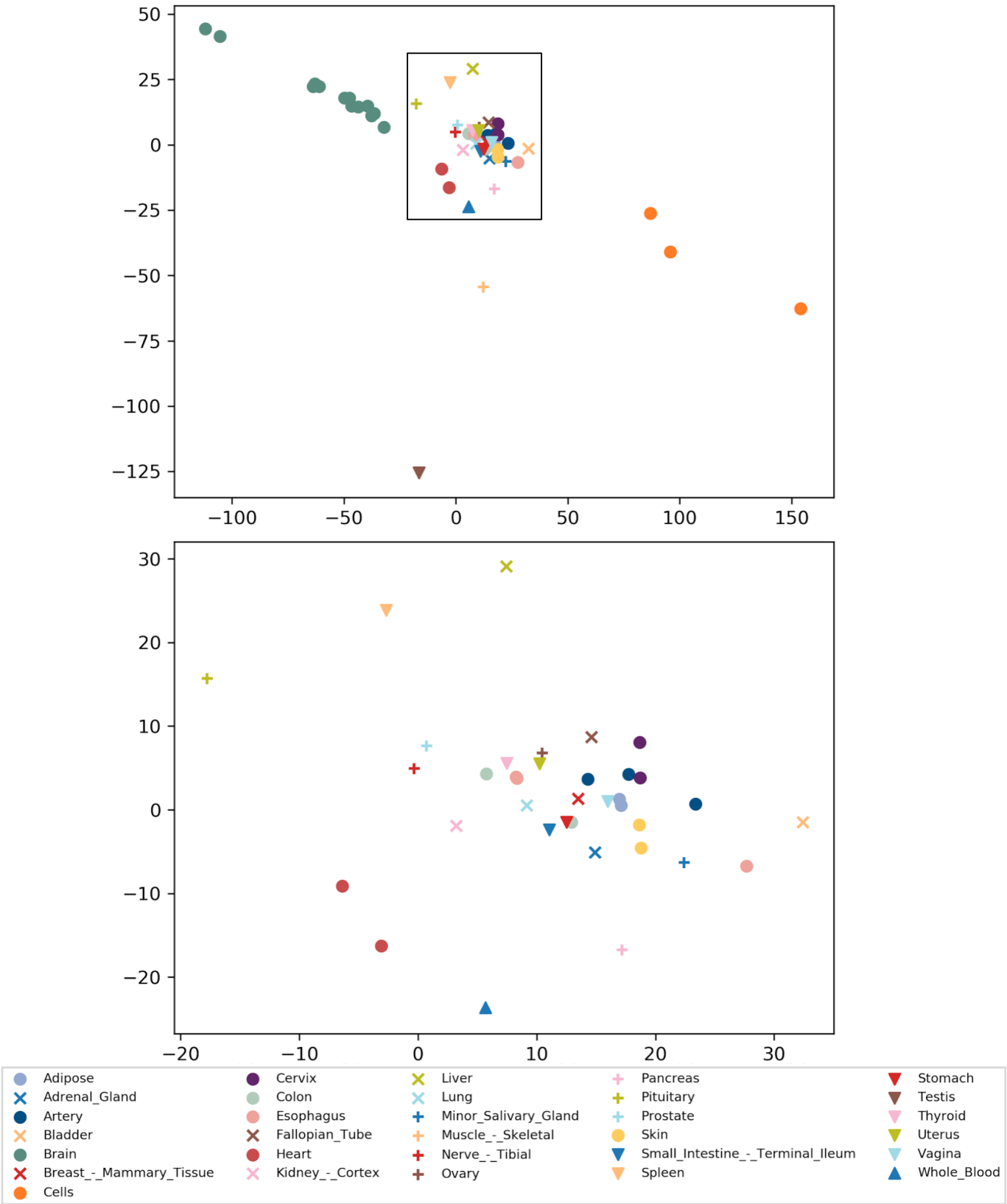


Figure 5.3 The top plot is a multidimensional scaling (MDS) analysis applied to 2-fold switch events. The bottom is a zoom plot of the region outlined with a rectangle on the top plot. A combination of colors and symbols was used to represent the tissues. All tissues of the same region were represented in the same color/symbol and in these cases only the prefix of the tissues was used in the legend (e.g. all brain regions were designated "brain" and were represented by green circles).

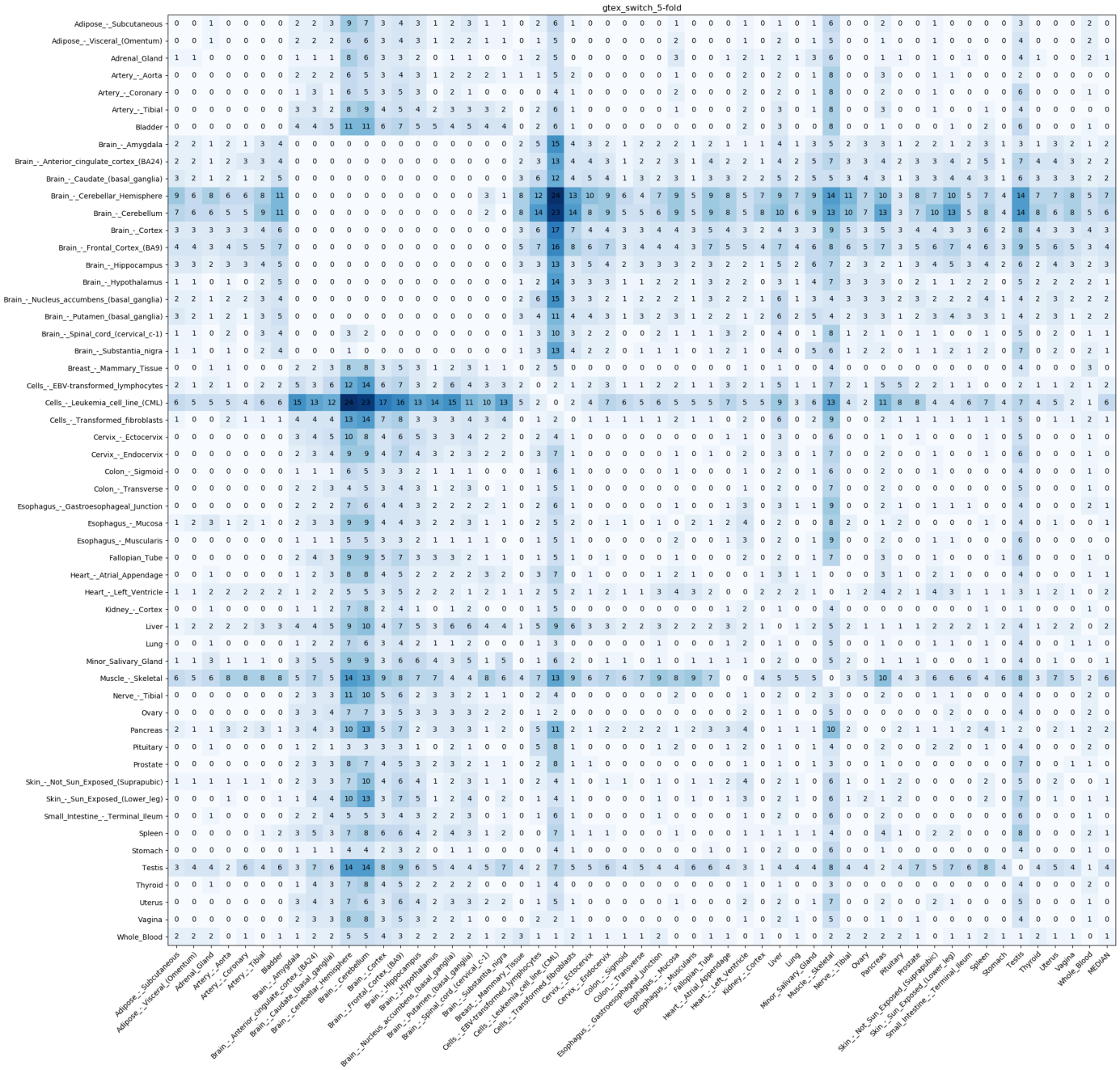


Figure 5.4 Number of 5-fold switch events for all pairs of tissues in GTEx dataset. Support criterion not used. The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

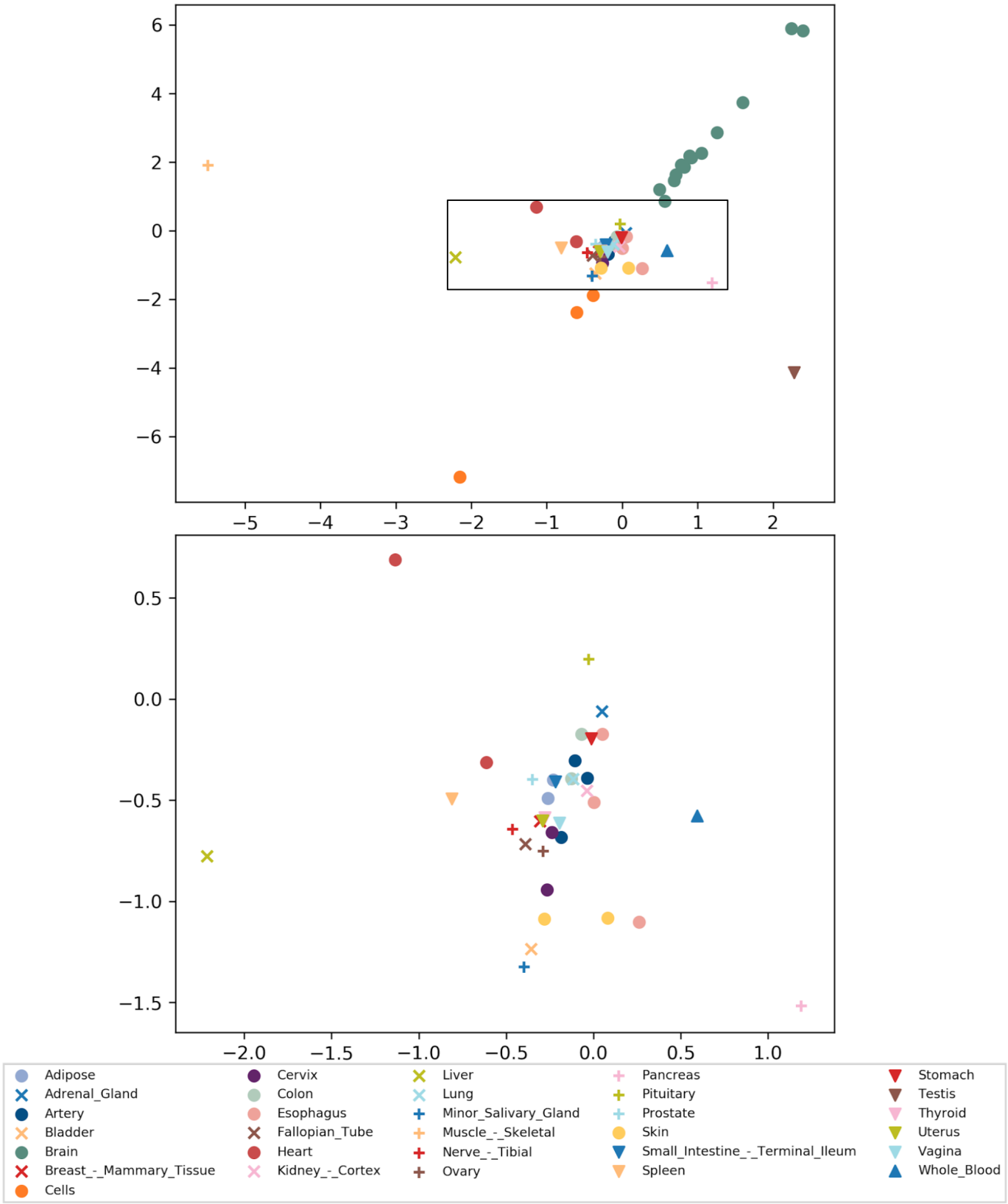


Figure 5.5 The top plot is a multidimensional scaling (MDS) analysis applied to 5-fold switch events. The bottom is a zoom plot of the region outlined with a rectangle on the top plot. A combination of colors and symbols was used to represent the tissues. All tissues of the same region were represented in the same color/symbol and in these cases only the prefix of the tissues was used in the legend (e.g. all brain regions were designated "brain" and were represented by green circles).

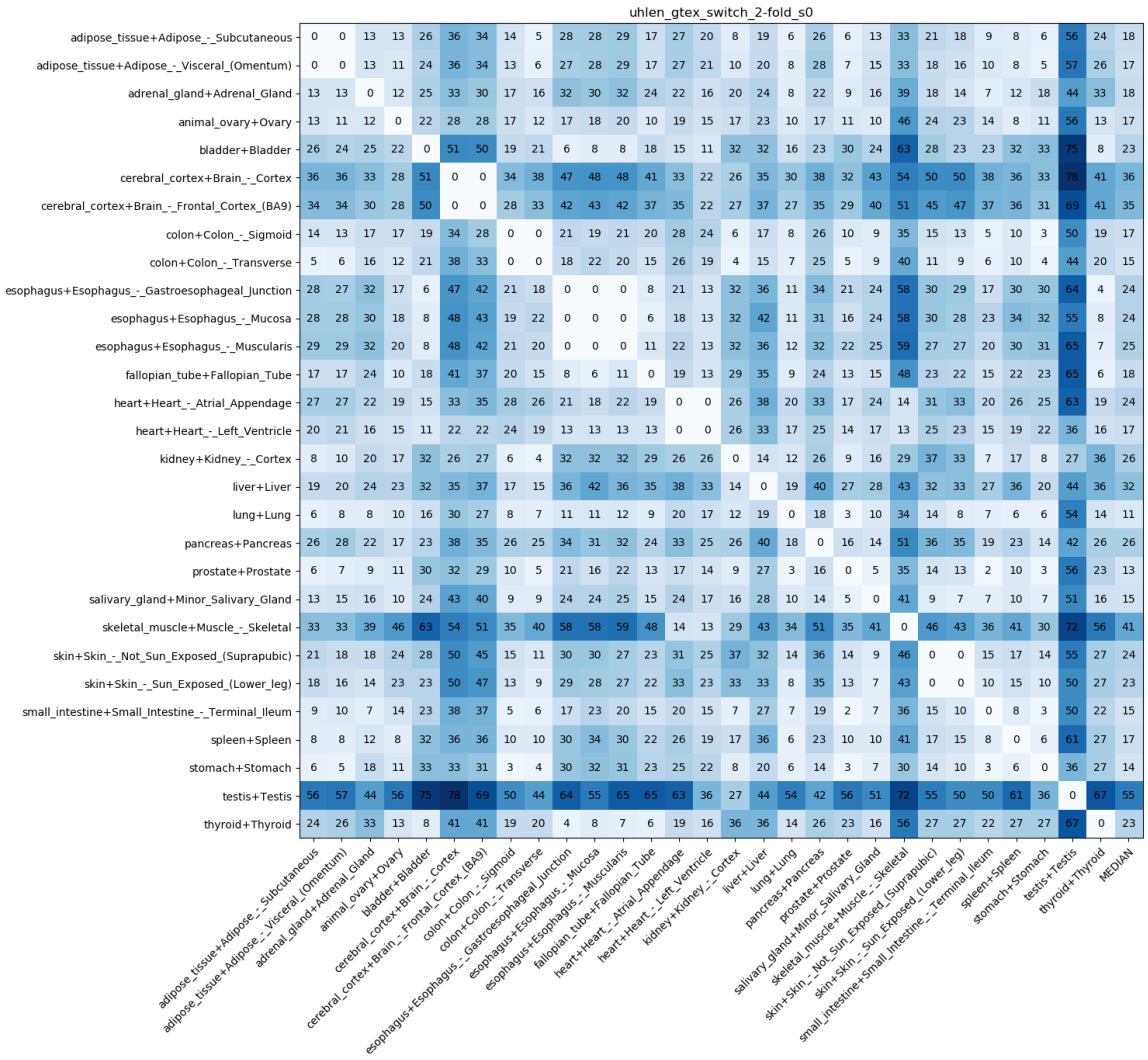


Figure 5.6 Number of common 2-fold switch events for all pairs of tissues in GTEx and Uhlen datasets. Support criterion not used (page 57). The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

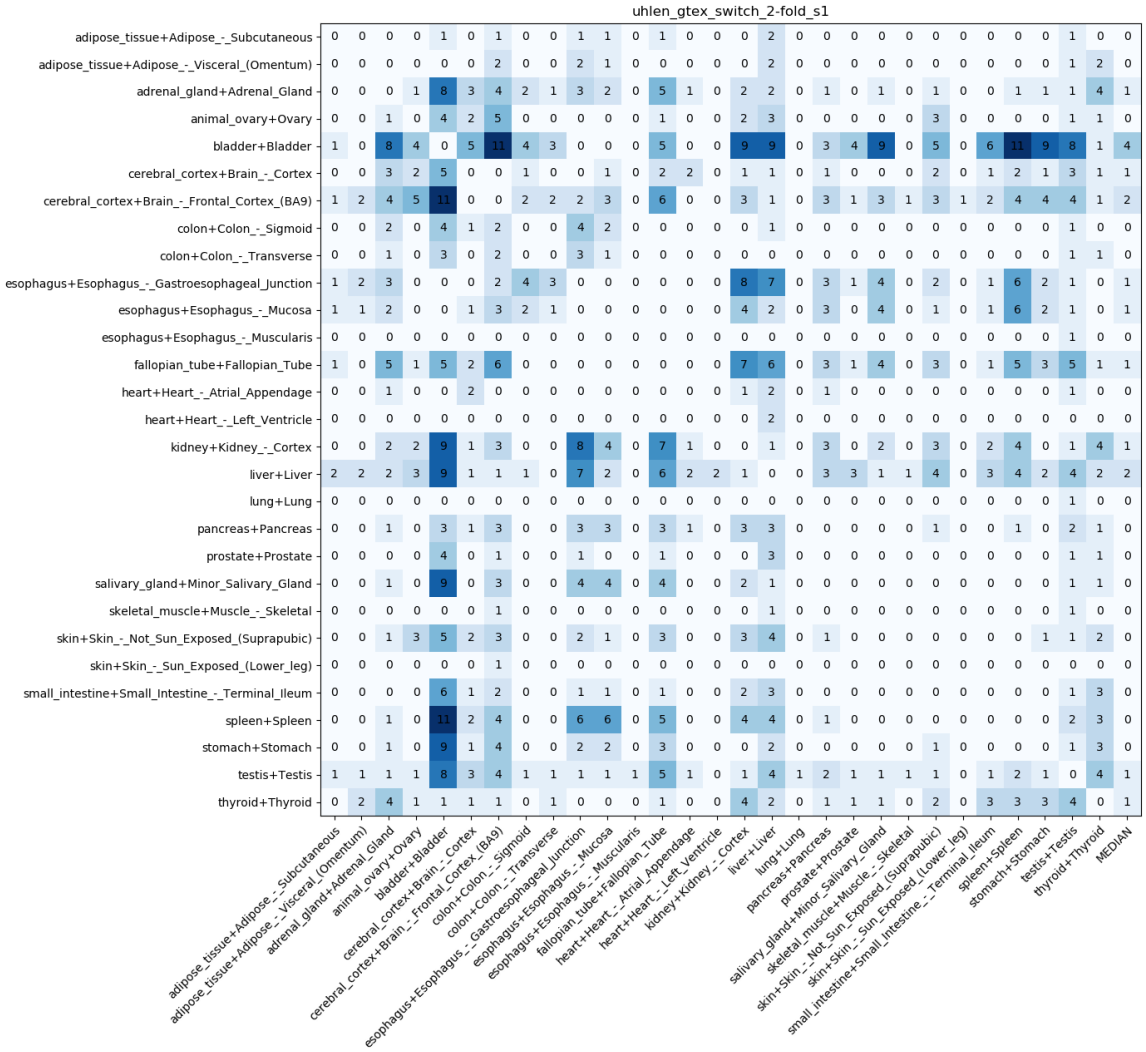


Figure 5.7 Number of common 2-fold switch events for all pairs of tissues in GTEx and Uhlen datasets. Support criterion used (page 57). The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

number of pairs of tissues with zero switch events and 19 out of 29 tissues had a median of zero 5-fold switches.

When support was used to define transcript dominance, only 20 out of 29 tissues being analysed had 5-fold switch events and even across them the number of switches was extremely low, being zero for most pairs of tissues (Figure 5.9). Frontal cortex was the tissue with the largest number of switches, having switch events across 17 tissue pairs, but the number of switches for a given pair never exceeded 1. In 16 of these cases, the switches involved the ENSG00000171992 gene (ENST00000307662 switching to ENST00000519664) and in one of the cases involved ENSG00000157368 (ENST00000429149 switching to ENST00000288098).

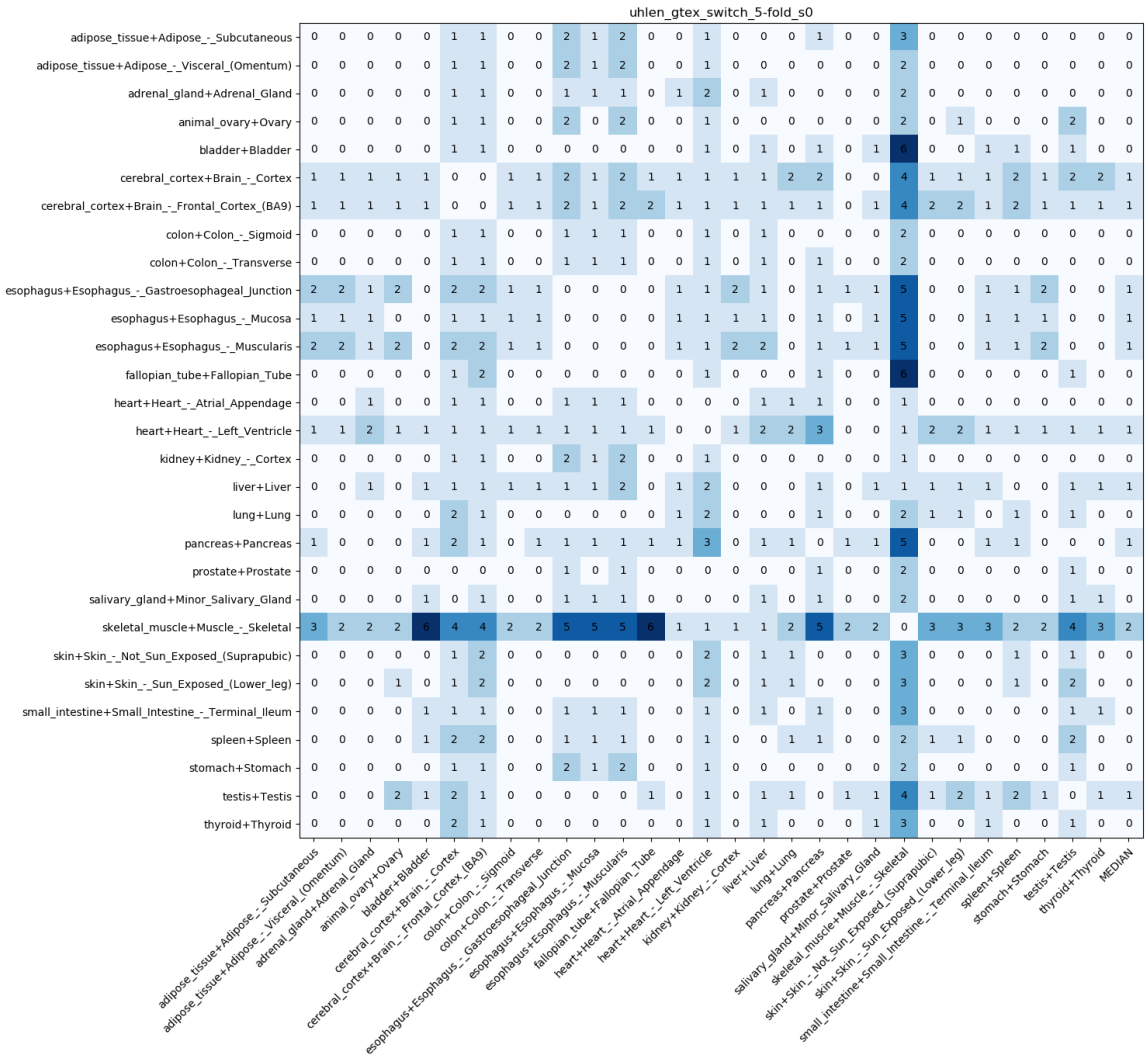


Figure 5.8 Number of common 5-fold switch events for all pairs of tissues in GTEx and Uhlen datasets. Support criterion not used (page 57). The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

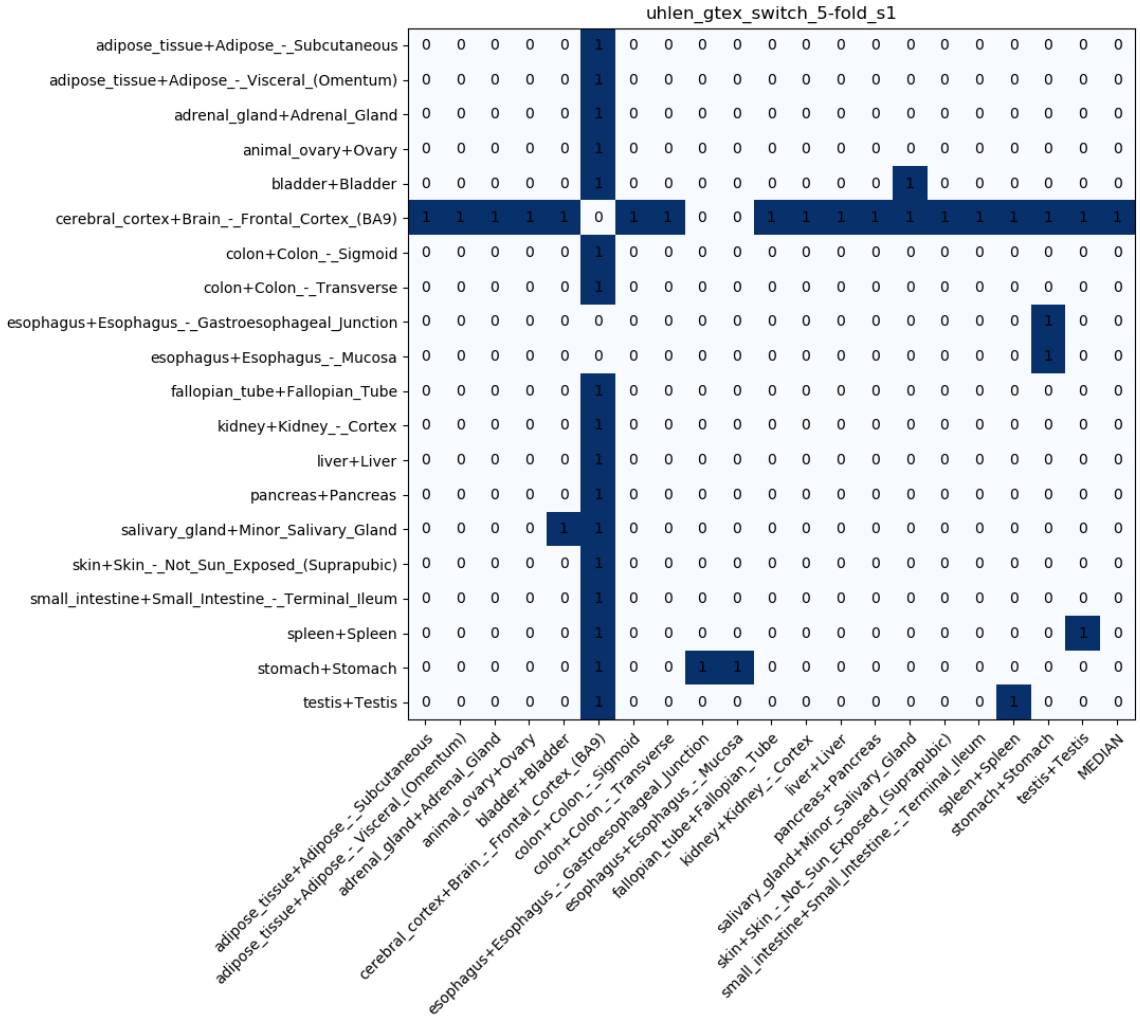


Figure 5.9 Number of common 5-fold switch events for all pairs of tissues in GTEx and Uhlen datasets. Support criterion used (page 57). The color scale is proportional to the number of switch events: the higher the number, the darker the blue tone.

5.2.4 Comparison of exons in switch transcripts

Like in the previous study, the exons of the pairs of transcripts that switch were compared. In this case, the transcripts compared were the ones from 2-fold switch events common to GTEx and Uhlen datasets (support criterion not used). It can be seen that most switch transcripts differ in a relatively small number of exons and the most common number of different exons was 4 between two switching transcripts (Figure 5.10), which is in accordance to what was observed before (Figure 4.15).

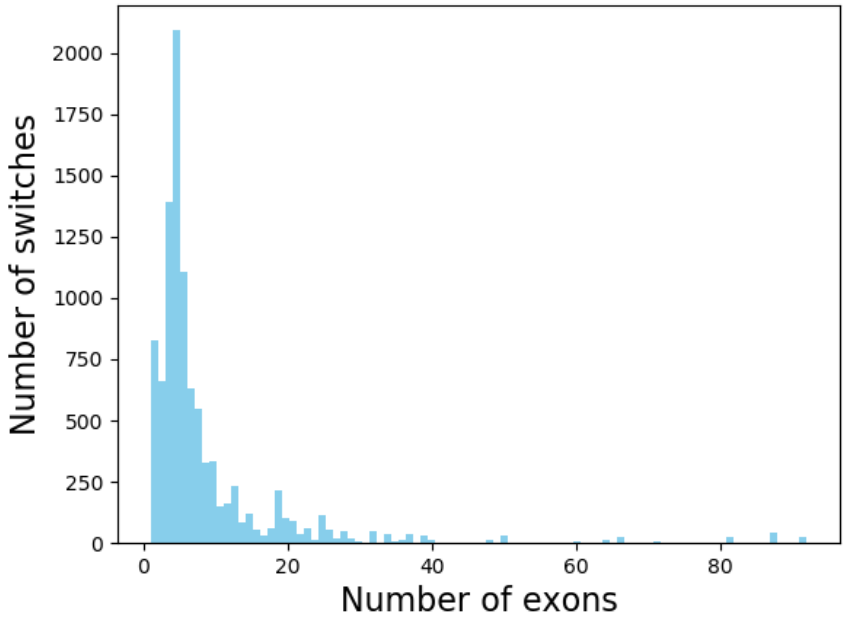


Figure 5.10 Distribution of the number of exons that differ between pairs of transcripts in a 2-fold switch event. On the x-axis is represented the number of exons that are different and on the y-axis is the number of pairs of transcripts (number of switches).

5.2.5 Alternative splicing types

To explore the changes driven by alternative splicing, the types of alternative splicing that occur between pairs of transcripts in 2-fold switch events common to GTEx and Uhlen datasets were analysed. The two most common types of alternative splicing found were alternative $3'$ and alternative $5'$ splice site selection, corresponding to 24.5% and 23.5% respectively. These were followed by alternative polyadenylation and alternative promoter, 15.0% and 11.8% respectively (Figure 5.11, blue columns). The first two cases account for 48% of the cases and both represent relatively small changes in the transcripts because both events are changes in a splice site of an exon (changes of less than one exon between the two

transcripts). These results are also similar to the ones obtained for Uhlen dataset (Figure 5.11, red columns). What is also similar between both is the relationship between splicing types, in other words, the ranking of their frequency. One of the reasons for these similarities is that the set of switches in common to GTEx and Uhlen datasets are of course a subset of Uhlen switches. With that said, some differences were also observed and the percentages of some splicing types differ between the two sets of switch events. The two splicing types that differ the most in frequency are alternative polyadenylation (21% for Uhlen and 15% for Uhlen/GTEx) and mutually exclusive exons (6.9% for Uhlen and 10.4% for Uhlen/GTEx).

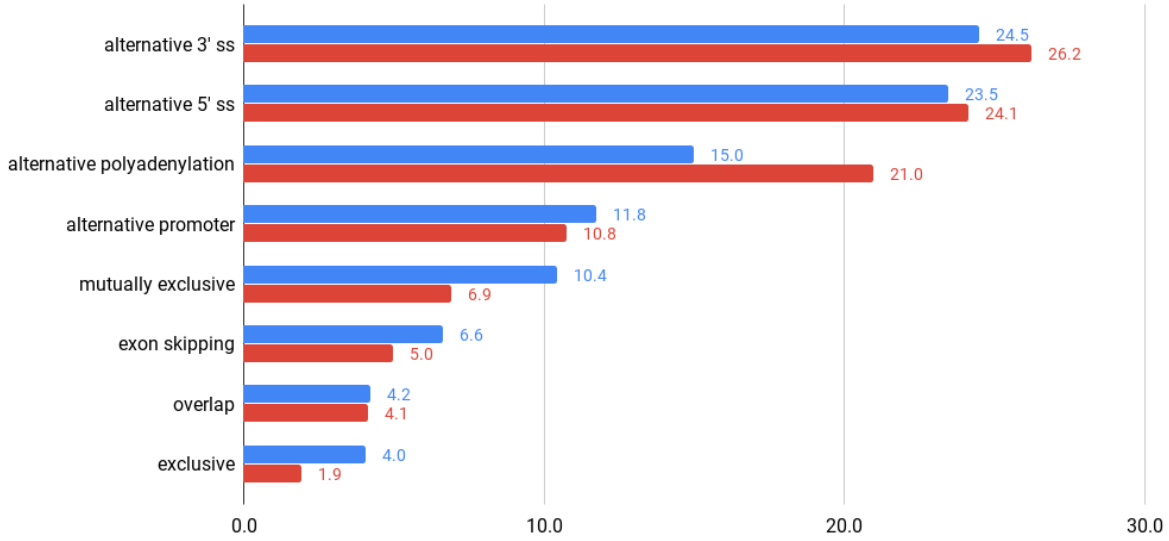


Figure 5.11 Comparison of the percentage of alternative splicing types occurring between transcripts in 2-fold switch events in Uhlen dataset (red) and the ones common to GTEx and Uhlen (blue). The types of alternative splicing are alternative 3′ splice site selection; alternative 5′ splice site selection; alternative polyadenylation; alternative promoter; mutually exclusive exons; exon skipping; overlapping exons; and exclusive exon. Besides the most common types of alternative splicing, two other categories were added: "overlap", for cases of overlapping exons that do not fit any of the other splicing categories; and "exclusive", for cases of exons that are exclusive to one of the transcript and do not fit any other splicing category.

5.2.6 Sequence identity

The impact of the exon changes in the sequence of the transcripts was evaluated by calculating the sequence identity of switch event isoforms using BLASTN [18] and using an implementation of the Needleman-Wunsch algorithm [144, 145], a global alignment algorithm, as described in the previous chapter.

It was observed that the sequence identity of pairs of switch transcripts is higher than 50% in most cases (Figure 5.12, Figure 5.13), although there are exceptions just like it has

been observed in the study in Chapter 4. With that said, the distribution here observed is different from the one previously reported (Figure 4.17, Figure 4.18). In particular, there is a peak around 60%, while in the previous case the peak was around 90%. This indicates that many of Uhlen's switch events involving highly similar transcripts are not found in GTEx, which might be related to the difficulty of the mapping and quantification software in attributing reads to the right transcript isoform in this situations, leading to the lack of consistency across datasets.

Again, it should be mentioned that the switch events here analysed are a subset of the ones previously analysed and this subset contains a significantly fewer number of elements.

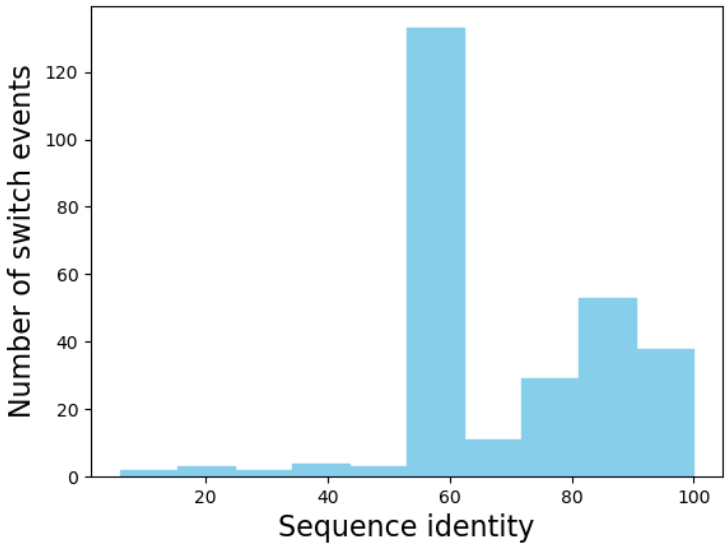


Figure 5.12 Distribution of the DNA sequence identity between pairs of switch transcripts calculated with BLASTN [18] for 5-fold switch events common to GTEx and Uhlen datasets.

5.2.7 Exon overlapping analysis

We noticed that in some cases the exons that differ between the switching transcripts are in fact overlapping. The overlapping exons were compared to evaluate to what extent they overlap each other. The percentage of overlap between the exons was calculated and it was observed that most exons overlap more than 95% (Figure 5.14). In the previous study, this analysis revealed something similar, with the distribution of overlap also having a peak at 95% overlap. This effect is even more pronounced than what was observed just for Uhlen dataset (Figure 4.20). In the case of Uhlen dataset, the distribution had a tail that decreased with decreasing overlap percentage, which is not observed in the current case.

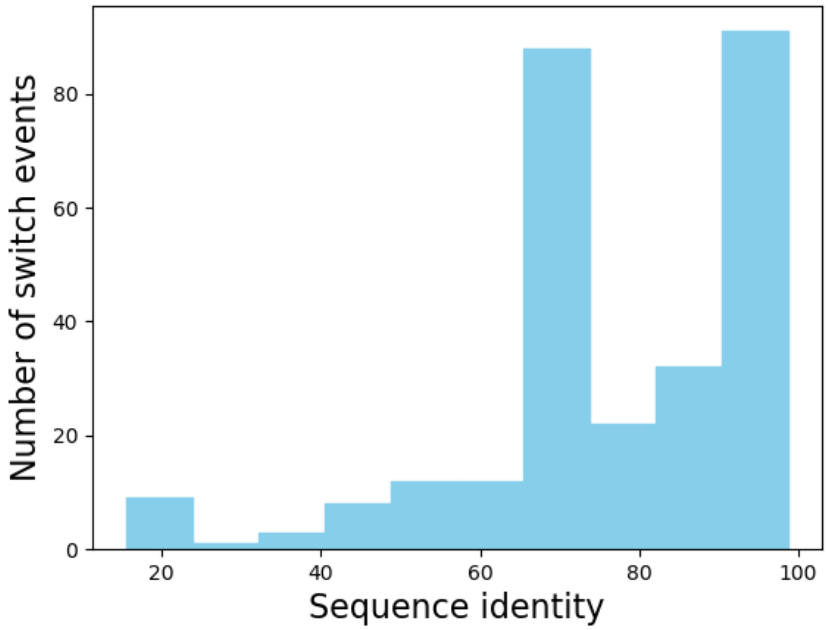


Figure 5.13 Distribution of the DNA sequence identity between pairs of switch transcripts calculated with a global aligner (needle) for 5-fold switch events common to GTEx and Uhlen datasets.

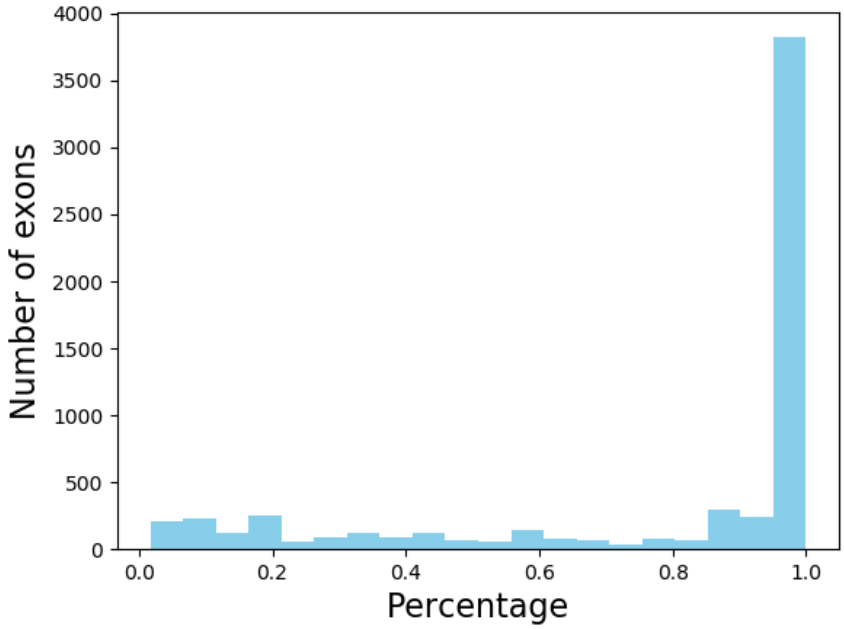


Figure 5.14 Distribution of the overlap percentage between exons of 2-fold switch events common to GTEx and Uhlen datasets.

5.2.8 Protein domain analysis

To determine how changes in exons translate into changes in protein function, the protein domains of each isoform in a switch event were compared as described in Chapter 4.

N domain changes	N switches 2-fold	% 2-fold	N switches 5-fold	% 5-fold
0	7204	87.2	240	92.0
1	619	7.5	17	6.5
2	276	3.3	2	0.8
3	101	1.2	2	0.8
≥4	57	0.7	0	0

Table 5.2 Number of common switch events in GTEx and Uhlen datasets with domain changes and respective percentages.

It was observed that in 87.2% of 2-fold switch events there were no protein domain changes between the isoforms. In the case of 5-fold switch events, the percentage is slightly higher, 92.0% of switch events have no domain changes (Table 5.2). In both cases, the values are higher than what was found just for Uhlen dataset, where around 80% of both 2- and 5-fold switches have no domain changes. This confirms that most switch events do not cause protein domain changes, indicating that is unlikely that the function of the isoforms is altered in a substantial way.

5.3 Examples

The examples presented in this section are 5-fold switch events common to both Uhlen and GTEx datasets where domain changes were found between the transcripts that switch. There were in total 21 switch events involving 5 genes (Table 5.3) that have transcript isoforms coding for different domains. All of these switches are described in this section. The protein domain identifiers used here are from Pfam database [140].

5.3.1 KIF1B - ENSG00000054523

The KIF1B gene encodes a motor protein that transports mitochondria and synaptic vesicle precursors. It is called kinesin family member 1B protein, which belongs to the kinesin family of proteins. The function of these proteins is intracellular transport and they are constituted by two elements. One element is a motor domain, that provides the power to move the protein and its cargo across a track-like system. The other element binds specific materials and is variable among members of this family [163].

KIF1B has a total of 12 annotated transcript isoforms, 7 of which are protein-coding. Two of these protein-coding isoforms were found to be dominant in different conditions (Table 5.4, Figure 5.15): ENST00000377093 (dominant in skeletal muscle) and ENST00000377086 (dominant in pancreas and cerebral cortex). These two isoforms have a 57.6% sequence

Gene name	Gene id	Transcripts id
KIF1B	ENSG00000054523	ENST00000377093 ENST00000377086
SLC35E4	ENSG00000100036	ENST00000343605 ENST00000451479
MARK4	ENSG00000007047	ENST00000262891 ENST00000300843
PTER	ENSG00000165983	ENST00000378000 ENST00000423462
PSD3	ENSG00000156011	ENST00000327040 ENST00000518315

Table 5.3 List of genes with 5-fold switch events common to Uhlen and GTEx datasets and respective dominant transcripts.

identity (Needleman-Wunsch) and differ in 3 protein domains that are not present in ENST00000377093 (skeletal muscle): PF12423, PF12473, and PF00169 (Figure 5.16, Figure 5.17). Of the 3 domains, PF12423 is an anterograde motor for the transport of mitochondria [164], PF12473 (DUF domain) is of unknown function and PF00169 (PH domain) is a pleckstrin homology domain that participates in the recruitment of proteins to different membranes, directing them to the appropriate cellular compartment [165]. The dominant isoform in skeletal muscle does not have these 3 domains but it still has 3 others: a kinesin, a kinesin-associated and a forkhead-associated domain. The kinesin domain enables the protein to move along microtubules [166] and the forkhead-associated domain is a phosphopeptide recognition domain that binds phosphothreonine-containing epitopes [167]. This indicates that this isoform is still functional because it contains the motor protein domain and is able to bind certain proteins. The fact that it is not able to bind and, consequently, transport mitochondria in an anterograde manner is most likely related with the role of mitochondria in skeletal muscle.

Tissue 1	Tissue 2	Transcript ID 1	Transcript ID 2
skeletal muscle	pancreas	ENST00000377093	ENST00000377086
skeletal muscle	cerebral cortex	ENST00000377093	ENST00000377086

Table 5.4 List of switch events common to Uhlen and GTEx datasets for KIF1B gene. "Tissue 1" and "Tissue 2" are the two switch event conditions and "Transcript ID 1" and "Transcript ID 2" are the correspondent transcript identifiers.

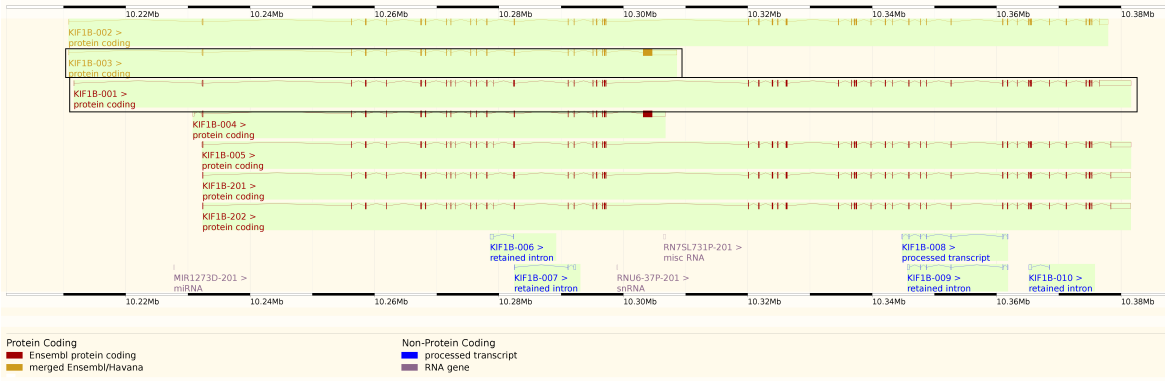


Figure 5.15 View of the transcript isoforms of KIF1B (ENSG00000054523) gene in Ensembl genome browser [19]. The two transcripts involved in switch events are outlined with black rectangles: KIF1B-003 and KIF1B-001 correspond to ENST00000377093 and ENST00000377086, respectively.

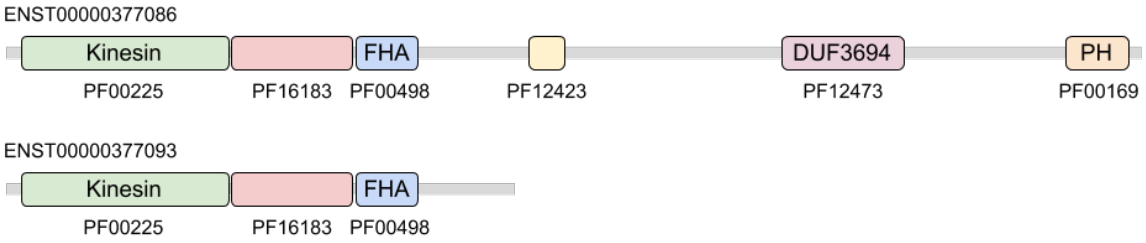


Figure 5.16 Protein domain struture for the two isoforms of KIF1B (ENSG00000054523) involved in switch events: ENST00000377093 (dominant in skeletal muscle) and ENST00000377086 (dominant in pancreas and cerebral cortex). The Pfam domain identifiers are displayed below the domains, which are represented by differently colored rectangles. Adapted from Pfam database [21].

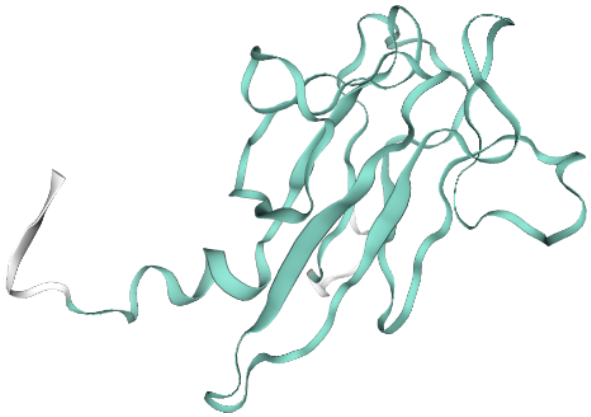


Figure 5.17 Structure of the FHA domain from KIF1B (2eh0 on PDBe [22]).

5.3.2 SLC35E4 - ENSG00000100036

SLC35E4 encodes the carrier family 35 member E4 protein, which is a transmembrane protein, a nucleotide-sugar transporter [168]. This gene has 3 isoforms, all of them protein-coding (Figure 5.18), and two of which are dominant in different conditions (Table 5.5): ENST00000343605 (dominant in skin and cerebral cortex) and ENST00000451479 (dominant in testis). These two isoforms have 52.6% sequence identity and differ in two protein domains that are not present in ENST00000451479 (testis): PF03151 and PF08449 (Figure 5.19). The PF03151 (TPT) domain is a transporter with specificity for triose phosphate and PF08449 (UAA) domain is a transporter with specificity for UDP-N-acetylglucosamine [169]. Since the ENST00000451479 isoform does not contain any of the annotated domains, it is possible that in testis, the SLC35E4 gene expresses a non-functional protein.

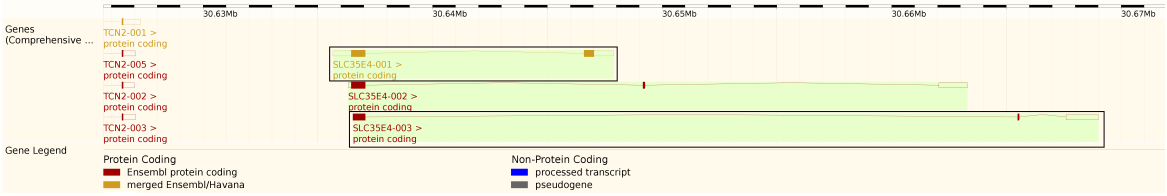


Figure 5.18 View of the transcript isoforms of SLC35E4 (ENSG00000100036) gene in Ensembl genome browser [19]. The two transcripts involved in switch events are outlined with black rectangles: SLC35E4-001 and SLC35E4-003 correspond to ENST00000343605 and ENST00000451479, respectively.

Tissue 1	Tissue 2	Transcript ID 1	Transcript ID 2
testis	skin	ENST00000451479	ENST00000343605
testis	cerebral cortex	ENST00000451479	ENST00000343605

Table 5.5 List of switch events common to Uhlen and GTEx datasets for SLC35E4 gene. "Tissue 1" and "Tissue 2" are the two switch event conditions and "Transcript ID 1" and "Transcript ID 2" are the correspondent transcript identifiers.

5.3.3 MARK4 - ENSG00000007047

MARK4 encodes the microtubule affinity-regulating kinase 4. These kinases phosphorylate microtubule-associated proteins, regulating the transition between stable and dynamic microtubules. The MARK4 protein is associated with the centrosome throughout mitosis and might be involved in the control of the cell cycle [170].

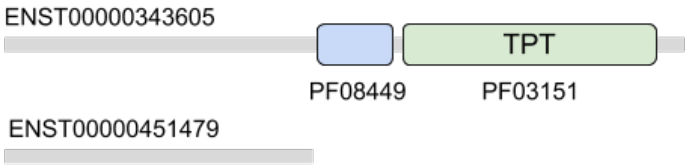


Figure 5.19 Protein domain struture for the two isoforms of SLC35E4 (ENSG00000100036) involved in switch events: ENST00000343605 (dominant in skin and cerebral cortex) and ENST00000451479 (dominant in testis). The Pfam domain identifiers are displayed below the domains, which are represented by differently colored rectangles. Adapted from Pfam database [23] with PF08449 added in the most likely location based on the exon structure.

MARK4 has 11 transcript isoforms, 6 of which are protein-coding. Two of the isoforms are dominant in different tissues (Table 5.6, Figure 5.20): ENST00000262891 (dominant in fallopian tube, skin, esophagus, small intestine, prostate, pancreas, thyroid and salivary gland) and ENST00000300843 (dominant in skeletal muscle). These two isoforms have 67.5% sequence identity and differ in only one protein domain (PF02149) which is not present in ENST00000300843 (Figure 5.21, Figure 5.22). PF02149 is a KA1 domain, whose function is not yet known [171], therefore it is difficult to predict what is the impact of the isoform switch in this case. Nevertheless, since both isoforms have a Pkinase domain (PF00069) and a ubiquitin-associated domain (PF00627), they are probably both functional because they have the kinase function, as well as the capacity to recognise ubiquitin [172], which is a signal added to incorrectly folded proteins so they can be degraded.

Tissue 1	Tissue 2	Transcript ID 1	Transcript ID 2
skeletal muscle	fallopian tube	ENST00000300843	ENST00000262891
skeletal muscle	skin	ENST00000300843	ENST00000262891
skeletal muscle	esophagus	ENST00000300843	ENST00000262891
skeletal muscle	small intestine	ENST00000300843	ENST00000262891
skeletal muscle	prostate	ENST00000300843	ENST00000262891
skeletal muscle	pancreas	ENST00000300843	ENST00000262891
skeletal muscle	thyroid	ENST00000300843	ENST00000262891
skeletal muscle	salivary gland	ENST00000300843	ENST00000262891

Table 5.6 List of switch events common to Uhlen and GTEx datasets for MARK4 gene. "Tissue 1" and "Tissue 2" are the two switch event conditions and "Transcript ID 1" and "Transcript ID 2" are the correspondent transcript identifiers.

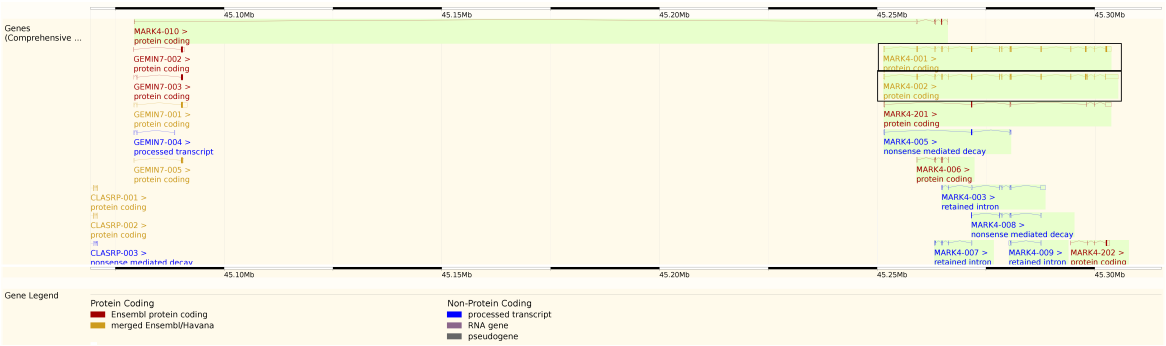


Figure 5.20 View of the transcript isoforms of MARK4 (ENSG00000007047) gene in Ensembl genome browser [19]. The two transcripts involved in switch events are outlined with black rectangles: MARK4-001 and MARK4-002 correspond to ENST00000262891 and ENST00000300843, respectively.

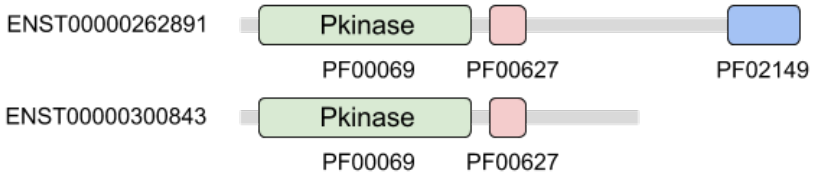


Figure 5.21 Protein domain struture for the two isoforms of MARK4 (ENSG00000007047) involved in switch events: ENST00000262891 (dominant in fallopian tube, skin, esophagus, small intestine, prostate, pancreas, thyroid and salivary gland) and ENST00000300843 (dominant in skeletal muscle). The Pfam domain identifiers are displayed below the domains, which are represented by differently colored rectangles. Adapted from Pfam database [24].

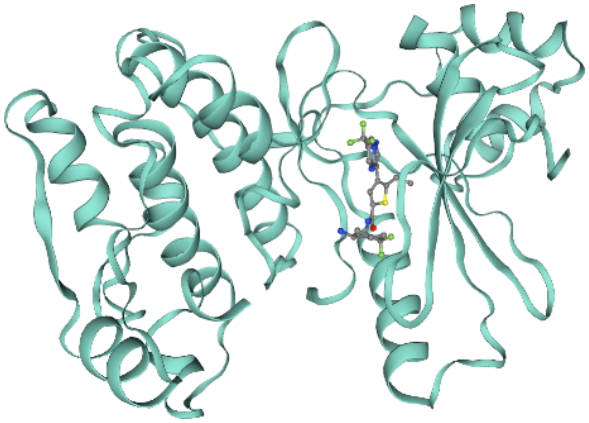


Figure 5.22 Structure of the Pkinase and UBA (PF00627) domains from MARK4 (5es1 on PDBe [22]).

5.3.4 PTER - ENSG00000165983

PTER encodes a phosphotriesterase-related protein, which is located in extracellular regions or secreted. Phosphotriesterase enzymes catalyze the conversion of aryl dialkyl phosphate

to dialkyl phosphate and aryl alcohol [173]. This gene has 5 annotated transcript isoforms of which 4 are protein-coding and it has two dominant transcripts (Table 5.7, Figure 5.23): ENST00000378000 (dominant in testis) and ENST00000423462 (dominant in skin). These two isoforms have 84.6% sequence identity and differ in one protein domain (PF01026) which is not present in ENST00000423462 (skin) (Figure 5.24). This particular domain is called TatD and it functions as a DNase [174], which means that the isoform expressed in skin does not have the function of DNA cleavage, although it might still be a functional enzyme because it contains the phosphotriesterase domain (PTE - PF02126).

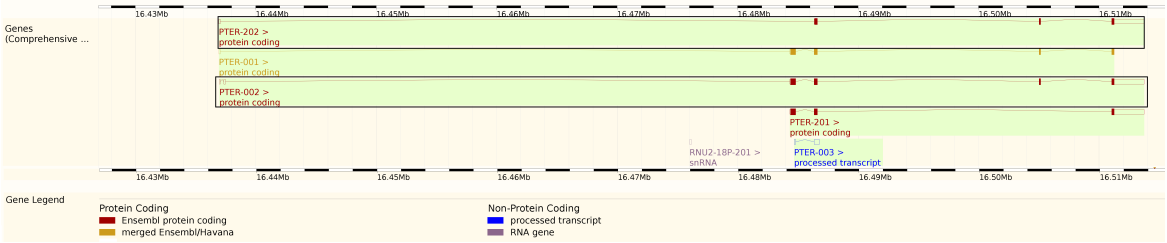


Figure 5.23 View of the transcript isoforms of PTER (ENSG00000165983) gene in Ensembl genome browser [19]. The two transcripts involved in switch events are outlined with black rectangles: PTER-002 and PTER-202 correspond to ENST00000378000 and ENST00000423462, respectively.

Tissue 1	Tissue 2	Transcript ID 1	Transcript ID 2
testis	skin	ENST00000378000	ENST00000423462

Table 5.7 List of switch events common to Uhlen and GTEx datasets for PTER gene. "Tissue 1" and "Tissue 2" are the two switch event conditions and "Transcript ID 1" and "Transcript ID 2" are the correspondent transcript identifiers.

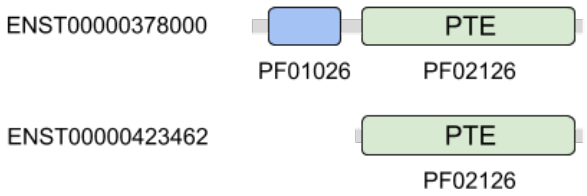


Figure 5.24 Protein domain struture for the two isoforms of PTER (ENSG00000165983) involved in switch events: ENST00000378000 (dominant in testis) and ENST00000423462 (dominant in skin). The Pfam domain identifiers are displayed below the domains, which are represented by differently colored rectangles. Adapted from Pfam database [25] with PF01026 added in the most likely location based on the exon structure.

5.3.5 PSD3 - ENSG00000156011

PSD3 encodes the "pleckstrin and Sec7 domain containing 3" protein, which is involved in intracellular signaling and can be a constituent of the cytoskeleton. This protein binds Phosphatidylinositol lipids in biological membranes and proteins, such as the protein kinase C. Pleckstrin homology domains participate in the recruitment of proteins to different membranes, directing them to particular cellular compartments or allowing the interaction between them and other components of signal transduction pathways [175].

This gene has 20 isoforms, 11 of which are protein-coding, and it has two dominant isoforms (Table 5.8, Figure 5.25): ENST00000327040 (dominant in liver, cerebral cortex and ovary) and ENST00000518315 (dominant in testis). These two isoforms have 15.5% sequence identity and differ in one protein domain only (PF15410) which is not present in ENST00000518315 (testis) (Figure 5.26). This domain is a pleckstrin homology domain and, as mentioned before, it participates in the recruitment of proteins to different membranes, directing them to the appropriate cellular compartment [165]. In the five switch event examples given in this section, this is the second case of a switch that controls the ability of a protein isoform to recruit and transport proteins. Lastly, it should be mentioned that, although lacking one protein domain, the isoform dominant in testis still has a Sec7 domain (PF01369), which is a guanine-nucleotide-exchange-factor, a protein that activates monomeric GTPases [176]. This suggests that this protein isoform is functional.

Tissue 1	Tissue 2	Transcript ID 1	Transcript ID 2
testis	liver	ENST00000518315	ENST00000327040
testis	cerebral cortex	ENST00000518315	ENST00000327040
testis	ovary	ENST00000518315	ENST00000327040

Table 5.8 List of switch events common to Uhlen and GTEx datasets for PSD3 gene. "Tissue 1" and "Tissue 2" are the two switch event conditions and "Transcript ID 1" and "Transcript ID 2" are the correspondent transcript identifiers.

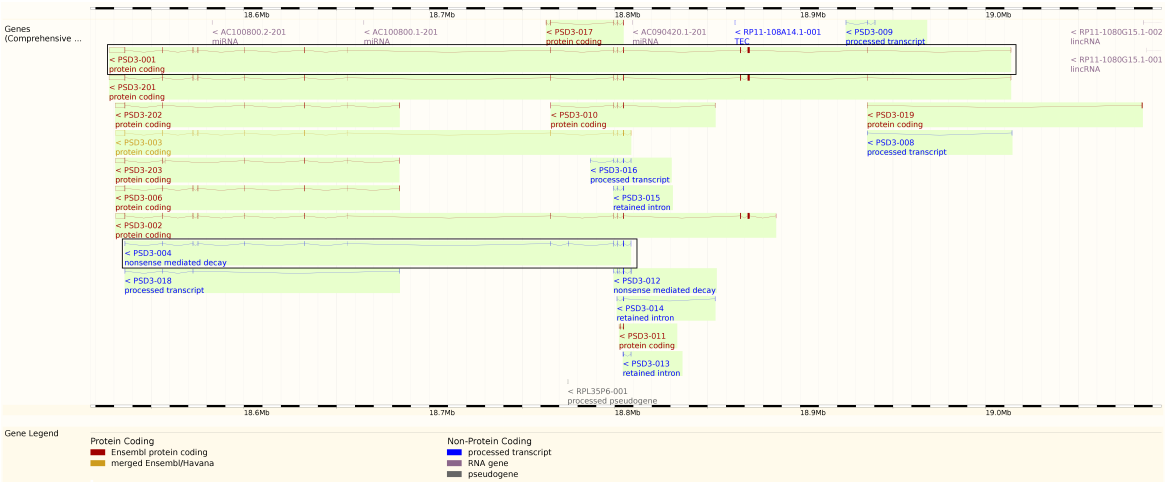


Figure 5.25 View of the transcript isoforms of PSD3 (ENSG00000156011) gene in Ensembl genome browser [19]. The two transcripts involved in switch events are outlined with black rectangles: PSD3-001 and PSD3-004 correspond to ENST00000327040 and ENST00000518315, respectively.

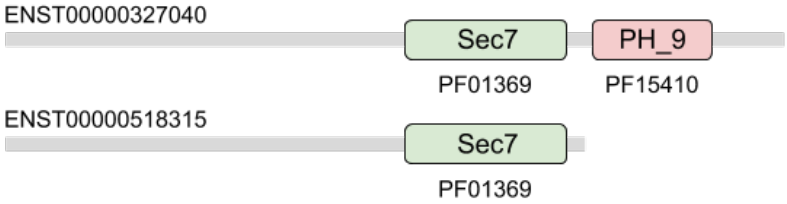


Figure 5.26 Protein domain struture for the two isoforms of PSD3 (ENSG00000156011) involved in switch events: ENST00000327040 (dominant in liver, cerebral cortex and ovary) and ENST00000518315 (dominant in testis). The Pfam domain identifiers are displayed below the domains, which are represented by differently colored rectangles. Adapted from Pfam database [26].

5.4 Discussion

The current study expands the work presented in the previous chapter by exploring transcript dominance and switch events on the GTEx dataset and by comparing the results. In this chapter, the results confirmed some of the findings from Chapter 4: most genes have a dominant transcript that is conserved across conditions and there are a few cases where the dominant transcript switches. With that said, it was noticed that using a dataset with considerably higher number of samples per tissue requires extra caution on the definition of transcript dominance, which was illustrated by the effect the criterion of support had on the number of dominant transcripts and, consequently, on the number of switch events. On one hand, the percentage of dominant transcripts in GTEx was on average lower than in

Uhlen dataset, indicating that there are fewer genes with dominant transcripts. On the other hand, the fewer cases of switch events reveal that when a transcript is dominant, tends to be dominant across all tissues where the gene is expressed.

The comparison between GTEx and Uhlen datasets showed that both datasets had a significant number switch events in common for matching tissues (Figure 5.6, Figure 5.8). However, this number was greatly decreased by the use of support to define dominant transcripts (Figure 5.7, Figure 5.9), which may be an extremely conservative definition in this case, requiring hundreds of samples to have the same major transcript in some tissues. Nevertheless, even in that case, 2-fold switch events were still observed for all comparable tissues of the 2 datasets. The same was not observed for 5-fold switch events.

The analysis of the transcripts involved in switch events confirmed what was showed in Chapter 4. The changes controlled by alternative splicing through switch events seem to be small, not disrupting protein domains in most cases. However, there were exceptions and some switch events control protein domain changes between switch transcripts. The five examples given at the end of the chapter were the only cases of 5-fold switch events common to Uhlen and GTEx datasets that involved protein domain changes. These examples showed that, although there are only a few of such cases, the genes controlled by alternative splicing in a switch-like manner can vary. This is evident by comparing the function of these genes, which is quite diverse: a kinase, a protein intracellular transporter, a transmembrane protein, a phosphotriesterase-related protein, and a pleckstrin protein. This shows that alternative splicing can drive changes through switch events in very different types of proteins, which shows that alternative splicing does not seem to have a single unique function. Also, in these 5 examples, there are cases from a single protein domain change up to 3 domain changes between the two switch isoforms, indicating that at least in some cases, alternative splicing can control drastic isoform changes. Therefore, although between 80% to 90% of switch events do not involve domain changes, there are a few cases where the changes can be significant.

Of the 5 cases of switches presented, there were 2 that had the same type of domain switching across conditions: a pleckstrin homology domain, which is involved in the recruitment of proteins to different membranes, relocating them to the appropriate cellular compartment. These examples show that switch events can potentially affect the localization of proteins in a cell.

It should also be mentioned that in the 5 examples given, there is an isoform that is dominant exclusively in one tissue and the other tissues express another isoform that is common to all of them. This indicates that switch events might appear as a consequence of the expression of an isoform specific to one tissue, but further analysis is required to shed

light on this particular question. There were tissues that were particularly more represented in these examples, these were skeletal muscle, testis and cerebral cortex. These tissues were also highlighted in the MDS analysis done in this chapter, appearing isolated from the others and participating in a significant number of switch events.

The reason why these tissues appeared represented in these examples might be related to function of the tissues and their regulation. Alternative splicing has long been known to be prevalent in testis, being critically important for several developmental pathways [177]. It is also known to be particularly widespread in the nervous system, where splicing patterns are very conserved, providing conserved functions to the tissues [178]. Lastly, in regards to skeletal muscle, it can be said that this tissue expresses a highly specialized proteome used for the metabolism of energy sources to mediate myofiber contraction, which is a result, not only of differential gene expression but also of specific alternative splicing patterns [179].

5.5 Methods

The methods used in this study were the same as the ones described in chapter 4. The cases where different or additional methods were used were described within the results section of this chapter (e.g. the comparison between Uhlen and GTEx datasets).

Chapter 6

Integrating transcriptomics and proteomics data in the study of alternative splicing

RNA-seq has allowed the investigation of the role of alternative splicing, particularly by enabling the quantification of different transcript isoforms of the same gene. Although mRNAs contain the information for the synthesis of proteins, it does not necessarily imply that functional protein products are produced. In fact, there have been some studies that explored the correlation of expression values between transcriptomics and proteomics data, revealing that the correlation is around 60% [155–157]. The work presented in this chapter is a different approach for combining transcriptomics and proteomics data to better understand alternative splicing. This integrative approach relies on perturbing the mRNA splicing patterns and checking if there are alterations on the composition of the proteome. To do so, RNA-seq data is integrated with SWATH-MS data, a DIA proteomics method, to investigate the impact of splicing events on the proteome. The strategy presented here relies on the depletion of PRPF8, a component of the core spliceosome U5 snRNP. After the perturbation, control, and PRPF8 knock-down samples were compared. Transcriptomic and proteomic changes for each detected isoform of each protein-coding gene were assessed by determining the correlation between the fold-change estimates obtained from RNA and protein data. This work also allowed to evaluate if including information on transcript relative abundances could improve the mentioned correlation.

The experimental work was done by Vi Wickramasinghe from the Hutchison Research Institute, and the proteomics data was generated and analysed by Dr. Yansheng Liu from ETH Zürich. The initial data analysis was done by Mar Gonzàlez-Porta.

A publication resulted from this work: Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R. and Wickramasinghe, V.O., 2017. Impact of alternative splicing on the human proteome. Cell reports, 20(5), pp.1229-1241.

6.1 Introduction

The contribution of alternative splicing to the diversity of proteins is not fully understood. The most common approaches for studying this subject are based on the identification of different isoforms of the same gene in a steady-state system, using MS [142, 180–182]. There have been studies that combined expression data from RNA-seq experiments with proteomics data aiming to reduce mapping noise [183, 184], however, these studies did not quantitatively analyse the impact of alternative splicing on protein diversity in a systematic manner. This study aimed at identifying changes in the composition of the proteome when selective perturbations were made to alternative splicing patterns. In this manner, it was possible to quantify the changes in a subset of transcripts affected by the induced perturbations.

RNA splicing is a complex process in which over one hundred proteins (splicing factors) participate, that together with snRNAs form the spliceosome. One of the splicing factors that participates in this process is PRPF8, a core spliceosomal component of the spliceosome [185, 186], which is part of the U5 snRNP and the B-complex, and it is one of the largest nuclear proteins. PRPF8 is highly conserved, occupying a central position in the catalytic core of the spliceosome, and taking part in crucial molecular rearrangements during splicing [185]. This protein acts as a scaffold during the assembly of the spliceosome, participating in the activation of the B-complex by recruiting the U4-U6·U5 tri-snRNP complex [187]. The method here presented analyses PRPF8 knock-down vs. control samples to understand how alternative splicing operates and how such perturbation affects the expression of alternative isoforms.

The modest correlation obtained between transcriptomics and proteomics data implies that some of the abundance variation observed at the protein level cannot be explained only by mRNA expression. The main strategies for detecting alternatively spliced isoforms at the protein level rely primarily on uniquely mapping peptides to support the identification of annotated isoforms [180, 142] or the identification of novel exon-exon junctions [188, 189]. However, understanding the role of alternative splicing is still a challenge. It is also hard to predict how differences in alternative splicing across different conditions actually affect the respective protein isoform expression levels.

The study described in this chapter is a collaborative effort aimed at understanding how differential splicing events manifest at the protein level. To do so, fold-changes obtained

using RNA-seq and SWATH-MS data were integrated using methods that rely not only on peptides that map uniquely to a specific isoform but also on peptides that map to multiple isoforms of the same gene. This is made possible by the use of RNA-seq expression data, which is used to guide the peptide assignment, enabling to increase the amount of usable proteomics data.

6.2 Results

6.2.1 Studying alternative splicing at the proteomic level

The perturbation made to splicing was the depletion of the core spliceosome U5 snRNP component PRPF8. This system had been previously validated for studying splicing at the mRNA level [190]. The induced transcriptomic and proteomic changes were subsequently assessed by RNA-seq and SWATH-MS, a data-independent acquisition (DIA) MS method which combines the coverage of conventional shotgun proteomics with the high reproducibility and accuracy of targeted protein profiling based on SRM (Figure 6.1). A total of 14,695 peptides (false discovery rate [FDR] 1%) were identified and quantified across three biological replicates for each condition, using SWATH-MS and the OpenSWATH software [191]. These peptides uniquely map to 2805 protein-coding genes and 1542 proteins of these genes had at least one peptide with altered expression after PRPF8 depletion. The reproducibility of SWATH-MS experiments was high, averaging a high correlation (R = 0.99, Pearson) between technical replicates, as well as biological replicates (R = 0.94).

The transcripts with altered splicing patterns, as well as the proteins with altered levels were found to be enriched in the same functional categories: translation, RNA splicing, mitotic cell cycle, and ubiquitination (Figure 6.2). On the other hand, the set of proteins with unchanged levels were enriched in transcription-related and ribosome biogenesis proteins [6].

6.2.2 Integrating RNA-Seq with SWATH/SRM mass spectrometry

The integration of transcriptomic and proteomic datasets was achieved by first identifying differential splicing events at the transcript level. To do so, a transcript-centric approach was used, relying on the quantification of all transcripts of a gene and identifying differences between conditions. The transcripts expression values were determined using MMSEQ [107], and cases of differential gene expression (DGE), as well as cases of differential transcript usage (DTU), were identified with MMDIFF [121]. Genes with DTU are cases with changes on the transcript relative abundances between conditions (Figure 6.1, left panel). 388 genes

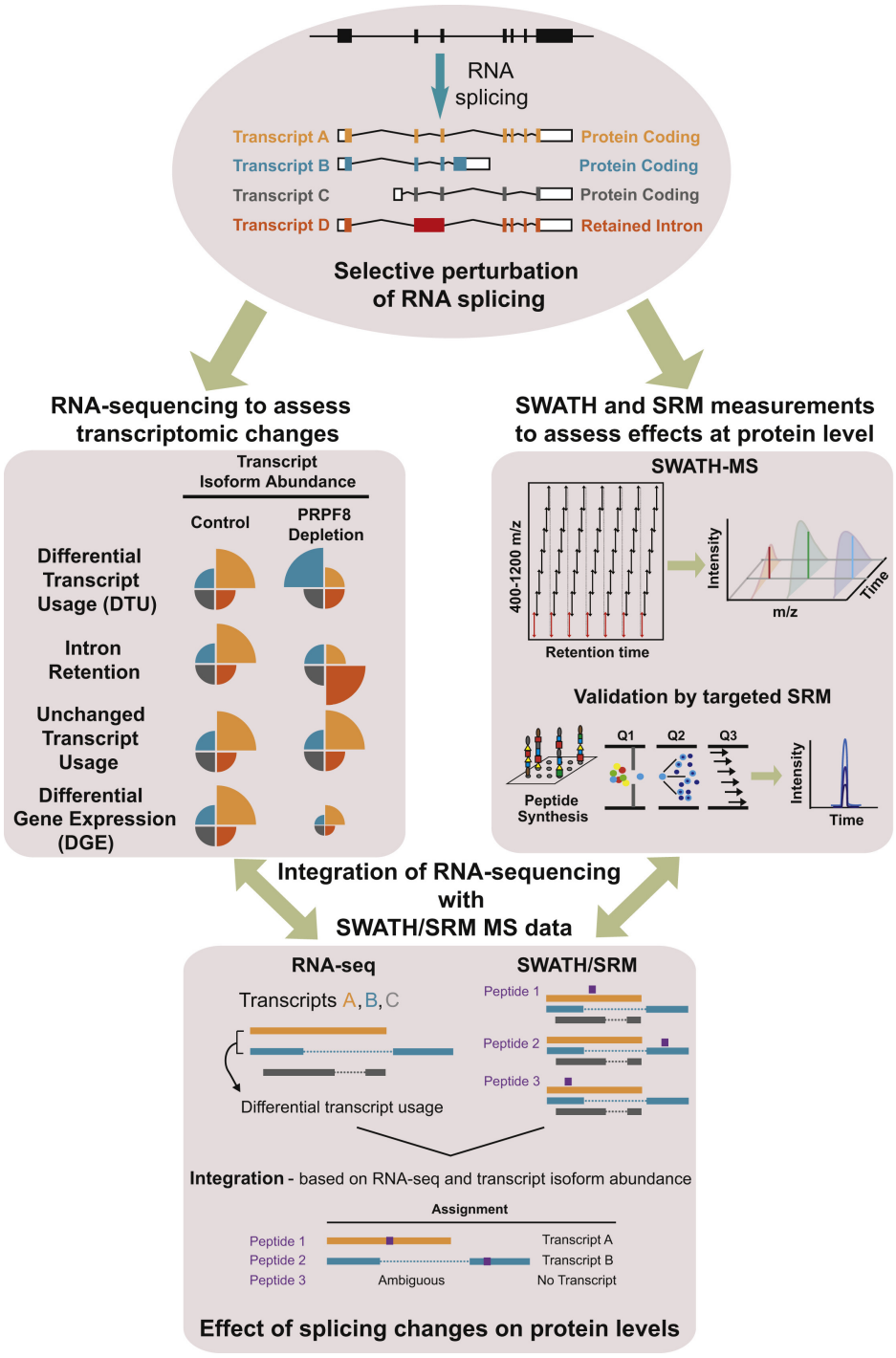


Figure 6.1 Framework to study the contribution of alternative splicing to proteomic composition and diversity. The alternative splicing process followed by a perturbation of RNA splicing is represented on the top. On the left, RNA-seq is used to assess transcriptomic changes. On the right, mass spectrometry is used to assess the effects at the protein level, first SWATH-MS was used, followed by SRM to validate the results in a targeted way. At the bottom, the data was integrated and the effects of splicing on protein levels were assessed (figure from [6]).

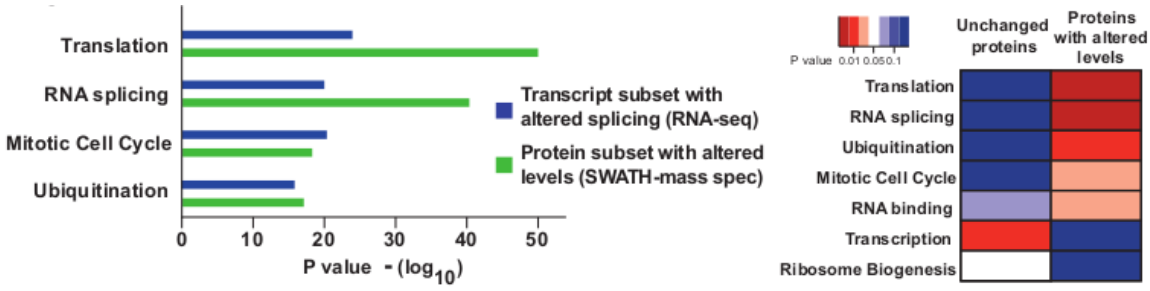


Figure 6.2 An analysis of the functional categories enriched in both transcripts with altered splicing patterns and proteins with altered levels. This analysis was done with DAVID [27]. On the left, it is shown that transcripts with altered splicing and proteins with altered levels are enriched in the same functional categories: translation, RNA splicing, mitotic cell cycle and ubiquitination. On the right, a similar analysis done with proteins detected by SWATH-MS shows that the proteins with unchanged levels after PRPF8 depletion are enriched in the categories of transcription and ribosome biogenesis. p-values are colour-coded (figure from [6]).

with DTU and 2021 genes with DGE were identified (out of 13,216 genes; expression threshold = 1 fpkm), when comparing controls to PRPF8 depleted conditions.

The first approach used was to consider only peptides that uniquely map to transcripts involved in DTU events, determined using RNA-seq data. The peptide expression levels were directly and exclusively associated with these isoforms. To determine the impact of PRPF8 depletion, the fold changes of the expression values were firstly calculated for transcripts and peptides, between controls and PRPF8 depleted samples. Then the correlation between fold changes of the two types of data was calculated, resulting in 0.49 and 0.51 for Spearman's and Pearson, respectively (65 peptides from 30 genes; p-value = 0.0169, Spearman; p-value = 0.0102, Pearson) (Figure 6.4 - A).

An alternative strategy to determine peptide fold changes for each isoform relied on determining fold changes for each peptide individually, obtaining the median fold change that mapped to each transcript. The results obtained were similar to the previous approach (Table 6.1).

It must be taken into account that uniquely mapping peptides make up a small proportion of the peptides detected by SWATH-MS (only 2974 out of a total of 14,665). Which means there are many peptides which map to multiple isoforms of the same gene.

6.2.3 Integrating the complete SWATH proteomic dataset

The strategy that was implemented to make use of the whole SWATH proteomic dataset took advantage of RNA-seq information to guide peptide assignments, particularly the most highly expressed transcript of each gene (major transcript) [1] (Figure 6.4 - B). Since lowly

DTU all transcripts + uniquely mapping peptides			
Correlation coefficient (SWATH)		Correlation coefficient (SRM)	
ρ	0.301	ρ	0.723
p-value	0.11	p-value	0.005
DTU all transcripts + all peptides			
Correlation coefficient (SWATH)		Correlation coefficient (SRM)	
ρ	0.273	ρ	0.667
p-value	0.003	p-value	0.004
DTU major transcripts + uniquely mapping peptides			
Correlation coefficient (SWATH)		Correlation coefficient (SRM)	
ρ	0.258	ρ	0.665
p-value	0.211	p-value	0.016
DTU major transcripts + all peptides			
Correlation coefficient (SWATH)		Correlation coefficient (SRM)	
ρ	0.425	ρ	0.682
p-value	0.0002	p-value	0.0047

Table 6.1 Summary of the determined correlation coefficients for cases of DTU and peptides detected in SWATH/SRM MS experiments, using an alternative method to calculate peptide fold-changes for each transcript (table from [6]).

Transcript set	Peptide set		I-over	A-over	Correlation	Agree. (%)
DTU All Transcripts and Uniquely Mapping Peptides						
transcripts (452)		transcript	30	30	ρ 0.487	Y, 21 (70)
	pep. (2974)	peptides	65	65	p-val 0.017	N, 9 (30)
genes (388)	genes (859)	genes	30	30		
DTU All Transcripts and All Peptides						
transcripts (452)		transcript	158	118	ρ 0.274	Y, 68 (57.6)
	pep. (14695)	peptides	700	530	p-val 0.0038	N, 50 (42.4)
genes (388)	genes (2805)	genes	128	116		
DTU Major Transcripts and Uniquely Mapping Peptides						
transcripts (291)		transcript	27	27	ρ 0.498	Y, 20 (74.1)
	pep. (2974)	peptides	61	61	p-val 0.017	N, 7 (25.9)
genes (263)	genes (859)	genes	27	27		
DTU Major Transcripts and All Peptides						
transcripts (291)		transcript	97	77	ρ 0.486	Y, 56 (72.7)
	pep. (14695)	peptides	481	419	p-val 1.97e−5	N, 21 (27.3)
genes (263)	genes (2805)	genes	84	75		

Table 6.2 Alternative strategies for the integration of differently used transcripts and peptides detected by SWATCH-MS. Abbreviations: 'I-over' - overlap between transcript and peptide datasets before depletion; 'A-over' - overlap between transcript and peptide datasets after depletion; 'Agree.' - percentage of agreement.; 'pep.' - peptides; 'p-val' - p-value (table from [6]).

expressed transcripts are less likely to be detected at the protein level due to the dynamic range of the mass spectrometer, only the major transcript of each gene was considered for peptide assignment, and cases that did not display DTU were discarded. 263 genes with major transcripts displaying DTU were identified (Table 6.2) and in some of these cases, the major transcript switched between conditions. In such cases, regions that allowed to distinguish these two transcripts were used to uniquely allocate peptides [6]. With this approach, it was possible to determine peptide fold changes for a total of 419 peptides which mapped to 75 genes that displayed DTU. The comparison of fold changes between mRNA and protein expression resulted in a correlation of 0.49 and 0.37 for Spearman's and Pearson, respectively (Figure 6.4 - B). It should be mentioned that the values obtained are similar to the ones obtained with uniquely mapping peptides only, even considering that was used a significantly larger dataset (419 peptides from 75 genes in contrast to 65 peptides from 30 genes). Alternatively, when both major transcripts and uniquely mapping peptides criteria were used, only 61 peptides from 27 genes met both criteria (0.50 and 0.52 for Spearman's and Pearson, respectively) (Figure 6.3).

Finally, a strategy that assigned peptides to all DTU cases regardless of their expression levels was used. This increased the dataset size to 530 peptides belonging to 116 genes but resulted in a decrease in the correlation to 0.27 and 0.21 for Spearman's and Pearson, respectively (Figure 6.4 - C).

6.2.4 Using SRM to validate SWATH-MS results

In this study, SRM was used on both control and PRPF8-depleted samples to validate the previously described findings. To increase the quantitative precision, a heavy isotope-labeled standard was used to spike the samples. Although SRM has higher sensitivity, it has the compromise of much lower analyte throughput when compared to SWATH. Therefore, it was only possible to determine peptide fold changes for 53 targeted peptides corresponding to 15 genes with major transcripts displaying DTU. The comparison of mRNA fold changes with protein expression, in this case, yielded correlations of 0.62 and 0.59 for Spearman's and Pearson (Figure 6.6 - B). When only peptides that map uniquely to transcripts involved in DTU were considered (35 peptides from 13 genes), there was a correlation increase to 0.78 and 0.71 (Figure 6.6 - A). Finally, when analysing major transcripts and uniquely mapping peptides corresponding to 33 peptides from 12 genes, the correlations were 0.73 and 0.70 for Spearman's and Pearson, respectively (Figure 6.5). These results indicate that changes in isoform usage across the human transcriptome are expressed at the proteome level.

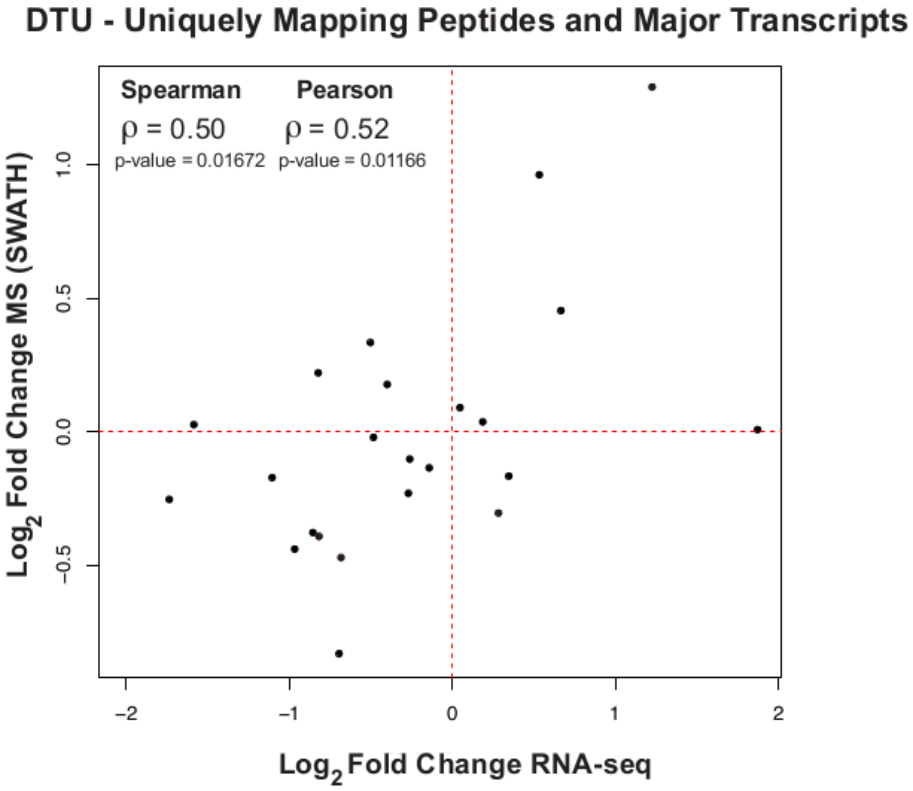


Figure 6.3 Correlation between MS and RNA-seq data fold changes for major transcripts and uniquely mapping peptides detected using SWATH-MS using the alternative approach for calculating fold changes (figure from [6]).

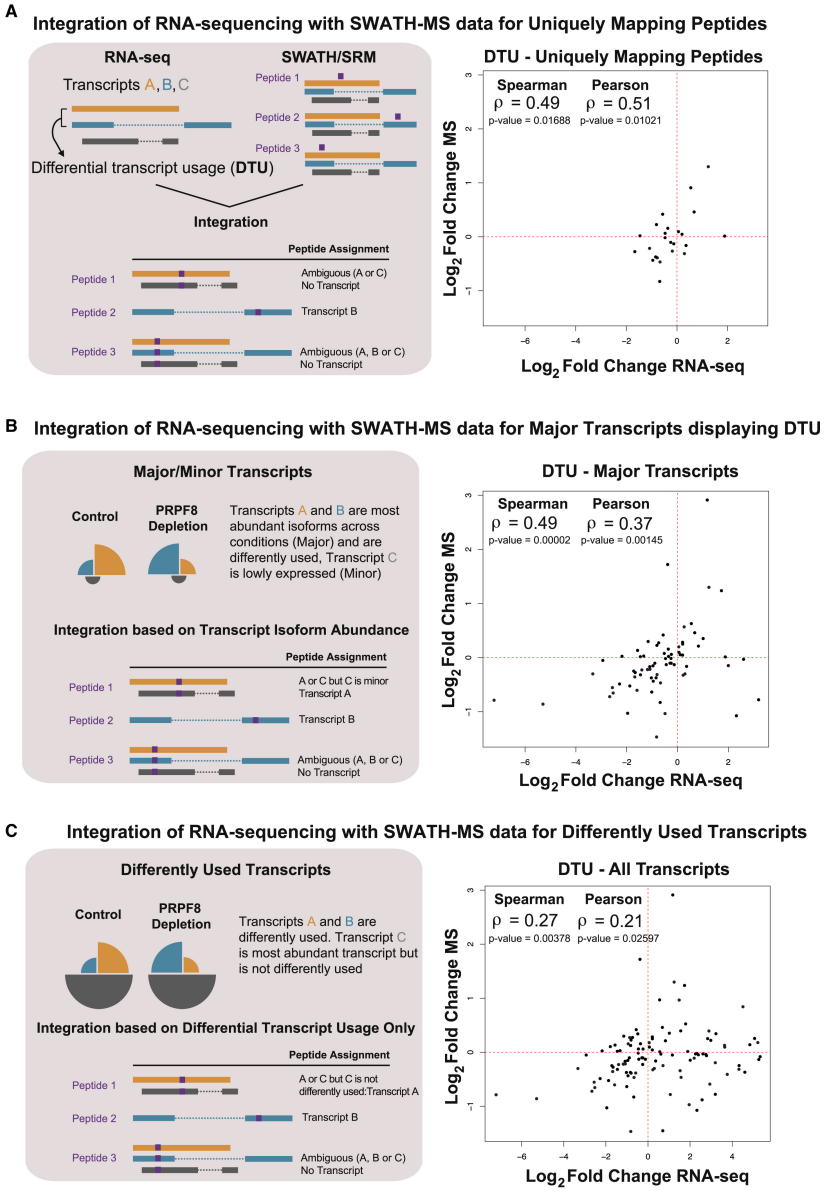


Figure 6.4 How changes in isoform usage manifest at the protein level (SWATH-MS data). A - comparison of fold changes in expression between differently used transcripts (DTU) and expression of peptides that uniquely map to them. B - comparison of fold changes in expression between DTU that are major transcripts and expression of corresponding peptides. C - comparison of fold changes in expression between all DTU transcripts and expression of corresponding peptides. Spearman's and Pearson correlation coefficients are in the top left corner of each plot (figure from [6]).

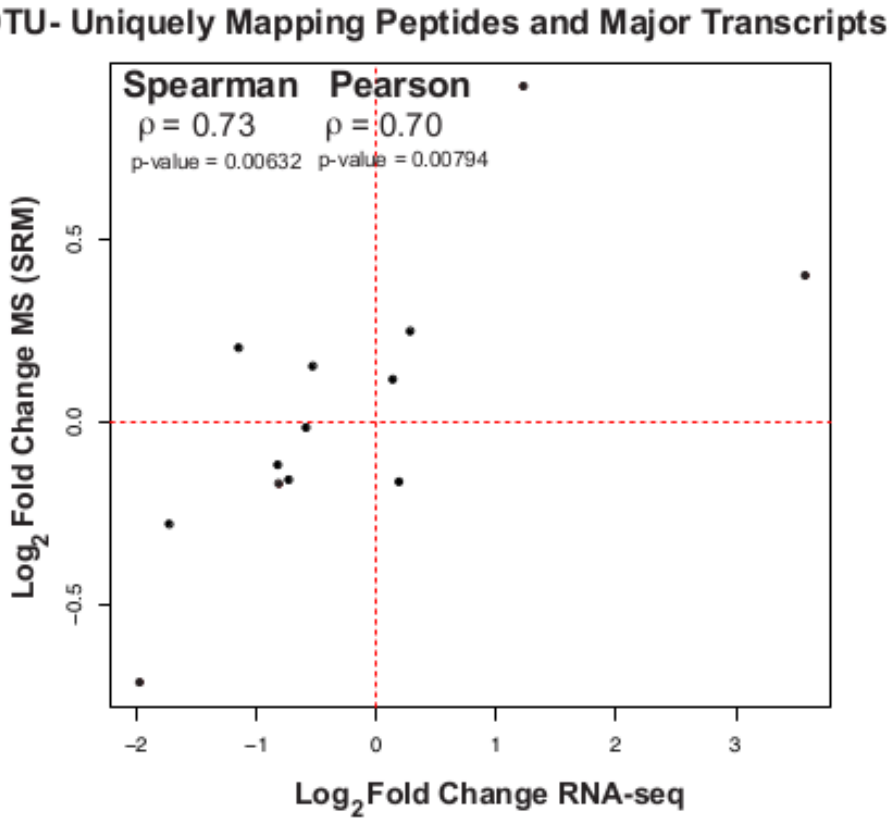


Figure 6.5 Correlation between MS and RNA-seq data fold changes for major transcripts and uniquely mapping peptides detected using SRM after PRPF8 depletion (figure from [6]).

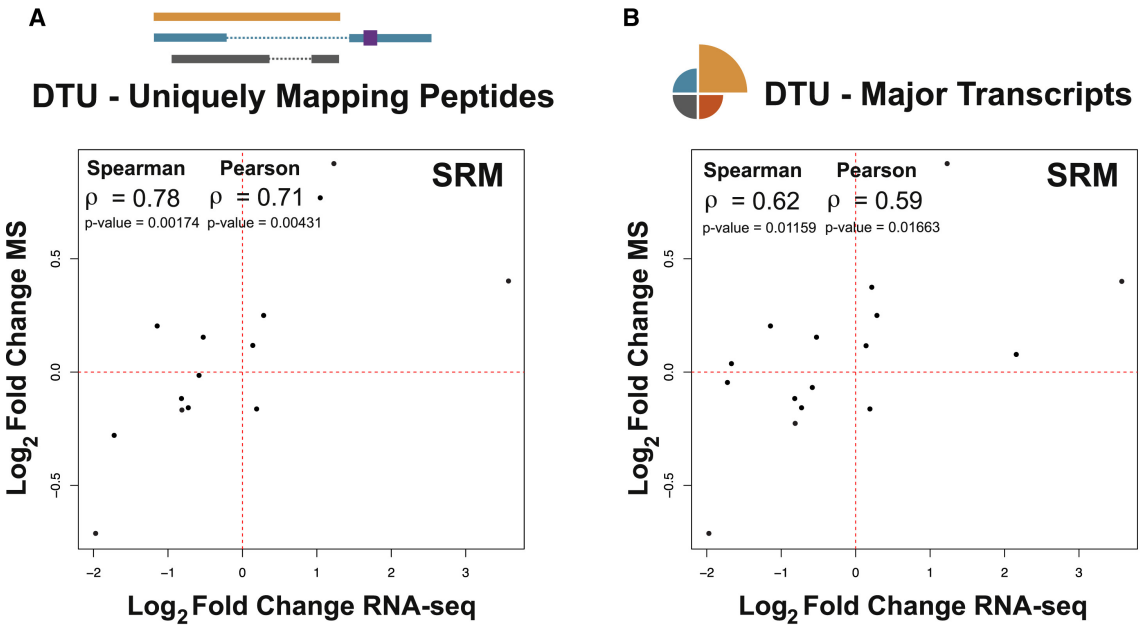


Figure 6.6 Validation of peptides using SRM-MS. A - comparison of fold changes in expression between differently used transcripts (DTU) and expression of peptides that uniquely map to them. B - comparison of fold changes in expression between major transcript that and expression of peptides that uniquely map to them (figure from [6]).

6.2.5 The effect of intron retention on protein levels

Intron retention is a type of alternative splicing which often plays a role in controlling gene expression [192, 193]. Therefore, the impact of intron retention on the proteome composition was assessed. It has been suggested that this type of alternative splicing could potentially affect transcripts from 75% of multi-exon genes [154]. Intron retention is known to cause changes in the transcripts that cause them not to be translated, specifically transcripts can be retained in the nucleus due to not being competent to be exported. They can also contain a premature stop codon, resulting in their degradation by nonsense-mediated decay (NMD), which can significantly affect transcript levels [154]. Here, the impact of retained introns on protein expression using a systematic approach was analysed.

Following PRPF8 depletion, peptide evidence of retained introns for 270 genes was found. The transcripts corresponding to these peptides were identified with DEXSeq [122] and it was observed that the expression of their corresponding proteins had decreased in comparison to the ones with no retained introns (Figure 6.8). The proportion of downregulated proteins was found to be higher in the genes with retained introns compared to the others, 161 out of 270 versus 231 out of 473, respectively (p-value = 0.0041).

Protein expression is also affected by the relative abundance of protein-coding transcript isoforms of a gene. After PRPF8 depletion, most downregulated proteins correspond to genes that have a higher relative abundance of non protein-coding transcripts (Figure 6.7 - A). Downregulated genes correspond to a relatively higher abundance of non protein-coding transcripts in comparison to upregulated proteins. It should be mentioned that after PRPF8 depletion, there were some genes with intron retention which expression levels did not change and even cases where it increased. This suggests that there might be some compensation mechanism at play. However, detecting a transcript with a retained intron does not imply that the protein levels of the same isoform are affected. In the case of upregulated proteins after PRPF8 depletion, the median of the relative abundance of protein-coding transcripts is higher than 0.9 (Figure 6.7), which means that less than 10% of the transcripts of this category of genes display intron retention, possibly explaining why the protein level is not affected. These results suggest that intron retention can affect both the human proteome and the transcriptome.

6.2.6 Alterations in transcript levels proportionally affect protein abundance

There were 2021 genes that displayed DGE after PRPF8 depletion and in the case of proteomics data, fold change information was obtained for 3057 peptides corresponding to 572 of these genes. The correlation between the two sets of data was calculated and 0.63 was obtained for Spearman's, increasing to 0.79 if only peptides with a significant fold change were considered (Figure 6.8 - B). When the same calculations were done for the cases of uniquely mapping peptides, 0.58 was obtained, and there was an increase to 0.76 when considering peptides with significant fold change (Figure 6.7 - B). In the case of genes that do not display DGE, the correlation coefficient was 0.29 (Figure 6.7 - C, D), which suggests that the changes in protein expression are being driven by changes in gene expression. Overall, the results suggest that in a alternative splicing disrupted system, a significant proportion of the variation at the protein level can be attributed to changes in mRNA levels.

6.3 Discussion

Protein abundance plays an important role in cellular function and it is closely related with transcript abundance, although this relationship is not fully understood. The abundance of mRNA and proteins are controlled by post-transcriptional and translational regulatory processes [159], which can make the understanding of the correlation between the transcriptome

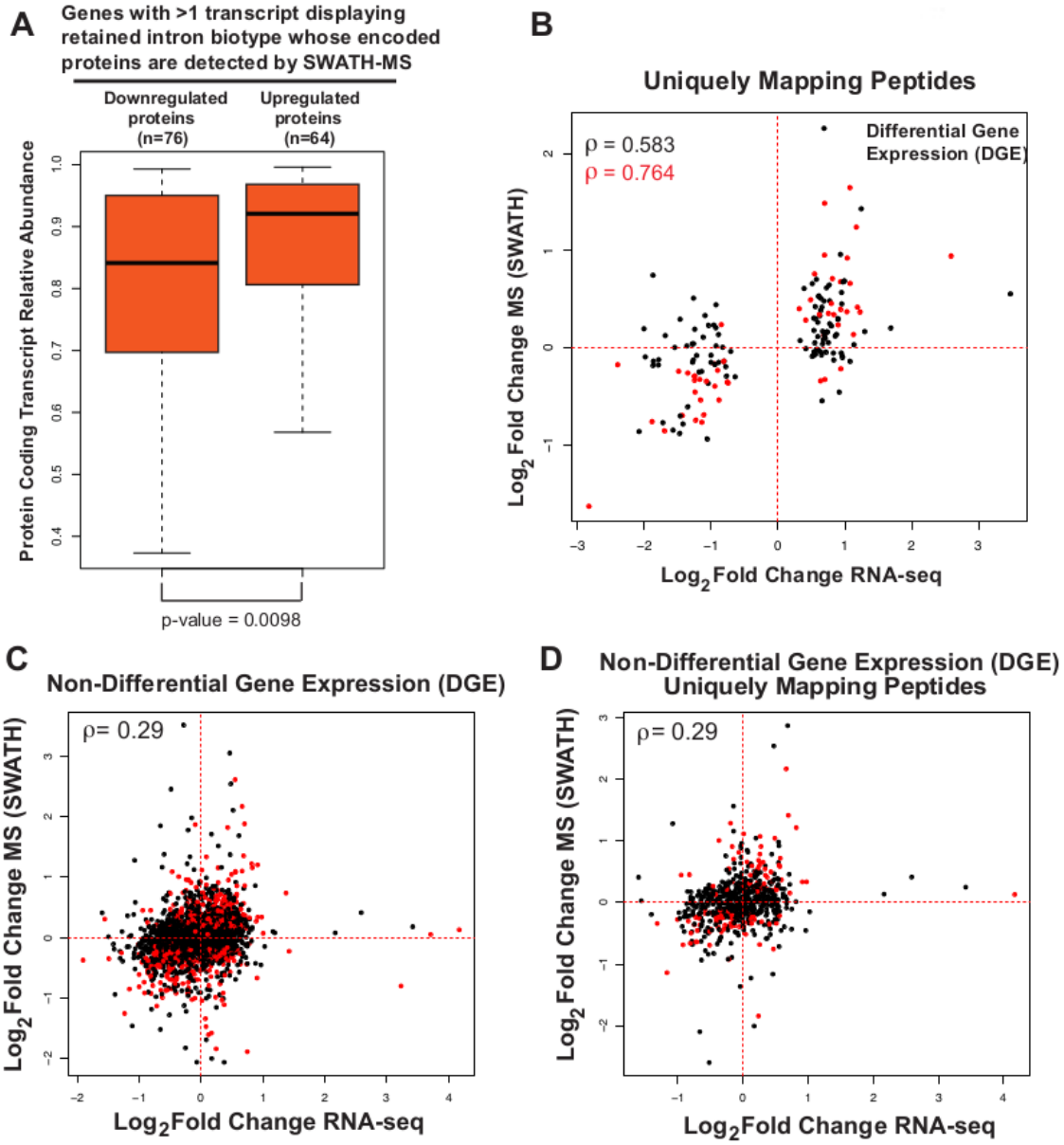


Figure 6.7 Intron Retention analysis. A - comparison of the relative abundance of protein coding transcripts between the set of downregulated and upregulated proteins for genes with more than one transcript displaying retained introns. B - scatter plot comparing expression changes in differentially expressed genes (x-axis) to expression changes in peptides that map uniquely to them (y-axis). C - scatter plot comparing expression changes of non-differentially expressed genes after PRPF8 depletion. D - scatter plot for the method using uniquely mapping peptides. In all scatter plots, the significantly differentially expressed genes are represented in red (adjusted p-value <0.1, t-test) and Spearman's correlation coefficient is shown in the top left (figure from [6]).

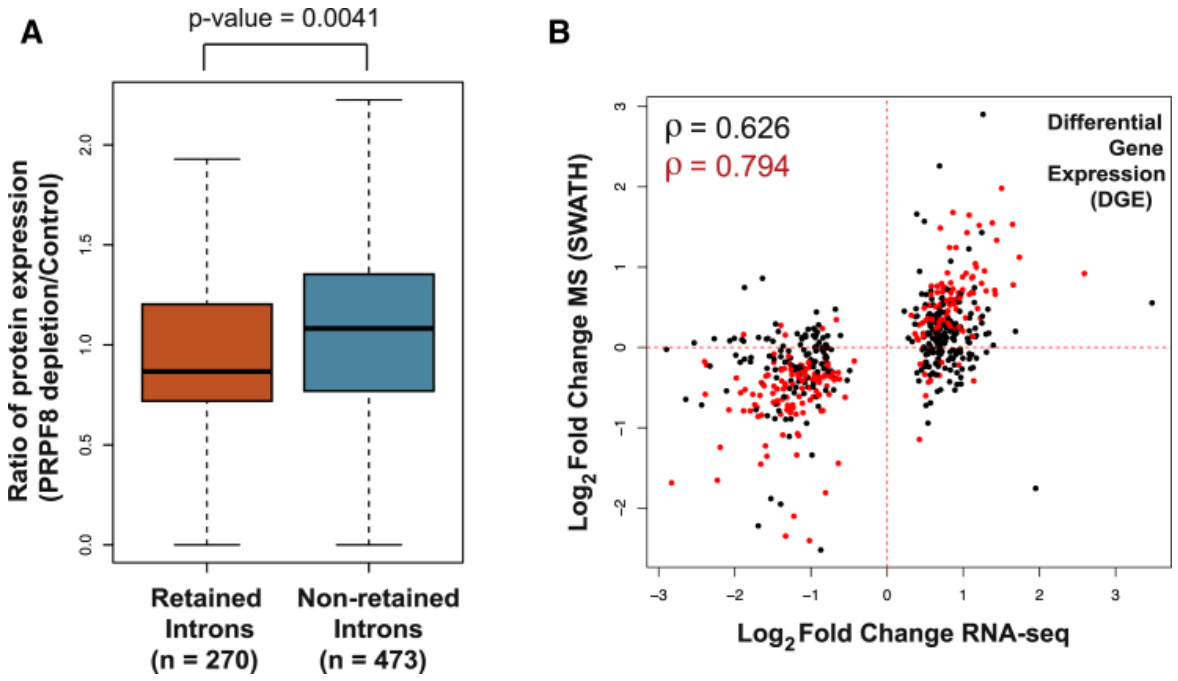


Figure 6.8 Effects of intron retention and differential gene expression (DGE) in the proteome. A - Boxplot of the ratio of protein expression between PRPF8 depletion and control conditions. The numbers on the bottom represent the number of transcripts with peptide evidence and the p-value is indicated on top (Wilcoxon test). B - Comparison between the fold changes between DGE cases and expression of peptides that map to them. In the top left corner is shown the Spearman's correlation coefficient and differentially expressed genes whose peptides change significantly in expression are indicated in red, with the respective coefficient also in red (adjusted p-value < 0.1, t-test, Holm method) (figure from [6]).

and the proteome difficult. Studies based on steady-state approaches have shown that there is a correlation between transcript and protein level. However, a substantial part of protein abundance cannot be explained solely by the transcript abundance [156, 157, 194, 195]. Other studies, based on perturbed states quantification, have shown that mRNA abundance plays a dominant role in the constitution of the proteome [196, 197].

The method described here suggests that a significant proportion of variation at the protein level can be attributed to mRNA levels. An increase in correlation coefficients between mRNA and protein levels from SWATH to SRM was observed, suggesting that a significant proportion of the protein isoform variance from the perturbed system can be explained by differences in RNA isoform usage. It also shows that there is a dependency on both the sensitivity of the MS method used and the identification of high confidence alternative splicing events at the transcript level. Integrating RNA-seq and quantitative SWATH and SRM mass spectrometry data, as well as using peptides that map to more than one transcript, provides useful information to study alternative splicing.

Transcript abundance is reliable information that can be used to assign peptides to isoforms. This indicates that the expression level of transcripts plays an important role in regulating protein abundance, which means that differential splicing events in minor transcripts will produce subtle changes that may be undetectable at the protein level. The inclusion of minor lowly expressed transcripts resulted in a sharp correlation drop, suggesting that these transcripts are increasing the noise levels at both mRNA and protein level, negatively affecting peptide assignment. This also indicates that the information from multi-mapping peptides can only be used if transcript abundance information is also taken into consideration. These results suggest that transcript expression levels have an important role in the regulation of protein isoform expression, supporting the idea that in the case of minor transcripts, differential splicing events induce subtle changes that do not drastically impact the proteome overall.

The results show there is a relationship between RNA expression and proteomic diversity. It is also shown that alternative splicing events that induce intron retention tend to lead to decreased protein abundance. On the other hand, alterations on DTU and gene expression change protein isoform abundances proportionally to transcript levels. Note that in this study the fraction of the whole proteome mass represented by the number of proteins identified is very high (>99.5%) [198], so the identified events are likely representative.

Overall, this study suggests that alternative splicing contributes to proteomic composition and diversity in humans to some degree, which is supported by a study based on ribosome occupancy as an indicator of translation output [159]. However, the extension of alternative splicing contribution to proteomic complexity is still unclear [4, 132, 199]. It is important

to notice that some of the difficulties in understanding this matter might be related to the technical aspects and limitations of both RNA-seq and MS. The level of confidence in alternative splicing events can significantly affect the results.

The methods used in this study to integrate RNA-seq and quantitative SWATH and SRM MS data show that it is possible to extract useful information from peptides that map to more than one transcript of the same gene if transcript abundance data is taken into account. This could provide the foundation for future studies on alternative splicing.

6.4 Methods

6.4.1 Analysis of RNA-Seq data

The RNA-seq data from both control siRNA-treated and PRPF8-depleted Cal51 cells was generated on an Illumina HiSeq2000 platform using 100 bp paired-end reads and the RNA was isolated from 3 and 4 independent experiments, respectively, as in a previous study from some of the same authors [190].

6.4.2 Read mapping and transcript quantification

The raw RNA-seq reads were mapped to the Ensembl v66 [134] reference transcriptome using Bowtie v0.12.7 [200]. The estimation of transcript expression levels was done using MMSEQ v1.0.7 [107] and MMDIFF [121] was used to determine differentially expressed genes and differential transcript usage. MMDIFF uses a Bayesian inference approach to determine the probability that two genes are differentially expressed or two transcripts are differentially used (posterior probability). In this case, a posterior probability of 0.85 was used as the significance threshold in the analysis of SWATH data and 0.9 in the case of SRM data.

The switch events for the set of genes that presented differential transcript usage were determined using SwitchSeq [136]. The switch events that involved transcripts that corresponded to protein isoforms with similar sequences were removed from the analyses.

6.4.3 Shotgun and SWATH-MS measurement

The peptides were measured on an AB Sciex 5600 TripleTOF mass spectrometer operated in DDA mode. The SWATH analysis was done in the same LC-MS/MS system.

6.4.4 Assignment of peptides to transcripts

The initial number of peptides detected using SWATH-MS was 16,779, which were detected across biological replicates for both control siRNA and PRPF8-depleted samples. These peptides were mapped against all the protein-coding transcripts from Ensembl v66 annotation. The peptides that mapped to more than one gene were then removed, leading to a set of 14,695 peptides, which corresponded to 2805 genes. The peptides were assigned to the transcripts as described in (Figure 6.1). Peptides that uniquely mapped to a specific transcript were directly assigned to it, which were a minority of cases: 2974 peptides mapping to 859 genes. Peptides mapping ubiquitously to more than one transcript of the same gene were assigned using information retrieved from the RNA-seq data. Two alternative strategies were used for peptide assignment. The first strategy relied on the abundance of transcript isoforms of each gene and only peptides mapping to the major transcript were taken into account. A major transcript is the most expressed transcript isoform of a gene. The major isoforms identified in control siRNA-treated or PRPF8-depleted samples were used for peptide assignment. On the other hand, the second approach did not use transcript expression levels data. In cases where a peptide mapped to multiple transcripts of the same genes, a peptide was only assigned to a particular transcript if the expression of this transcript had changed after PRPF8 depletion, regardless of its relative expression level. Finally, peptides that mapped to several differentially used transcripts were considered ambiguous and were discarded.

6.4.5 Integration of transcriptomic and proteomic data

The integration of transcriptomic and proteomic data was achieved by determining fold changes in transcript and peptide expression after PRPF8 depletion. The fold changes were obtained from RNA-seq and SWATH or SRM mass spectrometry experiments.

In the case of RNA-seq data, the fold changes were calculated from transcript expression values obtained with MMSEQ. The fold change for each transcript is the transcript expression median in PRPF8-depleted versus control siRNA-treated samples.

In the case of the proteomics data, the raw intensities for the peptides were quantile-normalized to enable comparison between samples. The median of observed intensities of each peptide was determined for both PRPF8-depleted and control siRNA-treated samples, and the fold change was calculated by dividing the two values. To calculate the peptide fold change for each transcript, first the intensities of the peptides that mapped to that transcript in each biological replicate were added up, then the median value of the summed peptide intensities of PRPF8 depletion cases was divided by the median value of the controls. This

resulted in one fold change per transcript. Both SWATH and SRM datasets were treated in the same manner.

An alternative approach to the one just described to determine peptide fold changes was also tested. For both PRPF8 depletion and controls, the fold change was determined individually for each peptide. Then, the median fold change of all peptides that mapped to each transcript was calculated (this approach yielded similar results).

The fold changes determined using transcriptomics and proteomics data were integrated as described in (Figure 6.1). The relationship between transcript and peptide fold changes was evaluated using Spearman correlation, as it has been suggested in [201]. Pearson correlation was also used for comparison.

Chapter 7

Conclusions

The understanding of alternative splicing has come a long way since was initially discovered in 1977. The identification of the components of the spliceosome [10] has enabled the characterization of the molecular mechanisms involved in splicing and many processes controlled by alternative splicing have also been identified, showing how differential isoform expression can affect cellular processes [9]. However, only high throughput techniques allow extensive analyses of the transcriptome and the proteome in a vast range of conditions, and that is what fully allows us to understand the splicing process. It allows us to determine how often alternative splicing is observed and assess the changes that occur between different conditions. The work presented in this thesis demonstrates how NGS technologies can be used in the understanding of the role of alternative splicing and its contribution to RNA and protein diversity.

All the studies here presented used RNA-seq data and showed how the quantification of transcripts and comparison across conditions can be used in different ways to understand the process of alternative splicing. These studies showed not only the potential of different approaches in this context but also some of the limitations.

The study described in Chapter 4 relied on comparing the relative expression of the transcripts of a gene across conditions to assess changes on alternative splicing. It showed that on average in a given tissue there are around 10,000 genes being expressed above 1 FPKM, which is close to half of the protein-coding genes in Ensembl human genome annotation [134]. This was observed in Uhlen dataset, as well as in GTEx dataset (Chapter 5), which is quite interesting considering that this latter dataset has a significantly higher number of samples and for a gene to be classified as expressed, it had to be expressed above 1 FPKM in all samples of a given tissue. This shows that there is consistency across samples of each tissue and that the number of protein-coding genes that are active in a specific condition is relatively small compared to the total number of genes. This result raises questions regarding

the function of genes that are not expressed or expressed at a low level, indicating that it is important to characterize and understand the conditions in which each gene is expressed.

This approach to study alternative splicing by analysing the relative expression levels of gene isoforms was used to first determine which genes have dominant isoforms and then to compare those dominant isoforms across tissues. In Chapter 4, it was shown that a large percentage of the expressed genes, a little more than two-thirds, had a dominant isoform. The idea that most genes have a single dominant transcript had been proposed before based on the evidence from transcriptomics studies which showed that a majority of genes express a main isoform at the RNA level. This is supported by what is found at the protein level, where only one isoform is usually detected for each gene. In the past, the results observed at the transcript level were obtained using relatively small datasets, which meant the generalization of these observations had to be done with extra caution. Therefore stringent criteria were used to define transcript dominance in Chapter 4 in an attempt to avoid taking incorrect conclusions regarding the number of dominant transcripts.

The transcript dominance analysis done on the GTEx study (Chapter 5) revealed that on average 59% of genes have a dominant transcript but in this case, the support criterion was not used to define transcript dominance. As discussed when the results were presented, the criterion of support might be too stringent in the case of datasets with a large number of samples per tissue, so it is acceptable to relax the dominance conditions in this situation. This shows how important is the definition of dominant transcript and how the methods must be carefully assessed because they can drastically change the results. Another example is the previous study done in the group [1], where the reported percentage of genes with dominant transcripts was 79%, illustrating again how dataset size and transcript dominance definition are linked and must be carefully evaluated. Moreover, using more stringent criteria with smaller datasets does not guarantee that the same results will be found in bigger ones, as it was shown by the differences encountered between Uhlen and GTEx results.

After obtaining the list of dominant transcripts, the switch events were determined. The number of switch events found between pairs of tissues was low, considering the total number of dominant transcripts. Not only was the number low, but also the changes observed between isoforms that switch were mostly small. The exon changes observed between switch transcripts do not translate into big differences in sequence identity between isoforms and, as a consequence, 80 to 90% of switch events do not involve protein domain changes. These results show that in most of these cases, alternative splicing controls subtle changes, which may not drastically alter protein functions. However, the changes induced by isoform switches were only inferred by comparing protein domains annotation, more extensive studies with proteomics data are needed to quantify and evaluate the effects of isoform switches.

The analysis of the impact of switch events was also done in Chapter 5 using the GTEx dataset. The number of switch events observed between pairs of tissues was again low, considering the number of expressed genes and dominant transcripts. On GTEx, the brain regions and testis were the normal tissues with the highest number of 2-fold switch events and the maximum number was 134 for testis. Again, a low number considering that 11,738 genes were found to be expressed in this tissue and 6214 of them had a 2-fold dominant transcript. This means that only around 2% of 2-fold dominant transcripts are switching in this particular case. This number really gives a different perspective on how common alternative splicing is generally thought to be and sheds light on its influence in tissue differentiation and tissue specificity. In regards to this observation, it is important to point out that alternative splicing seems to most likely have a modest role in the phenotypic variation observed between human tissues. Nevertheless, the consequences of this 2% change on dominant transcripts are not fully characterized or understood, and with the type of data used in the studies here presented, it is only possible to infer the possible consequences at the protein level and it is very difficult to extrapolate how specific cellular processes are affected.

Following the determination of switch events on GTEx dataset, the switches of the two studied datasets were compared. This comparison was made between pairs of matching tissues and, considering the number of switch events found was small, it was reassuring to find that for matching tissues, there were switch events in common between Uhlen and GTEx datasets. These are two distinct datasets with considerably different dimensions, which, as discussed before, affect transcript dominance, but still shared some of the same changes in dominant transcripts across tissues. This indicates that some switch-like splicing patterns are conserved across tissues.

To have a more detailed prespective of the type of genes switch events might affect, as well as the extent of changes controlled by this process, five examples were given at the end of Chapter 5. These examples were all the 5-fold switch events common to Uhlen and GTEx datasets whose isoforms had domain changes. The five selected genes had very distinct functions, suggesting that switch events are not exclusive to a specific type of gene or biological function. The number of domain changes encountered also varied, showing that the extent of the effect of alternative splicing through switch events might be significant, even if most of them cause subtle changes as was shown in Chapter 4 and 5.

These studies shed light on the role of alternative splicing in the cases where genes have dominant transcript isoforms, however, it is important to notice that around one-third of genes expressed in a given tissue does not have a dominant transcript. This means these genes express multiple isoforms at approximately the same level. Although these particular cases were not explored, alternative splicing might be enabling the simultaneous expression

of distinct isoforms with significantly different functions, which would mean that alternative splicing would have different functions in these situations compared to the switch event cases. However, this is of course speculation and the analyses presented in this thesis did not explore genes with no dominant transcript. Further work will be needed to study these cases.

Another question that can be raised is one regarding all the annotated isoforms that are expressed at low levels. Although these isoforms might have a minor contribution to the transcriptome, it is important to investigate if these isoforms are in fact artifacts or to identify the conditions in which they are expressed. This information can improve the annotation by helping to identify and delete redundant transcript isoforms and by providing additional knowledge that better characterizes each isoform. However, validating very lowly expressed isoforms might face technical limitations. Current RNA-seq technology might not have enough sensitivity to detect lowly expressed isoforms and, although MS proteomics could be used to identify the correspondent protein isoform, these methods can also have trouble identifying lowly expressed proteins.

The studies here presented tried to avoid sensitivity issue by focusing on highly expressed genes (>=1fpkm). However, there were still technical limitations that need to be considered. In terms of transcriptome characterization, the length of RNA-seq reads is still a limitation that can affect the identification and quantification of transcript isoforms. It can lead the mapping and quantification software to have troubles assigning reads to the correct transcripts, as it was illustrated by the analysis of the gene CUX1 in a case study in Chapter 4. In this particular case, Cufflinks and Kallisto quantified CUX1 isoforms quite differently, showing that the reads corresponding to this gene were assigned in a different manner. Although this limitation currently affects the quantification of isoforms of some genes, it will be overcome to a great extent over time with the increasing size of reads. Much longer reads are currently being generated in alternative platforms to Illumina, such as Pacific BioSciences [202] and Nanopore sequencing [203], and they will have a great impact on the quantification of RNA isoforms.

As shown in Chapter 6, alternative splicing can also be studied by characterizing the proteome but this approach also has some challenges. MS techniques are not able to identify the full set of proteins in a given condition. This can especially limit the ability to characterize alternative splicing events because isoforms of the same gene can be highly similar, and be distinguishable by one peptide only, which might not be identified in the experiments. The inability to identify more proteins is in part a consequence of the limited number of peptides that can be used to uniquely identify proteins of interest. Therefore, an increase in the coverage of MS techniques would also enable a better characterization of splicing events.

Besides the independent advances in RNA-seq and MS technologies, data from these distinct fields can be integrated in order to improve the identification of splicing events. In Chapter 6, an approach taking advantage of both technologies was presented. In this specific case, RNA-seq data was used to increase the number of peptides that could be used in the identification of proteins. In particular, major transcripts determined using RNA-seq data were used to assign peptides to the major isoform. This was an effective strategy, which allowed to infer that changes in splicing at the transcript level can have an equivalent effect at the protein level. It is possible that expanding this strategy to use abundances of all the expressed transcripts might allow the investigation of the coding potential of minor transcripts. This again can only be clarified in future studies. Another important aspect of the study presented in Chapter 6 was to show that studying perturbed systems can facilitate the understanding of certain processes that might not be as clear in steady-state systems.

Regarding the identification of isoforms of a gene, one of the most important pieces of data is the annotation, which, not only significantly affects this process but also has effects that propagate down the pipeline, affecting the quantification and consequently the conclusions taken from the study. Therefore, the quality of the annotation is paramount. In the case of the human genome, the quality of the annotation is high and keeps improving, being frequently updated. However, there are still cases of incomplete descriptions of gene or isoform functions, as well as unclear features such as exons with the same coordinates but different identifiers or exons with very high sequence overlap. It is important to be aware of these cases, particularly when comparing isoforms for the study of alternative splicing.

It is not clear what is the contribution of each annotated isoform to the transcriptome diversity in a given condition or in which conditions each isoform is actually expressed. The studies here presented, showed the genes and isoforms expressed in normal tissues, but there is still a large number of isoforms that are in the annotation and are not expressed in these cases. It is important to have a clear understanding of which conditions lead to the expression of these isoforms and studies must be designed to test if they are actually expressed. This will facilitate the comprehension of the annotation and make it more complete. Since the RNA-seq quantification methods rely to a very significant extent on the annotation, improving its quality will yield more accurate isoform quantifications, facilitating the assignment of reads to isoforms.

The annotation is of great importance in studies using NGS technologies but other factors will have a great impact on the potential applications of these types of data. One very important factor is the cost of RNA-seq, which keeps decreasing. This will allow the generation of larger datasets, with more replicates, increasing the robustness of this kind of study and making it applicable to a wider range of scenarios. It will also have an effect on

the applicability of this technology that will certainly increase in basic and applied research, but also in a clinical context. Therefore, methods such as the ones developed in these studies will have increased importance and very direct application in the future.

The tools developed in the studies here presented were applied to the basic understanding of switch events between normal human tissues. Although they were developed in this context, they can be used for the comparison of any other conditions (e.g. normal vs. disease) and help find answers to other problems.

In the future, the methods used in this thesis can be applied to analysing changes in dominant transcripts across species, which might be of interest because it can reveal how conserved is transcript dominance. Another possibility is the application of such methods to single-cell data, which is currently growing at a very fast pace and gives an extra level of resolution. It can also be interesting to analyse the transcripts of genes with no dominant isoform, understand which ones are protein-coding and explore which isoforms are expressed at the protein level. Lastly, MS and RNA-seq coupled datasets will be extremely useful in the study of alternative splicing and could possibly help to better understand the modest correlation between RNA and protein data observed in some studies [155–157].

There are currently a variety of tools to process NGS data that can be applied to solve different biological problems. Many of the tools are implemented in Python and range from big well established packages, such as biopython, to small individual projects available on GitHub. There are also a considerable number of libraries available in R, particularly in Bioconductor [204], such as DESeq that was used in the studies presented in this thesis. There is always a need for development of new tools and methods and the easy access to them greatly improves the advances of the field, alloing better software to be developed and facilitating the integration of existent tools.

Besides the development of new methods, the generation and availability of large datasets will have a significant impact on the fields of transcriptomics and proteomics. The increase in the number and size of datasets will also allow a broader application of artificial intelligence (AI) methods on NGS data. The advances in sequencing technologies combined with great data availability and AI will have a tremendous impact in the field of bioinformatics in the future. Although it is difficult to predict all the applications of AI in bioinformatics, as well as what will be discovered with its applications, it will definitely allow us to investigate a broad range of problems at multiple resolutions, integrate different types of data and allow us to better understand the very high level of complexity found in biological systems.

References

[1] M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, and A. Brazma, "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene," *Genome biology*, vol. 14, no. 7, p. R70, 2013.

[2] J. M. Rodriguez, A. Carro, A. Valencia, and M. L. Tress, "APPRIS webserver and webservices," *Nucleic acids research*, vol. 43, no. W1, pp. W455–W459, 2015.

[3] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "The genotype-tissue expression (gtex) project," *Nature genetics*, vol. 45, no. 6, p. 580, 2013.

[4] M. L. Tress, F. Abascal, and A. Valencia, "Alternative splicing may not be the key to proteome complexity," *Trends in biochemical sciences*, vol. 42, no. 2, pp. 98–110, 2017.

[5] A. Reyes and W. Huber, "Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues," *Nucleic acids research*, vol. 46, no. 2, pp. 582–592, 2017.

[6] Y. Liu, M. Gonzàlez-Porta, S. Santos, A. Brazma, J. C. Marioni, R. Aebersold, A. R. Venkitaraman, and V. O. Wickramasinghe, "Impact of alternative splicing on the human proteome," *Cell reports*, vol. 20, no. 5, pp. 1229–1241, 2017.

[7] L. Pray, "Gene expression." https://www.nature.com/scitable/topicpage/gene-expression-14121669, 2010. Visited on 2018-07-03.

[8] S. Clancy and W. Brown, "Translation: DNA to mRNA to Protein," *Nature Education*, vol. 1, no. 1, p. 101, 2008.

[9] H. Keren, G. Lev-Maor, and G. Ast, "Alternative splicing and evolution: diversification, exon definition and function," *Nature Reviews Genetics*, vol. 11, no. 5, p. 345, 2010.

[10] C. L. Will and R. Lührmann, "Spliceosome structure and function," *Cold Spring Harbor perspectives in biology*, p. a003707, 2010.

[11] E. R. Mardis, "Next-generation sequencing platforms," *Annual review of analytical chemistry*, vol. 6, pp. 287–303, 2013.

[12] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[13] S. Anders, "Gene expression." http://htseq.readthedocs.io/en/master/count.html, 2010. Visited on 2018-07-18.

[14] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms," *Nature biotechnology*, vol. 28, no. 5, p. 511, 2010.

[15] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature biotechnology*, vol. 34, no. 5, p. 525, 2016.

[16] F. Xie, T. Liu, W.-J. Qian, V. A. Petyuk, and R. D. Smith, "Liquid chromatography-mass spectrometry-based quantitative proteomics," *Journal of Biological Chemistry*, pp. jbc–R110, 2011.

[17] P. Picotti and R. Aebersold, "Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions," *Nature methods*, vol. 9, no. 6, p. 555, 2012.

[18] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC bioinformatics*, vol. 10, no. 1, p. 421, 2009.

[19] A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, *et al.*, "Ensembl 2016," *Nucleic acids research*, vol. 44, no. D1, pp. D710–D716, 2015.

[20] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "Dataset summary of analysis samples of gtex." https://gtexportal.org/home/tissueSummaryPage, 2017. Visited on 2019-01-06.

[21] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "Protein: Kif1b human (o60333)." http://pfam.xfam.org/protein/O60333, 2019. Visited on 2019-01-25.

[22] "Protein data bank: the single global archive for 3d macromolecular structure data," *Nucleic acids research*, vol. 47, no. D1, pp. D520–D528, 2018.

[23] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "Protein: S35e4 human (q6icl7)." http://pfam.xfam.org/protein/Q6ICL7, 2019. Visited on 2019-01-25.

[24] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "Protein: Mark4 human (q96l34)." http://pfam.xfam.org/protein/Q96L34, 2019. Visited on 2019-01-25.

[25] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "Protein: Pter human (q96bw5)." http://pfam.xfam.org/protein/Q96BW5, 2019. Visited on 2019-01-25.

[26] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "Protein: Psd3 human (q9nyi0)." http://pfam.xfam.org/protein/Q9NYI0, 2019. Visited on 2019-01-25.

[27] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2008.

[28] L. Pray, "Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100." https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397, 2008. Visited on 2018-07-03.

[29] E. S Lander, C. Chen, L. Linton, B. Birren, C. Nusbaum, M. C Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gaige, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, and L. Rowen, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, p. 860, 2001.

[30] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[31] Z. Abdellah, A. Ahmadi, S. Ahmed, M. Aimable, R. Ainscough, J. Almeida, C. Almond, A. Ambler, K. Ambrose, K. Ambrose, R. Andrew, D. Andrews, N. Andrews, T. Andrews, E. Apweiler, H. Arbery, B. Archer, G. Ash, K. Ashcroft, and L. Zembeck, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, p. 931, 2004.

[32] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. Davis, F. Doyle, C. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. Lajoie, S. Landt, b.-k. Lee, F. Pauli Behn, K. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, and E. Birney, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, p. 57, 2012.

[33] F. H. Crick, "On protein synthesis," in *Symp Soc Exp Biol*, vol. 12, p. 8, 1958.

[34] M. Cobb, "60 years ago, Francis Crick changed the logic of biology," *PLoS biology*, vol. 15, no. 9, p. e2003243, 2017.

[35] J. A. Shapiro, *Evolution: a view from the 21st century*. Pearson education, 2011.

[36] A. Kanhere and M. Bansal, "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes," *Nucleic Acids Research*, vol. 33, no. 10, pp. 3165–3175, 2005.

[37] S. Clancy, "DNA transcription," *Nature Education*, vol. 1, no. 1, p. 41, 2008.

[38] E. Rosonina, S. Kaneko, and J. L. Manley, "Terminating the transcript: breaking up is hard to do," *Genes & development*, vol. 20, no. 9, pp. 1050–1056, 2006.

[39] C. N. Cole and J. J. Scarcelli, "Transport of messenger RNA from the nucleus to the cytoplasm," *Current opinion in cell biology*, vol. 18, no. 3, pp. 299–306, 2006.

[40] A. Moghal, K. Mohler, and M. Ibba, "Mistranslation of the genetic code," *FEBS letters*, vol. 588, no. 23, pp. 4305–4310, 2014.

[41] L. Pray, "Eukaryotic genome complexity," *Nature Education*, vol. 1, no. 1, p. 96, 2008.

[42] G. Edwalds-Gilbert, "Regulation of mRNA Splicing by Signal Transduction," *Nature Education*, vol. 3, no. 9, p. 43, 2010.

[43] S. M. Rueter, T. R. Dawson, and R. B. Emeson, "Regulation of alternative splicing by RNA editing," *Nature*, vol. 399, no. 6731, p. 75, 1999.

[44] R. Iwasaki, H. Kiuchi, M. Ihara, T. Mori, M. Kawakami, and H. Ueda, "Trans-splicing as a novel method to rapidly produce antibody fusion proteins," *Biochemical and biophysical research communications*, vol. 384, no. 3, pp. 316–321, 2009.

[45] G. Parra, A. Reymond, N. Dabbouseh, E. T. Dermitzakis, R. Castelo, T. M. Thomson, S. E. Antonarakis, and R. Guigó, "Tandem chimerism as a means to increase protein complexity in the human genome," *Genome research*, vol. 16, no. 1, pp. 37–44, 2006.

[46] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA," *Cell*, vol. 12, no. 1, pp. 1–8, 1977.

[47] S. M. Berget, C. Moore, and P. A. Sharp, "Spliced segments at the 5′ terminus of adenovirus 2 late mRNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 8, pp. 3171–3175, 1977.

[48] C. P. Kala and B. S. Sajwan, "RNA interference–gene silencing by double-stranded RNA: The 2006 Nobel Prize for Physiology or Medicine," *Current Science*, vol. 91, no. 11, p. 1443, 2006.

[49] M. Chen and J. L. Manley, "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches," *Nature reviews Molecular cell biology*, vol. 10, no. 11, p. 741, 2009.

[50] J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker, "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays," *Science*, vol. 302, no. 5653, pp. 2141–2144, 2003.

[51] G. S. Wang and T. A. Cooper, "Splicing in disease: disruption of the splicing code and the decoding machinery," *Nature Reviews Genetics*, vol. 8, no. 10, p. nrg2164, 2007.

[52] J. M. Mudge, A. Frankish, J. Fernandez-Banet, T. Alioto, T. Derrien, C. Howald, A. Reymond, R. Guigó, T. Hubbard, and J. Harrow, "The origins, evolution, and functional potential of alternative splicing in vertebrates," *Molecular biology and evolution*, vol. 28, no. 10, pp. 2949–2959, 2011.

[53] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, *et al.*, "The Ensembl gene annotation system," *Database*, vol. 2016, 2016.

[54] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, p. 57, 2009.

[55] S. Haider and R. Pal, "Integrated analysis of transcriptomic and proteomic data," *Current genomics*, vol. 14, no. 2, pp. 91–110, 2013.

[56] F. Abascal, I. Ezkurdia, J. Rodriguez-Rivas, J. M. Rodriguez, A. del Pozo, J. Vázquez, A. Valencia, and M. L. Tress, "Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level," *PLoS computational biology*, vol. 11, no. 6, p. e1004325, 2015.

[57] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, p. 470, 2008.

[58] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. Thanaraj, and H. Soreq, "Function of alternative splicing," *Gene*, vol. 344, pp. 1–20, 2005.

[59] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, *et al.*, "The evolution of gene expression levels in mammalian organs," *Nature*, vol. 478, no. 7369, p. 343, 2011.

[60] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, *et al.*, "The human transcriptome across tissues and individuals," *Science*, vol. 348, no. 6235, pp. 660–665, 2015.

[61] J. Merkin, C. Russell, P. Chen, and C. B. Burge, "Evolutionary dynamics of gene and isoform regulation in Mammalian tissues," *Science*, vol. 338, no. 6114, pp. 1593–1599, 2012.

[62] A. Reyes, S. Anders, R. J. Weatheritt, T. J. Gibson, L. M. Steinmetz, and W. Huber, "Drift and conservation of differential exon usage across tissues in primate species," *Proceedings of the National Academy of Sciences of the United States of America*, p. 201307202, 2013.

[63] O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm, "Function of alternative splicing," *Gene*, vol. 514, no. 1, pp. 1–30, 2013.

[64] M. Long, E. Betrán, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, p. 865, 2003.

[65] E. Kim, A. Magen, and G. Ast, "Different levels of alternative splicing among eukaryotes," *Nucleic acids research*, vol. 35, no. 1, pp. 125–131, 2006.

[66] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Çolak, *et al.*, "The evolutionary landscape of alternative splicing in vertebrate species," *Science*, vol. 338, no. 6114, pp. 1587–1593, 2012.

[67] A. Busch and K. J. Hertel, "Evolution of SR protein and hnRNP splicing regulatory factors," *Wiley Interdisciplinary Reviews: RNA*, vol. 3, no. 1, pp. 1–12, 2012.

[68] J. Wang, P. J. Smith, A. R. Krainer, and M. Q. Zhang, "Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes," *Nucleic acids research*, vol. 33, no. 16, pp. 5053–5062, 2005.

[69] P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, *et al.*, "Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 22, pp. 8390–8395, 2006.

[70] R. F. Luco, M. Allo, I. E. Schor, A. R. Kornblihtt, and T. Misteli, "Epigenetics in alternative pre-mRNA splicing," *Cell*, vol. 144, no. 1, pp. 16–26, 2011.

[71] S. Naftelberg, I. E. Schor, G. Ast, and A. R. Kornblihtt, "Regulation of alternative splicing through coupling with transcription and chromatin structure," *Annual review of biochemistry*, vol. 84, pp. 165–198, 2015.

[72] A. Ameur, A. Zaghlool, J. Halvardson, A. Wetterbom, U. Gyllensten, L. Cavelier, and L. Feuk, "Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain," *Nature Structural and Molecular Biology*, vol. 18, no. 12, p. 1435, 2011.

[73] A. J. Matlin, F. Clark, and C. W. Smith, "Understanding alternative splicing: towards a cellular code," *Nature reviews Molecular cell biology*, vol. 6, no. 5, p. 386, 2005.

[74] Z. Wang, X. Xiao, E. Van Nostrand, and C. B. Burge, "General and specific functions of exonic splicing silencers in splicing control," *Molecular cell*, vol. 23, no. 1, pp. 61–70, 2006.

[75] P. Kafasla, I. Mickleburgh, M. Llorian, M. Coelho, C. Gooding, D. Cherny, A. Joshi, O. Kotik-Kogan, S. Curry, I. C. Eperon, *et al.*, "Defining the roles and interactions of PTB," *Biochemical Society Transactions*, 2012.

[76] N. Jelen, J. Ule, M. Živin, and R. B. Darnell, "Evolution of Nova-dependent splicing regulation in the brain," *PLoS genetics*, vol. 3, no. 10, p. e173, 2007.

[77] J.-A. Lee, Z.-Z. Tang, and D. L. Black, "An inducible change in Fox-1/A2BP1 splicing modulates the alternative splicing of downstream neuronal target exons," *Genes & development*, 2009.

[78] J. Ule, G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B. J. Blencowe, and R. B. Darnell, "An RNA map predicting Nova-dependent splicing regulation," *Nature*, vol. 444, no. 7119, p. 580, 2006.

[79] M. K. Sakharkar, V. T. Chow, and P. Kangueane, "Distributions of exons and introns in the human genome," *In silico biology*, vol. 4, no. 4, pp. 387–393, 2004.

[80] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Human molecular genetics*, vol. 15, no. suppl_1, pp. R17–R29, 2006.

[81] S. R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nature Reviews Genetics*, vol. 2, no. 12, p. 919, 2001.

[82] R. W. Carthew and E. J. Sontheimer, "Origins and mechanisms of miRNAs and siRNAs," *Cell*, vol. 136, no. 4, pp. 642–655, 2009.

[83] R. J. Taft, E. A. Glazov, T. Lassmann, Y. Hayashizaki, P. Carninci, and J. S. Mattick, "Small RNAs derived from snoRNAs," *Rna*, 2009.

[84] R. Li, H. Zhu, and Y. Luo, "Understanding the functions of long non-coding RNAs through their higher-order structures," *International journal of molecular sciences*, vol. 17, no. 5, p. 702, 2016.

[85] B. J. Blencowe, "Alternative splicing: new insights from global analyses," *Cell*, vol. 126, no. 1, pp. 37–47, 2006.

[86] Q. Xu, B. Modrek, and C. Lee, "Genome-wide detection of tissue-specific alternative splicing in the human transcriptome," *Nucleic acids research*, vol. 30, no. 17, pp. 3754–3766, 2002.

[87] R. Hrdlickova, M. Toloue, and B. Tian, "RNA-Seq methods for transcriptome analysis," *Wiley Interdisciplinary Reviews: RNA*, vol. 8, no. 1, p. e1364, 2017.

[88] M. Hu and K. Polyak, "Serial analysis of gene expression," *Nature protocols*, vol. 1, no. 4, p. 1743, 2006.

[89] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, *et al.*, "CAGE: cap analysis of gene expression," *Nature methods*, vol. 3, no. 3, p. 211, 2006.

[90] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, *et al.*, "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays," *Nature biotechnology*, vol. 18, no. 6, p. 630, 2000.

[91] C. W. Sugnet, K. Srinivasan, T. A. Clark, G. O'Brien, M. S. Cline, H. Wang, A. Williams, D. Kulp, J. E. Blume, D. Haussler, *et al.*, "Unusual intron conservation near tissue-regulated exons found by splicing microarrays," *PLoS computational biology*, vol. 2, no. 1, p. e4, 2006.

[92] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, *et al.*, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008.

[93] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome biology*, vol. 17, no. 1, p. 13, 2016.

[94] E. L. van Dijk, Y. Jaszczyszyn, and C. Thermes, "Library preparation methods for next-generation sequencing: tone down the bias," *Experimental cell research*, vol. 322, no. 1, pp. 12–20, 2014.

[95] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic acids research*, vol. 38, no. 12, pp. e131–e131, 2010.

[96] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic acids research*, vol. 40, no. 10, pp. e72–e72, 2012.

[97] M. L. Metzker, "Sequencing technologies-the next generation," *Nature reviews genetics*, vol. 11, no. 1, p. 31, 2010.

[98] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome biology*, vol. 14, no. 4, p. R36, 2013.

[99] A. E. Minoche, J. C. Dohm, and H. Himmelbauer, "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems," *Genome biology*, vol. 12, no. 11, p. R112, 2011.

[100] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[101] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature methods*, vol. 9, no. 4, p. 357, 2012.

[102] S. Anders, P. T. Pyl, and W. Huber, "HTSeq - a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[103] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[104] A. Roberts and L. Pachter, "Streaming fragment assignment for real-time analysis of sequencing experiments," *Nature methods*, vol. 10, no. 1, p. 71, 2013.

[105] N. Nariai, K. Kojima, T. Mimori, Y. Sato, Y. Kawai, Y. Yamaguchi-Kabata, and M. Nagasaki, "TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads," *BMC genomics*, vol. 15, no. 10, p. S5, 2014.

[106] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature methods*, vol. 14, no. 4, p. 417, 2017.

[107] E. Turro, S.-Y. Su, Â. Gonçalves, L. J. Coin, S. Richardson, and A. Lewin, "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads," *Genome biology*, vol. 12, no. 2, p. R13, 2011.

[108] J. Li, H. Jiang, and W. H. Wong, "Modeling non-uniformity in short-read rates in RNA-Seq data," *Genome biology*, vol. 11, no. 5, p. R50, 2010.

[109] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter, "Improving RNA-Seq expression estimates by correcting for fragment bias," *Genome biology*, vol. 12, no. 3, p. R22, 2011.

[110] R. Patro, S. M. Mount, and C. Kingsford, "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms," *Nature biotechnology*, vol. 32, no. 5, p. 462, 2014.

[111] A. Srivastava, H. Sarkar, N. Gupta, and R. Patro, "RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes," *Bioinformatics*, vol. 32, no. 12, pp. i192–i200, 2016.

[112] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature methods*, vol. 5, no. 7, p. 621, 2008.

[113] C. Evans, J. Hardin, and D. M. Stoebel, "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions," *Briefings in bioinformatics*, 2017.

[114] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome biology*, vol. 11, no. 3, p. R25, 2010.

[115] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[116] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, p. R106, Oct 2010.

[117] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[118] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[119] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC bioinformatics*, vol. 11, no. 1, p. 422, 2010.

[120] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nature biotechnology*, vol. 31, no. 1, p. 46, 2013.

[121] E. Turro, W. J. Astle, and S. Tavaré, "Flexible analysis of RNA-seq data using mixed effects models," *Bioinformatics*, vol. 30, no. 2, pp. 180–188, 2013.

[122] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome research*, vol. 22, no. 10, pp. 2008–2017, 2012.

[123] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *Journal of proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010.

[124] V. Vidova and Z. Spacil, "A review on mass spectrometry-based quantitative proteomics: targeted and data independent acquisition," *Analytica chimica acta*, vol. 964, pp. 7–23, 2017.

[125] M.-S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, *et al.*, "A draft map of the human proteome," *Nature*, vol. 509, no. 7502, p. 575, 2014.

[126] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, *et al.*, "Mass-spectrometry-based draft of the human proteome," *Nature*, vol. 509, no. 7502, p. 582, 2014.

[127] L. Martens and J. A. Vizcaíno, "A golden age for working with public proteomics data," *Trends in biochemical sciences*, vol. 42, no. 5, pp. 333–341, 2017.

[128] S. D. Patterson, "Data analysis—the Achilles heel of proteomics," *Nature biotechnology*, vol. 21, no. 3, p. 221, 2003.

[129] S. Gallien, E. Duriez, and B. Domon, "Selected reaction monitoring applied to proteomics," *Journal of Mass Spectrometry*, vol. 46, no. 3, pp. 298–312, 2011.

[130] R. Aebersold and M. Mann, "Mass-spectrometric exploration of proteome structure and function," *Nature*, vol. 537, no. 7620, p. 347, 2016.

[131] O. T. Schubert, L. C. Gillet, B. C. Collins, P. Navarro, G. Rosenberger, W. E. Wolski, H. Lam, D. Amodei, P. Mallick, B. MacLean, *et al.*, "Building high-quality assay libraries for targeted analysis of SWATH MS data," *Nature protocols*, vol. 10, no. 3, p. 426, 2015.

[132] B. J. Blencowe, "The relationship between alternative splicing and proteomic complexity," *Trends in biochemical sciences*, vol. 42, no. 6, pp. 407–408, 2017.

[133] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature genetics*, vol. 40, no. 12, p. 1413, 2008.

[134] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, *et al.*, "Ensembl 2012," *Nucleic acids research*, vol. 40, no. D1, pp. D84–D90, 2011.

[135] B. Taneri, B. Snyder, and T. Gaasterland, "Distribution of alternatively spliced transcript isoforms within human and mouse transcriptomes," *Journal of OMICS Research*, vol. 1, no. 1, pp. 1–5, 2011.

[136] M. Gonzàlez-Porta and A. Brazma, "Identification, annotation and visualisation of extreme changes in splicing from RNA-seq experiments with SwitchSeq," *bioRxiv*, p. 005967, 2014.

[137] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, *et al.*, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, 2015.

[138] L. Fagerberg, B. M. Hallstrom, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, *et al.*, "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," *Molecular & Cellular Proteomics*, pp. mcp–M113, 2013.

[139] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, *et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[140] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, *et al.*, "The Pfam protein families database: towards a more sustainable future," *Nucleic acids research*, vol. 44, no. D1, pp. D279–D285, 2015.

[141] I. Ezkurdia, J. M. Rodriguez, E. Carrillo-de Santa Pau, J. Vázquez, A. Valencia, and M. L. Tress, "Most highly expressed protein-coding genes have a single dominant isoform," *Journal of proteome research*, vol. 14, no. 4, pp. 1880–1887, 2015.

[142] I. Ezkurdia, A. del Pozo, A. Frankish, J. M. Rodriguez, J. Harrow, K. Ashman, A. Valencia, and M. L. Tress, "Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function," *Molecular biology and evolution*, vol. 29, no. 9, pp. 2265–2283, 2012.

[143] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[144] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[145] A. Levchuk, "Pairwise string alignment in Python (Needleman-Wunsch and Smith-Waterman algorithms)," 2017. Visited on 2018-06-27.

[146] J. P. O'rourke and S. A. Ness, "Alternative RNA splicing produces multiple forms of c-Myb with unique transcriptional activities," *Molecular and cellular biology*, vol. 28, no. 6, pp. 2091–2101, 2008.

[147] M. Buljan, G. Chalancon, S. Eustermann, G. P. Wagner, M. Fuxreiter, A. Bateman, and M. M. Babu, "Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks," *Molecular cell*, vol. 46, no. 6, pp. 871–883, 2012.

[148] J. D. Ellis, M. Barrios-Rodiles, R. Çolak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim, *et al.*, "Tissue-specific alternative splicing remodels protein-protein interaction networks," *Molecular cell*, vol. 46, no. 6, pp. 884–892, 2012.

[149] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature methods*, vol. 10, no. 3, p. 221, 2013.

[150] S. Das, D. Lee, I. Sillitoe, N. L. Dawson, J. G. Lees, and C. A. Orengo, "Functional classification of CATH superfamilies: a domain-based approach for protein function annotation," *Bioinformatics*, vol. 31, no. 21, pp. 3460–3467, 2015.

[151] V. Sangar, D. J. Blankenberg, N. Altman, and A. M. Lesk, "Quantitative sequence-function relationships in proteins based on gene ontology," *BMC bioinformatics*, vol. 8, no. 1, p. 294, 2007.

[152] Y. Z. Kurmangaliyev and M. S. Gelfand, "Computational analysis of splicing errors and mutations in human transcripts," *BMC genomics*, vol. 9, no. 1, p. 13, 2008.

[153] Y. Li, Y.-c. Bor, Y. Misawa, Y. Xue, D. Rekosh, and M.-L. Hammarskjöld, "An intron with a constitutive transport element is retained in a Tap messenger RNA," *Nature*, vol. 443, no. 7108, p. 234, 2006.

[154] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes transcriptomes," *Genome research*, pp. gr–177790, 2014.

[155] J. J. Li, P. J. Bickel, and M. D. Biggin, "System wide analyses have underestimated protein abundances and the importance of transcription in mammals," *PeerJ*, vol. 2, p. e270, 2014.

[156] E. Lundberg, L. Fagerberg, D. Klevebring, I. Matic, T. Geiger, J. Cox, C. Älgenäs, J. Lundeberg, M. Mann, and M. Uhlen, "Defining the transcriptome and proteome in three functionally different human cell lines," *Molecular systems biology*, vol. 6, no. 1, p. 450, 2010.

[157] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, and M. Mann, "Deep proteome and transcriptome mapping of a human cancer cell line," *Molecular systems biology*, vol. 7, no. 1, p. 548, 2011.

[158] R. J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, *et al.*, "Ensembl BioMarts: a hub for data retrieval across taxonomic space," *Database*, vol. 2011, 2011.

[159] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13, no. 4, p. 227, 2012.

[160] J. M. Rodriguez, J. Rodriguez-Rivas, T. Di Domenico, J. Vázquez, A. Valencia, and M. L. Tress, "APPRIS 2017: principal isoforms for multiple gene sets," *Nucleic acids research*, vol. 46, no. D1, pp. D213–D217, 2017.

[161] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[162] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "Genetic effects on gene expression across human tissues," *Nature*, vol. 550, no. 7675, p. 204, 2017.

[163] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, *et al.*, "Kinesin family member 1b." https://www.genecards.org/cgi-bin/carddisp.pl?gene=KIF1B, 2015. Visited on 2019-01-06.

[164] Y. Okada, H. Yamazaki, Y. Sekine-Aizawa, and N. Hirokawa, "The neuron-specific kinesin superfamily protein kif1a is a uniqye monomeric motor for anterograde axonal transport of synaptic vesicle precursors," *Cell*, vol. 81, no. 5, pp. 769–780, 1995.

[165] R. J. Haslam, H. B. Koide, and B. A. Hemmings, "Pleckstrin domain homology," *Nature*, vol. 363, no. 6427, p. 309, 1993.

[166] R. D. Vale, "The molecular motor toolbox for intracellular transport," *Cell*, vol. 112, no. 4, pp. 467–480, 2003.

[167] D. Durocher and S. P. Jackson, "The fha domain," *FEBS letters*, vol. 513, no. 1, pp. 58–66, 2002.

[168] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, *et al.*, "Solute carrier family 35 member e4." https://www.genecards.org/cgi-bin/carddisp.pl?gene=SLC35E4, 2015. Visited on 2019-01-06.

[169] D. L. Jack, N. M. Yang, and M. H. Saier, "The drug/metabolite transporter superfamily," *European Journal of Biochemistry*, vol. 268, no. 13, pp. 3620–3639, 2001.

[170] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, *et al.*, "Microtubule affinity regulating kinase 4." https://www.genecards.org/cgi-bin/carddisp.pl?gene=MARK4, 2015. Visited on 2019-01-06.

[171] J.-P. Tassan and X. Le Goff, "An overview of the kin1/par-1/mark kinase family," *Biology of the Cell*, vol. 96, no. 3, pp. 193–199, 2004.

[172] V. Su and A. F. Lau, "Ubiquitin-like and ubiquitin-associated domain proteins: significance in proteasomal degradation," *Cellular and molecular life sciences*, vol. 66, no. 17, pp. 2819–2833, 2009.

[173] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, *et al.*, "Phosphotriesterase related." https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTER, 2015. Visited on 2019-01-06.

[174] L. Holm and C. Sander, "An evolutionary treasure: unification of a broad set of amidohydrolases related to urease," *Proteins: Structure, Function, and Bioinformatics*, vol. 28, no. 1, pp. 72–82, 1997.

[175] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, *et al.*, "Pleckstrin and sec7 domain containing 3." https://www.genecards.org/cgi-bin/carddisp.pl?gene=PSD3, 2015. Visited on 2019-01-06.

[176] P. Chardin, S. Paris, B. Antonny, S. Robineau, S. Béraud-Dufour, C. L. Jackson, and M. Chabre, "A human exchange factor for arf contains sec7-and pleckstrin-homology domains," *Nature*, vol. 384, no. 6608, p. 481, 1996.

[177] D. J. Elliott and S. N. Grellscheid, "Alternative rna splicing regulation in the testis," *Reproduction*, vol. 132, no. 6, pp. 811–819, 2006.

[178] B. Raj and B. J. Blencowe, "Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles," *Neuron*, vol. 87, no. 1, pp. 14–27, 2015.

[179] K. Nakka, C. Ghigna, D. Gabellini, and F. J. Dilworth, "Diversification of the muscle proteome through alternative splicing," *Skeletal muscle*, vol. 8, no. 1, p. 8, 2018.

[180] P. Blakeley, J. A. Siepen, C. Lawless, and S. J. Hubbard, "Investigating protein isoforms via proteomics: a feasibility study," *Proteomics*, vol. 10, no. 6, pp. 1127–1140, 2010.

[181] M. Brosch, G. I. Saunders, A. Frankish, M. O. Collins, L. Yu, J. Wright, R. Verstraten, D. J. Adams, J. Harrow, J. S. Choudhary, *et al.*, "Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome," *Genome research*, 2011.

[182] M. L. Tress, B. Bodenmiller, R. Aebersold, and A. Valencia, "Proteomics studies confirm the presence of alternative protein isoforms on a large scale," *Genome biology*, vol. 9, no. 11, p. R162, 2008.

[183] G. Lopez-Casado, P. A. Covey, P. A. Bedinger, L. A. Mueller, T. W. Thannhauser, S. Zhang, Z. Fei, J. J. Giovannoni, and J. K. Rose, "Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case," *Proteomics*, vol. 12, no. 6, pp. 761–774, 2012.

[184] G. M. Sheynkman, M. R. Shortreed, B. L. Frey, and L. M. Smith, "Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq," *Molecular & Cellular Proteomics*, pp. mcp–O113, 2013.

[185] R. J. Grainger and J. D. Beggs, "Prp8 protein: at the heart of the spliceosome," *Rna*, vol. 11, no. 5, pp. 533–557, 2005.

[186] W. P. Galej, C. Oubridge, A. J. Newman, and K. Nagai, "Crystal structure of prp8 reveals active site cavity of the spliceosome," *Nature*, vol. 493, no. 7434, p. 638, 2013.

[187] "Comprehensive in vivo rna-binding site analyses reveal a role of prp8 in spliceosomal assembly,"

[188] A. Zhou, F. Zhang, and J. Y. Chen, "PEPPI: a peptidomic database of human protein isoforms for proteomics experiments," in *BMC bioinformatics*, vol. 11, p. S7, BioMed Central, 2010.

[189] X.-B. Xing, Q.-R. Li, H. Sun, X. Fu, F. Zhan, X. Huang, J. Li, C.-L. Chen, Y. Shyr, R. Zeng, *et al.*, "The discovery of novel protein-coding features in mouse genome based on mass spectrometry data," *Genomics*, vol. 98, no. 5, pp. 343–351, 2011.

[190] V. O. Wickramasinghe, M. Gonzàlez-Porta, D. Perera, A. R. Bartolozzi, C. R. Sibley, M. Hallegger, J. Ule, J. C. Marioni, and A. R. Venkitaraman, "Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5 prime splice site strength," *Genome biology*, vol. 16, no. 1, p. 201, 2015.

[191] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinović, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmström, L. Malmström, *et al.*, "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data," *Nature biotechnology*, vol. 32, no. 3, p. 219, 2014.

[192] M. Kalyna, C. G. Simpson, N. H. Syed, D. Lewandowska, Y. Marquez, B. Kusenda, J. Marshall, J. Fuller, L. Cardle, J. McNicol, *et al.*, "Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis," *Nucleic acids research*, vol. 40, no. 6, pp. 2454–2469, 2011.

[193] J. J.-L. Wong, W. Ritchie, O. A. Ebner, M. Selbach, J. W. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, *et al.*, "Orchestrated intron retention regulates normal granulocyte differentiation," *Cell*, vol. 154, no. 3, pp. 583–595, 2013.

[194] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng, *et al.*, "Estimating accuracy of RNA-Seq and microarrays with proteomics," *BMC genomics*, vol. 10, no. 1, p. 161, 2009.

[195] A. Ghazalpour, B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, *et al.*, "Comparative analysis of proteome and transcriptome variation in mouse," *PLoS genetics*, vol. 7, no. 6, p. e1001393, 2011.

[196] M. Jovanovic, M. S. Rooney, P. Mertins, D. Przybylski, N. Chevrier, R. Satija, E. H. Rodriguez, A. P. Fields, S. Schwartz, R. Raychowdhury, *et al.*, "Dynamic profiling of the protein life cycle in response to pathogens," *Science*, vol. 347, no. 6226, p. 1259038, 2015.

[197] M. S. Robles, J. Cox, and M. Mann, "In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism," *PLoS genetics*, vol. 10, no. 1, p. e1004047, 2014.

[198] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg, and R. Aebersold, "The quantitative proteome of a human cell line," *Molecular systems biology*, vol. 7, no. 1, p. 549, 2011.

[199] M. L. Tress, F. Abascal, and A. Valencia, "Most alternative isoforms are not functionally important," *Trends in biochemical sciences*, vol. 42, no. 6, pp. 408–410, 2017.

[200] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome biology*, vol. 10, no. 3, p. R25, 2009.

[201] T. Maier, M. Güell, and L. Serrano, "Correlation of mRNA and protein in complex biological samples," *FEBS letters*, vol. 583, no. 24, pp. 3966–3973, 2009.

[202] A. Rhoads and K. F. Au, "Pacbio sequencing and its applications," *Genomics, proteomics & bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015.

[203] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, no. 1, p. 239, 2016.

[204] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.