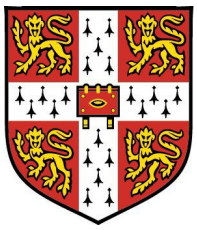
RNA sequencing for the study of splicing



Mar Gonzàlez-Porta

European Molecular Biology Laboratory
European Bioinformatics Institute

Darwin College

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

April 2014

To my mum.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated.

The text in this thesis does not exceed the specified word limit of 60,000 words as defined by the Biology Degree Committee.

April, 2014 Mar Gonzàlez-Porta

Abstract

Amongst the many processes that shape the final set of RNA molecules present in eukaryote cells, splicing emerges as the most prominent mechanism for message diversification. In recent years, applications of high throughput sequencing to RNA, known as RNA sequencing, have opened new avenues for the study of transcriptome composition, and have enabled further characterisation of such mechanism. In this thesis, I focus on the application of this technology to the study of human transcript diversity and its potential impact on the protein repertoire.

In the first results chapter, I explore the extent of transcriptome diversity by asking whether there is a preference for the production of specific alternative transcripts within each given gene. I show that while many alternative transcripts can be detected, the expression of most protein coding genes tends to be dominated by one single transcript (major transcript). Such findings are validated in the second chapter, and are further used to explore changes in splicing patterns in a disease context. By analysing healthy and tumor samples from kidney cancer patients, I show that most of the detected splicing alterations do not lead to big changes in the relative abundance of major transcripts, at least in a recurrent manner. In addition, I introduce a framework to visualise the most extreme changes in splicing and to evaluate their potential functional impact. In the third chapter, I investigate the role of spliceosome assembly dynamics on the regulation of splice site choice. I show that depletion of *PRPF8*, a core spliceosomal component, leads to the preferential retention of a subset of introns with weaker splice sites, and also introduces alterations in the rate of co-transcriptional splicing. Finally, in the last chapter, I explore the validation of changes in alternative transcript abundance at the protein level, through the integration of results derived from RNA sequencing datasets with those obtained from proteomics experiments.

Altogether, the findings described in this thesis provide a global picture on the extent of alternative splicing in the diversification of the transcriptome, expand current knowledge on the splicing reaction, and open new possibilities for the integration of transcriptomics and proteomics data.

Preface

I would like to start by thanking my supervisor Alvis Brazma, for giving me the opportunity to get involved in this experience. My PhD has been a long journey of almost four years, and I am extremely grateful for his support and guidance throughout all this time. I am also deeply indebted to Johan Rung, who behaved like a second supervisor to me, and to John Marioni, without whom half of this dissertation would have not been possible.

I cannot thank enough all the members of the Functional Genomics research team, who have made it easy to go to the office every day. Special thanks to Mitra, Liliana, Natalja, Gabry, Angela, Shyama, Wan, Jing, Aurora, Sandra and Yuedan for their continuous encouragement and interest in my research. Likewise, I shall not forget all of those with whom I have not necessarily shared offices, but who have always kept a smile on their faces when crossing me in the corridor.

I wish to acknowledge the members of my Thesis Advisory Committee, John Marioni, Jan Korbel and Simon Tavaré, for their insightful comments, as well as all my collaborators, including Adam Frankish, Jennifer Harrow, David Perera, Arthur Bartolozzi, Ashok Venkitaraman and Yansheng Liu, for the many thoughtful discussions. I would like to pay special thanks to Vi Wickramasinghe, for making the end of my PhD such a great experience. I am also thankful to all the people who have provided feedback along the path, specially Wolfgang Huber, Nicholas Luscombe, Roderic Guigó and Ernest Turró, and to all of those who have helped in improving this dissertation with their comments. Looking back, I am most grateful to all the people who contributed to my scientific growth and education, particularly Antoni Romeu, Julien Roux, Marc Robinson-Rechavi and Roderic Guigó. I also need to thank Gabriella Rustici for dynamising my PhD experience by in-

volving me in teaching. Similarly, I wish to thank all the people with whom I have been involved in this endeavor, including Angela Gonçalves, Mitra Barzine, Liliana Greguer and Laura Emery, as well as Francesco Paolo Casale for his great patience. Finally, I would like to express special thanks to Lynn French and Tracey Andrew, for making my bureaucratic life that much easier.

I have been extremely fortunate to share my time at the EBI with a great group of people, from which I would like to give special thanks to Robert, Alexa, Nidhi, Matthias, Senay, Sergio, Rita, Cathy and Afonso. Likewise, I cannot thank enough all those friends who have remained a part of my life, despite the long distances and the course of time.

Last, I am deeply grateful to all the people who have supported me in my personal life, specially my family, who have contributed the most in making this possible. In particular, I would like to thank my brother, for keeping the family conversations alive; my dad, for his routine visits and continuous care; and my mum, for pushing us all to continue ahead. Finally, I wish to thank Pol, for the countless trips and hours in Skype, which have now finally been replaced for conversations over dinner. I would not have made it this far without him.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When the first human genome sequence was declared completed in 2003, it established the order of the billions of letters that contain the instructions for the survival of each of the cells in our body [International Human Genome Sequencing Consortium, 2001, 2004; Venter et al., 2001], an exercise that was compared to reading the book of life[1]. In fact, the comparison of DNA to a book of instructions has been widely used as a metaphor to explain the flow of genetic information[2], and far from analogies, it has been recently shown that this molecule is able to encode for text and even more [Goldman et al., 2013]. Since then, vast efforts have been devoted to interpreting the content of the human genome sequence, best exemplified by the ENCODE project [ENCODE Project Consortium et al., 2012]. It has become clear from those that the flow of information is not as simple as it was initially envisioned [Crick, 1970], and that the book of instructions might actually look closer to those from the series "Choose Your Own Adventure"[3], in which one text can lead to different outcomes.

Overall, from the interpretation of the genetic information stored in the DNA to protein synthesis, many regulatory processes contribute to shape the final repertoire of molecules present in the cell. In multicellular organisms such as humans, those processes constitute the intricate basis upon which a large diversity of cell

[1] http://www.sanger.ac.uk/about/press/2003/030414.html
[2] http://ed.ted.com/lessons/dna-the-book-of-you-joe-hanson
[3] http://en.wikipedia.org/wiki/Choose_Your_Own_Adventure

types and states are derived from the same genetic message. In this context, the study of transcriptome composition has emerged as a valuable strategy to understand such diversity, and while the mere existence of an RNA molecule does not ensure the production of a functional protein product, this type of knowledge has helped in the identification of links between phenotype and genotype [Vogel and Marcotte, 2012]. Furthermore, recent technological advances have enabled the characterisation of RNA samples in a more reproducible and high throughput fashion than is yet possible for their final products [Wang et al., 2009].

In the present chapter, I provide an overview of the different regulatory processes that control the identity and abundance of the set of RNAs found in a given cell, with special emphasis on splicing and its potential for message diversification. In addition, I introduce key concepts on the use of RNA sequencing for the study of gene and transcript expression in a high throughput manner.

1.1 A day in the life of an mRNA

Amongst the diversity of identified RNA species, messenger RNAs (mRNAs) are regarded as those that contain the necessary information for the synthesis of proteins. The half-lives of mRNAs are short compared to those of other molecules in the cell (*e.g.* most mammalian mRNAs remain in the cell for approximately 9h, compared to 46h in the case of proteins [Schwanhausser et al., 2011]). Nonetheless, from birth to death, these molecules are controlled by complex regulatory systems that determine which messages are eventually expressed (**Figure 1.1**). In this section, I provide an overview of the key steps behind such control.

1.1.1 Transcription

Transcription is the first phase in determining the set of RNAs expressed in a given cell [Alberts et al., 2002]. During transcription, stretches of DNA (*i.e.* genes) are used as templates for the synthesis of complementary single stranded RNA molecules (*i.e.* transcripts). In eukaryote cells, such reaction can be catalysed by three different enzymes (*i.e.* RNA polymerases I, II and III), depending on the type of gene being targeted. RNA polymerase II is commonly regarded as the most prominent player in the transcription reaction, since it is responsible for the synthesis of RNAs derived from the majority of genes, including those that encode for proteins. On the other hand, RNA polymerase I and III are specifically

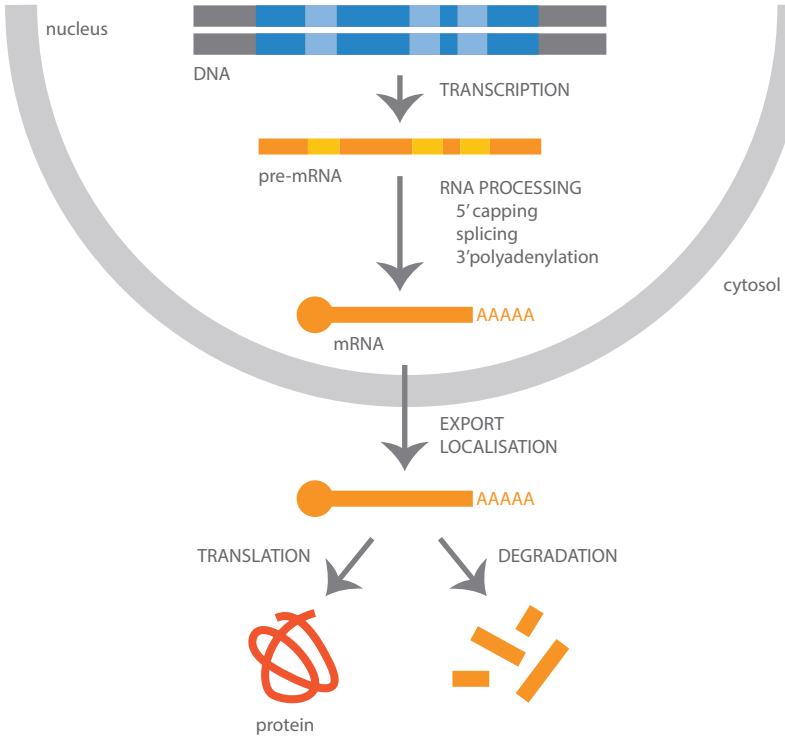


Figure 1.1| Key steps in the regulation of eukaryotic gene expression. mRNA expression starts with the nuclear transcription of specific DNA loci which contain the information required for the synthesis of proteins (*i.e.* genes). Following several processing steps, some of which occur co-transcriptionally, the transcription products are further transformed into mature mRNAs that can then be exported to the cytosol and localised to sub-cellular compartments. mRNA export is linked to strict quality control mechanisms, and both unprocessed RNAs and debris from the previous processes will be degraded previously to this stage. Once in the cytosol, mRNAs can be recognised by ribosomes and translated into proteins, and will be eventually degraded.

involved in the transcription of ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and several small RNAs [Paule and White, 2000].

Transcription by RNA pol II is a multi-step process that starts with the binding of several proteins to a regulatory region located upstream of the gene, known as the promoter [Fuda et al., 2009]. These proteins enable the subsequent assembly of the polymerase and the formation of the transcription initiation complex, and given that they participate in the recognition of the majority of promoters, they are commonly regarded as general transcription factors (TFs). Conversely, more spe-

cific TFs also exist, which are able to modulate the fate of the reaction by binding to DNA regions that promote or inhibit polymerase assembly (*i.e.* enhancers and silencers, respectively), hence contributing to the regulation of expression levels [Vaquerizas et al., 2009]. Following the assembly steps and further conformational rearrangements, RNA pol II releases from the large complex of proteins and abandons the promoter region (*i.e.* promoter clearance), thus entering the elongation phase [Kwak and Lis, 2013]. However, the transition to this stage is not immediate, and in most cases the polymerase remains at the promoter generating short truncated transcripts (*i.e.* abortive initiation). During elongation, RNA is synthesised from the transcription start site (TSS), and nucleotides are incorporated in a complementary basis in the 5′ to 3′ direction. Eventually, the polymerase transcribes through the cleavage and polyadenylation signals that mark the end of the gene, and it is released from the DNA template [Kuehner et al., 2011].

1.1.2 mRNA processing

All mRNA molecules undergo several modifications before they are exported to the cytosol, which include the addition of a 5′ cap, the polyadenylation of the 3′ end and the removal of introns via splicing [Darnell, 2013].

Shortly after the RNA pol II has entered the elongation phase, a methylated guanine nucleotide is added to its 5′ end through an enzymatic reaction (*i.e.* capping). Such cap not only delimits the 5′ end of transcripts, but also enables the distinction of mRNAs from other RNA species (*e.g.* RNA pol I and III produce uncapped RNAs). Furthermore, it protects the RNA molecule from degradation, and aids in the initiation of translation by enabling the binding of ribosomes to the mRNA. Similarly, the 3′ end of mRNAs is also modified through the addition of a polyAdenosine (polyA) tail, which is typically 200-250 nucleotides long in mammalian cells. Polyadenylation serves the function of extending the mRNA half-life, and given that it is a modification shared by all mRNA molecules, it is commonly exploited for their study (*i.e.* one of the most common RNA extraction protocols in use relies on the specific selection of polyA-tailed RNA species).

Splicing, however, is a much more complex reaction. During this process, some regions of the pre-mRNA are lost (*i.e.* introns), and the stretches of sequence that contain the necessary information for protein synthesis (*i.e.* exons) are brought together, as detailed later on in this chapter. As a result, a mature mRNA product is obtained.

1.1.3 The cytosolic life of mRNAs

Once a mature mRNA molecule is obtained, it undergoes selective export through the nuclear pore. mRNA export is linked to strict quality control mechanisms that ensure that immature RNAs remain in the nucleus [Porrua and Libri, 2013]. Those mechanisms rely on the recognition of protein complexes that accompany the RNA molecules (*i.e.* RNA binding proteins; RBPs), which act as markers on the completion status of the processing steps mentioned in the previous section. For example, the cap-binding and polyA binding-complexes serve as indicators of successful capping and polyadenylation reactions, respectively, and other protein complexes mark the end of the splicing in a similar fashion (*i.e.* the exon-junction complex - EJC; see next section). Conversely, the presence of RBPs involved in the execution of each of these steps marks the mRNA molecule as immature, hence preventing its export.

Unprocessed mRNAs, together with the remainders from the transcription and splicing reactions, will be eventually degraded by the exosome, a large complex of RNA exonucleases [Pérez-Ortín et al., 2013]. In the cases when they are erroneously exported, or when intact mRNAs become damaged in the cytosol, further quality control mechanisms prevent their translation. Most of these are intrinsic to the steps required for the initiation of protein synthesis, as is the case for the recognition of the 5′ cap and polyA tail by the translation initiation machinery. However, a separate surveillance system also exists, which actively seeks aberrant mRNAs for degradation, before efficient translation occurs. This system is referred to as nonsense-mediated decay (NMD) and specifically targets the presence of premature stop codons in the transcript, which might arise from errors in the splicing reaction [Pérez-Ortín et al., 2013]. During NMD, a first round of translation starts as soon as the 5′ end of the mRNA emerges from the nuclear pore, during which the exon-junction complexes that surround each splice-site are detached from the mRNA. Under the presence of nonsense codons, the mRNA remains bounded to such complexes and is rapidly degraded.

Following export, mRNAs are then localised within the cytosol according to the signals encoded in their 3′ UTR regions, and are eventually recognised by ribosomes and translated [Alberts et al., 2002]. Such binding of ribosomes to the mRNAs is in direct competition with mRNA decay, a process that starts as soon as transcripts are exported into the cytosol and which consists of the gradual short-

ening of the polyA tail. Once the polyA tail reaches a critical length, mRNAs are ultimately degraded, either through the continuation of the digestion from the 3′ end or through the removal of the 5′ cap (*i.e.* decapping) and subsequent 5′ to 3′ decay [Schoenberg and Maquat, 2012]. Alternatively, cytosolic polyadenylation can also occur, thus having a positive impact on the mRNA half life [Villalba et al., 2011]. Altogether, this evidences a critical role for both the 5′ and 3′ UTRs in regulating mRNA stability and translation efficiency.

1.2 The splicing reaction

Splicing was first discovered in 1977 by Phillip Sharp and Richard J. Roberts, who independently observed that genes could be encoded across split segments in the DNA. In their experiments, they hybridised adenoviral mRNAs with complementary single stranded DNA fragments, and following observation with electron microscopy (EM), they detected alternate double stranded and single stranded stretches in the resulting hybrid. They concluded that certain regions of the mRNA are removed during its maturation (*i.e.* the introns), hence bringing together separate parts of the RNA [Berget et al., 1977; Chow et al., 1977]. This observation was soon extended to all domains of life, even though splicing is most prevalent in eukaryotes (although some instances of splicing have been detected in prokaryotes, they lack the major pathway through which this process is achieved) [Alberts et al., 2002]. Both Sharp and Roberts were awarded a Nobel Prize in 1993 for their contribution.

In the next section, I describe further details regarding this RNA processing step, with emphasis on the key players and steps required for its completion. I also discuss the potential of this process for the diversification of the message encoded in the DNA, and outline the elements that contribute to its regulation.

1.2.1 Key players in mRNA splicing

In eukaryotes, splicing is most commonly carried out through the spliceosomal pathway, whereby a large complex of proteins and RNAs orchestrates the process of intron removal. Such complex is known as the spliceosome, and has been categorised as one of the most complicated machineries in the cell [Nilsen, 2003]. Two different types of spliceosomes have been identified (*i.e.* the major and minor), which differ in their components and the properties of the introns that they tar-

get. Specifically, the major spliceosome is involved in >99% of the splicing events, and is responsible for the removal of introns that harbour consensus splice site sequences (*i.e.* canonical splicing; **Figure 1.2a**) [Matera and Wang, 2014]. By contrast, a small set of introns display sequences that differ from the consensus ones, and are targeted instead by the minor spliceosome (*i.e.* non-canonical splicing) [Turunen et al., 2013].

In addition to spliceosomal introns, a separate class of introns that undergo splicing in a protein-independent fashion have also been detected. Such introns, referred to as self-splicing introns, are able to mediate the splicing reaction through rearrangements in the RNA structure by acting as ribozymes [Alberts et al., 2002]. While these are rarely detected in eukaryotes, they contribute to the splicing of some organelle genes and rRNAs, thus providing evidence that splicing is not limited to mRNAs (*i.e.* in fact, other non-coding RNAs such as tRNAs, micro-RNAs and long non-coding RNAs can also undergo such reaction). Their existence is considered to support the RNA world hypothesis, which states that self-replicating RNA constituted the initial building blocks of life [Robertson and Joyce, 2012].

Given the relevant contribution of the major spliceosome in catalysing the removal of the vast majority of introns, this section will further focus on the process of canonical splicing.

1.2.2 Splicing by the major spliceosome

The major spliceosome is a highly dynamic complex that undergoes significant conformational and compositional rearrangements during the several steps of the splicing reaction [Will and Luhrmann, 2011]. It is composed of five different small nuclear RNA molecules (snRNAs: U1, U2, U4, U5 and U6), as well as several hundreds of proteins. Each of these RNA molecules associate with several proteins and form complexes called small nuclear ribonucleoproteins (snRNP). Such snRNPs form the core of the spliceosome and are directly involved in the recognition of splice sites and branch-point sequences, as well as the catalysis of the splicing reaction. Indeed, one of the most striking features about the spliceosome is that the actual intron removal reaction is catalysed by RNA molecules rather than proteins [Fica et al., 2013]. More specifically, splicing is the result of two transesterification reactions, which are based on nucleophilic attacks between RNA nucleotides (**Figure 1.2b**) [Will and Luhrmann, 2011]. In the first reaction,

the 5′ exon is cleaved from the intron through a nucleophilic attack by the 2′ hydroxyl group of the A branch-point residue on the phosphate group of the GU dinucleotide at the 5′ splice site, thus forming a lariat intermediate. Next, the two exons become ligated via a similar reaction that involves the 3′ hydroxyl group of the previously released exon and the phosphate group of the last nucleotide of the 3′ end of the intron, thus causing the release of the intron in the form of a lariat.

However, before the actual intron removal can occur, the active catalytic site of the spliceosome needs to be created, an event that requires many changes in its composition and conformation (**Figure 1.2c**) [Matera and Wang, 2014]. The first step towards the accomplishment of the splicing reaction is splice site recognition. This task is achieved trough base pairing between the U1 and U2 snRNPS and the 5′ and 3′ splice sites, respectively, and results in the formation of complex E. Similarly, the branch-point sequence is also recognised by the U2 snRNP, which eventually interacts with the U1 snRNP and forms the pre-spliceosome (complex A), thus bringing together both splice sites. Further recruitment of a pre-assembled tri-snRNP formed by the U4, U5 and U6 snRNPs (the U4-U6·U5 tri-snRNP) leads to the creation of complex B, which then becomes activated as a result of several conformational and compositional changes. The activated complex B (complex B*) mediates the first catalytic step in the splicing reaction and leads to the creation of complex C, which contains the detached exon (exon 1) and the lariat intermediate (intron - exon 2). Following extra rearrangements, complex C then performs the second catalytic step and forms a post-spliceosomal complex, which contains the spliced exons and the lariat. The latter is eventually released, together with the remaining snRNPs, which are then recycled for the next round of splicing. Importantly, these steps are followed by the binding of a new complex of proteins to the newly created exon junction (*i.e.* the exon junction complex, ECJ), which mark the successful completion of splicing at that specific location and further contribute in shaping the fate of the mRNA molecule.

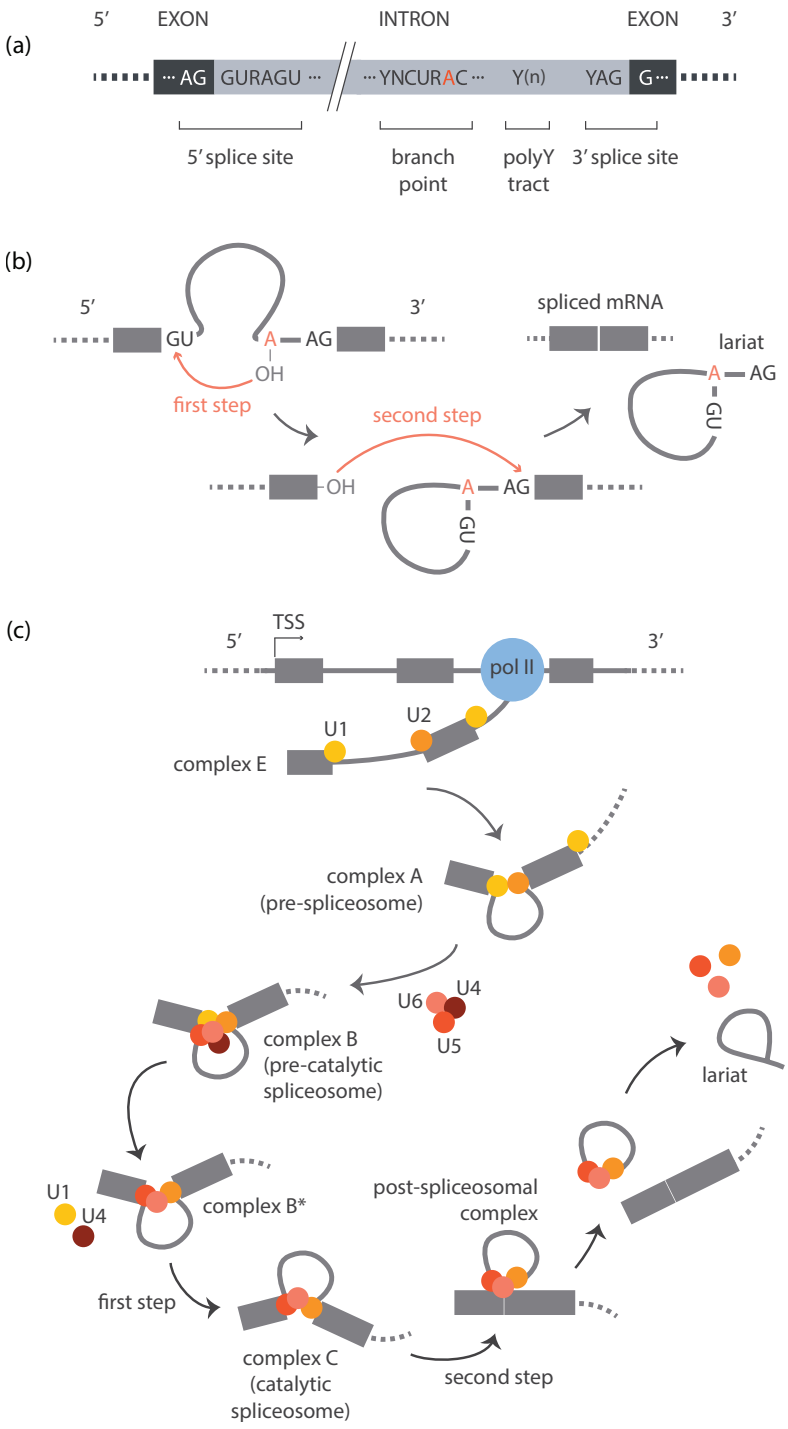

(a)
5′
EXON
INTRON
EXON
3′
···AG GURAGU ···
···YNCURAC···
Y(n)
YAG
G···
5′ splice site
branch point
polyY tract
3′ splice site
(b)
5′
GU
A
AG
3′
OH
first step
second step
spliced mRNA
lariat
A
AG
GU
OH
A
AG
GU
(c)
5′
TSS
pol II
3′
U1
U2
complex E
complex A
(pre-spliceosome)
U6
U4
U5
complex B
(pre-catalytic
spliceosome)
lariat
U1
U4
complex B*
post-spliceosomal
complex
first step
complex C
(catalytic
spliceosome)
second step

Figure 1.2| Splicing by the major spliceosome. Adapted from Will and Luhrmann [2011].
(a) Core splicing signals recognised by the major spliceosome. The spliceosome recognises several core signals in the pre-mRNA: the 5′ and 3′ splice sites, the branch-point sequence (typically located between 15 and 50 nucleotides upstream of the 3′ intron) and the polypyrimidine tract. Here, R indicates a purine (A or G), Y represents a pyrimidine (U or C) and N refers to any nucleotide. Introns that harbour these consensus sequences are referred to as U2-type introns, because they are recognised by the U2 snRNP.
(b) Steps in the splicing reaction. Splicing is the result of two transesterification reactions that involve the nucleotides from the pre-mRNA and snRNA molecules.
(c) Spliceosomal rearrangements during splicing. As transcription proceeds, several components of the spliceosome are transferred from the polymerase tail to the nascent pre-mRNA, thus facilitating the process of splice site recognition. Following this step, the spliceosome undergoes several compositional and conformational changes that lead to the formation of the catalytic site, the cleavage of the intron and the eventual release of the splicing products.

1.2.3 Diversifying the message: alternative splicing and beyond

During the splicing process, particular exons might become excluded from the final mRNA product, and similarly, some introns might fail to be removed, thus leading to the formation of alternative mRNA products from a given genomic locus (**Figure 1.3**). This process is known as alternative splicing and plays a key role in the diversification of the message encoded within a gene [Kornblihtt et al., 2013].

In humans, over 95% of multi-exon genes have been detected to result in more than one transcript, and such observation has been suggested as an explanation for the low number of genes compared to other lower eukaryotes (*e.g.* the human genome contains only ~30% more genes than that of *Caenorhabditis elegans*) [Pan et al., 2008; Wang et al., 2008]. Given the potential differences in biological function between the resulting alternative transcripts, alternative splicing can result in the generation of proteins with different biological function, structure, localisation and interaction capabilities [Keren et al., 2010; Nilsen and Graveley, 2010]. In this context, the detection of splicing products at the protein level confirms the potential of such process in increasing the protein repertoire [Tress et al., 2008b]. On the other hand, alternative splicing of pre-mRNAs can also contribute to the regulation of expression levels, through the formation of transcripts that will be targeted by the nonsense-mediated decay pathway, as well as non-coding mRNA

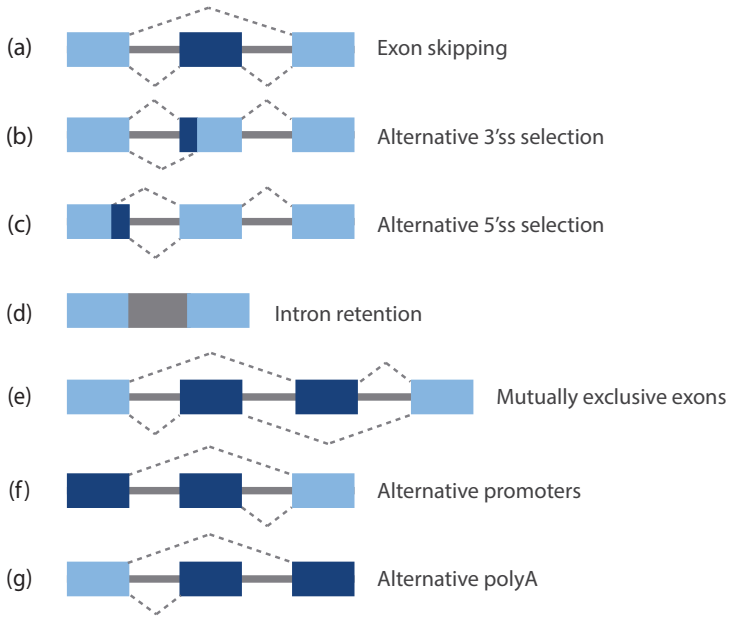


Figure 1.3 | Mechanisms for the formation of alternative transcripts from the same genomic locus. Differences in the execution of the splicing reaction can lead to message diversification (*a-e*). Similarly, the usage of alternative first and last exons also emerges as a common mechanism for the generation of alternative transcripts from the same gene (*f-g*). Adapted from Keren et al. [2010].

products that arise from intron retention events [McGlincy and Smith, 2008; Yap et al., 2012]. Finally, it has also been suggested that a considerable amount of the detected alternative splicing products result simply from noisy splicing and have no function at all [Melamud and Moult, 2009].

The biological significance of alternative splicing becomes evident in light of the detection of tissue-specific events [Buljan et al., 2012; Ellis et al., 2012; Merkin et al., 2012; Wang et al., 2008], as well as its relevant role in dynamic processes such as development [Kalsotra and Cooper, 2011] and cellular differentiation [Trapnell et al., 2010]. Hence, the annotation and functional characterisation of alternative mRNA products is an important task. A prominent effort towards that goal is the GENCODE project, which aims to annotate all evidence-based features in the human genome [Harrow et al., 2012]. Similarly, several methods have been devised in order to predict the impact of alternative splicing at the protein level (*e.g.* APPRIS [Rodriguez et al., 2013], AS-EAST [Shionyu et al., 2012], MAISTAS [Floris et al., 2011], AltAnalize [Emig et al., 2010]).

Nonetheless, message diversification does not stop with alternative splicing. For example, exons from different genes can be combined during the splicing reaction in a process called trans-splicing [Lasda and Blumenthal, 2011]. Similarly, RNA editing constitutes a separate mechanism through which the information contained within an mRNA can be altered, via the substitution, insertion or deletion of specific nucleotides [Maas, 2012]. Finally, post-translational modifications also contribute in shaping protein diversity. Amongst those, the removal of internal protein segments (*i.e.* inteins) emerges as an interesting example, given its analogy to alternative splicing [Volkmann and Mootz, 2013].

1.2.4 The regulation of splicing

Apart from the core splicing signals, further elements contribute to the definition of exon-intron boundaries and the regulation of splicing. This is the case for other cis-regulatory sequences that are typically present in the pre-mRNA (*i.e.* splicing regulatory elements, SREs), which can vary in terms of location and effect (**Figure 1.4a**) [Matera and Wang, 2014]. In general, SREs contribute to the recruitment of trans-acting splicing factors (SFs), a set of proteins that can act as repressors or activators of splicing, typically by influencing spliceosome assembly.

A prominent example of the role of SREs is their contribution to the recognition of exon-intron boundaries through a process called exon definition (**Figure 1.4b**) [De Conti et al., 2013]. Such mode of splice site recognition is especially common in higher eukaryotes, where intron size exceeds that of exons. This could lead to splicing errors because of the existence of cryptic splice sites. Hence, a class of SFs called SR proteins (Serine-Rich proteins) promote the binding of snRNPs to the splice sites located at both ends of the same exon, by binding to exonic splicing enhancers (ESE). As a result, a cross-exon recognition complex is formed, which will eventually lead to intron-spanning interactions through spliceosomal rearrangements. Conversely, in the case of short introns and in lower eukaryotes, intron definition emerges as the prevalent mode of splice site recognition (**Figure 1.4b**) [De Conti et al., 2013]. In this case, splice sites situated on both ends of the same intron are directly identified without the help of specific SFs.

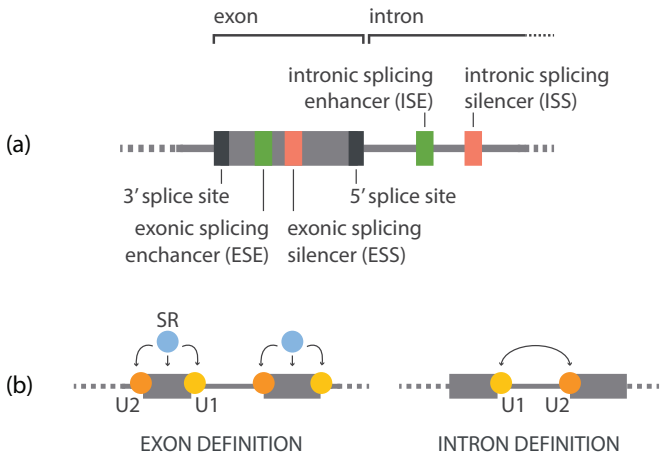


Figure 1.4 | The regulation of splicing.
(a) Cis-acting sequences involved in the regulation of intron removal. In addition to the core splicing signals (*i.e.* 5′ splice site, branch-point and 3′ splice site), several regulatory sequences shape the splicing decision by recruiting trans-acting splicing factors (SFs). Common SFs include SR proteins and hnRNPs, which typically promote and inhibit splicing, respectively. However, SRE activity is highly context dependent: albeit recruiting the same SF, the same sequence can have opposite roles, depending on whether it is located within exonic or intronic boundaries. Adapted from Matera and Wang [2014].
(b) Exon vs. intron definition. Intron definition is the prevalent mode of splice site recognition in lower eukaryotes, which consists of the pairing of the 5′ and 3′ splice site located at each end of the intron. In higher eukaryotes, the longer intron size could lead to the inclusion of cryptic splice sites, and exon definition is used instead. In this case, the recognition of the splice sites that surround a given exon is mediated by SR proteins, which promote the formation of a cross-exon recognition complex. Next, further spliceosomal rearrangements ensure intron-spanning interactions. Adapted from Ast [2004].

Together with the competition amongst splice sites and cis-acting SREs based on their sequence composition, the accessibility to those elements also plays a key role in alternative splicing regulation. Such accessibility can be influenced by the pre-mRNA secondary structure, chromatin arrangements and nucleosome positioning, and can be also dynamically controlled as a result of the coordination of the transcription and splicing processes [Brown et al., 2012; Plass and Eyras, 2014]. In humans, most of the splicing events occur before transcription termination, a phenomenon known as co-transcriptional splicing [Tilgner et al., 2012]. This implies that transcription elongation rates can have an impact on splice site choice: for example, a slow elongation will provide a window of opportunity for the recognition of weak splice sites, whilst a fast elongation will generally

promote the recognition of strong splice sites instead [Bentley, 2014]. Finally, the regulation of splicing decisions is not limited to the role of specific SFs, since fluctuations in the concentration of core components of the spliceosome are also known to influence the splicing outcome [Saltzman et al., 2011].

Overall, the above mentioned processes guarantee that splicing occurs in an accurate albeit flexible fashion. The accuracy of splicing is further increased by the many rearrangements that are required before the actual intron removal reaction can occur, and splicing errors are eliminated by the nonsense-mediated decay pathway. On the other hand, the accumulation of splice site mutations or the alteration in the function of spliceosomal components can lead to serious phenotypic consequences and, in fact, dysregulation of splicing has been linked to many diseases, including cancer [Ladomery, 2013; Padgett, 2012; Tazi et al., 2009].

1.3 Studying the transcriptome with RNA sequencing

In the past few years, RNA sequencing (RNA-seq) has become the method of choice for the study of transcriptome composition [Mortazavi et al., 2008; Wang et al., 2009]. Compared to microarrays, which constituted the first technology for the high throughput comparison of expression levels across conditions, RNA-seq offers a much bigger dynamic range to study gene expression patterns, and enables a much broader set of analyses without the need for intricate experimental designs [Malone and Oliver, 2011]. For example, besides standard differential gene expression analysis, popular applications of RNA-seq comprise the identification of novel transcribed regions, including fusion genes, the deconvolution of allele specific expression, and, as further explored in this thesis, the possibility to estimate transcript expression levels and to study differential splicing across conditions.

Since the introduction of the first sequencing machines in 2005, this technology has seen the rise and fall of many companies; however, following the acquisition of Solexa, Illumina's platforms have consolidated as the most commonly used. The reason behind such wide adoption of Illumina's systems is the large volume of information obtained from a typical sequencing run (*i.e.* sequencing depth), which, at a good ratio with the cost, compensates for the lower accuracy compared to other competitors [Mardis, 2013]. Thus, although microarrays can still be a

cheaper option to perform routine differential expression analysis at the gene level, the larger scope of applications and the decrease in the costs of sequencing (just announced to have reached the target of 1,000$ per human genome by Illumina while writing this thesis) explain the increasing popularity of RNA-seq.

In this section, I introduce the typical steps required to sequence a transcriptome with an Illumina platform, since this is the one that has been used to produce all the datasets analysed here. Moreover, I provide a detailed description of the most commonly used methods to study the transcriptome composition from RNA-seq data, with special emphasis on the analysis approaches used within the different chapters.

1.3.1 A typical sequencing workflow

The first step in transcriptome sequencing is library preparation, and consists of obtaining the starting material and converting it into a cDNA library that can be loaded into the sequencing machine (**Figure 1.5**) [van Dijk et al., 2014]. Following RNA extraction, the RNA species of interest are typically enriched through either polyA selection or ribodepletion. In both cases, the aim is to diminish the concentration of rRNAs, *i.e.* the most abundant species of RNA in the cell. With the first method, this is achieved through the use of oligo-dT beads, which enable the specific extraction of polyAdenylated RNAs, hence ensuring a good representation of mRNAs (**Figure 1.5** - step 1). Conversely, ribodepletion relies on the use of ribonucleases to specifically digest rRNAs, and has the advantage of not restricting the analyses to a specific type of RNA. Indeed, the term total RNA is typically used to refer to datasets produced with such protocol, while those obtained with the former method are commonly known as polyA-selected. Due to the simpler protocol and its lower price, polyA selection emerges as the most popular choice amongst the currently available RNA-seq datasets, with the notable exception of those studies aimed at characterising non-coding RNA species, which typically lack a polyA tail. The extracted RNA is then fragmented via hydrolysis with divalent cations and retro-transcribed into double stranded cDNA by using random hexamer primers, since the sequence of the obtained fragments is not known at this point (**Figure 1.5** - step 2). These steps are followed by the ligation of adapter sequences at both ends of each cDNA fragment (**Figure 1.5** - step 3). Such adapters satisfy two different purposes: on the one hand, they enable the hybridisation of those fragments into the flow cell,

where the sequencing takes place; on the other hand, they serve as primers for the sequencing reaction. Then, the resulting cDNA fragments are size-selected through gel electrophoresis to fit within the range required by the sequencing machine (typically 300-500 bp). Fragments outside this range will be missed; hence the existence of alternative protocols for the study of small RNAs [Zhuang et al., 2012]. Finally, the cDNA library is amplified by PCR.

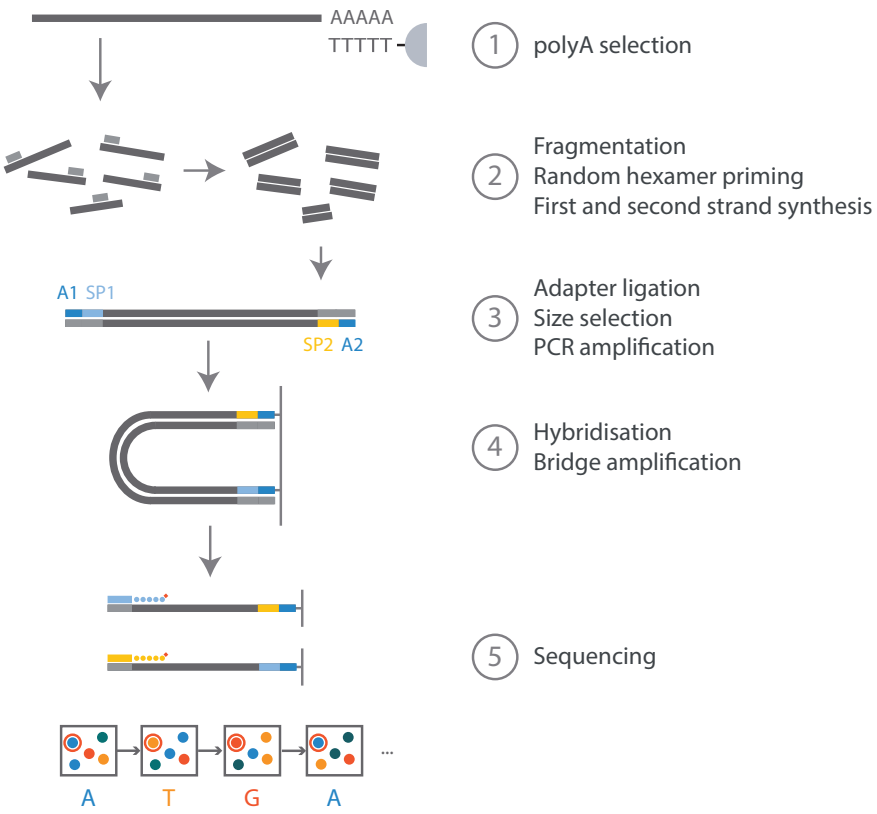


Figure 1.5| Overview of library preparation and sequencing steps in an Illumina platform. A typical paired-end workflow is illustrated here, which consists of ligating different adaptors at each end of the initial cDNA molecule. This enables sequencing each cDNA fragment from both ends, in two separate reactions, and has further advantages for the downstream bioinformatic analyses compared to single-end approaches. Adapted from Mardis [2013].

Once the library preparation procedure has finished, samples can be loaded into a flow cell for sequencing [Mardis, 2013]. Such flow cell is saturated with adapters that are complementary to the ones ligated at both ends of the cDNA fragments, consequently promoting the hybridisation of the denatured double

strand molecules. After this step, the starting material needs to be amplified once again in order to increase the signal for the sequencing reaction, this time through bridge amplification (**Figure 1.5** - step 4). Such process consists of the synthesis of fragments that are complementary to the hybridised cDNA molecules, which will in turn bend and hybridise with adjacent adapters, thus enabling subsequent rounds of synthesis. As a result, a large number of clusters with identical sequences will be formed, now ready to undergo sequencing. Illumina platforms rely on sequencing by synthesis to read the base pair composition of each cDNA cluster (**Figure 1.5** - step 5) [Bentley et al., 2008]. This reaction is based on the use of modified versions of the four bases, which differ from the standard nucleotides in the fact that they incorporate a reversible terminator, as well as a fluorescent dye. Hence, during each sequencing cycle, and following the addition of the necessary reagents, elongation will be blocked after a successful base incorporation, the identity of which can be recorded by measuring its fluorescent signal. Repetition of this process will lead to a set of images, which after interpretation with a base calling software, will be converted into a set of sequences or reads [Das and Vikalo, 2013]. Such reads represent the set of molecules expressed in the initial sample, and their length corresponds to the number of cycles performed during the sequencing reaction. Eventually, the obtained sequence information, together with the probability of a wrong base call at each given position of the read (*i.e.* Phred score), are stored in a plain text file in FASTQ format [Cock et al., 2010].

RNA-seq is not such an established technology as microarrays, and in spite of its many advantages, it still has some challenges. For example, regarding library preparation, it is known that the random hexamer priming step is not as random as initially proposed, since certain fragments have been observed to be preferentially converted to cDNA due to sequence composition [Hansen et al., 2010]. In the same category of sequence-dependent biases, the PCR amplification step has also been described to lead to differential amplification of fragments with higher or lower GC content [Benjamini and Speed, 2012], and it is known that failure to block the elongation reaction or to remove the fluorescent dye during the sequencing step can lead to wrong base calls [Metzker, 2010]. In most cases, the identification of such biases has been accompanied by the introduction of alternative protocols or analysis methods to overcome them. For example, several algorithms now try to take into account potential biases derived from the random

hexamer amplification step (*e.g.* Cufflinks [Trapnell et al., 2010], MMSEQ [Turro et al., 2011]). Alternative library preparation methods have also been proposed to account for PCR bias, whereby random barcodes (*i.e.* molecular identifiers) are used to quantify the absolute number of molecules [Shiroguchi et al., 2012]. Finally, some downstream analysis algorithms also incorporate information on the probability of a wrong base call at a given position of the read, as reported by the Phred score (*e.g.* Kim et al. [2013]).

On the other hand, alternative library preparation strategies can also add further information to the experiment. This is the case of strand-specific protocols, which are able to provide information on the strand from which each read originates [Levin et al., 2010]. Similarly, multiplexing emerges as a widely used approach to optimise the amount of data that can be obtained from each sequencing run, by enabling pooling of several samples into a single lane of the flow cell through the use of sequence identifiers (*i.e.* sample-specific barcodes) [Wong et al., 2013b]. Lastly, a very common strategy to overcome limitations on the read length and try to span larger regions consists of sequencing each cDNA fragment from both ends (*i.e.* paired-end sequencing, as opposed to the single-end strategy), which can be achieved through the use of modified adapters (**Figure 1.5** - step 3) [Mardis, 2013].

1.3.2 Read mapping strategies

The next step in a typical RNA-seq analysis pipeline consists of identifying, for each read, the genomic region from which it has originated. In RNA-seq, this task is equivalent to discovering the loci that are expressed in a given sample. In general, two different strategies exist to perform this task: on the one hand, reads can be aligned to the reference genome or transcriptome, provided that such information is available for the species of interest; on the other hand, they can be directly assembled into contigs (*i.e.* contiguously expressed regions) with the aim of reconstructing the set of expressed transcripts. The first strategy constitutes a much simpler approach, and it is typically the method of choice when working with model organisms.

Independently of the strategy used, read mapping is typically the most time consuming step of the analysis workflow, and the available tools make use of heuristic parameters such as the maximum number of allowed mismatches per read in order to speed up this task. Such processing can lead to information loss given a

decrease of quality at the 3' end of the read, which emerges as a common profile when working with Illumina platforms due to the increased difficulty in interpreting the fluorescent signal as sequencing cycles accumulate [Minoche et al., 2011]. Thus, in order to avoid such reads being discarded, it is often useful to first perform a quality control and pre-filtering step, whereby read sequences can be shortened (*i.e.* trimmed) in terms of their quality (*e.g.* Andrews [2010]). Similarly, reads with overall low quality can be also removed, in order to speed up the subsequent mapping process.

1.3.2.1 Alignment to the genome or transcriptome

A commonly used approach in the cases where a reference genome is available, is to align the reads directly to that sequence. Similarly, reads can be aligned to the transcriptome instead, provided that a good annotation exists. The main advantage of using this second strategy is that the alignment task is simplified due to the lack of intronic sequences; but this comes at the price of limiting the number of downstream analysis that can be performed (*e.g.* alignment to the transcriptome is not compatible with the identification of novel expressed regions nor the study of intronic expression levels). Thus, a good compromise is the use of hybrid approaches, as implemented in TopHat [Kim et al., 2013].

TopHat is a read mapping tool specially intended for RNA-seq data, since it enables alignment of the reads to the genome while taking into consideration the existence of splice junctions (**Figure 1.6**). It is based on Bowtie [Langmead and Salzberg, 2012], an independent algorithm for the alignment of short reads, and its main strength is the ability to detect exon-exon junctions without the need for any *a priori* knowledge on the annotation. However, the search can be simplified by providing such information, and in that case TopHat will first attempt to map the reads to the derived transcriptome. Those that fail to align will be subsequently queried against the genome (**Figure 1.6** - step 1). Alternatively, reads can also be mapped to the genome directly (**Figure 1.6** - step 2). In both cases, the goal is to assemble the initially aligned reads into exons, which might eventually become connected through spliced alignment (**Figure 1.6** - step 2). Reads that fail to align in this initial phase, as well as those that map with low alignment scores, are subsequently used to build a database of possible splice junctions, by splitting them into smaller segments and re-aligning those independently (**Figure 1.6** - step 3). In this context, a splice junction is reported whenever a read appears to span multi-

ple exons, *i.e.* in the cases in which an internal segment fails to align, or when two consecutive segments from the same read do not align contiguously on a given genomic locus. Next, the identified splice sites and their flanking sequences are concatenated into a novel transcriptome, which is then used to re-align the set of unmapped reads (**Figure 1.6** - step 4). In the case of paired-end data, each read is processed separately, and the alignments obtained are evaluated in a final phase by taking into account additional sources of information such as fragment length and orientation of the reads. Finally, all the information gathered during the mapping process is reported in SAM/BAM format [Li et al., 2009].

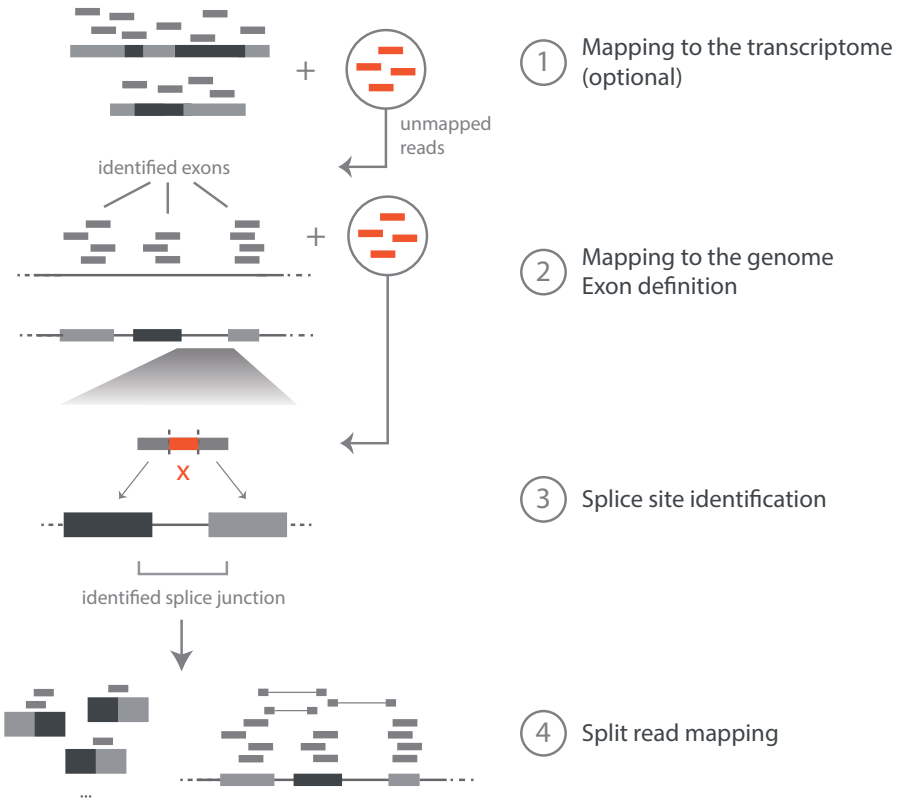


Figure 1.6 | Overview of the mapping algorithm implemented in TopHat. In the presence of an annotation file, TopHat uses a hybrid approach to uncover the genomic loci from which the detected reads could have originated. Alternatively, TopHat can directly align the reads to a reference genome. In both cases, the first step consists of identifying a set of expressed exons, and this is followed by the detection of splice junctions by using information from those reads that span multiple exons. Adapted from Kim et al. [2013].

1.3.2.2 *De novo* assembly

De novo assembly emerges as an advantageous strategy in the cases where the species of interest lacks a reference genome. Additionally, it can be used in situations where the genome composition of a given sample is expected to differ largely from that of the reference assembly (*e.g.* cancer samples). The goal here is to assemble the reads into sets of expressed regions (*i.e.* contigs), by relying on their overlap. Nonetheless, the short read length adds to the non-triviality of the task, and even though the use of paired-end data can simplify this process, lowly expressed regions are often difficult to solve. In terms of available software, Trinity [Grabherr et al., 2011] emerges as the most popular tool to perform this task; however, such methods are not used in this thesis and are covered elsewhere [Martin and Wang, 2011].

1.3.3 The estimation of expression levels

Once the reads have been assigned to a specific location in the genome or transcriptome, the next step in the RNA-seq analysis pipeline consists of estimating expression levels for the features of interest, typically genes and transcripts. Similarly to the scenarios encountered during the mapping step, the quantification of expression levels can be achieved by relying on existing information, but it can also be performed independently from any annotation, thus enabling *de novo* identification of transcribed regions (*i.e.* novel genes or unnanotated transcripts within known gene loci).

1.3.3.1 Gene expression levels

When working at the gene level, and provided a complete annotation exists, abundance estimation can be easily achieved by counting how many reads overlap each given locus (**Figure 1.7a**). Such count-based approach constitutes the starting point for many downstream analysis algorithms (*e.g.* DESeq2 [Love et al., 2014], DEXSeq [Anders et al., 2012]), and can be easily performed with the popular tool htseq-count [Anders et al., 2014]. However, despite this apparent simplicity, there are some challenges that need to be considered. First, in order not to over-estimate expression levels, reads that map to multiple locations in the genome, and which arise from repetitive or duplicated loci, need to be handled with care. In this situation, htseq-count adopts the most conservative approach and discards them, but other alternative strategies have been proposed in order to attempt to keep

the information from such multi-mapping reads. Generally, these consist of uniformly distributing them to all the mapped positions (*e.g.* Trapnell et al. [2010]), or probabilistically assigning them depending on the coverage at each mapping locus (*e.g.* Trapnell et al. [2010]; Turro et al. [2011]; first proposed by Mortazavi et al. [2008]). Second, special attention is required in the case of overlapping features. htseq-count offers several execution modes to deal with this scenario, even though in some cases reads remain ambiguously assigned (**Figure 1.7b**). Finally, despite not being intended for *de novo* quantification, htseq-count also gives the user some flexibility on how strictly the provided feature coordinates should be taken into account (**Figure 1.7b**).

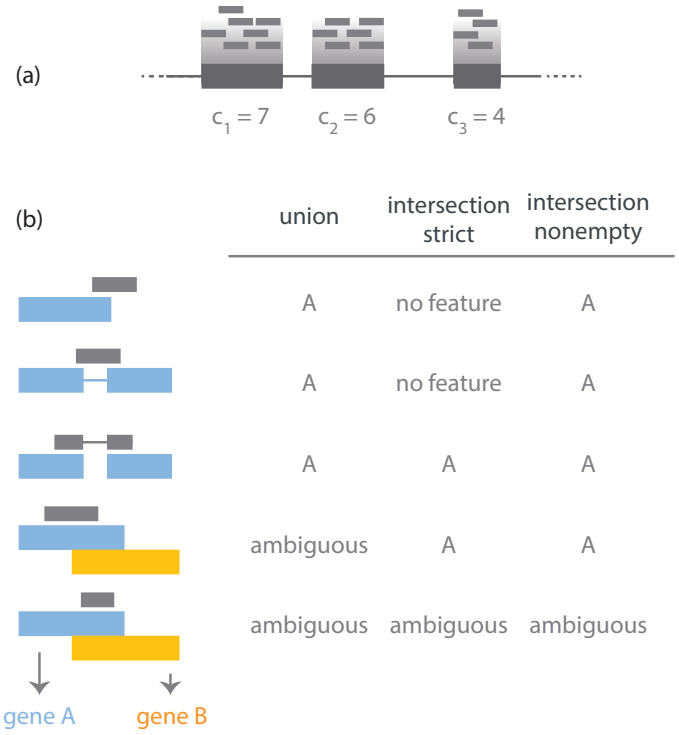


Figure 1.7 | Overview of htseq-count.
(a) Illustration of the read counting concept. Expression estimation with htseq-count consists of counting the reads that overlap with the features of interest. In this example, any reads that fall outside the grey areas will not be considered.
(b) The three different execution modes available in htseq-count. htseq-count provides different counting modes to rescue reads that do not strictly overlap with the provided coordinates. These modes differ in how strictly the annotation is taken into account and in the behaviour adopted in the case of overlapping features. Adapted from Anders et al. [2014].

Alternatively, gene expression can be calculated after estimation of transcript expression levels, by aggregating the corresponding individual transcript abundances, as implemented for example in Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011].

1.3.3.2 Transcript expression levels

On the other hand, the task of estimating expression levels becomes far more complicated when focusing on individual transcripts, since many reads will overlap with exons that are shared across multiple isoforms of the same gene. In this scenario, the question translates into attributing reads to specific transcripts, and further inference approaches are needed. The available algorithms typically rely on different sources of information in order to probabilistically estimate transcript expression levels, the most valuable one being those reads that map uniquely to one of the annotated transcripts within the loci. Moreover, reads that span two different exons (*i.e.* split reads) become especially informative. For example, splice junctions that involve cassette exons tend to provide unambiguous support for their inclusion or skipping. This is where paired-end information becomes most relevant: sequencing both ends of the initial cDNA fragment facilitates covering larger genomic regions, thus increasing the probability that a given read pair is mapped across different exons (*i.e.* spliced reads). Similarly, information on the fragment length distribution can also be used to deconvolute ambiguous assignments, by attributing a lower likelihood to those that would require extreme distances between the paired reads.

One of the most popular tools to estimate transcript expression levels is MISO [Katz et al., 2010], which formulates this task as a Bayesian inference problem [Beaumont and Rannala, 2004], whereby the goal is to find a probability distribution (the posterior) over transcript abundances (Ψ) given the observed RNA-seq data (**Figure 1.8**). Such distribution can be computed in terms of two quantities: the expectation about the value of Ψ before observation of the reads (the prior, set by MISO as a uniform distribution), and the probability of observing the data given a fixed value of Ψ (likelihood of the reads). Thus, following sampling across the space of Ψ values, all possible assignments of every read to each isoform are probabilistically evaluated, and this information is subsequently used to refine the search of the optimal set of Ψ values that best explain the observed data. Finally, an estimate of transcript abundances is obtained by calculating the mean over the

computed posterior distributions, and confidence intervals are also calculated as a measure of the certainty of the estimate.

Similar approaches are taken by other alternative methods, including Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011], both of which are also used in this thesis. The main difference among the mentioned methods relies on the implementation of the inference approach, as well as the type of input required. Similarly to MISO, Cufflinks requires the reads to be mapped to a reference genome, but relies on a frequentist approach to find the expression levels that best explain the observed data, which does not allow quantification of the uncertainty around the obtained expression estimates. On the other hand, MMSEQ adopts a Bayesian model similar to the one of MISO, but requires mapping to the transcriptome, which limits the scope of the downstream analysis that can be performed. Furthermore, both MMSEQ and Cufflinks accommodate known sequence biases in their models, and are also able to retain information from reads that map to multiple genes.

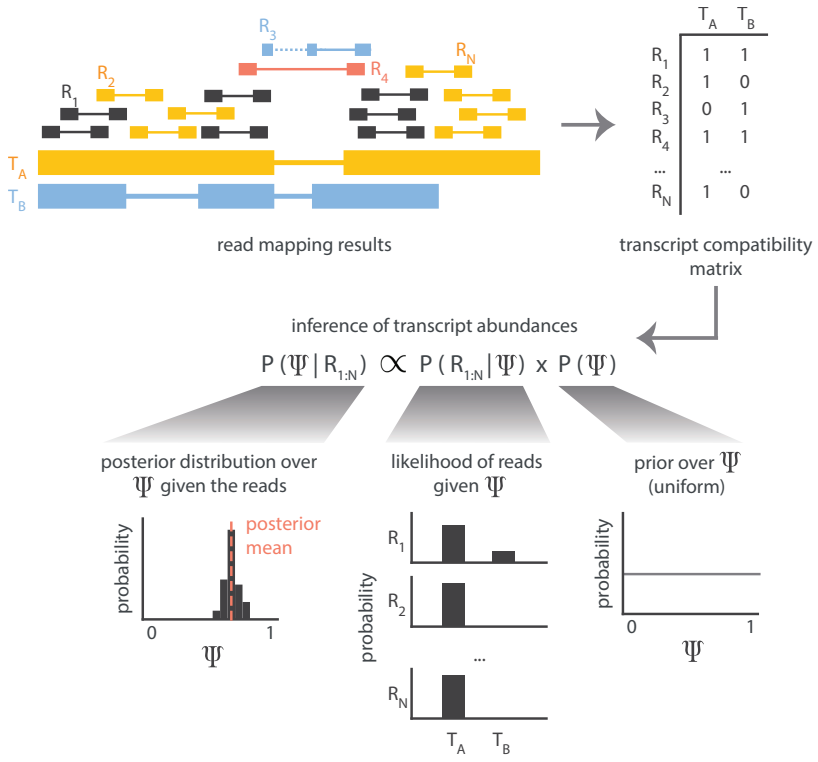


Figure 1.8| Overview of the analysis workflow implemented in MISO for the estimation of transcript abundances. After alignment to the genome, MISO evaluates the compatibility of each read with all the transcripts annotated within a given gene. For example, in the scenario depicted here both read 2 (R_2) and read N (R_N) can only be detected if transcript A (T_A) is expressed, whilst read 3 (R_3) uniquely supports transcript B (T_B). Inference of expression levels is then done by calculating a probability distribution (the posterior) over such expression (Ψ) given the reads. Following Bayes' rule, this distribution can be obtained from the product of two terms: the likelihood of obtaining the observed set of reads given a fixed value of Ψ and the expectation on the value of Ψ before observation of the data (the prior). Hence, the inference problem translates into sampling from the space of Ψ values and evaluating all possible assignments of each read to each transcript. For example, given the larger number of reads that support the expression of transcript A in comparison to transcript B, higher expression of the latter will be probabilistically penalised, and this will contribute to the preferential assignment of ambiguous reads to the former. Fragment length information can also be used to deconvolute ambiguous assignments, as it is the case for read 4 (R_4): assigning such read to transcript B would imply an unusual distance between the paired reads, hence increasing the likelihood that it can be explained by transcript A. Finally, the overall probability of observing the reads given the evaluated Ψ value is obtained by combining the information from all reads, and this information is further used to calculate the posterior distribution. Adapted from Wang et al. [2010].

1.3.3.3 *De novo* transcript identification

One of the main advantages of RNA-seq over microarrays is the possibility to gather information on novel expressed loci in a more high throughput manner. In this context, Cufflinks [Trapnell et al., 2010] emerges as one of the most popular tools to achieve this task, thus complementing its aforementioned quantification capabilities (**Figure 1.9**). By relying on the output provided by TopHat [Kim et al., 2013], the Cufflinks assembler first identifies the expressed loci (*i.e.* genes) present in a given sample. Then, for each of them, it evaluates the observed data in search of a set of incompatible reads, *i.e.* reads which have necessarily originated from different transcripts. This step is followed by the construction of an overlap graph, whereby each read represents a node and each edge is used to connect compatible reads. Finally, Cufflinks tries to identify the minimum set of paths (*i.e.* transcripts) that explain such graph.

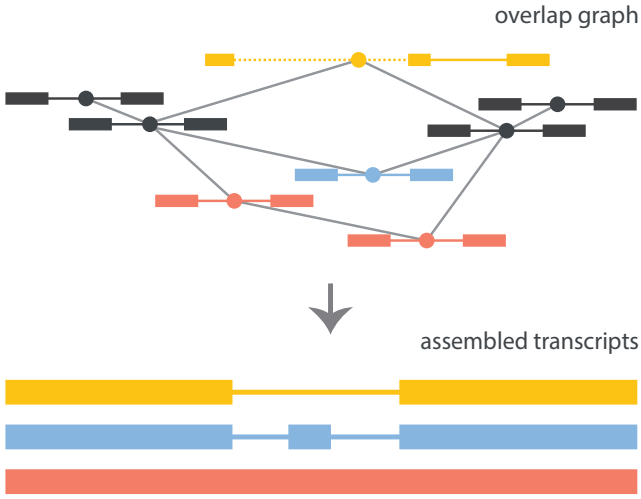


Figure 1.9 | Overview of the *de novo* transcript identification algorithm implemented in Cufflinks. Three different incompatible sets of fragments exist in the example depicted here (*i.e.* yellow, blue, red). Black reads represent those that are compatible with any of the sets. Following the construction of an overlap graph that indicates the possible connections amongst the observed fragments, Cufflinks assembles the data into the minimum set of paths required to explain such graph. The identified transcripts can then be used for subsequent downstream analyses, including estimation of transcript expression levels. Adapted from Trapnell et al. [2010].

Alternatively, Cufflinks can also be used in conjunction with the existing annotation (*i.e.* annotation based transcript assembly). In this scenario, the annotated transcripts are used to generate artificial data points that are combined with the observed data during the assembly process, hence serving as a guide. Following transcript assembly, the novelty of the obtained transcripts is evaluated by comparing them to the annotation, and those that differ are reported as novel.

Overall, and similarly to *de novo* read assembly, the task of transcript assembly is not a trivial one, although it becomes simplified by the existence of mappings to the genome. Similarly to the situation encountered in the former scenario, lowly expressed regions are difficult to analyse, given that in those cases the algorithm is less likely to find a unique solution for the constructed graph. Finally, alternative start and end sites also become difficult to characterise, since all the paths are extended to the maximum.

1.3.4 Read count normalisation

Independently of the quantification approach followed, the result from such step is going to be an estimate on the number of reads that can be attributed to a certain feature, further referred to as counts. These counts will be proportional to the expression levels of the feature of interest; however, they will also depend on the length of the feature and the sequencing depth of the experiment (*i.e.* the total number of sequenced reads). In addition, further experimental biases have also been detected to have an impact on the counts detected in certain loci, as it is the case for the previously mentioned sequence-dependent biases. Altogether, these observations illustrate the need for normalisation in order to enable the comparison of read counts across different samples and features.

One of the measures commonly used to report expression levels derived from RNA-seq data is the Reads/Fragments per Kilobase per Million mapped reads (RPKMs or FPKMs, in the case of single-end or paired-end data, respectively) [Mortazavi et al., 2008]:

$$\hat{\mu}_{ij} = \frac{k_{ij}}{N_j l_i} \cdot 10^9$$

where:

$\hat{\mu}_{ij}$ = normalised expression for gene i in sample j
k_{ij} = observed counts for gene i in sample j
N_j = total number of reads in sample j (sequencing depth)
l_i = length for gene i

Given that this measure takes into account both the length of the feature of interest and the total number of mapped reads in the dataset (*i.e.* sequencing depth), it has become established as an intuitive measure of expression levels. However, this method is based on the assumption that the overall RNA levels are similar across samples, and hence it might fail to properly estimate the normalisation factors in cases where the compared libraries differ in their composition [Robinson and Oshlack, 2010]. For example, let us imagine two samples that express a common set of genes at similar levels, and let us consider an extra small set of highly expressed genes in one of them. Since the sequencing step can be understood as a sampling process, where it is more likely to detect reads from genes with high expression levels, the signal from commonly expressed genes will be lower in the latter sample, provided that both are sequenced at a similar depth. Hence, using the above mentioned normalisation method would lead to the interpretation that most genes undergo changes in expression across conditions; whilst the observed differences could be better explained by the isolated differential expression of the few non-overlapping genes (**Figure 1.10**).

The above described scenario evidences the need for more robust normalisation methods, especially when the goal is to compare across libraries (*e.g.* in downstream analysis such as differential expression/splicing). An example of those methods is the one provided within the DESeq2 Bioconductor package [Love et al., 2014]. Such algorithm starts by calculating a geometric mean for each gene in order to capture the variability of the observed measurements across all the libraries (similar to obtaining a reference sample). Then, these values are used to normalise the initial counts, and finally, the library-specific normalisation factors are obtained from the median of the calculated ratios:

$$s_j = \underset{i:\ k_i^R \neq 0}{median} \frac{k_{ij}}{k_i^R}$$

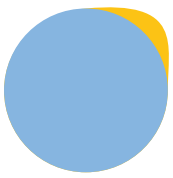
where:

s_j = size factor for sample j

k_{ij} = observed counts for gene i in sample j

k_i^R = geometric mean for gene i across the m samples: $(\prod_{v=1}^{m} k_{iv})^{1/m}$

Other tools (*e.g.* Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011]) enable also the correction of sequence-dependent biases, by attributing a weight to each position in the expressed loci based on its sequence context. The calculated weights are then used during the abundance inference step in order to model the non-uniform location of reads along the transcripts [Li et al., 2010; Roberts et al., 2011].



	sample A	sample B
gene 1	100	80
gene 2	100	80
...	...	...
gene 100	100	80
gene 101	0	2,000
total number of reads	10,000	10,000

Figure 1.10 | Limitations on the use of RPKMs for differential expression analysis. In this example, sample A and B express a common set of genes at similar levels; however, sample B also contains a highly expressed gene that is not present in the former. If both samples are sequenced at the same depth, the observed counts for the common set of genes will be lower in sample B, given the limited number of reads. In the context of differential expression analysis, the RPKM normalisation method would lead to the interpretation that all genes are differentially expressed, since it assumes homogeneity in library composition. In the scenario represented here such assumption does not hold, and the observed differences are better explained by the isolated differential expression of the gene unique to sample B. This highlights the need for more robust normalisation methods when comparison across libraries is attempted.

1.3.5 Differential expression

One of the most common uses of RNA-seq data is the assessment of differences in expression levels across conditions. Provided the corresponding counts have been obtained, such analysis can be performed both at the gene and transcript levels (differential gene/transcript expression), and one of the most popular tools to achieve that is the Bioconductor package DESeq2 [Love et al., 2014].

In general terms, DESeq2 relies on the use of Generalised Linear Models (GLMs) of the Negative Binomial (NB) family in order to address the significance of the detected changes in expression levels. The implemented analysis workflow first consists of normalising the observed counts in order to enable their comparison across libraries (**Figure 1.11** - step 1), as covered in the previous section. Next, for each gene, an estimate on the amount of variability that can be expected on the measurements from biological replicates is calculated (**Figure 1.11** - step 2), and finally, the differential expression test is performed (**Figure 1.11** - step 3).

As with any counting process, one would not expect the detected counts for a given gene to be exactly the same across all observations from a single condition. Hence, the underlying question in differential expression analysis is whether the counts observed across the two evaluated conditions are similar enough to be derived from the same distribution (null hypothesis), or whether they are better explained by two separate ones (alternative hypothesis). Given the nature of the data obtained from RNA-seq experiments, the Poisson distribution was first proposed to model noise intrinsic to the counting process [Marioni et al., 2008]. However, it was soon shown that while this approach works well for technical replicates, it underestimates the variability in measurements across biological replicates [Anders and Huber, 2010; Robinson et al., 2010]. As a result, the negative binomial distribution has been widely adopted to account for such over-dispersion:

$$k_{ij} = NB(\mu_{ij}, \sigma_{ij}^2)$$
$$\mu_{ij} = s_j q_{ij}$$

where:

k_{ij} = observed counts for gene i in sample j
μ_{ij} = distribution mean for gene i in sample j
σ_{ij}^2 = dispersion for gene i in sample j
s_j = size factor for sample j
q_{ij} = quantity proportional to the concentration of cDNA fragments for gene i in sample j

The identification of the amount of variation across biological replicates is an essential step in the aforementioned workflow, since it enables for the evaluation of the significance of any changes detected. However, because of the low number of replicates typically available in RNA-seq experiments, such variation cannot be directly calculated, and needs to be estimated from the data instead. Following the assumption that genes with similar expression levels have similar sample-to-sample variance, DESeq2 obtains gene-specific variance estimates by taking into account not only the observed dispersion for each given gene, but also that of all other genes. This is achieved by fitting a regression curve to the data (*i.e.* average normalised counts *vs.* observed dispersion), which is subsequently used to modify the observed dispersion values.

Finally, by further decomposing the mean into a function of independent variables (*i.e.* the covariates), it is possible to take all known sources of variation into account:

$$log_2(\mu_{ij}) = \sum_r x_{jr} \beta_{ir}$$

where:

μ_{ij} = mean for gene i in sample j
x_{jr} = independent variable r for sample j
β_{ir} = coefficient for gene i and variable r

Altogether, the algorithmic approach behind DESeq2 consists of fitting the model defined in the aforementioned equations for both the null and alternative hypotheses (reduced *vs.* full model, respectively), followed by the evaluation of the significance of the coefficient of interest (**Figure 1.11**).

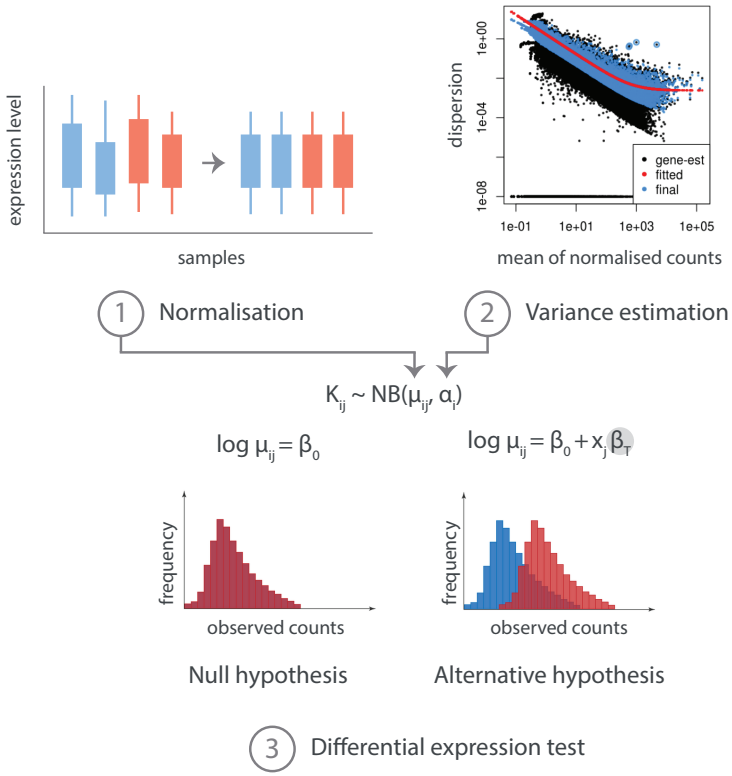


Figure 1.11 | Overview of the steps required for differential expression analysis using DESeq2. First, reads are normalised in order to enable comparison across libraries. Next, for each gene, an estimate of the amount of variability that can be expected across biological replicates is obtained. Given the typical low number of biological replicates in RNA-seq experiments, it is not possible to obtain such information directly from the data. Hence, DESeq2 relies on the observed dispersion from all genes instead (black dots), and by fitting a regression curve that explains the dependence of the dispersion on the mean (red line), further modifies the initial values through a process called shrinkage estimation of variability (blue dots). Finally, the obtained information is used to test the hypothesis that the observed counts originate from different distributions (alternative hypothesis). DESeq2 uses the negative binomial distribution to model both stochastic and biological noise, and it further relies on the use of GLMs to take all known sources of variation into account. Here, the independent variable x_j represents the experimental condition, and can be arbitrarily set to 0 in the case of controls and to 1 in the case of treated samples. Under the alternative hypothesis, this enables the existence of two different distributions that can explain potential differences in expression levels of the studied gene i. Conversely, under the null hypothesis, there is no need for such term, since in this case all counts arise from the same distribution. Hence, the general idea behind the differential expression test consists of deciding whether the inclusion of the variable x_j adds meaningful information to the model, and it translates into assessing the significance of the β_T coefficient (highlighted in grey).

Similarly to DESeq2, further available tools rely on the use of read counts in order to make assessments of differential gene expression (e.g. edgeR [Robinson et al., 2010] and baySeq [Hardcastle and Kelly, 2010]). These tools can also be used to study differential expression at the transcript level, but a common alternative approach consists of relying on the algorithms implemented in the framework of transcript abundance estimation, since those are able to take into account the uncertainty in the read assignment process. For example, this is the case with Cuffdiff2 [Trapnell et al., 2013] and MMDIFF [Turro et al., 2014], which can be executed following estimation of transcript expression levels with Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011], respectively. Moreover, and consistent with the Bayesian model adopted, MMDIFF is also able to make use of the uncertainty in the expression estimates, thus adding further sensitivity to the differential transcript expression analysis.

1.3.6 Differential splicing

In the previous section, I have discussed briefly several approaches for the assessment of differential transcript expression. However, differences in absolute transcript abundance are not necessarily indicative of differential splicing (**Figure 1.12a**), and alternative analysis strategies are preferred when the focus lies on the latter. In general terms, changes in splicing patterns can be assessed through the identification of either differential exon usage (DEU) or differential transcript usage (DTU) events (**Figure 1.12b** and **c**, respectively), with advantages inherent to both approaches. On the one hand, exon-centric analysis strategies are completely independent from isoform reconstruction efforts, thus avoiding the uncertainty intrinsic to that task. Furthermore, while those rely on the existing annotation, such dependence is limited to the exonic coordinates, and this approach still enables the indirect identification of novel transcripts (*i.e.* novel alternatively spliced isoforms). On the other hand, the results from such exon-centric analysis are often difficult to interpret, and in this context transcript-centric analysis strategies emerge as an attractive alternative.

Interestingly, in terms of algorithm development, much of the effort has been focused on the identification of differential transcript expression events, with limited availability of tools for the study of changes in splicing. Amongst those, the Bioconductor package DEXSeq [Anders et al., 2012], was the first tool to account for

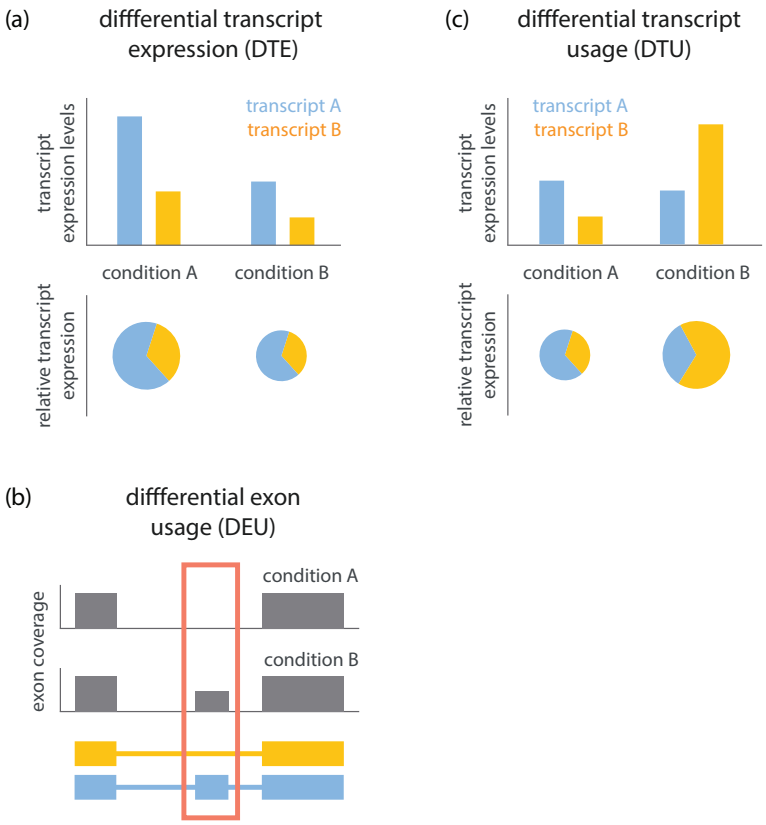


Figure 1.12 | Strategies for the study of changes in the abundance of alternative transcripts.
(a) Differential transcript expression concept. Differential transcript expression is analogous to differential gene expression, and does not necessarily imply differences in splicing.
(b) Differential exon usage concept. Differences in the read coverage for a given exon, relative to changes in the number of reads that overlap the other exons within the same gene, can be used as an indicator of differential splicing.
(c) Differential transcript usage concept. Differential transcript usage refers to those cases where there is a change in the transcript relative abundances, which is not necessarily linked to an overall change in expression levels. It constitutes the most direct strategy for the study of differential splicing.

biological variation in the analysis, a vital requirement for robust testing. Briefly, by relying on the same algorithmic principles as the aforementioned DESeq2, this method enables the identification of significant differences in the proportion of reads that overlap each exon, relative to the total number of reads that overlap the corresponding gene (DEU events; **Figure 1.12b**). On the other hand, the recently introduced tool MMDIFF [Turro et al., 2014] provides a method for the analysis

of DTU events (**Figure 1.12c**). MMDIFF is based on the use of Bayesian mixed models, whereby the uncertainty in transcript expression estimates can be incorporated into the regression models used for testing, thus improving the power to detect the events of interest. Altogether, the scarcity of tools to deconvolute differences in splicing from differences in expression, together with the lack of methods to infer the functional impact of the identified events, evidence that the computational pipelines for the analysis of RNA-seq data are still not completely established.

1.4 Aims of the thesis

The work presented in this thesis focuses on the use of RNA sequencing for the high throughput study of alternative transcript products in human samples. Overall, the goal is to improve the current understanding of splicing by addressing the following questions:

- What is the extent of transcriptome diversity? Are specific alternative transcripts preferentially produced within a given gene?
- How prevalent are changes in splicing patterns in cancer? How can we assess the potential functional impact of such changes?
- How do core spliceosomal factors participate in the regulation of splicing? What are the effects of disrupting such regulation on dynamic cellular processes such as cell division?
- Can the differential splicing events identified from transcriptomics data be recapitulated at the protein level?

The present chapter has provided an general introduction to the two central concepts behind this thesis, *i.e.* the splicing reaction and RNA sequencing. Further introductory remarks relevant to each of the aforementioned questions will be covered in the following chapters.

Chapter 2

The extent of transcriptome diversity

The number of transcripts annotated in the human genome exceeds by far the number of genes, and evidence collected from RNA-seq experiments during the last few years indicates that the vast majority of those transcripts can be detected as expressed across a range of conditions [Pan et al., 2008; Wang et al., 2008]. This chapter aims to investigate the extent to which transcriptome diversity is maintained in a given sample when considering transcript relative abundances within each gene. In the first part, I evaluate whether, in a given sample, gene expression tends to be dominated by one transcript, as opposite to observing similar expression levels for several of them. Next, I explore different analysis strategies in order to increase the robustness of the obtained results. Finally, I dissect further the biological impact of the reported findings by incorporating information on transcript biotypes.

All the computational analyses described here have been performed by myself under the supervision of Dr. Alvis Brazma. Dr. Adam Frankish and Dr. Jennifer Harrow from the Wellcome Trust Sanger Institute contributed to the results described in section 2.2.3, as detailed in the Methods, and also provided general feedback on the project.

Publications derived from this chapter

- Gonzàlez-Porta, Frankish, Rung, Harrow & Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology* 14, R70 (2013).
  The present chapter includes a relevant portion of text from this manuscript.
- (invited talk) Eukaryotic mRNA processing 2013, CSHL.
- (poster prize) Integrative RNA Biology 2013 (ISMB/ECCB satellite meeting), Berlin.
- (poster) Genomics Medicine in the Mediterranean 2013, Crete.
- (poster) European Conference on Computational Biology 2012, Basel.

2.1 Introduction

Although there are less than 22,000 protein coding genes known in the human genome, they are transcribed into over 140,000 different transcripts (Ensembl release 66 [Flicek et al., 2012]), more than 65% of which have protein coding potential and thus may contribute to protein diversity. Following the introduction of RNA-seq, significant evidence has accumulated showing that over 95% of multi-exon genes have several alternative splice-forms expressed [Pan et al., 2008; Wang et al., 2008], and that transcript expression is regulated [House and Lynch, 2008; Smith et al., 2008]. On the other hand, focusing on expressed sequence tag (EST) data, Taneri et al. [2011] predicted that there is a single dominant transcript per gene in primary tissues. Recently, the ENCODE project showed that indeed, most genes have a major transcript in cell lines, although at the same time noted that "genes tend to express many transcripts simultaneously, and as the number of annotated transcripts per gene grows, so does the number of expressed transcripts" [Djebali et al., 2012]. Despite these observations, it is still unclear if and to what extent major transcripts are dominating the transcriptome, and what proportion of transcript diversity is likely to contribute to protein diversity. In addition, given the notable differences in gene expression between primary tissues and cell lines [Lukk et al., 2010; Waks et al., 2011], transcriptome analysis in cell lines can be extended to primary tissues only to some extent.

The present chapter aims to characterise the potentially coding transcriptome from a functional perspective. By focusing on protein coding genes, I show that, in primary tissues, almost 85% of the total mRNA from protein coding loci originates from major transcripts (76% in cell lines). Notably, these major transcripts are not always the longest possible for the gene (40% of the major transcripts in primary tissues and 30% in cell lines are not the longest annotated), nor always include the longest coding DNA sequence (CDS; approximately 50% of the cases in both tissues and cell lines). I also observe that the ratio of the number of expressed transcripts to genes in primary tissues is on average 1.12, which corresponds to just over one transcript per gene. I further distinguish between: (1) *major transcript* - the transcript with the highest expression level for a given gene; and (2) *dominant transcript* - a major transcript that is expressed at a considerably higher level than any minor transcripts of the gene. I show that most protein coding genes in most conditions have one dominant transcript;

e.g. for almost 80% of the expressed genes in primary tissues the major transcript is at least twice as abundant as the next one. Furthermore, I detect a clear separation in expression levels between major and minor transcripts, as well as a considerably higher overlap for major *vs.* minor transcripts when comparing them to a set of transcripts predicted to be translated into functional proteins by entirely independent means [Rodriguez et al., 2013]. Finally, I observe that for almost 20% of the studied protein coding genes (n = 18,450) the major transcript does not code for a protein, and that this percentage is considerably higher in the nucleus than in the cytosol.

These analyses have been performed using three different computational methods [Katz et al., 2010; Trapnell et al., 2010; Turro et al., 2011], and additionally, where sufficient coverage exists, alternative transcript abundances were assessed directly from the reads spanning unique exon junctions. Furthermore, simulated data [Griebel et al., 2012] was used to confirm that the methods can reliably distinguish between two hypothetical alternative scenarios - one dominant transcript per gene *vs.* several transcripts per gene expressed at similar levels. All those methods produced a consistent outcome, supporting the robustness of the presented conclusions.

2.2 Results

In this chapter, I describe the analysis of two different RNA-seq datasets in order to quantify and analyse the overall contribution of major transcripts to the potentially coding mRNA transcriptome, in comparison to minor transcripts. These include data on 16 primary human tissues from the Illumina Body Map dataset (BM), further referred to as tissue dataset, as well as from 5 ENCODE cell lines, further referred to as cell lines dataset (see Methods and **Table A.1**). Moreover, the latter also include data from different cellular compartments (*i.e.* whole cell, cytosol and nucleus), thus making it possible to compare across them. All the results mentioned in this section refer to the average across all the samples of the tissue dataset unless otherwise indicated.

2.2.1 Most protein coding genes express one dominant transcript

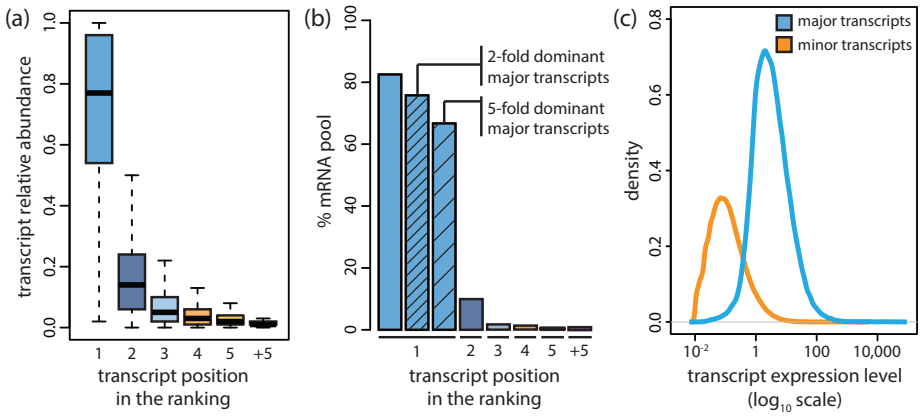
Similarly to the results reported in previous RNA-seq transcriptome studies, it is possible to detect more than one transcript for approximately 85% of the expressed genes (83.70% to 89.95%, SD = 1.84). In line with this observation, a total of 105,456 different transcripts can be detected as expressed in at least one tissue, which corresponds to approximately 90% of the studied transcripts (n = 117,759; see Methods). However, when quantifying all the annotated transcripts within a gene based on their relative abundance, the expression of most genes appears to be dominated by a single transcript in most conditions, rather than by a subset of similarly expressed transcripts (**Figure 2.1a** and **Figure B.1**). This observation is also supported by the fact that, in primary tissues, the ratio of the number of expressed transcripts to genes is 1.12 (0.98-1.40, SD = 0.11). Finally, analysis of the mRNA pool derived from protein coding loci revealed that major transcripts comprise approximately 85% of the coding mRNAs present in the cell (79.98% to 86.49%, SD = 2.17; **Figure 2.1b**).

In order to address the impact of these observations at the protein level, I plotted the distribution of expression levels for both major and minor transcripts (**Figure 2.1c**). This analysis led to the observation that minor transcripts have a tendency to be expressed below 1 FPKM, a threshold that has been suggested as the minimum expression required for protein detection [Hebenstreit et al., 2011; Vogel and Marcotte, 2012]. With the same goal in mind, I then evaluated the overlap between the major/minor transcript predictions obtained from the interpretation of the RNA-seq data and those obtained by an entirely independent method (APPRIS [Rodriguez et al., 2013]). Briefly, APPRIS aims at identifying principal isoforms amongst all the transcript annotated within a gene, by combining information on protein domains, structures and conservation across species. Such predictions were overlapped with two different sets of coding transcripts: on the one hand, recurrent major transcripts expressed above 1 FPKM; on the other hand, recurrent minor ones expressed below the same threshold. Focusing on common genes in these two scenarios (n = 6,082), the detected overlap was significantly higher for the first set compared to the second one (45.61% *vs.* 29.35%, respectively; Fisher's exact test p-value < $2.2 \cdot 10^{-16}$). This was also the case when taking into consideration only those genes that are expressed in all the tissues (n = 1,682, 59.93% *vs.* 22.06% overlap, respectively; Fisher's exact test

p-value < $2.2 \cdot 10^{-16}$). Altogether, these results suggest that major transcripts could be preferentially translated, even though minor ones may still play a functional role.

Next, I quantified major transcript dominance by calculating for every gene the ratio of the expression levels between the major transcript and the second most abundant one. Overall, such analysis revealed that 79% of the genes in the studied tissues (74.21% to 81.94%, SD = 2.16) have a two-fold dominant major transcript (*i.e.* expressed twice as much as the second most abundant one), and that for 56% of the genes (43.39% to 61.60%, SD = 3.50) the major transcript is five-fold dominant (**Figure 2.1d** and **Figure B.2**). This indicates that most genes tend to express one dominant transcript in a given sample. Similarly, dominant transcripts were estimated to account for most of the studied mRNA pool: 76.69% for a two-fold dominance (70.04% to 80.74%, SD = 3.48) and 67.47% for a five-fold dominance (59.97% to 73.83%, SD = 4.81; **Figure 2.1b**). GO enrichment analysis of genes that consistently express a five-fold dominant transcript across the 16 studied tissues indicated that they are functionally involved in metabolism and cellular respiration, protein transport and transcription regulation (n = 1,450; **Table B.1**). In addition, the comparison of the fraction of dominant *vs.* non-dominant major transcripts across different FPKM thresholds evidenced that the observed dominance is accentuated in highly expressed genes (**Figure 2.1e**). On the other hand, focusing on genes that tend to express several transcripts at a similar level, it was possible to identify 463 genes for which the major transcript was less than two-fold dominant in all the tissues analysed (only 17 for a five-fold dominance threshold). GO enrichment analysis of those revealed that they are involved in RNA splicing/processing, post-transcriptional regulation of gene expression and regulation of translation (**Table B.2**).

Similar analyses were also performed in the cell lines dataset, which includes different cellular compartments (**Figure B.1** and **Figure B.2**). In this context, I observed that major transcripts account for approximately 80% of the studied mRNA pool in the cytosol (77.20% to 83.66%, SD = 1.98; **Table B.3**), even though overall transcript dominance is less accentuated than in primary tissues: 69% (63.11% to 71.17%, SD = 2.40) of genes expressed a two-fold dominant major transcript and 42% (35.16% to 44.76%, SD = 2.90) a five-fold dominant one in the cytosol (**Table B.4**). Such differences could reflect higher transcription and splicing rates



(d)

	expressed genes	genes with a dominant major transcript	
		2-fold dominance	5-fold dominance
1 FPKM	10,410 (56.42%)	8,179 (78.51%)	5,864 (56.22%)
5 FPKM	4,671 (25.32%)	3,898 (83.64%)	3,077 (66.27%)
10 FPKM	2,486 (13.47%)	2,146 (86.54%)	1,794 (72.60%)

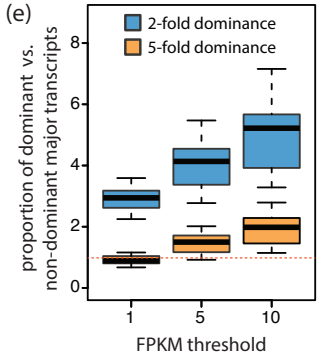


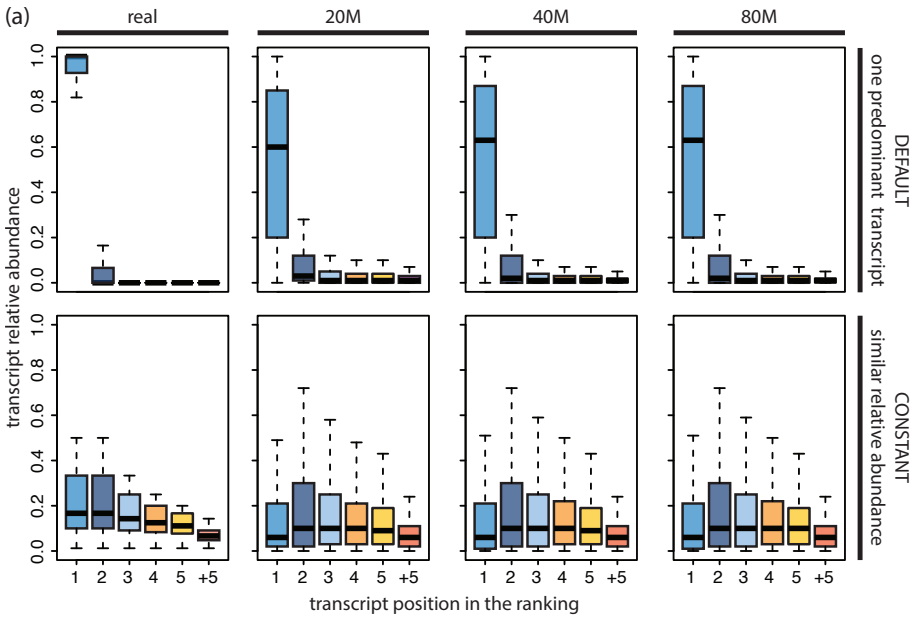
Figure 2.1| Most protein coding genes express one dominant transcript. All the results presented here correspond to the tissue data.
(a) Relative abundance of the subset of transcripts in each position of the ranking. For each gene, transcripts were ranked based on their relative abundances. There is generally one predominant transcript over the rest.
(b) Percentage of the studied mRNA pool explained by each category of transcripts. The mean percentage for all samples is represented here. Major transcripts represent approx. 85% of the studied mRNA population and were further classified into 2-fold and 5-fold dominant, depending on whether they are expressed twice or five times as much as the second most abundant transcript for the corresponding gene.
(c) Expression distribution for major and minor transcripts. A total of 31,902 transcripts are expressed above 1 FPKM in at least one tissue, including 26,641 different major transcripts.
(d) Average number of genes with dominant major transcripts. Different dominance ratios and gene expression thresholds were considered in the quantification.
(e) Proportion of dominant vs. non-dominant major transcripts. Values of the ratio above 1 indicate a higher proportion of dominant major transcripts with respect to non-dominant ones. Different dominance ratios and gene expression thresholds were considered for this analysis.

in cell lines, although they could also be due to technical variability between the two datasets. Given that these datasets have been generated by different laboratories, it is very difficult to distinguish between these two scenarios, but this question is revisited again in the next chapter, where the appropriate data is available.

2.2.2 The evaluation of different methods leads to a consistent outcome

Given that estimating transcript expression from short reads constitutes a challenging task, I performed additional analyses to test the reliability of the aforementioned observations. First, these included the simulation of different RNA-seq datasets to test whether the pipeline used can distinguish between two hypothesised scenarios: one dominant transcript per gene *vs.* similar expression levels of the different transcripts in each gene (see Methods). Such analysis led to the conclusion that the methods used here reliably discriminate between the two scenarios, even when taking into account different sequencing depths (**Figure 2.2a**). In addition, it evidenced that they are not biased towards the identification of a single transcript per gene and corroborated the previously described findings about transcript dominance (**Figure 2.2b**).

Figure 2.2 | Evaluation of two hypothetical scenarios on major transcript abundance.
(a) Relative abundance of the subset of transcripts in each position of the ranking for the simulated datasets, including different sequencing depths. Several datasets were simulated under the assumption of one predominant transcript (default) or similar expression levels for the transcripts annotated within each expressed gene (constant). In comparison to the real expression values used for the simulation (real), major transcript abundance is underestimated after the execution of the analysis pipeline (20M, 40M and 80M, corresponding to the different sequencing depths considered). This suggests that the analysis pipeline used is not biased towards the identification of a single transcript per gene.
(b) Number of dominant major transcripts for the simulated datasets. The results obtained when running the simulation under the hypothesis of a predominant transcript per gene (default) resemble those of the real datasets.



(b)

		expressed genes	genes with a dominant major transcript	
			2-fold dominance	5-fold dominance
DEFAULT	20M	6,102	4,603 (75.43%)	2,958 (48.48%)
	40M	5,817	4,471 (76.86%)	2,943 (50.59%)
	80M	6,081	4,722 (77.65%)	3,113 (51.19%)
CONSTANT	20M	13,533	4,309 (31.84%)	1,145 (8.46%)
	40M	13,585	4,288 (31.56%)	1,249 (9.19%)
	80M	13,622	4,313 (31.66%)	1,282 (9.41%)

Second, the evaluation of alternative methods to estimate transcript expression levels revealed a strong agreement in the major transcript predictions. For example, for highly expressed genes, up to 90% overlap was observed when using Cufflinks (87.78% to 92.59%, SD=1.62; **Table 2.1**). Furthermore, a considerable overlap with the predictions from MISO was also detected when using direct evidence from junction reads (*i.e.* 66.35% - 61.44% to 69.60%, SD=2.66; see Methods).

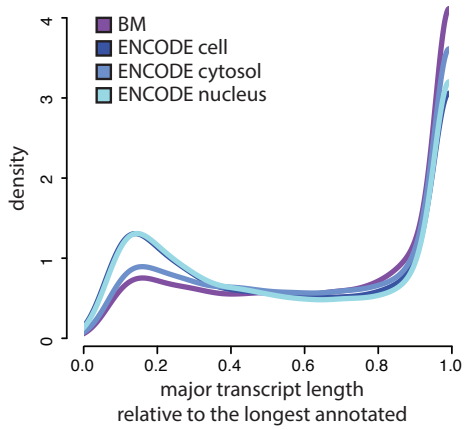
	MISO vs Cufflinks		MISO vs MMSEQ	
	mean correlation	**SD**	**mean correlation**	**SD**
1 FPKM	0.89	4.89e-03	0.70	2.36e-02
5 FPKM	0.93	5.33e-03	0.78	1.96e-02
10 FPKM	0.94	4.34e-03	0.81	1.91e-02

	MISO *vs.* Cufflinks		MISO vs MMSEQ	
	mean overlap (%)	**SD**	**mean overlap (%)**	**SD**
1 FPKM	84.29	1.14	54.66	1.54
5 FPKM	88.71	1.67	57.34	2.98
10 FPKM	90.43	1.62	56.50	4.12

Table 2.1| Consistency in the transcript abundance estimates across different software. There is in general a good agreement between the different methods evaluated, with the biggest differences detected when using MMSEQ, which requires mapping to the transcriptome (see Discussion).
(top) Average Pearson correlation coefficient in the expression estimates.
(bottom) Overlap in the major transcript predictions.

Third, the length of major transcripts was observed to be widely distributed (**Figure 2.3**). More specifically, in over 50% of the cases (50.98% to 55.46%, SD = 1.53) the identified major transcript did not correspond to the longest one annotated. The same trend was observed when taking into account CDS length: in approximately 50% of the genes (44.42% to 48.23%, SD = 1.12), the major transcript did not contain the longest CDS, thus not corresponding to the canonical transcript as annotated in UniProt [UniProt Consortium, 2012]. An example of such cases is the *AES* gene, which presents a ubiquitous major transcript that is shorter than the current reference (**Figure 2.4**).

(a)

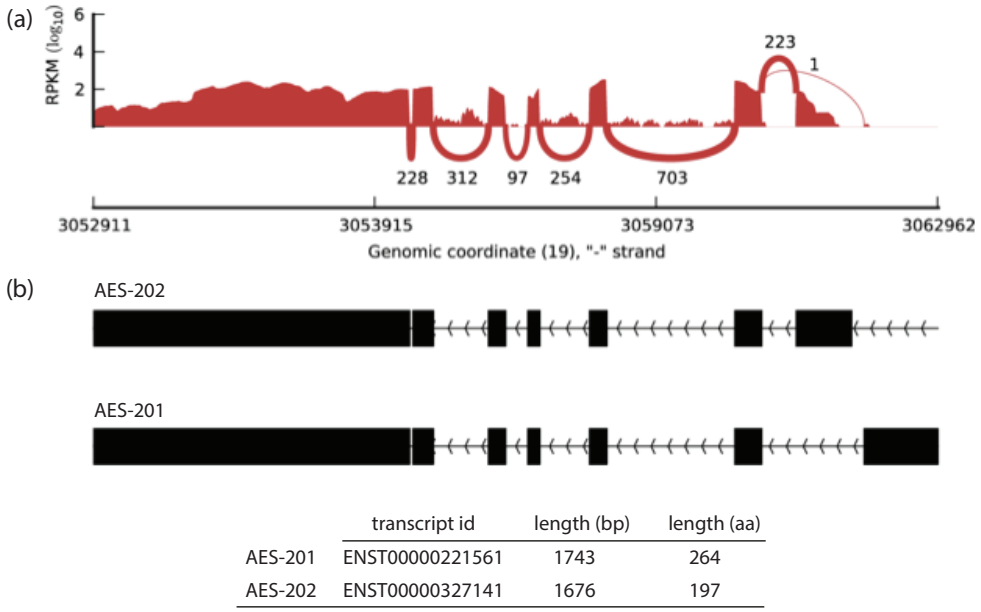


(b)

	expressed genes	major transcripts which are not the longest one annotated	%
BM	8654	5106	58.94
ENCODE cell	8527	5806	68.12
ENCODE cytosol	7877	5073	64.35
ENCODE nucleus	9177	6096	66.48

	expressed genes for which the major transcript has a CDS annotated	major transcripts for which the annotated CDS is not the longest	%
BM	7999	3714	46.35
ENCODE cell	7347	3650	49.74
ENCODE cytosol	7299	3532	48.38
ENCODE nucleus	7049	3564	50.61

Figure 2.3| Length distribution for major transcripts. Genes with several transcripts annotated and expressed above 1 FPKM were taken into account for the analysis.
(a) Length distribution for major transcripts.
(b) Number of major transcripts that correspond to the longest one annotated or that contain the longest CDS. Major transcript detection is not biased towards the longest one annotated.



	transcript id	length (bp)	length (aa)
AES-201	ENST00000221561	1743	264
AES-202	ENST00000327141	1676	197

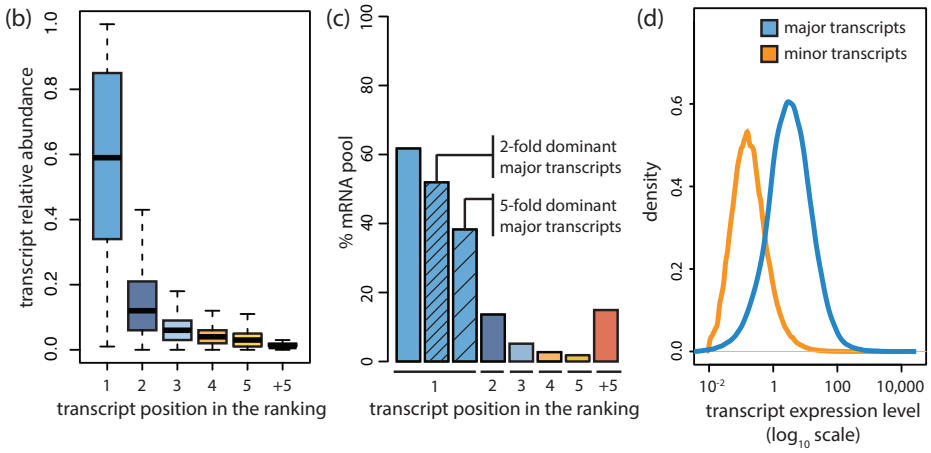
Figure 2.4| Example of non-canonical major transcript common to all the 16 tissues analysed.
(a) Read coverage for the gene AES (amino-terminal enhancer of split).
(b) Annotated transcripts structure and length. The transcript AES-202 is recurrently identified as major in all the tissues of the BM dataset, and it does not correspond to the longest CDS.

Finally, the impact of unnanotated transcripts in the above observations was addressed by performing *de novo* transcript quantification using Cufflinks (see Methods). As expected, such analysis led to the identification of a higher number of transcripts per gene (6.38 in GENCODE v11 *vs.* 12.84 using Cufflinks), but it is still possible to observe one dominant transcript for most of them (**Figure 2.5**).

(a)

	GENCODE	CUFFMERGE
genes	18,450	18,635
transcripts	117,759	239,350
ratio	6.38	12.84

% genes with *n* transcripts	GENCODE	CUFFMERGE
n=1	19.23	18.32
n=2	12.54	7.84
n=3	9.78	6.74
n=4	8.30	4.73
n=5	7.06	4.76
n>5	43.09	57.61



(e)

	expressed genes	genes with a dominant major transcript	
		2-fold dominance	5-fold dominance
1 FPKM	11,169 (59.94%)	7,399 (66.14%)	4,642 (41.47%)
5 FPKM	6,830 (36.65%)	4,669 (68.25%)	2,972 (43.68%)
10 FPKM	4,401 (23.62%)	3,052 (69.33%)	1,960 (44.61%)

Figure 2.5| Summary of Cufflinks *de novo* quantification results.
(a) Summary of the number of genes and transcripts considered in the study set (GENCODE) and those obtained after de novo transcript identification in the BM dataset using Cufflinks (Cuffmerge). As expected, the average number of transcripts per gene increases after *de novo* transcript quantification.
(b) Relative abundance of the subset of transcripts in each position of the ranking. It is possible to detect one predominant transcript over the rest, albeit with lower relative abundance than reported in that **Figure 2.1**.
(c) Percentage of the studied mRNA pool explained by each category of transcripts. Lowly abundant transcripts explain a larger fraction of the mRNA pool than in **Figure 2.1**, due to the increase in the average number of transcripts per gene.
(d) Expression distribution for major and minor transcripts. It is still possible to observe a clear separation between the expression levels of major *vs.* minor transcripts.
(e) Number of dominant major transcripts, taking into account all genes in the extended annotation (n = 18,635). Most major transcripts are expressed in a dominant fashion, and this becomes more evident for higher FPKM thresholds.

2.2.3 Major transcripts from coding genes do not always code for proteins

Functional classification of major transcripts revealed that, for 17% of protein coding genes expressed in primary tissues (15.26% to 20.64%, SD = 1.60), the major transcript lacks an annotated CDS as indicated by GENCODE. Taking into account expression levels, and focusing on cell lines data, major non-coding transcripts were observed to be more abundant in the nucleus, where they represent approximately 15% of the studied mRNA pool (12.99% to 16.66%, SD = 1.10, **Figure 2.6a**). Genes with major non-coding transcripts are expressed at higher levels in the nucleus, compared to those with major coding transcripts, while this trend is inverted in the cytosol (**Figure 2.6b**). In addition, non-coding major transcripts are less dominant than coding ones in both compartments (**Figure 2.6b**). Finally, analysis of the annotation revealed that these major non-coding transcripts correspond to retained introns and processed transcripts, which lack an open reading frame (see Methods). The latter are more prevalent in the cytosol, while the proportion of retained introns is higher in the nucleus (**Figure 2.6c**).

In order to evaluate the hypothesis that incomplete splicing could explain the higher proportion of major retained introns in the nucleus, I compared intron expression levels across cellular compartments (see Methods for details on the calculation of intron expression). As expected, intron expression was detected to be higher in the nucleus compared to the cytosol (**Figure 2.7a**). In addition, such analysis revealed a general trend in the location of major retained introns towards the transcriptional 3'-end (**Figure 2.7b**), which has been previously linked to the nonsense-mediated decay pathway (see Discussion). Interestingly, this trend is more accentuated in the cytosol than in the nucleus, where it could be masked by the higher intronic expression levels. Alternatively, the prevalence of retained introns as a major transcript could point to a functional mechanism, since genes with retained introns as the major transcript both in nucleus and cytosol were detected to be expressed at lower levels in the latter (**Figure 2.7c**; see Discussion). Those genes are associated with ribosomal components, consistent with previous findings indicating that introns regulate the expression of ribosomal proteins in yeast (**Table B.5**, see Discussion).

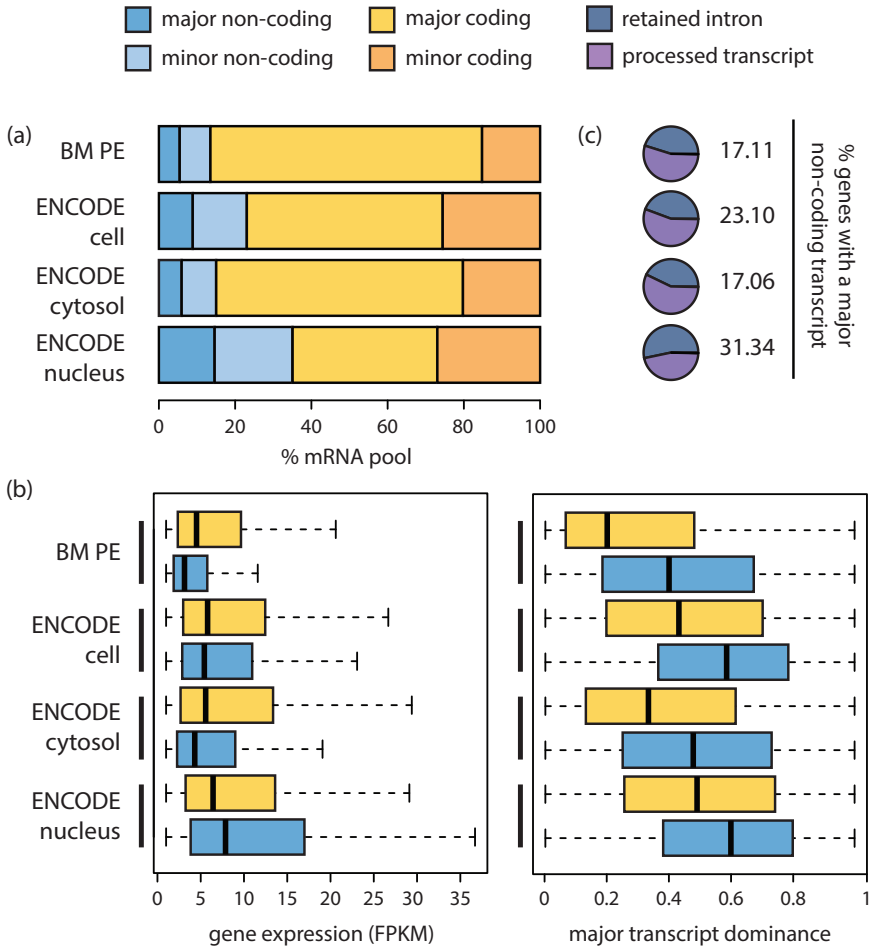


Figure 2.6 | Major non-coding transcripts in protein coding genes.
(a) Proportion of the mRNA studied represented by different categories of transcripts. Average proportions were calculated including all the samples from each dataset. Major non-coding transcripts are more abundant in the nucleus compared to the cytosol.
(b) Expression patterns across cellular compartments for major non-coding transcripts. Protein coding genes for which the most abundant transcript is non-coding are expressed at higher levels in the nucleus, whilst this trend becomes inverted in the cytosol (left). Major transcript dominance becomes reduced both when the major transcript is non-coding and in the nucleus (right).
(c) Transcript biotype categories for the major non-coding transcripts. Average proportions were calculated including all the samples from each dataset. Processed transcripts are more abundant in the cytosol, while retained introns represent the major fraction in the nucleus. Other minor categories that represented less than 1% of the transcripts were also identified, but are not visible in the plots.

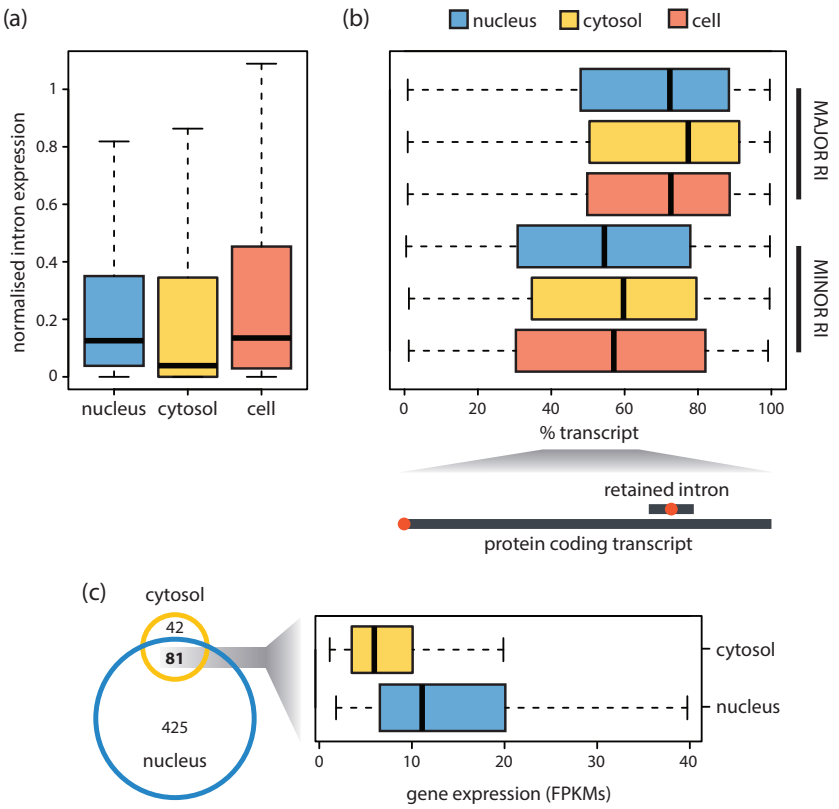


Figure 2.7 | Focus on retained introns.
(a) Normalised intron expression in different cellular compartments. FPKMs were calculated for all the introns and normalised by gene expression levels (see Methods). Intron expression is higher in the nucleus than in the cytosol (Wilcoxon test p-value < $2.2 \cdot 10^{-16}$).
(b) Location of the dominant retained introns within the context of protein coding transcripts. Genes for which the major transcript is a retained intron (RI) were initially considered in the analysis, and cases where the second most abundant transcript is protein coding and overlaps with the RIs were further selected. Similar criteria were applied to analyse minor RIs. The location of the RIs is obtained by measuring the distance from their centre to the transcriptional start of the overlapping coding transcript, as illustrated in the panel below the figure (red dots). Major RIs are preferentially located towards the transcriptional end of protein coding transcripts.
(c) Expression levels for genes with major retained introns. The number of genes for which the most abundant transcript is a RI is represented in the left. Amongst the genes with major RIs in both cellular compartments (n = 81), gene expression is higher in the nucleus (Wilcoxon test p-value < $2.2 \cdot 10^{-16}$).

On the other hand, the term *processed transcript* constitutes an ambiguous category. Manual inspection of a subset of processed transcripts that were consistently identified across all samples as the major transcript indicated that they could potentially be re-annotated to protein coding, nonsense-mediated decay or retained intron (**Table B.6**). Together, these observations suggest that the true proportion of non-coding major transcripts for protein coding genes may be lower than the current annotation suggests, in line with recent evidence pointing to the existence of peptides from non-coding RNAs [Hemberg et al., 2014].

2.3 Discussion

In this chapter, I combine RNA-seq data from different primary tissues, cell lines and cellular compartments to characterise the human protein coding transcriptome from a functional perspective. I show that, in a given condition, most protein coding genes not only express one major transcript, as recently observed by Djebali et al. [2012], but in most cases the major transcripts are dominating the transcriptome. The described findings are consistent across the two datasets studied here, and are supported by several quantification methods, including direct evidence from junction reads. Interestingly, the biggest differences in major transcript predictions were detected when using MMSEQ, and those even exceeded the ones obtained from direct interpretation of the raw data. Similarly to MISO, this software uses a Bayesian approach to infer transcript abundances, but requires mapping the reads to the transcriptome instead (see Introduction - section 1.3.3.2). Such differences expose the potential variability introduced during the mapping step and emphasise the need for further workflow evaluation studies similar to the ones undertaken through the RGASP project [Engstrom et al., 2013; Steijger et al., 2013]. On the other hand, transcript quantification from short read sequences is not a trivial task, and the current annotation is continuously updated to include novel transcripts. *De novo* transcript quantification revealed consistent results to those obtained when relying exclusively on the annotation, although major transcript dominance was less accentuated due to the higher proportion of genes with a large number of transcripts. Nonetheless, it is reassuring to observe that the detected single transcript dominance becomes stronger for highly expressed genes, for which transcript prediction and quantification have been reported to be more reliable [Steijger et al., 2013]. Furthermore, similar observations have been made at the single-cell level following publication of our

manuscript [Marinov et al., 2014; Shalek et al., 2013; Yan et al., 2013], and in the long term, longer reads will shed more light on the topic.

The observation that a non-negligible fraction of protein coding genes express a major non-coding transcript came as a surprising finding. However, non-coding major transcripts are more prevalent in the nucleus, specifically in the case of retained introns, and it was also possible to detect higher intronic expression levels in this compartment. Altogether, these observations could reflect incomplete splicing, consistent with existing evidence indicating that unspliced or incompletely spliced mRNAs are not exported to the cytosol [Porrua and Libri, 2013]. In addition, the described results show that retained introns are preferentially located towards the transcriptional end of transcripts. This observation has been previously linked to the nonsense-mediated decay pathway [Kurmangaliyev and Gelfand, 2008], a surveillance mechanism that ensures the degradation of unspliced transcripts when they are transported to the cytosol (see Introduction - The cytosolic life of mRNAs) [Pérez-Ortín et al., 2013]. Nevertheless, several cases of functionally relevant retained introns have been described, either as a mechanism to target mRNA molecules (*e.g.* [Buckley et al., 2011]), produce alternative protein products (*e.g.* [Li et al., 2006]) or to regulate expression levels (*e.g.* [Wong et al., 2013a; Yap et al., 2012]). In this context, I observe that genes with retained introns as the major transcript in both nucleus and cytosol are expressed at considerably lower levels in the latter, which could point to a regulatory role. Finally, I also detect that those genes are associated to ribosomal components, which is consistent with previous findings indicating that introns regulate the expression of ribosomal proteins in yeast [Parenteau et al., 2011].

Overall, it is difficult to predict the impact of the described observations at the protein level. There have been several studies addressing the relationship between protein and transcript levels, which in general point at a modest, but not insignificant correlation (*i.e.* the best estimates point at a range of 56%-64% correlation [Li et al., 2014; Lundberg et al., 2010; Nagaraj et al., 2011]). Translational efficiency, mRNA and protein turnover rates are likely to have an impact on protein levels [Vogel and Marcotte, 2012]. In addition, alternative splicing not only has an impact on the protein repertoire, but also contributes to the control of expression levels [McGlincy and Smith, 2008; Yap et al., 2012] and transcript localisation [Keren et al., 2010; Nilsen and Graveley, 2010], which brings in other

potential roles for minor transcripts. On the other hand, proteomics studies also show that detecting a protein is unlikely unless there are at least a certain number of RNA molecules per cell [Ramakrishnan et al., 2009]. This may be partly due to insufficient sensitivity of the methods used; nevertheless, the clear separation in expression levels between major and minor transcripts, together with the higher overlap of the former with an independent set of transcripts predicted to be translated into functional products, suggest that the abundance of proteins derived from minor transcripts is likely to be lower than the one from dominant ones, and that minor transcripts could simply result from noisy splicing [Melamud and Moult, 2009].

Further information on the functional impact of the observations described in this chapter can be obtained from comparative analyses. Alternative splicing is a widespread process among eukaryotes, where it has been regarded as a mechanism for the diversification of protein sequence, structure and function, hence conferring flexibility in the regulation of functional products across different tissues or developmental stages [Nilsen and Graveley, 2010]. Comparative studies based on the analysis of exon usage patterns have provided evidence for the existence of lineage-specific splicing patterns [Barbosa-Morais et al., 2012; Merkin et al., 2012], and whether similar results can be observed from the evaluation of trends in major transcript expression across species remains an open question. This analysis would not only complement past efforts for the identification of evolutionary conserved splicing programs, but would also help in assessing the functional relevance of major transcripts, providing extra evidence for function for those with a high degree of conservation. Similarly, it would also be interesting to evaluate major transcript dominance in the context of gene properties that have been shown to correlate with alternative splicing. For example, the number of transcripts annotated per gene has been shown to correlate positively with gene age, length and the number of cassette exons [Roux and Robinson-Rechavi, 2011], but the relationship of these features with major transcript expression patterns has so far been unexplored. Altogether, the research questions proposed here would contribute in improving the current mechanistical understanding of splicing.

Knowledge on major transcripts can be used to build a catalogue for the reference transcriptome, by focusing on dominant transcripts that are recurrent in many

samples. In this context, a closer inspection of such set of transcripts revealed cases where they do not contain the longest CDS, a criteria often used in resources like UniProt to define a reference transcript [UniProt Consortium, 2012]. These results reinforce previous observations on the limitations of the current definitions [Tress et al., 2008a], and point to potential advantages of taking into account this type of data. Similarly, such information can also help in improving the existing annotation, as exemplified here with the re-annotation of a subset of major processed transcripts that were consistently identified across all samples. Moreover, identification of changes in the major transcript across conditions can lead to relevant findings, as shown in the next chapter. Finally, such observations may also help proteome analysis by prioritising the candidate proteins that are more likely to be present in a given sample, as further explored in Chapter 5.

In conclusion, the discovery of alternative splicing and many different classes of non-coding RNAs, together with the establishment of RNA-seq, revealed that the number of transcripts exceeds many times the number of genes in the human genome. This has been used to argue that alternative splicing possibly explains the low number of genes compared to what was believed before it was sequenced [Nilsen and Graveley, 2010]. Despite such transcriptome diversity, the results described here show that most protein coding genes express one dominant transcript in a given condition, which, when combined, comprise most of the potentially coding mRNA transcriptome. Hence, it is tempting to hypothesize that although some minor transcripts may play a functional role, the major ones are likely to be the main contributors to the proteome.

2.4 Computational methods

All the computational analyses described in this chapter have been performed by myself, except for the re-annotation of major processed transcripts, which has been carried out by Dr. Adam Frankish, as detailed below.

Datasets and mapping

Analyses were based on the Illumina Body Map (BM) dataset and a subset of ENCODE cell lines [ENCODE Project Consortium et al., 2012] (ArrayExpress accession ids: E-MTAB-513 and E-GEOD-26284, respectively), jointly covering

a total of 21 different tissues and cell lines, as well as different cellular compartments (see **Table A.1**). Raw FASTQ files were retrieved from the European Nucleotide Archive[1]. In addition to the publicly available datasets, two RNA-seq experiments were generated using the Flux Simulator [Griebel et al., 2012]. On the one hand, the scenario of one dominant transcript per gene was simulated by running this tool with the default parameters. On the other hand, a custom `.pro` file with `AFREQ_EXP=1000` was used to resemble similar expression of the several transcripts annotated within each given gene.

FASTQ files in the BM dataset were filtered before mapping by trimming the last five nucleotides of all reads. Raw data were mapped to the human genome and transcriptome (Ensembl 66 [Flicek et al., 2012]) using Bowtie v0.12.7 [Langmead et al., 2009] and TopHat v1.3.3 [Trapnell et al., 2009], respectively.

Gene and transcript study sets

Gene and transcript annotations used in the analyses correspond to those in GENCODE v11 [Harrow et al., 2012]. This annotation is the result of the combination of both computational and experimental evidence, as well as manual curation efforts, and has been shown to offer a good compromise in terms of complexity when compared to existing alternatives [Wu et al., 2013].

All the analyses discussed in this chapter are based on protein coding genes. Those genes for which at least one of the annotated transcripts was shorter than 300 bp were removed from the analyses (n = 1,638), since they would be lost during the size selection step in the RNA-seq experiment (see Introduction - A typical sequencing workflow). In total, the study set comprises 18,450 protein coding genes, of which 14,902 have more than one transcript annotated.

Biotype definition for transcripts derived from protein coding genes

The transcript biotypes used in this chapter were obtained from the GENCODE annotation, and are defined as follows:

- *Protein coding*: if the transcript contains a CDS.
- *Nonsense-mediated decay (NMD)*: if the transcript contains a CDS but has one or more splice junctions >50bp downstream of the stop codon.

[1] `http://www.ebi.ac.uk/ena/`

- *Retained intron*: if the transcript has an intronic sequence compared to a reference variant and has no strong evidence for function.
- *Processed transcript*: if the transcript does not contain a CDS and does not fulfil the previous criteria.

Counting reads overlapping exonic and intronic regions

Exonic coordinates were retrieved from the annotation and used to define intronic regions. Formally, the definition of intron encompasses those regions that are located inside genic boundaries and are not overlapped by any exon in any annotated transcript. The number of reads overlapping known exons and introns was computed using dexseq-count (DEXSeq v1.5.5 [Anders et al., 2012]) and converted to FPKM values with custom scripts.

Estimating gene and transcript expression levels

For each gene, expression levels were calculated as the average FPKMs of all expressed exons. Transcript abundances were obtained using three different tools: MISO v0.4.1 [Katz et al., 2010], Cufflinks v1.3.0 [Trapnell et al., 2010] and MMSEQ v0.10.0 [Turro et al., 2011]. MISO and Cufflinks take as input alignments to the genome, while MMSEQ requires mapping to the transcriptome, thus the need to use two different mapping strategies (see Methods - Datasets and mapping). In all three cases, the expression estimates were based on the existing transcript annotation, cancelling any option for *de novo* inference, and were converted to transcript relative abundances when necessary. In this chapter, I refer to the results obtained by MISO and I use a default FPKM threshold of 1 to consider a gene/transcript as *expressed*, a threshold that has been suggested as the minimum expression required for protein detection [Hebenstreit et al., 2011; Vogel and Marcotte, 2012]. In addition, higher expression thresholds are also included in the analyses (5 and 10 FPKM), since transcript quantification has been reported to be more reliable for those [Steijger et al., 2013]. Finally, a transcript is considered as *detected* independently of its expression level, provided that the corresponding gene is expressed.

mRNA pool estimates

mRNA pool estimates were calculated as introduced by Ramskold et al. [2009]. Briefly, the fraction of the studied mRNA pool that can be explained by the expression of major transcripts can be represented as the ratio of the sum of FPKMs

for major transcripts *vs.* the sum of FPKMs for all the transcripts in the study set. All transcripts encoded within protein coding genes were taken into account in the calculation, independently of their transcript biotype. Mitochondrial genes in the study set were discarded (n = 11 in the study set), since they are present multiple times in the cell and could bias the quantification.

Using direct evidence from junction reads

Genes that are part of the study set (see Methods - Gene and transcript study sets) and for which all of the annotated transcripts can be uniquely identified by at least one splice junction (n = 2,306) were considered in this analysis. Major transcripts were then identified based on coverage evidence, by quantifying the number of reads supporting each junction and taking the average in case of several splice junctions. For each sample, the overlap with MISO was calculated.

De novo transcript discovery using Cufflinks

Cufflinks v1.3.0 [Trapnell et al., 2010] was used to discover novel transcripts in each tissue from the BM dataset and all the obtained annotations were merged using cuffmerge [Trapnell et al., 2010]. Next, the subset of transcripts that overlaps with known protein coding genes was used for analysis, and those genes with transcripts shorter than 300 bp were filtered out (see Methods - Gene and transcript study sets). A summary of the number of genes and transcripts identified can be found in **Figure 2.5a**.

Gene Ontology enrichment analyses

Gene Ontology (GO) enrichment analyses were performed with the DAVID and WebGestalt [Huang et al., 2009a,b; Wang et al., 2013]. The reference population was defined by our gene study set (see Methods - Gene and transcript study sets), and an adjusted p-value of 0.05 was used as a threshold for the identification of significant GO terms (Benjamini and Hochberg correction [Benjamini and Hochberg, 1995]).

Re-annotating major processed transcripts

Analysis performed by Dr. Adam Frankish.

Processed transcripts that were recurrently identified as the major isoform in a consistent manner across all the studied samples were manually curated by using

information from the Zmap annotation interface[1] and re-annotated as indicated in **Table B.4**.

[1] http://www.sanger.ac.uk/resources/software/zmap

Chapter 3

The prevalence of splicing changes in cancer

Splicing plays a central role in the expression of most genes, and its dysregulation is known to contribute to the pathogenesis of several diseases, including cancer [Padgett, 2012; Pedrotti and Cooper, 2014]. In this context, alterations in the splicing reaction have been associated with oncogenesis, tumour suppression and metastasis [Hagen and Ladomery, 2012; Kaida et al., 2012]. The goal of the present chapter is to investigate the prevalence of changes in splicing patterns in the most common type of kidney cancer, *i.e.* clear cell Renal Cell Carcinoma (ccRCC). The analyses described here have been performed as part of the CAGEKID project (CAncer GEnomics of the KIDney), a broad effort aimed at improving the current understanding of this type of cancer through the combination of data from genomics, transcriptomics and epigenomics experiments (DNA-seq, RNA-seq and methylation arrays, respectively). In the first part of the chapter, I perform a global analysis of the splicing changes that underlie ccRCC. Those are further dissected in the second part, with special emphasis on the most extreme and recurrent events. I next pursue integration of the described results with those obtained from other types of data generated within the same Consortium. Finally, I expand these analyses to a separate RNA-seq dataset in order to investigate the resemblance of splicing patterns in primary tissues and cell lines, linking the derived findings to the observations from Chapter 2.

The results from all the computational analyses reported here have been produced by myself under the supervision of Dr. Alvis Brazma. Further analyses

performed by Dr. Johan Rung from EMBL-EBI and Dr. Louis Letourneau from McGill University have also been included in sections 3.2.1 and 3.2.4, and their contribution has been emphasised in the Methods.

Publications derived from this chapter

- Gonzàlez-Porta & Brazma. Identification, annotation and visualisation of extreme changes in splicing from RNA-seq experiments with SwitchSeq. *bioRxiv*. 10.1101/005967.
- (submitted) Scelo*, Riazalhosseini*, Greger*, Letourneau*, Gonzàlez-Porta* et al. Whole-genome sequencing reveals variation in the genomic landscape of clear cell Renal Cell Carcinoma in Europe.
- (poster) Systems biology: Global regulation of gene expression 2014, CSHL.

*shared first authors

3.1 Introduction

Numerous studies have evidenced the role of alternative splicing in disease development, including cancer (*e.g.* see Padgett [2012]; Pedrotti and Cooper [2014] for a review), where its dysregulation has been linked to oncogenesis, tumour suppression and metastasis [Hagen and Ladomery, 2012; Kaida et al., 2012]. In general, changes in splicing patterns are known to contribute to disease either by directly affecting the function of proteins (e.g. protein-protein interactions, sub-cellular localization or catalytic ability [Keren et al., 2010; Nilsen and Graveley, 2010]), or by subtly regulating gene expression levels [McGlincy and Smith, 2008; Yap et al., 2012]. Moreover, several mechanisms that could lead to the disruption of splicing have been identified, including mutations in core elements of the splicing machinery, alterations on the concentration and function of Splicing Factors and mutations in regulatory regions associated with splice site selection [Tazi et al., 2009].

In this chapter, I aim at investigating the changes in splicing patterns associated with clear cell Renal Cell Carcinoma (ccRCC), the most common and aggressive type of kidney cancer [Jonasch et al., 2012]. ccRCC tumours develop in the renal parenchyma and represent 70-80% of renal cancers, which are in general one of the most commonly diagnosed cancers in adults (*i.e.* they represent the 7th most common cancer in Europe and account for 3% of adult malignancies worldwide) [Chow et al., 2010]. Key clinical needs for ccRCC include the identification of new biomarkers and new therapeutic targets, as suggested by the fact that this is one of the few tumour types for which there are currently no biomarkers in routine clinical use [Jonasch et al., 2012]. Recently, several large scale sequencing projects, including initiatives from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have shed light on the genomic alterations that underlie this type of cancer [Sato et al., 2013; Scelo et al.; TCGA Network, 2013]. The main findings derived from this large scale projects include the consolidation of previous knowledge on the role of the tumour suppressor gene *VHL* in ccRCC development (with mutations detected in approximately 90% of spontaneous tumours), as well as on the relationship between the allelic loss of chromosome 3p and disease pathogenesis [Brugarolas, 2013]. Interestingly, the *VHL* gene is located at this locus, as it is the case for other driver genes that have been identified amongst the most commonly mutated ones (*i.e. PBRM1*, *SETD2*

and *BAP1*; all of them tumour suppressors) [Brugarolas, 2013]. In addition, these studies have reinforced the idea that ccRCC is a metabolic disease, with many genomic aberrations affecting several metabolic pathways [Linehan et al., 2010]. However, with much of the focus being placed on the genomics scale, a thorough characterisation of the splicing alterations that affect this type of cancer is still missing, in spite of data availability (*i.e.* RNA-seq data for hundreds of patients was generated in all the three aforementioned studies). Valletti et al. [2013] and Zhao et al. [2013] relied on the identification of differentially expressed exons to assess exon skipping events, by using microarrays and the previously mentioned TCGA RNA-seq data, respectively, but this is a limited approach that focuses only on the study of a subset of splicing events.

Through the analysis of RNA-seq data for 45 matched ccRCC samples (*i.e.* tumour and healthy peripheral tissue from the same patient), I show here that splicing patterns are largely altered in ccRCC, with almost 40% of the expressed genes being affected (n = 7,842 out of 19,944 genes). Consistent with the results reported in Chapter 2, I identify gene expression to be dominated by one transcript in most cases, both in tumour and healthy samples, and I detect a bigger overlap in major transcripts for the latter condition. Following these observations, I estimate that only ~8% of the genes with altered splicing manifest significant changes in major transcript dominance across conditions (n = 602 genes). Thus, I conclude that despite the large number of genes that undergo differential splicing events, most of those do not lead to big changes in the abundance of major transcripts, at least in a recurrent fashion. Further study of switch events in each sample pair (*i.e.* changes in the identity of the major transcript for a given gene across conditions) confirmed the idea that most of the big changes in splicing are patient-specific. This behaviour becomes even more extreme for those events that involve dominant transcripts (*i.e.* major transcripts that are expressed at a considerably higher level in comparison to the second most abundant one from the same gene). Nonetheless, following integration with genomic data for a set of 36 matched samples, I observe that pathways are widely disrupted when combining information across multiple regulatory layers. For example, from such integrative analysis, I do not detect major transcriptomic alterations for *VHL*, but I do observe switch events for *PBRM1* and *BAP1* in several patients. Interestingly, for *PBRM1*, these do not overlap with those cases in which the gene is mutated. Finally, I explore overall differences in the splicing patterns of primary tissues

vs. cell lines by analysing RNA-seq data from a set of 6 ccRCC cell lines.

The analyses described in this chapter have been performed as part of the CAGEKID project (Cancer Genomics of the Kidney), one of the aforementioned large scale sequencing efforts that aimed at characterising this disease from a genomic, transcriptomic and epigenetic perspective. This project is part of ICGC and comprises 14 partners from 6 EU countries and Russia, including our team at the EBI, which has led the interpretation of the transcriptomic data.

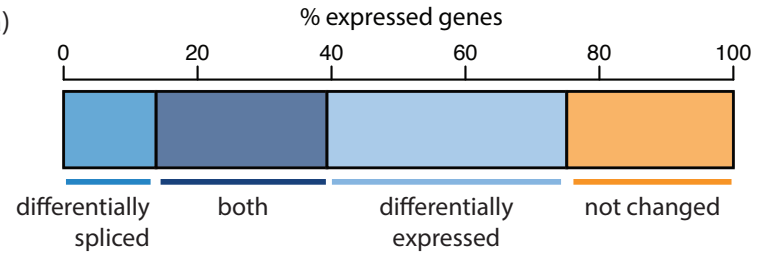
3.2 Results

In spite of the diversity of data generated within the context of CAGEKID, my analyses rely primarily on the available matched RNA-seq samples. This includes a set of 45 patients diagnosed with ccRCC, from whom both tumour and healthy kidney tissue were extracted and sequenced with good quality. In all the cases, revision by multiple pathologists ensured a 70% content of tumour cells in the ccRCC samples. RNA-seq was performed on polyA-selected extracts using the Illumina HiSeq 2000 platform (100 bp paired-end reads; **Table A.2**). In addition, DNA-seq data for 36 of the 45 matched samples was also produced with the same platform (100 bp paired-end reads), thus enabling the integration of genomic and transcriptomic evidence. Finally, I also had access to RNA-seq data from a separate set of 6 ccRCC cell lines (**Table A.2**). Even though those samples were not formally generated as part of CAGEKID, they were sequenced by one of the participating laboratories following the same protocols, thus providing a good scenario for the comparison of splicing patterns in primary tissues *vs.* cell lines.

3.2.1 Splicing is largely altered in ccRCC

First, genes with significant changes in splicing in ccRCC *vs.* healthy tissues were identified by running DEXSeq on the 45 matched RNA-seq samples (see Methods). Out of the 19,944 genes detected as expressed, a total of 7,842 genes were predicted to manifest significant differences in exon usage (padj < 0.01; **Figure 3.1a**), the vast majority of which are protein coding genes (n = 7,182). GO enrichment analysis of this set of genes revealed that they are involved in processes such as cellular respiration, the formation of cell junctions and the regulation of cell cycle and apoptosis, amongst other functions (**Figure 3.1b**).

(a)



(b)

GO term	adjusted p-value
intracellular protein transport	2.75E-014
protein biosynthesis	8.83E-012
kinase activity	1.33E-012
cellular respiration	2.52E-008
generation of precursor metabolites and energy	1.61E-008
cytoskeleton organization	8.35E-006
actin binding	1.57E-006
DNA repair	8.81E-005
cell death	5.46E-005
cell cycle	4.35E-005
RNA splicing, via transesterification reactions	9.6E-003
tight junction	6.3E-003
oxidative phosphorylation	5.7E-003
mitochondrial respiratory chain	5.4E-003
posttranscriptional regulation of gene expression	3.5E-003
regulation of apoptosis	2.1E-003
adherens junction	1.6E-003

pathway	adjusted p-value
metabolic pathways	2.72E-191
pathways in cancer	1.82E-036
oxidative phosphorylation	1.10E-028
focal adhesion	4.37E-027
tight junction	1.90E-025
cell cycle	2.79E-018
MAPK signalling pathway	8.05E-018
spliceosome	6.00E-017
PPAR signalling pathway	1.87E-016
adherens junction	2.27E-016
RNA transport	4.20E-016
ErbB signalling pathway	8.82E-015
renal cell carcinoma	1.34E-011
p53 signalling pathway	1.31E-010
cell adhesion molecules (CAMs)	6.21E-009
mRNA surveillance pathway	6.91E-009
gap junction	7.12E-009
RNA degradation	1.16E-008
mTOR signalling pathway	6.41E-008
drug metabolism - cytochrome P450	1.13E-007
apoptosis	4.35E-006
drug metabolism - other enzymes	2.51E-005
VEGF signalling pathway	4.67E-005
TGF-beta signalling pathway	4.95E-005

(c)

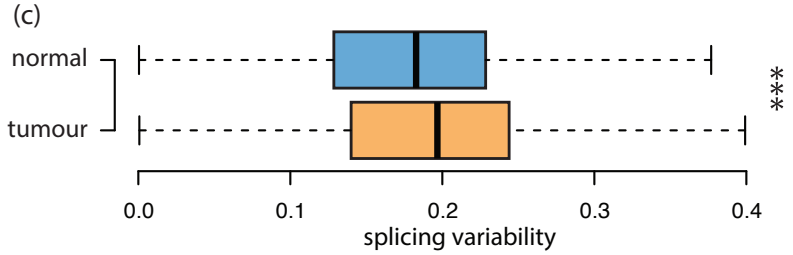


Figure 3.1 | Splicing patterns in ccRCC tumours *vs.* healthy matched samples. *(a) Percentage of genes that are differentially spliced, differentially expressed or both amongst the set of expressed genes.* A large proportion of the expressed genes present differences in either splicing, expression or both.
(b) GO and pathway enrichment analysis for the set of differentially spliced genes. Genes with significant differences in splicing have a role in cellular respiration, energy metabolism, the formation of cell junctions and cell cycle progression. Representative terms for 77 enriched pathways are shown here.
(c) Splicing variability distribution for healthy and tumour samples. Transcript relative abundances within each given gene are more variable across tumours than across normal samples (p-value $< 2.2 \cdot 10^{-16}$).

Similar results were obtained following pathway enrichment analysis, which expectedly pointed at an alteration of pathways involved in cancer and renal cell carcinoma (**Figure 3.1b**). Overall, the above results suggest that, in almost 40% of the expressed genes, alterations in splicing are broad enough to be detected in a general comparison of normal *vs.* tumour samples. Such a large fraction of altered genes is consistent with the high number of differentially expressed genes detected in the same sample set (n = 12,231; padj < 0.01; see Methods). Notably, there is a significant overlap between these two sets (p-value < $2.2{\cdot}10^{-16}$; common genes = 5,090; **Figure 3.1a**).

On a separate note, transcript expression levels were detected to be more variable within tumours compared to normal samples (p-value < $2.2 \cdot 10^{-16}$, **Figure 3.1c**; see Methods). Given the positive correlation between the number of annotated transcripts and splicing variability (r_S = 0.75; p-value < $2.2 \cdot 10^{-16}$), part of this increase in variability could be explained by a higher number of expressed transcripts per gene in tumours (**Figure 3.2**).

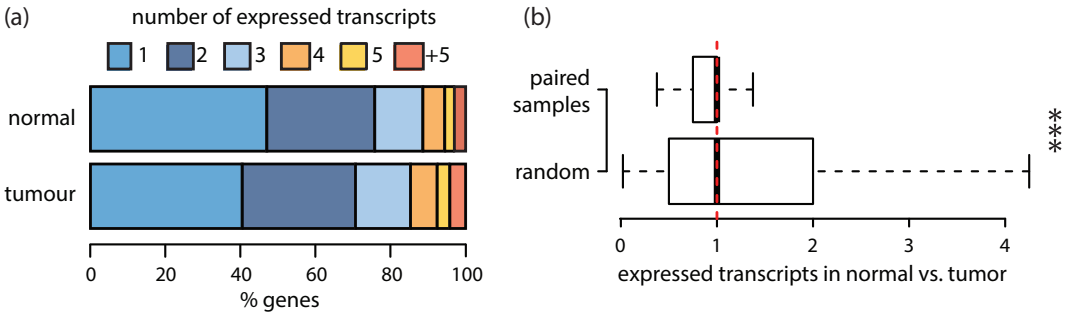
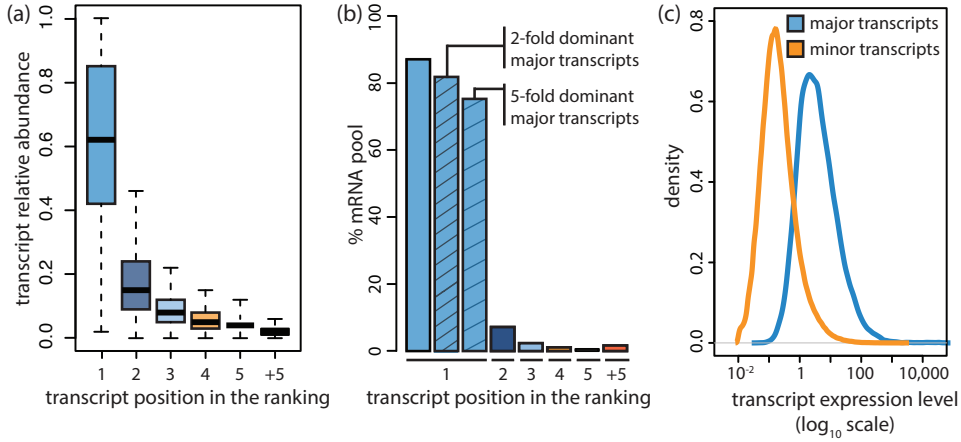


Figure 3.2| Number of expressed transcripts per gene in normal *vs.* tumour samples.
(a) Percentage of genes with each given number of expressed transcripts. 1 FPKM was used as the expression threshold, following the analysis workflow introduced in Chapter 2.
(b) Ratio of the number of expressed transcripts in paired samples compared to a randomised set. Tumours tend to express a higher number of transcripts per gene (p-value < $2.2 \cdot 10^{-16}$).

Finally, and consistent with the results described in Chapter 2, gene expression was observed to be dominated by one transcript in most cases, both in tumours and in healthy tissue samples (**Figure 3.3**). The integration of information across biological replicates evidenced that genes commonly express one major transcript (**Figure 3.4a**); nonetheless, it was also possible to identify several major transcripts for a considerable fraction of genes. Notably, the number of different major transcripts per gene was detected to be higher in tumours, in line with the previous observations on the overall larger number of expressed transcripts in this condition. Furthermore, major transcripts were detected to be recurrently expressed in both conditions (*e.g.* 90% of the genes express the same major transcript in >50% of the samples where they are expressed; **Figure 3.4b**). Lastly, the overlap in major transcripts was identified to be higher within healthy tissue samples than within tumours (**Figure 3.4c**), consistent with the higher splicing variability previously detected for the latter.

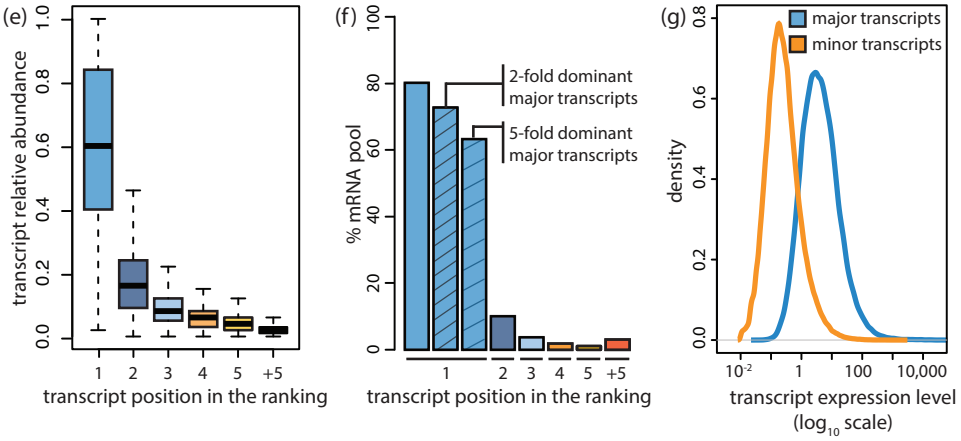
NORMAL



(d)

	expressed genes	genes with a dominant major transcript	
		2-fold dominance	5-fold dominance
1 FPKM	10,368 (56.19%)	6,977 (67.31%)	4,222 (40.74%)
5 FPKM	5,549 (30.07%)	3,934 (70.91%)	2,627 (47.34%)
10 FPKM	3,336 (18.08%)	2,445 (73.28%)	1,715 (52.34%)

TUMOUR



(h)

	expressed genes	genes with a dominant major transcript	
		2-fold dominance	5-fold dominance
1 FPKM	10,802 (58.55%)	7,115 (65.88%)	4,187 (38.79%)
5 FPKM	6,491 (35.18%)	4,502 (69.41%)	2,878 (44.42%)
10 FPKM	3,937 (21.34%)	2,836 (72.07%)	1,922 (48.89%)

Figure 3.3 | Most protein coding genes express one dominant transcript in both normal and tumour samples.
(a,e) Relative abundance of the subset of transcripts in each position of the ranking.
(b,f) Percentage of the studied mRNA pool explained by each category of transcripts.
(c,g) Expression distribution for major and minor transcripts.
(d,h) Average number of genes with dominant major transcripts.

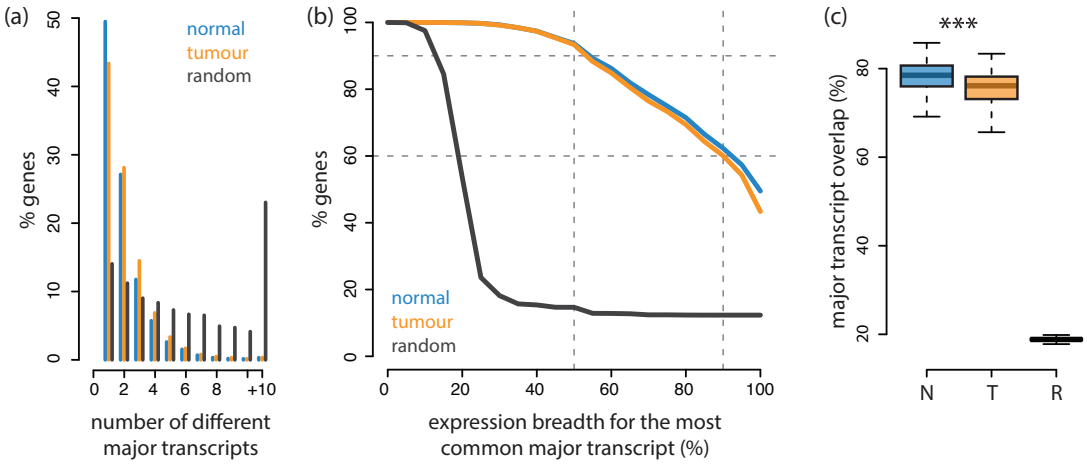


Figure 3.4 | Major transcript expression patterns in ccRCC tumours *vs.* healthy matched samples.
(a) Number of major transcripts detected for each gene in each of the evaluated conditions. Genes have been stratified with regard to the number of different major transcripts that can be detected across biological replicates, and a random distribution was obtained by randomising the set of major transcripts expressed in each sample. One single major transcript can be detected for most genes; however, it is also possible to observe more than one major transcript for a large percentage of cases.
(b) Expression breadth for major transcripts. Expression breadth is measured by the number of samples where a given transcript is detected as major, relative to the number of samples where the corresponding gene is identified as expressed. For each gene, the most common major transcript was considered for the analysis. Major transcripts tend to be recurrently detected in both normal and tumour samples, and there is a negative correlation between the number of annotated transcripts and the observed expression breadth (r_S: -0.44, p-value < $2.2{\cdot}10^{-16}$).
(c) Overlap in the set of major transcripts across healthy and tumour samples. There is higher consistency in major transcript identity in normal samples (p-value < $2.2 \cdot 10^{-16}$). N: normal; T: tumour.

3.2.2 Large and recurrent changes in splicing are rare

As a next step, and in order to address the magnitude of the above reported splicing alterations, I evaluated the extent of changes in the dominance and identity of major transcripts across conditions. Analysis of major transcript dominance from a patient-centric perspective revealed that, amongst the set of expressed genes, the number of genes with a dominant transcript differs significantly between the healthy and tumour tissues in all the 45 sample pairs (padj < $2.2 \cdot 10^{-16}$ in all the comparisons; see Methods). However, from a gene-centric perspective, only 602 genes show significant changes in major transcript dominance across conditions (see Methods), which represents a small fraction of the set of genes with reported alterations in splicing (*i.e.* 7.68% of the 7,842 genes previously identified). Even though the first observation is consistent with the broad alterations described in the previous section, the results from the gene-centric analysis suggest that most of those alterations do not lead to big changes in the abundance of major transcripts, at least in a recurrent fashion.

On the other hand, ~50% of the genes with altered splicing were predicted to undergo changes in the identity of their major transcript in at least one sample pair; cases that are further referred to as *switch events* (**Figure 3.5a**; see Methods). Switch frequency decreases rapidly as the stringency on their expression breadth is increased, suggesting that genes do not tend to be recurrently switched (*e.g.* ~65% of genes that undergo switch events do so in less than 50% of the patients where they are expressed; **Figure 3.5b-d**). Nonetheless, a small fraction of broadly switched genes can be detected, which are dominated by those expressed in a low number of samples (**Figure 3.5b-c**). Additional filtering of switch events that involve 2-fold and 5-fold dominant transcripts in each condition revealed a decrease in the proportion of genes being affected (*i.e.* 25% and 2% of genes in the study set, respectively; **Figure 3.5a**), as well as in the number of samples involved (**Figure 3.5d**). Altogether, these findings add to the idea that most of the big changes are not recurrently detected.

In the context of switch events, an interesting set of genes are those that are recurrently switched. An arbitrary threshold of >50% expression breadth was chosen to investigate such cases, leading to a total of 1,112 protein coding genes that fulfil this criteria, which were further analysed based on the potential

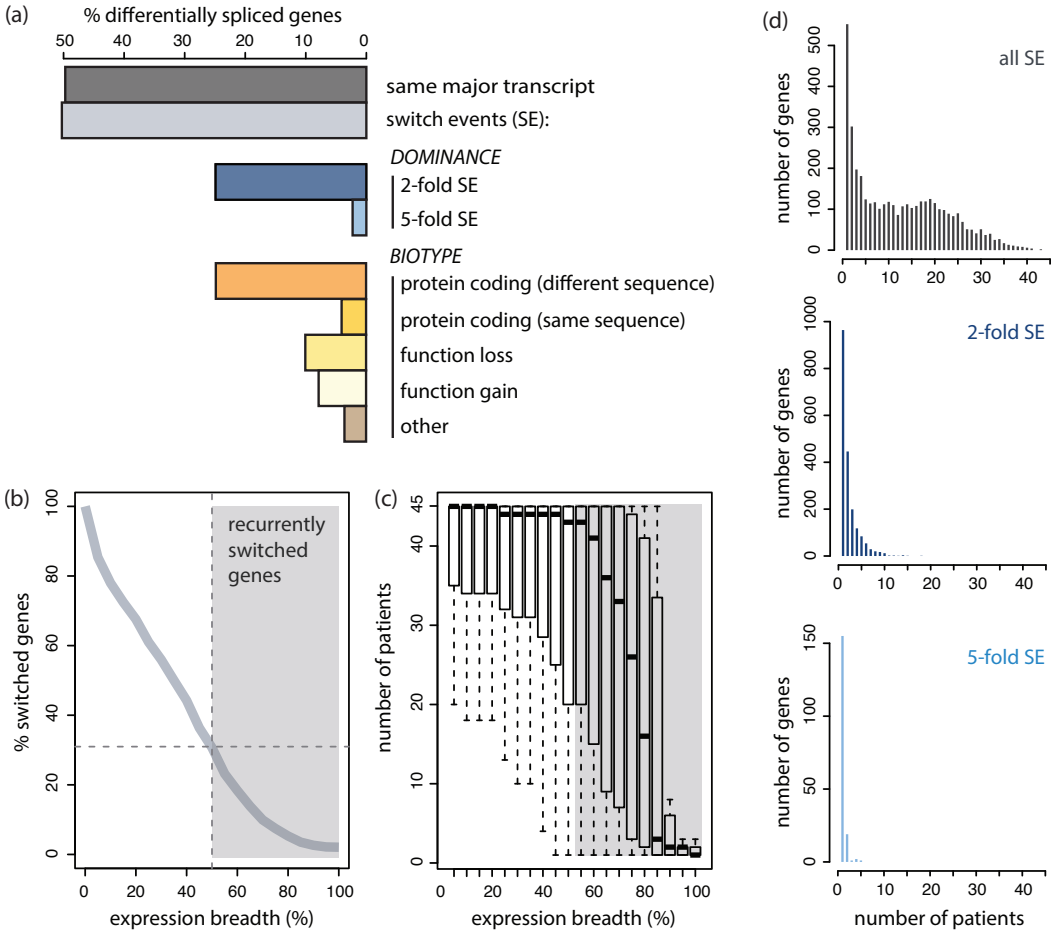


Figure 3.5| Switch events between tumours *vs.* healthy matched samples in protein coding genes.
(a) Percentage of differentially spliced genes involved in switch events. Half of the genes with altered splicing present changes in the identify of major transcripts across conditions in at least one sample pair. This percentage decreases when considering events that involve dominant transcripts. Most commonly, switch events in protein coding genes involve two different coding transcripts that differ in their protein sequence.
(b) Proportion of switched genes relative to their expression breadth. Genes are rarely switched in all the samples where they are expressed.
(c) Number of samples involved in switch events relative to switch expression breadth. Recurrently switched genes tend to be expressed in a lower number of samples.
(d) Frequency of switch events relative to the number of samples involved in the switch. In most cases, genes are not recurrently involved in switch events. This is specially true for events that involve dominant transcripts.

functional implications of the associated switch events. This information can be inferred from the transcript biotype information available in Ensembl [Flicek et al., 2012], which allows the classification of transcripts derived from protein coding genes into four main categories: protein coding, nonsense-mediated decay, retained intron and processed transcript (see Chapter 2 - Methods for a description). Thus, switch events from the set of recurrently switched genes can be classified depending on whether they involve (i) a protein coding transcript in each condition, (ii) a protein coding transcript in the healthy tissue but not in the tumour (function loss), (iii) the reverse situation (function gain), or (iv) any other combination of the above mentioned biotypes (other).

This classification system revealed that, in about half of the cases (*i.e.* 49%), those events involve two different protein coding transcripts which differ in their coding sequence (**Figure 3.5a**). *PPP2R4*, a Ser/Thr phosphatase that participates in the negative control of cell division and growth, stands out as a clear example in this category of events (**Figure 3.6a**). Notably, the recurrent major transcript in healthy samples is classified as the principal isoform by APPRIS [Rodriguez et al., 2013], while this is not the case for the one recurrently detected in tumours. A similar assessment can be obtained based on the crystal structure of the corresponding protein, where the smaller peptide that would derive from the major transcript detected in tumours is predicted to be non-functional by MAISTAS [Floris et al., 2011]. Following with the functional classification of switch events, it is also possible to identify cases where the two transcripts involved encode the same protein and consequently differ only in their UTR sequences. For example, a major transcript with longer UTRs was recurrently detected in tumour samples for the *ASPA* gene (**Figure 3.6b**), an N-acetylaspartate (NAA) scavenger. Alternatively, the splicing factor *SRSF6* and the transcriptional repressor *PATZ1* are clear examples within the function loss and function gain categories, respectively (**Figure 3.6c** and **Figure 3.6d**). Finally, switch events that involve dominant transcripts also constitute an interesting category, even though they are not as broadly detected (**Figure 3.5d**). For example, *FGFR2*, for which splicing alterations had already been reported in the context of ccRCC [Zhao et al., 2013], was detected to undergo a switch event in 22 out of 41 patients, which in 7 cases involved 2-fold dominant and coding transcripts in both normal and tumour samples.

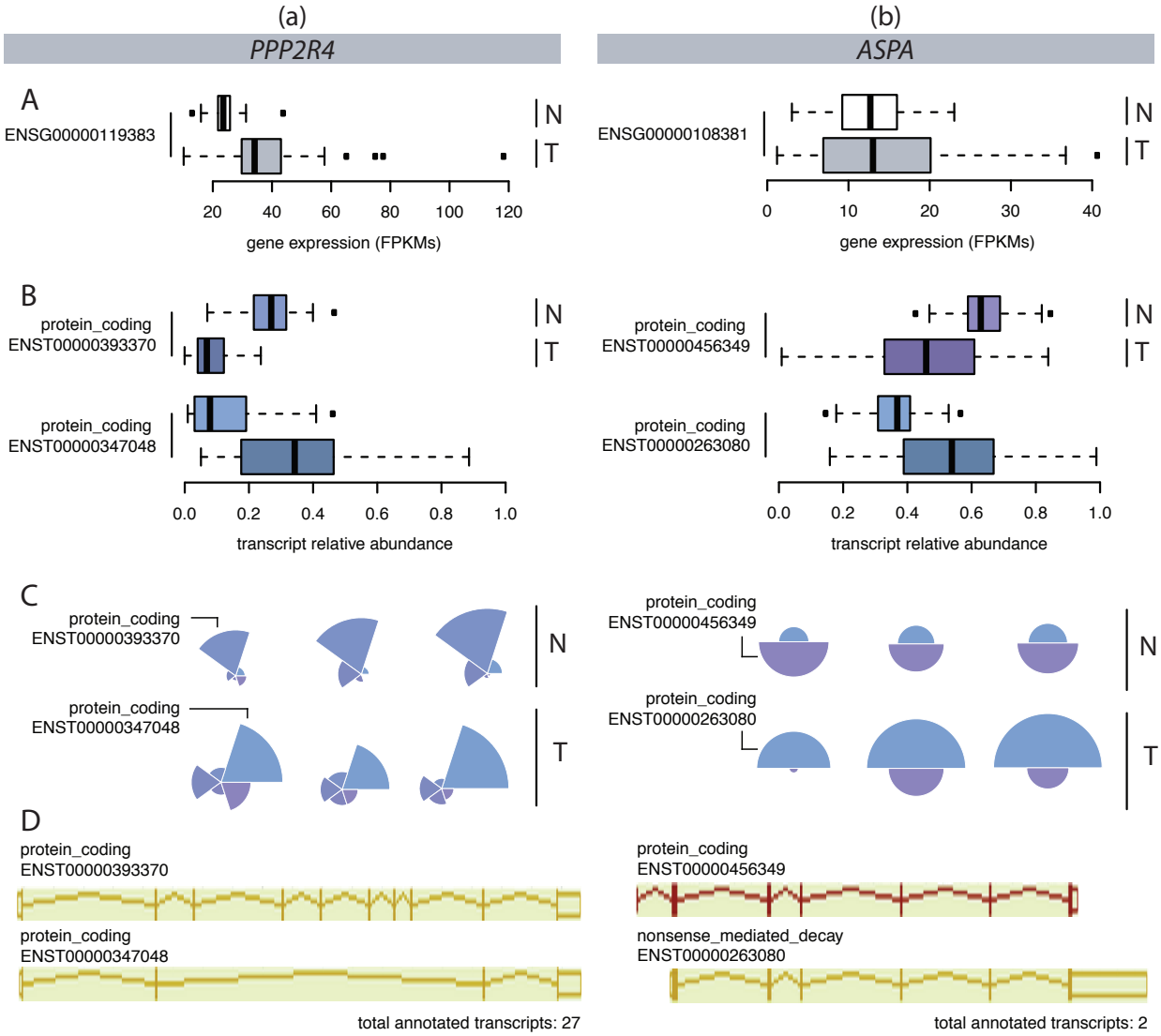


Figure 3.6 | Examples of recurrent switch events. Each panel contains information on the summarised gene and transcript expression profiles in tumours *vs.* healthy samples (A and B, respectively), as well as sample-specific expression profiles for a subset of three patients (C). The latter plot constitutes a novel visualisation strategy, where each sample is represented as a pie chart, and each of the slices corresponds to a transcript. The size of each slice is proportional to the transcript expression level, and the overall size of the plot is proportional to the gene expression level, thus allowing for comparisons across samples. Exon and intron structures for the highlighted transcripts are also included in D. None of the genes depicted here is differentially expressed.
(a) A switch between two protein coding transcripts for the PPP2R4 gene.
(b) A switch event between two protein coding transcripts that differ only in the UTRs for the ASPA gene.

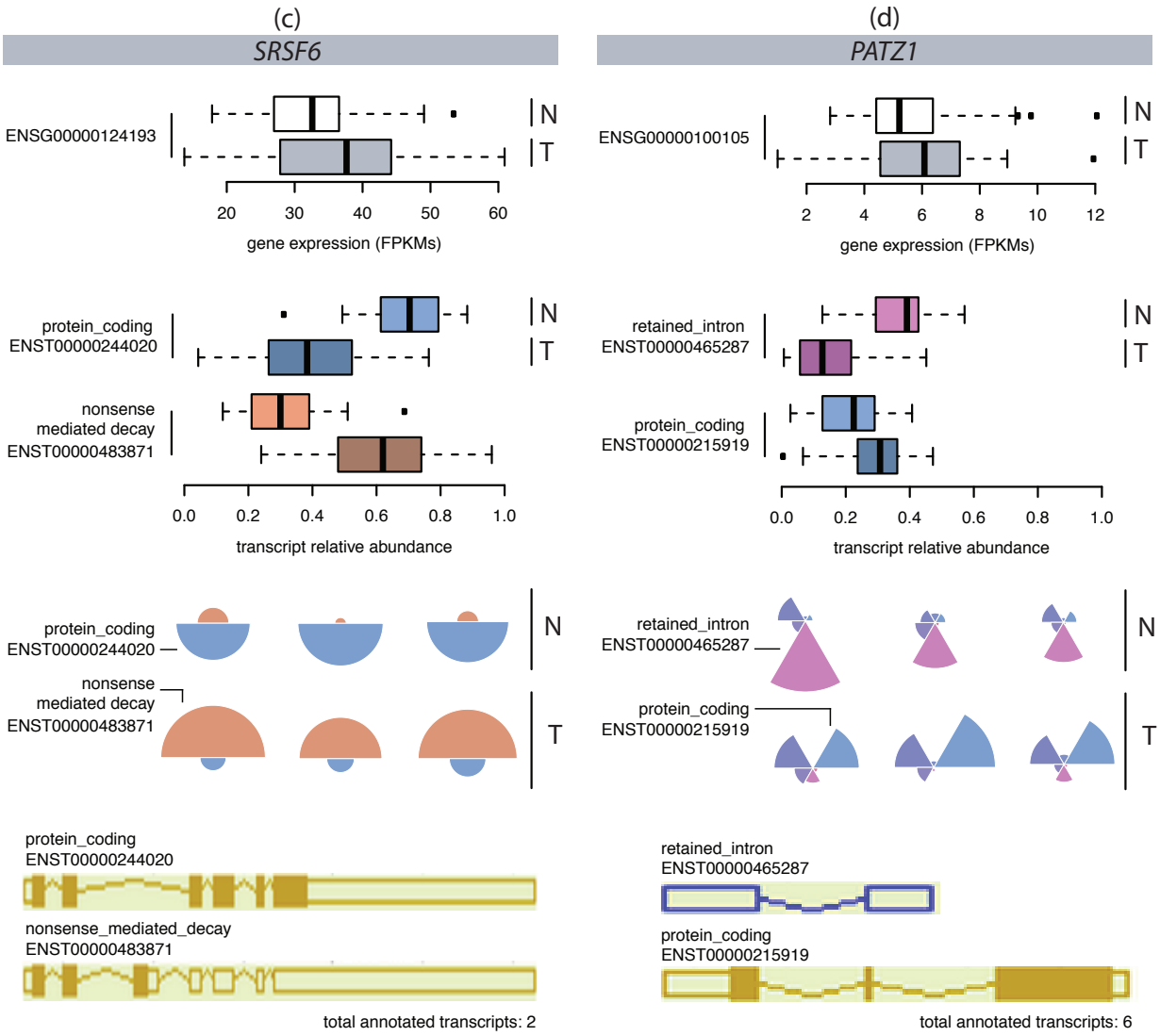


Figure 3.6 | Examples of recurrent switch events (continued).
(c) A switch event from a protein coding transcript to a nonsense-mediated decay one for the SRSF6 gene.
(d) A switch event from a retained intron to a protein coding one for the PATZ1 gene.

3.2.3 Identification, annotation and visualisation of switch events with SwitchSeq

The analysis workflow followed in the previous section, based on the detection and interpretation of switch events that affect protein coding genes, has been automated within the tool SwitchSeq [Gonzàlez-Porta and Brazma, 2014]. Briefly, this tool facilitates the interpretation of the changes in splicing reported by tools like DEXSeq [Anders et al., 2012] and MMDIFF [Turro et al., 2014], by letting the user focus on the most extreme cases (*i.e.* switch events).

SwitchSeq input consists of a matrix of normalised transcript-level counts, as well as several annotation files that can be easily obtained with the provided wrapper tool (**Figure C.1**). Based on the expression data provided by the user, the tool identifies the most abundant transcript within each gene and detects cases where its identity changes across conditions, here referred to as switch events. The detected events are then annotated by incorporating information from several public resources (Ensembl, APPRIS, UniPDB), and the results are reported in a self-contained HTML report, as well as in txt and json format. Furthermore, SwitchSeq produces plots for the visualisation of the identified events, similar to those depicted in **Figure 3.6**, which can be easily accessed from the report. Such plotting capabilities have been implemented in an independent R package (tviz), and hence can be used independently of any SwitchSeq execution.

3.2.4 Pathways appear as broadly disrupted when integrating different layers of information

Switch events that are not as commonly shared across patients are also interesting, specially if further alterations have been reported for the same gene in the same patient. With this question in mind, and focusing on the subset of 36 matched samples for which there is both RNA-seq and DNA-seq data available, I integrated sample-specific information on switch events with the results from other analyses performed in the context of CAGEKID, namely somatic mutations and gene expression changes. Out of the 585 genes with non-silent somatic mutations identified (see Methods), 226 overlap with those reported in previous studies [Sato et al., 2013; TCGA Network, 2013], and from those, a subset of 42 also undergo switch events in at least one patient. Such patient-specific alterations, including changes in gene expression, can be visualised in **Figure 3.7**. Notably,

VHL, which has been reported as the most frequently mutated gene in ccRCC, is not part of this subset of genes. On the other hand, *PBRM1*, the second most mutated gene in this type of cancer, is affected by switch events in 5 patients, which, interestingly, do not overlap with those in which this gene is mutated. Other genes that have been reported as significantly mutated in ccRCC and which are also affected by switch events include *PTEN*, *MTOR* and the recurrently switched *BAP1* [Sato et al., 2013; TCGA Network, 2013].

The same strategy can be used to investigate patterns in the alterations that affect components of relevant pathways in ccRCC (see section 3.2.1). For example, in the well characterised *VHL/HIF* pathway (**Figure 3.8**), such a visualisation approach makes it easy to discriminate genes that are predominantly altered at the genomic level (*e.g. VHL*) *vs.* cases in which transcriptomic alterations prevail, either as changes in gene expression (*e.g. EGLN3*, *SLC2A1*, *TGFB1*), splicing (*e.g. PIK3R*, *HIF1A*, *AKT2*) or both (*e.g. VEGFA*, *ETS1*). Similar patterns can be observed when focusing on the *PI(3)K-AKT-MTOR* pathway (**Figure C.2**), which also plays a relevant role in ccRCC, and the focal adhesion pathway (**Figure C.3**), located upstream of the former. Overall, such analyses evidence that, even though most genes are not altered in a recurrent fashion, pathways are broadly disrupted when combining information across multiple layers of information.

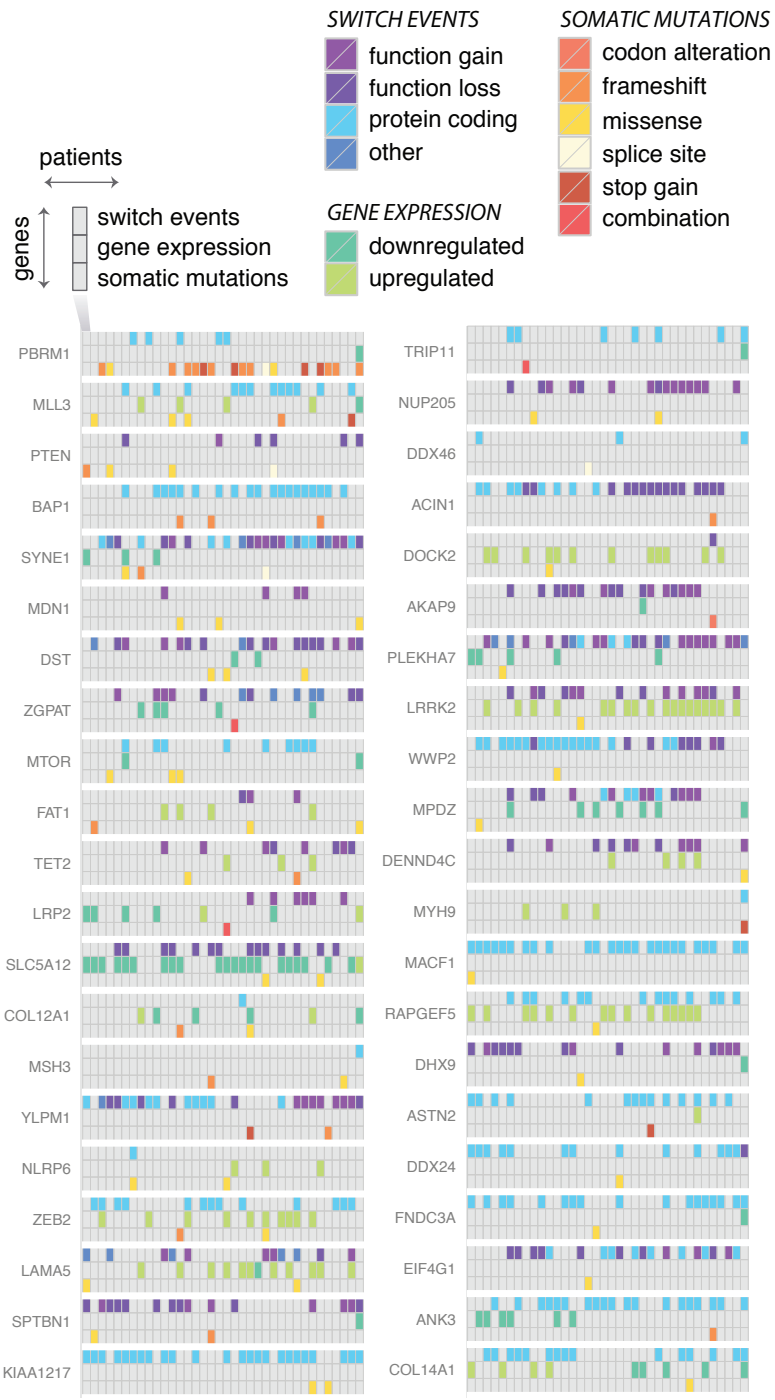


Figure 3.7| Patient-specific landscape of alterations for a subset of the genes with recurrent somatic mutations. A set of 42 genes with recurrent somatic mutations across three different studies and which undergo a switch event in at least one sample are represented here. For each gene, patient specific information was obtained by comparing the healthy and tumour samples and has been summarised in three tracks: switch event information (top), gene expression differences (middle) and confirmed somatic mutations (bottom). A gene is reported as either up or down-regulated in a given patient if there is at least a 2-fold change in its expression levels across conditions (*i.e.* 2-fold $\log_{10}$ FPKMs). Genes have been sorted regarding the total number of somatic mutations identified.

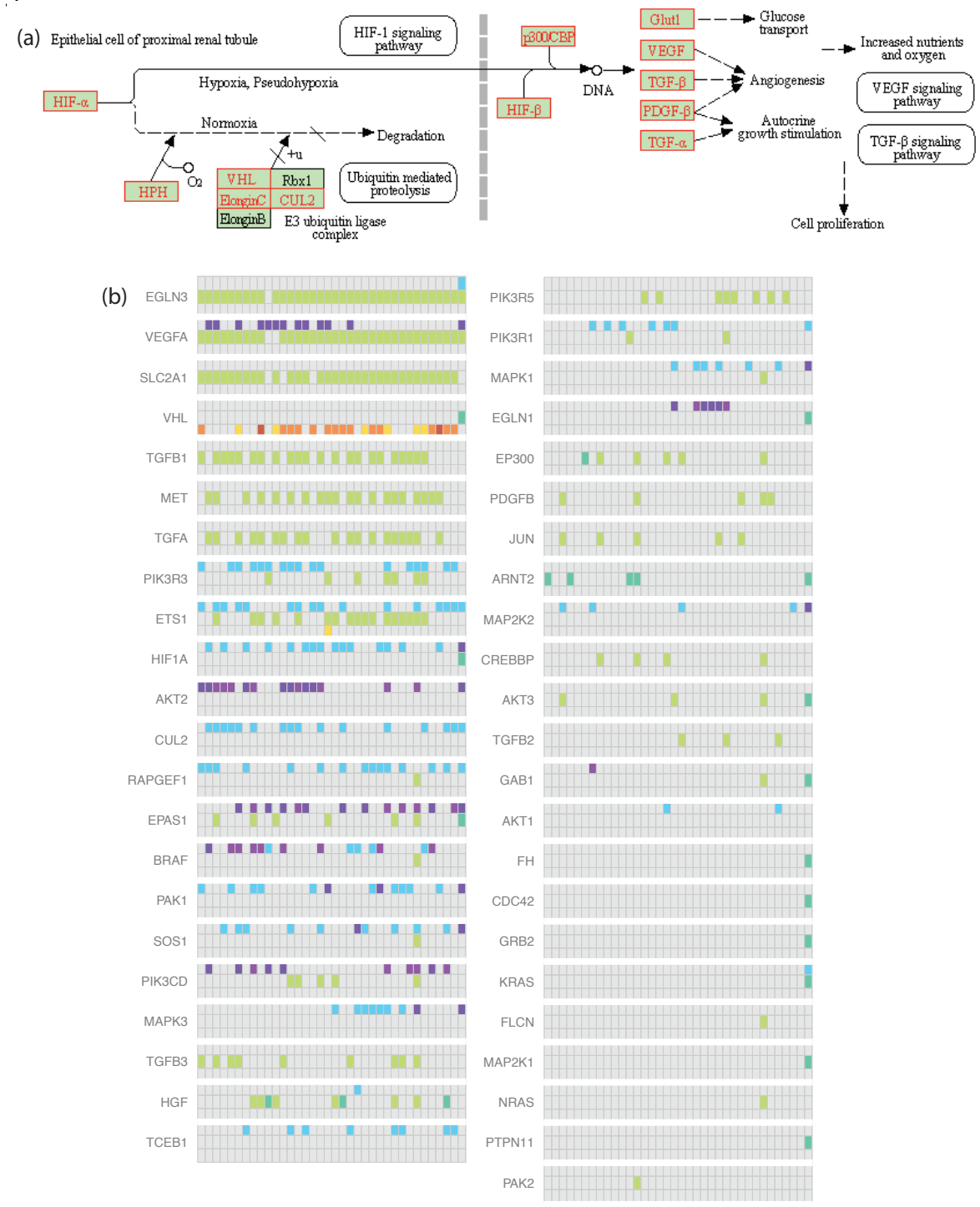
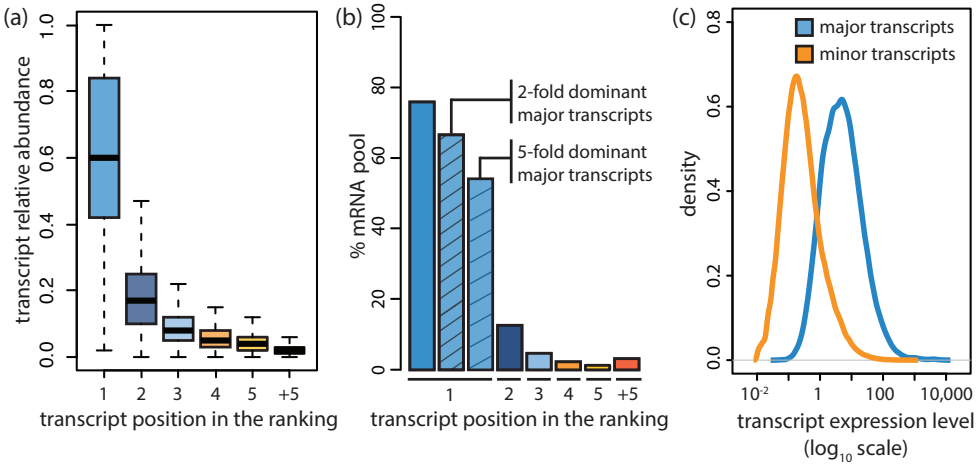


Figure 3.8 | Summary of detected alterations in the *VHL/HIF* pathway.
(a) The VHL/HIF *pathway.* Genes which have been identified as altered in any of the three considered analyses (*i.e.* switch events, gene expression changes and confirmed somatic mutations) have been highlighted in red.
(b) Patient-specific landscape of alterations for genes in the VHL/HIF *pathway.* Each gene is represented by three tracks as described in **Figure 3.7**, and have been sorted based on the total number of alterations of any kind. Only those previously highlighted in **a** are included here.

3.2.5 Splicing patterns in cancer cell lines are different from those in primary tissues

Cell lines are widely used as cancer models, and thus, it is important to understand how they differ from primary tissues. Several studies have evidenced differences in their transcriptional programs, reporting alterations in metabolic pathways, cell division and communication (*e.g.* Auman and Howard [2010]; Ertel et al. [2006]; van Staveren et al. [2009]), but changes in splicing have been so far unexplored. In this section, I will focus on the comparison of splicing patterns in primary tissues *vs.* cell lines by analysing a set of 6 ccRCC cell lines.

In general terms, and consistent with the results reported for primary tissues both here and in Chapter 2, ccRCC cell lines also express one dominant transcript per gene (**Figure 3.9**). Nonetheless, there are differences in the percentage of the mRNA pool that can be attributed to dominant transcripts (**Figure 3.9** *vs.* **Figure 3.3**), which constitute a first hint of differences in the splicing programs. For example, 2-fold dominant transcripts account for 83% of the mRNA pool in healthy tissues, while this percentage decreases to 74% and 67% for tumours and cell lines, respectively. In the same line, clustering analysis based on transcript relative abundances revealed that cell lines behave differently from primary tissues in terms of splicing patterns (**Figure 3.10**). Notably, matched samples clustered by disease status rather than by patient, thus indicating that the alterations described in the previous sections dominate above inter-individual variability. Closer inspection of splicing differences with DEXSeq revealed that there are 8,831 genes with at least one exon differentially used between cell lines and tumour samples. This represents 26% of the expressed genes (n = 33,982), a lower percentage than the one obtained when comparing healthy *vs.* tumour samples. Pathway enrichment analysis of the set of differentially spliced genes suggested a role in metabolism, cell cycle control and several signalling pathways (**Table C.1**), consistent with previous findings regarding differences in gene expression [Ertel et al., 2006].



(d)

	expressed genes	genes with a dominant major transcript	
		2-fold dominance	5-fold dominance
1 FPKM	9,226 (50.01%)	5,962 (64.62%)	3,429 (37.16%)
5 FPKM	5,936 (32.18%)	4,017 (67.66%)	2,480 (41.79%)
10 FPKM	3,949 (21.40%)	2,733 (69.21%)	1,759 (44.54%)

Figure 3.9 | Most protein coding genes express one dominant transcript in cell lines. Panels to be be interpreted as in **Figure 3.3**.

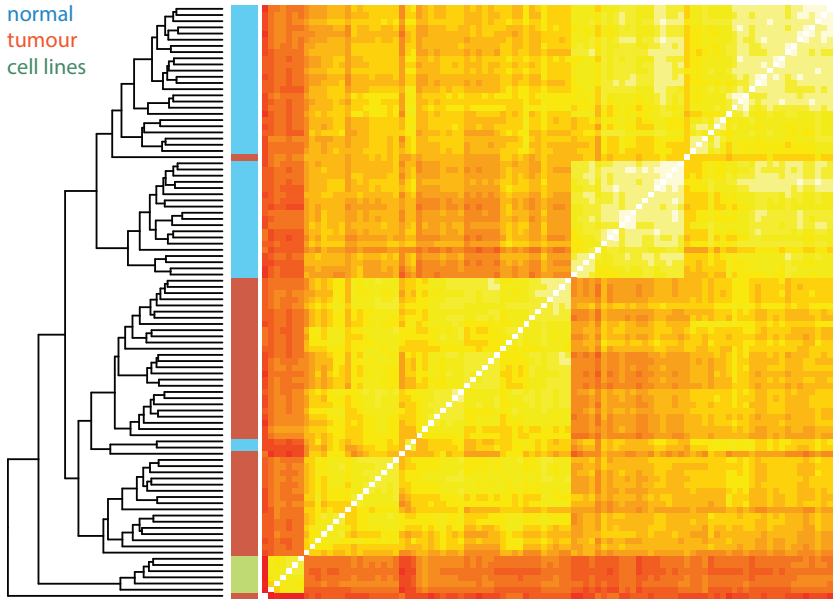


Figure 3.10 | Splicing patterns in cancer cell lines *vs.* primary tissues. Samples were clustered based on transcript relative abundances. There is a clear separation between cell lines (green) and primary tissues, as well as between matched healthy (blue) and tumour samples (red). This suggests that cancer cell lines differ from primary tumours in their splicing patterns and that disease status prevails over inter-individual variability.

Moreover, the extent to which ccRCC cell lines resemble primary tumours in terms of switch events was also investigated. In this context, a potentially interesting set of genes are those which had been previously identified as recurrently switched (n = 1,112). Focusing on the most common switch event for each gene, the overlap in the identity of the involved transcripts and the most recurrent major transcript in cell lines was evaluated, thus leading to four possible different scenarios: (i) cases where the most recurrent major transcript in cell lines coincides with the one detected in tumours; (ii) cases where it coincides with the one identified in normal samples; (iii) cases where there is no overlap or where there are several recurrent transcripts that can be ambiguously classified; and (iv) cases where the corresponding gene is not expressed in cell lines. Such classification revealed that recurrently expressed major transcripts in ccRCC cell lines overlap significantly more often with those detected in tumours than with those detected in healthy tissues (**Figure 3.11a**; Chi-square test, p-value = $5.531 \cdot 10^{-08}$). For example, the previously highlighted gene *PPP2R4* has consistent splicing patterns between cell lines and tumours (**Figure 3.11b**). However, this situation represents only 32% of the cases, thus suggesting that, most of the time, major transcript expression patterns are not conserved. Surprisingly, *SRSF6*, for which a clear switch event has been reported earlier in this chapter, emerges as an example of the divergence in splicing patterns between cell lines and tumours (**Figure 3.11c**). Finally, 19% of the genes that had been identified as recurrently involved in switch events are not detected as expressed in cell lines.

3.3 Discussion

The results described in this chapter constitute the first in-depth characterisation of the splicing alterations that underlie ccRCC, the most common type of renal cancer [Jonasch et al., 2012]. Formerly, most of the efforts towards the characterisation of this disease have focused on the genomic level, with recent significant advances derived from three different large scale sequencing projects [Sato et al., 2013; Scelo et al.; TCGA Network, 2013]. The few existing attempts to understand the potential disruption of splicing programs in ccRCC have mostly been based on the use of differential exon expression as a measure of exon skipping [Valletti et al., 2013; Zhao et al., 2013]. However, such approach does not cover the totality of annotated splicing events, as evidenced by the differences in the number of genes reported to undergo differences in splicing (*i.e.* hundreds in previous

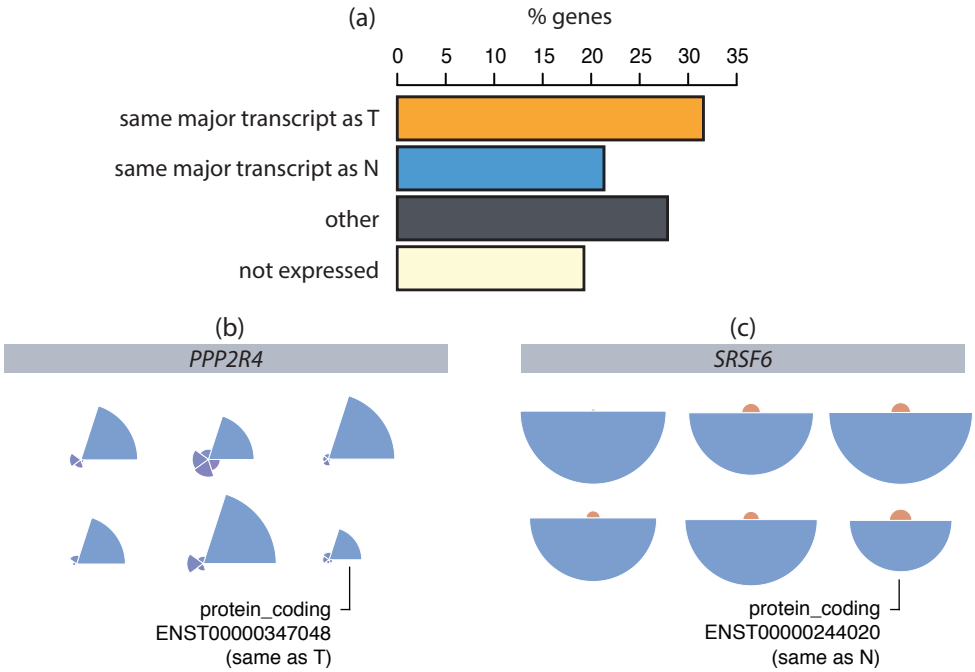


Figure 3.11 | Switch events in cancer cell lines *vs.* primary tissues.
(a) Overlap in the identity of the most recurrent major transcript in cell lines and those involved in recurrently switch events. Major transcripts in ccRCC cell lines overlap significantly more often with those detected in tumours than with those detected in healthy tissues (Chi-square test, p-value = $5.531 \cdot 10^{-08}$). N: normal; T: tumour. *(b-c) Sample-specific transcript usage in the six cell lines analysed for a subset of genes previously identified to undergo switch events.* The major transcripts detected in cell lines tend to coincide with those detected in primary tumours (*e.g. PPP2R4*, see **Figure 3.6a**); however there are also genes for which the most abundant transcript in cell lines corresponds to the one detected in healthy samples (*e.g. SRSF6*, see **Figure 3.6c**).

studies *vs.* thousands in the present chapter). Thus, I show that splicing patterns are largely altered in this type of cancer, consistent with the wide alterations already described at the gene level [Sato et al., 2013; Scelo et al.; TCGA Network, 2013]. Furthermore, these observations link to the identification of the mRNA processing pathway as commonly disrupted in ccRCC [Sato et al., 2013].

GO and pathway enrichment analysis of the set of genes that undergo differential splicing in ccRCC points to a significant alteration of metabolic pathways, including oxidative phosphorilation, the citrate cycle, glycolisis and the pentose phosphate pathway. This is suggestive of a general metabolic shift and consistent

with the Warburg effect, which constitutes a hallmark of ccRCC [Linehan et al., 2010]. Moreover, such analyses identified a dysregulation of the focal adhesion pathway, which contains genes that act upstream of *PI(3)K*. Abnormalities of the extra-cellular matrix play key roles in tumour formation and progression [Cox and Erler, 2011] either by affecting downstream growth-promoting pathways [Levental et al., 2009] or by contributing to angiogenesis and metastasis by influencing the tumour microenvironment [Lu et al., 2012]. Together with the recurrently altered *PI(3)K-AKT-MTOR* signalling pathway [Sato et al., 2013; TCGA Network, 2013], this novel finding highlights the relevance of this signalling cascade as a therapeutic target for ccRCC.

In spite of the large number of differentially spliced genes identified, I observe that big changes in splicing do not tend to be recurrently detected. These observations are consistent with the abundant molecular heterogeneity identified in ccRCC [Gerlinger et al., 2012; Martinez et al., 2013] and, more generally, with the high levels of inter-tumour variability described for cancer [Cusnir and Cavalcante, 2012; Shibata and Shen, 2013]. By focusing on major transcripts, I estimate that, amongst the genes with differential splicing, only a small subset manifest also differences in major transcript dominance between the healthy and tumour tissues. Following the introduction of the concept of *switch event* as a means to explore the most extreme changes in splicing, I detect that most of these events can be detected only in a small number of patients (*e.g.* ~65% of genes that undergo switch events do so in less than 50% of the samples where they are expressed). Nonetheless, it is also possible to identify interesting cases of recurrent switch events. For example, I predict a non-functional major transcript in tumours for *PPP2R4*, a Ser/Thr phosphatase involved in the negative control of cell division and growth. Similarly, I detect a switch between a protein coding transcript and a nonsense-mediated decay one for the splicing factor *SRSF6*, which has been reported as an oncogene [Michal et al., 2013]. Interestingly, this is consistent with existing knowledge that levels of core components of the splicesome (SmB/B') can be auto-regulated through feedback loops that rely on NMD [Saltzman et al., 2011]. Overall, these analyses evidence that many of the detected splicing differences arise from lowly abundant transcripts, which could potentially result from noisy splicing. Nonetheless, similarly to what has been discussed in Chapter 2, one cannot rule out the possibility that minor transcripts may have a functional impact on the cell's fate. mRNA and protein levels are

only correlated up to some extent, and hence it is possible that minor transcripts are still translated. Even if this is not the case, they might still play a functional role in the regulation of expression levels [McGlincy and Smith, 2008; Yap et al., 2012]. Thus, evaluating the functional relevance of the annotated transcripts is one of the challenging tasks ahead. In this context, incorporating information on the conservation of protein sequences and features across species could help in identifying potentially functional minor transcripts, and the disruption of key protein domains and structures compared to major transcripts could indicate functional defects [Rodriguez et al., 2013].

Beyond the general comparison of normal *vs.* tumour samples, patient-specific analysis of transcriptomic and genomic alterations allowed for a more clear picture of the complexity underlying this type of cancer. For example, I show that *VHL* is predominantly altered at the genomic level. On the other hand, I observe that *PBRM1*, the second most commonly mutated gene in ccRCC, is also affected by switch events in a small number of patients. This is also the case for *BAP1*, identified amongst the most mutated genes in this type of tumour, and *MTOR*, a component of the *PI(3)K-AKT-MTOR* pathway that constitutes a strong therapeutic target for ccRCC [Audenet et al., 2012]. In addition, I observe that despite the low recurrence of the detected splicing changes, pathways are broadly altered when aggregating information across different regulatory layers. This evidences the need for integrative analysis in complex diseases such cancer, and highlights the importance of sample-specific comparisons in uncovering such complexity.

Finally, following analysis of a set of 6 ccRCC cell lines, I predict overall differences in the splicing patterns of cell lines *vs.* primary tissues when clustering samples based on transcript relative abundances. Several studies had already evidenced differences in transcriptional programs for cell lines (*e.g.* Auman and Howard [2010]; Ertel et al. [2006]; van Staveren et al. [2009]), but changes in splicing patterns have remained uncharacterised. Focusing on genes that had been previously identified as recurrently switched, I show that recurrently expressed major transcripts in ccRCC cell lines overlap significantly more often with those detected in tumours than those from healthy tissues. However, in most cases (*i.e.* 68%), major transcript expression patterns are not conserved between cell lines and tumours, as exemplified by *SRSF6*. Overall, such observations

reinforce the idea that transcriptome analysis in cell lines can be extended to primary tissues only to some extent [Lukk et al., 2010].

Altogether, the findings described in this chapter constitute the tip of the iceberg for a landscape of large transcriptomic alterations in ccRCC. By analysing alternative splicing patterns, it has been possible to detect changes that would not have been picked up by a gene-level analysis, which might as well constitute actionable therapeutic targets [Tang et al., 2013]. Nonetheless, there are still many more avenues to cover in order to gain a deep understanding of the ccRCC transcriptome, including novel transcription and novel transcripts, as well as potential alterations affecting the expression of long non-coding RNAs (lncRNAs) and micro-RNAs (miRNAs). RNA-seq has undoubtedly facilitated the study of the transcriptome, and as the costs of sequencing continue to fall down, having access to a more integrated picture through the combination of different layers of information (*e.g.* genomics, transcriptomics, epigenetics, etc.) will not only enable the discovery of novel biomarkers in research, but will soon constitute a reality in the clinics.

3.4 Computational methods

All the computational analyses described in this chapter have been performed by myself, with the exception of the differential expression and somatic mutation analyses, which have been carried out by Dr. Johan Rung and Dr. Louis Letourneau, as detailed below.

Datasets and mapping

RNA from 45 healthy kidneys and matched ccRCC tumours was extracted and polyA-selected following Illumina's standard protocols. Sequencing was performed on an Illumina HiSeq 2000 platform (100 bp paired-end reads; **Table A.2**). In addition, DNA-seq data for 36 of the 45 matched samples was also produced with the same platform (100 bp paired-end reads). In all the cases, revision by multiple pathologists ensured a 70% content of tumour cells in the ccRCC samples. Finally, RNA from a separate set of 6 ccRCC cell lines was extracted and prepared for sequencing following the same Illumina's protocols and platform (100 bp paired-end reads; **Table A.2**).

Raw RNA-seq reads were initially trimmed down to 95 nucleotides from the 3'

end using PRINSEQ v0.19.5 [Schmieder and Edwards, 2011] and mapped to the reference genome (Ensembl 66 [Flicek et al., 2012]) using TopHat v2.0.6 [Kim et al., 2013]. The DNA-seq data were analysed by Dr. Mark Lathrop's group at McGill University and Genome Quebec Innovation Centre (see Integrative analyses).

Gene level analyses

Analyses performed by Dr. Johan Rung:
Gene counts were estimated using HTSeq v0.5.3p7 [Anders et al., 2014], using the `--intersection-nonempty` and `--stranded=no` parameters. Genes with zero counts for more than nine tumour or normal samples were filtered out, and the rest were considered as expressed. A paired test for differential expression was performed for the 45 matched sample pairs using edgeR [Robinson et al., 2010], and significance was assessed with an FDR threshold of 0.01 (Benjamini & Hochberg correction [Benjamini and Hochberg, 1995]). Gene FPKMs were obtained from the calculated counts using custom scripts.

Analyses performed by myself:
Gene expression levels of ccRCC cell lines *vs.* primary tumours were compared with edgeR as detailed above. Following the previous strategy, only genes with non-zero counts for five or less samples were considered as expressed.

GO and pathway enrichment analyses were performed with DAVID [Huang et al., 2009a,b] and WebGestalt [Wang et al., 2013], and significance was assessed with an adjusted p-value of 0.01 (Benjamini & Yekutieli correction [Benjamini and Yekutieli, 2001]). All the information on gene function was retrieved from Genecards [Safran et al., 2010] unless otherwise indicated in the text.

Transcript level analyses

Differential splicing and splicing variability

DEXSeq v1.7.0 [Anders et al., 2012] was used to identify genes that undergo differential exon usage across the two studied conditions (*i.e.* healthy kidney *vs.* ccRCC samples). An FDR threshold of 0.01 was used to assess significance (Benjamini & Hochberg correction [Benjamini and Hochberg, 1995]).

Splicing variability was calculated following the method introduced by Gonzàlez-

Porta et al. [2012]. Briefly, for each gene, the relative abundances of the annotated transcript isoforms for all the individuals in the study set (normal or tumour) are represented in a multi-dimensional space, with the number of dimensions corresponding to the number of transcripts. Then, the mean Hellinger distance to the centroid of all the data points (individuals) is taken as a measure of splicing variability (*i.e.* the higher the observed dispersion, the bigger the variability in transcript relative abundances within individuals of the study set).

Major transcripts and switch events

Relative abundances for the transcripts annotated in Ensembl 66 were obtained using MISO v0.4.1 [Katz et al., 2010]. Transcript FPKMs were then obtained by multiplying those relative abundances with the corresponding gene FPKMs. Similarly to the nomenclature introduced in Chapter 2, the most abundant transcript within each gene is referred to as *major transcript*.

Major transcript dominance was assessed by calculating, for every gene, the ratio of the expression levels between the second most abundant transcript and the major one. For each sample, the number of expressed genes with a dominant transcript (*i.e.* 2-fold dominant major transcript) was calculated. Then, the difference in the numbers obtained for the normal and tumour samples from each given patient was assessed with a McNemar's test (patient-centric analysis). Similarly, for each gene in the set of differentially spliced genes, the difference in major transcript dominance between all normal *vs.* tumour samples was interrogated with a Wilcoxon test (gene-centric analysis). In both analyses, p-values were corrected for multiple testing using the Benjamini-Hochberg method [Benjamini and Hochberg, 1995].

Changes in major transcript identity across conditions were referred to as *switch events* (see **Figure 1.12c** for an illustration). A given switch event was classified as *dominant* if both transcripts involved are expressed in a dominant fashion. Formally, given a gene G, a pair of transcripts I_k and I_l and two paired samples S_i and S_j from patient P, we say that gene G undergoes an x-fold switch between transcripts I_k and I_l in samples S_i and S_j, if G is expressed in both S_i and S_j and the ratio of the expression of I_k to I_l is at least x in S_i and no more than $1/x$ in S_j. In addition, switch events were also classified as *recurrent* if they were present in >50% of the samples in which the corresponding gene was expressed.

Cluster analysis

Spearman correlation coefficients were calculated by comparing transcript relative abundances across samples in a pairwise fashion and used to cluster samples using the R function `heatmap.2`.

Integrative analyses

Analyses performed by Dr. Louis Letourneau:
Non-silent somatic mutations for a total set of 585 genes were identified with Genome MuSiC v0.4 [Dees et al., 2012] (FDR < 0.2) and validated by pooled amplicon sequencing on an Illumina MiSeq or by Sanger sequencing.

Analyses performed by myself:
A gene is reported as either up or down-regulated for a given patient if the observed difference in expression levels across matched samples is more than 2-fold (*i.e.* >2-fold $\log_{10}$ FPKMs).

Pathway information was retrieved from KEGG [Kanehisa and Goto, 2000; Kanehisa et al., 2014] and reproduced with the permission of Kanehisha Laboratories.

Chapter 4

The regulation of splicing by core spliceosomal factors

Gaining a comprehensive knowledge of all the factors involved in intron removal and their specific contribution in shaping splicing decisions is the key to understanding the basis of this reaction and how its disruption might contribute to pathogenesis [Barash et al., 2010]. This chapter describes a collaborative effort to further characterise the function of the splicing factor *PRPF8*, and to dissect the role of core spliceosomal components in the regulation of the splicing reaction. In the first part, I describe the phenotypic effects obtained following knock-down of this splicing factor. Next, I link such effects to alterations in splicing, and I identify common characteristics of the set of genes that are most affected by such down-regulation. Finally, I evaluate the impact of *PRPF8* knock-down on the rate of splicing as a potential explanation for the observed effects.

All the computational analyses have been performed by myself under the supervision of Dr. John Marioni from EMBL-EBI, unless otherwise indicated in the Methods. The experimental work has been carried out by Dr. Vi Wickramasinghe, Dr. David Perera and Arthur Bartolozzi in Dr. Ashok Venkitaraman's laboratory at the Hutchison Research Institute. Finally, Dr. Christopher W. Sibley from Dr. Jernej Ule's laboratory at University College London produced and analysed the iCLIP dataset.

Publications derived from this chapter

- (submitted) Wickramasinghe*, Gonzàlez-Porta* et al. PRPF8 abundance dictates the patterns of alternative and constitutive messenger RNA splicing.

*shared first authors

4.1 Introduction

The splicing reaction is a multi-step process in which over one hundred proteins take part, typically referred to as splicing factors (SFs). Together with several small nuclear RNAs (snRNAs), these proteins form a large protein complex - the spliceosome - which has been widely investigated in terms of composition, function and structure (*e.g.* reviewed in Hoskins and Moore [2012]; Matera and Wang [2014]). Amongst the collection of SFs that participate in splicing, the core spliceosomal component *PRPF8* stands out because of its location at the centre of the spliceosome [Galej et al., 2013; Grainger and Beggs, 2005]. This SF is part of the U5 small nuclear ribonucleoprotein particle (snRNP) and the B-complex, and has been suggested as a master regulator of splicing [Grainger and Beggs, 2005]. More specifically, *PRPF8* is known to act as a scaffold during spliceosomal assembly, and to be involved in the activation of the B-complex through the recruitment of the the U4-U6·U5 tri-snRNP ([Li et al., 2013]; see Introduction - The splicing reaction). Notably, this SF has been established as a candidate gene for autosomal dominant retinitis pigmentosa, a genetic disease that affects specifically retinal cells and that is characterized by a reduction in splicing activity [Tanackovic et al., 2011]. Furthermore, similarly to other SFs, it has been detected to play an essential role in cell division [Neumann et al., 2010]. However, despite its key involvement in the splicing reaction, its clinical relevance and its requirement for proper cell cycle progression, much of the structure and mode of action of *PRPF8* remain uncharacterised.

In this chapter, I describe the work carried out in collaboration with Dr. Ashok Venkitaraman's and Dr. Jernej Ule's teams in order to further elucidate the function of *PRPF8* in the context of its associated mitotic arrest phenotype. Following down-regulation of this SF with small interference RNAs (siRNAs), we confirmed and characterised the resulting disruption in cell cycle progression. Moreover, by introducing properly spliced mRNAs into *PRPF8* knock-down (KD) cells, we show that such phenotype can be attributed to splicing defects, rather than to a direct role of *PRPF8* in cell division, as reported for other splicing factors [Hofmann et al., 2010]. Further results from Sm iCLIP experiments indicate that spliceosomal binding is affected at a global scale, hence pointing to an overall disruption of splicing. We further generated and analysed an RNA-seq dataset in order to compare splicing patterns in control *vs. PRPF8* KD cells, and consistent with the

above results, we observed higher intronic expression levels in the latter. We also detect differences in splicing for a subset of the expressed genes (n = 3,388 out of 13,216 protein coding genes) that are enriched for those encoding mitotic cell cycle factors, thus linking the identified alterations to the observed phenotype. Focusing on intronic reads, we identify a subset of introns that remain preferentially unspliced after *PRPF8* KD, and find that they harbour weaker splice sites and GC composition similar to that of adjacent exons, characteristics that have been previously linked to intron retention [Amit et al., 2012; Sakabe and de Souza, 2007]. Mini-gene experiments further show that increasing splice site strength can compensate for the down-regulation of *PRPF8*, hence suggesting that the kinetic competition across splicing signals could constitute an explanation for the observed alterations. Finally, and consistent with this kinetic competition model, we predict and validate changes in the rate of splicing following *PRPF8* KD by relying on novel computational analyses based on intronic reads. Altogether, these findings highlight *PRPF8* as a key player for the completion of splicing in temporally constrained processes such as cell division and reveal an intimate role of this SF in regulating the dynamics of splicing.

4.2 Results

The findings described here are the result of a collaborative effort, which comprised the use of a variety of experimental and computational approaches for the purposes stated above. Even though the present chapter includes an overall description of the project, the emphasis has been placed on the analysis of the RNA-seq data, in which I have been most involved.

4.2.1 *PRPF8* knock-down causes accumulation of cells in mitosis

Following down-regulation of *PRPF8* (**Figure D.1**), it was possible to observe that cells with reduced expression of this splicing factor did not progress through the cell cycle and tended to accumulate in mitosis (**Figure 4.1a**). Moreover, *PRPF8* knock-down (KD) cells manifested mitotic abnormalities that could provide an explanation for the failure to exit this phase, as exemplified by the severe chromosome misalignment observed from chromosome alignment assays (**Figure 4.1b**). Finally, knock-down of other B-complex components also led to a mitotic arrest phenotype, while this was not generally the case for A and C-complex components (**Figure D.2**; see Discussion).

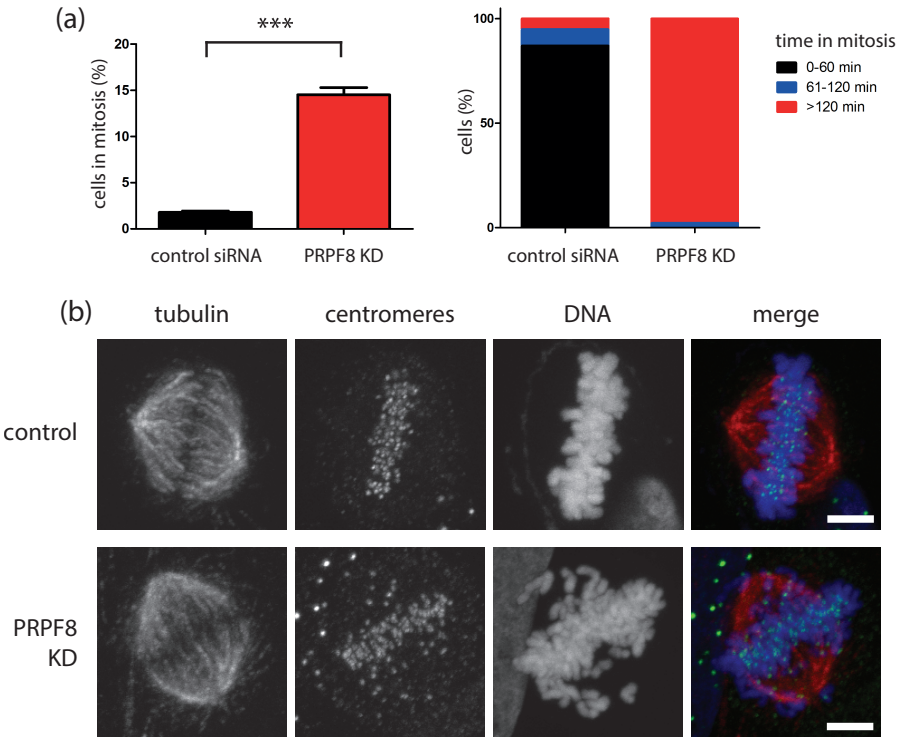


Figure 4.1| Characterisation of the mitotic arrest phenotype obtained after *PRPF8* KD.
(a) Results from the cell cycle analyses. Cells with reduced expression of this splicing factor tend to accumulate in mitosis (p-value < 0.001).
(b) Results from the chromosome alignment assays. Chromosome alignment assays measure whether chromosomes are aligned correctly at the metaphase plate. Compared to controls, *PRPF8* KD cells display lagging chromosomes.
Figure provided by Dr. Vi Wickramasinghe.

4.2.2 The mitotic arrest phenotype is driven by disruptions in splicing

Given that *PRPF8* is a component of the catalytic core of the spliceosome, it is likely that the observed phenotype is caused by alterations in splicing. In order to evaluate this hypothesis, *PRPF8* KD cells were treated with a cytosolic extract from control cells that contained polyA-selected mRNAs. Such treatment led to partial phenotype recovery in *PRPF8* KDs (**Figure 4.2**), suggesting that the lack of properly spliced mRNA is affecting cell cycle progression. Similar results were obtained from further Sm iCLIP experiments, a technique that allows for the study of spliceosomal protein-RNA interactions along the genome at nucleotide resolution [Briese et al.; König et al., 2010]. These experiments evidenced lower spliceosomal

binding in *PRPF8* KD cells, thus pointing to global defects in the splicing reaction (**Figure 4.3** and **Figure D.3**). Notably, most of the differences in spliceosome occupancy emerged at the 5′ splice site, consistent with the preferential interaction of *PRPF8* with this splicing signal [Li et al., 2013].

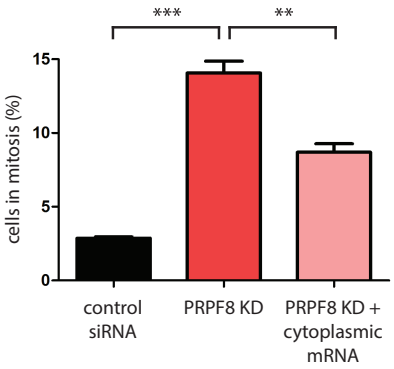


Figure 4.2 | Phenotype recovery experiments in *PRPF8* KD cells. Cytoplasmic mRNAs were isolated from control cells and introduced in *PRPF8* KDs. Such treatment revealed a decrease in the proportion of cells arrested in mitosis, suggesting that the observed alterations in cell cycle progression are caused by splicing defects (control *vs. PRPF8* KD: p-value < 0.001; *PRPF8* KD *vs. PRPF8* KD + cytoplasmic mRNA: p-value < 0.01). *Figure provided by Dr. Vi Wickramasinghe.*

To gain further insights into the splicing alterations that led to the observed mitotic arrest, the transcriptome of *PRPF8* KD and control cells was sequenced on an Illumina HiSeq2000 using 100 bp paired-end reads (4 and 3 biological replicates, respectively; see Methods and **Table A.3**). Analysis of this RNA-seq dataset revealed that intronic expression levels are higher in *PRPF8* KDs compared to control cells (p-value < $2.2 \cdot 10^{-16}$; **Figure 4.4a** and **Figure 4.4b**; see Methods), thus pointing to an overall misregulation of splicing patterns, as initially suggested by the iCLIP experiments. Following such observation, DEXSeq was used to detect intron retention events (see Methods), and this analysis led to the identification of a set of 2,086 protein coding genes that contain at least one retained intron. Similarly, from the execution of DEXSeq on annotated exons it was possible to detect a set of 1,921 protein coding genes with significant differences in exon usage across conditions (FDR < 0.01; see Methods). These results suggest that down-regulation of *PRPF8* not only impairs intron removal, but also affects alternative splicing patterns. Notably, there is a significant overlap between this

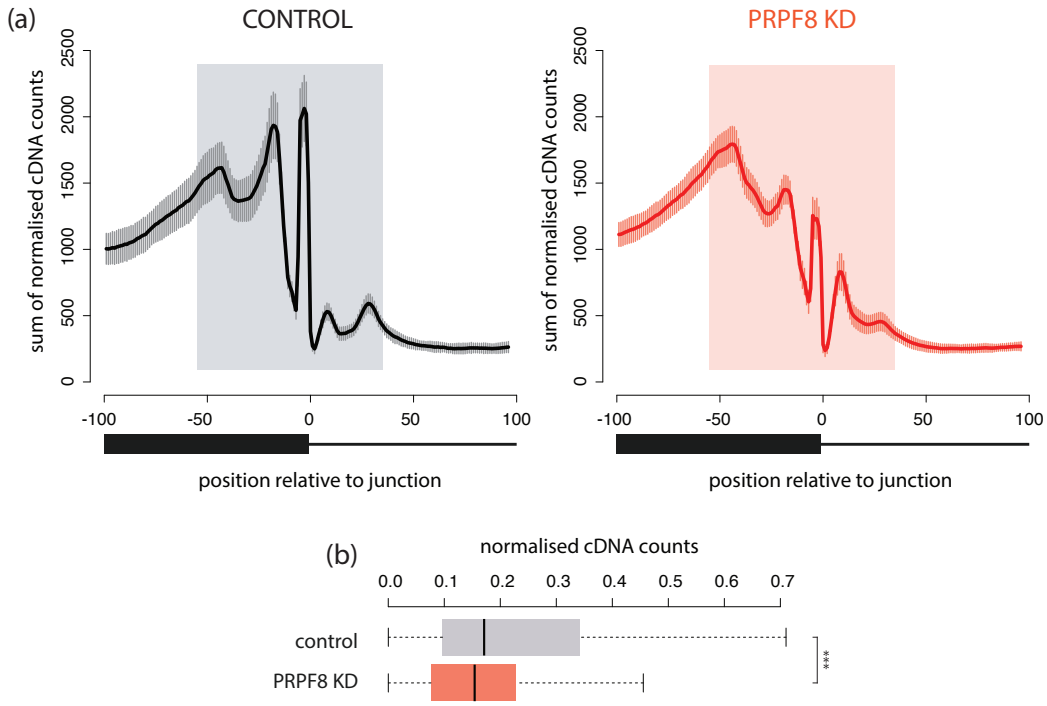
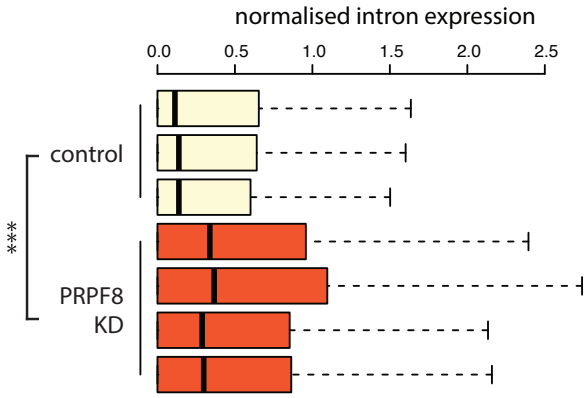


Figure 4.3| Results from Sm iCLIP experiments for the 5′ splice site.
(a) RNA maps for control and PRPF8 KD samples. RNA maps provide a summarised view of the detected spliceosomal binding sites across all exon-intron junctions, where each peak corresponds to different RNA-binding proteins [Briese et al.]. Such maps suggest that the down-regulation of *PRPF8* leads to broad differences in the studied RNA-protein interactions. Mean + SD across replicates have been plotted for each condition (see Methods).
(b) Changes in the distribution of normalised cDNA counts between control and PRPF8 *KD samples.* Normalised cDNA counts in the shaded regions from **a** have been plotted here for all exon-intron junctions. The overall decrease in spliceosomal binding in *PRPF8* KDs evidences that the previously detected differences are not caused by a small subset of genes (p-value $< 2.2 \cdot 10^{-16}$, Wilcoxon test).

set of genes and those previously detected to display retained introns (n = 637; p-value $< 2.2 \cdot 10^{-16}$). Most importantly, the set of genes with reported alterations in splicing (either in the form of intron retention or differential exon usage events) were detected to constitute only a subset of the expressed protein coding genes (n = 3,388 out of 13,216; expression threshold = 1 FPKM; **Figure 4.4c**; see Methods), suggesting a certain degree of specificity in the mode of action of *PRPF8*.

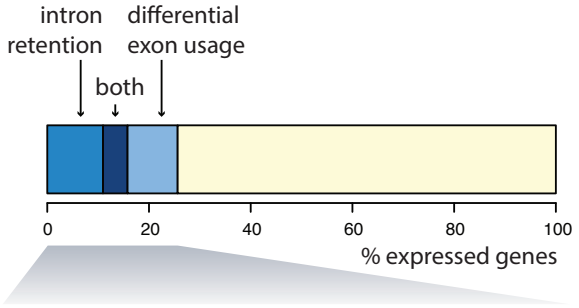
(a)



(b)

	mapped reads (M)	exons (%)	exon-intron boundaries (%)	intron bodies (%)	total introns (%)
control	17.24	86.53	4.59	4.85	9.44
	22.00	85.70	4.70	5.41	10.12
	25.76	85.79	4.72	5.36	10.07
PRPF8 KD	20.62	78.88	5.49	11.18	16.68
	12.84	79.12	5.51	10.84	16.35
	24.97	80.21	5.48	9.95	15.43
	24.63	80.12	5.48	10.12	15.60

(c)



GO term	adjusted p-value
RNA processing	1.23E-20
RNA splicing	1.14E-10
ubiquitin conjugation pathway	3.66E-13
ubiquitin-dependent protein catabolic process	3.83E-10
cell cycle	1.05E-11
mitosis	5.73E-08
cell division	4.70E-08

Figure 4.4| Characterisation of splicing alterations in *PRPF8* KD cells.
(a) Intron expression levels in PRPF8 KD vs. control cells. Down-regulation of *PRPF8* leads to an increase in intronic expression (p-value < $2.2 \cdot 10^{-16}$). Intron expression levels were normalised to take into account gene expression and library size as specified in Methods.
(b) Percentage of reads mapping to exons and introns in PRPF8 KD vs. control cells. The proportion of reads mapping to both exon-intron boundaries and intron bodies increases after *PRPF8* KD. Paired reads are counted once.
(c) Gene Ontology enrichment analysis for protein coding genes with altered splicing after PRPF8 KD. Genes that manifest significant differences in splicing patterns after *PRPF8* KD (n = 3,388) are involved in cell division and mitosis, amongst other functions.

Altogether, the above described sets of genes were determined to be involved in cell division and mitosis through Gene Ontology enrichment analysis (**Figure 4.4c**), thus linking these findings to the observed phenotype. Focusing on the common genes from these two sets, the predicted splicing changes were further validated by RT-PCR in a subset of genes with a role in cell division (**Figure 4.5a** and **Figure 4.5b**). These included *ASPM*, observed to undergo an exon skipping event; *CDC23*, for which an alternative terminal exon was detected; and *CDC20* and *NUDC*, both predicted to contain retained introns. Furthermore, for both *ASPM* and *CDC23*, the observed differences were big enough to lead to a change in the identity of major transcripts across conditions (**Figure 4.5c**). Finally, such changes could also be validated at the protein level by Western blot in the cases where antibodies were available (**Figure 4.5d**). Overall, these findings confirm the hypothesis that *PRPF8* is required for proper splicing of genes involved in mitosis.

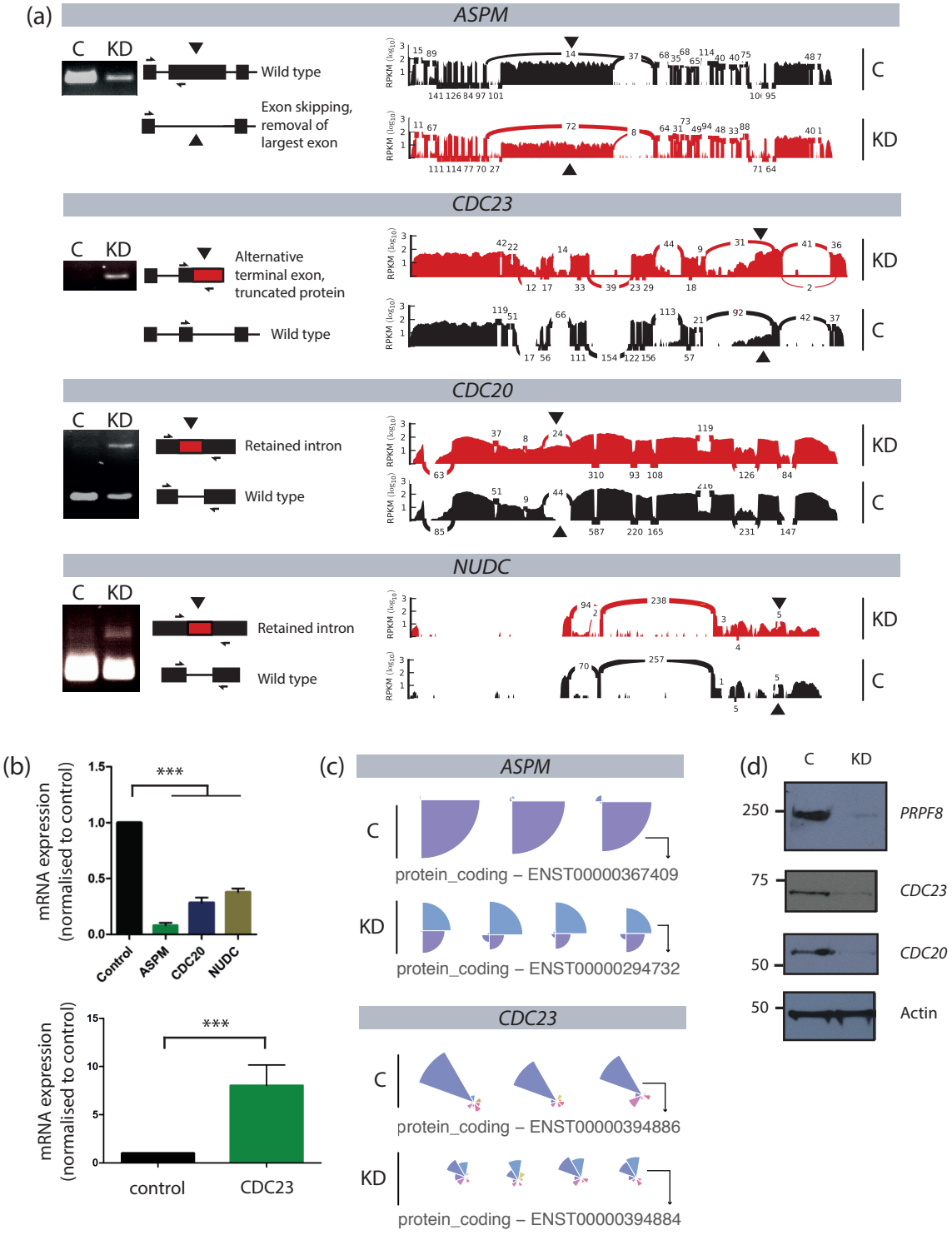

(a)
ASPM
C
KD
Wild type
Exon skipping, removal of largest exon
RPKM (log10)
CDC23
Alternative terminal exon, truncated protein
Wild type
CDC20
Retained intron
Wild type
NUDC
Retained intron
Wild type
(b)
mRNA expression (normalised to control)
***
Control
ASPM
CDC20
NUDC
control
CDC23
(c)
ASPM
protein_coding – ENST00000367409
protein_coding – ENST00000294732
CDC23
protein_coding – ENST00000394886
protein_coding – ENST00000394884
(d)
C
KD
250
PRPF8
75
CDC23
50
CDC20
50
Actin

Figure 4.5| Validation of the predicted splicing changes in a subset of genes with a role in cell division.
(a) RT-PCR validation results. Each panel includes information on the read coverage in the corresponding genomic region in a control *vs. PRPF8* KD sample (right) and the RT-PCR validation results (left). Genes have been selected to represent a variety of splicing events.
(b) qRT-PCR validation results. The relative abundance for the amplification products labelled with arrows in **a** is represented here. *ASPM* and *CDC23*: p-value < 0.001; *CDC20* and *NUDC*: p-value < 0.01.
(c) Estimated transcript expression levels. Both *ASPM* and *CDC23* are predicted to undergo a switch event in *PRPF8* KD *vs.* control samples (*i.e.* different major transcripts are detected in each condition; see Chapter 3 - Methods).
(d) Western blot results. Predicted alternative splicing events in genes for which antibodies were available could be successfully validated at the protein level.
Figures with experimental results have been provided by Dr. Vi Wickramasinghe.

4.2.3 Modulating splice site strength can compensate for the down-regulation of *PRPF8*

Further characterisation of the most extreme intron retention events suggested that those introns that are not as efficiently removed after *PRPF8* KD tend to have weaker 5′ splice sites (**Figure 4.6a**; p-value = 0.0220, Wilcoxon test). Motif enrichment analysis on the same set of genes revealed consistent results: while the top identified motifs correspond to the consensus 5′ splice site sequence both for retained (RI) and non-retained introns (NRI), the percentage of targets with such motif varies significantly between the two categories (**Figure 4.6b**; 68.50% for RI *vs.* 91.03% for NRI; p-value < $2.2 \cdot 10^{-16}$, Fisher's exact test). Such differences were not observed at the 3′ splice site (**Figure 4.6c** and **d**). Finally, mini-gene experiments further demonstrated that increasing splice site strength can revert the splicing defects observed in *PRPF8* KD cells (**Figure 4.7**).

On a separate note, analysis of GC composition for the most extreme RI revealed that, contrary to NRI, the former display GC content similar to that of adjacent exons (p-value < $2.2 \cdot 10^{-16}$, Wilcoxon test; **Figure 4.8a**), as well as higher GC content overall (p-value = 0.0055, Wilcoxon test; **Figure 4.8b**).

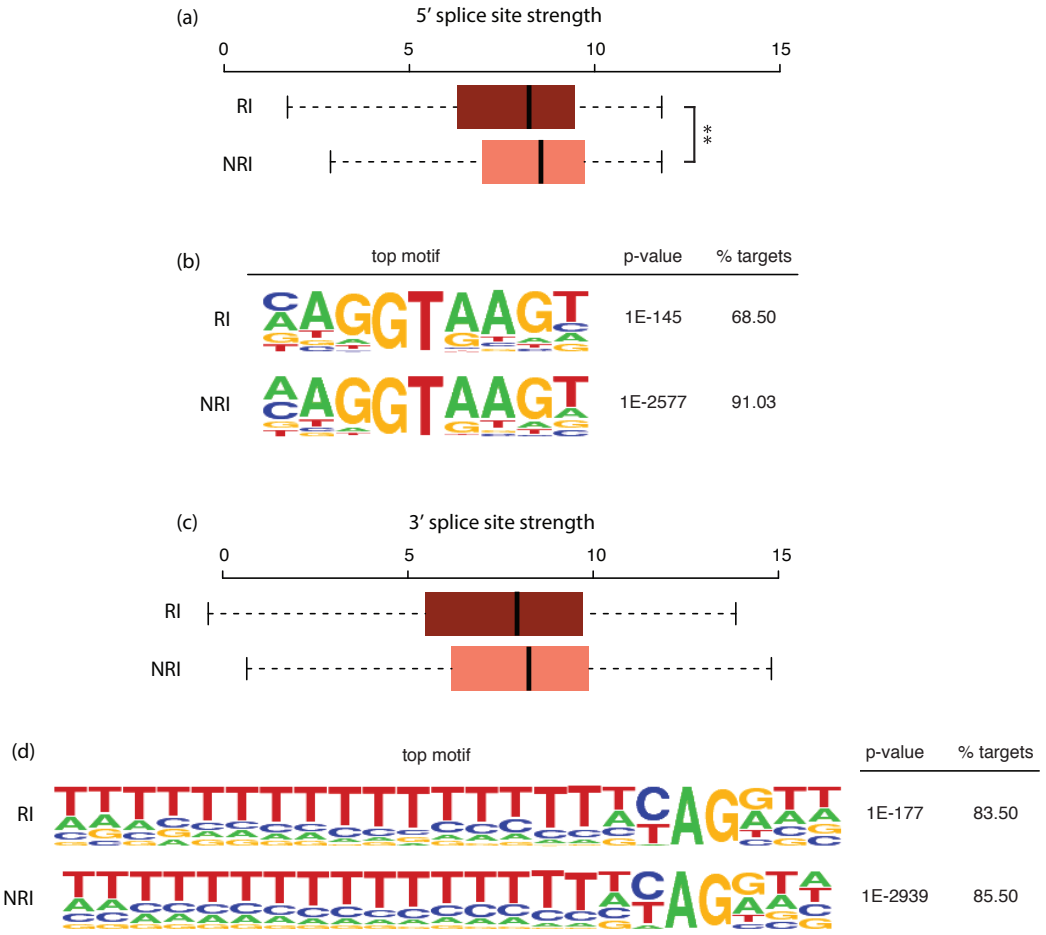


Figure 4.6 | Splice site strength for the top retained introns following *PRPF8* KD. A set of 200 retained introns (RI) were selected based on fold-change differences (see Methods), and non-retained introns (NRI) within the same set of genes were used as a contrast.
(a) Differences in 5′ splice site strength. Top retained introns have weaker 5′ splice sites compared to non-retained introns from the same set of genes (p-value = 0.0220, Wilcoxon test).
(b) Results from the motif enrichment analysis on the 5′ splice site. The top motif identified corresponds to the consensus 5′ splice site sequence both for RI and NRI. However, the percentage of targets displaying such sequence differs significantly between the two sets (p-value < $2.2{\cdot}10^{-16}$, Fisher's exact test).
(c) Differences in 3′ splice site strength. No significant differences were detected between RI *vs.* NRI (p-value = 0.2068, Wilcoxon test).
(d) Results from the motif enrichment analysis on the 3′ splice site. The findings in **c** are supported by these results.

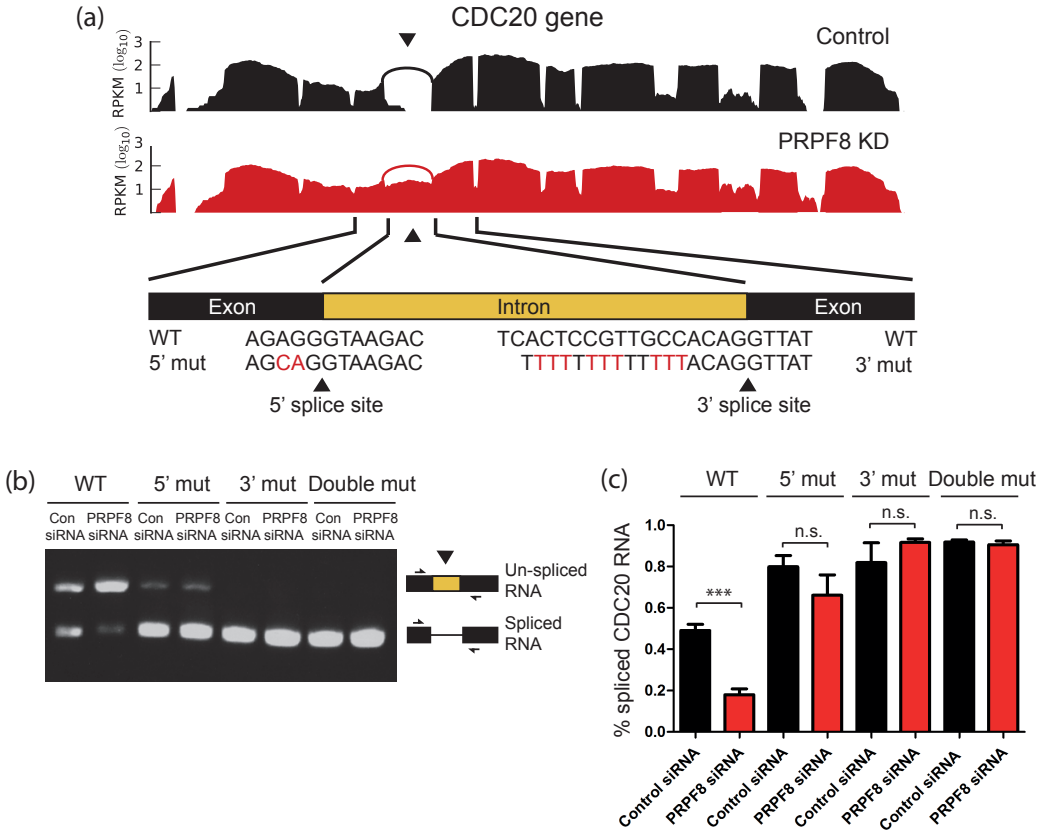


Figure 4.7| Mini-gene experiments for the *CDC20* gene.
(a) Illustration of the experimental design. The *CDC20* gene had been previously identified to contain a retained intron following *PRPF8* KD (see **Figure 4.5**). In this experiment, the splicing efficiency of such retained intron was evaluated with several mini-gene constructs, which contained either the wild-type splice sites, stronger 5′ and 3′ splice sites (5′ mutants and 3′ mutants, respectively), and the combination of both (double mutants).
(b) RT-PCR results. Splicing of the intron of interest can be rescued by the introduction of stronger 5′ and 3′ splice sites, as well as by the combination of both. Most importantly, splicing recovery is obtained independently of the condition studied (*i.e.* control and *PRPF8* KD). The same results can be observed in **c**.
(c) qRT-PCR results. Mean + standard error are shown for triplicate reactions of each experiment.
Figure provided by Dr. Vi Wickramasinghe.

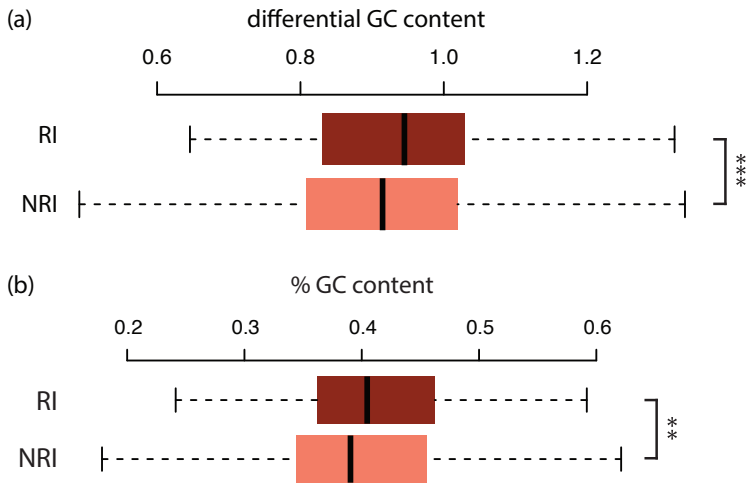


Figure 4.8 | GC content of top retained *vs.* non-retained introns.
(a) Differential GC content. Differential GC content was calculated by dividing the GC content of each intron to the average of its adjacent exons (see Methods). Compared to NRI, the GC content in retained ones is closer to that of neighbouring exons (p-value $< 2.2 \cdot 10^{-16}$, Wilcoxon test).
(b) Overall GC content. Top retained introns have slightly higher GC content than non-retained ones (p-value = 0.0055, Wilcoxon test).

4.2.4 *PRPF8* knock-down has an impact on the rate of co-transcriptional splicing

The above results suggest that the down-regulation of *PRPF8* leads to increased kinetic competition across splice sites; however, they do not provide an explanation for what could be mediating such increase. In this context, an alteration in the rate of splicing following *PRPF8* KD emerges as a plausible cause for such differences: splicing is known to occur co-transcriptionally for most human genes [Tilgner et al., 2012], and thus any changes in the co-existence of this reaction with transcriptional elongation could result in an increased competition across splice sites.

Given that the analysed RNA-seq libraries are polyA-selected, it is theoretically possible to obtain information on the kinetics of the splicing reaction from intronic reads. More specifically, under a scenario of co-transcriptional splicing, introns located towards the 5′ end of transcripts are more prone to be spliced out, since they are transcribed first. Conversely, introns located towards the 3′ end are more likely to be detected in the RNA-seq experiment, given that splicing has

not yet finalised after the addition of the polyA tail. Altogether, this suggests that a difference in the number of reads that map to the first *vs.* last introns of each given transcript could be used as an indicator of co-transcriptional splicing (**Figure 4.9a**). Similarly, differences in such proportion across the two studied conditions (*i.e.* control *vs. PRPF8* KD) could serve as an indicator of changes in the dynamics of splicing.

The results obtained following the above analysis strategy revealed that introns located towards the 3′ end of transcripts are generally expressed at higher levels compared to those situated towards the 5′ end (**Figure 4.9b** and **Figure D.4**), thus suggesting that splicing is co-transcriptional for most genes. Interestingly, there is a negative correlation between the differences in expression of first *vs.* last introns and transcript length, which indicates that co-transcriptional splicing is more prevalent in longer transcripts (average r_S = -0.24, p-value < $5 \cdot 10^{-4}$ in all samples). On the other hand, such differences are less accentuated after down-regulation of *PRPF8*, hence pointing at changes in splicing efficiency (**Figure 4.9b**; p-value = $5.34 \cdot 10^{-9}$). Further experimental validation on a subset of the previously highlighted genes with a prominent role in cell cycle progression showed that, indeed, splicing is co-transcriptional overall and less efficient after *PRPF8* KD (**Figure 4.10**). Specifically, splicing recovery was considerably delayed in *PRPF8* KDs compared to normal cells for *CDC20* and *CDC23*, and even more extremely, it could not be detected during the time frame of the experiment for *ASPM*.

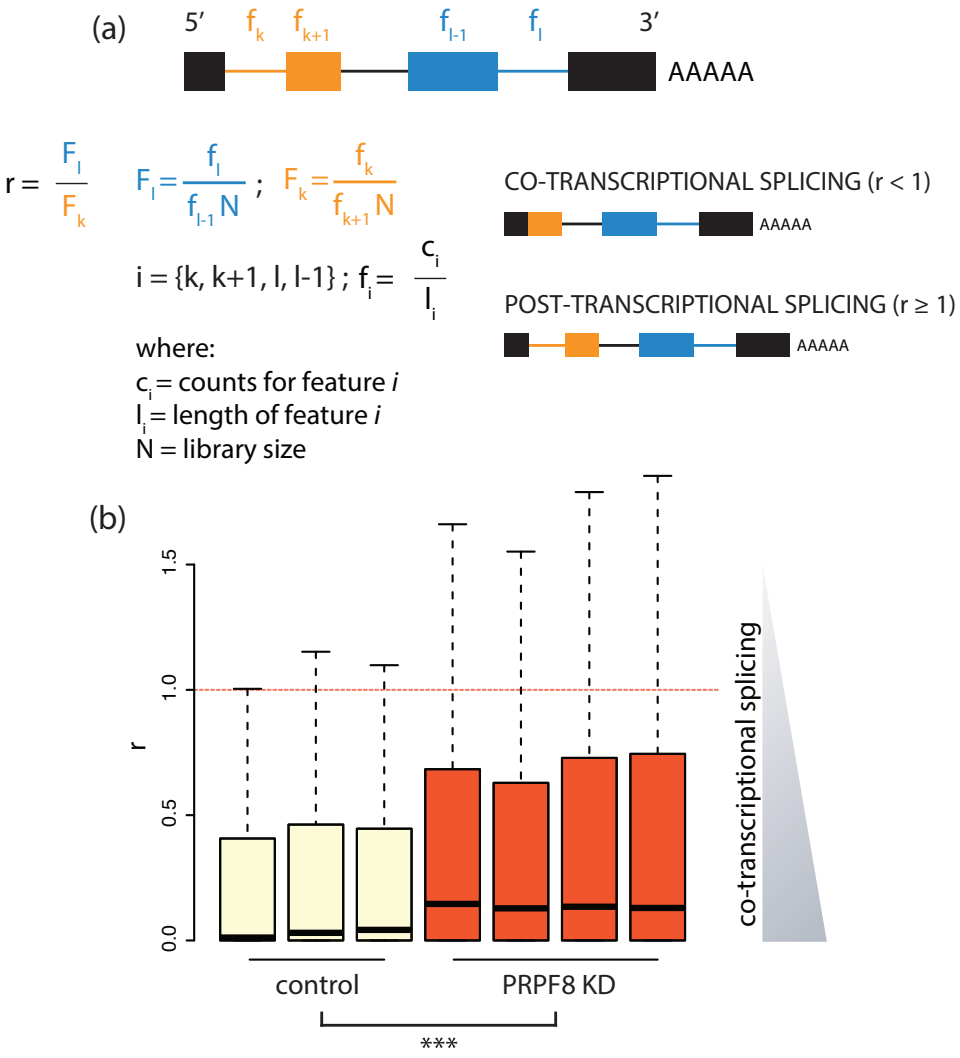


Figure 4.9 | Co-transcriptional splicing evidence from intronic reads.
(a) Intronic reads contain potential information on the dynamics of splicing. Introns located towards the 5′ end of transcripts have a higher probability of being spliced out under a scenario of co-transcriptional splicing, but this is not the case if splicing occurs post-transcriptionally. A co-transcriptional splicing ratio was calculated by comparing the intronic coverage of the first *vs.* last introns in each transcript. Such coverage was normalised to take into account intron length and the expression levels of adjacent exons.
(b) Distribution of the co-transcriptional splicing ratio in control vs. PRPF8 *KD samples.* Co-transcriptional splicing dominates in control samples as indicated by the low ratio values. *PRPF8* KDs present higher ratios, suggesting disruptions in the rate of co-transcriptional splicing (p-value = $5.34 \cdot 10^{-9}$). Only genes with one transcript annotated in Ensembl 66 were considered for this analysis (n = 2,366; see Methods).

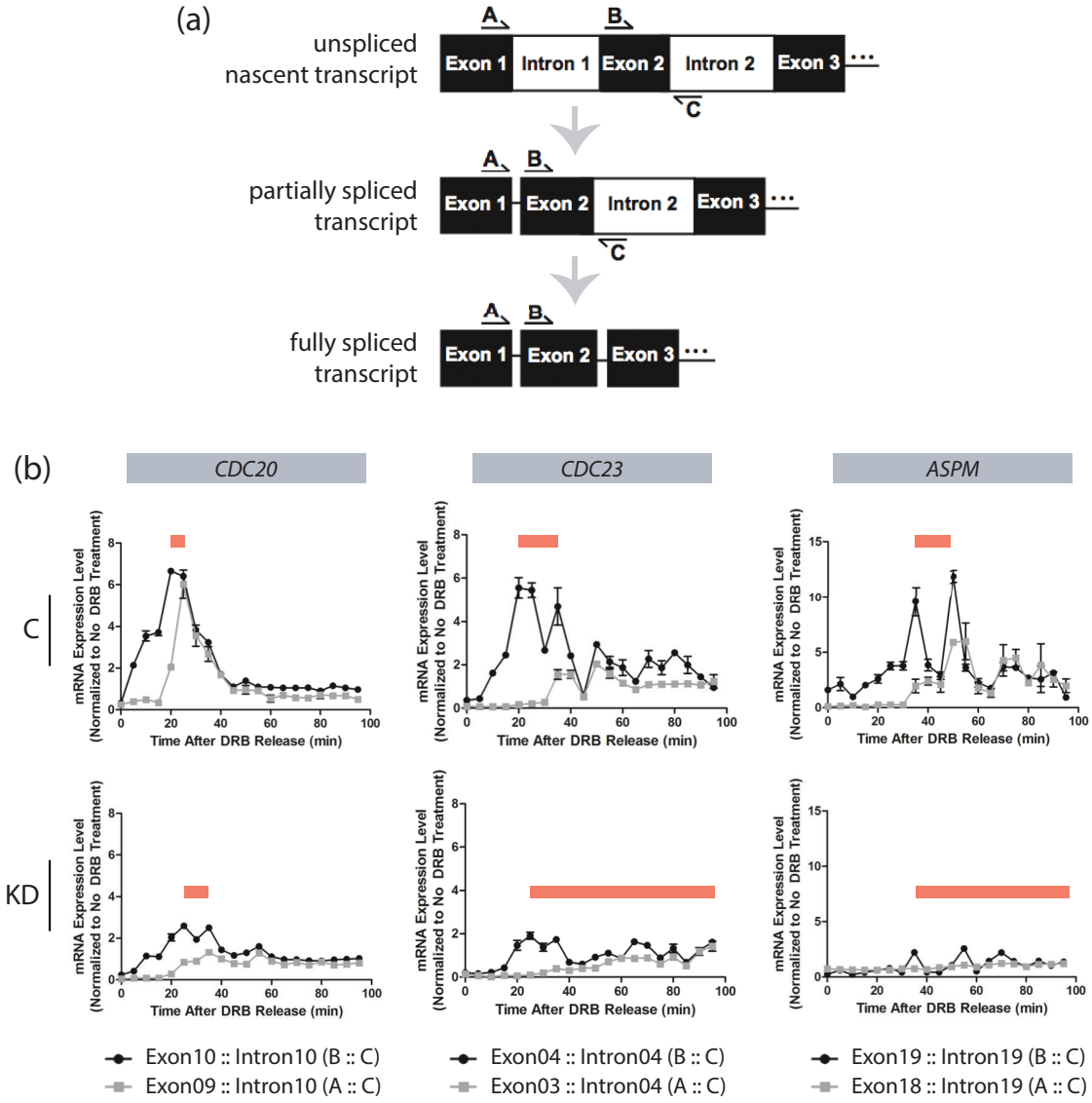

(a)
unspliced nascent transcript
partially spliced transcript
fully spliced transcript
A
B
C
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
(b)
CDC20
CDC23
ASPM
C
KD
mRNA Expression Level (Normalized to No DRB Treatment)
Time After DRB Release (min)
Exon10 :: Intron10 (B :: C)
Exon09 :: Intron10 (A :: C)
Exon04 :: Intron04 (B :: C)
Exon03 :: Intron04 (A :: C)
Exon19 :: Intron19 (B :: C)
Exon18 :: Intron19 (A :: C)

Figure 4.10 | Experimental validation of the predicted differences in the rate of co-transcriptional splicing after *PRPF8* KD.
(a) Illustration of the experimental design. Rates of transcription and splicing can be measured *in vivo* by treating cells with DRB, a reversible kinase inhibitor [Ardehali and Lis, 2009; Singh and Padgett, 2009]. This drug targets specifically a kinase that is involved in the phosphorilation of the carboxy-terminal domain (CTD) of RNA pol II, thus blocking its entry into productive elongation. Following DRB release, newly synthesised transcripts can be detected with the indicated primers. Specifically, the primer pair BC provides information on the rate of transcription, as it measures the production of pre-mRNAs. On the other hand, the primer pair AC will anneal to both unspliced and partially spliced transcripts, but it will fail to generate a PCR product in the former due to the large size of the enclosed intron. Thus, it can be used as an indicator of the rate of splicing. Eventually, once the intron that is targeted by primer C is removed, no amplification product will be detected.
(b) Results of the DRB experiments for a set of genes with a role in cell division. Following treatment of cells with DRB, the recovery of transcription (black) and splicing (gray) were measured as described in **a**. In general, recovery of splicing can be detected previously to the completion of transcriptional recovery, hence suggesting that intron removal occurs in a co-transcriptional fashion. *PRPF8* KD cells manifest less accentuated recovery rates overall, as well as a slower recovery of splicing. More extremely, for the *CDC23* and *ASPM* genes, such recovery was largely impaired. For all panels, mean + standard error are shown for triplicate reactions of each experiment.
Figure provided by Dr. Vi Wickramasinghe.

4.3 Discussion

The present chapter describes a collaborative effort to further characterise the function of the splicing factor *PRPF8* and, more generally, to dissect the role of core spliceosomal components in regulating the splicing reaction. Similarly to previous studies that combined RNA interference and imaging techniques for the large-scale identification of genes required for cell division, the findings presented here indicate that the expression of this SF is necessary to exit mitosis [Neumann et al., 2010]. Moreover, they reveal that the down-regulation of other B-complex components can also lead to a mitotic arrest phenotype, whilst this is not the case for components of the A and C complexes, which emphasises the role of the B-complex in initiating the catalysis of intron removal. In spite of this B-complex specificity, the reported observations show that knocking-down *PRPF8* results in the strongest effects, consistent with its location at the centre of the spliceosome and its proposed role as a master regulator of splicing [Galej et al., 2013; Grainger and Beggs, 2005]. Moreover, the treatment of *PRPF8* KD cells with

cytosolic extracts from control samples leads to a decrease in the proportion of cells arrested in mitosis, thus indicating the the observed phenotype is driven by defects in splicing. These findings are further corroborated by the results from Sm iCLIP experiments, which point to an overall decrease in spliceosomal RNA-protein interactions after *PRPF8* down-regulation.

Consistent with the above observations, RNA-seq experiments show an overall increase in intronic expression levels. In addition, analysis of exon usage patterns reveals further differences in alternative splicing, adding up to existing evidence in indicating that changes in the concentration of core components of the spliceosome can regulate this process [Saltzman et al., 2011]. Overall, analysis of the RNA-seq data allowed for the detection of alterations in splicing for a subset of the expressed protein coding genes (n = 3,388 out of 13,216). Notably, these genes are enriched for cell division regulators and mitotic factors, hence linking the observed splicing defects to the mitotic arrest phenotype. Further exploration of a set of introns especially sensitive to *PRPF8* KD revealed that they present typical characteristics of retained introns, including weaker splice sites [Sakabe and de Souza, 2007] and lower differential GC content [Amit et al., 2012]. Particularly, the results from the computational analyses indicate that these differences only exist at the 5′ splice site, in line with existing evidence that points to the preferential interaction of *PRPF8* with this splicing signal [Li et al., 2013]. However, experimental validation with mini-gene experiments showed that, in *PRPF8* KD cells, intron removal can be promoted by increasing both 5′ and 3′ splice site strength. Altogether, these findings suggest that the splicing alterations detected following a decrease in the concentration of this SF are likely to arise from the kinetic competition between splice sites, a well known mechanism of splicing regulation [House and Lynch, 2008; Wang and Burge, 2008].

Finally, analysis of intronic reads in control samples evidenced that slicing occurs mostly in a co-transcriptional fashion in the studied set of genes. This is consistent with existing knowledge suggesting that splicing occurs co-transcriptionally for most human genes [Bentley, 2014], and in particular, it recapitulates previous observations from Tilgner et al. [2012] that were based on the analysis of junction reads. Moreover, such finding shows that it is possible to gain biological knowledge from the intronic data generated in an RNA-seq experiment, often considered to be noise. Interestingly, co-transcriptional splicing was detected

to be more prevalent in longer transcripts, where the increased timespan of the transcription reaction could provide an opportunity for splicing to start prior to the addition of the polyA tail. Indeed, in yeast, in which the average gene length is shorter than in humans, the RNA pol II has been detected to pause at the 3′ terminal end of transcripts and downstream of introns, hence providing extra time for co-transcriptional splicing to occur [Alexander et al., 2010; Carrillo Oesterreich et al., 2010]. Most importantly, co-transcriptional splicing was detected to be less prevalent after *PRPF8* KD, thus providing an explanation for the detected kinetic competition amongst splice sites.

Altogether, the findings described in this chapter show that fluctuations in the concentration of core spliceosomal factors can result in changes in the dynamics of splicing, which can in turn manifest as changes in splice site choice. Even though the kinetics of spliceosomal rearrangements had been already identified to play a role in shaping splicing decisions [Query and Konarska, 2004; Saltzman et al., 2011; Yu et al., 2008], the interplay between this process and the dynamics of splicing had not been characterised to date, hence evidencing the novelty of the described findings.

4.4 Computational methods

All computational analyses have been performed by myself unless otherwise stated. Experimental analyses have been performed by Dr. Vi Wickramasinghe, Dr. David Perera and Arthur Bartolozzi. Dr. Christopher W. Sibley carried out the Sm iCLIP experiments and part of the derived data analysis. Given the computational focus of this thesis, details on the experimental methods followed have been covered elsewhere [Wickramasinghe et al.].

Analysis of the Sm iCLIP data

Analyses performed by Dr. Christopher W. Sibley:
Sm iCLIP datasets for 3 control and 3 *PRPF8* KD samples were generated following the protocol described in Briese et al. Each biological replicate was sequenced twice, hence resulting in a total of 12 samples. The obtained reads were mapped to the human genome and analysed with the iCount software[1], following the steps detailed in König et al. [2010]. bed files that contain information on

[1] http://icount.biolab.si/

spliceosome occupancy along the genome were generated as a result, and further used for downstream analysis.

Analyses performed by myself:
iCLIP counts were normalised by library size to account for variations in sequencing depth across samples. Similarly, for each junction, counts were divided by the maximum value to allow for the comparison across different features. Normalised counts were eventually used to produce **Figure 4.3**.

Datasets and mapping

RNA from control and *PRPF8* KD Cal51 cell lines was extracted, polyA-selected and prepared for sequencing following standard Illumina protocols. Sequencing was performed on a HiSeq 2000 machine (100 bp paired-end reads), including 3 and 4 biological replicates for control and *PRPF8* KD samples, respectively (**Table A.3**). On average, 25M reads per sample were obtained.

Due to the high quality of the reads, raw data were directly mapped to the human genome (Ensembl 66 [Flicek et al., 2012]) using TopHat v1.3.3 [Trapnell et al., 2009] with the following options: `--max-multihits 1 --no-novel-juncs --min-isoform-fraction 0.0 --GTF $gtf_file`.

Intron counts and normalised intron expression

Intron coordinates were obtained with custom scripts based on bedtools v2.17.0 [Quinlan and Hall, 2010], by considering exons from transcripts annotated as protein coding in Ensembl 66. Overlapping exons were merged and the longest exon or combination of exons was kept. Intronic expression levels were then obtained with dexseq-count (DEXSeq v1.7.0 [Anders et al., 2012]) and normalised as follows:

$$\hat{\mu}_i = \frac{c_i / l_i}{c_g / l_g} \cdot \frac{10^9}{N}$$

where:

$\hat{\mu}_i$ = normalised expression for intron i
c_i = counts for intron i
l_i = length for intron i
c_g = counts for gene g
l_g = length for gene g
N = library size

Identification of differentially used exons

DEXSeq v1.7.0 was used to identify genes with differential exon usage across the two studied conditions. Specific options include `--aggregate=no` for the preparation of the annotation and `--paired=yes --stranded=no` to count reads that overlap exons. An FDR threshold of 0.01 was used to assess the significance of the detected fold-changes (Benjamini & Hochberg p-value correction [Benjamini and Hochberg, 1995]).

Identification of retained introns

Differential intron usage was assessed with DEXSeq v1.7.0 with the same options as indicated above. Due to the overall low number of intronic reads compared to exonic ones, library size factors were not inferred from intron counts, but instead the previously calculated ones were re-used. Differentially used introns with a positive log_2 fold-change (FDR < 0.01, Benjamini & Hochberg p-value correction) were defined as retained introns (RI), whilst those that did not fulfil this criteria were classified as non-retained (NRI). The most extreme intron retention events (*i.e.* top 200) were selected for subsequent analyses based on fold-change information.

Splice site strength was calculated with the MaxEntScan algorithm [Yeo and Burge, 2004], and motif enrichment analysis was performed with Homer v2 [Heinz et al., 2010] using a common set of background sequences. The calculation of differential GC content in introns *vs.* adjacent exons was performed similarly to Amit et al. [2012], by discarding 20 nucleotides (nt) from each end of the introns and 3 nt from the exons in order to account for splicing signals.

Gene and transcript-level analyses

A gene was defined as expressed if it is detected above 1 FPKM in any given sample. Gene Ontology enrichment analyses were performed with DAVID [Huang et al., 2009a,b] and WebGestalt [Wang et al., 2013], using an adjusted p-value threshold of 0.01 (Benjamini & Yekutieli correction [Benjamini and Yekutieli, 2001]) and the set of expressed genes as a background (n = 13,216). Differential gene expression was assessed with DESeq v1.10.1 [Anders and Huber, 2010], and significant changes were evaluated following the criteria previously used for the analysis of differential splicing (FDR < 0.01, Benjamini & Hochberg correction).

Transcript expression estimates were calculated with MISO v0.4.1 [Katz et al., 2010], using the annotation from Ensembl 66.

Co-transcriptional splicing ratio

Co-transcriptional splicing was evaluated by comparing intronic expression levels for the first *vs.* last introns of each transcript. Given the potential complexity introduced by alternative splicing events, this analysis was limited to genes which had only one transcript annotated in Ensembl 66. Genes with only one intron, those shorter than 300 bp and those that overlap with other genes in either strand were also discarded. Altogether, such filtering led to a final set of 2,366 genes (*i.e.* transcripts) that could be considered in the analysis. Following the identification of the first and last introns of each transcript in the study set, a co-transcriptional splicing ratio was calculated using the formula detailed in **Figure 4.9a**.

Chapter 5

The impact of splicing at the protein level

While mRNAs contain the necessary information for the synthesis of proteins, their mere existence does not ensure the production of functional protein products. Hence, many studies have focused on establishing links between the transcriptome and the proteome, with most of the emphasis placed at the gene level (*e.g.* Fu et al. [2009]). In the present chapter, I investigate the possibility of recapitulating differential splicing events at the protein level, by referring to proteomics data for the same control and *PRPF8* knock-down samples as studied in Chapter 4. As a first step, I assess the feasibility of such integration by focusing on the most extreme changes in splicing. I then proceed to evaluate the correlation between the fold-change estimates obtained from the RNA-seq and proteomics experiments, whilst expanding the analysis to a larger set of genes. Finally, I investigate whether such correlation can be improved by incorporating extra information on transcript relative abundances from the RNA-seq experiment, and I explore the extent to which differences in gene expression levels could be influencing the obtained results.

The findings described here constitute the first results from an ongoing collaborative project. All the computational analyses have been performed by myself, under the supervision of Dr. John Marioni from EMBL-EBI. All the experimental work has been carried out by Dr. Vi Wickramasinghe from the Hutchison Research Institute, and the proteomics data has been generated and analysed by Dr. Yansheng Liu from ETH Zürich.

Publications derived from this chapter

- This is an ongoing project.

5.1 Introduction

Establishing a correlation between the transcriptome and the proteome is not straightforward, with initial genome-wide estimates of ~40% correlation obtained from the quantification of gene expression levels with microarrays [Maier et al., 2009; de Sousa Abreu et al., 2009]. Following the introduction of RNA-seq, this technology was proven to correlate better with protein levels [Fu et al., 2009], leading to higher estimates of 56-64% correlation (Li et al. [2014]; Lundberg et al. [2010]; Nagaraj et al. [2011]). Nonetheless, these findings still imply that 40% of the variation in protein abundance cannot be attributed to mRNA expression, and both biological processes (*i.e.* post-transcriptional regulation, translational efficiency, mRNA and protein turnover...) and limitations inherent to the technologies used have been proposed as an explanation for such observation [Haider and Pal, 2013; Lundberg et al., 2010].

On the other hand, when taking into account the diversity of the transcriptome in terms of alternatively spliced transcripts, the reconciliation of mRNA and protein levels becomes even more difficult. There have been several efforts to detect proteins derived from alternatively spliced transcripts, which rely primarily on the detection of peptides that uniquely support annotated isoforms (*e.g.* Blakeley et al. [2010]; Ezkurdia et al. [2012]; Leoni et al. [2011]) or novel exon-exon junctions (*e.g.* Xing et al. [2011]; Zhou et al. [2010]). Further approaches are based on the incorporation of expression data, including evidence from RNA-seq experiments, to prioritise the set of junctions to consider when interpreting the proteomics datasets and hence reduce mapping noise [Gloria et al., 2012; Ning and Nesvizhskii, 2010; Sheynkman et al., 2013; Tanner et al., 2007]. However, in spite of these attempts, understanding the prevalence of alternative splicing at the protein level in a genome-wide scale still remains a challenge. Similarly, it is not easy to predict the extent to which the detected differences in alternative splicing across two given conditions might ultimately lead to differences in protein expression levels.

In the present chapter, I describe a collaborative effort aimed at assessing the impact of differential splicing events at the protein level. This is achieved by integrating fold-changes obtained by RNA-seq and SWATH-MS, a new proteomics method that enables the quantification of a large fraction of the proteome with

high accuracy [Liu et al., 2013]. I introduce here a novel pipeline for the integration of these two types of data, which relies not only on the information from peptides that map uniquely to one transcript isoform, but also on that from the ones that map ubiquitously to several transcripts of the same gene. This is achieved by using knowledge from the RNA-seq experiment in order to guide the peptide assignments, and enables to expand the set of usable data. By executing such pipeline on the *PRPF8* knock-down *vs.* control samples studied in Chapter 4, I am able to show that a large fraction of the most extreme changes in splicing can be recapitulated at the protein level (*i.e.* 9 out of 11 switch events that overlap with the proteomics dataset are supported by peptide evidence). Moreover, when expanding such analysis to the larger set of differentially used transcripts, I predict a 43% correlation between mRNA and protein fold-changes, and I observe that such correlation increases up to 78% when focusing on differential splicing events that affect major transcripts. Finally, I show that the estimated correlation remains stable after discarding differentially expressed genes, which strongly supports the idea that protein expression levels can be regulated by changes in transcript usage.

5.2 Results

In order to attempt proteomics validation of the splicing changes predicted from RNA-seq data, our collaborators generated a SWATH-MS dataset that comprises both *PRPF8* knock-down (KD) and control samples (3 biological replicates in each condition; CAL51 cell lines), thus matching the conditions studied in Chapter 4. Specifically, in the previous chapter, mapping to the genome enabled the study of differential splicing events from an exon-centric perspective, as well as the identification of cases of intron retention. Here such data are re-analysed in a transcript-centric fashion to facilitate the integration with the proteomics data (see Methods). The results from such analyses, together with those obtained from the SWATH-MS dataset, are summarised in **Table 5.1**.

RNA-seq	
differential transcript usage	668 transcripts (540 genes)
switch events	142 genes
differential gene expression	1,607 genes

SWATH-MS	
initial number	16,779 peptides
filter 1: remove peptides that map to >1 genes	14,648 peptides (2,800 genes)
peptides that map to only one transcript	2,964 peptides (856 genes)
filter 2: subset based on fold-change significance	2,901 peptides (1,275 genes)
peptides that map to only one transcript	537 peptides (297 genes)

Table 5.1 | Summary of the results obtained from the RNA-seq and SWATH-MS datasets. Details on the analysis workflow used have been described in the Methods section.

5.2.1 Extreme changes in splicing can be detected at the protein level

Amongst the genes detected to undergo differential splicing (n = 540 genes with differential transcript usage, DTU; see Methods), an interesting set are those cases in which the identity of major transcripts changes across conditions, previously introduced in this thesis as switch events (SE). Such events constitute the most extreme and prevalent changes in splicing, and thus constitute a good set to test whether changes in transcript isoform levels can be recapitulated at the protein scale. In this context, peptides that map simultaneously to both major transcripts of a given switch contain ambiguous information, but those that map to either of the major transcripts provide valuable evidence for the validation of such event (**Figure 5.1**).

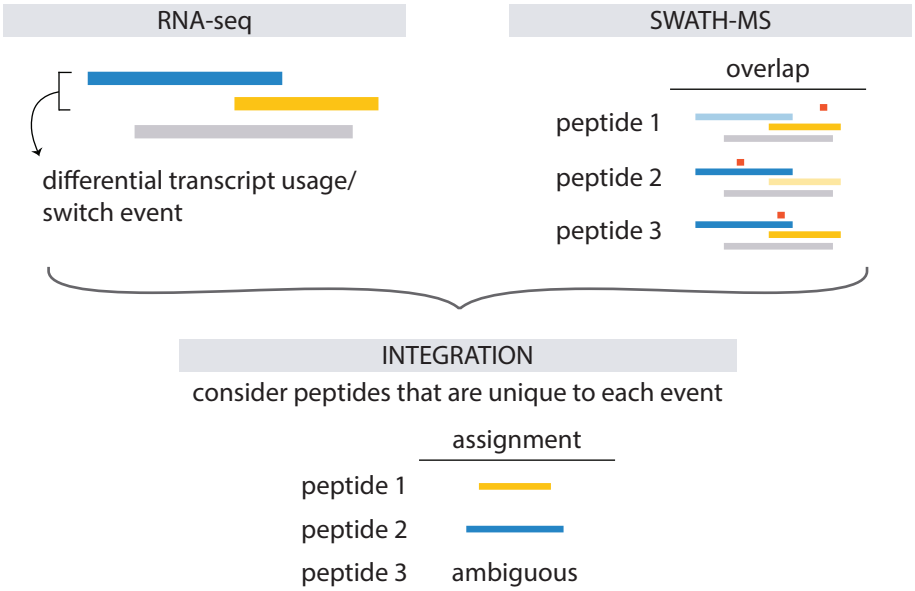


Figure 5.1 | Integration of the RNA-seq and SWATH-MS data. The RNA-seq and SWATH-MS data were first analysed separately as detailed in the Methods section. Integration of the results of such analyses was then performed by focusing on the set of peptides that map uniquely to the differential splicing events under consideration (*i.e.* differential transcript usage/switch events; all events/events that affect major transcripts). Finally, the correlation between transcript and peptide fold-changes was evaluated. In this plot, protein coding transcripts are represented by coloured lines, and red boxes correspond to the aligned peptides.

Following the logic described above, the integration of the results from the RNA-seq and SWATH-MS datasets revealed a set of 27 switched genes with enough peptide evidence available to attempt validation (see Methods). In this set, 57% of the detected peptides were consistent with the predicted SE, and a total of 14 events could be classified as supported by the proteomics data (**Table 5.2**). Notably, the validation rate increased drastically when further filtering the peptide data based on the significance of the calculated fold-changes: in this scenario, ~90% of the peptides with a significant fold-change supported the predicted SE, and 9 out of 11 events could be successfully recapitulated (**Table 5.3** and **Figure 5.2**). For example, for *TMPO*, a protein involved in the structural organisation of the nuclear envelope, a SE between two protein coding transcripts was strongly supported by several peptides (**Figure 5.2**). Similarly, two different peptides that support each of the transcripts from the switch were identified for *HNRNPK*, an hnRNP believed

to play a role in cell cycle progression (**Figure 5.2**). In both cases, evidence for alternative protein isoforms had already been reported [Ezkurdia et al., 2012].

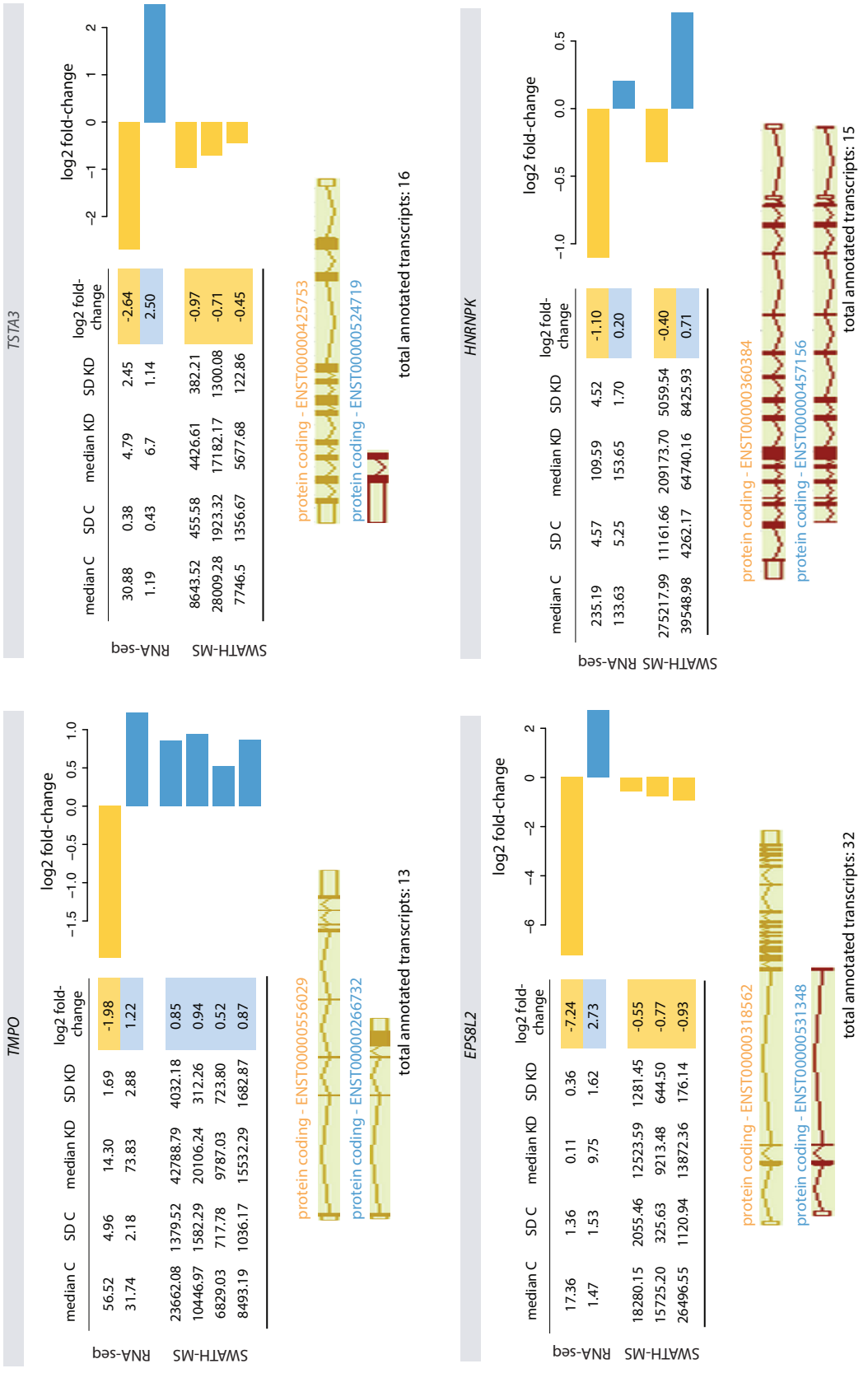
gene name	number of peptides that support the switch	number of peptides that do not support the switch	is the switch event validated?
TMPO	8	0	Y
EPS8L2	5	0	Y
TSTA3	3	0	Y
NUP98	3	1	Y
FXR1	3	2	Y
ACOT9	3	0	Y
HNRNPK	3	1	Y
PFN2	2	1	Y
C9orf5	1	0	Y
XPNPEP1	1	0	Y
HELLS	1	0	Y
HP1BP3	1	0	Y
MOV10	1	0	Y
CPNE1	1	0	Y
NOP56	2	3	N
UQCRC1	0	6	N
AP1B1	0	1	N
CSNK2A1	0	2	N
GOSR2	0	1	N
ABCB6	0	1	N
COMMD4	0	2	N
BRD4	0	1	N
FAM98B	0	3	N
GMPPB	0	2	N
SSNA1	0	1	N
ASRGL1	2	2	A
PDLIM7	1	1	A
total number of peptides	**41**	**31**	

Table 5.2 | Validation of switch events. For each of the 27 genes with usable peptide evidence, the available peptides are classified as either supportive or contrary to the observed switch event (SE). Specifically, a given peptide is considered to support the SE if it suggests an increase in the abundance of the protein isoform expected in *PRPF8* KDs, or a decrease in the one expected in controls. Y: yes; N: no; A: ambiguous.

gene name	number of peptides that support the switch	number of peptides that do not support the switch	is the switch event validated?
TMPO	4	0	Y
TSTA3	3	0	Y
EPS8L2	3	0	Y
HNRNPK	2	0	Y
PFN2	1	0	Y
C9orf5	1	0	Y
HELLS	1	0	Y
ACOT9	1	0	Y
MOV10	1	0	Y
AP1B1	0	1	N
CSNK2A1	0	1	N
total number of peptides	**17**	**2**	

Table 5.3| Validation of switch events when filtering by peptide fold-change significance. When subsetting peptides based on their fold-change significance, the vast majority support the trend marked by the RNA-seq data. None of the genes listed here are detected as differentially expressed, with the exception of *HELLS*. Y: yes; N: no; A: ambiguous.

Figure 5.2| Examples of switch events with supporting peptide evidence. Fold-changes represent changes in expression levels for *PRPF8* KD *vs.* control samples (yellow: down-regulated in *PRPF8* KD; blue: up-regulated in *PRPF8* KD). Barplots feature information from the two transcripts involved in the switch (RNA-seq), as well as from the overlapping peptides (SWATH-MS; **Table 5.3**). Tables include further details on the median and the standard deviation of the estimates obtained with each technology (FPKMs for RNA-seq, peptide intensities for SWATH-MS). For each gene, genomic maps for the highlighted transcripts are depicted in the bottom panels. C: control; KD: *PRPF8* KD.


TSTA3
median C SD C median KD SD KD log2 fold-change
30.88 0.38 4.79 2.45 -2.64
1.19 0.43 6.7 1.14 2.50
8643.52 455.58 4426.61 382.21 -0.97
28009.28 1923.32 17182.17 1300.08 -0.71
7746.5 1356.67 5677.68 122.86 -0.45
RNA-seq
SWATH-MS
log2 fold-change
protein coding - ENST00000425753
protein coding - ENST00000524719
total annotated transcripts: 16
HNRNPK
median C SD C median KD SD KD log2 fold-change
235.19 4.57 109.59 4.52 -1.10
133.63 5.25 153.65 1.70 0.20
27521.99 11161.66 20917.70 5059.54 -0.40
39548.98 4262.17 64740.16 8425.93 0.71
protein coding - ENST00000360384
protein coding - ENST00000457156
total annotated transcripts: 15
TMPO
median C SD C median KD SD KD log2 fold-change
56.52 4.96 14.30 1.69 -1.98
31.74 2.18 73.83 2.88 1.22
23662.08 1379.52 42788.79 4032.18 0.85
10446.97 1582.29 20106.24 312.26 0.94
6829.03 717.78 9787.03 723.80 0.52
8493.19 1036.17 15532.29 1682.87 0.87
protein coding - ENST00000556029
protein coding - ENST00000266732
total annotated transcripts: 13
EPS8L2
median C SD C median KD SD KD log2 fold-change
17.36 1.36 0.11 0.36 -7.24
1.47 1.53 9.75 1.62 2.73
18280.15 2055.46 12523.59 1281.45 -0.55
15725.20 325.63 9213.48 644.50 -0.77
26496.55 1120.94 13872.36 176.14 -0.93
protein coding - ENST00000318562
protein coding - ENST00000531348
total annotated transcripts: 32

5.2.2 RNA-seq fold-change estimates correlate with those obtained from SWATH-MS

Next, I investigated whether the integration strategy introduced in the previous section can be extended to the general validation of DTU events. More specifically, I evaluated the impact of the predicted changes in splicing at the protein level on the basis of the correlation between transcript and peptide fold-changes. In this context, a positive correlation would indicate that the predicted changes in splicing contribute to determining the detected fluctuations in protein levels, whilst a lack of correlation would support the contrary.

Quantitative integration of the fold-changes from the two technologies used in this chapter can be easily achieved by focusing on the peptides that map to only one transcript. Given the size of the proteomics dataset analysed here, it was possible to identify a small number of peptides that map uniquely to transcripts involved in DTU events (n = 48 peptides from 21 genes), from which a Spearman correlation coefficient of 0.41 was obtained when comparing fold-changes (p-value = $3.67 \cdot 10^{-3}$; **Table 5.4**). Further focusing on peptides with a significant fold-change led to a bigger decrease in the amount of available information (n = 15 peptides from 6 genes), which was not enough to obtain a correlation estimate. Hence, it is evident that even though uniquely mapping peptides constitute a valuable source of information, they represent only a minority of the cases, and therefore most of the data are lost when following such integration approach.

Alternatively, I evaluated a novel analysis strategy that enables incorporating information from the larger set of ubiquitous peptides. Briefly, such strategy relies on the use of the fold-change information from those peptides that overlap uniquely with each given DTU event (*i.e.* peptides that support more than one event are not considered for analysis, similarly to the approach followed in the previous section for the validation of SE; see **Figure 5.1**). This is based on the assumption that changes in protein expression levels will be preceded by a change in transcript expression levels. Implementation of such analysis approach resulted in a Spearman correlation coefficient of 0.31 when taking into account all the available peptides (n = 336 from 80 genes; p-value = $9.45 \cdot 10^{-9}$; **Table 5.4**), and a correlation of 0.43 when focusing on those with a significant fold-change (n = 101 from 40 genes; p-value = $9.04 \cdot 10^{-6}$; **Table 5.4**).

5.2.3 The correlation estimates can be improved by incorporating information on major transcripts

Given the observation that the correlation estimates improve when filtering out peptides with non-significant fold-changes, I next evaluated whether similar effects could be obtained by adopting an analogous approach for the RNA-seq data. Such strategy consisted on focusing on those DTU events that involve major transcripts, and discarding those that affect lowly abundant ones. Notably, this resulted in an increase of the observed correlation coefficients, with a value of 0.48 when taking into account all the peptides (n = 270 peptides from 54 genes; p-value < $2.20 \cdot 10^{-16}$; **Table 5.4**) and 0.78 when focusing on peptides with a significant fold-change (n = 66 peptides from 27 genes; p-value = $1.47 \cdot 10^{-14}$; **Table 5.4**).

	all DTU events			
	number of peptides	**number of genes**	r_S	**p-value**
peptides that map to only one transcript	48	21	0.41	3.67E-03
all peptides	336	80	0.31	9.45E-09
subset based on fold-change significance	101	40	0.43	9.04E-06

	DTU events in major transcripts			
	number of peptides	**number of genes**	r_S	**p-value**
all peptides	270	54	0.48	<2.20E-16
subset based on fold-change significance	66	27	0.78	1.47E-14

Table 5.4 | Validation of differential transcript usage events. Peptide fold-changes (FCs) were correlated with RNA-seq FCs for those transcripts involved in differential transcript usage events (DTU; see **Figure E.1** for the corresponding scatter plots). Two different integration strategies are evaluated here: either taking into account all the DTU events (top) or only those that affect major transcripts (bottom). Incorporating information on major transcripts leads to an increase in the observed correlation coefficients.

5.2.4 The detected correlation is not driven by differences in gene expression levels

The results obtained in the previous sections indicate that there is a correlation between the fold-changes predicted by RNA-seq and those obtained by SWATH-MS, thus supporting the hypothesis that the detected changes in protein abundance could arise from changes in transcript usage. Nonetheless, DTU is not isolated from differential gene expression, and it is possible that the observed agreement is primarily driven by overall differences in gene expression levels. In order to dissect the impact of such events, I repeated the above analyses by focusing on the DTU events from those genes that are not differentially expressed (n = 457 genes; see Methods). In general, the results obtained are consistent with those previously reported, with comparable correlation coefficients (see **Table 5.4**). Similarly, only one of the genes predicted to undergo a SE was also detected as differentially expressed (*i.e. HELLS*; see **Table 5.3**). Overall, these results indicate that differences in gene expression levels are not the main cause behind the observed correlation.

	all DTU events (excluding DGE)			
	number of peptides	**number of genes**	r_S	**p-value**
peptides that map to only one transcript	47	20	0.40	5.81E-03
all peptides	283	70	0.37	1.24E-10
subset based on fold-change significance	84	35	0.55	6.85E-08

	DTU events in major transcripts (excluding DGE)			
	number of peptides	**number of genes**	r_S	**p-value**
all peptides	245	47	0.48	7.37E-16
subset based on fold-change significance	61	24	0.80	1.13E-14

Table 5.5| Validation of differential transcript usage events after excluding differentially expressed genes. The analysis strategy followed is analogous to that described in **Table 5.3**, and the corresponding scatter plots can be found in **Figure E.2**. The observed correlation is not biased by changes in gene expression levels.

5.3 Discussion

In the present chapter, I describe a novel analysis approach for the validation of differential splicing events at the protein level. Importantly, the findings reported here rely on the use of SWATH-MS, a novel proteomics method that has not yet been applied to the study of protein isoforms [Gillet et al., 2012]. Specifically, and unlike targeted proteomics [Domon and Aebersold, 2010], peptide quantification with SWATH-MS is performed in a data-independent fashion, hence enabling the detection of a larger number of peptides. Furthermore, this technology also offers higher reproducibility rates than traditional shotgun methods [Domon and Aebersold, 2010], as demonstrated by the fact that SWATH-MS results correlate closely with those obtained from targeted proteomics experiments (*i.e.* 98% correlation; Liu et al. [2013]).

Previous attempts to integrate transcriptomics and proteomics information at the transcript level have focused on the detection of protein isoforms, rather than in the validation of predicted changes in splicing (*e.g.* Blakeley et al. [2010]; Ezkurdia et al. [2012]; Leoni et al. [2011]). Such integration has been limited by the use of uniquely mapping peptides, and further studies have relied on the existence of a matched RNA-seq dataset to refine the identification of the set of expressed proteins in a given sample (*e.g.* Sheynkman et al. [2013]). The approach introduced in this chapter relies on a similar strategy; however it differs from the aforementioned studies not only in the main goal of the analysis, but also in the amount of usable peptide information. By assuming that changes in protein expression are more likely to arise from those transcripts for which mRNA levels have already been detected to fluctuate, I show here that it is possible to retain information from ubiquitous peptides, *i.e.* those that map to more than one transcript within a given gene. Following filtering of instances of differential splicing predicted from the RNA-seq data based on peptide availability, it was possible to validate a large proportion of switch events (*i.e.* 9 out of 11 events), hence indicating that extreme changes in splicing can be recapitulated at the protein level. Subsequently, the application of this analysis strategy to the entire set of differentially used transcripts led to a 43% correlation between the RNA-seq and SWATH-MS fold-changes. Further prioritisation of peptide assignments based on information about major transcripts resulted in an improved correlation of 78%, which exceeds the one reported in previous studies [Li et al., 2014; Lundberg

et al., 2010; Nagaraj et al., 2011]. Altogether, these findings evidence that the transcript relative abundances derived from RNA-seq experiments can assist in the interpretation of proteomics data, and support the idea that differential splicing events in minor transcripts correspond to subtle changes that do not have a strong impact on protein levels.

Reasons for the lack of correlation amongst the observed mRNA and protein levels most likely include the effects of a variety of cellular processes that are not taken into account in the experiments performed here (*e.g.* post-transcriptional regulation, translational efficiency, mRNA and protein turnover...), as well as limitations inherent to the technologies used. Regarding the former, it is possible that the importance of the previously mentioned processes has been underestimated, and that transcription has a less decisive role in shaping mRNA expression levels than expected. However, recent observations by Li et al. [2014] strongly contradict this scenario, suggesting that mRNA levels could explain ~81% of the variance in protein levels, and indicating that further work is needed to improve the current understanding of the remaining sources of variation that might affect protein levels. On the other hand, proteomics studies show that observing a protein is unlikely unless there are at least a certain number of RNA molecules per cell [Ramakrishnan et al., 2009]. In this context, it is difficult to distinguish between transcriptional noise and insufficient sensitivity of the methods used, and it is expected that further technological advances will result in a higher throughput, lower detection limits, and better correlation estimates. Finally, protein folding, localisation and post-translational modifications all have prominent roles in establishing the function of proteins, which cannot be assessed from their mere detection. Thus, trying to predict which changes in mRNA levels will eventually have a functional impact *vs.* what proportion can be attributed to transcriptional and splicing noise constitutes a much more complicated question.

Overall, even though the results described here are mostly exploratory, they show that it is possible to attempt validation of the changes in splicing predicted from RNA-seq data by relying on information from SWATH-MS experiments. Future steps of this ongoing project will include the automation of the described analysis workflow, and the development of a testing approach to expand the validation of differential splicing beyond switch events. Eventually, all this information will be used to revisit the biological insights described in Chapter 4, and to further

understand the consequences of knocking-down *PRPF8*. In the longer term, the described strategy can be extended in order to take into consideration not only the information derived from major transcripts, but that from all the expressed transcripts in a given sample. This would allow for a more accurate evaluation of the true extent of alternative splicing at the protein level, and would help in assessing the coding potential of minor and novel transcripts. However, such advantages would come at the cost of increasing the complexity of peptide assignment. This could be compensated by having access to larger amounts of data, and it is expected that future technological advances will allow for an even more high throughput integration.

5.4 Computational methods

All the computational analyses described in this chapter have been carried out by myself, unless otherwise indicated in the text. The experimental work has been performed by Dr. Vi Wickramasinghe and Dr. Yansheng Liu.

Analysis of the RNA-seq data

The RNA-seq dataset used here corresponds to the one analysed in the previous chapter of the thesis, and includes 3 control samples and 4 *PRPF8* KDs (CAL51 cells; see Chapter 4 - Methods and **Table A.3**). Raw reads were directly mapped to the transcriptome with Bowtie v0.12.7 [Langmead et al., 2009], using Ensembl v66 as a reference [Flicek et al., 2012]. Following the estimation of transcript expression levels with MMSEQ v1.0.7 [Turro et al., 2011], its companion tool MMDIFF [Turro et al., 2014] was used to identify both differentially expressed genes (n = 1,607 genes; **Table 5.1**) and differentially used transcripts (n = 668 transcripts from 540 genes; **Table 5.1**). A posterior probability of 0.9 was used as the significance threshold in both analyses. Switch events within the set of genes identified to undergo differential transcript usage were identified with SwitchSeq (Gonzàlez-Porta and Brazma [2014]; n = 150 genes). Switch events that involved major transcripts with identical protein sequences were removed from the analyses (n = 8; final set of switched genes = 142; **Table 5.1**).

RNA-seq fold-changes were calculated from the transcript-level expression estimates. For each transcript, the fold-change represents the median transcript ex-

pression in *PRPF8* KD *vs.* control samples.

Analysis of the SWATH-MS data

Protein expression levels were assessed with SWATH-MS [Liu et al., 2013] for a set of 3 control and 3 *PRPF8* KD samples. An initial set of 16,779 peptides were detected and mapped against all the protein coding transcripts annotated in Ensembl v66, including those with a nonsense-mediated decay biotype. Removal of peptides that mapped to more than one gene led to a set of 14,648 peptides (corresponding to 2,800 genes), which was used for downstream analysis (**Table 5.1**).

Raw peptide intensities were first quantile-normalised in order to enable comparison across samples. For each peptide, the observed intensities across the biological replicates in each condition were summarised by using the median, and a fold-change was obtained by dividing the value obtained for *PRPF8* KD *vs.* control samples. Fold-change significance was assessed with a t-test, and a p-value of 0.1 was used as the significance threshold.

Integration of RNA-seq and SWATH-MS fold-change estimates

The fold-changes derived from these two technologies were integrated as described in **Figure 5.1**. Spearman correlation was used to evaluate the relationship between transcript and peptide fold-changes, as suggested by Maier et al. [2009]. Information on gene functions was retrieved from Genecards [Safran et al., 2010], unless otherwise indicated in the text.

Chapter 6

Conclusions

Splicing is only one of the many processes that shape the final set of RNA molecules present in eukaryote cells; nonetheless, it emerges as the most prominent mechanism for message diversification. Recently, the introduction of RNA sequencing has facilitated the study of this mechanism, and has allowed for a thorough characterisation of transcriptome composition. In this thesis, I have focused on the application of this technology to the study of human transcript diversity, as well as its potential impact on the protein repertoire.

Most protein coding genes express one dominant transcript in a given condition

High throughput sequencing technologies have contributed significantly to the efforts for the identification and annotation of expressed loci within the human genome, as illustrated by the ENCODE and GENCODE projects, respectively [ENCODE Project Consortium et al., 2012; Harrow et al., 2012]. These have resulted in high quality annotations, characterised by the existence of a diversity of gene types and transcript isoforms. However, despite these efforts, the contribution of each of the annotated transcripts to the overall transcriptome diversity in a given sample had remained largely uncharacterised. The findings derived from Chapter 2 provide a first answer to this question, by showing that gene expression tends to be dominated by one transcript in a given condition. In the future, evaluation of major transcript expression patterns across conditions and species will not only allow for the characterisation of splicing programs, but might also contribute to improving the existing annotation. For example, information about recurrent ma-

jor and minor transcripts could be displayed along with transcript structures in order to inform about the prevalence of the annotated features, and major transcript dominance could be used to identify sets of genes that are prone to express several transcripts simultaneously. Furthermore, assessing the impact of unnannotated features will be of particular interest: even though novel transcripts have been reported to be expressed at lower levels (*e.g.* Djebali et al. [2012]), they might still represent the most abundant transcript in a specific set of genes. Eventually, a bigger challenge will be to understand the effect of major and minor transcripts at the protein level, and to evaluate possible functional differences among alternative transcripts. Integrative analysis of RNA-seq and proteomics data, together with a better understanding of the structural and evolutionary properties that might differentiate a given set of transcripts, will be essential towards accomplishing these goals.

Alterations in splicing are widespread in kidney cancer, even though extreme changes are rare

The applications of RNA sequencing go far beyond the improvement of the existing annotation. In a clinical context, such technology opens the door for an unprecedented characterisation of the changes in transcriptome composition that might govern a specific disease, offering new possibilities for biomarker discovery and drug design [Costa et al., 2013]. For example, the findings reported in Chapter 3 show that the splicing process is broadly altered in ccRCC; however, analysis of major transcript expression patterns evidences the subtlety of most of the detected alterations. In the future, it will be interesting to investigate whether similar results are observed for other cancer types. Specifically, identification of sets of genes that commonly undergo extreme changes in splicing without being differentially expressed is likely to result in novel findings, since these effects have been largely ignored in the past. Consistency in the detected switch events across tumour types could provide extra evidence for a functional role of the involved transcripts, whilst high variability in major transcripts would be supportive of a general dysregulation of splicing programs. Lastly, given enough sample size, exploration of the relationship between switch events and mutation status at the patient level might provide evidence for the existence of cis-regulated events, and correlation of alterations in splicing with clinical variables might prove useful for

biomarker discovery and tumour stratification.

PRPF8 abundance dictates the patterns of alternative and constitutive mRNA splicing

The impact of alternative transcript products in disease can be further understood by gaining a deeper understanding of the role of the many proteins involved in the splicing reaction. In Chapter 4, I show that fluctuations in the concentration of *PRPF8* can result in alterations of the rate of splicing, which in turn can cause changes in splice site choice and lead to alternative splicing events. Even though the chain of events investigated in this chapter corresponds to a somehow artificial scenario, changes in the concentration and function of spliceosomal components are known to occur in development and cell cycle progression [Lane et al., 2013; Park et al., 2004], and have also been reported across tissues and in disease contexts [Grosso et al., 2008; Matera and Wang, 2014]. *PRPF8* mutations are found in autosomal dominant retinitis pigmentosa [Tanackovic et al., 2011], and it remains to be seen whether those also affect the kinetic competition equilibrium between weak and strong splice sites, just as observed when this gene is downregulated. Ultimately, knowledge on the mechanistic basis of the splicing reaction is essential to predict the effects of perturbations of the system, an exercise regarded as deciphering the splicing code [Barash et al., 2010].

Extreme changes in splicing can be detected at the protein level

Proteins derived from alternative splicing products have been detected in a variety of studies and contexts (*e.g.* Ezkurdia et al. [2012]; Leoni et al. [2011]). The main challenge in these analyses has been the limited number of peptides that can be used to uniquely identify relevant proteins, and past studies have relied on the use of RNA-seq datasets to create sample-specific databases that would optimise peptide assignment (*e.g.* Sheynkman et al. [2013]). In Chapter 5, I evaluated a similar approach with the goal of validating differential splicing events across two conditions of the same experiment. The findings derived from the first exploratory analyses indicate that it is possible to retain information from ubiquitously mapping peptides, by incorporating previous knowledge on major

transcripts. Moreover, they suggest that the changes in splicing detected at the transcriptomic level are likely to have an effect in the protein repertoire. In the future, expanding the analysis workflow to consider the abundances of all the expressed transcripts in a given sample would allow for the inspection of the coding potential of minor transcripts and unnanotated isoforms. Altogether, the knowledge derived from such integrative analyses will ultimately help in understanding the impact of alternative splicing at the protein level.

Concluding remarks

Previous to the establishment of RNA sequencing, microarrays constituted the only option for the genome-wide analysis of expression levels. Despite not being initially intended for the study of alternative transcripts, adequate probe design strategies allowed for the detection of alternative splicing events [Lee and Roy, 2004]. However, compared to sequencing, such an approach offers more limited throughput and a lower dynamic range in terms of expression levels of the targeted molecules [Zhao et al., 2014]. On the other hand, there are also challenges linked to the use of RNA sequencing for the characterisation of the transcriptome, the main ones associated with the short read length [Hooper, 2014]. In this context, transcript reconstruction remains a difficult, yet not impossible task. While exon-centric analysis approaches have been devised in order to circumvent the existing limitations, further technological developments will be crucial to overcome them. Indeed, during the timespan of this PhD, significant advances have already been achieved in terms of read length, evolving from 50 nucleotides a few years ago to more than 150 bases nowadays. The past years have also seen the rise of Pacific Biosciences and nanopore sequencing as alternatives to Illumina platforms for the study of full-length transcripts [Sharon et al., 2013]. Hence, it is a matter of time for transcript-level analyses to become central to any transcriptomics workflow.

Overall, RNA sequencing is only one of the many applications of high throughput sequencing. Novel techniques such as single cell sequencing are being established, and the routine application of this technology in the clinical sector is becoming closer to a reality. Furthermore, the rapid decrease in the costs of sequencing, together with the continuous increase in the amount of data obtained from each run, will contribute even further to the democratisation of sequencing [Shendure

and Ji, 2008]. Altogether, these facts evidence that the era of data production is just at its beginning, and that high throughput sequencing will continue to revolutionise many fields of basic and applied research.

Full list of publications

Gonzàlez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome Research* 22, 528–38 (2012).

Roux, J., Gonzàlez-Porta, M. & Robinson-Rechavi, M. Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs. *Nucleic Acids Research* 40, 5890–900 (2012).

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology* 14, R70 (2013).

Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–11 (2013).

't Hoen, P. et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology* 31, 1015–22 (2013).

Gonzàlez-Porta, M. & Brazma, A. Identification, annotation and visualisation of extreme changes in splicing from rna-seq experiments with switchseq. *bioRxiv* (2014). 10.1101/005967.

(submitted) Scelo, G.⋆, Riazalhosseini, Y.⋆, Greger, L.⋆, Letourneau, L.⋆, Gonzàlez-Porta, M.⋆ et al. Whole-genome sequencing reveals variation in the genomic landscape of clear cell Renal Cell Carcinoma in Europe.

(submitted) Wickramasinghe, V.⋆, Gonzàlez-Porta, M.⋆ et al. PRPF8 abundance dictates the patterns of alternative and constitutive messenger RNA splicing.

(submitted) Ferreira, P.⋆, Monlong, J.⋆, Gonzàlez-Porta, M.⋆, Barann, M.⋆ et al. Small changes in the big picture: genetic fine-tuning of transcriptome variation

⋆shared first authors

across human populations.

(submitted) Ballester, B. et al. Multi-species, multi-transcription factor binding atlas highlights the conserved control of tissue-specific pathways.

(in preparation) Diaz-Muñoz, M.D. et al. HuR-dependent regulation of mRNA splicing is essential for B cell antibody response.

Appendix A

RNA-seq datasets used in the thesis

Table A.1| RNA-seq data used in Chapter 2. All samples are polyA-selected. Paired reads are counted once.

ArrayExpress experiment id	ENA experiment id	ENA sample id	Source	Sample description	Sequencing platform	Read length (bp)	Sequencing depth (M reads)
E-MTAB-513	ERP000546	ERR030872	Body Map 2.0	thyroid	Illumina HiSeq 2000	50x2	81.91
E-MTAB-513	ERP000546	ERR030873	Body Map 2.0	testes	Illumina HiSeq 2000	50x2	81.84
E-MTAB-513	ERP000546	ERR030874	Body Map 2.0	ovary	Illumina HiSeq 2000	50x2	80.95
E-MTAB-513	ERP000546	ERR030875	Body Map 2.0	white blood cell	Illumina HiSeq 2000	50x2	81.22
E-MTAB-513	ERP000546	ERR030876	Body Map 2.0	skeletal muscle	Illumina HiSeq 2000	50x2	82.11
E-MTAB-513	ERP000546	ERR030877	Body Map 2.0	prostate	Illumina HiSeq 2000	50x2	82.33
E-MTAB-513	ERP000546	ERR030878	Body Map 2.0	lymph node	Illumina HiSeq 2000	50x2	82.08
E-MTAB-513	ERP000546	ERR030879	Body Map 2.0	lung	Illumina HiSeq 2000	50x2	79.30
E-MTAB-513	ERP000546	ERR030880	Body Map 2.0	adipose	Illumina HiSeq 2000	50x2	77.30
E-MTAB-513	ERP000546	ERR030881	Body Map 2.0	adrenal	Illumina HiSeq 2000	50x2	74.47
E-MTAB-513	ERP000546	ERR030882	Body Map 2.0	brain	Illumina HiSeq 2000	50x2	73.51
E-MTAB-513	ERP000546	ERR030883	Body Map 2.0	breast	Illumina HiSeq 2000	50x2	75.86
E-MTAB-513	ERP000546	ERR030884	Body Map 2.0	colon	Illumina HiSeq 2000	50x2	82.44
E-MTAB-513	ERP000546	ERR030885	Body Map 2.0	kindey	Illumina HiSeq 2000	50x2	80.40
E-MTAB-513	ERP000546	ERR030886	Body Map 2.0	heart	Illumina HiSeq 2000	50x2	82.92
E-MTAB-513	ERP000546	ERR030887	Body Map 2.0	liver	Illumina HiSeq 2000	50x2	80.05
E-GEOD-26284	SRP007461	SRR307897	ENCODE	cell_gm12878	Illumina GAII	70x2	32.21
E-GEOD-26284	SRP007461	SRR307898	ENCODE	cell_gm12878	Illumina GAII	70x2	47.00
E-GEOD-26284	SRP007461	SRR315330	ENCODE	cell_hela-s3	Illumina GAII	70x2	36.94
E-GEOD-26284	SRP007461	SRR315331	ENCODE	cell_hela-s3	Illumina GAII	70x2	37.45
E-GEOD-26284	SRP007461	SRR307926	ENCODE	cell_hepg2	Illumina GAII	70x2	41.14
E-GEOD-26284	SRP007461	SRR307927	ENCODE	cell_hepg2	Illumina GAII	70x2	41.16

E-GEOD-26284	SRP007461	SRR307905	ENCODE	cell_huvec	Illumina GAII	70x2	28.60
E-GEOD-26284	SRP007461	SRR307906	ENCODE	cell_huvec	Illumina GAII	70x2	35.87
E-GEOD-26284	SRP007461	SRR315336	ENCODE	cell_k562	Illumina GAII	70x2	35.75
E-GEOD-26284	SRP007461	SRR315337	ENCODE	cell_k562	Illumina GAII	70x2	38.67
E-GEOD-26284	SRP007461	SRR307899	ENCODE	cytosol_gm12878	Illumina GAII	70x2	42.69
E-GEOD-26284	SRP007461	SRR307900	ENCODE	cytosol_gm12878	Illumina GAII	70x2	36.58
E-GEOD-26284	SRP007461	SRR315334	ENCODE	cytosol_hela-s3	Illumina GAII	70x2	39.46
E-GEOD-26284	SRP007461	SRR315335	ENCODE	cytosol_hela-s3	Illumina GAII	70x2	33.63
E-GEOD-26284	SRP007461	SRR307928	ENCODE	cytosol_hepg2	Illumina GAII	70x2	37.95
E-GEOD-26284	SRP007461	SRR307929	ENCODE	cytosol_hepg2	Illumina GAII	70x2	37.39
E-GEOD-26284	SRP007461	SRR307917	ENCODE	cytosol_huvec	Illumina GAII	70x2	40.59
E-GEOD-26284	SRP007461	SRR307918	ENCODE	cytosol_huvec	Illumina GAII	70x2	37.08
E-GEOD-26284	SRP007461	SRR387661	ENCODE	cytosol_k562	Illumina GAII	70x2	124.83
E-GEOD-26284	SRP007461	SRR387662	ENCODE	cytosol_k562	Illumina GAII	70x2	88.45
E-GEOD-26284	SRP007461	SRR315297	ENCODE	nucleus_gm12878	Illumina GAII	70x2	42.04
E-GEOD-26284	SRP007461	SRR315298	ENCODE	nucleus_gm12878	Illumina GAII	70x2	38.98
E-GEOD-26284	SRP007461	SRR315332	ENCODE	nucleus_hela-s3	Illumina GAII	70x2	24.67
E-GEOD-26284	SRP007461	SRR315333	ENCODE	nucleus_hela-s3	Illumina GAII	70x2	36.51
E-GEOD-26284	SRP007461	SRR307915	ENCODE	nucleus_hepg2	Illumina GAII	70x2	34.18
E-GEOD-26284	SRP007461	SRR307916	ENCODE	nucleus_hepg2	Illumina GAII	70x2	28.53
E-GEOD-26284	SRP007461	SRR307909	ENCODE	nucleus_huvec	Illumina GAII	70x2	38.83
E-GEOD-26284	SRP007461	SRR307910	ENCODE	nucleus_huvec	Illumina GAII	70x2	37.69
E-GEOD-26284	SRP007461	SRR315299	ENCODE	nucleus_k562	Illumina GAII	70x2	38.23
E-GEOD-26284	SRP007461	SRR315300	ENCODE	nucleus_k562	Illumina GAII	70x2	35.68

Table A.2| RNA-seq data used in Chapter 3. All samples are polyA-selected. Paired reads are counted once.

Sample id	Sample description	Sequencing platform	Read length (bp)	Sequencing depth (M reads)
B00EUNU.BC08M3ACXX.6	cell line	Illumina HiSeq 2000	100x2	44.79
B00EUNZ.BC08T1ACXX.2	cell line	Illumina HiSeq 2000	100x2	72.84
B00EUNX.BD0C6HACXX.6	cell line	Illumina HiSeq 2000	100x2	36.04
B00EUNV.BC08T1ACXX.2	cell line	Illumina HiSeq 2000	100x2	48.21
B00EUNY.C11LHACXX.1	cell line	Illumina HiSeq 2000	100x2	83.30
B00EUNW.AC081RACXX.2	cell line	Illumina HiSeq 2000	100x2	66.74
B00E4GT.AC01UUABXX.8	normal	Illumina HiSeq 2000	100x2	87.96
B00E4GY.AC07UWACXX.3	normal	Illumina HiSeq 2000	100x2	51.07
B00E4GZ.AC07UWACXX.2	normal	Illumina HiSeq 2000	100x2	50.77
B00E4H2.BC07B5ACXX.6	normal	Illumina HiSeq 2000	100x2	46.91
B00E4H3.BC07B5ACXX.5	normal	Illumina HiSeq 2000	100x2	48.86
B00E4H4.BC07B5ACXX.4	normal	Illumina HiSeq 2000	100x2	46.11
B00E4H5.BC07B5ACXX.6	normal	Illumina HiSeq 2000	100x2	47.99
B00E4H7.BC07B5ACXX.5	normal	Illumina HiSeq 2000	100x2	45.04
B00E4H9.BC07B5ACXX.4	normal	Illumina HiSeq 2000	100x2	40.38
B00E4HO.AC07UWACXX.1	normal	Illumina HiSeq 2000	100x2	46.74
B00E4HP.BC07B5ACXX.7	normal	Illumina HiSeq 2000	100x2	43.61
B00E4HQ.AC07UWACXX.3	normal	Illumina HiSeq 2000	100x2	41.48
B00E4HR.AC07UWACXX.2	normal	Illumina HiSeq 2000	100x2	56.66
B00E4HT.AC07UWACXX.1	normal	Illumina HiSeq 2000	100x2	45.20
B00E4HV.BC07B5ACXX.7	normal	Illumina HiSeq 2000	100x2	35.24
B00E4I7.BC0GGFACXX.3	normal	Illumina HiSeq 2000	100x2	23.86
B00E4I9.AD0P8HACXX.1	normal	Illumina HiSeq 2000	100x2	45.22

B00E4IA.AD0P8HACXX.2	normal	Illumina HiSeq 2000	100x2	83.25
B00E4IB.AC0ML7ACXX.1	normal	Illumina HiSeq 2000	100x2	38.89
B00E4ID.BC0GGFACXX.3	normal	Illumina HiSeq 2000	100x2	62.59
B00E4IF.AD0P8HACXX.1	normal	Illumina HiSeq 2000	100x2	73.96
B00E4IG.AD0P8HACXX.2	normal	Illumina HiSeq 2000	100x2	45.18
B00E4IH.AC0ML7ACXX.1	normal	Illumina HiSeq 2000	100x2	92.38
B00E4IV.BC0GGFACXX.4	normal	Illumina HiSeq 2000	100x2	52.47
B00E4IW.BC0GGFACXX.4	normal	Illumina HiSeq 2000	100x2	35.61
B00E4JH.BC0GGFACXX.2	normal	Illumina HiSeq 2000	100x2	43.38
B00EXZC.BC0MLYACXX.4	normal	Illumina HiSeq 2000	100x2	41.17
B00EXZD.BC0MLYACXX.3	normal	Illumina HiSeq 2000	100x2	34.86
B00EXZV.BC0MLYACXX.3	normal	Illumina HiSeq 2000	100x2	61.76
B00EXZX.BC0MLYACXX.3	normal	Illumina HiSeq 2000	100x2	37.23
B00EXZY.BC0MLYACXX.4	normal	Illumina HiSeq 2000	100x2	54.07
B00EY0G.BC0MLYACXX.5	normal	Illumina HiSeq 2000	100x2	39.03
B00EY0H.BC0MLYACXX.4	normal	Illumina HiSeq 2000	100x2	44.71
B00EY0I.C0MP5ACXX.3	normal	Illumina HiSeq 2000	100x2	40.27
B00EY0Q.AC0RMCACXX.1	normal	Illumina HiSeq 2000	100x2	31.48
B00EY0S.AC0MFTACXX.1	normal	Illumina HiSeq 2000	100x2	41.40
B00EY0T.C0MP5ACXX.5	normal	Illumina HiSeq 2000	100x2	48.04
B00EY0U.AC0MFTACXX.2	normal	Illumina HiSeq 2000	100x2	51.27
B00EY0V.AC0MFTACXX.1	normal	Illumina HiSeq 2000	100x2	46.29
B00EY0Y.C0MP5ACXX.3	normal	Illumina HiSeq 2000	100x2	62.11
B00EY17.C0MP5ACXX.4	normal	Illumina HiSeq 2000	100x2	46.53
B00EY1D.C0LM0ACXX.6	normal	Illumina HiSeq 2000	100x2	52.45
B00EY1E.C0LM0ACXX.6	normal	Illumina HiSeq 2000	100x2	32.17
B00EY1G.AC0RMCACXX.4	normal	Illumina HiSeq 2000	100x2	27.97

B00F3DY.D16U7ACXX.4	normal	Illumina HiSeq 2000	100x2	45.21
B00E4HA.AC07UWACXX.1	tumor	Illumina HiSeq 2000	100x2	40.96
B00E4HB.BC07B5ACXX.7	tumor	Illumina HiSeq 2000	100x2	60.39
B00E4HE.BC07B5ACXX.6	tumor	Illumina HiSeq 2000	100x2	32.95
B00E4HF.BC07B5ACXX.5	tumor	Illumina HiSeq 2000	100x2	34.61
B00E4HG.BC07B5ACXX.4	tumor	Illumina HiSeq 2000	100x2	40.30
B00E4HH.BC07B5ACXX.6	tumor	Illumina HiSeq 2000	100x2	32.81
B00E4HJ.BC07B5ACXX.5	tumor	Illumina HiSeq 2000	100x2	33.86
B00E4HL.BC07B5ACXX.4	tumor	Illumina HiSeq 2000	100x2	32.24
B00E4HY.AC081RACXX.2	tumor	Illumina HiSeq 2000	100x2	54.14
B00E4HZ.BC08T1ACXX.2	tumor	Illumina HiSeq 2000	100x2	90.44
B00E4I0.BC08M3ACXX.6	tumor	Illumina HiSeq 2000	100x2	78.46
B00E4I1.BC08M3ACXX.6	tumor	Illumina HiSeq 2000	100x2	82.11
B00E4I3.AC07UWACXX.2	tumor	Illumina HiSeq 2000	100x2	39.71
B00E4I5.AC07UWACXX.2	tumor	Illumina HiSeq 2000	100x2	44.10
B00E4IJ.BC0GGFACXX.3	tumor	Illumina HiSeq 2000	100x2	43.88
B00E4IL.AD0P8HACXX.1	tumor	Illumina HiSeq 2000	100x2	48.99
B00E4IM.AD0P8HACXX.2	tumor	Illumina HiSeq 2000	100x2	44.66
B00E4IN.AC0ML7ACXX.1	tumor	Illumina HiSeq 2000	100x2	36.63
B00E4IP.BC0GGFACXX.3	tumor	Illumina HiSeq 2000	100x2	49.23
B00E4IR.AD0P8HACXX.1	tumor	Illumina HiSeq 2000	100x2	42.38
B00E4IS.AD0P8HACXX.2	tumor	Illumina HiSeq 2000	100x2	48.99
B00E4IT.AC0ML7ACXX.1	tumor	Illumina HiSeq 2000	100x2	28.99
B00E4J1.BC0GGFACXX.5	tumor	Illumina HiSeq 2000	100x2	58.12
B00E4J4.BC0GGFACXX.4	tumor	Illumina HiSeq 2000	100x2	63.60
B00E4J5.BC0GGFACXX.2	tumor	Illumina HiSeq 2000	100x2	32.92
B00E4J7.C0LM0ACXX.6	tumor	Illumina HiSeq 2000	100x2	82.45

B00E4JB.BC0MLYACXX.2	tumor	Illumina HiSeq 2000	100x2	43.76
B00EXZH.BC0MLYACXX.2	tumor	Illumina HiSeq 2000	100x2	33.27
B00EXZI.BC0MLYACXX.1	tumor	Illumina HiSeq 2000	100x2	22.40
B00EXZR.C0MP5ACXX.3	tumor	Illumina HiSeq 2000	100x2	38.68
B00EXZU.BC0MLYACXX.5	tumor	Illumina HiSeq 2000	100x2	39.16
B00EY02.BC0MLYACXX.5	tumor	Illumina HiSeq 2000	100x2	47.20
B00EY0J.C0MP5ACXX.5	tumor	Illumina HiSeq 2000	100x2	57.85
B00EY0M.C0MP5ACXX.4	tumor	Illumina HiSeq 2000	100x2	49.45
B00EY0N.C0MP5ACXX.4	tumor	Illumina HiSeq 2000	100x2	42.18
B00EY0O.C0MP5ACXX.3	tumor	Illumina HiSeq 2000	100x2	39.75
B00EY0R.AC0MFTACXX.2	tumor	Illumina HiSeq 2000	100x2	55.80
B00EY0Z.AC0RMCACXX.3	tumor	Illumina HiSeq 2000	100x2	32.13
B00EY16.C0MP5ACXX.5	tumor	Illumina HiSeq 2000	100x2	38.65
B00EY18.AC0MFTACXX.1	tumor	Illumina HiSeq 2000	100x2	42.81
B00EY19.C0MP5ACXX.5	tumor	Illumina HiSeq 2000	100x2	36.94
B00EY1A.C0MP5ACXX.4	tumor	Illumina HiSeq 2000	100x2	47.42
B00EY1B.AC0MFTACXX.2	tumor	Illumina HiSeq 2000	100x2	34.23
B00EY1K.C0LM0ACXX.3	tumor	Illumina HiSeq 2000	100x2	29.58
B00EY1L.C0LM0ACXX.7	tumor	Illumina HiSeq 2000	100x2	37.01

Table A.3| RNA-seq data used in Chapters 4 and 5. All samples are polyA-selected. Paired reads are counted once.

Sample id	Sample description	Sequencing platform	Read length (bp)	Sequencing depth (M reads)
C1	CAL51 control	Illumina HiSeq 2000	100x2	18.97
C2	CAL51 control	Illumina HiSeq 2000	100x2	26.19
C3	CAL51 control	Illumina HiSeq 2000	100x2	29.24
KD1	CAL51 PRPF8 KD	Illumina HiSeq 2000	100x2	24.66
KD2	CAL51 PRPF8 KD	Illumina HiSeq 2000	100x2	18.42
KD3	CAL51 PRPF8 KD	Illumina HiSeq 2000	100x2	28.80
KD4	CAL51 PRPF8 KD	Illumina HiSeq 2000	100x2	28.95

Appendix B

Supplementary Material for Chapter 2

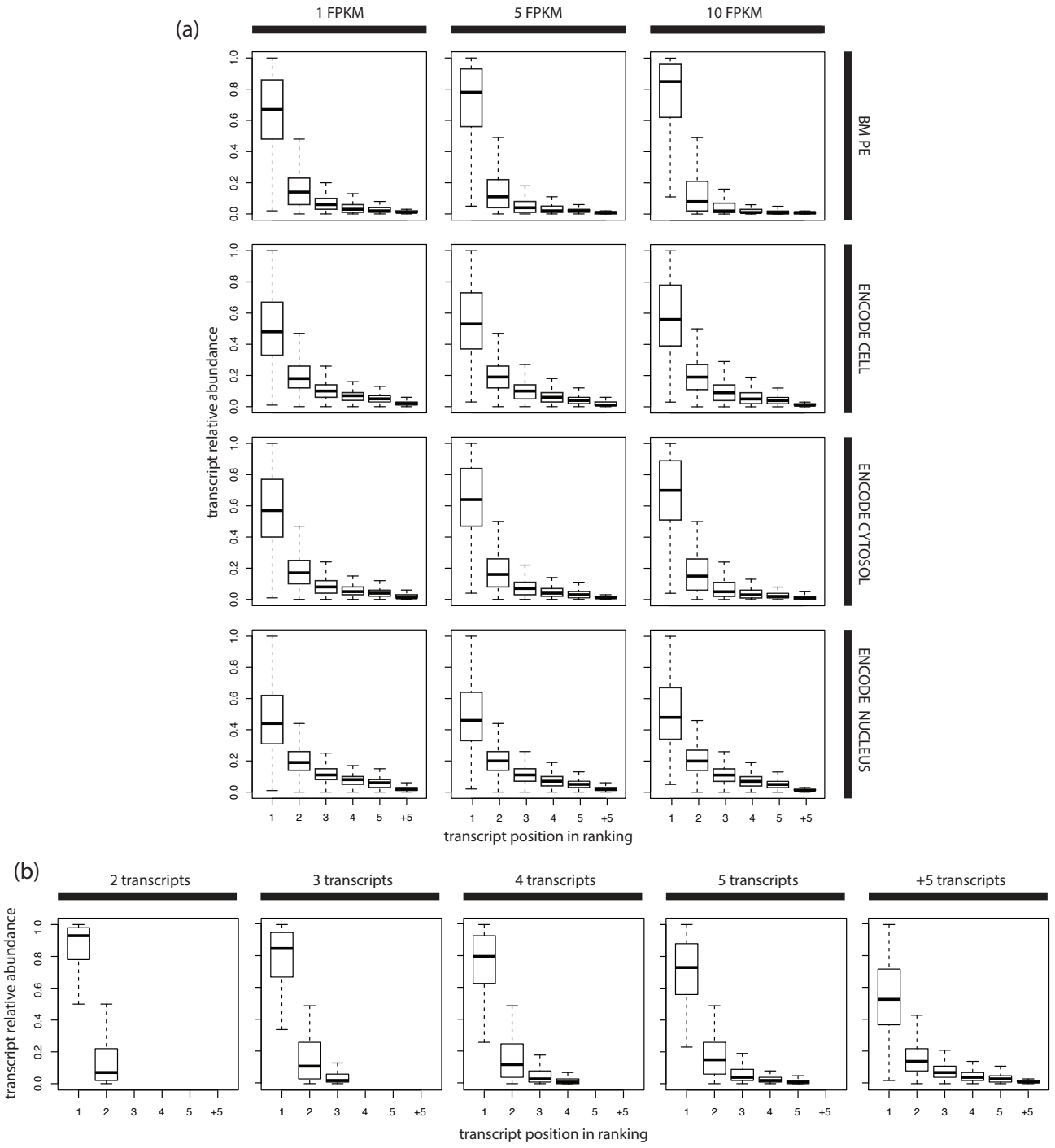


Figure B.1| Transcript relative abundances across all the studied datasets.
(a) Relative abundance of the subset of transcripts in each position of the ranking for all datasets, including different expression thresholds. Only genes with more than one annotated transcript are represented here.
(b) Relative abundance of the subset of transcripts in each position of the ranking for the BM dataset, separating genes by the number of annotated transcripts. The expression threshold used here is 1 FPKM.

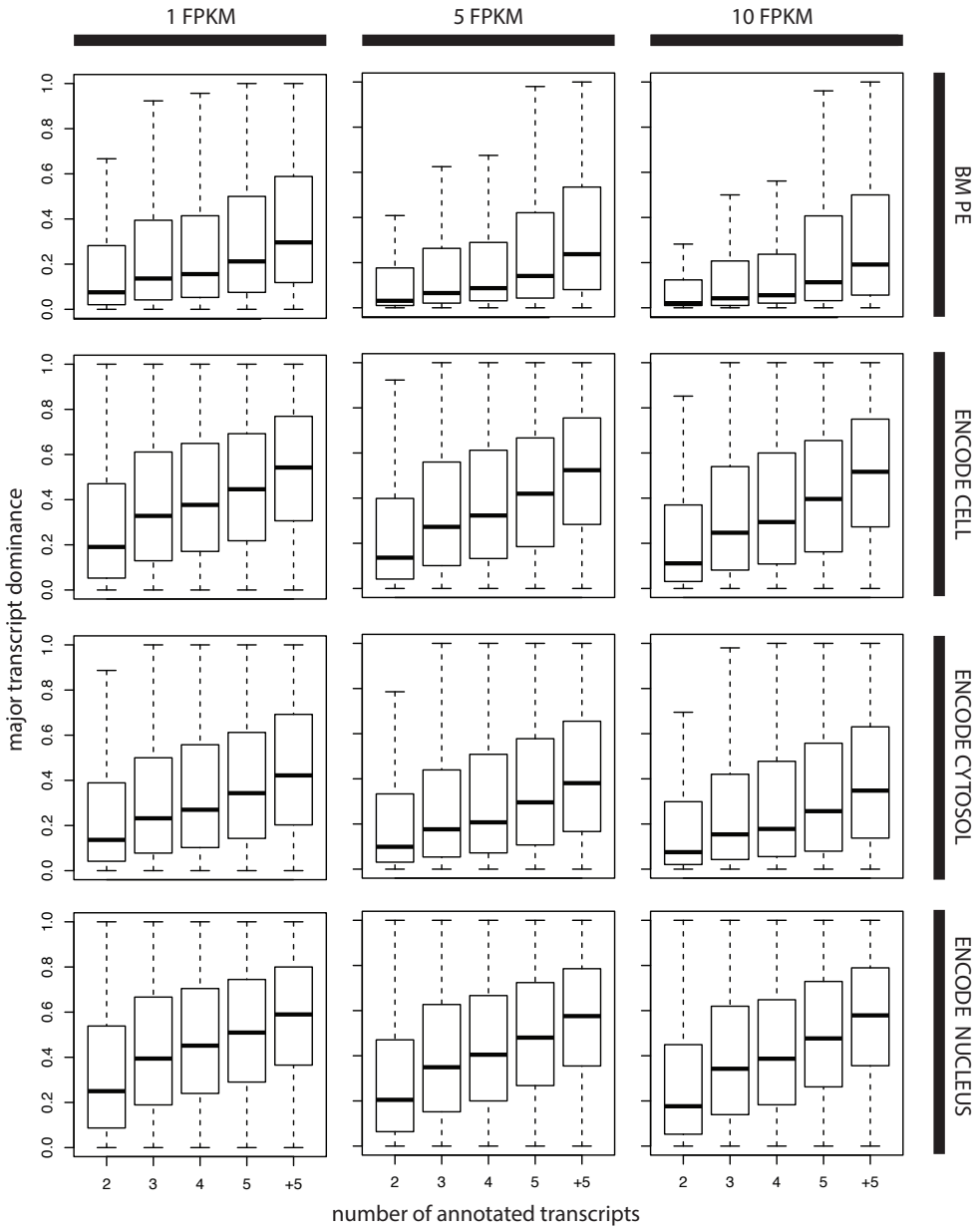


Figure B.2| Major transcript dominance across all the studied datasets. Major transcript dominance is measured by the ratio of expression between the major transcript and the second most abundant one.

term	padj
protein transport	6.85E-08
Transcription	1.46E-06
transcription regulation	7.27E-06
establishment of protein localization	1.73E-05
protein transport	2.19E-05
small GTPase mediated signal transduction	2.65E-05
respiratory chain	8.53E-05
transcription	2.34E-04
ubl conjugation pathway	2.50E-04
oxidative phosphorylation	4.36E-04
cellular macromolecule catabolic process	9.00E-04
macromolecule catabolic process	1.21E-03
electron transport	2.51E-03
mitochondrial ATP synthesis coupled electron transport	2.96E-03
regulation of transcription	3.11E-03
GTPase activity	3.36E-03
respiratory chain	3.44E-03
intracellular transport	6.93E-03
Oxidative phosphorylation	2.34E-02

Table B.1| GO enrichment analysis for recurrent 5-fold dominant genes.

term	padj
regulation of translational initiation	2.23E-03
mrna processing	3.83E-03
translation initiation factor activity	4.25E-03
mrna splicing	6.20E-03
RNA splicing	6.63E-03
RNA binding	7.65E-03
posttranscriptional regulation of gene expression	1.83E-02
translation regulation	2.28E-02
regulation of translation	2.31E-02
mRNA processing	2.51E-02
mRNA metabolic process	2.95E-02
RNA processing	3.39E-02
spliceosome	3.50E-02

Table B.2| GO enrichment analysis for genes that tolerate splicing.

	transcript position in the ranking					
	1	**2**	**3**	**4**	**5**	**>5**
			ENCODE CELL			
1 FPKM	71.36	14.06	5.62	2.91	1.74	4.31
5 FPKM	75.29	12.57	4.86	2.45	1.43	3.40
10 FPKM	80.14	10.52	3.90	1.91	1.08	2.45
			ENCODE CYTOSOL			
1 FPKM	81.05	10.54	3.40	1.63	0.95	2.44
5 FPKM	83.88	9.39	2.86	1.32	0.74	1.81
10 FPKM	86.77	8.04	2.30	1.02	0.55	1.31
			ENCODE NUCLEUS			
1 FPKM	63.49	16.20	7.53	4.15	2.54	6.09
5 FPKM	66.79	14.96	6.87	3.74	2.27	5.36
10 FPKM	71.71	12.97	5.89	3.17	1.90	4.36

Table B.3| mRNA pool estimates for the cell line dataset.

	expressed genes	**genes with a dominant major transcript**			
		2-fold dominance		**5-fold dominance**	
			ENCODE CELL		
1 FPKM	9770	6550	66.83	4061	41.29
5 FPKM	4952	3475	71.51	2334	49.01
10 FPKM	2787	2029	74.67	1449	54.73
			ENCODE CYTOSOL		
1 FPKM	9281	6139	66.29	3619	39.16
5 FPKM	5076	3590	71.31	2352	47.09
10 FPKM	3041	2259	74.81	1601	53.39
			ENCODE NUCLEUS		
1 FPKM	9923	5936	60.32	3182	32.62
5 FPKM	5796	3620	63.70	2085	37.53
10 FPKM	3516	2268	65.85	1391	41.34

Table B.4| Number of dominant transcripts in the cell line dataset.

term	padj
ribosomal protein	1.48E-04
protein biosynthesis	1.87E-04
ribosomal subunit	3.07E-04
structural constituent of ribosome	3.20E-04
ribosome	3.23E-04
translation	1.99E-03
ribosome	3.23E-03
small ribosomal subunit	2.03E-02
transit peptide	3.18E-02
mRNA transport	3.68E-02

Table B.5| GO enrichment analysis for genes with a major retained intron both in the nucleus and the cytosol.

Gene id	Transcript id	HGNC symbol	Reannotated biotype
ENSG00000029364	ENST00000031146	SLC39A9	protein coding
ENSG00000090054	ENST00000486910	SPTLC1	removed (artefact)
ENSG00000097033	ENST00000370558	SH3GLB1	removed (merged)
ENSG00000101150	ENST00000474176	TPD52L2	retained intron
ENSG00000103549	ENST00000562110	RNF40	protein coding (NMD)
ENSG00000112159	ENST00000487831	MDN1	protein coding (NMD)
ENSG00000114999	ENST00000460450	TTL	processed transcript
ENSG00000116288	ENST00000469225	PARK7	protein coding
ENSG00000122696	ENST00000496760	SLC25A51	lncRNA
ENSG00000130703	ENST00000471817	OSBPL2	retained intron
ENSG00000136908	ENST00000495270	DPM2	retained intron
ENSG00000136935	ENST00000475407	GOLGA1	protein coding (NMD)
ENSG00000143149	ENST00000463610	ALDH9A1	processed transcript
ENSG00000148396	ENST00000467838	SEC16A	retained intron
ENSG00000149823	ENST00000527646	VPS51	procesed transcript
ENSG00000156502	ENST00000497254	SUPV3L1	retained intron
ENSG00000162775	ENST00000487146	RBM15	protein coding
ENSG00000166582	ENST00000472570	CENPV	retained intron
ENSG00000173812	ENST00000310837	EIF1	possible protein coding
ENSG00000185305	ENST00000502271	ARL15	possible protein coding
ENSG00000198917	ENST00000467582	C9orf114	protein coding
ENSG00000213995	ENST00000470164	CARKD	protein coding or NMD

Table B.6| Re-annotation of major processed transcripts.

Appendix C

Supplementary Material for Chapter 3

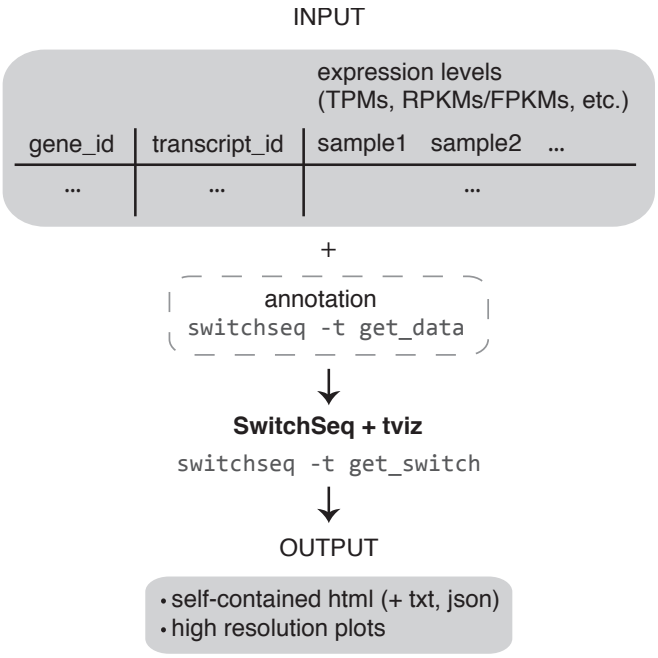
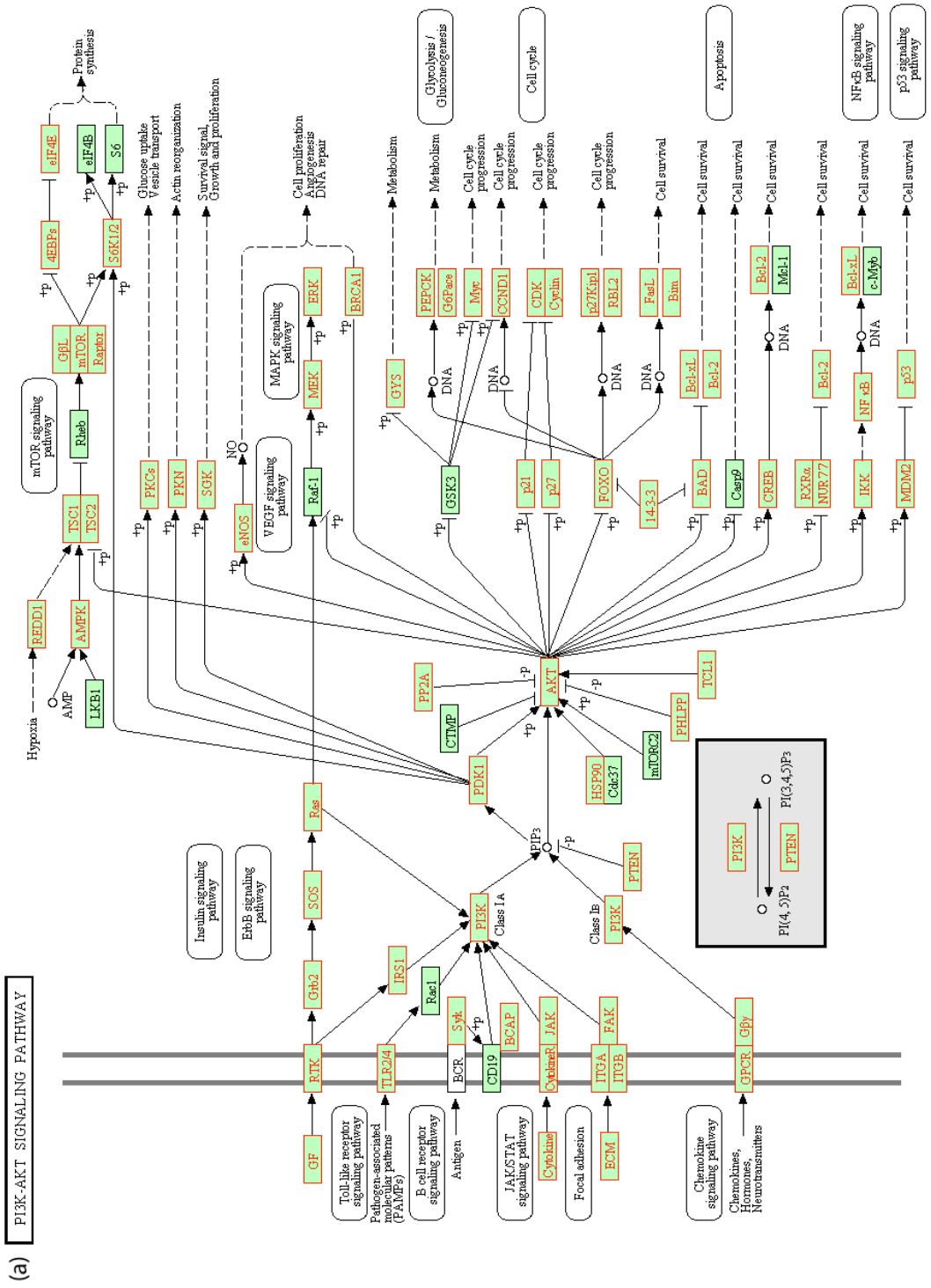


Figure C.1| SwitchSeq analysis workflow. Further information on installation and execution instructions can be found on the project website (`https://github.com/mgonzalezporta/SwitchSeq`).


(a)
PI3K-AKT SIGNALING PATHWAY
Hypoxia
AMP
LKB1
REDD1
AMPK
TSC1
TSC2
Rheb
GβL
mTOR
Raptor
mTOR signaling pathway
4EBPs
S6K1/2
eIF4E
eIF4B
S6
Protein synthesis
PKCs
PKN
SGK
Glucose uptake
Vesicle transport
Actin reorganization
Survival signal
Growth and proliferation
NO
eNOS
VEGF signaling pathway
MAPK signaling pathway
MEK
ERK
Raf-1
BRCA1
Cell proliferation
Angiogenesis
DNA repair
GYS
Metabolism
PEPCK
G6Pase
Glycolysis / Gluconeogenesis
Myc
CCND1
CDK
Cyclin
Cell cycle progression
Cell cycle
p27Kip1
RBL2
FasL
Bim
Cell survival
Apoptosis
GSK3
p21
p27
FOXO
14-3-3
BAD
Bcl-xL
Bcl-2
Casp9
CREB
Mcl-1
RXRα
NUR77
IKK
NF-κB
c-Myb
MDM2
p53
NF-κB signaling pathway
p53 signaling pathway
DNA
PP2A
CTMP
PDK1
AKT
PHLPP
TCL1
HSP90
Cdc37
mTORC2
PIP3
PTEN
PI3K
Class IA
Class IB
Ras
SOS
Grb2
IRS1
Rac1
TLR2/4
Syk
BCAP
CD19
JAK
FAK
ITGA
ITGB
RTK
GPCR
Gβγ
GF
Insulin signaling pathway
ErbB signaling pathway
Toll-like receptor signaling pathway
Pathogen-associated molecular patterns (PAMPs)
B cell receptor signaling pathway
Antigen
JAK/STAT signaling pathway
Cytokine
Focal adhesion
ECM
Chemokine signaling pathway
Chemokines, Hormones, Neurotransmitters
PI(4,5)P2
PI(3,4,5)P3


(b)
VEGFA
PTK2
FGF1
DDIT4
PGF
CCND1
LAMA4
VWF
PCK2
FGF11
EGFR
SYK
MYC
MLST8
ITGA5
PDGFD
OSMR
FN1
COL4A1
PCK1
CCND3
SGK3
CSF1R
G6PC
MTCP1
EGF
YWHAZ
NGF
SGK2
FLT1
CCND2
TSC1
COL1A2
COL5A2
IL2RG
MET
FGFR2
ITGA4
CDKN1A
TLR2
EIF4EBP1
ANGPT2
KDR
ITGB7
IL4R
PRKCZ
COL1A1
PPP2R5B
GNB1
COL6A2
IL7R
PIK3R3
IFNAR2
AKT2
NGFR
PRLR
GNB5
CSF1
LPAR2
RBL2
FGFR3
COL4A3
NR4A1
CSF3R
GNG11
THBS3
CREB5
COL3A1
YWHAH
GNB4
JAK3
IL3RA
GNGT2
COL4A2
THBS2
TNC
LAMA5
LAMC1


BCL2
VEGFC
EIF4E
FGF2
COL6A3
TLR4
RPS6KB2
ITGB4
IL7
PKN3
SOS1
EPHA2
ITGB6
PDGFRB
CREB3L3
EFNA5
PIK3CD
LAMB1
LAMB3
LAMA1
PDGFA
HGF
MTOR
EFNA1
ATF2
TEK
MAPK3
GNB2
COL5A3
PIK3R5
NOS3
GNG2
PDGFRA
PIK3AP1
LPAR6
EFNA3
PIK3R1
FLT4
COL4A4
MAPK1
COL5A1
KITLG
ITGA11
ITGA1
LAMC2
ITGA3
IL2RB
LAMC3
THBS1
PTEN
ITGB1
KIT
F2R
GNG7
INSR
PKN1
ITGB8
FGF7
IL2RA
GHR
MAP2K2
ITGA9
COL27A1
PDGFB
IL6
THBS4
ITGAV
SPP1
ITGA7
MDM2
ITGB3
BCL2L11
COL4A5
YWHAE
AKT3
GYS1
FGFR4
IRS1

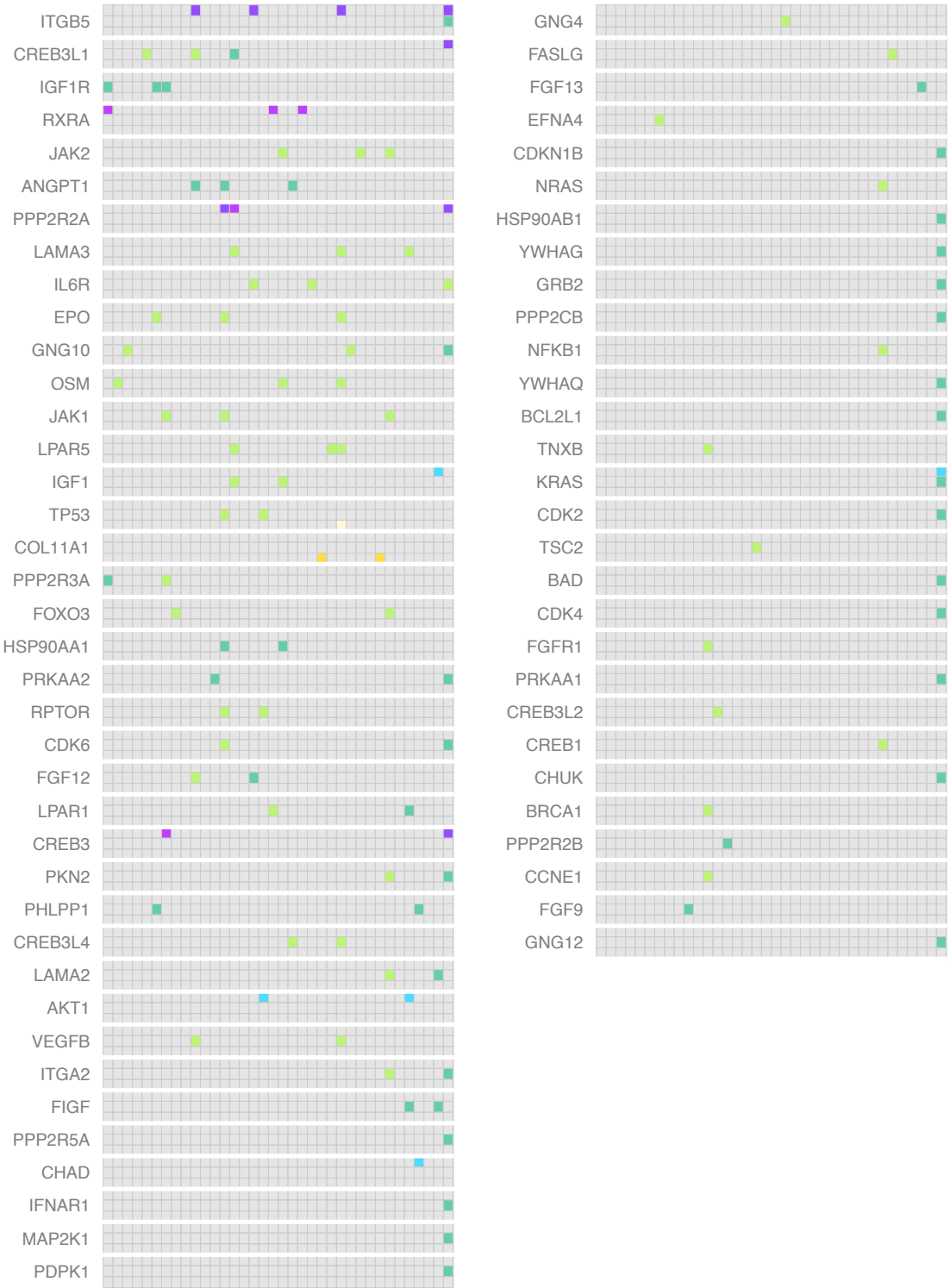
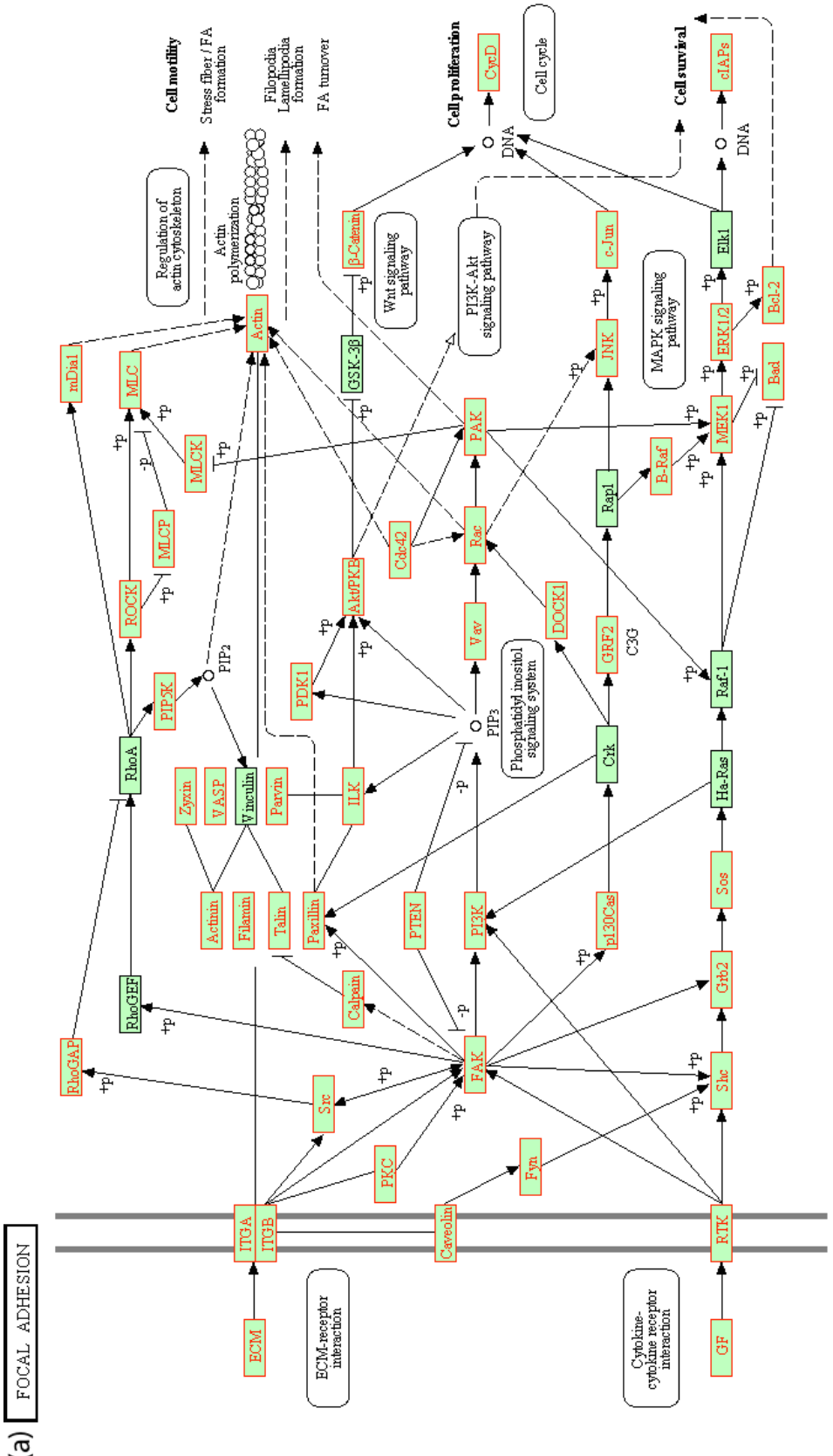

ITGB5
CREB3L1
IGF1R
RXRA
JAK2
ANGPT1
PPP2R2A
LAMA3
IL6R
EPO
GNG10
OSM
JAK1
LPAR5
IGF1
TP53
COL11A1
PPP2R3A
FOXO3
HSP90AA1
PRKAA2
RPTOR
CDK6
FGF12
LPAR1
CREB3
PKN2
PHLPP1
CREB3L4
LAMA2
AKT1
VEGFB
ITGA2
FIGF
PPP2R5A
CHAD
IFNAR1
MAP2K1
PDPK1
GNG4
FASLG
FGF13
EFNA4
CDKN1B
NRAS
HSP90AB1
YWHAG
GRB2
PPP2CB
NFKB1
YWHAQ
BCL2L1
TNXB
KRAS
CDK2
TSC2
BAD
CDK4
FGFR1
PRKAA1
CREB3L2
CREB1
CHUK
BRCA1
PPP2R2B
CCNE1
FGF9
GNG12

Figure C.2| Summary of detected alterations in the *PI(3)K-AKT-MTOR* signalling pathway.
(a) The PI(3)K-AKT-MTOR *signalling pathway.* Genes which have been identified as altered in any of the three considered analyses (*i.e.* switch events, gene expression changes and confirmed somatic mutations) have been highlighted in red.
(b) Patient-specific landscape of alterations for genes in the PI(3)K-AKT-MTOR *signalling pathway.* Each gene is represented by three tracks as described in **Figure 3.7**, and only those previously highlighted in **a** have been included. Genes have been sorted based on the total number of alterations of any kind.


(a) FOCAL ADHESION
ECM
ECM-receptor interaction
ITGA
ITGB
Caveolin
RTK
GF
Cytokine-cytokine receptor interaction
FAK
PKC
Src
Fyn
RhoGAP
RhoGEF
RhoA
ROCK
MLCP
MLC
MLCK
PIP2
PIP3K
mDia1
Actin
Regulation of actin cytoskeleton
Actin polymerization
Cell motility
Stress fiber / FA formation
Filopodia Lamellipodia formation
FA turnover
Zyxin
VASP
Vinculin
Parvin
ILK
Actinin
Filamin
Talin
Paxillin
Calpain
PDK1
Akt/PKB
GSK-3β
β-Catenin
Wnt signaling pathway
Cell proliferation
CycD
Cell cycle
DNA
PI3K-Akt signaling pathway
Cdc42
Rac
PAK
Vav
PTEN
PI3K
Phosphatidyl inositol signaling system
DOCK1
Crk
p130Cas
GRF2
C3G
Rap1
B-Raf
JNK
c-Jun
MAPK signaling pathway
Cell survival
cIAPs
Elk1
ERK1/2
MEK1
Raf-1
Ha-Ras
Sos
Grb2
Shc
Bad
Bcl-2
DNA


(b)
VEGFA
PTK2
CAV1
FYN
PGF
CCND1
LAMA4
CAV2
VWF
BCAR1
EGFR
BIRC3
ITGA5
PDGFD
FN1
VAV1
COL4A1
RAC2
CCND3
PPP1R12A
DIAPH1
EGF
FLT1
CCND2
ACTN4
COL1A2
COL5A2
MET
SRC
CAPN2
ITGA4
MAPK9
MAPK10
KDR
ITGB7
COL1A1
MYL5
COL6A2
PIK3R3
AKT2
ACTN1
COL4A3
RAPGEF1
BRAF
FLNA
SHC1
THBS3
FLNB
COL3A1
RAC3
PAK1
COL4A2
THBS2
PARVG
TNC
LAMA5
LAMC1
BCL2
VEGFC
COL6A3
PPP1R12B
ITGB4
SOS1
ITGB6
MAPK8
SHC2
PDGFRB
PIK3CD
LAMB1
LAMB3
LAMA1
PDGFA
HGF
MAPK3
COL5A3
PIK3R5
PDGFRA
PIK3R1
FLT4
COL4A4
ZYX
MAPK1
COL5A1
ITGA11
ITGA1
LAMC2
ITGA3
TLN2
LAMC3
THBS1

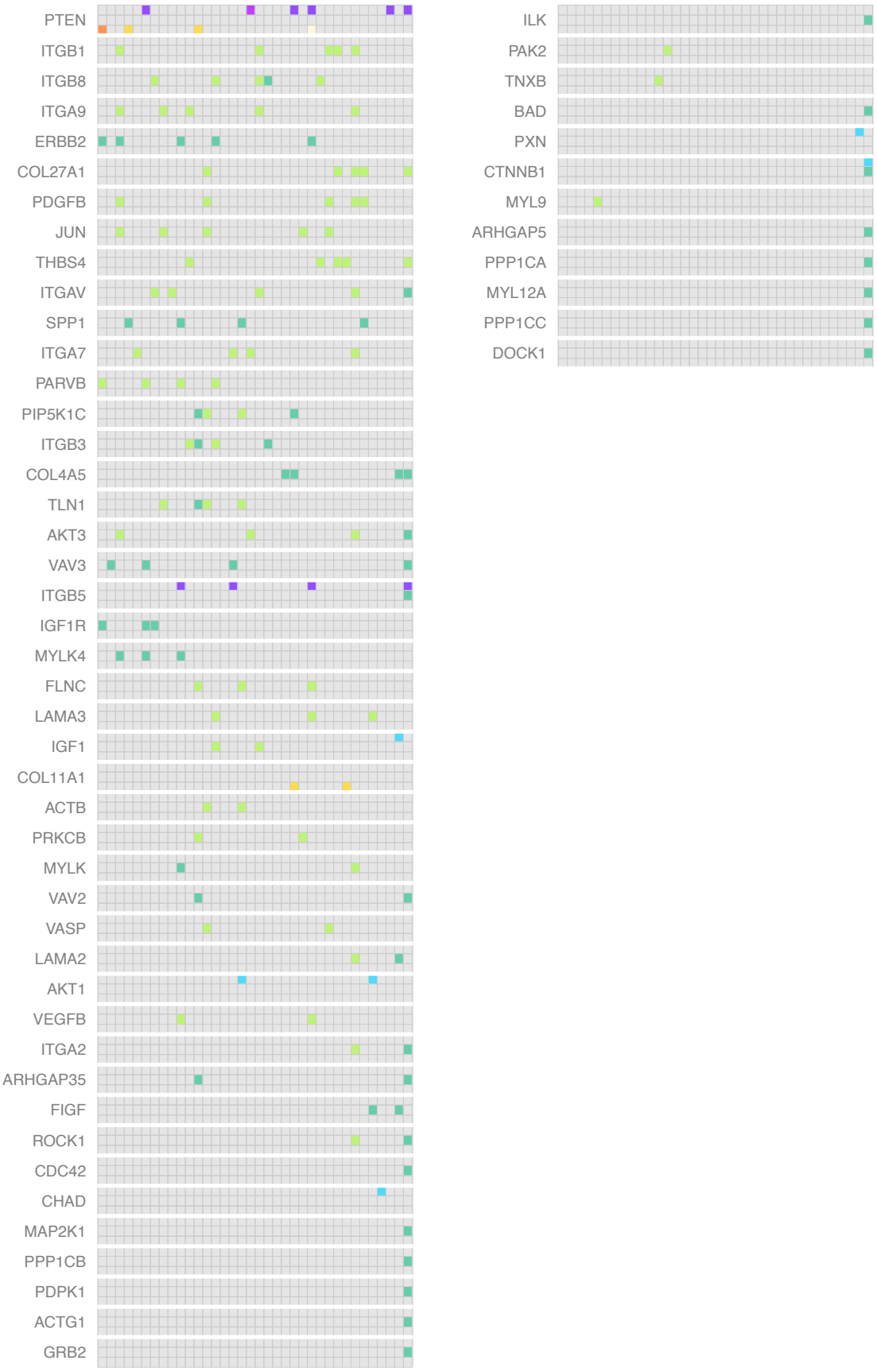

PTEN
ITGB1
ITGB8
ITGA9
ERBB2
COL27A1
PDGFB
JUN
THBS4
ITGAV
SPP1
ITGA7
PARVB
PIP5K1C
ITGB3
COL4A5
TLN1
AKT3
VAV3
ITGB5
IGF1R
MYLK4
FLNC
LAMA3
IGF1
COL11A1
ACTB
PRKCB
MYLK
VAV2
VASP
LAMA2
AKT1
VEGFB
ITGA2
ARHGAP35
FIGF
ROCK1
CDC42
CHAD
MAP2K1
PPP1CB
PDPK1
ACTG1
GRB2
ILK
PAK2
TNXB
BAD
PXN
CTNNB1
MYL9
ARHGAP5
PPP1CA
MYL12A
PPP1CC
DOCK1

Figure C.3| Summary of detected alterations in the focal adhesion pathway.
(a) The focal adhesion pathway. Genes which have been identified as altered in any of the three considered analyses (*i.e.* switch events, gene expression changes and confirmed somatic mutations) have been highlighted in red.
(b) Patient-specific landscape of alterations for genes in the focal adhesion pathway. Each gene is represented by three tracks as described in **Figure 3.7**, and only those previously highlighted in **a** have been included. Genes have been sorted based on the total number of alterations of any kind.

pathway	padj
Metabolic pathways	2.25E-127
Pathways in cancer	2.16E-049
Cell cycle	1.79E-034
MAPK signaling pathway	1.91E-028
Ubiquitin mediated proteolysis	1.32E-027
Regulation of actin cytoskeleton	5.06E-027
RNA transport	1.13E-026
Spliceosome	3.87E-026
Wnt signaling pathway	2.97E-023
Renal cell carcinoma	1.72E-019
Apoptosis	2.27E-019
Focal adhesion	4.78E-018
Adherens junction	2.29E-017
ErbB signaling pathway	7.36E-017
TGF-beta signaling pathway	3.40E-016
p53 signaling pathway	1.17E-015
mRNA surveillance pathway	1.84E-013

Table C.1| Pathway enrichment analysis for differentially spliced genes in cell lines *vs.* tumour samples.

Appendix D

Supplementary Material for Chapter 4

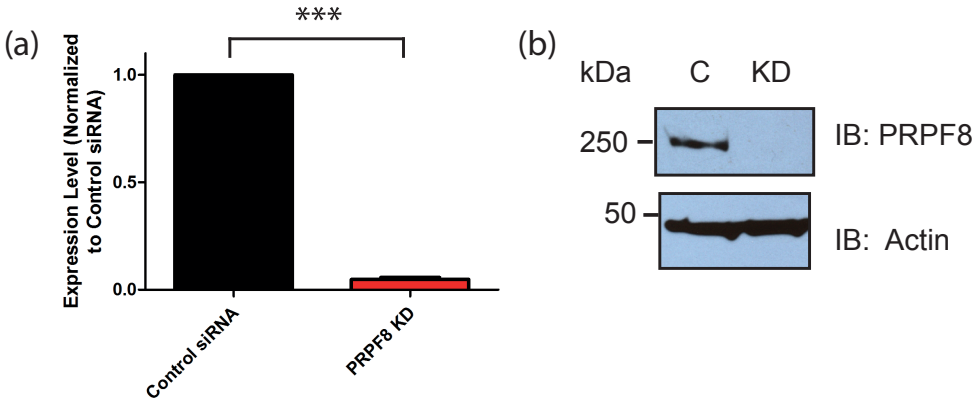


Figure D.1| Validation of *PRPF8* down-regulation. Cal51 cells were treated with either *PRPF8* siRNAs (*PRPF8* knock-down; KD), or all-stars siRNAs (control; C).
(a) qRT-PCR results. PRPF8 expression levels were efficiently down-regulated (p-value < 0.001).
(b) Western blot validation. The knock-down can also be recapitulated at the protein level.
Figure provided by Dr. Vi Wickramasinghe.

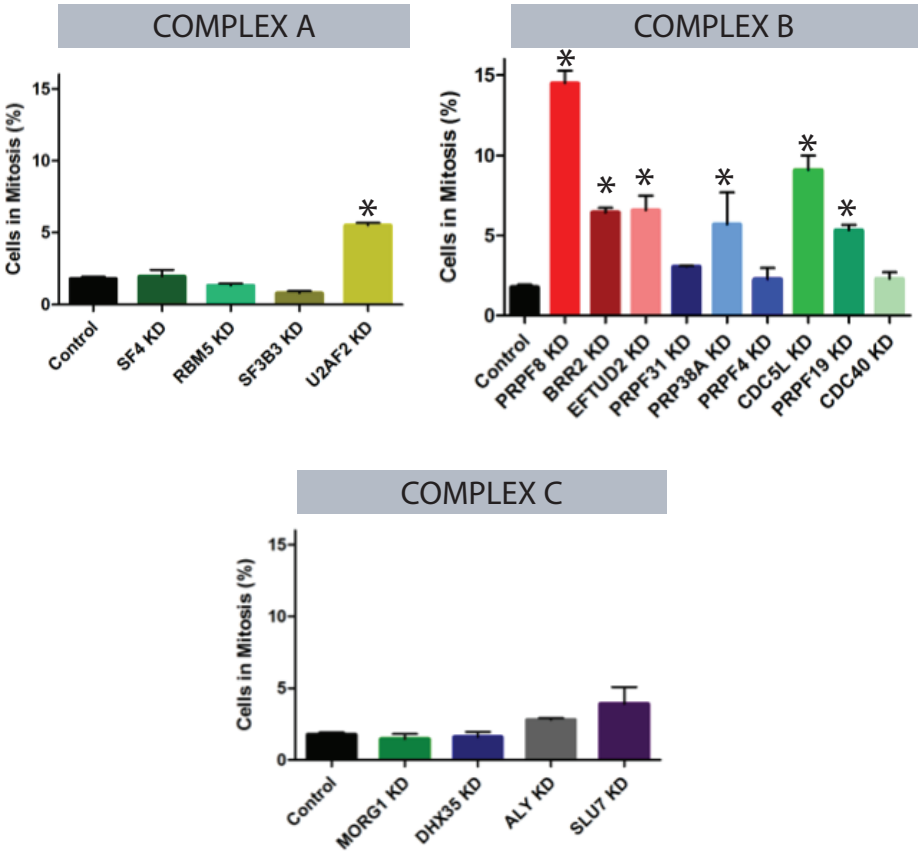


Figure D.2| Cell cycle analysis following knock-down of splicing factors from several spliceosomal complexes. Compared to other components of the A and C-complexes of the spliceosome, knock-down (KD) of those from complex B generally results in the accumulation of cells in mitosis. More specifically, KD of *PRPF8* has the biggest effects on the measured phenotype. Complex A: not significant (n.s.) except for *U2AF2* $p < 0.05$. Complex B: *PRPF8* $p < 0.001$; *BRR2*, *EFTUD2* $p < 0.01$; *PRPF31* n.s.; *PRPF38A* $p<0.05$; *PRPF4* n.s.; *CDC5L* $p<0.01$; *PRPF19* $p<0.05$; *CDC40* n.s. Complex C: n.s.
Figure provided by Dr. Vi Wickramasinghe.

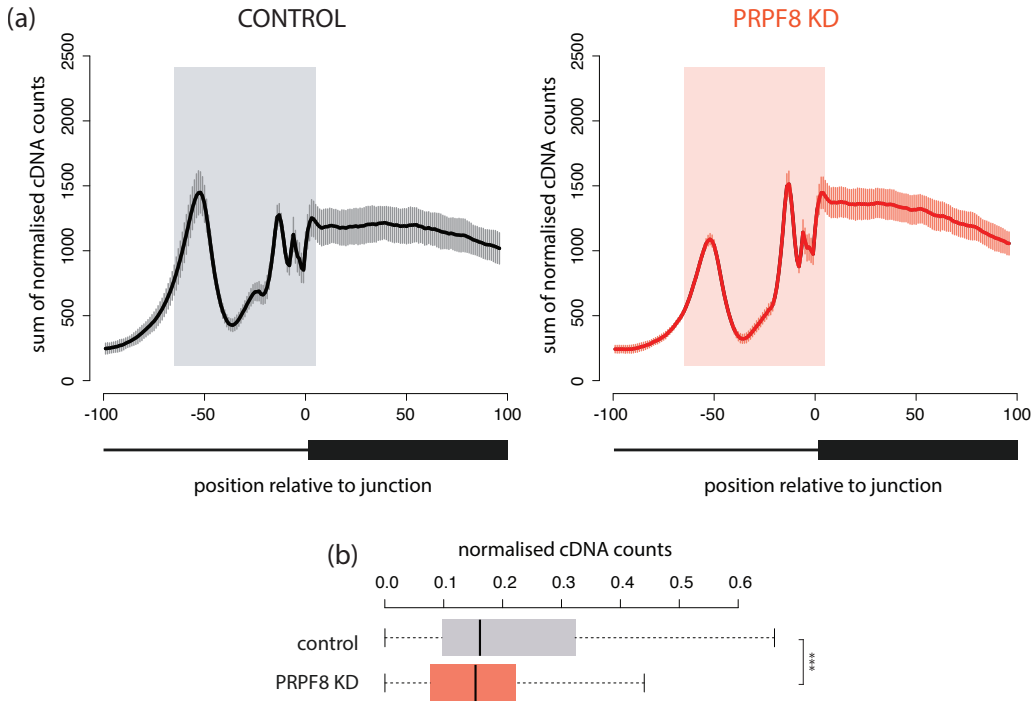


Figure D.3 | Results from Sm iCLIP experiments for the 3′ splice site. Plots to be interpreted as in **Figure 4.3**.
(a) RNA maps for control and PRPF8 *KD samples.*
(b) Changes in the distribution of normalised cDNA counts between control and PRPF8 *KD samples.*

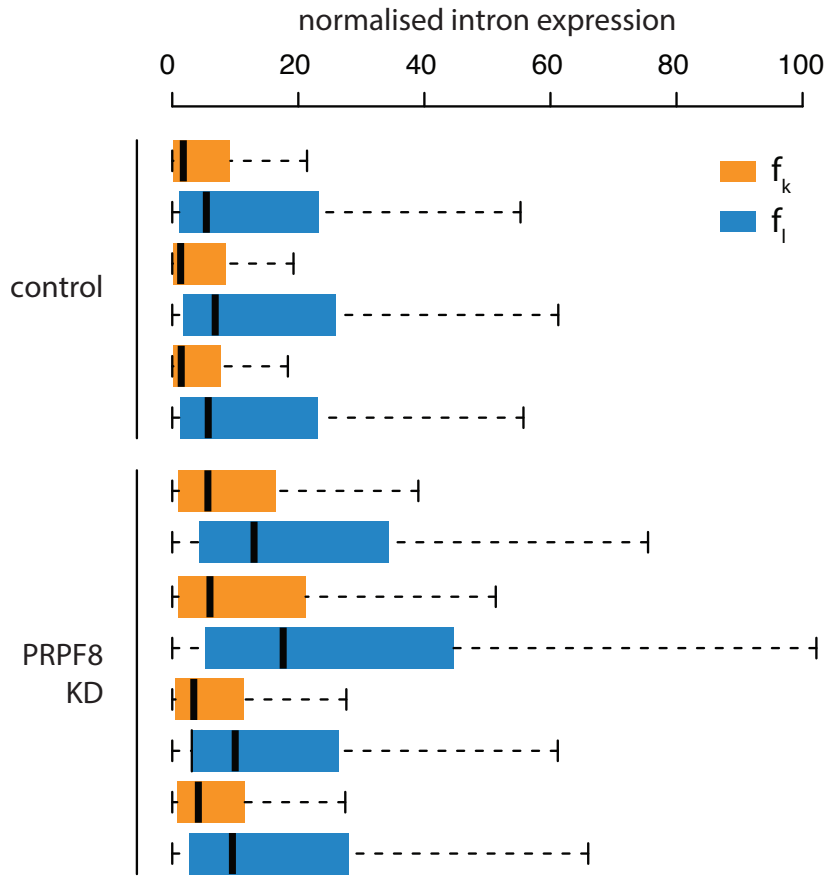


Figure D.4| Differences in intron expression between first and last introns. Intron counts were normalised to take into account intron length and both counts and length of adjacent internal exons (see Chapter 4 - Methods). In general, last introns (f_l) have higher expression levels than the ones located towards the 5′ end of the same transcript (f_k).

Appendix E

Supplementary Material for Chapter 5

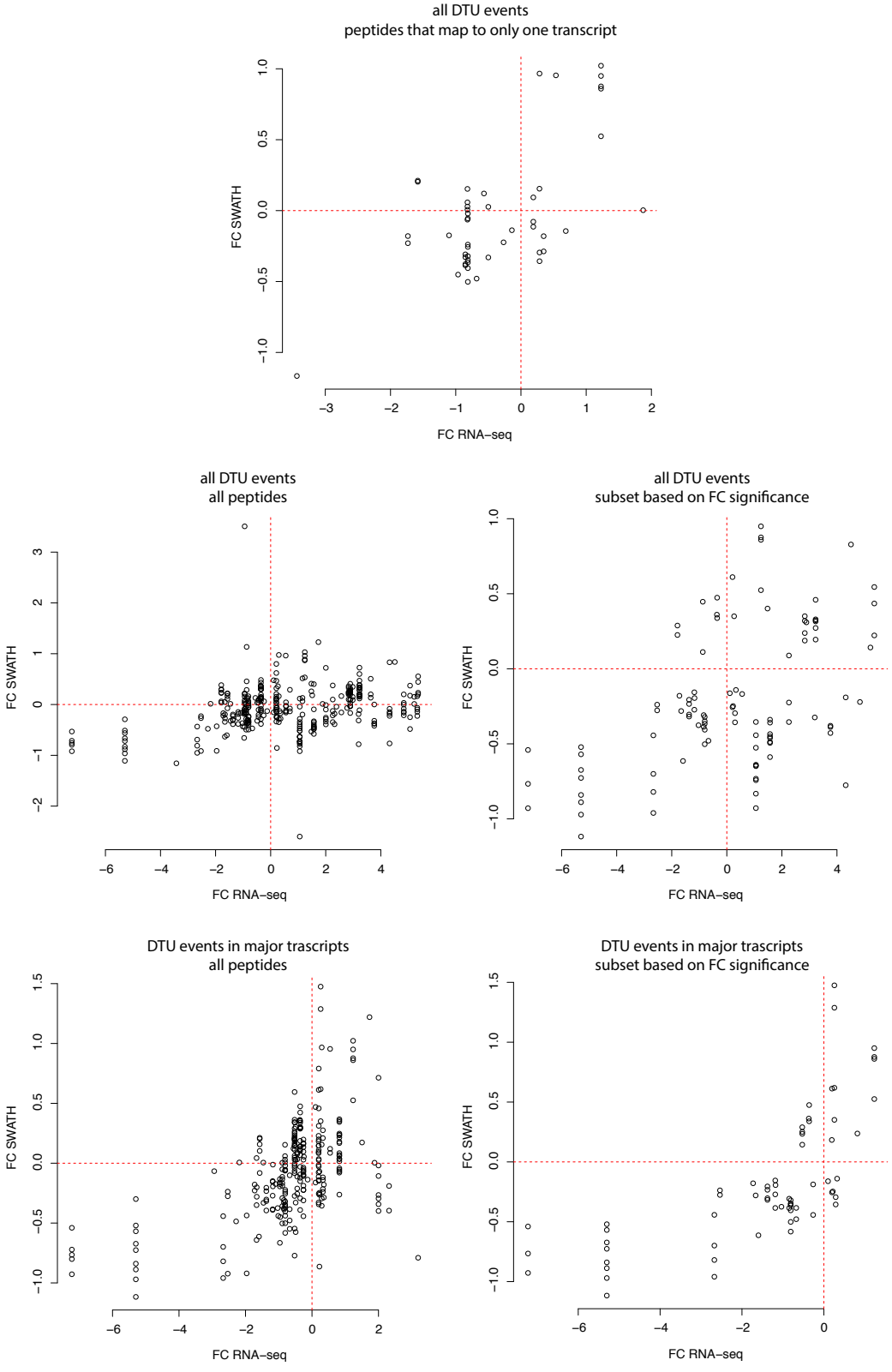


Figure E.1| RNA-seq and SWATH-MS fold-change estimates for the differentially used transcripts with peptide evidence.

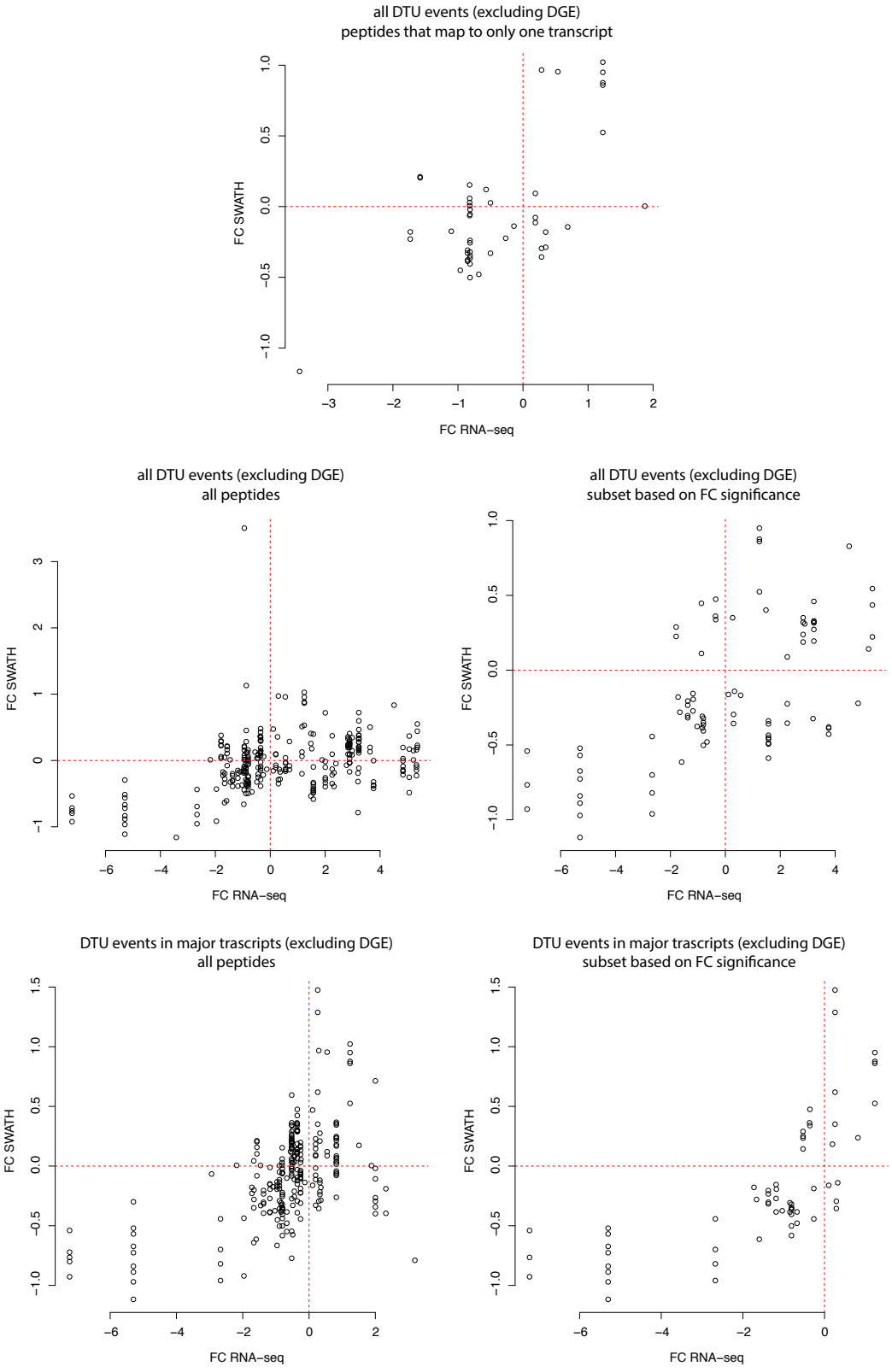


Figure E.2| RNA-seq and SWATH-MS fold-change estimates for the differentially used transcripts with peptide evidence after excluding differentially expressed genes.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P. 2002. *Molecular Biology of the Cell*. Garland, 4th edition. 2, 5, 6, 7

Alexander RD, Innocente SA, Barrass JD, and Beggs JD. 2010. Splicing-dependent RNA polymerase pausing in yeast. *Molecular Cell* **40**: 582–93. 112

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Galit L, Burstein D, Schwartz S, Postolsky B, et al.. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports* **1**: 543–556. 96, 111, 114

Anders S and Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**: R106. 30, 115

Anders S, Pyl PT, and Huber W. 2014. HTSeq: A python framework to work with high-throughput sequencing data. *bioRxiv*. 10.1101/002824. 21, 22, 89

Anders S, Reyes A, and Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Research* **22**: 2008–2017. 21, 33, 58, 78, 89, 113

Andrews S. 2010. Fastqc: A quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. 19

Ardehali M and Lis J. 2009. Tracking rates of transcription and splicing in vivo. *Nature Structural & Molecular Biology* **16**: 1123–1124. 110

Ast G. 2004. How did alternative splicing evolve? *Nature Reviews Genetics* **5**: 773–782. 13

Audenet F, Yates D, Géraldine C, Cussenot O, and Rouprêt M. 2012. Genetic pathways involved in carcinogenesis of clear cell renal cell carcinoma: genomics towards personalized medicine. *BJU International* **109**: 1864–1870. 87

Auman J and Howard M. 2010. Colorectal cancer cell lines lack the molecular heterogeneity of clinical colorectal tumors. *Clinical Colorectal Cancer* **9**: 40–47. 82, 87

Barash Y, Calarco J, Gao W, Pan Q, Wang X, Shai O, Blencowe B, and Frey B. 2010. Deciphering the splicing code. *Nature* **465**: 53–59. 93, 135

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al.. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)* **338**: 1587–1593. 55

Beaumont MA and Rannala B. 2004. The Bayesian revolution in genetics. *Nature Reviews Genetics* **5**: 251–61. 23

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* **57**: 289–300. 59, 89, 90, 114

Benjamini Y and Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**: e72. 17

Benjamini Y and Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**: 1165–1188. 89, 115

Bentley DL. 2014. Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* **15**: 163–75. 14, 111

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al.. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–9. 17

Berget S, Moore C, and Sharp P. 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 3171–3175. 6

Blakeley P, Siepen J, Lawless C, and Hubbard S. 2010. Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* **10**: 1127–1140. 119, 129

Briese M, Rot G, Wang Z, König J, Sim AY, Levitt M, Gorup C, Zupan B, Saieva L, Pellizzoni L, et al.., Genome-wide analysis of spliceosomal protein-rna interactions and their atp-dependent remodeling. In preparation. 97, 99, 112

Brown S, Stoilov P, and Xing Y. 2012. Chromatin and epigenetic regulation of pre-mRNA processing. *Human Molecular Genetics* **21**: R90–R96. 13

Brugarolas J. 2013. PBRM1 and BAP1 as novel targets for renal cell carcinoma. *Cancer Journal* **19**: 324–332. 65, 66

Buckley P, Lee M, Sul J, Miyashiro K, Bell T, Fisher S, Kim J, and Eberwine J. 2011. Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. *Neuron* **69**: 877–884. 54

Buljan M, Chalancon G, Eustermann S, Wagner G, Fuxreiter M, Bateman A, and Babu M. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular Cell* **46**: 871–883. 11

Carrillo Oesterreich F, Preibisch S, and Neugebauer K. 2010. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Molecular Cell* **40**: 571–581. 112

Chow L, Gelinas R, Broker T, and Roberts R. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* **12**: 1–8. 6

Chow WH, Dong LM, and Devesa SS. 2010. Epidemiology and risk factors for kidney cancer. *Nature Reviews Urology* **7**: 245–57. 65

Cock P, Fields C, Goto N, Heuer M, and Rice P. 2010. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**: 1767–1771. 17

Costa V, Aprile M, Esposito R, and Ciccodicola A. 2013. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *European Journal of Human Genetics* **21**: 134–42. 134

Cox TR and Erler JT. 2011. Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. *Disease models & mechanisms* **4**: 165–178. 86

Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561–563. 1

Cusnir M and Cavalcante L. 2012. Inter-tumor heterogeneity. *Human Vaccines & Immunotherapeutics* **8**: 1143–1145. 86

Darnell J. 2013. Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. *RNA* **19**: 443–460. 4

Das S and Vikalo H. 2013. Base calling for high-throughput short-read sequencing: dynamic programming solutions. *BMC Bioinformatics* **14**: 129. 17

De Conti L, Baralle M, and Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews* **4**: 49–60. 12

Dees N, Zhang Q, Kandoth C, Wendl M, Schierding W, Koboldt D, Mooney T, Callaway M, Dooling D, Mardis E, et al.. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome Research* **22**: 1589–1598. 91

van Dijk E, Jaszczyszyn Y, and Thermes C. 2014. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* **322**: 12–20. 15

Djebali S, Davis C, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al.. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108. 39, 53, 134

Domon B and Aebersold R. 2010. Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology* **28**: 710–721. 129

Ellis J, Miriam B, Colak R, Irimia M, Kim T, Calarco J, Wang X, Pan Q, Dave O, Kim P, et al.. 2012. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell* **46**: 884–892. 11

Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin B, and Albrecht M. 2010. AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Research* **38**: W755–W762. 11

ENCODE Project Consortium, Bernstein B, Birney E, Dunham I, Green E, Gunter C, and Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. 1, 56, 133

Engstrom P, Steijger T, Sipos B, Grant G, Kahles A, Consortium R, Alioto T, Behr J, Bertone P, Bohnert R, et al.. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* **10**: 1185–1191. 53

Ertel A, Verghese A, Byers S, Ochs M, and Tozeren A. 2006. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Molecular Cancer* **5**. 82, 87

Ezkurdia I, del Pozo A, Frankish A, Rodriguez J, Harrow J, Ashman K, Valencia A, and Tress M. 2012. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Molecular Biology and Evolution* **29**: 2265–2283. 119, 123, 129, 135

Fica S, Tuttle N, Novak T, Li N, Lu J, Koodathingal P, Dai Q, Staley J, and Piccirilli J. 2013. RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**: 229–234. 7

Flicek P, Amode M, Barrell D, Beal K, Brent S, Denise C, Clapham P, Coates G, Fairley S, Fitzgerald S, et al.. 2012. Ensembl 2012. *Nucleic Acids Research* **40**: D84–D90. 39, 57, 75, 89, 113, 131

Floris M, Raimondo D, Leoni G, Orsini M, Marcatili P, and Tramontano A. 2011. MAISTAS: a tool for automatic structural evaluation of alternative splicing products. *Bioinformatics* **27**: 1625–1629. 11, 75

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, et al.. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**: 161. 117, 119

Fuda N, Ardehali M, and Lis J. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**: 186–192. 3

Galej W, Oubridge C, Newman A, and Nagai K. 2013. Crystal structure of prp8 reveals active site cavity of the spliceosome. *Nature* **493**: 638–643. 95, 110

Gerlinger M, Rowan A, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al.. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine* **366**: 883–892. 86

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, and Aebersold R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics* **11**: O111.016717. 129

Gloria L, Covey P, Bedinger P, Mueller L, Thannhauser T, Zhang S, Fei Z, Giovannoni J, and Rose J. 2012. Enabling proteomic studies with RNA-Seq: the proteome of tomato pollen as a test case. *Proteomics* **12**: 761–774. 119

Goldman N, Bertone P, Chen S, Dessimoz C, Emily L, Sipos B, and Birney E. 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**: 77–80. 1

Gonzàlez-Porta M, Calvo M, Sammeth M, and Guigó R. 2012. Estimation of alternative splicing variability in human populations. *Genome Research* **22**: 528–538. 89

Gonzàlez-Porta M and Brazma A. 2014. Identification, annotation and visualisation of extreme changes in splicing from rna-seq experiments with switchseq. *bioRxiv*. 10.1101/005967. 78, 131

Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al.. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652. 21

Grainger R and Beggs J. 2005. Prp8 protein: at the heart of the spliceosome. *RNA* **11**: 533–557. 95, 110

Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, and Sammeth M. 2012. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research* **40**: 10073–10083. 40, 57

Grosso A, Gomes A, Nuno B, Caldeira S, Thorne N, Grech G, von Lindern M, and Maria C. 2008. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Research* **36**: 4823–4832. 135

Hagen RM and Ladomery MR. 2012. Role of splice variants in the metastatic progression of prostate cancer. *Biochemical Society Transactions* **40**: 870–4. 63, 65

Haider S and Pal R. 2013. Integrated analysis of transcriptomic and proteomic data. *Current Genomics* **14**: 91–110. 119

Hansen KD, Brenner SE, and Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**: e131. 17

Hardcastle TJ and Kelly KA. 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**: 422. 33

Harrow J, Frankish A, Gonzalez J, Tapanari E, Diekhans M, Kokocinski F, Aken B, Barrell D, Zadissa A, Searle S, et al.. 2012. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research* **22**: 1760–1774. 11, 57, 133

Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, and Teichmann SA. 2011. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology* **7**: 497. 41, 58

Heinz S, Benner C, Spann N, Bertolino E, Lin Y, Laslo P, Cheng J, Murre C, Singh H, and Glass C. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell* **38**: 576–589. 114

Hemberg M, Prabakaran S, Chauhan R, Winter D, Tweedie-Cullen R, Dittrich C, Hong E, Gunawardena J, Steen H, Steen J, et al.., Peptides from rnas classified as non-coding. Poster at the "Systems Biology: Global regulation of gene expression meeting", CSHL. 53

Hofmann J, Husedzinovic A, and Gruss O. 2010. The function of spliceosome components in open mitosis. *Nucleus* **1**: 447–459. 95

Hooper JE. 2014. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Human Genomics* **8**: 3. 136

Hoskins A and Moore M. 2012. The spliceosome: a flexible, reversible macromolecular machine. *Trends in Biochemical Sciences* **37**: 179–188. 95

House A and Lynch K. 2008. Regulation of alternative splicing: more than just the ABCs. *The Journal of Biological Chemistry* **283**: 1217–1221. 39, 111

Huang DW, Sherman B, and Lempicki R. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**: 1–13. 59, 89, 115

Huang DW, Sherman B, and Lempicki R. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**: 44–57. 59, 89, 115

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921. 1

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945. 1

Jonasch E, Futreal P, Davis I, Bailey S, Kim W, Brugarolas J, Giaccia A, Kurban G, Pause A, Frydman J, et al.. 2012. State of the science: an update on renal cell carcinoma. *Molecular Cancer Research* **10**: 859–880. 65, 84

Kaida D, Schneider-Poetsch T, and Yoshida M. 2012. Splicing in oncogenesis and tumor suppression. *Cancer Science* **103**: 1611–6. 63, 65

Kalsotra A and Cooper TA. 2011. Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics* **12**: 715–29. 11

Kanehisa M and Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**: 27–30. 91

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, and Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**: D199–D205. 91

Katz Y, Wang E, Airoldi E, and Burge C. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7**: 1009–1015. 23, 40, 58, 90, 115

Keren H, Galit L, and Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* **11**: 345–355. 10, 11, 54, 65

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**: R36. 18, 19, 20, 26, 89

Kornblihtt A, Schor I, Alló M, Dujardin G, Petrillo E, and Muñoz M. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology* **14**: 153–165. 10

Kuehner J, Pearson E, and Moore C. 2011. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews Molecular Cell Biology* **12**: 283–294. 4

Kurmangaliyev YZ and Gelfand MS. 2008. Computational analysis of splicing errors and mutations in human transcripts. *BMC Genomics* **9**: 13. 54

Kwak H and Lis JT. 2013. Control of transcriptional elongation. *Annual Review of Genetics* **47**: 483–508. 4

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner D, Luscombe N, and Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology* **17**: 909–915. 97, 112

Ladomery M. 2013. Aberrant alternative splicing is another hallmark of cancer. *International Journal of Cell Biology* **2013**: 463786. 14

Lane KR, Yu Y, Lackey PE, Chen X, Marzluff WF, and Cook JG. 2013. Cell cycle-regulated protein abundance changes in synchronously proliferating HeLa cells include regulation of pre-mRNA splicing proteins. *PloS One* **8**: e58456. 135

Langmead B and Salzberg S. 2012. Fast gapped-read alignment with bowtie 2. *Nature Methods* **9**: 357–359. 19

Langmead B, Trapnell C, Pop M, and Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25. 57, 131

Lasda E and Blumenthal T. 2011. Trans-splicing. *Wiley Interdisciplinary Reviews* **2**: 417–434. 12

Lee C and Roy M. 2004. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biology* **5**: 231. 136

Leoni G, Le Pera L, Ferrè F, Raimondo D, and Tramontano A. 2011. Coding potential of the products of alternative splicing in human. *Genome Biology* **12**. 119, 129, 135

Levental KR, Yu H, Kass L, Lakins JN, Egeblad M, Erler JT, Fong SF, Csiszar K, Giaccia A, Weninger W, et al.. 2009. Matrix crosslinking forces tumor progression by enhancing integrin signaling. *Cell* **139**: 891–906. 86

Levin J, Yassour M, Adiconis X, Nusbaum C, Thompson D, Friedman N, Gnirke A, and Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7**: 709–715. 18

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and Subgroup GPDP. 2009. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. 20

Li J, Jiang H, and Wong WH. 2010. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology* **11**: R50. 29

Li JJ, Bickel PJ, and Biggin MD. 2014. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**: e270. 54, 119, 129, 130

Li X, Zhang W, Xu T, Ramsey J, Zhang L, Hill R, Hansen KC, Hesselberth JR, and Zhao R. 2013. Comprehensive in vivo RNA-binding site analyses reveal a role of Prp8 in spliceosomal assembly. *Nucleic Acids Research* **41**: 3805–18. 95, 98, 111

Li Y, Bor Y, Misawa Y, Xue Y, Rekosh D, and Hammarskjold M. 2006. An intron with a constitutive transport element is retained in a tap messenger RNA. *Nature* **443**: 234–237. 54

Linehan W, Srinivasan R, and Schmidt L. 2010. The genetic basis of kidney cancer: a metabolic disease. *Nature Reviews Urology* **7**: 277–285. 66, 86

Liu Y, Huttenhain R, Surinova S, Gillet L, Mouritsen J, Brunner R, Navarro P, and Aebersold R. 2013. Quantitative measurements of n-linked glycoproteins in human plasma by SWATH-MS. *Proteomics* **13**: 1247–1256. 120, 129, 132

Love MI, Huber W, and Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv.* 10.1101/002832. 21, 28, 30

Lu P, Weaver VM, and Werb Z. 2012. The extracellular matrix: a dynamic niche in cancer progression. *The Journal of cell biology* **196**: 395–406. 86

Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, and Brazma A. 2010. A global map of human gene expression. *Nature Biotechnology* **28**: 322–324. 39, 88

Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenas C, Lundeberg J, Mann M, and Uhlen M. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology* **6**: 450. 54, 119, 129

Maas S. 2012. Posttranscriptional recoding by RNA editing. *Advances in protein chemistry and structural biology* **86**: 193–224. 12

Maier T, Guell M, and Serrano L. 2009. Correlation of mRNA and protein in complex biological samples. *FEBS letters* **583**: 3966–3973. 119, 132

Malone JH and Oliver B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology* **9**: 34. 14

Mardis ER. 2013. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* **6**: 287–303. 14, 16, 18

Marinov G, Williams B, Ken M, Schroth G, Gertz J, Myers R, and Wold B. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research* **24**: 496–510. 54

Marioni J, Mason C, Mane S, Stephens M, and Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**: 1509–1517. 30

Martin J and Wang Z. 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics* **12**: 671–682. 21

Martinez P, Birkbak N, Gerlinger M, Nicholas M, Burrell R, Rowan A, Joshi T, Fisher R, Larkin J, Szallasi Z, et al.. 2013. Parallel evolution of tumour subclones mimics diversity between tumours. *The Journal of Pathology* **230**: 356–364. 86

Matera A and Wang Z. 2014. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology* **15**: 108–121. 7, 8, 12, 13, 95, 135

McGlincy N and Smith C. 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences* **33**: 385–393. 11, 54, 65, 87

Melamud E and Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Research* **37**: 4873–4886. 11, 55

Merkin J, Russell C, Chen P, and Burge C. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**: 1593–1599. 11, 55

Metzker M. 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**: 31–46. 17

Michal C, Regina G, Siegfried Z, Andersen C, Thorsen K, Ørntoft T, Mu D, and Karni R. 2013. The splicing factor SRSF6 is amplified and is an oncoprotein in lung and colon cancers. *The Journal of Pathology* **229**: 630–639. 86

Minoche AE, Dohm JC, and Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology* **12**: R112. 19

Mortazavi A, Williams B, Kenneth M, Schaeffer L, and Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–628. 14, 22, 27

Nagaraj N, Wisniewski J, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, and Mann M. 2011. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology* **7**: 548. 54, 119, 130

Neumann B, Walter T, Hériché J, Bulkescher J, Erfle H, Conrad C, Rogers P, Poser I, Held M, Liebel U, et al.. 2010. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**: 721–727. 95, 110

Nilsen T. 2003. The spliceosome: the most complex macromolecular machine in the cell? *BioEssays* **25**: 1147–1149. 6

Nilsen T and Graveley B. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463. 10, 54, 55, 56, 65

Ning K and Nesvizhskii AI. 2010. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics* **11 Suppl 11**: S14. 119

Padgett R. 2012. New connections between splicing and human disease. *Trends in Genetics* **28**: 147–154. 14, 63, 65

Pan Q, Shai O, Lee L, Frey B, and Blencowe B. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**: 1413–1415. 10, 37, 39

Parenteau J, Durand M, Morin G, Gagnon J, Lucier J, Wellinger R, Chabot B, and Elela S. 2011. Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147**: 320–331. 54

Park JW, Parisky K, Celotto AM, Reenan RA, and Graveley BR. 2004. Identification of alternative splicing regulators by RNA interference in Drosophila. *Proceedings of the National Academy of Sciences* **101**: 15974–9. 135

Paule MR and White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Research* **28**: 1283–98. 3

Pedrotti S and Cooper TA. 2014. In Brief: (mis)splicing in disease. *The Journal of Pathology* **233**: 1–3. 63, 65

Pérez-Ortín JE, Alepuz P, Chávez S, and Choder M. 2013. Eukaryotic mRNA decay: methodologies, pathways, and links to other stages of gene expression. *Journal of Molecular Biology* **425**: 3750–75. 5, 54

Plass M and Eyras E. 2014. Approaches to link RNA secondary structures with splicing regulation. *Methods in Molecular Biology* **1126**: 341–356. 13

Porrua O and Libri D. 2013. RNA quality control in the nucleus: the angels' share of RNA. *Biochimica et Biophysica Acta* **1829**: 604–611. 5, 54

Query CC and Konarska MM. 2004. Suppression of multiple substrate mutations by spliceosomal prp8 alleles suggests functional correlations with ribosomal ambiguity mutants. *Molecular Cell* **14**: 343–54. 112

Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–2. 113

Ramakrishnan SR, Vogel C, Prince JT, Li Z, Penalva LO, Myers M, Marcotte EM, Miranker DP, and Wang R. 2009. Integrating shotgun proteomics and mRNA

expression data to improve protein identification. *Bioinformatics* **25**: 1397–403. 55, 130

Ramskold D, Wang ET, Burge CB, and Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology* **5**: e1000598. 58

Roberts A, Trapnell C, Donaghey J, Rinn JL, and Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology* **12**: R22. 29

Robertson MP and Joyce GF. 2012. The origins of the RNA world. *Cold Spring Harbor Perspectives in Biology* **4**. pii: a003608. 7

Robinson M, Davis M, and Smyth G. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140. 30, 33, 89

Robinson MD and Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**: R25. 28

Rodriguez J, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J, Lopez G, Valencia A, and Tress M. 2013. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research* **41**: D110–D117. 11, 40, 41, 75, 87

Roux J and Robinson-Rechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome research* **21**: 357–363. 55

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, et al.. 2010. GeneCards Version 3: the human gene integrator. *Database* **2010**: baq020. 89, 132

Sakabe NJ and de Souza SJ. 2007. Sequence features responsible for intron retention in human. *BMC Genomics* **8**: 59. 96, 111

Saltzman A, Pan Q, and Blencowe B. 2011. Regulation of alternative splicing by the core spliceosomal machinery. *Genes & Development* **25**: 373–384. 14, 86, 111, 112

Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Aiko S, Nagae G, Suzuki H, et al.. 2013. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature Genetics* **45**: 860–867. 65, 78, 79, 84, 85, 86

Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzàlez-Porta M, et al.., Whole-genome sequencing reveals variation in the genomic landscape of clear cell renal cell carcinoma in europe. Under review in Nature Communications. 65, 84, 85

Schmieder R and Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–4. 89

Schoenberg D and Maquat L. 2012. Regulation of cytoplasmic mRNA decay. *Nature Reviews Genetics* **13**: 246–259. 6

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, and Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337–342. 2

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al.. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**: 236–40. 54

Sharon D, Tilgner H, Grubert F, and Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology* **31**: 1009–1014. 136

Shendure J and Ji H. 2008. Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–45. 136

Sheynkman G, Shortreed M, Frey B, and Smith L. 2013. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular & Cellular Proteomics* **12**: 2341–2353. 119, 129, 135

Shibata M and Shen M. 2013. The roots of cancer: stem cells and the basis for tumor heterogeneity. *BioEssays* **35**: 253–260. 86

Shionyu M, Takahashi Ki, and Go M. 2012. AS-EAST: a functional annotation tool for putative proteins encoded by alternatively spliced transcripts. *Bioinformatics* **28**: 2076–2077. 11

Shiroguchi K, Jia T, Sims P, and Xie X. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 1347–1352. 18

Singh J and Padgett R. 2009. Rates of in situ transcription and splicing in large human genes. *Nature Structural & Molecular Biology* **16**: 1128–1133. 110

Smith D, Query C, and Konarska M. 2008. "Nought may endure but mutability": spliceosome dynamics and the regulation of splicing. *Molecular Cell* **30**: 657–666. 39

de Sousa Abreu R, Penalva L, Marcotte E, and Vogel C. 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems* **5**: 1512–1526. 119

van Staveren WCG, Solís DYW, Hébrant A, Detours V, Dumont JE, and Maenhaut C. 2009. Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells? *Biochimica et Biophysica Acta* **1795**: 92–103. 82, 87

Steijger T, Abril J, Engstrom P, Kokocinski F, Consortium R, Abril J, Akerman M, Alioto T, Ambrosini G, Antonarakis S, et al.. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**: 1177–1184. 53, 58

Tanackovic G, Ransijn A, Thibault P, Abou Elela S, Klinck R, Berson E, Chabot B, and Rivolta C. 2011. PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. *Human Molecular Genetics* **20**: 2116–2130. 95, 135

Taneri B, Snyder B, and Gaasterland T. 2011. Distribution of alternatively spliced transcript isoforms within human and mouse transcriptomes. *J OMICS Res* **1**: 1–5. 39

Tang J, Lee J, Hou M, Wang C, Chen C, Huang H, and Chang H. 2013. Alternative splicing for diseases, cancers, drugs, and databases. *The Scientific World Journal* **2013**: 703568. 88

Tanner S, Shen Z, Ng J, Florea L, Guigó R, Briggs S, and Bafna V. 2007. Improving gene annotation using peptide mass spectrometry. *Genome Research* **17**: 231–239. 119

Tazi J, Bakkour N, and Stamm S. 2009. Alternative splicing and disease. *Biochimica et Biophysica Acta* **1792**: 14–26. 14, 65

TCGA Network. 2013. Comprehensive molecular characterisation of clear cell renal cell carcinoma. *Nature* **499**: 43–49. 65, 78, 79, 84, 85, 86

Tilgner H, Knowles D, Johnson R, Davis C, Chakrabortty S, Djebali S, Curado Ja, Snyder M, Gingeras T, and Guigó R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research* **22**: 1616–1625. 13, 106, 111

Trapnell C, Hendrickson D, Sauvageau M, Goff L, Rinn J, and Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**: 46–53. 33

Trapnell C, Pachter L, and Salzberg S. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111. 57, 113

Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M, Salzberg S, Wold B, and Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511–515. 11, 18, 22, 23, 24, 26, 29, 33, 40, 58, 59

Tress M, Wesselink J, Frankish A, López G, Goldman N, Loytynoja A, Massingham T, Pardi F, Whelan S, Harrow J, et al.. 2008a. Determination and validation of principal gene products. *Bioinformatics* **24**: 11–17. 56

Tress ML, Bodenmiller B, Aebersold R, and Valencia A. 2008b. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology* **9**: R162. 10

Turro E, Astle WJ, and Tavaré S. 2014. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* **30**: 180–8. 33, 34, 78, 131

Turro E, Su SY, Gonçalves A, Coin LJM, Richardson S, and Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology* **12**: R13. 18, 22, 23, 24, 29, 33, 40, 58, 131

Turunen J, Niemea E, Verma B, and Frilander M. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdisciplinary Reviews* **4**: 61–76. 7

UniProt Consortium. 2012. Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Research* **40**: D71–D75. 46, 56

Valletti A, Gigante M, Palumbo O, Carella M, Divella C, Sbisà E, Tullo A, Picardi E, D'Erchia AM, Battaglia M, et al.. 2013. Genome-wide analysis of differentially expressed genes and splicing isoforms in clear cell renal cell carcinoma. *PloS One* **8**: e78452. 66, 84

Vaquerizas J, Kummerfeld S, Teichmann S, and Luscombe N. 2009. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10**: 252–263. 4

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al.. 2001. The sequence of the human genome. *Science* **291**: 1304–1351. 1

Villalba A, Coll O, and Gebauer F. 2011. Cytoplasmic polyadenylation and translational control. *Current Opinion in Genetics & Development* **21**: 452–457. 6

Vogel C and Marcotte E. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**: 227–232. 2, 41, 54, 58

Volkmann G and Mootz H. 2013. Recent progress in intein research: from mechanism to directed evolution and applications. *Cellular and molecular life sciences* **70**: 1185–1206. 12

Waks Z, Klein AM, and Silver PA. 2011. Cell-to-cell variability of alternative RNA splicing. *Molecular Systems Biology* **7**: 506. 39

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, and Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476. 10, 11, 37, 39

Wang J, Duncan D, Shi Z, and Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research* **41**: W77–83. 59, 89, 115

Wang X, Wu Z, and Zhang X. 2010. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *Journal of Bioinformatics and Computational Biology* **8 Suppl 1**: 177–192. 25

Wang Z and Burge C. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**: 802–813. 111

Wang Z, Gerstein M, and Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57–63. 2, 14

Wickramasinghe V, Gonzàlez-Porta M, et al.., Prpf8 abundance dictates the patterns of alternative and constitutive messenger rna splicing. Submitted. 112

Will CL and Luhrmann R. 2011. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology* **3**. pii: a003707. 7, 10

Wong J, Ritchie W, Ebner O, Selbach M, Wong J, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al.. 2013a. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**: 583–595. 54

Wong KH, Jin Y, and Moqtaderi Z. 2013b. Multiplex Illumina sequencing using DNA barcoding. *Current Protocols in Molecular Biology / edited by Frederick M. Ausubel et al.* **Chapter 7**: Unit 7.11. 18

Wu P, Phan J, and Wang M. 2013. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics* **14 Suppl 11**. 57

Xing X, Li Q, Sun H, Fu X, Zhan F, Huang X, Li J, Chen C, Shyr Y, Zeng R, et al.. 2011. The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. *Genomics* **98**: 343–351. 119

Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al.. 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology* **20**: 1131–1139. 54

Yap K, Lim Z, Khandelia P, Friedman B, and Makeyev E. 2012. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes & Development* **26**: 1209–1223. 11, 54, 65, 87

Yeo G and Burge C. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* **11**: 377–394. 114

Yu Y, Maroney P, Denker J, Zhang X, Dybkov O, Luhrmann R, Jankowsky E, Chasin L, and Nilsen T. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. *Cell* **135**: 1224–1236. 112

Zhao Q, Caballero O, Davis I, Jonasch E, Tamboli P, Yung W, Weinstein J, for TCGA research network KS, Strausberg R, and Yao J. 2013. Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas. *Clinical Cancer Research* **19**: 2460–2472. 66, 75, 84

Zhao S, Fung-Leung W, Bittner A, Ngo K, Liu X, and Zhang S. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**: e78644. 136

Zhou A, Zhang F, and Chen JY. 2010. PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. *BMC Bioinformatics* **11 Suppl 6**: S7. 119

Zhuang F, Fuchs RT, and Robb GB. 2012. Small RNA expression profiling by high-throughput sequencing: implications of enzymatic manipulation. *Journal of Nucleic Acids* **2012**: 360358. 16