A Systems Biology Study of Alternative Splicing Regulations and Functions

by

Seyed Alborz Mazloomian

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies
(Bioinformatics)

The University of British Columbia
(Vancouver)

June 2017

Abstract

Alternative splicing is highly appreciated as a major contributor to cellular complexity, and its dysregulation has been associated to several diseases. Despite being the focus of numerous studies in recent years, there remains much unknown about functions and regulations of alternative splicing in mammalian systems. Here, I take a systems biology approach to study alternative splicing using high-throughput sequencing data.

In Chapter 2, I use tissue-specific high-throughput libraries of *Drosophila melanogaster* to explore the potential inter-relation of RNA editing and alternative splicing. I first develop a pipeline to accurately detect editing events. Next, I find regions where editing and splicing are likely to influence each other, and report conserved RNA structures that can mediate the inter-relation.

In Chapter 3, I study functions of Cyclin dependent kinase 12 (*CDK12*) using human cell line data. I show that *CDK12* influences the differential usage of alternative last exon. Additionally, the results demonstrate that *CDK12* modulates the expression of DNA damage response genes, and increases the tumorigenicity of breast cancer cells by down-regulating the long isoform of *DNAJB6* gene.

Finally, in chapter 4, I first present a review of methods that search for underlying mechanisms explaining variations between high-throughput measurements of two biological conditions. Next, I introduce our RNA-seq data derived from progressively inhibiting splicing-related proteins at multiple concentrations of pharmaceuticals, and I discuss how the reviewed methods should be adopted to benefit most from our type of data.

Our systems biology research provides new insights on how the studied components of the splicing machinery contribute to splicing functions and regulations, and these findings can help to improve our understanding of related diseases.

Lay Summary

Genes contain information that determine what cells should do at specific times, and they are sometimes referred to as the blueprint for life. Alternative splicing is a mechanism through which multiple products are generated from a single gene, and these products (e.g. proteins) can have different functions; therefore, the mechanism expands the capacity of genes. Disruption in alternative splicing has been associated to many genetic diseases. However, the mechanism is not fully understood. Fortunately, recent advances in technology have brought new opportunities to better investigate this mechanism. In this thesis, I study how the alternative splicing mechanism and its functions are regulated by some genes and cellular machineries using data generated by new sequencing technologies. The selected genes and mechanisms are known to play important roles in human diseases such as cancers. Our findings can help to improve our understanding of the alternative splicing mechanism and related diseases.

Preface

A version of Chapter 2 has been published in *RNA biology* journal [1]. I performed the analysis and wrote the manuscript under the supervision of Professor Irmtraud Meyer.

A version of Chapter 3 has been accepted for publication in *Nucleic Acids Research* journal [2]. I am a co-first author of the paper with Jerry F. Tien. This Chapter was supervised by Professor Shah and Professor Morin. I performed the computational analysis of the data, made figures and wrote the manuscript. Jerry F. Tien designed and performed experiments, analyzed data, made figures and wrote the manuscript. S.-W. Grace Cheng designed and performed experiments. Ali Bashashati and James Xu performed the data analysis. Christopher S. Hughes, Christalle C.T. Chow, Leanna T. Canapi, Arusha Oloumi, Genny Trigo-Gonzalez, Vicky C.-D. Chang, Stella S. Chun performed experiments. Professor Samuel Aparicio assisted in research design and data interpretation. Professor Gregg Morin conceived the project, designed experiments, and wrote the manuscript. Professor Shah supervised the computational analysis, assisted in research design and data interpretation.

Chapter 4 has not been submitted for publication yet. The project was supervised by Professor Sohrab Shah and Professor Samuel Aparicio. I performed the analysis, made figures and wrote the chapter. All the small compounds used for inhibiting the studied proteins affecting alternative splicing have been developed by Takeda Pharmaceutical Company Limited.

[1] Mazloomian, A. & Meyer, I.M., 2015. Genome-wide identification and characterization of tissue-specific RNA editing events in *D. melanogaster* and their potential role in regulating alternative splicing. *RNA biology* **12** (12): 1391-1401.

[2] Tien, J.F. *, Mazloomian, A. *, Cheng, S., Hughes, C.S., Chow, C., Canapi, L.T., Oloumi, A., Trigo-Gonzalez, G., Bashashati, A., Xu, J. et al., 2017. CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. to appear in *Nucleic acids research*. (*: co-first authors).

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my deep gratitude to my supervisors Professor Irmtraud Meyer and Professor Sohrab Shah for their guidance and encouragement. I would like to thank my supervisory committee, Professor Wyeth Wasserman, Professor Steven Jones, and Professor Samuel Aparicio for their help and feedback during my PhD studies. I am also very thankful for my collaborations with Professor Gregg Morin and Professor Samuel Aparicio.

I extend my sincere thanks to Meyer Lab members, Shah Lab members, Morin Lab members, and all my other friends for their discussions and for creating pleasing environments.

I would like to thank the University of British Columbia and Bioinformatics Training Program for their generous four-year fellowship funding.

Finally, my heartfelt thanks to my parents and my three brothers, for all the support, encouragement, and love. Specific to my PhD research, I would like to thank Amin for all the great helps and mentorship.

Dedicated to:

Baba & Maman

Chapter 1

Introduction

Complexity of molecular responses is created through the interplay between biological mechanisms. Hundreds and thousands of genes, RNA molecules and proteins communicate through signalling pathways to provide appropriate responses to environmental stimuli. Disruption in any of these cellular processes including transcription, translation, DNA repair, cell division and cell adhesion can initiate disease development. Consequently, investigating the inter-relations and interactions between these mechanisms is one key step towards deciphering their regulations and functions.

Alternative splicing, a mechanism thorough which multiple products are generated from a single gene, is highly appreciated as a major contributor to cellular complexity [3]. Many disease mutations have been associated to mis-regulation of alternative splicing [4, 5]. As a result, this process has become a critical topic for thorough research. Fortunately, recent advances in technology [6, 7] has brought new opportunities to better investigate alternative splicing and the related mechanisms such as transcription, RNA editing, and poly-adenylation.

In this thesis, I took a systems biology approach to study the inter-relation between alternative splicing and two other mechanisms, RNA editing, and phosphorylation of splicing related proteins by Cyclin Dependent Kinase 12 (*CDK12*). I also studied the global consequence of intervening with the splicing machinery. Based on the systems biology perspective, understanding properties of a system requires simultaneously modelling components of systems and integrating results of multiple types of experiments. The modelling

can benefit from the experiments in biological backgrounds when different conditions are screened, or it can benefit from intervening with the system and knocking down parts of the system to monitor its influence on the other parts.

I start by studying the inter-relation of RNA editing and alternative splicing in a model organism, *Drosophila melanogaster* in tissue specific data sets. Model organisms have been broadly studied for understanding biological machineries and development of therapeutics [8–10]. Fewer repeat regions, fewer overlapping genes, and smaller number of transcripts per gene on average in *Drosophila melanogaster* compared to human [11, 12] make the computational study of editing and splicing in *Drosophila melanogaster* easier and less error prone. Thus, *D. melanogaster* brings great opportunities to study the inter-relation of these two mechanisms in a context less complex than human, and its interpreted results provide a test bed to explore various hypotheses.

Furthermore, I study the influence of *CDK12* on the regulation of alternative splicing through inhibiting *CDK12* in human cell lines. I use data sets where *CDK12* expression was manipulated to quantify and model its influence on global RNA processing. Moreover, I show that how the proper gathering of information from multiple cell lines leads to understanding the main functions of a protein. Human cell lines present closer estimates to the rules governing cellular responses in human and are broadly being used to carefully investigate findings [13, 14].

Finally, I investigate how inhibiting components of the splicing machinery impacts cellular responses when the inhibition level is gradually increased. With the development of pharmacologic agents, there is the opportunity to systematically interfere with the spliceosome components to inhibit their functions in human. By progressively increasing drug levels and measuring responses and studying response curves, one can develop more accurate assumptions regarding the primary and secondary effects of disrupted components.

In parallel to my focus on the main topic of this thesis, how splicing relates to its components and other mechanisms, the thesis is designed to cover different steps which should be taken for understanding a mechanism and targeting it in relevant diseases.

1.1 Alternative splicing

In this section, I will briefly present the current knowledge of the splicing mechanism, functions and regulations of alternative splicing, and also the computational approaches developed to study alternative splicing.

1.1.1 Splicing mechanism

Splicing is a mechanism responsible for removing introns from a pre-mRNA molecule and merging exons together [15, 16]. The process is carried out by spliceosomes in the nucleus, where the splicesosome can cooperate and couple with other RNA processing machineries such as transcription [17]. A typical exon is on average about one order of magnitude shorter than an average-size intron in human [18] (few hundred and few thousand nucleotides in a typical exon and intron, respectively). Thus, exon recognition remains non-trivial for spliceosomes.

Spliceosomes employ the information in some regulatory conserved sequence motifs to accomplish RNA splicing [19]. These complex macromolecular machines identify intron boundaries with the help of the 5' and 3' conserved sequences [19]. More specifically, there exist a highly conserved GU di-nucleotide at the 5' end (splice donor site) and a conserved AG di-nucleotide at the 3' end (splice acceptor site) of introns. Some of the other conserved informative sequences in a primary sequence are the branch point located close to the acceptor site followed by a pyrimidine rich region [18–20]. Mutations in these conserved sequences can change open reading frames and result in degradation of transcripts, or producing incorrect amino acids and non-functional proteins.

Through detection of conserved sequence motifs by small nuclear ribonnucleoproteins (snRNPs) of the spliceosomal machinery, two transesterification steps are carried out [21]. In the first transesterification step (figure 1.1.A), RNA molecules in snRNPs interact and detect conserved motifs to trigger transesterification steps. Once the region is identified, the 2' hydroxyl of the branch point adenine nucleotide in the intron attacks the 5' splice site and cuts the sugar phosphate backbone of the pre-RNA molecule (figure 1.1.B). Subsequently, the end of the intron covalently bonds to the adenine nucleotide and forms a lariat structure. In the second transesterification step, the spliceosome goes to a conforma-

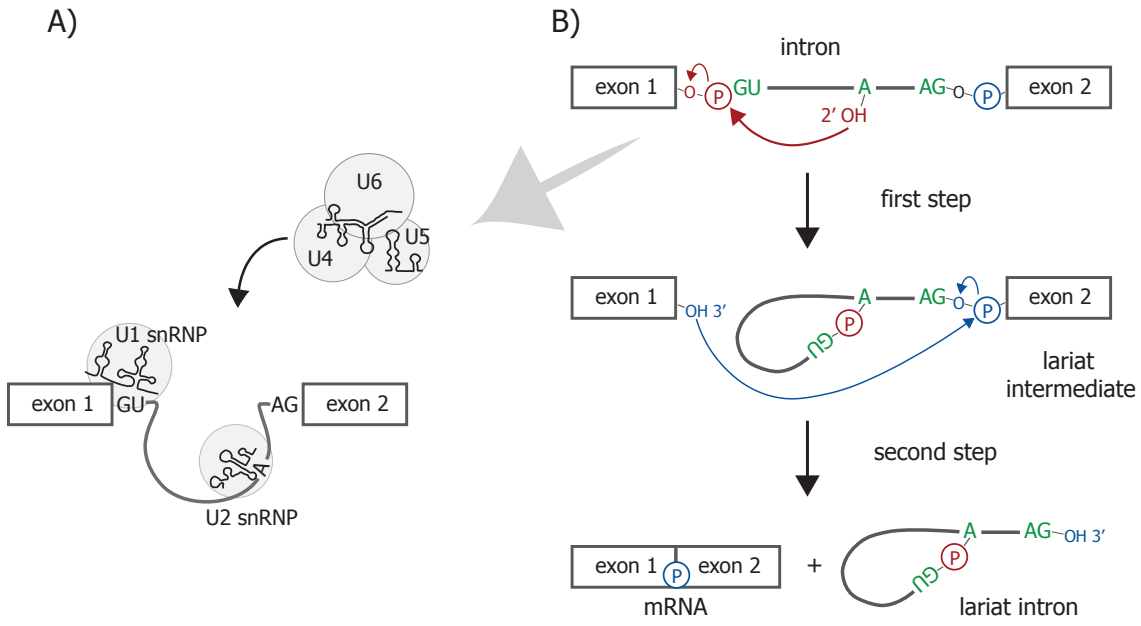

Figure 1.1: The two transesterification steps of the splicing mechanism. **(A)** Before the catalytic reactions of splicing, snRNA molecules in the spliceosome interact with the pre-mRNA molecule. These interactions followed by conformational changes in the splicing machinery initiate splicing. **(B)** The transesterification steps carried out by the spliceosome. The first step reactions are shown in red, and the second step reactions in blue. Conserved consensus motifs are shown in green, and the circled P's represent phosphates. Figure modified from [21] and [22].

tional rearrangement to bring the exons together, and guides the 3' hydroxyl group of the detached exon to react with the 5' end of the other exon. Finally, the two exons are merged into a continuous sequence and the lariat is released and degraded [21].

The spliceosome is composed of five snRNPs (U1, U2, U4, U5, and U6) and hundreds of other protein components [21, 23]. The two transesterification reactions required for the splicing mechanism cannot completely explain the necessity for such a complicated machinery. Some of the spliceosomal proteins are required to avoid making defective mRNAs, and some others link splicing to transcription, or other post-splicing events such as mRNA transport [24].

1.1.2 Functions and regulations of alternative splicing

Alternative splicing (AS), a process by which multiple transcripts are produced from a single pre-mRNA molecule is one major cause of cellular complexity [25]. Splicing patterns are determined by cell-types, developmental stages, or external stimulus [26]. Moreover, studying the alternatively spliced genes reveals that the process is most important where a differential processing is critical and a high level of diversity is required, especially in brain [27, 28]. Brain-specific AS events play crucial roles in neuronal differentiation and development, regulating protein-protein interactions, and regulating transcription networks [29, 30]. The accessibility and interactions of cis-regulatory sites with trans-acting proteins can modify splice site selection. As a result, the final splicing product is not always uniquely defined, and spliceosome decisions dictate the final conformation when alternative junction choices are available. Based on the consequences of such decisions, AS events are classified into multiple types, as illustrated in Figure 1.2.

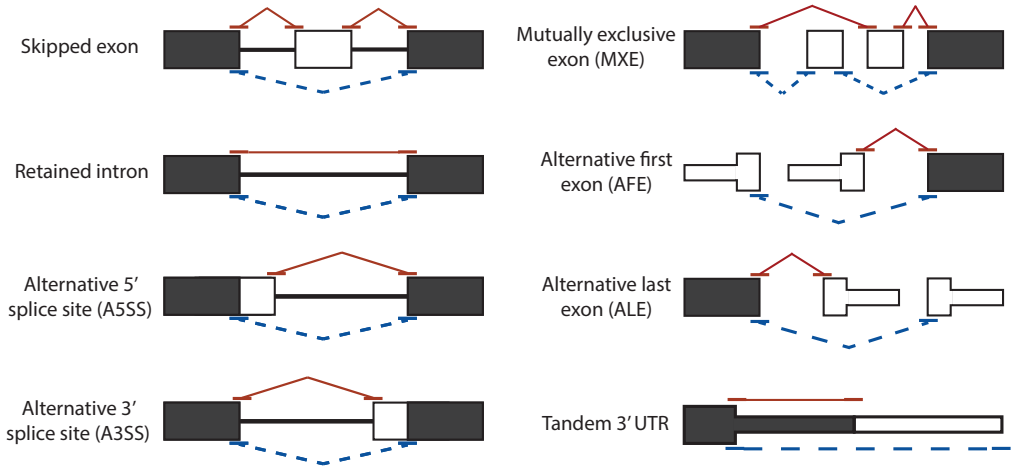

Figure 1.2: Types of alternative splicing defined by the alternative choices that spliceosome can make. The black boxes represent constitutive exons and the white boxes represent alternative regions whose inclusion depend on splicing choices. Only in the "Retained intron" type, an intron is contained in the final product. The red solid lines and the blue dashed lines illustrate the two possible patterns of splicing for each class. Figure modified from [31].

Specific features of RNA regions qualify them as candidates of each AS type. For example, skipped exons are usually shorter than constitutive exons and are flanked by long intronic regions [32]. Besides, the number of nucleotides in skipped exons is often multiples of three in order to prevent a change of reading frame and the introduction of a premature stop codon [32]. On the other hand, retained introns tend to be short and possess weak splicing signals around their junctions [33]. Finally, alternative 3' and 5' splice sites are mainly evolved from constitutive exons after introducing mutations that could create competitive splice sites [33].

The alternative splicing mechanism is evolutionarily conserved and is observed abundantly in multicellular eukaryotic organisms [34]; however, its prevalence increases in more behaviorally complex species such as humans [35]. The prevalence of alternatively spliced genes grows from $\sim$25% of the genes in *C. elegans* and $\sim$60% in *Drosophila melanogaster* to $\sim$95% of the genes in human [36]. Also, the relative abundance of splicing types changes among organisms. As an example, in lower metazoans intron retention is common while the relative abundance of skipped exons grows for more complex species [37].

Because the AS mechanism is conserved, one promising way to evaluate related hypotheses would be using model organisms. In the second chapter of this thesis, we chose to investigate the regulation of alternative splicing in *Drosophila melanogaster*. *Drosophila melanogaster* shares a large amount of its genetic content with human has been broadly used to improve our understanding of many cellular mechanisms including alternative splicing [38]. For example, Reiter *et al* showed $\sim$77% of human disease genes have statistically significant related sequences in *Drosophila melanogaster* [39]. According to FLYBASE [40] (release: May 24, 2016), the genome of *Drosophila melanogaster* contains $\sim$17,700 genes of which $\sim$13,900 are protein coding. These 13,900 protein coding genes encode $\sim$30,400 protein coding isoforms in total (an average of $\sim$2.2 isoforms per gene) manifesting the potential of AS regulation in *D. melanogaster*. In particular, more than 40% of the genes are alternatively spliced and there exist a set of highly complex genes ($\sim$50 genes) each encoding over 1000 isoforms [41]. Similar to human, different patterns of splicing are detected in *Drosophila melanogaster* [38, 41]. Dscam (Down syndrome cell adhesion molecule) is an example of a gene displaying complex AS patterns. The gene

contains 20 constitutive and 95 alternatively spliced exons [42]. Combinatorial assembly of observed local splicing patterns in Dscam can potentially encode more than 38,000 isoforms; a number greater than the number of genes in the entire *D. melanogaster*'s genome. Dscam encodes an axon guidance receptor and this huge level of complexity seems essential for its functional roles as an axon guidance receptor [42].

In addition to the interesting features of splicing in *Drosophila melanogaster* that resembles AS in human, a massive volume of publicly available data makes *Drosophila melanogaster* a promising candidate model to study alternative splicing. UCSC genome browser [43, 44] provides a genome annotation of *Drosophila melanogaster* and an alignment of its genome to 14 other Drosophila species. This alignment enables benefiting from evolutionary information and comparative methods. Besides, FLYBASE is a rich growing source of information on gene expression, genes interactions, observed phenotypes, and also genome features gathered from thousands of papers [40]. Additionally, there exist experimental data sets generated by different experimental pipelines including RNA-seq and chromatin immunoprecipitation (ChIP) in the MODENCODE project [45]. These experiments are replicated on different tissues and through various developmental stages, and the information can be used to study cellular behaviors in a condition-specific manner.

Apart form the cis-acting regulatory sites discussed, there are two other main classes of such primary sequence signals known as splicing silencers and splicing enhancers [46, 47]. Moreover, these sites and the corresponding trans-acting proteins have the tendency to decrease (in the case of silencers) or increase (in the case of enhancers) the probability of a neighboring intron to be spliced. These signals can be intronic or exonic (Exonic Splicing Silencers (ESS)/Enhancers (ESE), Intronic Splicing Silencers (ISS)/Enhancers (ISE)). These regulatory motifs can function as a silencer or an enhancer depending on their location in a pre-mRNA sequence. For instance, if G triplets occur in introns, they act as enhancers and in exonic context they usually act as silencers [31].

Recent studies have significantly improved our understanding of the AS mechanism. Apart from the regulatory elements discussed, splicing is found to be modulated by other factors such as transcription rates, pre-mRNA structures, histone marks and nucleosome positioning [48]. Among these, of special interest to my focus in this thesis are transcription rates and pre-mRNA structures, as they are more related to the splicing components I

investigate.

RNA splicing generally happens co-transcriptionally [49, 50]. Being co-transcriptional provides further opportunities to regulate the splicing mechanism. For example, C terminal domain (CTD) of RNA polymerase II (RNA poly II) helps in the recruitment of splice factors [50]. Additionally, when RNA pol II elongates slowly, weak 3' splice sites acquire higher chance of being properly processed without competing with stronger downstream 3' splice sites [50]. Also, the transcription rate can influence formation of alternative structures which in turn can affect splicing, as discussed in the following [51].

Pre-mRNA structures act as additional regulators of alternative splicing [52]. Several mechanisms in which the formation of structure influences splicing patters have been reported in literature [53]. A considerable amount of studies reported that the substrate selection of RNA binding proteins depends not only on the primary sequence of a target, but also on the target structure [52], and clearly this conformation dependence exists for many of the major SR proteins (proteins with RNA binding motifs important in AS regulations) as well. As a simple example, an RNA sequence which is detected by the spliceosome machinery (*e.g.* 5' and 3' splice sites) becomes inaccessible because of its base-parings with the other parts of the sequence in the human *MAPT* gene (microtubule-associated protein tau). Meanwhile, the splicing machinery requires well-suited distance between the splicing consensus motifs in order to carry out splicing in a certain way. Pre-mRNA structures have the tendency to decrease the effective distance between the conserved motifs, and thus resulting in a modified pattern of splicing. These findings are well supported by further computational evidence as well. Pervouchine *et al* [54] searched for pairs of complementary sequences around splice sites that can form stable hairpin structures. By studying mammalian protein-coding genes, they identified hundreds of such pairs where the energy of suggested structures could modify the pattern of splicing. In an earlier study, Meyer and Miklos [55] analyzed the alignment of 11 human genes to other vertebrates and found conserved double-stranded structures in coding regions. They showed that among codons encoding the same amino acid (due to degeneracy of genetic code) there is a selective pressure towards those leading to a more appropriate double-stranded structure. Also, in a case study, Meyer and Miklos [55] predicted secondary-structures of regions around exon 12 of the human CFTR (cystic fibrosis transmembrane regulator) gene for the wild

type sequence and also for the sequences carrying synonymous mutations. They showed that secondary-structure of sequences with high (experimentally evaluated) splicing efficiencies are more similar to each other than to those sequences with low splicing efficiencies. These studies suggest a global regulation of alternative splicing by the formation of different structures and consequently, taking the secondary-structure of pre-mRNAs into consideration facilitates understanding splicing patterns [53].

A considerable number of human diseases have been linked to aberrant splicing events [56]. Genomic mutations can create or destroy splicing sites or splicing enhancers and silencers, and in this way they sometimes alter the splicing patterns [52, 56]. For example, the severity of spinal muscular atrophy is affected by the creation of an ESS (Exonic Splicing Silencer) [57]. Also, mutations in genes involving the splicing mechanism have been linked to some human diseases such as retinitis pigmentosa [58]. After the uncovering of the cause of these and many other diseases, splicing events became important therapeutic targets.

The importance of AS regulation can be illustrated by an example involving *tau* protein. A gene located on chromosome 17, *MAPT*, encodes *tau* which is a microtubule-associated protein required for the polymerization and stability of axonal transport in neurons [26]. Through the alternative splicing of exons 2, 3 and 10, six protein isoforms are produced in the adult human brain. There exist three patterns of splicing which involve exons 2 and 3; and exon 10 is included or skipped independently [58]. Exon 10 and three other exons (9, 11, and 12) encode four microtubule-binding domains. Depending on the inclusion or exclusion of exon 10, the N-terminal of the resulted protein can have 3 or 4 microtubule-binding domains. In a normal human brain, the abundance of isoforms including exon 10 is equal to isoforms where exon 10 is spliced out, and the ratio of these two sets of isoforms seem to be important for neuronal function [26, 58]. Furthermore, it has been shown that mutations in *tau* protein cause neuro-degeneration accounting for fronto-temporal dementia and parkinsonism. Further analysis of the pre-mRNA structure revealed that a stem-loop structure could be formed involving the 5' exon/intron junction of exon 10 and the accessibility of the splice site is regulated by the structural configuration of this region [53] (See Figure 1.3). Mutations that destabilize the helix structure enhance the accessibility of the region and result in the increase of the set of isoforms

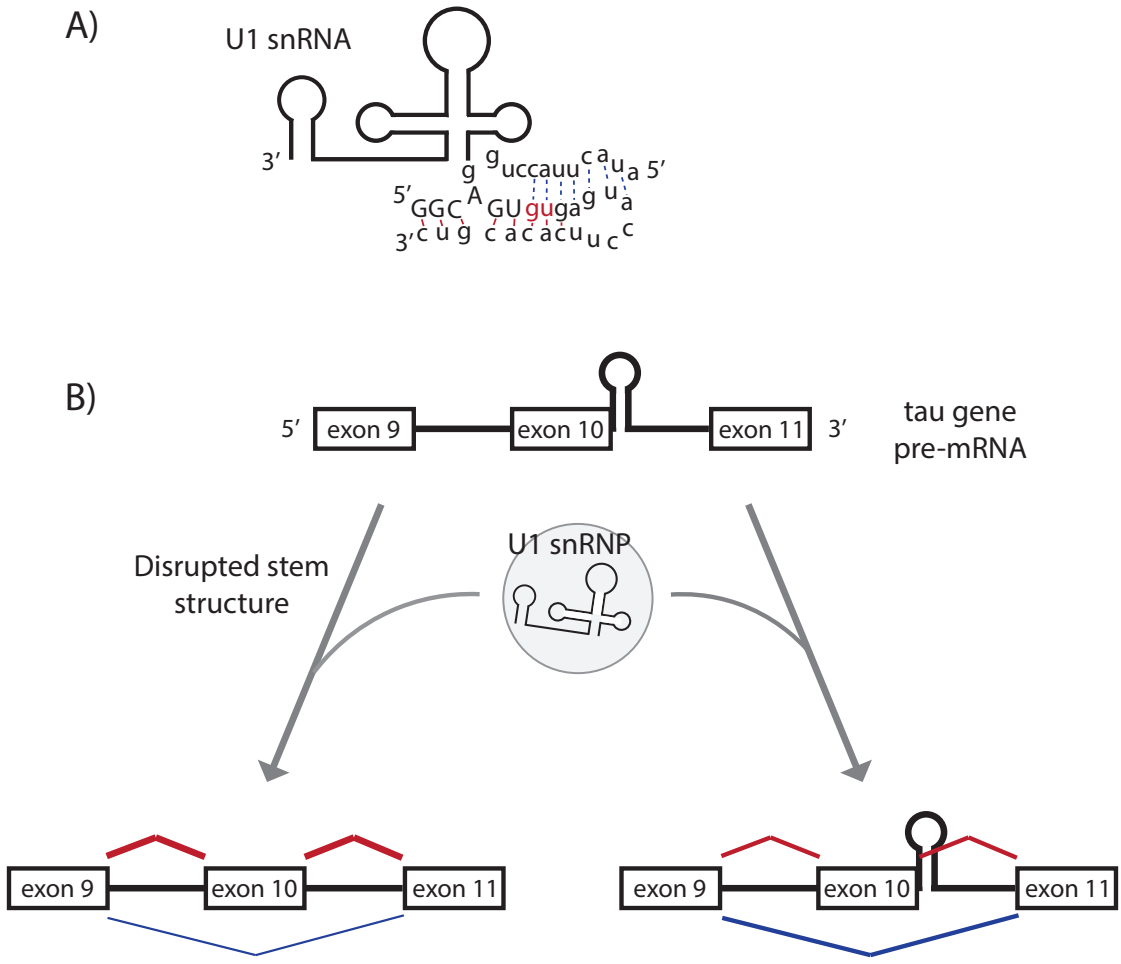

Figure 1.3: Alteration of alternative splicing in a human disease (modified from [53]). A hairpin-loop structure plays an important role in regulating the inclusion of exon 10 in the human *tau* protein. **(A)** The U1 snRNA structure and its potential interaction with the exon-intron hairpin-loop structure, revealed by NMR studies. Exonic and intronic nucleotides of the hairpin-loop are shown by uppercase and lowercase letters, respectively. If the structure is formed, the interaction of the region and U1 snRNA (the dashed line) does not happen properly, and U1 snRNA cannot detect the region. **(B)** Mutations in the primary sequence disrupt the hairpin-loop structure and increase the recognition of the region by U1 snRNA. As a result, more transcripts will include exon 10 (Abundance of the isoforms are presented by the thickness of red and blue lines in the two conditions); the condition leads to frontotemporal dementia and parkinsonism.

that include exon 10. Accordingly, therapeutic agents have been suggested to stabilize the stem-loop configuration.

Genome-wide studies have broaden our understanding of regulations and functions of alternative splicing; however, considering the diversity of cis-acting elements and the huge number of corresponding trans-acting factors, there is still so much unknown regarding position dependent and context dependent regulations and functions of alternative splicing [59]. Moreover, the role and importance of factors such as non-coding RNAs and dsRNAs, and upstream pathways are being more appreciated based on recent studies [60]; accordingly, the discovery of many new AS regulators are anticipated [61]. Finally, to get closer to understanding the splicing code, we need to investigate genes that are affected by specific splicing factors [62].

Despite the numerous studies on alternative splicing, some fundamental questions are yet to be answered. For example, it is still not clear what percentage of observed isoforms are essential for regulating cellular responses [49], or what are the relevance of coupling alternative splicing and other mechanisms such as transcription? [49] and how abundant and functionally relevant are the coupling of alternative splicing with other mechanisms? In this thesis, I investigate some of these questions.

1.1.3 Computational identification of alternative splicing using RNA-seq data

Despite being remarkably helpful in improving our understanding of the AS mechanism, properly interpreting RNA-seq data is challenging. Short reads are sometimes mis-aligned, especially when they originate from the repetitive regions of genomes (*e.g.* more than 50% of the human genome constitute repetitive elements [63]) or when they harbor multiple sequencing errors. Besides, sequencing biases due to non-uniform sampling of sequencing machines should be appropriately addressed [64, 65]. Additionally, many of the isoforms share common parts, making prediction of reads' origins nontrivial. Understanding and modeling these and other potential issues help avoid misleading conclusions.

The importance of the AS mechanism and the inherent complexity of studying AS using RNA-seq data have motivated the development of several computational methods [66–

76]. Alamancos *et al* published a comprehensive review of the proposed methods along with their strengths and limitations [77]. Conceptually, these methods can be classified into two groups: The first group contains algorithms that model the problem at the isoform level and assess the differential usage of entire isoforms; and the second group contains methods that model AS at local regions (*e.g.* an exon or an intron) without being concerned to the other parts of transcripts. Another criteria that differentiate methods is whether they are restricted to the existing transcriptome annotations or they detect *de novo* AS events as well.

Some of the methods performing differential splicing analysis at the isoform level are CUFFDIFF2 [71], BITSEQ [72] and MISO [73]. All three methods take aligned reads in addition to annotation files as input and provide information on the differential regulation of genes and isoforms. Apart from distinct statistics being used in these methods, there are some other clear distinctions as well. CUFFDIFF2 can detect *de novo* isoforms and AS events. Both BITSEQ and CUFFDIFF2 allow incorporating biological replicates, while MISO cannot. CUFFDIFF2 assigns a *p-value* to candidate events, MISO reports a Bayes factor, and BITSEQ uses a one sided Bayesian test to rank the genes based on their probability of being up or down regulated.

The event based differential alternative splicing analysis includes methods such as MISO [73], DEXSEQ [74], DSGSEQ [75], and DIFFSPLICE [76]. MISO can be applied in both isoform-based and event based analysis and therefore is placed in both categories. All methods accept aligned read files and identify differential regulation in local regions of transcripts. Among them, DIFFSPLICE is the only method that does not rely on annotation files. It constructs alternatively spliced modules (ASMs) using the aligned reads which represent regions where transcripts diverge. Accordingly, it is able to identify complex splicing events. On the other hand, MISO is able to distinguish between 8 different pre-annotated types of splicing (those shown in figure 1.2). In contrast, DEXSEQ and DSGSEQ are specialized for only one type of AS event, skipped exon. All the methods except MISO incorporate information from multiple replicates. MISO assigns a Bayes Factor values to each of the identified events as a measure of confidence, DEXSEQ reports corrected *p-values*, DSGSEQ uses Negative Binomial statistics to rank AS candidates and finally DIFFSPLICE outputs events under a given false discovery rate by considering its

introduced test statistics.

Isoform-based methods model the AS problem in more detail and take into account all potential transcripts and all reads aligned to the genes under investigation. Therefore, properly solving these models can provide helpful clues on global regulation of isoforms. However, sequencing biases with regard to non uniform sampling, as well as shared regions among transcripts complicate the inference problem. In situations where the genes constitute many alternatively spliced isoforms, these methods encounter problems [78]. On the other hand, local event based methods resolve this issue by only considering reads aligned to a small region of interest. Clearly these methods disregard information from many reads and are especially error prone when local events are short [78]. The appropriate method should be selected according to the research question, type and amount of available data (*e.g.* read length and read depth), and also the completeness of existing annotations for the species of interest.

1.2 RNA editing by ADAR proteins

In the second chapter of this thesis, I investigate the inter-relation between alternative splicing and RNA editing. Here, I briefly summarize what we already know about the editing mechanism.

1.2.1 Mechanism and abundance of A-to-I RNA editing

RNA editing is a widespread molecular mechanism which modifies transcripts [79]. The mechanism was first discovered in 1986 in trypanosomes, where nucleotide insertions cause reading frame shifts [80]. However, in mammals, the most frequent type of RNA editing is A-to-I conversion carried out by ADAR (Adenosine Deaminase that Act on RNA) proteins through deamination process (Figure 1.4). Most cellular mechanisms interpret inosine as guanosine, including splicing and translation. Therefore, in RNA-seq data also adenosines will be presented as guanosines. Some cellular factors (e.g., Tudor staphylococcal nuclease involved in RNA interference), however, can distinguish inosine from guanosine (as shown in *Xenopus laevis* [81]).

Different ADAR genes show non-identical behavior based on their distinct conserved

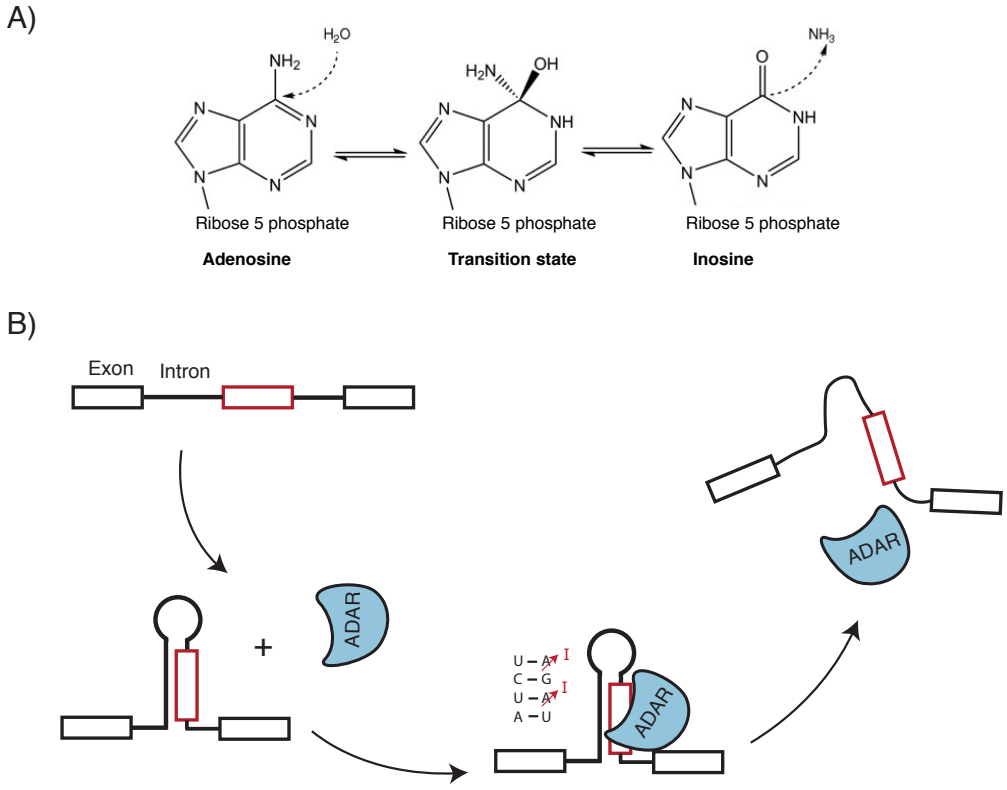

Figure 1.4: A-to-I mechanism carried out by ADAR proteins. **(A)** Chemical process of deamination through which an adenosine is converted to an inosine (Part A from [82]). **(B)** ADARs target double-stranded structures in pre-mRNA molecules. Many of these structures are formed by base pairings between exons and flanking introns and usually upon ADAR binding, multiple nucleotides are converted until the structure is destabilized and ADAR is released [83] (more details in the text).

protein domains [84]. All members of the ADAR gene family share protein domains essential for RNA binding and catalytic activities (Figure 1.5). The dsRNA (double-stranded RNA) binding domains in the N-terminal region of ADAR genes fulfill the recognition and binding of the protein to the substrate. The highly conserved catalytic domain carries out the deamination process in all ADAR proteins. Additionally, there are some other domains which make human ADARs work uniquely [85]. The functional impact of Z-DNA binding domains in ADAR1 is still unclear, but one hypothesis is that Z_α domain local-

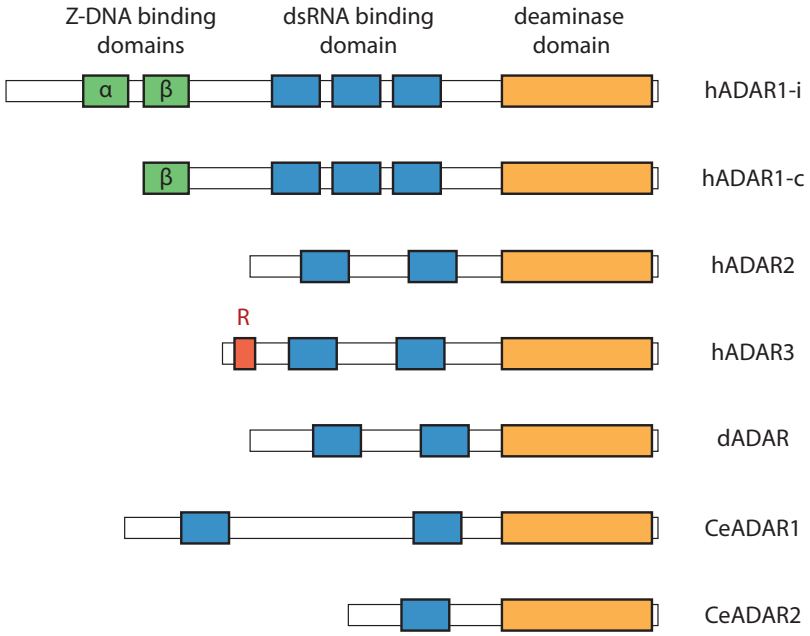

Figure 1.5: Organization of domains in ADAR proteins (from [85]). Domains that are identified in ADAR family members are shown for three genes in human genome (ADAR1 encodes two expressed isoforms), two genes in *C. elegans* and one gene in *D. melanogaster*. The deaminase domain and the dsRNA binding domains are common in species, whereas there are other domains specific to some of the genes.

izes ADAR1 at genes being transcribed, which enables ADAR1 to more efficiently use intronic regions required for editing, before they are removed [85]. Finally, ADAR3 has been shown to bind to single stranded RNA with the aid of its arginine-rich RNA binding domain (R-domain) [85].

A remarkable number of RNA editing events were found in different species indicating their significant potential to contribute to the regulation of other cellular mechanisms. RADAR [86] (Rigorously Annotated Database of A-to-I RNA editing) database gathered $\sim$5,000 editing sites in fly, $\sim$9,000 editing sites in mouse and over 2.5 million, a surprisingly huge number, editing sites in human (version 2, update: December 24, 2014).

The identified editing sites occur in both exons and introns. Based on RADAR annotations intronic editing happens $\sim$20 times more often than exonic editing in human

($\sim$2,000,000 intronic sites compared to $\sim$100,000 sites in coding regions and untranslated regions). In Drosophila, however, exonic sites are $\sim$1.5 times more observed compared to intronic sites ($\sim$2700 exonic sites compared to $\sim$1750 intronic sites). The dominance of intronic sites observed only in human can be at least partially explained by the prevalence of repetitive elements in human genome that can form dsRNA structures served as ADAR targets. Meanwhile, it should be noted that most of the RNA-seq libraries are enriched for the mRNA molecules where most introns are removed. Thus, the ratio of detected intronic events presents only a lower bound of genuine intronic targets.

In the second chapter, I study the reciprocal influence of RNA editing and alternative splicing in *D. melanogaster*. Considering the large number of identified A-to-I RNA editing events in *Drosophila melanogaster* in addition to the valuable publicly available data discussed before, we use *Drosophila melanogaster* data to study ADAR mechanisms, as well as alternative splicing. *D. melanogaster* has one ADAR gene (dADAR), and ADAR2 is the most similar gene to it among vertebrate ADARs [85]. DADAR is highly expressed in the central nervous system, and similar to vertebrates, its expression shows temporal regulation [87]. In recent years, there has been an increasing amount of studies on the importance and abundance of RNA editing in *Drosophila melanogaster* [88, 89].

1.2.2 Functions of RNA Editing

ADARs require double-stranded RNA regions to perform the deamination process [84]. Double-stranded RNAs are composed of hydrogen bonds that form between pairs of complementary nucleotides (A-U, C-G, and G-U) in an RNA molecule. In primary transcripts, these regions are typically formed by local RNA secondary-structure features such as hair-pins and they can be very long ($>$500 nucleotides). Once an appropriate double-stranded region is found, ADARs bind a base-paired adenosine and edit it without being very specific about the primary sequence surrounding the substrate [90]. In other words, the requirement for a double-stranded structural context is much more important than the primary nucleotide composition in specifying a potential ADAR binding site [91]. Somewhat surprisingly, this key feature has not yet been directly exploited in most RNA editing prediction programs [92, 93].

One of the key features of ADAR-derived RNA editing is that even in the same cell, the editing of two transcripts of the same gene does not necessarily involve identical RNA editing sites, but only the same double-stranded region which seems to be necessary and sufficient requirement for RNA editing to have the desired functional effect. Many of the known double-stranded regions serving as ADAR binding sites are formed between exonic sequences and complementary intronic sequences [94] (known as editing site complementary sequences). This supports the idea that editing usually precedes splicing [95]. Also, for many editing sites, the levels of pre-mRNA editing and mRNA editing correlate well in *Drosophila melanogaster* showing that RNA editing can happen co-transcriptionally [89]. A well-studied example is the editing of RNA structures formed between inverted *Alu* repeats in human transcripts [96]. *Alu* repeats constitute more than 10% of the human genome and can readily form double-stranded region and thus potential RNA editing sites by binding to their inverted copies in the same primary transcript. When one site is edited, other adenosine nucleotides in the same double-stranded region have a high chance of also being edited by the same ADAR protein; this may result in the conversion of several adenosines in a small region [97, 98].

Several functions of RNA editing have been identified so far. I briefly review some of these functions in the following.

RNA editing generally destabilizes the structure of its targets [83]. The function of an RNA molecule is mainly determined by its structure [99]. In recent years, the crucial role of RNA structure in regulating other cellular mechanisms has become more clear [100, 101]. Blow *et al* [83] studied several editing sites in the human transcriptome to investigate the global effect of RNA editing on the stability of the target's structure. By predicting the secondary-structure of editing regions before and after the corresponding editing events, the authors illustrate that the abundance of edited A:U matches (which is changed to a G:U mismatch) reduces the stability of the target molecule. Alteration in the structure of a target molecule changes the way it interacts with other molecules or the way it responds to cellular machineries.

Diversifying protein products of a single gene is another known function attributed to RNA editing [102]. Non-synonymous modifications in coding regions of transcripts produce protein products with altered functionalities. Most of these editing events oc-

cur in the genes involving rapid electrical and chemical neuro-transmission in *Drosophila melanogaster* [94], which are strongly expressed in central nervous systems. Stark *et al* [103] observed that the high conservation in coding regions of 12 aligned *Drosophila* genomes continues to hundreds of nucleotides after the stop codon, and argued that the editing of stop codons in these genes could be a candidate underlying mechanism in such cases in order to generate two isoforms which are significantly different.

ADAR proteins can also affect gene expression through the editing of miRNA molecules and their targets [104]. MiRNAs bind to their complementary 3' untranslated regions of messenger RNAs and suppress gene expression by preventing translation or causing target degradation. Several studies have revealed that a notable amount of editing events occur in 3' UTRs in human and mouse [104, 105]. The large number of editing events in 3'UTRs indicates the potential effect of RNA editing on post-transcriptional gene silencing. Additionally, some of the editing events are known to happen in miRNA sequences, and this could be considered as another way of affecting gene expression by ADAR proteins [104].

Severe phenotypes have been associated with the deficiency of ADARs. *Drosophila* deficient for ADAR shows severe neurological disorders such as locomotor incoordination and temperature sensitive paralysis [106]; *mice* deficient for ADAR1 have a heterozygous embryonic lethal phenotype [107], and in humans, variations of RNA editing have been linked to neurological and psychiatric disorders [108]. These severe phenotypes also make it challenging to investigate ADAR functions.

Figure 1.6 illustrates an example of RNA editing in protein coding regions of $5-HT_{2C}R$. 5-HT receptors are G-protein coupled receptors that cross the cell membrane several times and play roles in signal transduction. A double-stranded structure is formed between exon 5 and intron 5 of its pre-mRNA molecule which serves as an editing substrate. Accordingly, five sites in exon 5 of $5-HT_{2C}R$ undergo A-to-I editing. If all the sites are edited, the G-protein-coupling activity of the corresponding protein will be hugely different from the corresponding unedited protein[109]. As a consequence, RNA editing of these genes have been associated with some psychiatric disorders such as depression [109].

Although studies suggested some primary sequence features and also proteins that affect ADAR activity in specific target regions, the general regulation of RNA editing is

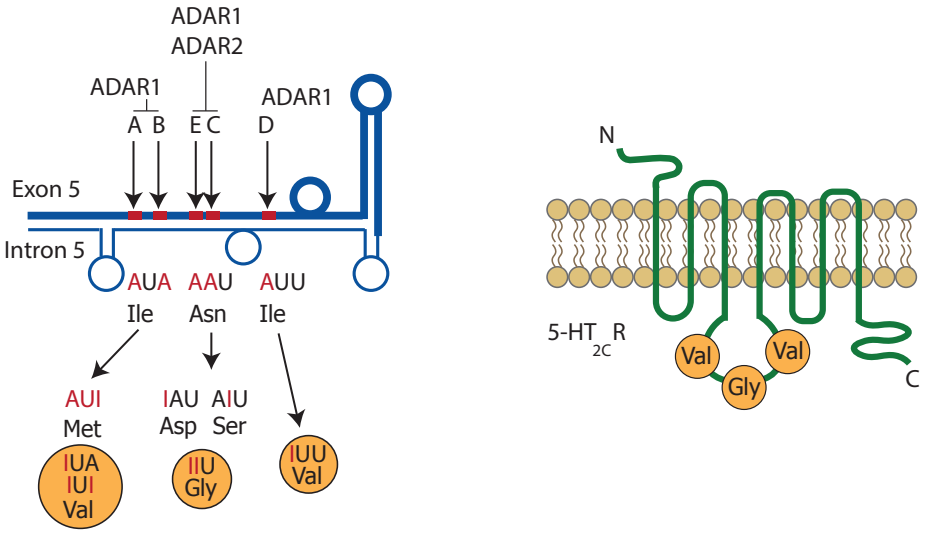

Figure 1.6: Modification of the amino acid sequence of the human $5-HT_{2C}R$ through RNA editing (figure from [109]). A part of the pre-mRNA molecule of $5-HT_{2C}R$, a transmembrane receptor, is shown in this figure. ADAR1 and ADAR2 target 5 sites (A, B, C, D, and E) of exon 5 to produce proteins with highly modified properties. These 5 sites are embedded within a hairpin structure formed by base-pairings of exon 5 (thick blue line) and the exonic complementary sequences of intron 5 (thin blue line).

unclear. Inverted copy sequences in proximity of a region increase the editing probability of that region, probably by having the potential to form the double-stranded region required for ADAR binding; in support of this, *Alu* repeats were observed to constitute the majority of ADAR targets in human transcripts [110, 111]. Moreover, some short primary sequence preferences have been observed for ADAR proteins in human [112, 113], mouse [93] and fly [114]. On the other hand, few RNA-binding proteins have so far been shown to suppress the editing levels of specific targets [115]. The *SFRS9* gene, which encodes a splicing factor, represses the editing of the *cyFIP2* gene. This could be the result of competition between the two proteins for common substrates or due to the protein-protein interaction between ADAR2 and *SFRS9* [115]. The level of ADAR expression is another regulatory factor, despite it not usually correlating well with the level of RNA editing [116].

Moreover, considering the huge number of identified editing sites, there is still much to be discovered and understood regarding the molecular mechanisms and functional roles of RNA editing. The way RNA editing interacts with and affects other mechanisms is still unclear [90]. For example, given the abundance of RNA editing events in noncoding RNAs, and the growing evidence for the influence of RNA editing on gene expression, more detailed study of how editing affects RNA interference seems promising [84]. Additionally, recent studies suggest that alternative splicing and RNA editing mechanisms have the potential to influence each other [95, 117]. Considering the co-occurrence of RNA editing and alternative splicing in same genes[114, 117], we study their potential inter-relation in this study.

1.2.3 Computational detection of RNA editing

The number of detected RNA editing sites has grown rapidly since the development of RNA-seq technologies. Sequencing machines are able to generate hundreds of millions of reads with a much lower cost compared to Sanger sequencing. As a result, the large number of reads aligned to a single location makes the detection of RNA editing sites with a low level of editing much easier. In the following, I discuss some of the computational methods proposed to encounter potential errors when using RNA-seq data. Table 1.1 summarizes these methods.

Author/year	Strategy	ADAR features incorporated	confidence measure?	Reference
Peng *et al* (2011)	Using thresholds	None	No	[104]
Danecek *et al* (2012)	Using thresholds	Vicinity of targets	No	[98]
Li *et al* (2012)	Likelihood model	None	Log-likelihood ratios	[113]
Guiliany *et al* (2012)	Bayesian model	None	Probability based	[118]
Laurent *et al* (2013)	Thresholds/Random forest	None	Ranked based	[114]
Zhang *et al* (2015)	Mutual information based	Randomness of editing	Ranked based	[119]

Table 1.1: Summary of the methods proposed to identify editing events.

Early methods of identifying editing events by high-throughput sequencing data were threshold based, mainly for their simplicity [104, 111, 120]. The major concern when applying empirically determined thresholds is that the margin value of passing a threshold is not considered in making the final conclusions [118]. Li *et al* [120] claimed the discovery of thousands of editing events for each of the twelve possible conversions, most of which

were repeatedly reported as being the consequence of sequencing artifacts [121, 122]. One of the convincing arguments was that most of those events were predicted to occur on either ends of the reads where the probability of error is much higher [121]. Accordingly, More stringent filters were utilized in the following studies to prevent the abundance of false positive predictions.

Likelihood models offer the ability to quantify the significance of predictions. To overcome the shortcomings of threshold based models, Bahn *et al* [113] applied a statistical approach. By considering the quality score of the aligned reads and the position of nucleotides in a read, the authors proposed a model to compute the likelihood of a site being edited with ratio r. Then, in order to find the ratio of editing, this likelihood function is maximized with respect to r. Finally, the confidence in predictions of editing ratios is assessed by comparing them against the null hypothesis using a log likelihood ratio test.

A study by Guiliany *et al* [118] involved the jointly modeling of whole genome sequencing and RNA-seq data using mixture models. This model, called *Auditor*, requires DNA and RNA base counts as input, and calculates the probability of editing at each position in the genome. To benefit most from data, transcriptotype (mRNA genotype) is modeled as a function of genotype using a transition matrix to present the probability of observing an specific transcriptotype given a defined genotype. The transition values can be learned by the expectation maximization method. When the pipeline is coupled with MUTATION-SEQ [123], a method to detect somatic mutations, the enzymatic modifications carried out by ADARs can be effectively distinguished from other types of observed discrepancies.

In more recent studies, other machine learning approaches have also been proposed. Laurent *et al* [114] developed a method based on multiple rounds of detection and validation to adjust the applied thresholds. They also applied the Random Forest method to train a classifier based on true positive and true negative events found in their validations, and also to assess the importance of different features. In an interesting and completely different approach, Zhang *et al* [119] used the fact that if a read covers two nearby SNPs (single nucleotide polymorphisms), the two variable positions will have a fixed allelic linkage; however the fixed linkage breaks in the case of an RNA editing event coupled with an SNP due to inherent randomness in editing. Accordingly, they compute the mutual

information (MI) for observed variants. If the MI value deviates significantly from the distribution of publicly available MI values for SNPs, it would be considered as an editing event.

Other known ADAR features can be incorporated to improve the computational power of existing pipelines. As an example the fact that ADARs edit multiple sites within a small region has been incorporated in a pipeline introduced by Danecek *et al* [98] to extend the list of detected editing sites. As discussed, one of the requirements of ADAR targets is that they must be dsRNA regions. In the following, I briefly explain methods developed to computationally detect structural regions within RNA transcripts.

Predicting tertiary structures is computationally hard, and experimentally costly and time consuming. Fortunately, RNA secondary-structure are also informative for uncovering functional roles of RNA molecules [51]. Conceptually, RNA secondary-structure prediction methods are classified into two main categories: energy-based methods and evolution-based methods.

Energy based methods are commonly established upon the idea that the ultimate structured RNA molecule is the one minimizing the overall free Gibbs energy. RNAFOLD [124], MFOLD [125] and SFOLD [126, 127] are some of the methods that try to solve this optimization method by introducing time and memory efficient algorithms. These methods are fast and work for sequences of thousands of nucleotides, however they have some limitations. Several assumptions of these methods are violated *in vivo* and the accuracy of them rapidly drops by sequence length for sequences longer than few hundred nucelotides *in vivo*. First, *in vivo*, proteins and other molecules interact with a folding transcript and impose further folding constraints. Second, the folding time for RNA molecules is finite and the structure may never reach the optimum energy point; and finally, there exist uncertainties in the experimentally measured parameters (stacking energies, energy of bulges, etc) required by these methods.

The second class of RNA secondary-structure prediction methods rely on a completely different assumption. The basic idea is that homologous sequences diverge through evolution in a way that the functionally important structures are preserved. In other words, if a base pair is functionally important, then although the primary sequence of the corresponding nucleotides may diverge, the changes always happen in a way that the pairing

potential is maintained (C:G base-pair changes to A:U and not A:G). Based on this concept, methods such as TRANSAT [128], EVOFOLD [129], RNA-DECODER [130, 131] and PFOLD [132] search for the linkage between divergence of pairs of bases. Given a set of aligned sequences and an evolutionary tree, Most of the proposed methods apply phylo-SCFG (stochastic context free grammar) to model and score the evolution of paired columns and unpaired columns statistically and find an optimal solution based on the resulting conservation scores. Moreover, the flexibility of phylo-SCFGs allows assigning prior probabilities to predictions and also capturing additional hypotheses on secondary-structures. For example, if functional structures form in coding regions, apart from the structural restrictions, the amino acid sequences should also be preserved. In other words, only the third codon position can freely hold structural information because it usually does not change the amino acid, in contrast to the first and the second positions. RNA-DECODER is the only method that properly models these different evolutions and in the case of RNA editing, because many of the editing events happen in coding genes, the method could be helpful by incorporating coding information as well.

1.3 Phosphorylation by Cyclin Dependent Kinase 12 (*CDK12*)

In the third chapter of this thesis, I study how *CDK12* influences the regulation of RNA processing and specifically alternative splicing. In this section, I review our current understanding of *CDK12* mechanisms and functions.

1.3.1 *CDK12* is a protein kinase

One other mechanism that contributes to expanding genome repertoire is post-translational phosphorylatoin. Over 500 protein kinases have been annotated that perform the phosphorylation process [133]. This large family of regulatory enzymes supplies cells with an additional level of regulation in order to control most cellular processes [134]. The functionalities of these enzymes are crucial for determining cell fate; accordingly, impaired functions of kinases have been attributed to diseases including multiple cancer types..

Kinases have been investigated thoroughly as they constitute attractive targets for ther-

apeutics. Besides, studies have shown the potential to develop highly selective drugs for kinases. For instance, matinib is one of them with a high success rate in chronic-phase CML (Chronic myelogenous leukemia) patients [135]. Therefore, considering the general role of kinases in regulating many signalling pathways and the small number targeting agents designed so far, the future investigation to target other members of the family seems promising and essential [135].

One class of regulatory kinases are cyclin dependent kinases (CDKs). CDKs are inactive when they are in their monomeric form, and form holoenzymes with their cyclin partners for activation [136]. Although initial studies conducted based on their cyclin domain confirmed their involvement in cell cycle regulation, CDKs are known to be engaged in a variety of other mechanisms such as transcription, splicing and DNA repair [137–139]. Similar to other kinases, impaired CDKs are hallmark of several diseases. *CDK12* is one member of this family of enzymes which associates with *Cyclin K* to become active [140].

Cyclin dependent kinase 12 is evolutionarily conserved [141]. Human *CDK12* is a large protein (1,490 amino acids) [142] located on chromosome 17. The *Drosophila melanogaster*orthologue is 41% identical to the human *CDK12*, and the *C. elegans* orthologue shows 53% identity [143]. Among the human genes, *CDK13* has a very similar kinase domain, but other that that looks different [143]. The RS domain (domains rich in alternating arginine and serine residues) of *CDK12* is usually observed in SR proteins, proteins known to play crucial roles in the regulation of pre-mRNA splicing. Furthermore the protein is found to be co-localized with the splicing machinery and the hyperphosphorylated form of *RNA pol II* [144].

1.3.2 Functions of *CDK12*

CDK12 is involved in the regulation of transcription elongation. The protein helps in the productive elongation of *RNA pol II* by phosphorylating C terminal domain (CTD) of RNA polymerase II, as shown both in human and *Drosophila* [140, 141, 145]. An expression microarray study showed that the phosphorylation only modulates the transcription of a small set of genes, primarily long genes with many exons [146]. Some of these target genes are involved in genome stability including *BRCA1* (Breast and ovarian cancer

type 1 susceptibility protein 1), *ATR* (Ataxia telangiectasia and Rad3-related) and *FANCI* (Fanconi anemia complementation group I).

In addition, *CDK12* also contributes actively to the regulation of alternative splicing. The idea is supported by several evidence. First, as explained before, *CDK12* proteins contain RS domains. RS domains are usually observed in SR proteins and are believed to be important for recruiting proteins of the splicing machinery [147]. Second, over 30 splicing proteins interact *CDK12*, including *SRSF1* and *U2AF2*, and several 3'-end formation factors [148]. Finally, a study by Chen *et al* [149] illustrated that the expression of *CDK12* can modulate the splicing pattern of a synthetic E1A minigene; and Rodrigues *et al* [150] found that *CDK12* is essential for regulating the splicing activity carried out by *HOW* protein. Further genome-wide investigation seems necessary to uncover general regulation of AS modulated by *CDK12*.

Similar to many other protein kinases, functional and structural properties of *CDK12* qualify it as a promising drug target. *CDK12* is one of few genes recurrently mutated in ovarian cancer and these mutations are usually mutually exclusive with mutations in *BRCA1* or *BRCA2*, two of the most abundant mutated genes in ovarian cancer [151]. Disruption of *CDK12* has also been observed in breast and gastric cancers [152]. Furthermore, *CDK12* over-expression has been associated with poor prognosis power and a higher risk of tumour recurrence [144]. Thus, *CDK12* seems to be an attractive drug target for investigation, and better understanding of its general effect on RNA processing could help advancement in therapeutics.

In chapter 3, I explore how CDK12 regulates alternative splicing and gene expression, and how target genes are selected at a genome-wide scale.

1.4 Research contributions

In the following, I briefly summarize my research questions and my objectives in the three main chapters of this dissertation (Figure 1.7).

In chapter 2, my main research question was how RNA editing regulates splicing patterns. Based on the importance of RNA editing and alternative splicing mechanisms in diversifying gene products, as described in this section, I investigated the potential inter-

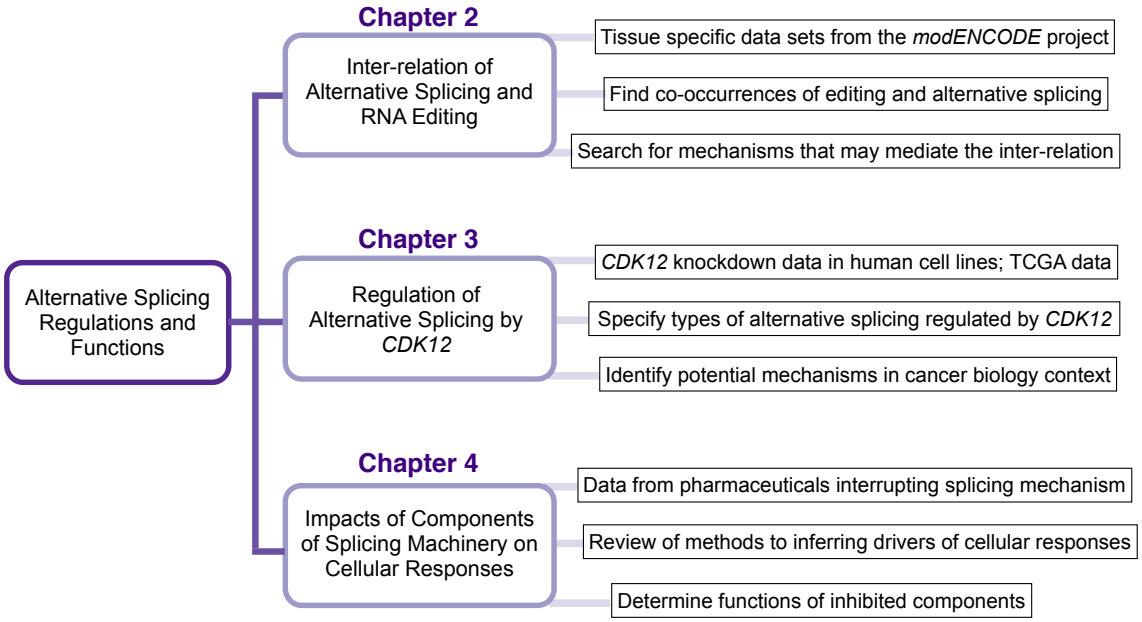

Figure 1.7: A diagram of my research presented in this dissertation. The figure shows data sets I use and the analyses I perform in the 3 main chapters of this study.

relation between these two mechanisms. Considering the existing evidence on some cases where editing regulates splicing [95], I hypothesized this regulation happens more frequently than what was known previously.

My main goal was to find local regions where splicing patterns is modulated by RNA editing and also to uncover the mechanism of regulation. I hypothesized splicing and editing can compete for common targets in local regions, or editing can influence splicing through modifying sequence motifs and structural features in a genome-wide scale. I addressed this problem in the context of *Drosophila melanogaster*, using tissue-specific RNA-seq data from the MODENCODE project.

In Chapter 3, the research question I explored is how *CDK12* contributes to the regulation of alternative splicing. Despite the growing number of studies investigating functions of *CDK12*, the mechanism through which CDK12 contributes to cancer development and progression is still unclear. I hypothesized that a part of functional roles of *CDK12* in cancer biology occurs through regulation of alternative splicing. Therefore, I performed

a genome-wide analysis of splicing and expression regulation by *CDK12* using knock-down and control libraries of breast cell line data. I also examined if my findings could be generalized in tumour cells using The Cancer Genome Atlas (TCGA) ovarian data [153].

In Chapter 4, my main objective was to assess how different methods in the literature can be employed to provide mechanistic insights when a gene is systematically inhibited by pharmaceutical agents at different levels. The inhibited genes are splicing related genes and the data are generated to study their contributions to splicing regulation, and better understanding the splicing mechanism as a complex machinery. I summarized appropriate methods in the literature, and compared their advantages and limitations. I also examined the usefulness of the data using one of the appropriate methods, and finally discussed how the methods in the literature should be adopted to properly benefit and extract information from this type of data.

Chapter 2

Genome-wide Identification and Characterisation of Tissue-specific RNA Editing Events in *Drosophila melanogaster* and their Potential Role in Regulating Alternative Splicing

2.1 Introduction

Recent studies suggest that alternative splicing and RNA editing mechanisms have the potential to influence each other [95, 117]. Obviously, RNA editing can directly modify splicing patterns by editing primary sequence motifs required such as splice sites, splicing enhancers or silencers [95, 154]. Other studies in human and fly suggest that many of the editing sites occur in transcripts encoding RNA-binding proteins that play roles in alternative splicing. This may alter the expression, efficiency or binding properties of these proteins which may in turn affect the splicing of many genes [114, 117]. On the other hand, different ADAR isoforms have different editing efficiencies [87], so the splicing machinery also has the potential to influence RNA editing. It thus seems obvious to hypothesise

that there are feedback loops between RNA editing and alternative splicing waiting to be discovered.

In the past few years, thousands of editing sites have been discovered by calling A-to-G differences between the reference genome and the transcriptome reads in human [104, 111, 113], mouse [98], and fly [38, 89] using RNA-seq data. One key challenge when analysing RNA-seq is to discriminate true editing events from artifacts [104, 111, 113] as explained in Chapter 1; RNA-seq data require sophisticated and statistical data analysis methods for reliably detecting RNA editing events.

Fortunately, the large number of experimentally confirmed A-to-I RNA editing events in *Drosophila melanogaster* and the considerable amount of publicly available data make the fly a promising model organism to study ADAR mechanisms. In recent years, there has been an increasing amount of studies on the importance and abundance of RNA editing in this organism [38, 87, 89]. *Drosophila melanogaster* has one ADAR gene (dADAR), and among vertebrate ADARs, ADAR2 is the most similar gene to dADAR [85]. In fly, dADAR is highly expressed in the central nervous system, and similar to vertebrates, its expression shows tight temporal regulation [87].

Here, we use tissue-specific high-throughput data sets of *Drosophila melanogaster* from the MODENCODE project [155] to identify RNA editing events in multiple tissues. To achieve this, we introduce a new computational analysis pipeline to accurately identify editing events and to distinguish genuine editing events from sequencing and mapping artifacts. In our analysis of the resulting, predicted cases of RNA editing, we search for cases of differential exon usage between pairs of different tissues to identify regions where RNA editing and alternative splicing may influence each other. Finally, in order to discover potential molecular mechanisms underlying this interplay, we identify many cases of evolutionarily conserved RNA secondary-structures that have the potential to regulate alternative splicing via RNA editing.

2.2 Materials and methods

2.2.1 Data set

To study tissue-specific RNA editing events, we selected tissue-specific RNA-seq libraries of *Drosophila melanogaster* from the MODENCODE project [45, 155]. These libraries correspond to paired-end, strand-specific RNA-seq reads of 74–120 nucleotides length. The strand-specificity of the reads allows us to assess the correct conversion types in overlapping or incompletely annotated parts of the genome [98], whereas the paired-ends improve the alignment of reads to repeat-rich regions of the genome which would otherwise easily result in incorrectly aligned reads or the false positive prediction of SNPs or RNA editing sites. The 29 selected libraries are classified into 10 tissues (Table.2.1). Some of these libraries are extracted from multiple tissues. For each library there exist two to five technical replicates. All libraries derive from the *OregonR* strain of *Drosophila melanogaster* which is, however, not the strain of the *Drosophila melanogaster* reference genome.

Dataset	Tissue	Dataset	Tissue	Dataset	Tissue
MOD4241	Head	MOD4266	Ovaries	MOD4259	Digestive system
MOD4242	Head	MOD4247	Accessory glands	MOD4256	Central nervous system
MOD4243	Head	MOD4249	Testes	MOD4257	Central nervous system
MOD4245	Head	MOD4250	Carcass	MOD4260	Fat body
MOD4246	Head	MOD4252	Carcass	MOD4267	Fat body
MOD4248	Head	MOD4254	Carcass	MOD4268	Fat body
MOD4263	Head	MOD4258	Carcass	MOD4261	Imaginal discs
MOD4264	Head	MOD4251	Digestive system	MOD4262	Salivary glands
MOD4265	Head	MOD4253	Digestive system	MOD4269	Salivary glands
MOD4244	Ovaries	MOD4255	Digestive system		

Table 2.1: Tissue specific data sets selected from the MODENCODE project. The IDs of the selected libraries and the tissues from which these libraries are sampled are shown in this table. The data contain 29 libraries from 10 tissue types.

Since we do not have genomic DNA sequencing reads in our data, it is essential to align the short transcriptome reads to the reference genome of the *OregonR* strain when searching for DNA/RNA discrepancies; otherwise, genomic differences between the genome of the *OregonR* strain and the *D. melanogaster*'s reference genome could be misinterpreted

as RNA editing events. We therefore generate an annotation for the *OregonR* genome by aligning the genome of the *OregonR* strain to the *D. melanogaster*'s reference genome. We first use MUMMER [156, 157] to find a set of consecutive matches of at least 20 nucleotides long. Next, we align the remaining parts between these matches using the NEEDLEMAN-WUNSCH algorithm [158] with default parameter values. Finally, we convert the coordinates of the reference annotation of *Drosophila melanogaster* in ENSEMBL [159] to the corresponding coordinates of the resulting *OregonR* genome.

2.2.2 Prediction pipeline

Figure 2.1 gives an overview of the steps of our computational analysis pipeline for identifying RNA editing events using multiple RNA-seq libraries and the reference genome as input. Considering the potential challenges in reliably detecting RNA editing events [160], we designed a probabilistic pipeline to achieve the following in an efficient manner: (1) filter variants against artifacts due to mapping and sequencing errors; (2) explicitly capture ADAR-specific features such as the requirement for double-stranded region to distinguish RNA editing events from other types of observed variants; and (3) leverage the statistical power derived from the size and number of our input data sets. In the following, we briefly explain the steps of our pipeline.

We use TOPHAT2 [163] to align short reads to the genome in a splice-aware manner. We allow up to five mismatches in the alignment step to permit TOPHAT2 to successfully align reads that have been RNA-edited multiple times. Next, we employ PICARD-TOOLS (http://broadinstitute.github.io/picard) to remove duplicates from each technical replicate. These duplicate reads may be generated during RT-PCR as a result of amplification bias [111, 113]. Finally, technical replicates are merged and positions showing DNA/RNA discrepancies are extracted for further analysis.

Our analysis pipeline combines a set of statistical and deterministic filters that apply two sets of threshold values, one set called the flexible set and one set called the stringent set (Figure 2.1). By employing these two sets of threshold values and leveraging the large size of the input data, it is possible to simultaneously lower the false positive and the false negative error rates. If only the stringent threshold values were used to distinguish genuine

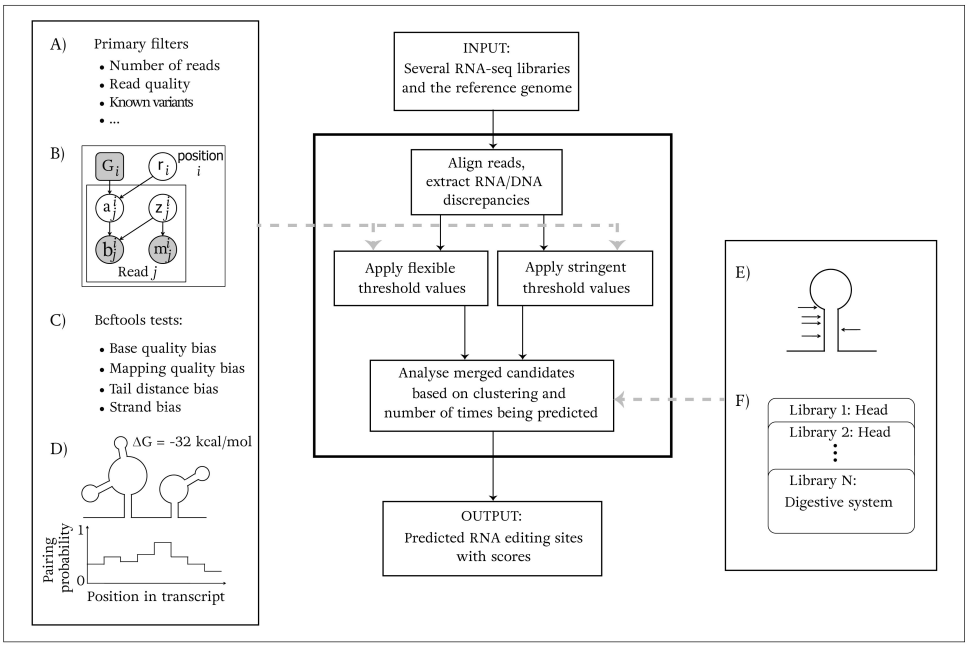

Figure 2.1: Outline of the computational analysis pipeline for identifying editing events from multiple RNA-seq libraries. The input consists of several RNA-seq libraries and the reference genome. As shown, first, reads are aligned and RNA/DNA mismatches are extracted. Then, two sets of values (for flexible and stringent filtering) are used for several filters (**(A)**-**(D)**) to remove potential experimental artifacts. Finally, our pipeline considers clustering of identified candidates and the number of times they are detected in multiple libraries to output a final set of predicted editing events. **(A)**-**(D)** show the statistical tests and filters used in our pipeline. **(A)** A set of primary filters used to assess the initial requirements for candidate sites. **(B)** The statistical graphical model (modified from [161]) that we use to find the maximum editing ratio, and to compute a log likelihood ratio score. Shaded circles are the random variables that are observed in data and unshaded circles are the ones that are inferred. The rounded square is fixed to represent the reference genotype. a is a binary variable which indicates whether or not a read aligned to a position comes from an edited molecule. z is also a binary variable that indicates whether the read is aligned correctly. The editing ratio of position i is presented with node r; and nodes m and b present mapping and base qualities. **(C)** Statistical tests in SAMTOOLS/BCFTOOLS [162] to check the potential biases in reads. **(D)** The energy of local structures and base pairing probabilities of nucleotides in close vicinity of candidate sites are used to ensure the structural requirements of candidates are met. **(E)** We use the fact that editing events occur in clusters to improve our predictions. **(F)** For less confident sites, the site requires to be detected in multiple libraries in order to be reported in our final set.

RNA editing sites from mapping and sequencing errors, filtering the potential artifacts could result in discarding many true RNA editing events, i.e., a high false negative error rate. On the other hand, using only the set of flexible threshold values could lead to an increased false positive error rate. To overcome this issue, our pipeline combines the two sets of threshold values. The potential editing sites that pass the flexible threshold values are only reported in the final output if they are detected in multiple samples and are close to other predicted sites.

After the alignment step, a set of primary filters are applied to reduce the identified DNA/RNA discrepancies and remove those that are likely to be due to mapping and sequencing errors (Figure 2.1.A). These primary filters examine, for example, the number of reads covering a candidate RNA editing site, the read and mapping qualities of the input data, and also the distance to both ends of the read. In addition, any known variants listed in the ENSEMBL fly variant files are removed. Some of these variants may correspond to genuine RNA editing events – similar to what has been observed in human SNP data bases [102] – yet we decided to be conservative and to remove all known variants in the absence of any corresponding DNA sequencing reads.

Figure 2.1.B shows the graphical model we use to compute the maximum likelihood editing ratio, and to apply a log-likelihood ratio test. The model is a modification of the model introduced in SNVMIX2 [161]. The original model considers both mapping and base qualities of the reads and takes uncertainties of bases and alignments into account. We took a part of the model and added a new node (shown with "r_i" in the figure) that presents the editing ratio (r). This can take values ranging from 0 (not edited) to 1 (always edited) with uniform prior. We model a_j^i (which indicates whether read j aligned to position i comes from an edited RNA molecule) to have a Bernoulli distribution with its parameter set to the editing ratio r. The conditional probability distribution for the other three nodes ("z", "b" and "m") are the original ones used in SNVMIX2 [161]. Using this statistical model, the null hypothesis of a position having an RNA editing level of zero is compared to the hypothesis of the position being edited with the inferred maximum likelihood level of editing. More precisely, for each candidate position i, we compute the following log-likelihood ratio score for position i:

$$\text{score}(i) = \text{argmax}_{\alpha} \log \frac{P(D_i \mid M_i, B_i, r_i = \alpha)}{P(D_i \mid M_i, B_i, r_i = 0)} \tag{2.1}$$

where r_i is the editing ratio; D_i presents the observed reads overlapping position i; and B_i and M_i are the base and the mapping qualities of reads, respectively, overlapping position i.

In the following step, the pipeline applies SAMTOOLS/BCFTOOLS [162] tests to identify and remove positions that are discovered as a result of potential biases. These tests have been used in the literature to improve the quality of variant calls [98]. Base quality and mapping quality tests gauge the bias of the corresponding scores between the reads showing the reference allele and the reads showing the variant allele. Two additional other tests evaluate the strand bias and the tail distance bias. The strand bias gauges the bias between the distribution of the strand of reference reads and the distribution of the strand of non-reference reads. The tail distance bias investigates whether nucleotide reads from one allele tend to occur closer to read ends compared to nucleotide reads from the other allele.

Unlike most other existing prediction methods for RNA editing sites, our analysis pipeline explicitly utilises the requirement for the existence of double-stranded regions in potential ADAR target regions [164] to further improve our predictions (Figure 2.1.D). Long double-stranded regions constitute perfect potential target sites for ADARs [84], and structured regions also recruit ADARs to nearby sites that are not in the same double-stranded region [101]. Consequently, the stability of potential structures has been used to rank output candidates [92], although edited double-stranded region have been observed to have a wide range of stabilities [97]. Also, the vicinity of complementary nucleotide regions which allows the formation of RNA secondary-structures was used to improve prediction results [93]. Most of the double-stranded regions bound by ADAR have been shown to correspond to *intramolecular* interactions, i.e., RNA structure features in the same transcript [83]. We therefore use local RNA secondary-structure prediction algorithms in our pipeline. We employ RNAFOLD [125] on a sequence interval of 200 nucleotides length around each candidate editing site to calculate the minimum-free-energy (MFE) RNA structure predicted for this region. We use the corresponding minimum free

energy as an indicator of the stability of all potential local RNA structures that can be formed in that region. Additionally, as ADAR binding and editing predominantly happens in double-stranded regions [84], we use RNAPLFOLD [165] to estimate the probability of a potential RNA editing site being in a double-stranded region. For this we examine sequence intervals of five nucleotides length around the candidate editing site. Finally, the two sets of thresholds (stringent and flexible) introduced above are applied to these potential RNA editing sites in order to incorporate structural information in our pipeline.

It is well known that ADAR tends to edit several sites in the same double-stranded region upon binding [97] which we explicitly judge by our analysis pipeline (Figure 2.1.E). In addition, we expect true RNA editing events to show up in several libraries due to the large amount of input reads. To use these features, we first include all candidate editing sites that pass the stringent threshold values. Any remaining candidate sites are then added if: a) The same position passes the stringent threshold values in another sample, or b) the position has been predicted (passes the flexible threshold values) at least twice and there is another identified site showing the same conversion type within a distance of 25 nucleotides.

To summarise, by using a large number of samples as input, by explicitly capturing ADAR specific features and requirements and by combining two distinct sets of threshold values, we create an analysis pipeline that has a low false positive as well as a high true positive rate (see results section below). Assuming similar mutation rates for transitions and similar mutation rates for transversions, we can use the ratio of A-to-G conversions in our predictions to estimate the false positive ration [98, 166]. Based on this estimate, we chose a set of pipeline parameters that result in a decent overall number of predictions and also a high ratio of A-to-G conversion type. Details of the pipeline including parameter values are explained in Appendix. A

2.2.3 Finding alternatively spliced exons

To find alternatively expressed exons between pairs of tissues, we use DEXSEQ [74]. DEXSEQ applies a generalised linear model to detect exonic regions that are differentially expressed between two conditions. We consider libraries from the same tissue as

replicates as required by DEXSEQ. Furthermore, we only consider genes that show an expression higher than a pre-defined threshold in both conditions in our analysis (by applying a threshold on expression predicted by CUFFLINKS [167]). Additionally, we discard genes for which many of the exonic parts are predicted to be alternatively used, keeping only those genes for which the number of alternatively used exons is smaller than $\max(2, 1/4 \cdot \text{number of exons})$. The main reason for doing so it so focus our analysis of the potential interplay between alternative splicing and RNA editing on genes that are more likely to be regulated locally.

2.3 Results

2.3.1 Our pipeline accurately distinguishes genuine editing sites from SNPs, and sequencing and mapping artifacts

The set of sites predicted by our pipeline is highly enriched in A-to-G conversions. Figure 2.2.A shows the number of unique RNA editing sites identified for each of the twelve possible types of DNA/RNA differences after applying our pipeline to the combined data set comprising all 29 libraries. We find 3680 unique conversion sites in multiple tissues of *Drosophila melanogaster* of which 2879 (78.2%) correspond to A-to-G conversions. Assuming similar A-to-G and G-to-A mutation rates as well as similar rates of sequencing and mapping errors for these two types of transitions, we can estimate the false positive error rate of our predictions. Of the 3680 sites in our set, 112 of them are G-to-A conversions. By assuming that up to 112 of these A-to-G detected sites are false positive predictions, we estimate the false positive rate to be at most 3.9% ($112/2879$).

Figure 2.2.B shows the extent of overlap of the 2879 RNA editing sites identified by us and those of two other genome wide studies in *Drosophila melanogaster* by Graveley *et al* [38] and Laurent *et al* [114]. In contrast to our study, Graveley *et al* analyse RNA-seq data sets of the MODENCODE project from different developmental stages, i.e., their read samples do not overlap our tissue-specific data sets. In another high-throughput genome wide study of RNA editing in *D. melanogaster*, Laurent *et al* employ single molecule sequencing for data generation. As Figure 2.2.B shows, the overlap of sites predicted by

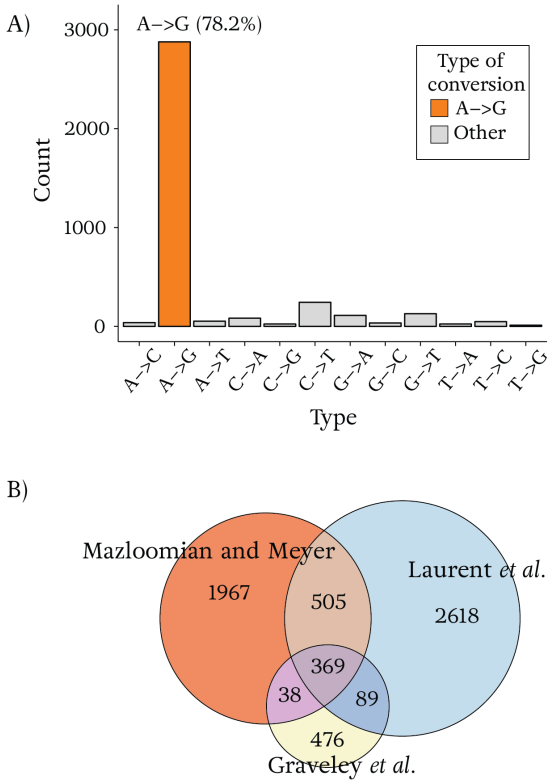

Figure 2.2: Types of identified conversions and the overlap of A-to-G conversions with other high-throughput studies. **(A)** Number of different types of conversions identified by our analysis pipeline. Most of the identified sites correspond to A-to-G conversion. **(B)** Venn diagram showing the overlap between our study and two other high-throughput studies by Graveley *et al* [38] and Laurent *et al* [114].

this and both previous studies is not very high (369/2879 (13%)), yet the overlap between our sites and each study separately, especially Laurent *et al*, is considerable (874/2879 (30%) Laurent *et al*, 407/2879 (14%) Graveley *et al*), implying that a reassuring third (912/2879 (32%)) of our RNA editing sites have been detected by either of these earlier studies, while still adding a large number (1967) of new potential RNA editing to the existing *Drosophila melanogaster* annotation.

Apart from the obvious differences in the sampled cells and the transcripts that may

be highly expressed in one cell and not the other cell, effects such as random sampling in high-throughput studies, sequencing errors and other challenges in distinguishing editing sites from artifacts [160] can account for the observed differences in detected sites. One of the key features of ADAR-derived RNA editing is that even in the same cell, the editing of two transcripts of the same gene does not necessarily involve identical RNA editing sites, but only the same double-stranded region which seems to be necessary and sufficient requirement for RNA editing to have the desired functional effect. Furthermore, differences in the proposed pipelines in these studies (a different set of thresholds, tests and engaged features) could at least partially account for some of the observed variations in results.

Among all output sites of our pipeline, 45% (1288/2879) have been predicted by at least one of four existing RNA-seq studies of RNA editing in *D. melanogaster* [38, 89, 114, 166], and at least 14% (400/2879) have been experimentally validated. In summary, the number of previously validated and identified sites combined with the low estimated false positive error rate indicates that most of our reported sites are likely to be genuine RNA editing sites.

2.3.2 Characterisation of identified RNA editing sites

As would be expected when analysing libraries of RNA-seq data (i.e., reads derived from mRNAs), most (40%, 1149/2879) our RNA editing sites occur in coding regions. Figure 2.3.A represents the distribution of RNA editing sites (which obviously derive from the transcriptome) onto different types of genomic regions. The abundance of RNA editing events in coding regions when analysing pre-mRNAs of the fly genome has been reported earlier [89]. Editing in coding regions can cause non-synonymous changes. These may alter the sequence of a protein (and possibly its length) and also change the protein's structure and function.

The next class of genomic regions with a large number of identified sites are 3' untranslated regions (3' UTRs). Editing of 3' UTRs may alter gene expression by changing nucleotides in target sequences, e.g., of miRNAs. On the other hand, binding of ADAR to a target region can also prevent miRNAs and other molecules from binding [104]. Indeed, we find that 165 of our editing sites overlap known miRNA target regions. Another

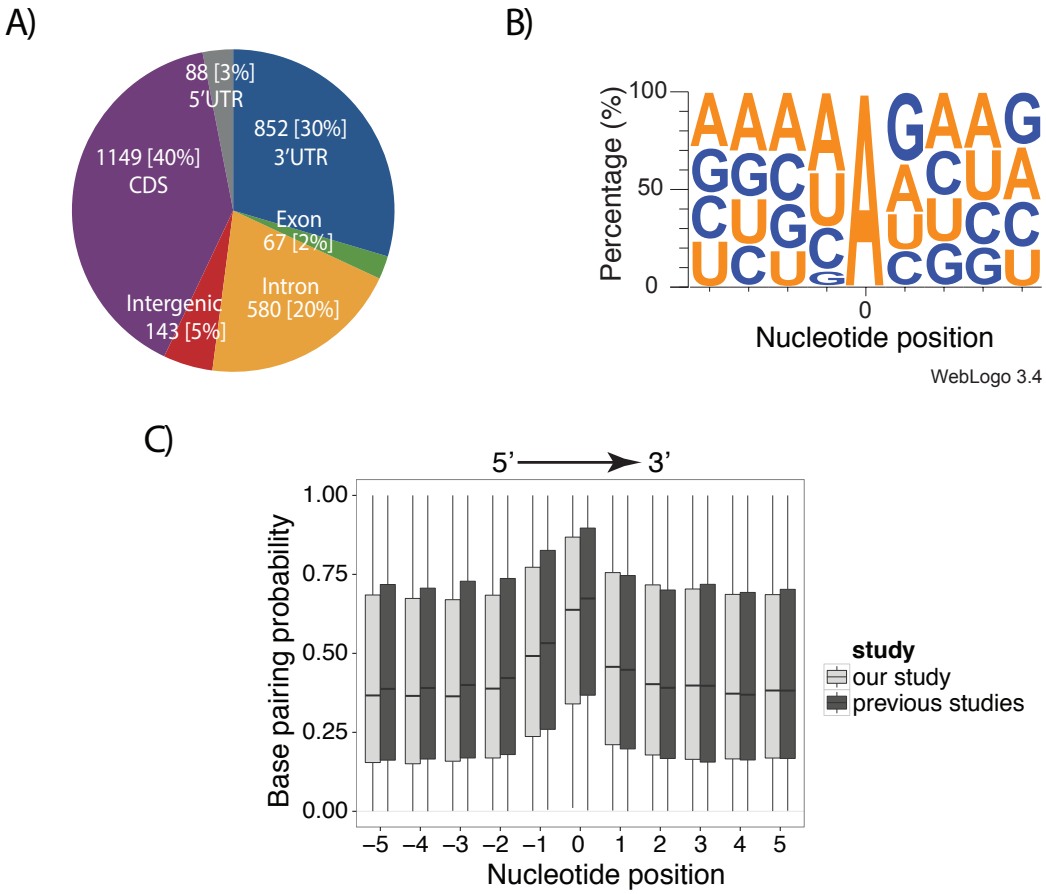

Figure 2.3: Characterisation of the identified editing sites. **(A)** Number and percentage of identified sites in different genomic regions. Coding regions contain more sites than other regions. **(B)** The frequency of each nucleotide at each position relative to the predicted editing sites. Guanosine is depleted at the exact 5' position of editing sites. **(C)** Average base pairing probabilities computed using RNAPLFOLD [168] for regions close to ADAR targets for sites predicted in our study, and previous studies [38, 89, 114, 166]. Positions -1 to 1 show higher average pairing probabilities compared to other loci. Using structural features in our pipeline may bias our predictions towards sites with higher base pairing probabilities around reported sites; however a similar pattern has also been observed when considering sites predicted in previous studies. Part **(B)** is generated using WEBLOGO [http://weblogo.berkeley.edu/logo.cgi].

mechanism for altering gene expression patterns is to directly edit the miRNA molecules themselves or by interfering with miRNA processing [169–171]. We find 6 editing sites in 4 miRNA molecules: *mir-4971* (1 site), *mir-2a-2* (2 sites), *mir-4961* (2 sites), and *mir-4956* (1 site). These miRNA editing sites have the potential to influence miRNA processing and targeting.

Although our data derives from spliced transcripts, i.e., mRNAs (polyA enriched), we find 580 editing sites (20%) in genomic regions that are annotated as being intronic. The prevalence of editing in retained introns has already been reported [89]. Editing in introns can happen when the editing site falls into an editing site complementary sequence (ECS) which forms a double-stranded region with a region in an adjacent exon [94]. RNA editing may then lead to changes in the local RNA secondary structure which may result in the exon being retained [55]. Via this molecular mechanism, RNA editing thus has the potential to alter splicing patterns by changing local RNA secondary-structure.

Our remaining sites overlap intergenic regions, 5' UTRs and exons of non-coding genes. Sites classified as intergenic may be due to an incomplete annotation of the *Drosophila melanogaster* genome. The number of editing sites in the other two classes is small, but may have interesting biological consequences.

We took advantage of the large number of predicted RNA editing sites to investigate the primary sequence and structural binding preferences of ADAR. In agreement with earlier studies [114], we find that a guanosine directly adjacent in the 5' position of an adenosine decreases the chance of the adenosine being edited (see figure 2.3.B). Analysing the estimated base-pairing probabilities of small regions around the predicted RNA editing sites using RNAPLFOLD [165], we find that the two nucleotides directly adjacent to the site are the most important to be base-paired in ADAR target regions (2.3.C).

Analysing different tissue-specific data, we find that RNA editing happens in multiple tissues of *D. melanogaster*, predominantly in head. We highlight the number of DNA/RNA mismatches for four tissues in figure 2.4. The majority of detected editing sites occur exclusively in head and central nervous system. In other tissues, RNA editing is rare. Reassuringly, we find that in heavily edited tissues most of the predicted sites are A-to-G conversions that can be attributed to ADAR activity; the false positive rate of our analysis is thus low, conversely, in other tissues the estimated error rate is higher (figure 2.5.A).

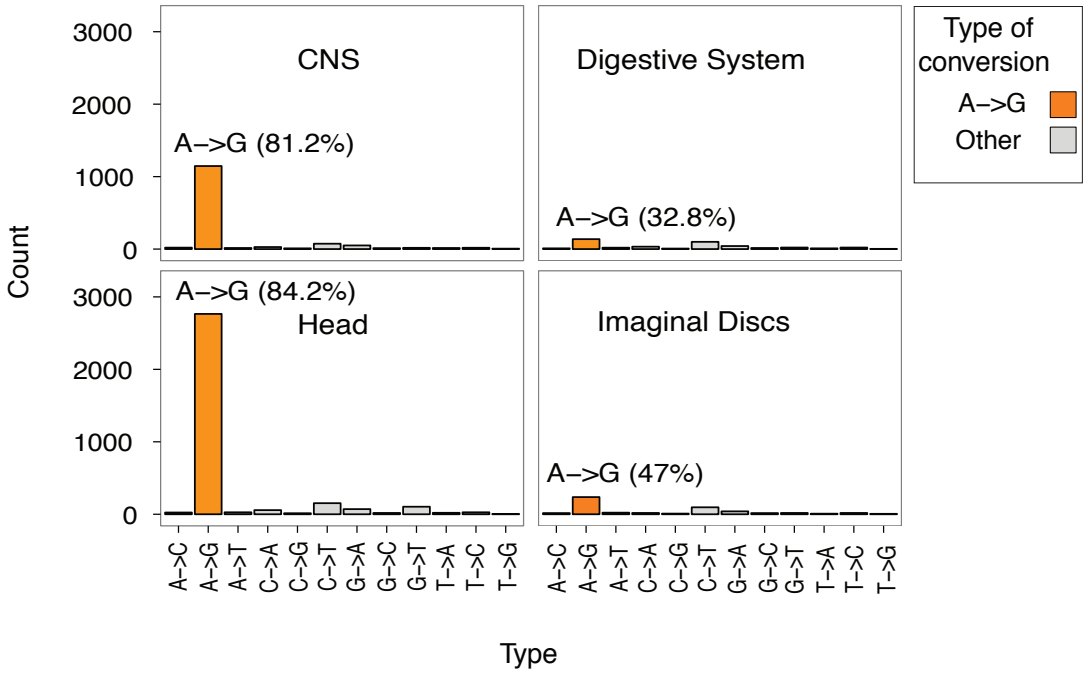

Figure 2.4: The number of all 12 types of conversions for four tissues of our study: central nervous system (CNS), digestive system, head, and imaginal discs. Head and CNS contribute most to the list of our predictions.

Editing patterns differ considerably between different types of tissues. Figure 2.5.B illustrates the relative overlap between sets of predicted sites in the ten studied tissues. Generally, different pairs of tissues do not share most of their editing events. One obvious candidate for regulating RNA editing is the expression of the ADAR gene itself. We find that ADAR expression is highest in head and central nervous system (CNS), but that the gene is also expressed in other tissues (figure 2.5.C). Over-expression of ADAR in head and CNS is in agreement with the number of detected sites in these tissues, however, a higher expression of ADAR in one tissue compared to the other, does not necessarily imply a greater level of RNA editing; thus, as suggested before [116], the level of ADAR expression alone cannot explain how RNA editing levels are regulated.

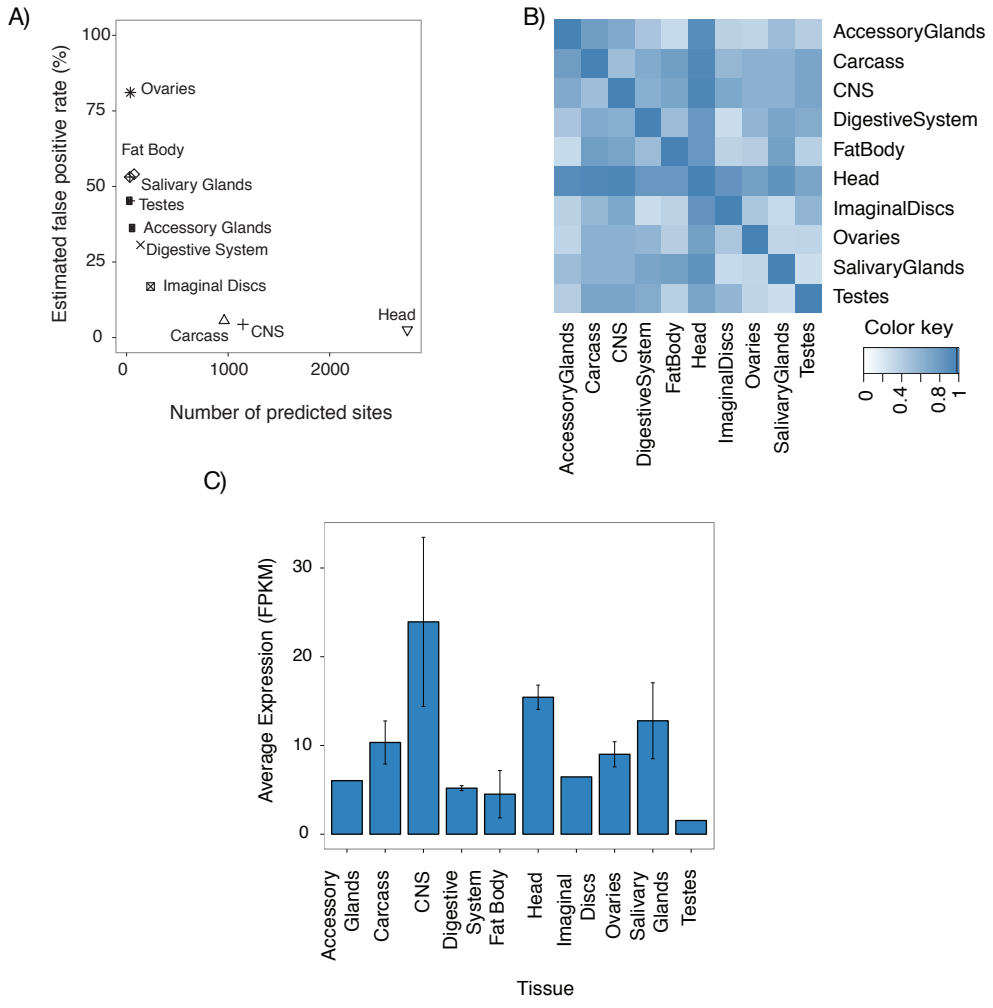

Figure 2.5: Comparing the editing mechanism in different tissues of *Drosophila melanogaster*. **(A)** Estimated error rate versus the number of predicted sites in different tissues of our study. **(B)** Percentage of overlapping sites between pairs of tissues as encoded by color shading. To compute the overlap ratio, the number of common sites between pairs of tissues is divided by the smaller number of detected sites between corresponding tissues. **(C)** Average expression of dADAR in tissues of the MODENCODE project. Expression values are measured in FPKM (fragments per kilobase of transcript per million fragments mapped) unit using CUFFLINKS [167]. Although dADAR expression is highest in CNS (central nervous system) and head, but the gene is expressed in other tissues as well.

Our functional enrichment analysis using DAVID [172] confirms that edited genes are involved in ion transport (Benjamini Hochberg (BH) adjusted p-value: $2 \cdot 10^{-13}$), gated channel activity (BH adjusted p-value: $3 \cdot 10^{-8}$) and cell-cell signalling (BH adjusted p-value: $8 \cdot 10^{-8}$), the well known functions of ADAR targets [94, 173]. Additionally, functional annotation clustering using DAVID [172] identifies a cluster of genes involved in locomotory behaviour (BH adjusted p-value: $2 \cdot 10^{-3}$) and similar genes which is in agreement with the phenotype associated with ADAR knock-down flies [106, 174].

2.3.3 Evidence for cross-regulation of RNA editing and alternative splicing and the potential underlying regulatory mechanism

As discussed in the introduction, there already exists some evidence for an inter-relation between alternative splicing and RNA editing mechanisms. Leveraging the large number of selected tissue-specific data sets used in our study, we decided to investigate the reciprocal effect between alternative splicing and RNA editing in much greater details and to discover potential underlying regulatory mechanisms. Alternative splicing and RNA editing both play key roles in diversifying gene products and in fine tuning gene expression on RNA level. It would thus be of great conceptual importance to identify potential mechanisms of their cross-regulation.

We find that a gene with a greater number of known isoforms has a higher chance of being edited. Figure 2.6.A illustrates the positive correlation ($R_{Pearson} = 0.33$, p-value $< 2 \cdot 10^{-15}$) between the number of annotated isoforms and the number of predicted RNA editing sites in our study. One would expect longer genes to have a higher probability of being edited and to also have more splice variants (based on the larger number of exons). In order to test if the correlation observed in our data can be explained by gene length *alone*, we grouped genes according to their lengths and calculated the average number of known isoforms per group, once for the sub-group of edited and once for the complementary sub-group of un-edited genes (Figure 2.6.B). Although we find that longer genes tend to contain more editing sites, edited genes have a significantly greater number of known isoforms than un-edited genes (Figure 2.6.C). Other features such as exon lengths, intron lengths, and nucleotide bias may also affect the number of editing sites in genes.

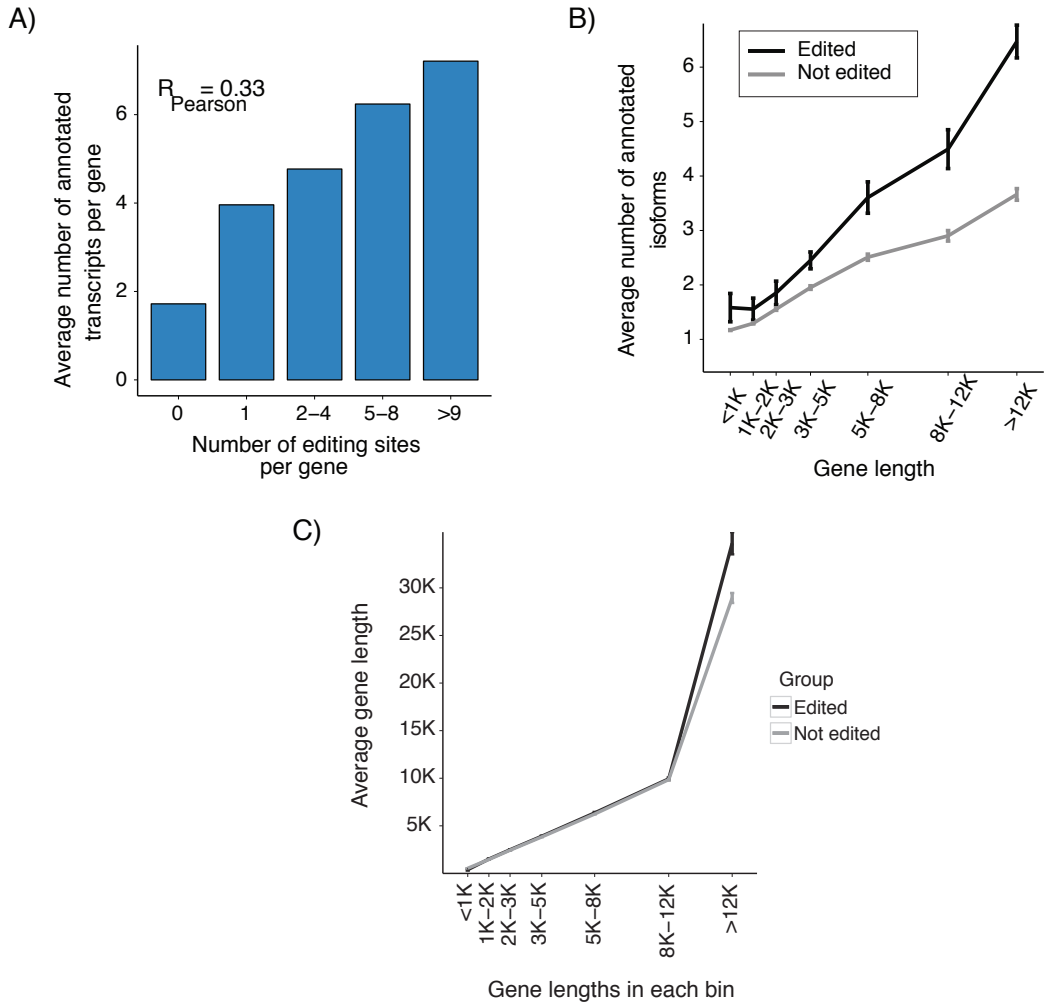

Figure 2.6: There is a positive correlation between genes that are targets of RNA editing and genes that are alternatively spliced. **(A)** The number of annotated isoforms *vs.* the number of predicted sites in our study. The number of detected sites is found to be greater in genes that express more annotated isoforms. **(B)** We group genes based on their length and compare the average number of annotated isoforms for genes of similar length between those that are edited and those that are un-edited genes. For genes with similar length, edited genes have a higher chance of being alternatively spliced. **(C)** Here we tested whether genes in the same length bins from edited and un-edited groups have similar lengths. The plot shows that for most of our bins, average gene lengths is almost equal for edited and un-edited group.

Even more interestingly, we find that editing events tend to preferentially occur near exons with multiple splicing donor/acceptor sites (χ^2 test, p-value $< 2 \cdot 10^{-15}$). For this, we classify exons (including UTR exons) into two groups, those with multiple known acceptor and/or donor sites and those with unique acceptor and donor sites. Within each group, we count the number of RNA editing sites and normalise by the combined lengths of all exons in that group. Based on the resulting numbers, RNA editing sites are 3.2 times more likely to occur in exons with multiple splicing donor/acceptor sites compared to those with unique acceptor and donor sites (χ^2 test, p-value $< 2 \cdot 10^{-15}$, this p-value is calculated for the null hypothesis of a 1:1 ratio). To further confirm our findings that are based on our set of predicted RNA editing events, we repeated the same analysis for all sites reported by four existing high-throughput studies of RNA editing in *Drosophila melanogaster* [38, 89, 114, 166] and find again that RNA editing is 1.9 times more likely to occur in exons with multiple acceptor/donor sites (χ^2 test, p-value $< 2 \cdot 10^{-15}$).

We then identified 244 regions where RNA editing and tissue-specific alternative splicing can have reciprocal effect (Appendix. A). For this, we searched for RNA editing sites in and around exons (between -150 and +150 around each exonic part) that are alternatively spliced when comparing expression for pairs of tissues using DEXSEQ [74]. Figure 2.7 shows an example of a region that is predicted to be highly edited and observed to be alternatively spliced. The figure shows that many more editing sites are predicted in the head tissue (blue arrows) compared to digestive system (red arrow). This is also true for the exonic region that is not predicted to be alternatively used (E002).

One reason for the alternative splicing of the 3' exon could be the formation of double stranded structure; or the binding of ADAR could prevent splicing machinery from detecting splicing signals and splice out the last exonic part. We should mention that the predicted editing level is low even in head tissue, and low editing level and random sampling may have caused the editing events not to be predicted in the digestive system samples. Dedicated follow-up experiments are required to understand how the two mechanisms affect each other.

To discover potential mechanisms regulating the interplay between alternative splicing and RNA editing, we also searched for statistically significant conserved RNA secondary-structure features in the vicinity of exons where we found RNA editing and alternative

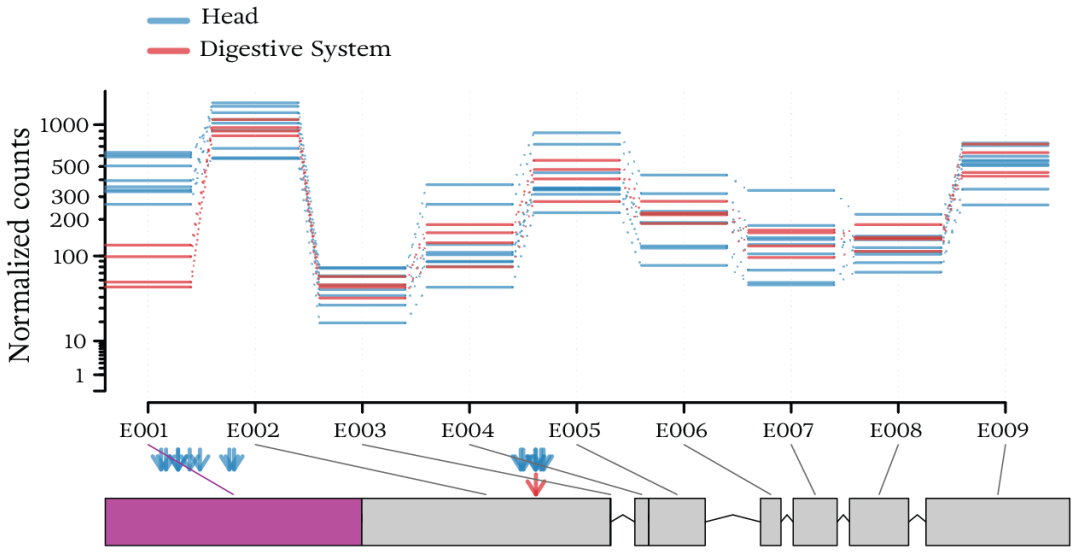

Figure 2.7: An example of a region where RNA editing and alternative splicing may affect each other. Rectangles at the bottom represent exonic parts of gene *CG5850* located on the reverse strand of the left arm of chromosome 2. Exonic parts are numbered by E001, E002, ..., E009, where E001 is the 3' most exonic part and E009 is the 5' most exonic part. The *Y* axis shows the number of reads aligned to each exonic bin, normalised by library size. Blue lines correspond to the number of reads from the libraries of the head tissue and red lines correspond to libraries from the digestive system tissue. The purple rectangle shows the rectangle that is predicted to be alternatively expressed between the two tissue types. In this region, multiple arrows are shown for identified editing sites for head (blue arrows) and digestive system (red arrows). Figure generated using DEXSEQ [74]

splicing to co-occur. For this, we employed TRANSAT [128] on input alignments of 15 fly species downloaded from UCSC [176] (We also added *OregonR* sequence to the alignment; see Appendix. A for more details) around splice sites of alternatively spliced exonic parts where editing sites are also predicted (extended by 150 nucleotides on either side, a total of 167 regions). There already exist quite a few computational methods to predict evolutionarily conserved RNA secondary-structure [132, 168, 177]. These programs, however, expect the input alignment to contain one more or less *global* secondary-structure,

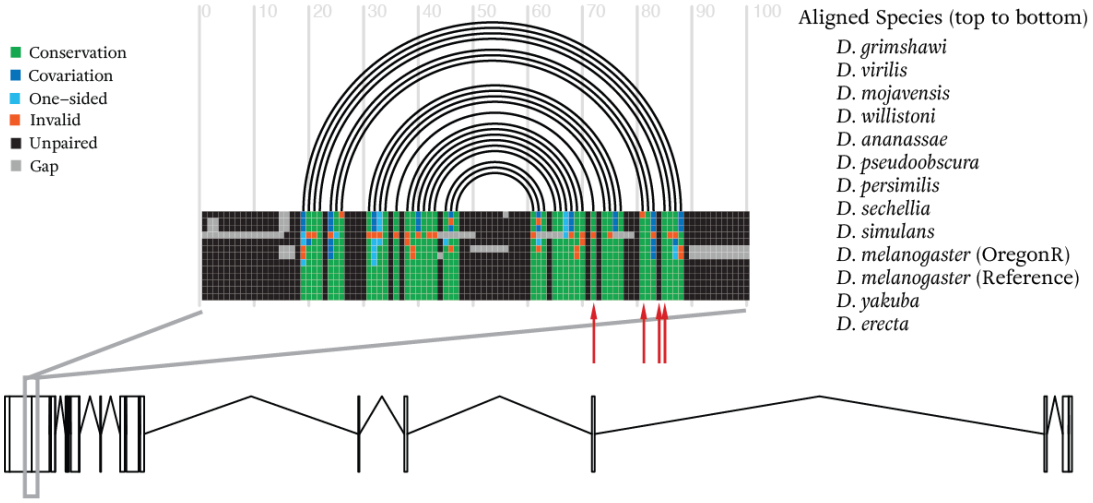

Figure 2.8: An example of a region where a conserved RNA secondary structure feature detected by investigating editing events can potentially influence alternative splicing. Rectangles at the bottom of the figure show exonic parts of *Cip4* gene located on the reverse strand of the left arm of chromosome 3. The figure shows the structure predicted using RNAALIFOLD in a region of 100 nucleotides around the splice site of an exonic region which is predicted to be alternatively used between tissues. Red arrows show predicted editing sites. Black arcs indicate alignment columns that are predicted to be base-paired, and black columns correspond to un-paired nucleotides. Green squares within the alignment show valid base-pairs and orange squares invalid base-pairs. Dark blue squares represent valid base-pairs with two-sided mutations (compared to the most common base-pair in the pair of columns), probably in order to retain base-pairing potential. Likewise, light blue colour represents single mutations to retain base-pairing potential. The existence of multiple compensatory mutations provides evidence for its functional importance throughout evolution. Figure generated using R-CHIE [175]

i.e., a structure spanning the entire alignment. As there *a priori* no reason to expect secondary-structure features relevant for RNA editing to involve the entire transcript – especially not longish fly pre-mRNAs *in vivo* – we use TRANSAT as this program has been specifically designed to identify *local*, conserved RNA secondary-structure features such as the double-stranded regions needed for ADAR binding and RNA editing. TRANSAT method takes a set of aligned sequences and an evolutionary tree as input; extracts potential helices in the alignment, and assigns a p-value to each of these helices. For 96 of the 167 regions (57%) where alternative splicing and RNA editing co-occur in our data we find one or more conserved RNA secondary-structure features (when we filter helices with p-value greater than 0.05 and helices shorter than 8 nucleotides). Figure 2.8 shows an example of these regions and the corresponding, conserved RNA secondary-structure detected by RNAALIFOLD [168] in this region. Multiple compensatory mutations for conserved base-pairs provide evolutionary evidence for a likely functional role of this double-stranded region. The list of the identified regions is presented in Appendix. A. Finally, we applied RNAALIFOLD to assess the stability of the global structures in these regions. The list of the identified regions sorted based on the energy of the predicted global structure by RNAALIFOLD can be found in Appendix. A.

2.4 Discussion

We identify 2879 A-to-I RNA editing sites in different tissues of *Drosophila melanogaster* with high precision. More than half of these have not been identified previously. The high ratio of A-to-G conversion type among the detected DNA/RNA discrepancies shows that most of our predictions are anticipated to be true editing events and not the result of experimental or computational artifacts. Also, our study suggests that other types of possible RNA editing apart from A-to-I RNA editing are very rare or do not happen at all in the investigated tissues of *D. melanogaster*.

Furthermore, our results show that editing occurs in multiple tissues, with many of the sites being edited exclusively in brain and central nervous system where ADAR expression is also higher than in other tissues. Moreover, patterns of editing differ significantly between tissues, implying a tissue-specific underlying regulatory mechanism.

Our study demonstrates how the appropriate use of ADAR specific features enhances the detection of RNA editing events when DNA reads are not available. A previous study by Ramaswami *et al* [166] shows that evolutionary information can be used to detect editing sites in the absence of DNA reads. Here, we explicitly capture ADAR specific features - in particular the requirement for the formation of local RNA secondary structures around target sites and clustering of editing sites - in addition to utilising large number of selected data sets to distinguish editing events from artifacts and SNPs.

We identify more than 200 regions exist where RNA editing and alternative exon usage between tissues co-occur when comparing libraries. Many of the identified regions have been identified in multiple pair-wise comparison of tissues. Studies showed the co-occurrence of RNA editing and alternative splicing in same genes [114, 117], similar to what we find in this analysis. Solomon *et al* reported the enrichment of editing events in cassette exons in human, although they reported most of the sites are far from exon boundaries. We here show that editing events tend to happen much more abundantly in exons with multiple known acceptor or donor sites, or 3' and 5' UTRs that contain alternative splicing potential. Further, we find 96 regions around splice sites with significant statistical evidence for the overlap of evolutionarily conserved, local RNA secondary-structures. The actual formation of these RNA structure features *in vivo* is supported by both computational RNA secondary-structure prediction programs and predicted RNA editing sites.

RNA editing thus has the potential to regulate alternative splicing via changes of local RNA secondary structures. This suggests a potential, tissue-specific molecular mechanism of regulation for alternative splicing whose potential mediation via changes of local RNA structure we showed earlier [55].

Overall, we find strong evidence for our hypothesis that RNA editing and alternative splicing mechanisms directly influence each other in specific regions of the transcriptome. Both, RNA editing and alternative splicing are abundant in the CNS and are both known to be temporally and spatially regulated [87]. Also, target genes of the two mechanisms correlate well. These mechanisms may influence each other in several ways. First, the splicing machinery may compete with ADAR for common substrates. This is plausible given that RNA editing and splicing can happen at the same time co-transcriptionally in *Drosophila melanogaster* [89, 178]. Targeting of a specific location by one machinery lim-

its the simultaneous access of the other machinery and can thereby affect its functionality. Second, considering the potential importance of RNA secondary structures in regulating alternative splicing, ADAR may edit and thereby alter local secondary structures which can in turn change exon usage. Blow *et al* showed earlier that RNA editing of double-stranded regions has the overall effect of destabilising these features. Finally, editing of splicing silencers and enhancers or splice site motifs could additionally affect splicing.

Based on our results, the type of local co-regulation through changing RNA structures happens predominantly within exons with multiple acceptor or donor sites. In these regions, the primary sequence splicing signals may be weak, and these weak signals can prevent the splicing machinery from always making the same decision.

The formation and RNA editing-mediated modification of local RNA secondary structures therefore has the potential to significantly alter splicing patterns in these genes as local RNA structure features can be "encoded" in a transcript-specific way. In fact, the necessity for encoding RNA structure features that are involved in regulating the alternative splicing of their own transcript may explain why introns tend to be longer in more complex organisms: these RNA structure features are (at least partly) encoded in introns thus imposing no undue additional evolutionary constraints on the protein-coding exons.

Previous studies suggest that the dominant way in which editing regulates splicing is by editing RNA-binding proteins [117]. This would, however, imply a more indirect and global way of regulating alternative splicing and could not easily happen in a gene-specific way. Our results support a gene-specific mechanism where alternative splicing can be directly regulated via tissue-specific changes of RNA editing. Also, one of the roles of pre-mRNA sequences may be to not only encode amino-acid information, but also RNA secondary-structure motifs that determine the correct splicing patterns in a tissue-specific way. Detailed follow-up experiments, e.g., ADAR knockdowns and mutational studies of specific genes, are now required to experimentally confirm our results.

Chapter 3

The Regulation of Alternative Last Exon Splicing by *CDK12* Promotes the Oncogenic Potential of Breast Cancer Cells

3.1 Introduction

Cyclin-dependent kinases (CDKs) and their activating cyclin partners integrate numerous signal transduction pathways to regulate a variety of critical cellular processes [138, 179]. *CDK12* (*CRK7*, *CrkRS*) is one of several CDKs that regulate transcription through the differential phosphorylation of the C-terminal domain (CTD) of *RNA Polymerase II* [137]) as discussed in Chapter 1. There is still much unkonwn regarding how *CDK12* regulates alternative splicing and gene expression at a genome-wide scale.

The Cancer Genome Atlas (TCGA) project identified recurrent somatic alterations in *CDK12* in 13% of breast cancers and 5% of ovarian cancers [153, 180–182]. *CDK12* mutations are commonly nonsense mutations or impair *CDK12* kinase activity [183], and are frequently coupled with loss of heterozygosity [180, 184]. Recent studies show that *CDK12* functions in maintaining genome stability. In cell-based assays, depletion of

CDK12 is associated with defects in DNA damage response (DDR) and decreases expression of genes involved in the homology-directed repair (HDR) pathway [146, 151, 152, 183, 185]. Though it is generally classified as a tumor suppressor gene from its role in DDR, additional evidence indicates that *CDK12* may have pro-oncogenic functions in breast cancers. *CDK12* is located on chromosome 17, 165-267 kb proximal to *HER2* (*ERBB2*), an oncogene that is frequently amplified in breast cancers. *CDK12* is co-amplified with *HER2* in 27-92% of breast tumors or tumor cell lines [186–194]. Similar to *HER2*, over expression of *CDK12* also correlates with high proliferative index and grade 3 tumor status based on tissue microarrays of invasive breast carcinomas [134]. It is noteworthy that in about 13% of *HER2*$^+$ (*HER2*-amplified) breast tumors, the amplification breakpoint resides in the *CDK12* allele and likely results in the functional loss of one *CDK12* allele [185]. In related observations, recurrent *CDK12-HER2* gene fusions in gastric cancers result in impaired *CDK12* protein levels [195]. It is currently unknown how alterations in *CDK12* contribute to the myriad of changes seen in breast tumors. Overall, these data suggest *CDK12* may have oncogenic roles in cancer progression, but the mechanisms underlying this effect have not been explored.

To address the oncogenic roles of *CDK12*, we performed a comprehensive and systematic genomic and proteomic analysis of *CDK12* function in a breast cancer cell line with genomic amplification of *CDK12*. We sought to determine if the role of *CDK12* in tumorigenesis and DDR was related to its hypothesized ability to regulate splicing or AS in addition to its role in transcription. Instead of having a general effect on transcription or splicing, we found that *CDK12* regulated the expression and AS of a distinct set of mRNAs in a cell type-specific manner. Furthermore, *CDK12* predominantly regulated only the alternative last exon (ALE) sub-type of AS. Functionally, events regulated by *CDK12* potentiated tumorigenic processes, indicating that aberrant *CDK12* expression can have oncogenic properties.

3.2 Materials and methods

3.2.1 Data

The RNA-seq data consists of biological triplicates of SK-BR-3 and 184-hTERT cells treated with *CDK12* siRNA-1 or scrambled siRNA. The libraries contain on average 103 $\pm$ 12 million paired end 75 nucleotides strand-specific reads (mean $\pm$ s.d.). The paired-end reads were aligned to the reference genome (*hg19* reference genome downloaded from UCSC genome browser [196]) using GSNAP [197]. The corresponding gene annotation file was downloaded from ENSEMBL [159]. The "novel splicing" parameter of GSNAP was enabled to allow the discovery and use of novel junctions in the alignment step. In the final step, duplicate reads were removed using SAMTOOLS [162, 198]. The procedure resulted in an average of $\sim$92% successfully aligned reads.

To assess our findings in an independent data set, we downloaded the RNA-seq data published in a previous study [145]. The data contain two control and two *CDK12* shRNAs in two replicates from HCT-116 cells. These libraries consisted of $\sim$14 million to $\sim$48 million single end un-stranded 50 nucleotides reads. On average 84% of the reads were successfully aligned to the *hg19* reference genome using TOPHAT2 [163]. Because of the small number and short lengths of these reads, duplicate reads were kept in the aligned files and reads with mapping quality of less than 10 were removed.

3.2.2 Differential gene expression and alternative splicing analysis

DESEQ2 [199] was applied to detect genes that were differentially expressed in *CDK12* siRNA-treated libraries as compared to control libraries. Given a table of raw read counts for genes in the genome, DESEQ2 applies a statistical model to compare counts between the two conditions, and it calculates a fold change for each gene and assigns a statistical measure of confidence for differential regulation of the gene. The gene read counts required as input by DESEQ2 was provided using HTSEQ-COUNT [200] by setting "mode" parameter to "union". Genes with adjusted *p-values* < 0.01 form the list of confident differentially expressed genes between siRNA-treated and control conditions in each cell line. CUFFLINKS [201] was also used to quantify gene and isoform expressions.

Given a ranked list of genes, GSEA [202] applies a statistical method to identify pathways for which the genes involved in that pathway are over-represented at the top or bottom of the ranked list. For the RNA-seq data, genes were filtered for having very few ($<$100 on average) reads aligned to them and the remaining genes ($\sim$12,000 in SK-BR-3 and 184-hTERT cell lines) were sorted based on the estimated fold changes. For the global proteome data, the sign of the fold change was multiplied by the inverse of the FDR rate and the genes were sorted based on the corresponding values. Here, FDR values represented the statistical significance of evidence for the differential expression of proteins. The GSEA pre-ranked analysis assigned a normalized enrichment score (NES) representing the extent of over-representation of genes of a pathway at the top or bottom of the ranked list. All of the 1,454 GO (Gene Ontology) gene sets from the Molecular Signature Database (MSigDB) [202] were used. Gene sets having fewer that 15 or more than 500 genes common to each list were filtered out. GSEA was applied in classic enrichment statistics mode with 1,000 permutations.

The ENRICHMENTMAP plugin [203] in CYTOSCAPE [204] was used to make enrichment map plots. A *p-value* cut-off to 0.005 and FDR *q-value* cut-off of 0.01 was applied. These are output *p-values* and FDR *q-values* from the GSEA analysis that represent the statistical significance of evidence for pathways being enriched for in the top up-regulated or top down-regulated genes. The clustering feature of ENRICHMENTMAP with default parameters was used to cluster gene sets that share common genes, and cluster names were manually curated based on the contained pathways.

The MISO package was used to investigate the regulation of alternative splicing (AS) by *CDK12*. The MISO package [73] applies a statistical framework to distinguish eight different types of annotated AS and processing events. These events are skipped exons, mutually exclusive exons, retained introns, alternative 3' and 5' splice sites, alternative first and last exons, and tandem 3' UTRs. The method takes a pair of samples as input and reports a $\Delta\Psi$ value between the two samples for each annotated event. The Ψ (Percent Spliced In) value represents the fraction of inclusion of one isoform when two isoforms are being considered in an splicing event (a value between 0 and 1). The method also reports a Bayes Factor (BF) value to quantify the support for the model where Ψ value is altered between the two samples compared to the alternative model of no difference in Ψ

value between the two samples.

For each cell line, siRNA-treated and control samples were randomly paired and events with BF values ≥ 20 and $|\Delta\Psi|$ values ≥ 0.1 were selected. To form the most confident set of detected splicing events, events were required to have been predicted in all three pairwise comparisons for each cell line (SK-BR-3 and 184 hTERT). To compare the MISO analyses in SK-BR-3 and 184-hTERT cells with an independent data set previously published [145], a smaller BF threshold value of 10 was applied to allow the discovery of more events (considering the smaller number of reads present in the Liang *et al* [145] data).

3.2.3 TCGA data analysis

High-grade serous ovarian cancer data generated by the TCGA Research Network (http://cancergenome.nih.gov/) were analyzed for AS events. A total of 70 raw fastq case files were downloaded from the Cancer Genome Hub (CGHub) repository (https://cghub.ucsc.edu); 14 cases have *CDK12* alterations, including 7 cases with point mutations, 3 cases with homozygous deletions, and 4 cases with amplifications (Appendix. B). The remaining 56 cases were control tumor samples with no reported alterations in *CDK12*. Additionally, these control samples did not have alterations in *CDK13*, *BRCA1*, *BRCA2*, *PALB2*, and *BRIP1* genes. These criteria were applied to exclude tumors with alterations that may potentially phenocopy the effects of *CDK12* alterations.

Fastq files were aligned to the *hg19* reference genome with the same gene annotation file and parameters used for datasets from SK-BR-3 and 184-hTERT cell lines using GSNAP. MISO was used for pairwise comparisons between two different datasets (e.g. *CDK12*-mutated cases vs. cases with no *CDK12* alterations). The analysis of AS events in the TCGA data was restricted to the union set of ALE events detected in 184-hTERT and SK-BR-3. This list comprised of 133 ALE events (predicted in all 3 pairwise comparisons for each cell line with BF > 20 and $|\Delta\Psi| > 0.1$) with 23 events common between the two cell lines.

To analyze the effects of *CDK12* point mutations, each of the 7 *CDK12*-mutated cases was paired with 2 random unique control cases (without *CDK12* alterations), generating 14 total comparisons for MISO analysis. The number of times that each of the 133 ALE

events were detected in these 14 comparisons (with BF $\geq$ 20) was calculated. As a control, 7 other random cases without *CDK12* alteration were selected and paired with the same 14 control samples. The control experiment was replicated three times to determine if the identified ALE events were over-represented when comparing *CDK12*-altered cases with other random cases. *P-values* were calculated using the Mann-Whitney U test. Similar analyses were performed for cases with *CDK12* amplifications and homozygous deletions. Due to the smaller number of *CDK12*-amplified and -deleted cases, each sample was randomly paired with 4 control cases rather than 2.

3.2.4 Motif analysis

The 3'UTRs (3' untranslated regions) of ALE (alternative last exon) events regulated by *CDK12* were searched to identify polyadenylation motifs based on published Position Weight Matrices (PWMs) [205]. The number of identified ALE events in SK-BR-3 cells was small; therefore, to expand the list and include more potential targets in the motif analysis, all ALE events that were identified in at least 2 of the 3 pairwise comparisons when using the thresholds of $|\Delta\Psi| > 0.1$ and BF > 10 were included. For each MISO ALE event representing two isoforms, the best two overlapping isoforms from ENSEMBL gene annotations that explained the ALE event in the RNA-seq data was determined. This was done by taking into account: isoform expressions computed by CUFFLINKS, the overlap between the MISO last exon and the ENSEMBL annotated isoform, and the percentage of exon-exon junctions from each candidate Ensembl isoform that were verified in RNA-seq data. The positive samples comprised all the genes for which ALE events were predicted in SK-BR-3 cell lines. These genes were divided into two groups: genes with over-expressed proximal last exons and genes with over-expressed distal last exons after *CDK12* depletion. Negative samples contained genes that were annotated to have ALE events in MISO annotations, but were not predicted to be regulated by *CDK12*. Genes for which the total FPKM expression value of the two isoforms was smaller than 0.5, were filtered out. Positive and negative samples were split into bins of similar UTR lengths and for each positive sample UTR, 20 negative samples from the same UTR length bin were randomly selected.

To count motif abundance, background nucleotide frequencies in the extracted regions

were determined, and then using the PWM for each motif, the log odds score of a sequence being a motif hit compared to being randomly generated according to background frequencies was calculated. The sequence with the maximum of these log odds scores for each motif was identified. All the sequences for which the computed score was above 80 percent of the maximum score were counted as motif hits. For the count-based motif analysis, the number of hits in each region was normalized by the length of the region, and the Mann-Whitney U test was used to compare the normalized hits in positive samples to the normalized hits in negative samples. The Benjamini-Hochberg correction was used to correct the calculated *p-values* for multiple testing. For the distance-based motif analysis, the distance of each motif hit to the 5' and to the 3' end of the 3'UTR was calculated. The significance (Benjamini-Hochberg corrected) of the difference between the calculated distances for hits in the positive samples and negative samples was calculated.

3.3 Results

3.3.1 *CDK12* regulates alternative last exon splicing of genes with long transcript and many exons

To explore the function of *CDK12* in splicing regulation, we performed mRNA sequencing (RNA-seq) on SK-BR-3 cells treated with a scrambled siRNA control or siRNA directed to *CDK12* (achieving 8- and 7-fold reduction in *CDK12* mRNA and protein, respectively). SK-BR-3 cells are a *HER*2^+ epithelial breast cancer cell line where *CDK12* is co-amplified with *HER2*. As a result, SK-BR-3 cells over-express *CDK12* protein [185]. We also performed RNA-seq on *CDK12* siRNA-treated 184-hTERT cells, an immortalized normal mammary epithelial cell line that does not over-express *CDK12*. In our RNA-seq libraries, the transcriptome was deeply sequenced (103 ± 12 million reads per sample) in order to enable the identification of low level alternative splicing events.

To find differentially spliced events, we used the MISO package [73], which applies a statistical framework to distinguish eight different types of annotated AS events in pairwise RNA-seq comparisons. From three independent pairs of *CDK12* siRNA:scrambled siRNA samples, we identified 102 AS events common to all SK-BR-3 samples and 86 AS events

common to all 184 hTERT samples (Figure 3.1.A). The regulation of specific AS events by *CDK12* was cell type-specific and only 23 AS events were common to both datasets (Figure 3.1.B). However, the mechanism of regulation appears conserved: 86% and 79% of AS events observed in *CDK12*-depleted SK-BR-3 and 184-hTERT cells, respectively, were alternative last exon (ALE) splicing. Moreover, all 23 AS events common to both cell lines were ALE events. ALE events regulated by *CDK12* had an average MISO $|\Delta\Psi|$ value of 0.23 ± 0.09 (Figure 3.1.C).

The cell type-specific effects we observed (Figure 3.2.A) are likely not an indirect result of low gene expression in either cell type; genes that were regulated by *CDK12* in only one cell type were similarly expressed in the other cell type (Figure 3.2.B). Genes with ALE events regulated by *CDK12* were expressed with an average FPKM (fragments per kb of exon per million fragments mapped) value of 12 and 15 in SK-BR-3 and 184-hTERT cells, respectively. For the 23 genes common to both cell lines, the average FPKM value was 15 and 16 in SK-BR-3 and 184-hTERT cells, respectively. Genes with SK-BR-3-specific ALEs had an average FPKM of 15 in 184-hTERT cells, and genes with 184-hTERT-specific ALEs had an average FPKM of 16 in SK-BR-3 cells.

To further explore the universality of this type of regulation, we performed MISO analysis on published RNA-seq data of HCT-116 cells (derived from colorectal cancer) treated with *CDK12* shRNAs [145]. The experiments in HCT-116 were performed in duplicates with two different shRNA constructs. Consistent with our findings in SK-BR-3 and 184-hTERT cells, ALE events accounted for 33% and 41% of all AS types in HCT-116 cells for each of the two shRNAs, respectively (Figure 3.3.A). Common AS events resulting from treatment with *CDK12* siRNA-1 (SK-BR-3 and 184-hTERT cells) and either of the two shRNAs (HCT-116) were all ALEs (n = 9, Figure 3.3.B).

The regulation of AS by *CDK12* is largely cell type-specific, but the preponderance of ALE events suggests the regulated genes may possess a common feature. When compared to the total set of protein coding genes, genes whose ALEs were regulated by *CDK12* had significantly longer transcripts and contained a greater number of exons (Figure 3.4.A). It was previously reported that genes transcriptionally regulated by *CDK12* generally had longer transcripts [146]. In our analysis, we found that genes with ALE events regulated by *CDK12* were significantly longer than those transcriptionally regulated by *CDK12* (Fig-

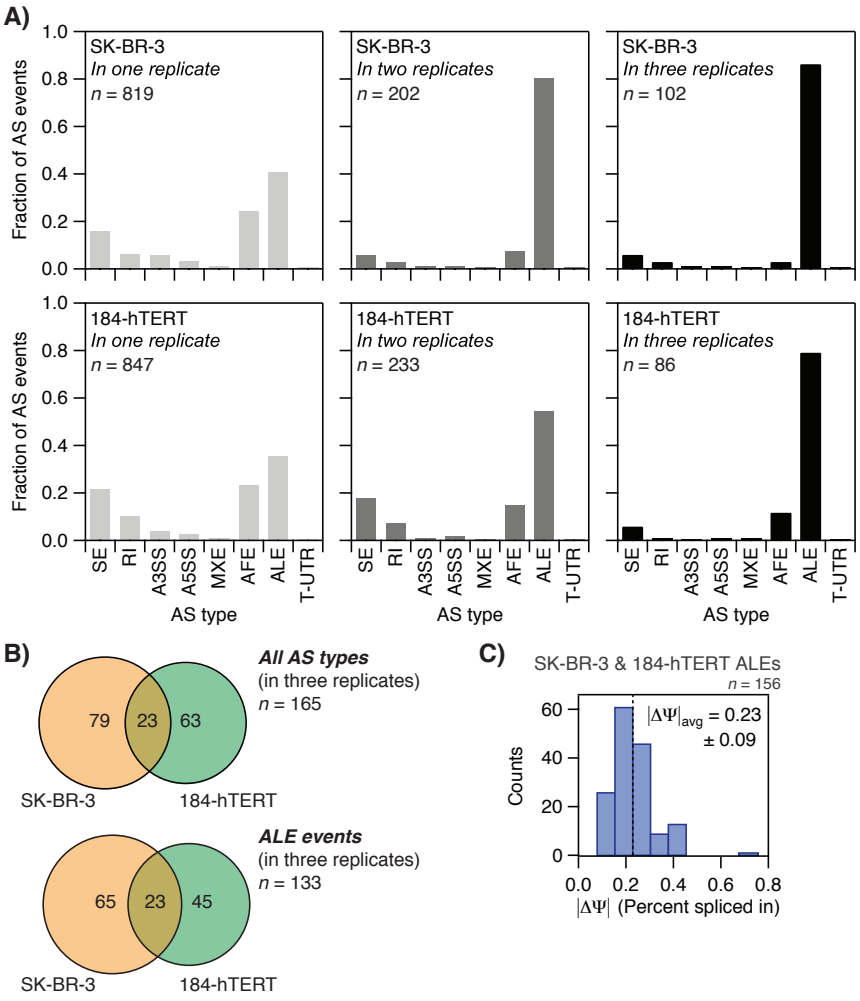

Figure 3.1: *CDK12* regulates alternative last exon (ALE) splicing. **A.** MISO analysis identified AS events that resulted from depletion of *CDK12* in SK-BR-3 and 184-hTERT cells (Bayes Factor ≥ 20, $|\Delta\Psi| \geq 0.1$). AS events present in all three RNA-seq replicates were primarily alterative last exon splicing in both cell types. SE, skipped exons; RI, retained introns; A3SS, alternative 3' splice sites; A5SS, alternative 5' splice sites; MXE, mutually exclusive exons; AFE, alternative first exons; ALE, alternative last exons; T-UTR, tandem 3' untranslated regions. **B.** The majority of AS events are cell type-specific, and events common to both SK-BR-3 and 184-hTERT cells are all ALEs. **C.** Distribution of $|\Delta\Psi|$ values for ALE events (total n = 156) regulated by *CDK12* in SK-BR-3 (n = 88) and 184 hTERT (n = 68) cells.

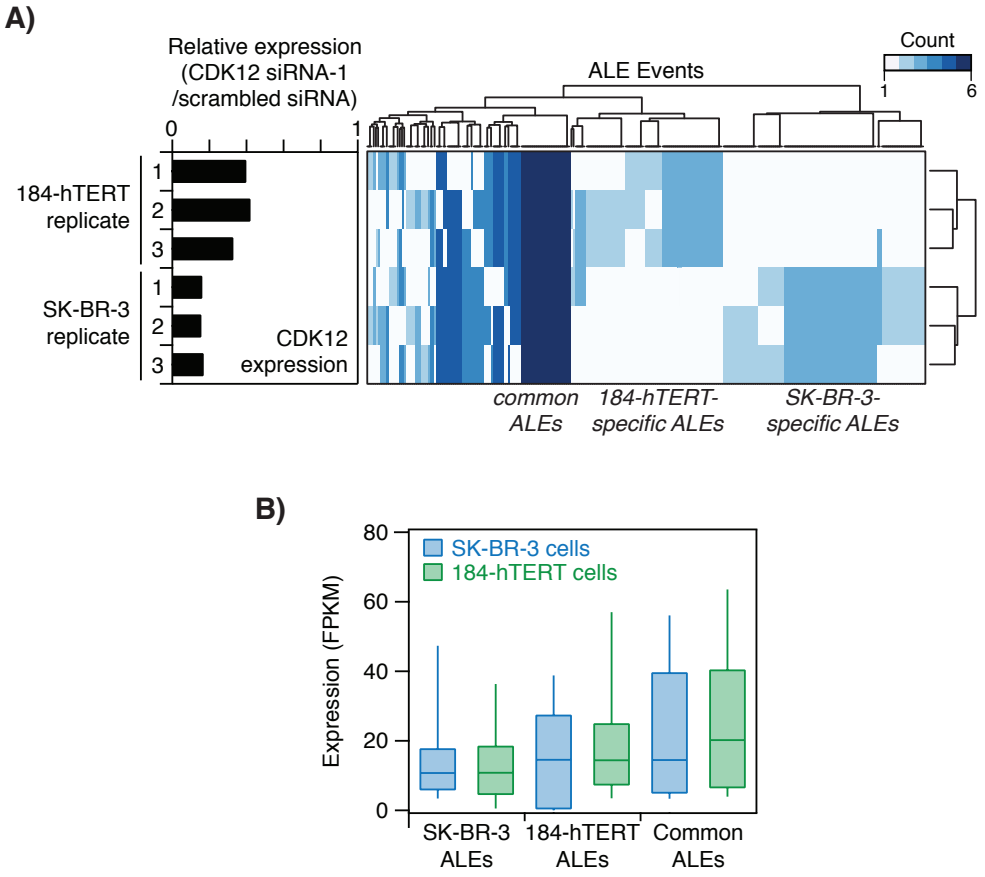

Figure 3.2: ALE regulation by *CDK12* is cell type-sepecific. **A.** *CDK12*-regulated alternative last exon (ALE) events are both common and unique between SK-BR-3 and 184-hTERT cells. Biological replicates assist in classifying cell type-specific and common ALE events. For this analysis, a lower statistical threshold was applied (Bayes Factor ≥ 10, $|\Delta\Psi| \geq 0.1$) to increase the number of ALE events for the heat map. **B.** Distribution of FPKM values in SK-BR-3 cells (blue boxes) or 184-hTERT cells (green boxes) for genes with *CDK12*-regulated ALE events specific to SK-BR-3 cells, specific to 184-hTERT cells, or common to both cell types.

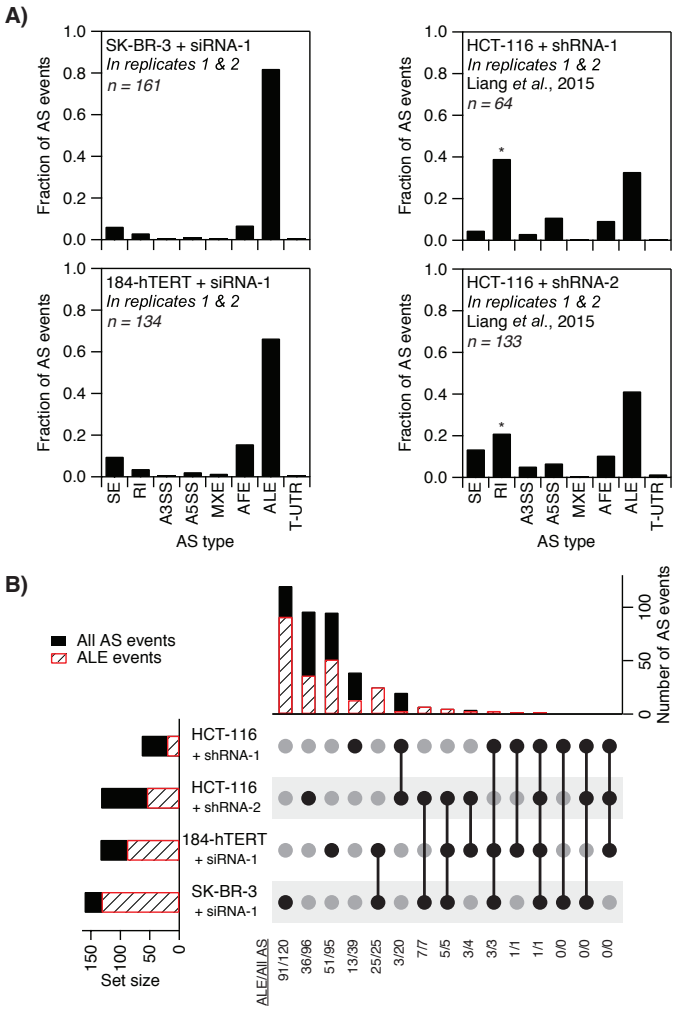

Figure 3.3: Regulation of ALE splicing is a universal function of *CDK12*. **A.** Comparison of AS events identified by MISO in SK-BR-3 and 184-hTERT cells (treated with *CDK12* siRNA-1, this study), and HCT-116 cells (treated with two shRNA constructs, Liang et al [145]). Only two replicates of the SK-BR-3 and 184-hTERT RNA-seq were used in order to match the conditions of the HCT-116 RNA-seq data. In the HCT-116 experiment RNA-seq was performed on total RNA after depletion of rRNA. The RNA was not enriched for mRNA, which could explain the enrichment of retained introns (denoted by asterisks) observed in the HCT-116 data versus the SK-BR-3 and 184-hTERT data. **B.** Intersection set analysis showing number of AS and ALE events common to SK-BR-3, 184-hTERT, and HCT-116 cells. Top: set sizes of each group are shown. Bottom: numbers of AS and ALE events in each intersection group. Graph created using the UpSetR package in R (https://cran.r-project.org/web/packages/UpSetR/).

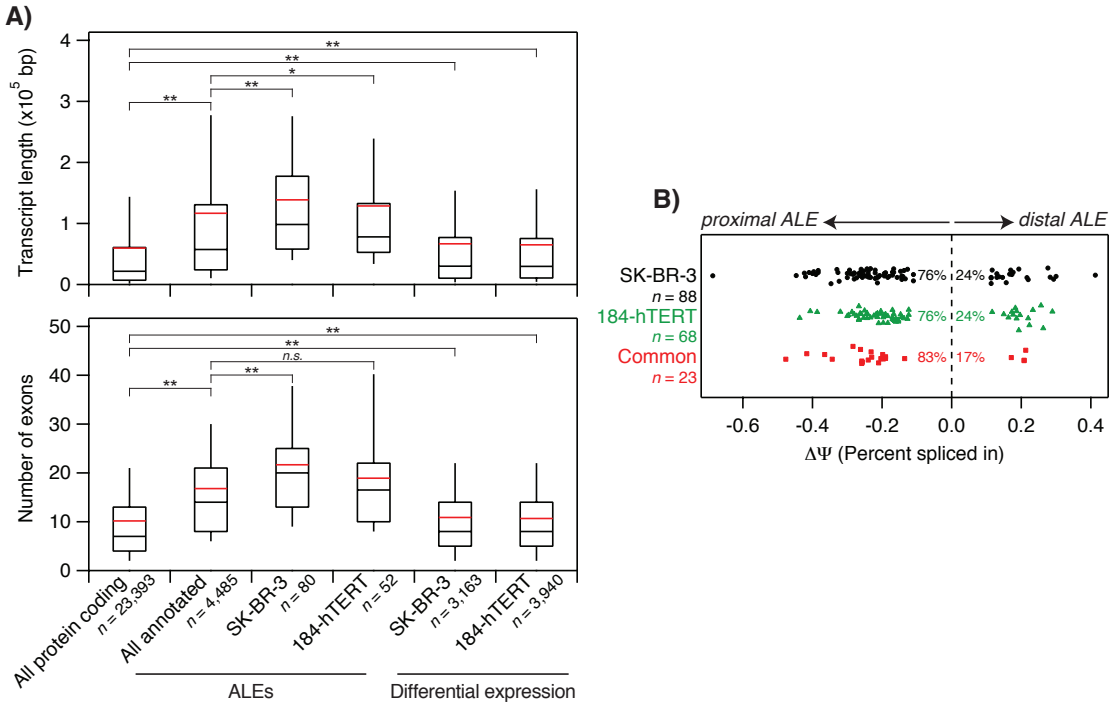

Figure 3.4: *CDK12* regulates ALE splicing of genes with long transcripts and a large number of exons. **A.** Distributions of gene transcript length and number of exons. All protein coding genes are compared to all genes with annotated ALE events and genes regulated by *CDK12* (ALE splicing or differential expression in SK-BR-3 and 184-hTERT cells). Red lines represent the means. Pairwise statistical comparisons performed using the Kolmogorov-Smirnov test (*p $<$ 0.0005, **p $<$ 1 x 10-6, n.s. not significant). **C.** Depletion of *CDK12* generally results in the utilization of proximal ALEs (negative $\Delta\Psi$ values).

ure 3.4.A). Genes with ALE events regulated by *CDK12* were also longer than all genes with annotated ALEs. Notably, of all genes with annotated ALEs, only 3% with transcripts longer than the average were regulated by *CDK12* in SK-BR-3 or 184-hTERT cells. In other words, only a subset of long genes was regulated by *CDK12*-dependent ALE splicing, suggesting additional gene-specific factors that direct AS by *CDK12*.

In 76% of ALE events, *CDK12* depletion resulted in the enrichment of mRNA isoforms utilizing the proximal ALE (Figure 3.4.B). When considering only ALE events common

to both SK-BR-3 and 184 hTERT cells, the proximal ALE was utilized more in 83% of the cases. These results were independently validated by performing qRT-PCR (by Christalle Chow and Jerry Tien) on a select number of ALE events in SK-BR-3 and 184-hTERT cells depleted of *CDK12*, with good correlation of $\Delta\Psi$ values between the MISO and qRT-PCR data (Appendix. B). These observations were also not due to off-target effects; we obtained similar results with a different *CDK12* siRNA construct (*CDK12* siRNA-2; Appendix. B), but not with siRNA constructs targeting *CDK9* or *CDK13*.

Furthermore, the immunoprecipitation experiments carried (by Christopher S. Hughes and Jerry Tien) determined that *CDK12* interactome is enriched in spliceosomal proteins. *CDK12*-interacting proteins were highly enriched for RNA splicing function (Figure 3.5), and could be generally classified into core spliceosome components (pre-catalytic complexes A and B, and the associated Prp19 complex) and regulators of constitutive and alternative splicing (SR proteins, RBM proteins, and hnRNPs) (Figure 3.5.C) [206]. The interactions between *CDK12* and hnRNPs were sensitive to nuclease treatment and were therefore likely dependent on RNA intermediates, such as the pre-mRNA upon which hnRNPs are assembled. By contrast, interactions between *CDK12* and core spliceosome and SR proteins were largely unaffected by nuclease treatment. The universality of interactions between *CDK12* and core spliceosome components was further supported by immunoprecipitation experiments in HEK-293T cells [145, 207], Jurkat T-cells [208], and HeLa cells [148, 209]; however, many of the regulatory splicing components differ across cell types. This could be a product of cell type-specific regulation or differences in experimental methodology. Together, these results suggest that *CDK12* is a bona fide component of the splicing machinery.

While the regulation of ALE usage by *CDK12* can be achieved through its association with regulatory splicing factors, it could also be an indirect product of transcription termination processes (such as alternative polyadenylation) initiated by termination signals in the 3' untranslated regions (UTRs) [211]. To address this possibility, we searched for polyadenylation motifs in the 3'UTRs of proximal and distal ALEs that were regulated by *CDK12* (Figure 3.6). We observed no differences in the distribution and density of polyadenylation motifs in ALEs regulated by *CDK12*, as compared to ALEs unaffected by *CDK12* function. This observation further suggests that the regulation of ALE usage

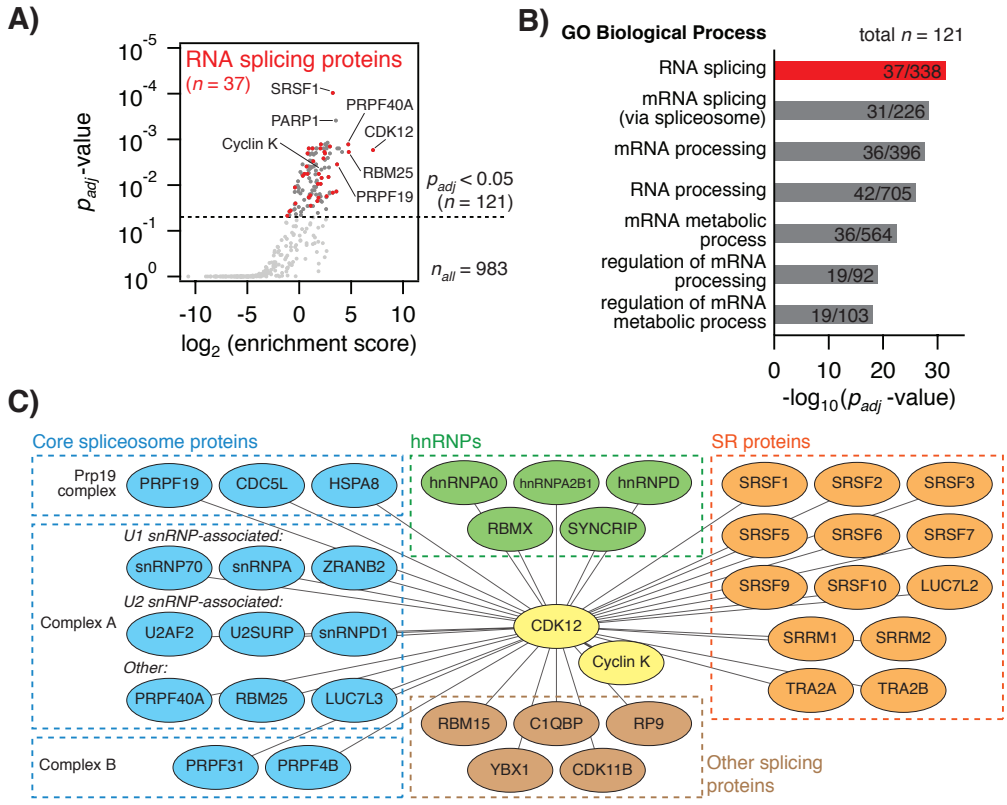

Figure 3.5: *CDK12* interacts with the RNA splicing machinery. **A.** Immunoprecipitation of FLAG-*CDK12* and mass spectrometry was used to identify 121 *CDK12*-interacting proteins in SK-BR-3 cells (enrichment score > 0, *padj* < 0.05). **B.** Interacting proteins were highly enriched for RNA splicing functions as determined by Gene Ontology (GO) analysis [210]. **C.** *CDK12*-interacting splicing proteins can be generally divided into core spliceosome proteins (blue) and regulatory splicing factors (green, orange, and brown).

by *CDK12* occurs through a splicing mechanism rather than through gene-specific recruitment of polyadenylation factors.

3.3.2 Tumors defective in *CDK12* function exhibit mis-regulation of ALE splicing

Alterations in *CDK12* have been described in numerous tumor types, including breast, ovarian, uterine, gastric, and bladder cancers [153, 181, 185, 194, 195]. The TCGA consortium has performed large-scale analyses on collections of tumor samples, including RNA-seq for 311 cases of ovarian serous cystadenocarcinoma [153]. *CDK12* is recurrently altered in 6% of these cases (Figure 3.7.A). Tumors containing the *CDK12* mutations are notably not amplified for *HER2*, and previous studies demonstrated that these ovarian cancer mutations impair the kinase activity of *CDK12 in vitro* [152, 183]. Therefore, these samples are well suited to explore the consequences of modulating *CDK12* function in a tumor setting.

To generalize the regulation of ALE events by *CDK12* to tumor cells, we used the MISO package to perform pairwise comparisons of tumor samples containing *CDK12* alterations to tumor samples without *CDK12* alterations (Figure 3.7.B). For this analysis, we utilized data from four types of available TCGA RNA-seq samples [153]: tumors with *CDK12* point mutations (n = 7), tumors with homozygous *CDK12* deletions (n = 3), tumors with genomic amplification of *CDK12* (n = 4), and tumors with no alterations in *CDK12* (n = 56 control samples). We queried the point mutation, deletion, amplification, and control samples for the occurrence of the 133 ALE events that resulted from *CDK12* depletion in SK-BR-3 and 184-hTERT cells. Each ALE event in *CDK12*-mutated tumors was found in 49% of comparisons on average (point mutation:control), as compared to 23% of control (control:control) comparisons (Figure 3.7.B i). When considering only the 23 events common to both SK-BR-3 and 184-hTERT cells, each ALE event was found in 71% and 27% of mutation and control comparisons on average, respectively. Similar trends were obtained with tumors containing homozygous *CDK12* deletions (Figure 3.7.B ii), demonstrating that these ALE events were identified more frequently in tumors impaired in *CDK12* function.

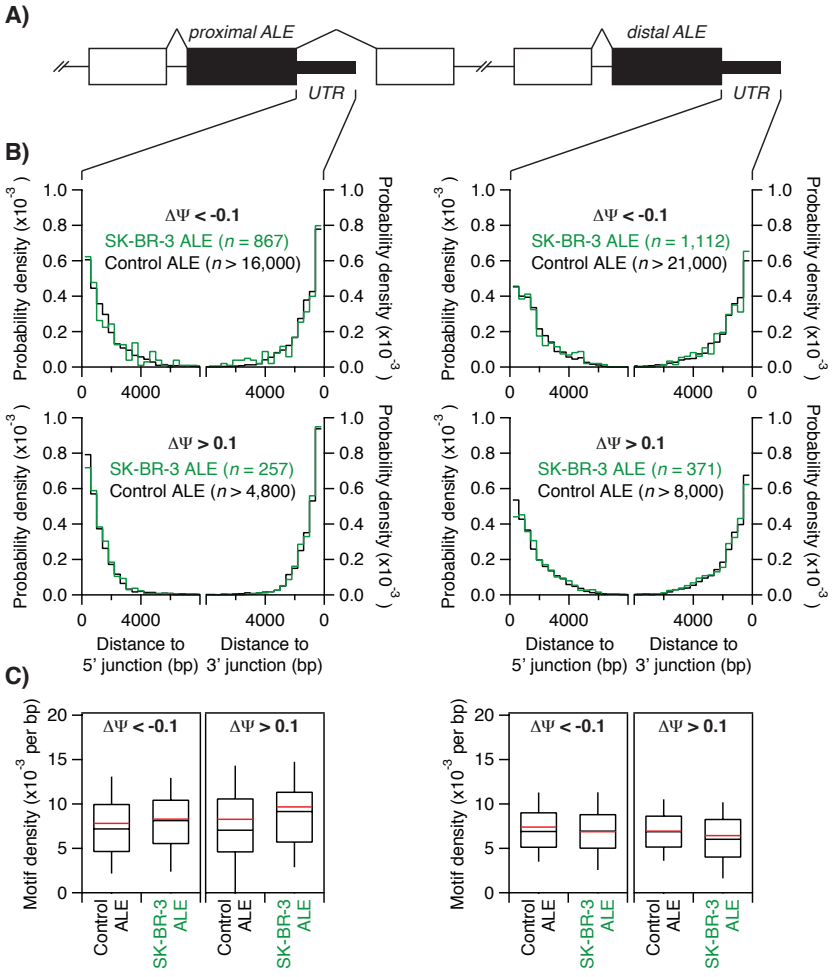

Figure 3.6: The 3'UTR of ALEs regulated by *CDK12* do not feature unique patterns of polyadenylation motifs. Distribution plots are shown for ALEs regulated by *CDK12* (green lines and boxes) and control ALEs (black lines and boxes). ALE events are divided into those that result in greater usage of the proximal ALE ($\Delta\Psi < -0.1$) and those that favor the distal ALE ($\Delta\Psi > 0.1$). **A.** Analysis of polyadenylation motifs was performed on the 3'UTRs of proximal and distal ALEs regulated by *CDK12*. **B.** Distributions of distances of polyadenylation motifs from the 5' and 3' junctions of 3'UTRs. The n values represent total numbers of polyadenylation motifs identified. **C.** Distributions of the densities of polyadenylation motifs in the 3'UTRs. The differences between the distributions of SK-BR-3 ALEs and control ALEs are not statistically significant in all comparisons (Mann-Whitney U test, Benjamini-Hochberg corrected $p > 0.05$).

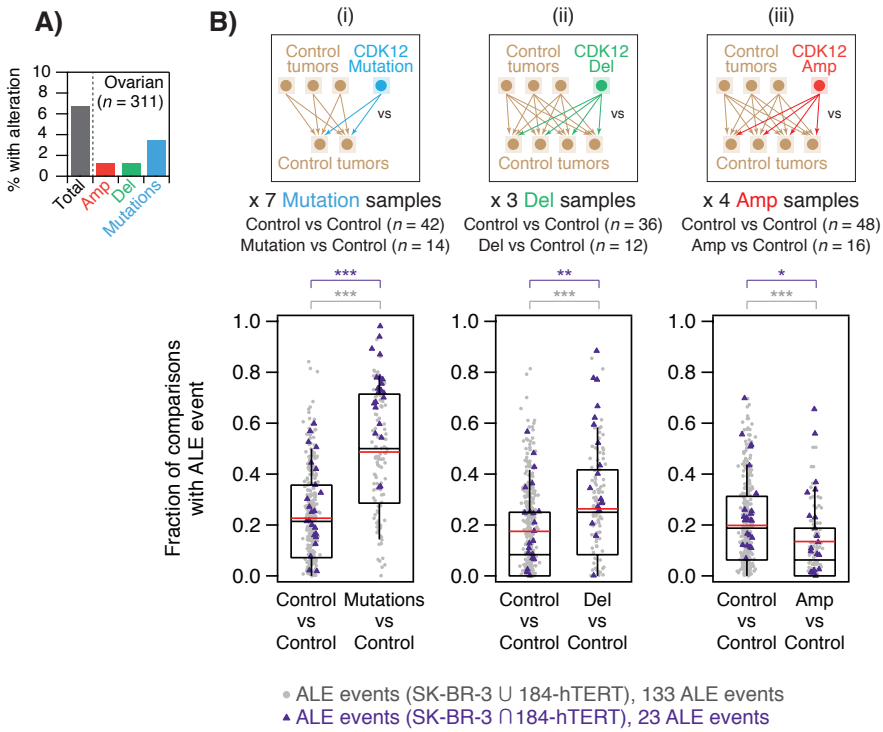

Figure 3.7: Alterations in *CDK12* correlate with mis-regulation of ALE splicing in ovarian tumor samples. **A.** *CDK12* is recurrently altered in ovarian serous cystadenocarcinomas [153]. From this dataset, RNA-seq data was available for tumors containing *CDK12* point mutations (blue, n = 7), homozygous deletions (green, n = 3), and amplifications (red, n = 4). **B.** Using the MISO package, changes in AS (Bayes Factor $\geq$ 20) were determined based on the following comparisons: (i) *CDK12* point mutation vs. control, (ii) *CDK12* deletion vs. control, and (iii) *CDK12* amplification vs. control. Changes in *CDK12*-regulated AS events were compared to AS events found in control vs. control comparisons. To obtain a similar number of comparisons in each scenario, each point mutation sample (i) was compared to two unique control samples (n = 14 comparisons), while each deletion (ii) and amplification sample (iii) was compared to four unique control samples (n = 12 and 16 comparisons, respectively). Control vs. control comparisons were likewise paired, and additionally performed in triplicate (n = 36, 42, or 48 comparisons). A total of 133 ALE events were queried, representing the events found in either the SK-BR-3 or 184-hTERT experiments (grey circles). We also queried 23 ALE events common to both SK-BR-3 and 184-hTERT cells (purple triangles). Red lines represent the means. The significances of comparisons (grey and purple lines) were determined using the Mann-Whitney U test (*p $<$ 0.05, **p $<$ 0.005, ***p $<$ 1 x 10^{-5}). The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at: http://cancergenome.nih.gov/.

In breast cancers, *CDK12* is commonly co-amplified with *HER2*. Similarly, the four ovarian tumor samples with *CDK12* amplifications also contain *HER2* amplifications. Unlike cases containing *CDK12* point mutations or deletions, the queried ALE events were found less frequently in tumors amplified for *CDK12* (13% of amplification:control and 20% of control:control comparisons; Figure 3.7.B iii). These findings with the *CDK12*-amplified samples mirror our results in SK-BR-3 cells, where the ALE events were identified after depletion of *CDK12* from an over-expressed state. Together, these results suggest that mis-regulation of ALE splicing occurs due to aberrations in *CDK12* and support a functional role of *CDK12* alterations in tumor development in ovarian tumors.

3.3.3 Regulation of gene expression by *CDK12* is gene- and cell type-specific but modulates a core set of common pathways

The role of *CDK12* in regulating ALE splicing of long transcripts occurs in multiple cell and tumor types; however, only a small subset of these regulated genes are common to multiple cell types. To address the question if *CDK12* also regulated cell type-specific gene transcription we evaluated the effects of *CDK12* on global gene expression.

We analyzed the triplicate *CDK12* siRNA and control siRNA RNA-seq data from SK-BR-3 and 184-hTERT cells using DESEQ2 [199, 212]. The analysis found that depletion of *CDK12* resulted in modest changes in gene expression (Figure 3.8.A). In SK-BR-3 cells, 3,163 statistically significant ($padj < 0.01$) events were evenly divided into up-regulated (50%, mean fold change = 1.5) and down-regulated (50%, mean fold change = -1.5) genes. Of these events, only 386 exhibited more than a 2 fold change in gene expression (Figure 3.8.C). Depletion of *CDK12* in 184-hTERT cells resulted in slightly more differential expression events (n = 3,940 with $padj < 0.01$). Again, events were differentially expressed in both directions (49% up-regulated an average 1.7 fold; 51% down-regulated an average 1.6-fold). Only 678 changed more than 2 fold in expression. Of these genes, 37 were differentially expressed in both cell lines (Figure 3.8.C). These analyses contrast with a previous study in HCT-116 cells, which reported that 98% of differentially expressed genes were down-regulated after *CDK12* depletion [145]. Taken together, our observations suggest that similar to the regulation of ALE splicing, regulation of gene

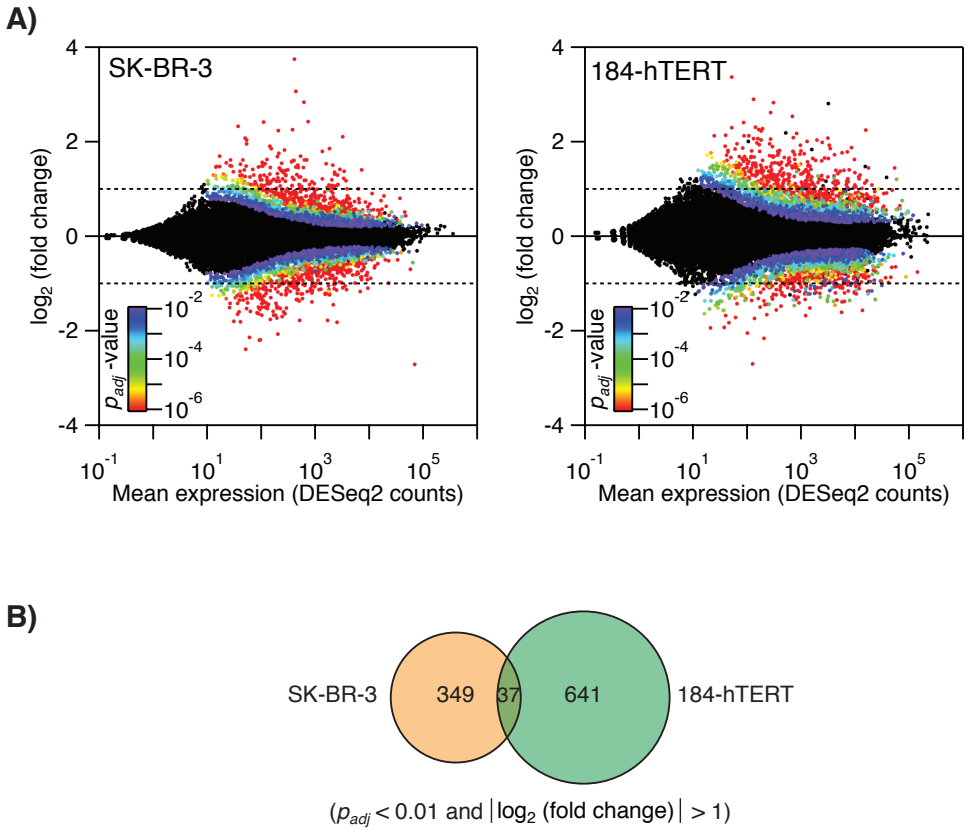

Figure 3.8: *CDK12* differentially regulates gene expression in a cell type-specific manner. **A.** Differential gene expression analysis by RNA-seq following *CDK12* depletion in SK-BR-3 and 184-hTERT cells. Mean expression (DESEQ2 counts) is plotted against fold change (*CDK12* siRNA-1 versus scrambled siRNA). Dotted lines delineate events with $|foldchange| > 2$. Events with $padj < 0.01$ are colored. **C.** Few differential gene expression events with padj < 0.01 and $|foldchange| > 2$ are common between SK-BR-3 and 184-hTERT cells.

expression by *CDK12* is highly gene- and cell type-specific.

While the regulation of individual genes by *CDK12* was cell type-specific, an examination of the affected cellular pathways using Gene Set Enrichment Analysis (GSEA) [202] offered additional insight. We found that in both SK-BR-3 and 184-hTERT cells, loss of *CDK12* altered similar pathways. Identification of these pathways also support previously

reported functions of *CDK12* [145, 146, 151, 152, 183, 185, 213, 214]. Namely, depletion of *CDK12* resulted in the down-regulation of genes involved in RNA splicing and processing, cell cycle progression, and regulation of DNA damage response pathways in both cell types (Figure 3.9). Since these processes were previously reported in different cell types, they appear to represent universal functions of *CDK12*.

The pathway analysis also aided in determining cell type-specific properties of *CDK12*. Depletion of *CDK12* in SK-BR-3 cells decreased expression of genes associated with mitochondrial function (Figure 3.9). This change was not observed in 184-hTERT cells. Instead, depletion of *CDK12* in 184-hTERT cells increased expression of genes associated with the plasma membrane or related to development and extracellular activity (Figure 3.9). In general, *CDK12* expression both increased and decreased the expression of genes in various pathways in 184-hTERT cells, but primarily up-regulated pathways in SK-BR-3 cells (Figure 3.9). Taken together, these results demonstrate that while transcriptional regulation by *CDK12* is largely gene- and cell type-specific, common cellular processes are modulated by *CDK12* activity amongst different cell types.

We next sought to determine how changes in gene expression due to *CDK12* function manifest at the protein level to affect the expressed phenotype of SK-BR-3 cells. Global proteomics experiment was performed (by Grace Cheng, Christalle Chow, Jerry Tien, and Christopher Hughes) to quantify alterations in protein expression after depletion of *CDK12* in SK-BR-3 cells, and we compared the results to the matching RNA-seq data (Figure 3.10). We found that the proteome data represented a smaller subset of the transcriptome data (Figure 3.10.A). Of the 11,072 expressed genes in the RNA seq data (defined as FPKM $\geq$ 1), 7,031 (64%) were identified at the protein level by mass spectrometry (Figure 3.10.A).

Moreover, similar to the transcriptome data, only a small proportion of proteins were differentially expressed (n = 444, $padj < 0.01$) in SK-BR-3 cells after depletion of *CDK12* (Figure 3.10.B). There was a high correlation in the fold change values of the 197 genes that were differentially expressed in a statistically significant manner in both the transcriptome and proteome datasets (Figure 3.10.C). We note that 242 genes were changed at the protein level and not at the mRNA level, and that 1,136 mRNAs were changed at the transcriptome level and not at the protein level. Pathway analyses demonstrated that the core

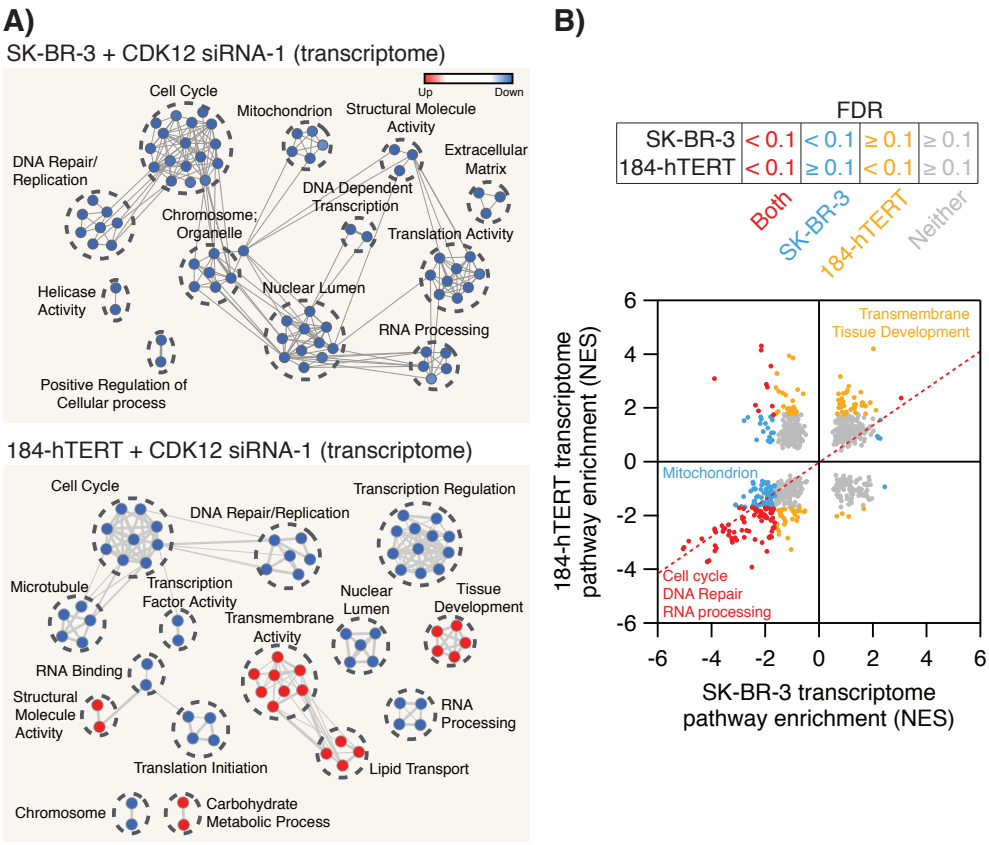

Figure 3.9: *CDK12* regulates the expression of a core set of genes and pathways. **A.** Enrichment maps from GSEA analysis of differential gene expression resulting from *CDK12* depletion in SK-BR-3 and 184-hTERT cells. **B.** For each pathway, GSEA pre-ranked analysis assigned a normalized enrichment score (NES) representing the extent of over-representation of genes of a pathway at the top or bottom of a ranked list. Positive and negative NES values represent up- and down-regulated pathways, respectively. For each pathway, NES values in SK-BR-3 and 184-hTERT are shown. Red markers represent NES values significant in both cell lines (FDR < 0.1). The dotted red line shows the general trend of these points. Blue and yellow markers represent NES values only significant in SK-BR-3 and 184-hTERT cells, respectively.

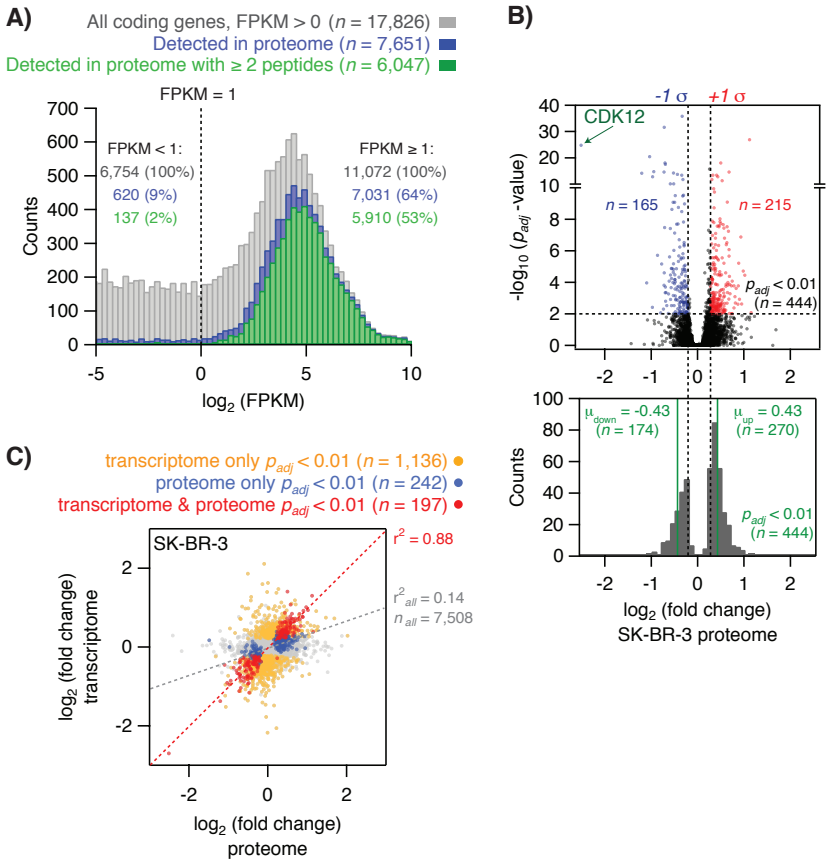

Figure 3.10: Differential protein expression due to *CDK12* regulation represents a subset of differential gene expression events. **A.** Histogram of RNA-seq expression values (FPKM) for all coding genes and genes with corresponding proteins detected by mass spectrometry with ≥ 1 unique peptides (blue bars) or ≥ 2 unique peptides (green bars). **B.** Top: volcano plot of the global proteome analysis in SK-BR-3 cells. Dotted horizontal line denotes point at which *padj* = 0.01. Dotted vertical lines lineate events with $|foldchange| > 1$ s.d. (σ) from the mean. Bottom: distribution of fold change values for all differential protein expression events with $padj < 0.01$. Green vertical lines denote mean fold change (μ) values for up- and down-regulation. Dotted lines are the $\pm$ 1 σ lines extended from the top plot. **C.** Correlation of fold change values from global transcriptome and proteome analysis in SK-BR-3 cells ($r^2_{all} = 0.14$, p $< 10^{-5}$). Events with significant fold change values ($padj < 0.01$) in both datasets are shown in red ($r^2 = 0.88$, $p < 10^{-5}$). Events significant only in the transcriptome and proteome are colored yellow and blue, respectively.

functions of *CDK12* (e.g., RNA processing and DNA damage response) were all observed in the proteomics experiment (Appendix. B). Functions specific to SK-BR-3 cells, such as the involvement of mitochondrial processes, were also found at the protein level. However, the regulation of proteins involved in cell cycle and cell division, which was prominent in the transcriptome data, was absent in the proteome data. This is likely a result of mRNA-independent means of regulating protein expression and turnover, and may also be cell type-specific.

3.3.4 *CDK12* can modulate the expression of DNA damage response genes in SK-BR-3 cells through alternative splicing

Based on microarray differential gene expression analysis, it was proposed that *CDK12* regulates the expression of DNA damage repair genes [146]. Our analysis suggests that AS may be a significant mechanism of regulation by *CDK12*, especially for genes with long transcripts and many exons. One such example we identified in our SK-BR-3 RNA-seq data was the gene encoding the *ATM* (Ataxia Telangiectasia Mutated) protein. *ATM* is a key regulatory kinase that responds to DNA double-strand breaks and initiates DNA repair pathways [215]. The canonical isoform of *ATM* is a 350 kDa protein translated from a 13,147-bp transcript containing 63 exons (Figure 3.11.A). Along with many other DDR genes, treatment of SK-BR-3 and 184-hTERT cells with *CDK12* siRNA resulted in a down-regulation of *ATM* mRNA expression (Appendix B). Specific to SK-BR-3 cells, however, *CDK12* regulates the expression of *ATM* through ALE splicing (Figure 3.11.B and C). By examining expression of individual *ATM* exons (Figure 3.11), and as confirmed by qRT-PCR (Figure 3.11.C), *CDK12* depletion resulted in a 1.3-fold down-regulation of most of the exons. However, the terminal exon and 3'UTR were down-regulated more than 4-fold. This was in contrast to 184-hTERT cells, where *CDK12* depletion resulted in a 1.4-fold down-regulation across the entire length of the *ATM* gene.

These data indicate that in SK-BR-3 cells, the expression of full-length *ATM* isoform could be regulated through AS in addition to direct transcriptional control. Using a monoclonal antibody targeting *ATM* residues 980-1,512 (exons 20-30), it was confirmed at the protein level that the expression of full-length *ATM* was decreased 3-fold after *CDK12*

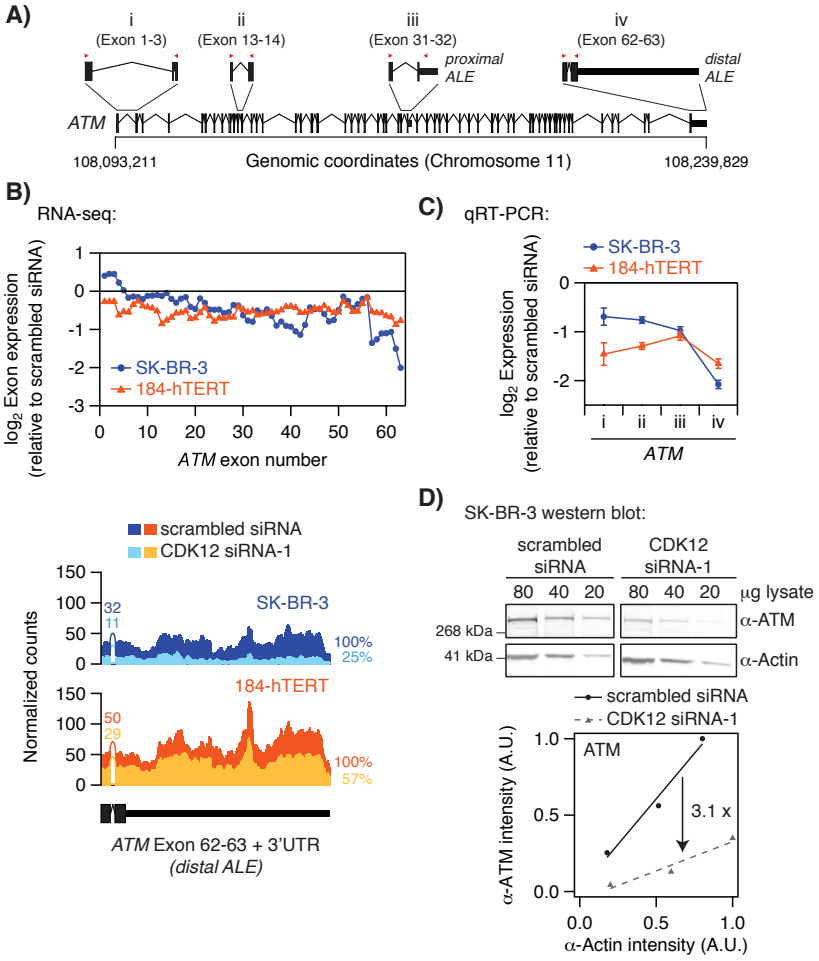

Figure 3.11: *CDK12* regulates the expression of full-length *ATM* through ALE splicing in SK-BR-3 cells. **A.** Exon structure of the canonical *ATM* isoform, corresponding to Ensembl transcript *ENST00000278616*. Primers for qRT-PCR in **(C)** were designed to target four exon junctions (i-iv) and are shown as red arrowheads. **B.** Top: relative expression of each *ATM* exon after *CDK12* depletion in SK-BR-3 (blue circles) and 184-hTERT (orange triangles) cells. Bottom: normalized read counts for the 3' end of the canonical *ATM* isoform after *CDK12* depletion in SK-BR-3 (dark and light blue traces) and 184-hTERT (dark and light orange traces) cells. **C.** Validation of RNA-seq exon expression analysis by qRT-PCR. Expression levels were determined for the four regions of *ATM* (i-iv, shown in **(A)**) after *CDK12* depletion in SK-BR-3 (blue circles) and 184-hTERT (orange triangles) cells. Error bars denote the 99% confidence interval range. **D.** Relative quantification of full-length *ATM* protein expression due to *CDK12* depletion by western blot analysis.

depletion (by Jerry Tien, Christalle Chow, and Leanna Canapi; Figure 3.11.D). These observations suggest that *CDK12* can modulate the protein expression of full-length *ATM* by altering the ratio of different *ATM* splice isoforms. These results demonstrate that AS is an additional mechanism by which *CDK12* can control DNA repair pathways.

3.3.5 *CDK12* down-regulates the long isoform of *DNAJB6* and increases the tumorigenicity of breast cancer cells

Pathway analysis of differential gene and protein expression suggests that some *CDK12* functions are conserved across cell types. In addition to cell type-specific regulation described above, we identified common ALE events that were regulated by *CDK12* in multiple cell lines. From our experiments with SK-BR-3 and 184-hTERT cells, and the available datasets from HCT-116 cells [145], we found that loss of *CDK12* is frequently associated with changes in ALE splicing of the *DNAJB6* (DnaJ homolog subfamily B member 6, MRJ) gene (in SK-BR-3 and 184-hTERT, $\Delta\Psi_{avg} = 0.21$). In our analysis of TCGA RNA-seq data for tumors containing homozygous *CDK12* deletions (12 deletion:control pairs), the *DNAJB6* ALE event was found in 92% of comparisons on average, as compared to 44% of control (36 control:control pairs) comparisons (Fisher's exact test $p = 0.006$).

Unlike the long genes that were regulated in a cell type-specific manner, *DNAJB6* encodes two small protein isoforms (36 and 27 kDa) from transcripts containing 10 and 8 exons, respectively (Figure 3.12.A). The short isoform of the *DNAJB6* protein (*DNAJB6-S*) is a HSP40 family cytosolic chaperone with implicated roles in Huntington's disease [216, 217]. By contrast, ALE splicing introduces a nuclear localization signal in the long isoform of *DNAJB6* (*DNAJB6-L*), and therefore it operates primarily in the nucleus. Increased nuclear localization of *DNAJB6-L* has been reported to mitigate tumorigenicity and metastasis in breast and esophageal cancer cells [218, 219].

Our RNA seq data showed that through ALE splicing, higher *CDK12* expression in SK-BR-3 cells reduced the expression of *DNAJB6-L* (Figure 3.12.B), consistent with *CDK12* functioning as an oncogene. We tested this hypothesis using MDA-MB-231 cells, a highly invasive triple-negative breast cancer cell line where *DNAJB6-L* had been previously shown to decrease cell migration potential [218]. Global proteome pathway analysis

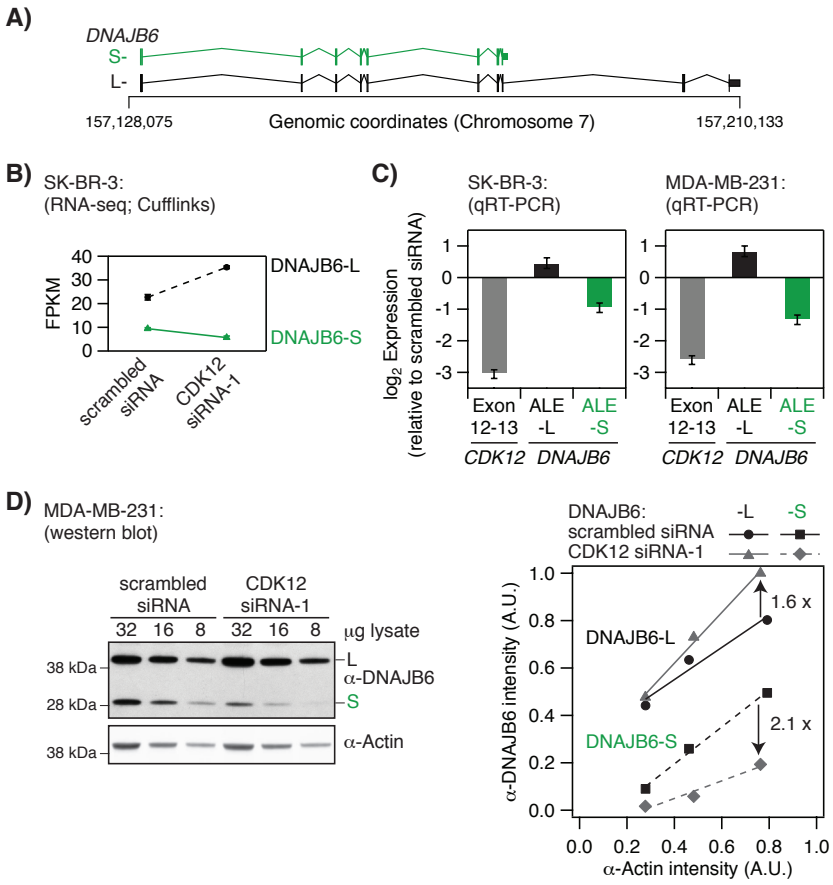

Figure 3.12: *CDK12* down-regulates the long isoform of *DNAJB6* through ALE splicing. **A.** Exon structure of the long (-L) and short (-S) isoforms of *DNAJB6*, corresponding to ENSEMBL transcripts *ENST00000262177* and *ENST00000429029*, respectively. **B.** Quantification of *DNAJB6-L* and *DNAJB6-S* transcript levels (FPKM) after *CDK12* depletion in SK-BR-3 cells by RNA-seq using CUFFLINKS. Error bars represent s.d. **C.** Validation of changes in *DNAJB6-L* and *DNAJB6-S* transcript expression after *CDK12* depletion in SK-BR-3 and MDA-MB-231 cells by qRT-PCR. Error bars denote the 99% confidence interval range. **D.** Relative quantification of changes in *DNAJB6-L* and *DNAJB6-S* protein expression due to *CDK12* depletion in MDA-MB-231 cells by western blot analysis.

of *CDK12*-depleted MDA-MB-231 cells largely resembled results from SK-BR-3 cells, with the exception of the down-regulation of cell cycle and cell division proteins that was not seen in SK-BR-3 cells (Appendix. B). This analysis further supported the use of MDA-MB-231 cells to examine the effects of *CDK12* on tumorigenicity. Using qRT-PCR and western blot analysis, we confirmed that MDA-MB-231 cells treated with *CDK12* siRNA increased gene and protein expression of *DNAJB6-L* (and decreased expression of *DNAJB6-S*) as compared to a scrambled siRNA control (Figure 3.12.C and D).

To examine the cellular phenotype associated with *CDK12* expression, a scratch wound assay and live cell imaging of MDA-MB-231 cells were used (by Grace Cheng and Jerry Tien) as a functional test for cell migration. The experiments show that the ability of MDA-MB-231 cells to invade is correlated with *CDK12* expression and inversely correlated with the expression level of *DNAJB6-L*, and suggest that *CDK12* can increase the tumorigenicity of an invasive breast cancer cell line, likely through ALE splicing of the *DNAJB6* gene [2].

3.4 Discussion

We showed that *CDK12* regulates ALE splicing in a cell type specific manner. Prior to this study, the global effect of *CDK12* on AS was uncharacterized, and opposing conclusions had been made regarding its role in gene expression. While several studies proposed that *CDK12* specifically affects a small number of genes [146, 220], another report suggested a global up-regulation of transcription [145]. Here, we applied stringent criteria, combining RNA-seq datasets in biological triplicates from two different cell lines to identify AS and differential gene expression events with high confidence.

We found that the regulation of ALE splicing and differential gene expression by *CDK12* was limited to a small subset of genes and the nature of this regulation was highly cell type-specific. In 184-hTERT cells, *CDK12* both up- and down-regulated the expression of genes and pathways. Using the same statistical criteria in SK-BR-3 cells, *CDK12* both up- and down-regulated genes, but the most significantly affected pathways were all down-regulated after *CDK12* depletion. Down-regulation of pathways in SK-BR-3 cells is consistent with the role of *CDK12* in increasing the rate of transcription elongation.

Importantly, our proteomic analysis of SK-BR-3 cells suggests that not all *CDK12*-mediated transcriptional regulation manifests at the protein level. For example, pathways relating to cell cycle and cell division were down-regulated in the transcriptome of SK-BR-3 cells after *CDK12* depletion, but not in the proteome. These results could reflect additional layers of regulation at the protein level, including the modulation of translation, post-translational modifications, and protein turnover/proteolysis. An additional factor to explain this observation could be a dominant effect of *HER2* over-expression on many pathways [221]. Consistent with this idea, loss of *CDK12* significantly down-regulates cell cycle and cell division proteins in MDA-MB-231 cells, which do not have *HER2* amplification.

In general, we found that *CDK12* regulates ALE splicing of genes with long transcripts and high numbers of exons. This trend was significantly more pronounced in ALE splicing events regulated by *CDK12*, rather than in differential gene expression events as previously reported for HeLa cells [146]. Furthermore, in a majority of events, native *CDK12* promoted the splicing of the longer mRNA isoform.

The simplest model for *CDK12* regulation of pre-mRNA processing is that *CDK12* increases the processivity and/or rate of elongation to achieve successful splicing of one exon to the next exon. In the absence of *CDK12*, this splicing event is reduced due to decreased processivity and transcription defaults to termination and polyadenylation of what then becomes the last exon (the proximal ALE). However, this simple model cannot explain all our major observations. For instance, it is unclear how the proximal ALE is selected amongst all the exons within a long transcript. Notably, we did not observe any difference in the density of polyadenylation motifs in the 3'UTRs of *CDK12*-regulated ALEs.

It is also not known how *CDK12* achieves regulation of only a small subset of genes that differs depending on cell type. This is possibly accomplished by the various tissue-specific splicing regulatory factors that associate with *CDK12* or by signal transduction processes that regulate the action of *CDK12* and/or its interacting proteins. The processivity and elongation model also does not explain ALE splicing to promote the shorter mRNA isoform, as observed with a minority of genes (20%). One such gene, *DNAJB6* , is regulated by *CDK12* in multiple cell types and tumors, suggesting a gene-specific reg-

ulation that differs from the possible length-dependent regulation common to other ALE events. Therefore, it is probable that regulation of AS by *CDK12* also requires additional splicing factors such as the SR proteins, hnRNPs, and RNA processing factors identified in our immunoprecipitation experiments. Future studies should be aimed at determining the precise role of these regulatory proteins in *CDK12*-dependent regulation of splicing.

Our results shows that *CDK12* regulates the DNA damage response through multiple mechanisms. One of the most consistently reported functions of *CDK12* has been the regulation of the DDR. Differential expression of specific DDR genes was first identified by microarray analysis [146], and changes in DDR pathways were determined from transcriptome analysis [145]. Furthermore, *CDK12* depletion was found to be synthetic lethal with PARP inhibition [151, 183, 185]. This behavior is reminiscent of the sensitivity of *BRCA1*/*BRCA2*-deficient tumors to PARP inhibitors [222–224], suggesting that *CDK12* may be specifically involved in the HDR pathway. Indeed, ovarian tumors containing *CDK12* mutations exhibited down-regulation of several HDR genes [152].

In all cell types we examined, *CDK12* regulated gene and protein expression of components of the DDR pathway. Furthermore, our RNA-seq data for SK-BR-3 cells suggest that *CDK12* may be a key regulator of HDR through ALE splicing of *ATM*, a master regulating kinase that directly responds to DNA damage. The splicing-dependent regulation of *ATM* in SK-BR-3 cells was independent of transcriptional regulation, whereas in 184-hTERT cells there was modest transcriptional regulation of *ATM*. By compiling our data and those on gene regulation from the literature [145, 146] it is apparent that gene regulation and AS regulation by *CDK12* is both cell type specific and gene specific. Furthermore, while *CDK12* alters the transcription of some genes, it can also modulate the splicing of functional isoforms of DDR genes.

In line with our findings, experiments exploring the effect of loss-of-function mutations in *CDK12* on the DDR suggest that *CDK12* is a tumor suppressor gene. However, several observations show that *CDK12* can also function as an oncogene. This is particularly pertinent in breast cancers, where *CDK12* is frequently co-amplified with the *HER2* oncogene. Over-expression of *CDK12* is correlated with aggressive tumor behaviour and poor survival [134, 180, 182]. Notably, these properties also apply to the small fraction of tumors where *CDK12* is amplified but *HER2* is not, suggesting an oncogenic potential

independent of *HER2* [144].

Our RNA-seq experiments examining a breast cancer cell line over-expressing *CDK12* (SK-BR-3 cells) identified AS splicing events that could promote tumorigenesis. These events were also found in our analysis of TCGA RNA-seq data of ovarian tumors containing *CDK12* amplifications. One notable AS event regulated by *CDK12* and identified in multiple cell types and tumors was the ALE splicing of *DNAJB6*. Recent studies show that the long isoform of *DNAJB6* (*DNAJB6-L*) suppresses cell migration and invasion in MDA-MB-231 cells [218]. While the mechanism driving this activity was unclear, it was dependent on the ALE splicing and subsequent nuclear localization of *DNAJB6-L*.

Using the same MDA-MB-231 cell line model, we showed that *CDK12* expression is inversely correlated with ALE splicing of *DNAJB6-L*. The ability of cancer cells to migrate and invade is a fundamental mechanism underlying tumorigenesis and metastasis [225]. MDA-MB-231 cells can seed tumors in mouse models, and increasing *DNAJB6-L* expression decreases tumor growth and metastasis in athymic mice [218]. Therefore, the ability of *CDK12* over-expression to down-regulate *DNAJB6-L* through ALE splicing represents a specific cellular mechanism by which *CDK12* can increase the tumorigenicity of breast cancer cells. This could be a significant factor contributing to the progression of *HER2*$^+$ breast cancers, where *CDK12* is co-amplified in 27-92% of cases [186–194].

In this study, we applied a comprehensive genomic and proteomic approach to define the cellular functions of *CDK12* and to investigate its oncogenic properties. We showed that in multiple cell lines, *CDK12* regulated a core set of cellular processes including RNA processing and DNA repair. We also found that *CDK12* regulated ALE splicing, primarily of genes with long transcripts and a large number of exons. While this regulation mechanism is present in multiple cell lines, the affected genes are highly cell type-specific. In SK-BR-3 cells, *CDK12* modulated ALE splicing to promote the generation of full-length *ATM*, a key component of DNA repair associated with tumorigenesis. *CDK12* also regulated splicing of *DNAJB6*, whose nuclear localization attenuates tumor invasion. In MDA-MB-231 cells, *CDK12* promoted tumor migration and invasion in a dose-dependent manner. Together, these results show how loss of *CDK12* can disrupt DNA repair, but also demonstrate an AS-dependent mechanism by which *CDK12* over-expression can increase the tumorigenicity of breast cancer cells.

Chapter 4

Investigating Cellular Responses upon Inhibiting Components of Splicing Machinery

4.1 Introduction

Alternative splicing is precisely regulated through complex interactions of a large number of proteins, RNA molecules, and environmental stimuli [26]. The complex interplay between components of this machinery is essential to maintain cell functions. Consequently, a considerable number of genetic diseases has been linked to mutations that impair splicing. For instance, more than 15% of disease causing genetic mutations are believed to disturb splicing [226].

Disruption of splicing has been involved in many diseases including: Growth hormone deficiency, Parkinsons disease, Cystic fibrosis, Retinitis pigmentosa, Spinal muscular atrophy, and also several types of cancer [5, 227, 228]. Usually, genes important for tumor biology involved in processes such as cell cycle regulation and apoptosis are regulated by alternative splicing [228]. In this case, aberrant splicing events in genes with specific functions can lead to uncontrolled growth and survival of cells [229, 230]. These aberrant splicing events are usually a consequence of mutations in components of splicing ma-

chinery, 3' and 5' splice sites, or splicing silencers and enhancers. As a result, splicing mechanism contributes to the development and progression of tumors [231].

Since the recognition of cancer specific splice variants, splicing is now being appreciated as a potential therapeutic target [232]. Conceptually, two strategies are being investigated. The first strategy is trying to interfere with the components of splicing machinery, and if the components are more crucial for tumor cells compared to normal cells, then the interruption may show therapeutic advantage [233]. Nevertheless, because the spliceosome components modulate the splicing of an extensive number of genes, the corresponding drugs may display cytotoxic effects. As an alternative, in the second strategy, tumor specific splicing events are targeted directly [233]. This approach is expected to have less off-target effects, but it is necessary to identify the key splicing events to establish better treatment potentials.

A primary step to the development of splicing related therapeutics is understanding how components of spliceosome contribute to the regulation of splicing, and uncovering how they interact to maintain balance between isoforms. In general, splicing machinery can be modelled as a dynamic system of interactions. To understand regulations of this system, we can interfere with the system from multiple points (e.g., inhibiting one protein at a time) and evaluate the system's response. Next, systematically integrating the results of these measurements leads to developing a model, capable of explaining our observations and predicting system's responses in further conditions. Already, several methods have been proposed to infer genetic interactions and relations using perturbation screens [234–236].

Advancement in developing pharmacological agents improved the opportunities in systematic study of biological systems. Despite the growing evidence of the importance of splicing mechanism in maintaining normal cellular functions, there remains much unknown about its regulation in mammalian systems. Most of our current understanding of the spliceosome is determined through studying model organisms. Only recently, with the development of pharmacological agents, we acquired the opportunity to systematically interfere with spliceosome components at different levels to inhibit their functions in human. In other words, several inhibition levels can be experimented through applying different dosage of pharmacological agents in order to investigate gradual changes in cellular re-

sponses. Moreover, following the inhibition, RNA-seq enables us to gauge corresponding changes in a genome-wide scale. Finally, we can cluster response patterns to determine groups of genes that may undergo similar regulations. For instance, genes that show monotonically increasing or decreasing response patterns have a higher chance of being primary targets of inhibited proteins.

When a gene is inhibited, we are interested to know what are the direct targets of it, which pathways undergo differential regulations as a result, and what are the main regulators of the observed differences. Here, I first briefly review the methods proposed to identify primary pathways and genes that are more probable to trigger differential regulation of usually a large number of other genes that are observed in genome-wide RNA-seq studies. Following that, I present the data set where multiple components of the splicing machinery are inhibited using small compounds, and finally, I show some preliminary results on how our data can help to understand functions of these components.

4.2 Identifying Pathways and genes contributing most to cellular responses: A short review

Human cells are remarkably complex systems with thousands of genes whose interactions are organized in order to maintain appropriate responses based on a given condition. One important goal in biology is to understand, predict, and ideally advantageously manipulate emergent responses of these complex systems [237].

To gain a mechanistic insight on how an stimulus drives a cellular response, how a tissue differentially regulates genes, or how a disease state deviates from a normal state, one needs to interpret the measurable differences between conditions. RNA-seq is one of the encouraging tools that provides good opportunities to study complex cellular mechanisms. By simultaneously measuring transcript abundance, RNA-seq provides snapshots of a cell status in a particular condition. When two conditions are compared using RNA-seq snapshots, hundreds or thousands of genes may exhibit alterations in transcript abundance. Some changes are direct consequence of the modified condition (which are of most interest), some are secondary effects of the direct targets, and some others may be due to errors or inherent stochasticity in RNA-seq sampling.

An additional source of knowledge to complement RNA-seq measurements is the information on known gene interactions stored in knowledge bases [238–243]. The number of these interactions grow rapidly. As an example, the number of non-redundant physical interactions in BIOGRID [242, 243] data base increased over 10 times since ten years ago. It should be noted that many of the stored interactions are context specific and do not always apply to a particular cell state, some are inferred from high-throughput experiments with lower confidence, and some may be reported in very few studies. Thus, a careful strategy should be designed to benefit most from these knowledge bases, while filtering irrelevant information.

In a study by Rolland et.al *et al* [244], it was estimated that among all possible binary interactions (direct interactions) between human proteins, fewer than 10% of them are known. This estimation does not consider the impact of alternative splicing and could be optimistic, and thus highlights the limitation of our knowledge in this research area. Therefore, when investigating cellular responses, the potential novel interactions could also be taken into account. Here, our focus is on methods that are knowledge-driven, and we do not consider the problem of inferring novel interactions from RNA-seq data.

Having access to prior knowledge of interactions and RNA-seq measurements, a central question is therefore, what are the pathways or genes whose differential regulation in an experiment contribute most to the observed variations. In this short review, we explain the two alternative approaches, but only focus on the methods that aim to provide rules and mechanisms governing the observed differences.

We group methods proposed to interpret variation between conditions into two broad classes. The first class constitutes approaches that try to find gene sets or pathways that are enriched by differentially expressed genes. These methods usually assess whether the genes of a specific gene set are over-represented in a given set of N differentially expressed genes compared to a randomly selected set of N genes. In a conceptually different approach, methods in the second class are designed to infer pathways or a small set of upstream genes driving cascades of changes that lead to the observed measurements.

The first class of methods aim to summarize a list of identified differentially transcribed genes into smaller sets of genes that are somehow connected: either they participate in a same pathway, or they take part in related functions. A large number of methods have

been proposed based on this point of view [245, 246]. Khatri *et al* [247] categorized these approaches into 3 smaller groups. In the first group, a threshold is applied to select a set of genes showing significant alterations, and then, the set of the selected genes are statistically assessed to detect their over-representation in predefined pathways or gene sets [248–252]. The second group incorporates the fold change magnitudes of the gene expression values in the statistical evaluation, as an improvement [202, 253, 254]. Finally, the last group modifies pathway scores in a way to account for interactions between pathway genes as well [255, 256].

Although gene set and pathway enrichment based methods were successful in organizing results and highlighting affected functions and pathways, they turn to usually fall short in predicting driver genes. These techniques become ineffective in spotting genes that govern the usually large number of genes that undergo differential regulation. In other words, these methods cannot provide insights on the underlying mechanisms controlling the transition between the two given conditions [257]. Furthermore, another limitation of these methods is that they only present transcriptional view of the variations; however, many of the interactions do not directly influence changes in transcript abundance [257].

To address the limitations of the pathway and gene set enrichment based methods, a second group of methods focus on detecting a small number of driver genes or few pathways, whose mis-regulation elucidate a mechanistic explanation of measured variations (Figure. 4.1). These methods that rely on the quality and quantity of knowledge bases for the existence and direction of interactions in regulatory networks are explained in the following.

Among the methods we discuss here that try to perform mechanistic inference, LPIA (Latent Pathway Identification Analysis) [257] is the most similar one to pathway enrichment based methods. Similar to previous methods, the output of the algorithm constitutes pathways; however, here the pathways are scored based on their potential to initiate cascades of changes in other pathways. More specifically, The method first constructs a network of pathways, one node for each pathway of a given knowledge base. Next, it assigns weighted edges between pathways. For each pair of pathways (pair of nodes in the generated network), the assigned weight reflects the number of GO (Gene Ontology) terms that are common between the genes of the two pathways, and also it reflects the number

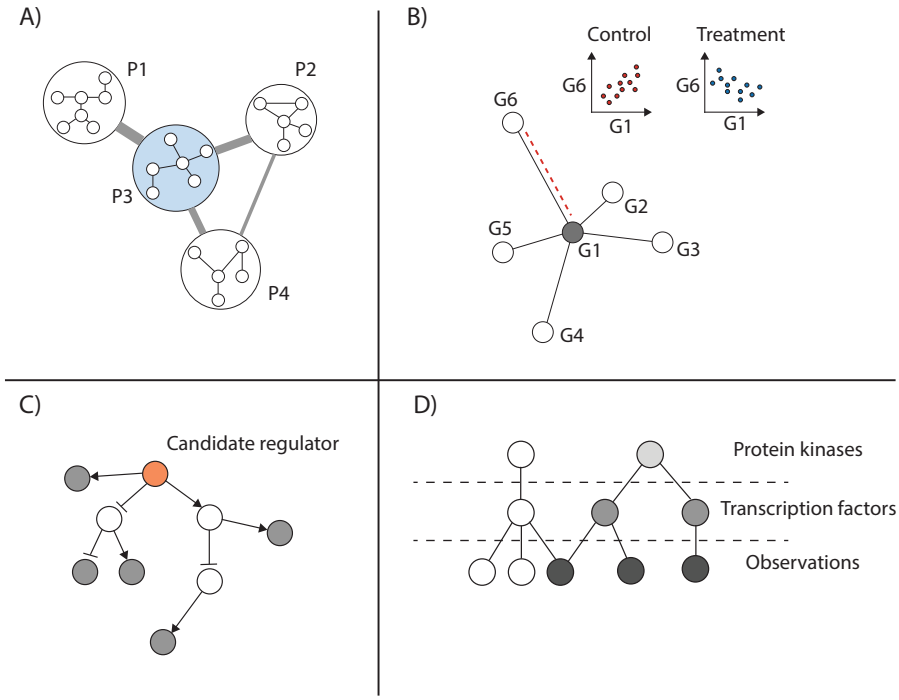

Figure 4.1: Methods proposed to perform mechanistic inference using high-throughput sequencing data. **A.** LPIA method [257] searches for pathways central to a set of disrupted pathways that have the potential to initiate alterations in other pathways. P_i's represent pathways consisting of sets of genes. Edge thickness displays how pathways are believed to be inter-related. For instance, the pathway shown in blue (P3), will be reported as an upstream regulatory pathway based on its connections. **B.** DEMAND method [258] searches for dysregulated interactions between genes. The joint probability density of the expression of interacting genes are compared between the two conditions (here between the genes G1 and G6), and genes whose interactions are significantly altered are reported as upstream causal genes. **C.** The third group of methods use the direction and sign of interactions to compare the predicted versus observed changes upon disruption of a given gene (here the gene shown in orange). Shaded circles display genes whose transcript abundance are observed and are expected to alter. **D.** The last group of methods perform multiple levels of inference in order to connect several regulatory levels. For example, based on the observed changes (black nodes), active transcription factors are identified (dark grey nodes), and in the next level of the analysis, candidate regulatory kinases are determined (the light grey node). For more information see the main text.

of common genes between the pathways that are differentially expressed. As a result, the assigned weights express both the prior belief on how much the two pathways are related, and also the context dependent (based on the experiment) measure of their interactions. In the final step, the pathways that are more central in the constructed network are reported as being potential causal pathways.

A main advantage of this method compared to pathway enrichment analysis is that here, the genes in the top reported pathways not only show differential regulation, but can describe observed changes in other pathways as well. On the other hand, one limitation of LPIA is that it does not benefit from interaction of genes and their directions to increase the confidence of causal inference. Additionally the amount of observed variations of gene expression could improve the scoring scheme.

To find genes driving transcriptional transition between the two conditions, DEMAND [258] (DEtecting Mechanism of Action by Network Dysregulation) searches for genes whose known interactions are significantly dysregulated. The method was primarily proposed to identify mechanism of action (MoA) of a compound, defined as targets essential to cause the pharmacological effect of a compound.

The underlying assumption in DEMAND method is that if a gene belongs to the MoA of a compound, then its direct targets are more likely to be dysregulated compared to random genes. As a result, DEMAND evaluates the changes in joint gene expression probability density of candidate genes using Kullback-Leibler divergence (KLD). Estimating joint probability distributions may require many samples; however, DEMAND is claimed to efficiently detect the corresponding changes by applying KLD. In the final step, the evaluated dysregulations between a gene and its neighbors are combined and a p-value is assigned to candidate genes. The method has been successfully applied to classify compounds with similar functions and targets.

The first limitation of DEMAND method is that in only considers first order neighbours (directly connected to the gene of interest) without considering the direction of regulation. Incorporating these additional information may improve the algorithm. Also, the assumption of expected alterations of gene expression can be violated when the regulation does not happen at transcriptional level. These issues are addressed in the next class of methods.

The next group of methods consolidate direction and sign of known interactions with differential expression analysis to discriminate upstream causal genes from others [259–266]. In most of these methods, a directed graph is constructed by putting one node for each network entity. The nodes represent transcripts, proteins, small molecules and compounds. Interactions between these entities are compiled from various data bases. Interactions should be ideally signed and directed, showing the direction of the regulation and whether the regulation is activation or inhibition. In the inference step, the common framework is to predict the expected change of downstream genes using signed directed paths, and then scoring candidates by comparing the expected and measured values.

In one of these methods, Chindelevitch *et al* [259] evaluated the expected direction of changes for measured entities (transcript abundance) downstream of a candidate gene, assuming the candidate gene is disrupted. In their evaluation, they considered directions and signs of shortest paths from the candidate genes to those measured values. Next, they introduced a scoring scheme based on rewarding and penalizing correct and incorrect predictions, accordingly. In the final step, they compared the computed scores to randomized situations, in order to assign p-values to each of the candidate upstream regulators. Several improvements have been applied in IPA (Ingenuity Pathway Analysis) approach [261]. The technique takes advantage of edge weights (indicating the confidence in edge direction) as well, and additionally, it determines interactions between upstream regulators that are relevant to explaining variations. Zarringhalam *et al* implemented similar ideas in a Bayesian framework [262]. Their proposed approach is however, limited to direct interactions (paths of length one). The authors also incorporated context dependence of edges in their study. Applying a similar statistical inference to genes connected with longer paths rapidly increases the computational complexity of the problem, and the information carried in cascades of interactions is inevitably ignored. Finally, several algorithms of using edge weights, fold change values, and type of paths between nodes (*i.e.* only shortest *vs.* all paths) were compared by Jaeger *et al* [263].

A clear limitation of these methods is their strong dependence on the quality of prior networks generated from available knowledge bases. Sign and direction information is scarce; meanwhile, many of such information heavily rely on the context. As a consequence, some of the methods use knowledge bases not publicly available (for example

from Ingenuity Inc. (http://www.ingenuity.com) or Selventa Inc. (http://www.selventa.com) with more curated information. This limitation may be temporary given the huge amount of high-throughput data generated these days; however, providing a well studied interaction network along with a high quality high-throughput data seem essential to benchmark the proposed methods.

Finally, the last group of methods in our classification contains methods that apply one layer/type of regulator-target detection at each step. Lefebvre *et al* constructed a network consisting of context specific transcription factors and their binding sites in human B-cell, and introduced a method called MARINA (MAster Regulator INference algorithm) to specify context specific transcription factors with regulatory roles [267]. Genes are ranked based on their down or up regulation magnitude, and the method evaluates if targets of a candidate transcription factor are enriched in top or bottom of the ranked list using GSEA [202]. The algorithms proposed in [268, 269] also investigate this layer of regulation, but they use different knowledge bases to build the initial graph, and they apply different enrichment techniques. In addition, these methods incorporate protein-protein interaction data bases to detect key proteins involving the differential regulation, as an additional layer. As a final step, EXPRESSION2KINASE method [270] employs kinase enrichment analysis to find kinases that potentially phosphorylate the input list of detected proteins in the previous layer of the analysis.

Similar to some methods discussed before, the strength of these methods also strongly depend on the quality of the generated influence graphs. The methods are very informative in providing clear insights in mechanisms underlying the regulation at different regulatory levels; however approaches that integrate information from multiple regulatory layers [260, 261] may be more sensitive in detecting weaker signals.

Most of the methods we discussed here are proposed to efficiently explain what are the minimal genes or pathways necessary for a cell to make a transition from state A to state B. However, it would be helpful to take several other snapshots of the transient states between the initial and final states as an additional guideline. For instance, when we want to study function of a gene as a component of a machinery, instead of only comparing the two states where the gene is active (control experiment) and when it is knocked down, it would be helpful to investigate the situation where the gene is 50% active. In the remaining part of

this chapter, we talk about response curves of cells upon inhibiting genes at different levels using pharmaceutical compounds. We present preliminary results showing that clustering of response patterns helps to identify gene functions, and in the discussion section, we explain how we think this type of data can be incorporated in causal reasoning inference algorithms.

4.3 Analyzing genes expression and splicing through inhibiting splicing components at multiple levels: preliminary results

So far, we have reviewed methods enabling extracting biological knowledge when two RNA-seq experiments are compared, regarding how to identify a smaller number of genes or pathways essential to attain the observed results. These two experiments, for example, may originate from a disease state and a normal state, or a knock down experiment and the analogous control experiment. In addition to the samples from the two conditions, other measurements from the intermediate conditions can also be informative to better understanding and tracking of variations. For instance, data may provide measurements at multiple time points of the transition, or when different inhibition levels are imposed.

In this section, we present results of analyzing a data set consisting of RNA-seq experiments of targeting proteins with pharmaceuticals. Target proteins of these pharmaceuticals are known to directly or indirectly influence splicing. We show that by appropriately using our data, we get consistent results when we investigate different cell lines, or different compounds that target the same protein.

4.3.1 Materials and methods.

Our data consists of samples from multiple concentrations of three pharmaceuticals. The first compound, T3, targets CDC-like kinases (CLKs) and has been shown to have a high specificity to CLK1-3 protein isoforms [271]. The two other compounds (T-202 and T-595) target EIF4A3 (Eukaryotic translation initiation factor 4A-III) protein.

EIF4A3 is known to play roles in translation initiation, splicing and ribosome assem-

bly [272, 273]. EIF4A3 data consists of RNA-seq libraries generated from 5 inhibition levels of EIF4A3 protein, each being applied with two different pharmaceutical compounds (T-202 and T-595) in two cell lines. These two cell lines are Hela (derived from cervical cancer) cell line and HCT-116 (derived from human colon carcinoma) cell line. In total, EIF4A3 drug data comprises 22 RNA-seq libraries: 2 control libraries, in addition to 20 drug treated libraries. Additionally, we have RNA-seq libraries from 3 different siRNAs directed to EIF4A3 in Hela cell line and also a corresponding control RNAi experiment. CLKs are also known to contribute to the regulation of splicing. Especially, phosphorylation by CLK proteins is required for SR proteins to facilitate their cooperation in splicing mechanism [274]. CLK data consists of libraries generated from treating HCT116 and hTert cell lines at three different concentrations of T3 compounds and two control libraries (one for each cell line). For CLK data, we use stranded libraries previously published by Funnell *et al* [271]. Figure 4.2 summarizes our drug RNA-seq libraries and our analysis workflow. These drugs have been developed by Takeda Pharmaceutical Company Limited and their specificity and efficacy were previously investigated [271, 275, 276] and the RNA-seq experimental procedures were previously explained [271].

The paired-end reads of our libraries were aligned to the reference genome (*hg19* reference genome downloaded from UCSC genome browser [43]) using GSNAP [197]. The corresponding gene annotation file was downloaded from ENSEMBL [159]. We enabled "novel splicing" parameter of GSNAP. Following the alignment step, duplicate reads were removed using SAMTOOLS [162, 198]. Next, gene and isoform abundance were computed by employing CUFFLINKS [167] package, resulting multiple FPKM values assigned to each gene based on the number of inhibition levels. The computed FPKM values for a cell line and a compound were combined to form gene responses upon multiple treatments.

Next, we applied WGCNA (Weighted correlation network analysis) [277] to cluster genes exhibiting correlated response patterns. We filtered genes and isoforms with FPKM value <1. Moreover, we only considered genes for which the maximum expression level is at least 50% larger than the smallest observed expression value in a set of experiments performed by changing compound levels. This was done to remove genes with small variations across treatments. Next, To determine gene functions, we applied GO enrichment analysis for gene clusters using BINGO [278]. BINGO takes a list of genes as input

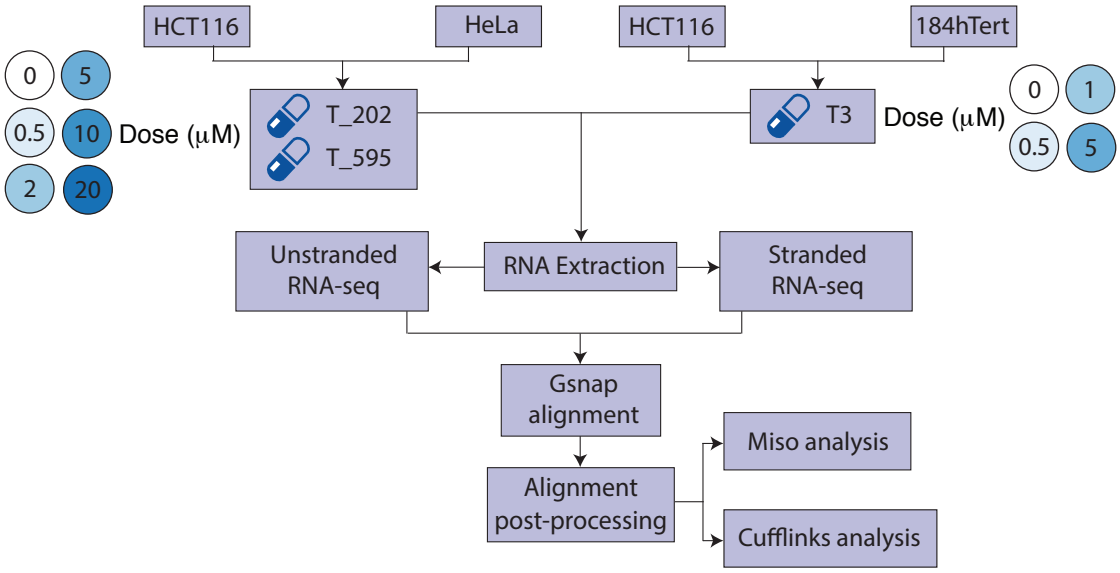

Figure 4.2: Our systematic approach to study proteins via gradual inhibition. For each protein two cell lines were treated with multiple concentration of pharmaceuticals for 6 hours. Stranded paired-end RNA-seq libraries were generated for CLK and unstranded paired-end RNA-seq data for EIF4A3 protein. Reads were aligned using GSNAP [197]; MISO [73] and CUFFLINKS [167] analyses were performed on each data set separately.

and examines the over-representation of genes in GO sets within those lists and reports a FDR (false discovery rate) value for each GO set. We consider GO terms with FDR value smaller than 0.05 as being statistically significant. Finally, to summarize enriched GO terms, we cluster them using ENRICHMENTMAP plugin [203] of CYTOSCAPE [204].

We used the MISO package [73] to find differential splicing events when drug treated RNA-seq samples were compared to control (no treatment) samples. As explained in previous chapters, MISO detects and differentiate 8 types of splicing events by applying a statistical framework. These event types are: skipped exons (SE), retained introns (RI), alternative 3'/5' splice sites (A3SS/A5SS), tandem 3' UTRs, mutually exclusive exons (MXE), alternative first exons (AFE), and also alternative last exons (ALE). The method assumes two potential splice variants for each event, and assigns a Ψ value (percent spliced in) to one of the isoforms in the two given conditions. Additionally, it reports a BF (Bayes

factor) value as a measure of confidence of them being differentially expressed. We filter events with $|\Delta\Psi|$ value smaller than 0.1 or events with BF value smaller than 20.

4.3.2 Results

Inhibiting our target proteins using pharmaceuticals imposes dose dependent splicing regulations. First, we investigated the regulation of alternative splicing upon increasing inhibition levels of the proteins. Figure 4.3 illustrates the number of identified differentially spliced events when treated samples were compared to control samples. All three inhibitors cause the increase of detected splicing events at higher compound concentrations as compared to lower concentrations. Moreover, the type of regulation is maintained in the different cell lines inspected and also with the two drugs targeting EIF4A3.

The results also suggest the distinct contribution of the two proteins in regulating different splicing types. Although the same database of splicing events was used when applying MISO pipeline, the proportion of splicing types regulated by proteins are different. While 4 AS types are almost equally abundant in EIF4A3 detected events, CLKs seem to predominantly regulate SE type. Additionally, a much larger number of AS events happen to be influenced by CLK inhibition. The results of these experiments can be utilized to determine genes and splicing regions that are more sensitive to disruption of a gene function. Events detected at lower drug levels may help in uncovering cis regulatory motifs related to a protein.

Inhibiting proteins with pharmaceutical compounds partially reproduces the results of knockdown experiments. We next sought to determine whether treating cells with drugs reproduces the results of knock down experiments with siRNAs. To compare the results, we took advantage of EIF4A3 data for which we have 3 different siRNAs targeting EIF4A3 transcripts, and the corresponding control siRNA. We paired data from each EIF4A3 siRNA to the control siRNA data and detected splicing events using the MISO package [73]. Events showing BF value ≥ 10 and $|\Delta\Psi| \geq 0.1$ for the three knockdown:control comparisons were reported, and the overlaps between them are represented in Figure 4.4.A. There are $\sim$31% of all events that are observed in at least two of the three knockdown:control comparisons.

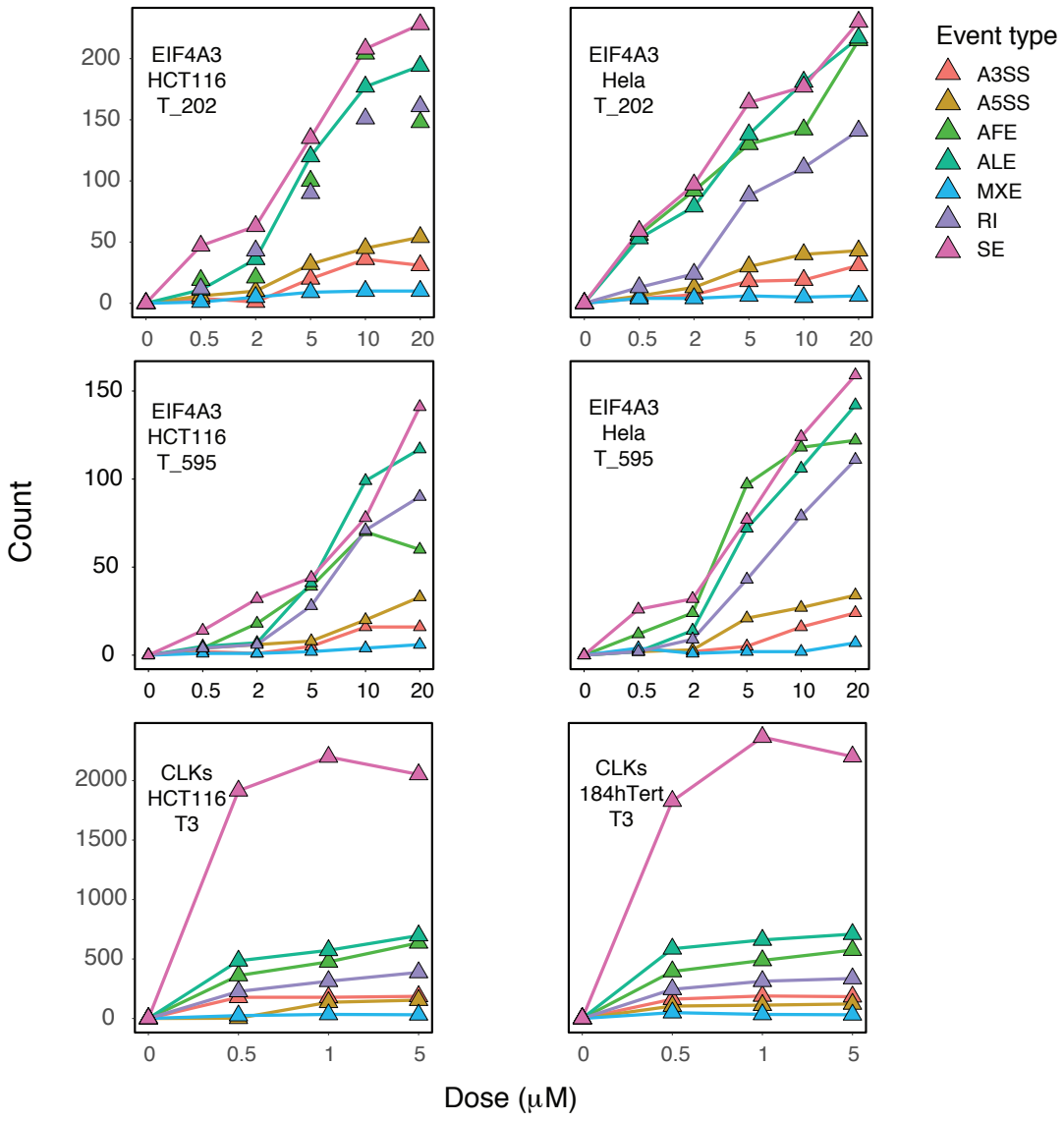

Figure 4.3: Splicing response patterns upon increasing inhibitor levels. The figure presents the number of splicing types at multiple inhibitor levels for three compounds. EIF4A3, and CLK proteins are inhibited at 5, and 3 different levels, respectively. For all types of splicing, the number of differentially spliced events generally increases by increasing the inhibitor level. Similar patterns of AS regulations are observed when EIF4A3 is inhibited in the two different cell lines and using the two distinct drugs; this pattern is different from the pattern of events undergoing AS regulation upon CLK inhibition

Figure 4.4.B shows a Venn diagram of the overlap between events found in the siRNA knockdown experiments and the drug inhibition experiments. We classified the identified events in the T-202 drug inhibition data into "low dose" and "high dose" groups based on the concentration of the treatment at which the events were predicted. Events detected at drug concentrations of 0.5 μM and 2 μM were classified as "low dose" events and the ones predicted at higher drug concentrations were classified as "high does" events. The majority of events detected in "low dose" group were also detected in "high dose" group (60%), however, only 38% of the events in the union of the three siRNAs were also identified in the drug inhibition experiments. The different specificity of siRNAs and the drug, their distinct off-target effects added to the uncertainties and the inevitable noise in RNA-seq data can explain some of the sources of the observed dissimilarity. This shows the importance of performing independent experiments to characterize more confident or more sensitive events as opposed to events less dependent on the gene of interest or potential artifacts.

Genes showing monotonic responses in different cell lines account for similar functions. In order to assess the potential of utilizing multiple level inhibition data in exploring gene and drug functions, we clustered gene response patterns in each cell line and compound in EIF4A3 inhibition data, using WGCNA [277]. Figure 4.5 represents clusters sorted based on the number of genes in them. For 3 out of 4 of our compound-cell line pair experiments, there exist two dominant clusters consisting of genes following a general monotonically increasing or monotonically decreasing patterns. The two patterns is indeed observed in the remaining compound-cell line experiment as well constituting two of the top four dominant clusters.

Here, we assume primary targets of a protein tend to show monotonic responses upon increasing inhibition much more than a set of randomly selected genes. Secondary effects, or random genes are expected to receive the inhibition signal at a lower amount, only after several other rounds of regulations were imposed on the signal. Based on this idea, we performed GO enrichment analysis by applying BINGO [278] to uncover primary functions of EIF4A3 protein in the four sets of libraries that we have.

Parts A and B in Figure 4.6 show the enriched GO terms found by BINGO method for the set of monotonically increasing and monotonically decreasing genes. For this analysis, we used the libraries provided by applying the first compound to the HCT116 cell

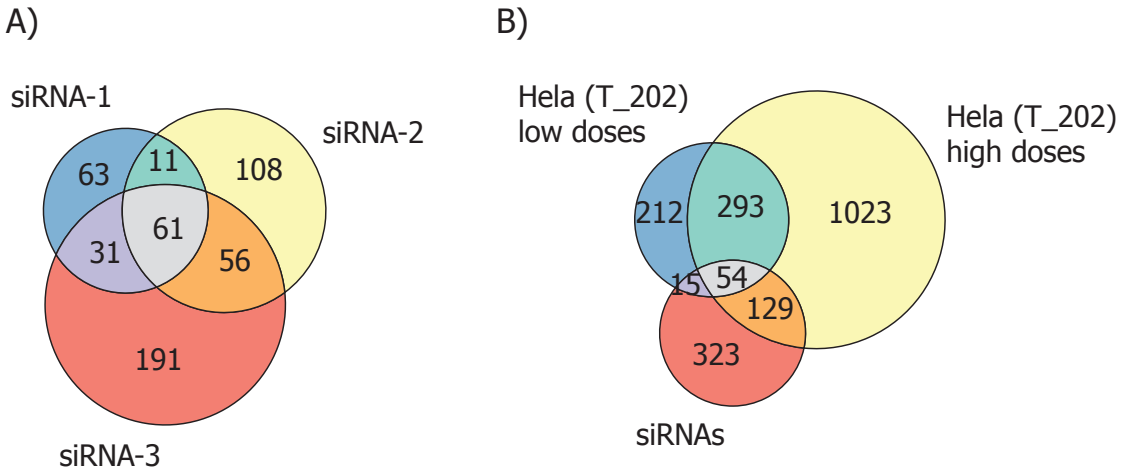

Figure 4.4: The overlap of the splicing events detected in the inhibitor and siRNA experiments. **A.** A venn diagram showing the overlap of the detected events between the three different siRNA experiments where EIF4A3 was knocked down. **B.** The overlap between the results of knocking down EIF4A3 with siRNAs, inhibiting EIF4A3 with low drug concentrations, and inhibiting EIF4A3 with high drug concentrations. Almost 38% of the siRNA knockdown events were also detected in the drug inhibition experiments.

line. The GO terms (one node per each term) are clustered based on common genes in them which are also present in the list of monotonically changing genes. The enriched terms for up-regulated genes include: regulation of metabolic process, regulation of transcription and gene expression, DNA damage response, regulation of kinase activity and some others. Similarly, the list of enriched terms for down-regulated genes include: regulation of cell cycle, protein localization, regulation of signal transduction and also some common enriched terms with GO terms for up-regulated genes such as the regulation of metabolic processes and gene expression.

Next, we analyzed the other EIF4A3 RNA-seq libraries generated using the other compound or in the other cell line to check if terms and functions associated with EIF4A3 could also be retrieved in the other datasets. We replicated the GO enrichment analysis in the 3 remaining combination of compounds-cell lines (2 compounds and 2 cell lines), and found

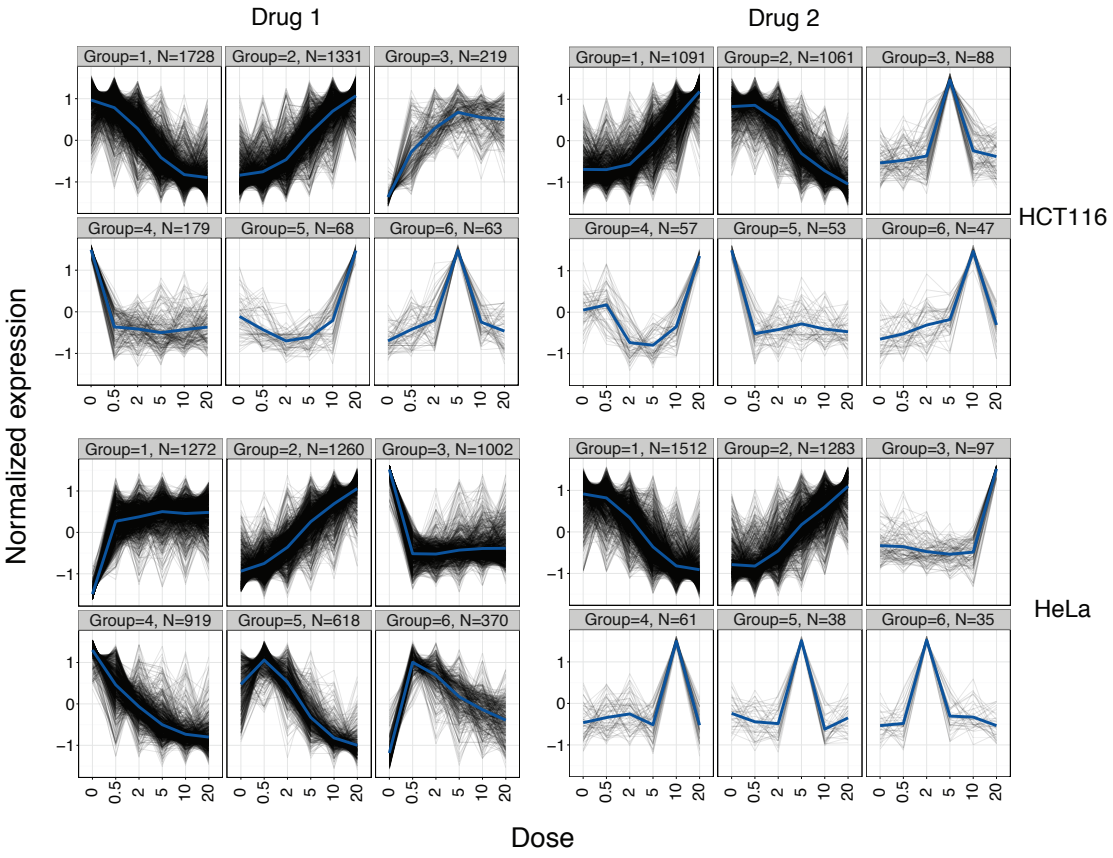

Figure 4.5: Clustering of expression response patterns upon inhibiting EIF4A3. Results are presented for distinct compounds in two cell lines. Gene expression values are clustered using WGCNA [277]. For each case, only the six clusters with the largest number of genes are shown. In each one of our experiments, two of the largest clusters can be attributed to genes showing mostly monotonically increasing or decreasing responses. The blue line demonstrates the consensus response pattern for each of the clusters (by connecting average values of gene expression at different compound levels).

many common enriched terms among the four experiments. For instance, parts C and D in Figure 4.6 illustrates the -log10 false discovery rate for the top 15 GO terms identified in the list of genes showing monotonically increasing expression patterns in HCT116 cell line, when the cells were treated by the first compound. Additionally, their existence or absence in the other data sets are also presented in the figure. Out of the 15 top enriched terms in HCT116 cell lines treated with compound 1, 14 of them are also detected in the three other libraries for up-regulated genes (part B). For the terms enriched in the list of down-regulated genes, 12 out of 15 were detected in all three other data sets, and another term was detected in two of the other data sets as well.

To further assess if the type of our data can uncover functions of a targeted protein, we inspected a previously known function of EIF4A3 in our data. EIF4A3 is known to be a core component of exon junction complex, an important member of nonsense mediated decay (NMD) mechanism. Through NMD, mRNA molecules that contain premature stop codons are eliminated before being translated. Inhibiting EIF4A3 intervenes with NMD, thus the isoforms that are supposed to undergo NMD are expected to be expressed more.

To analyze the consequence of EIF4A3 inhibition on NMD using our data, we first extracted $\sim$14,000 isoforms known to undergo NMD from ENSEMBL data base. Next, similar to our gene expression analysis, we clustered isoform expressions for the isoforms having average FPKM value ≥ 1 and median value ≥ 0. Figure 4.7 shows WGCNA clustering results. Unlike gene expression clusters where we usually found two dominant clusters with both up- and down-regulated genes, here for all the experiments, we only found 1 dominant cluster that predominantly contains genes with monotonically increasing expression patterns. Moreover, we confirmed that the observed pattern cannot be associated to the up-regulation of the corresponding genes (results not shown), and therefore, the up-regulation may be mainly attributed to in-activation of NMD process. Thus, the clustering of isoform expression patterns confirms a known function of EIF4A3.

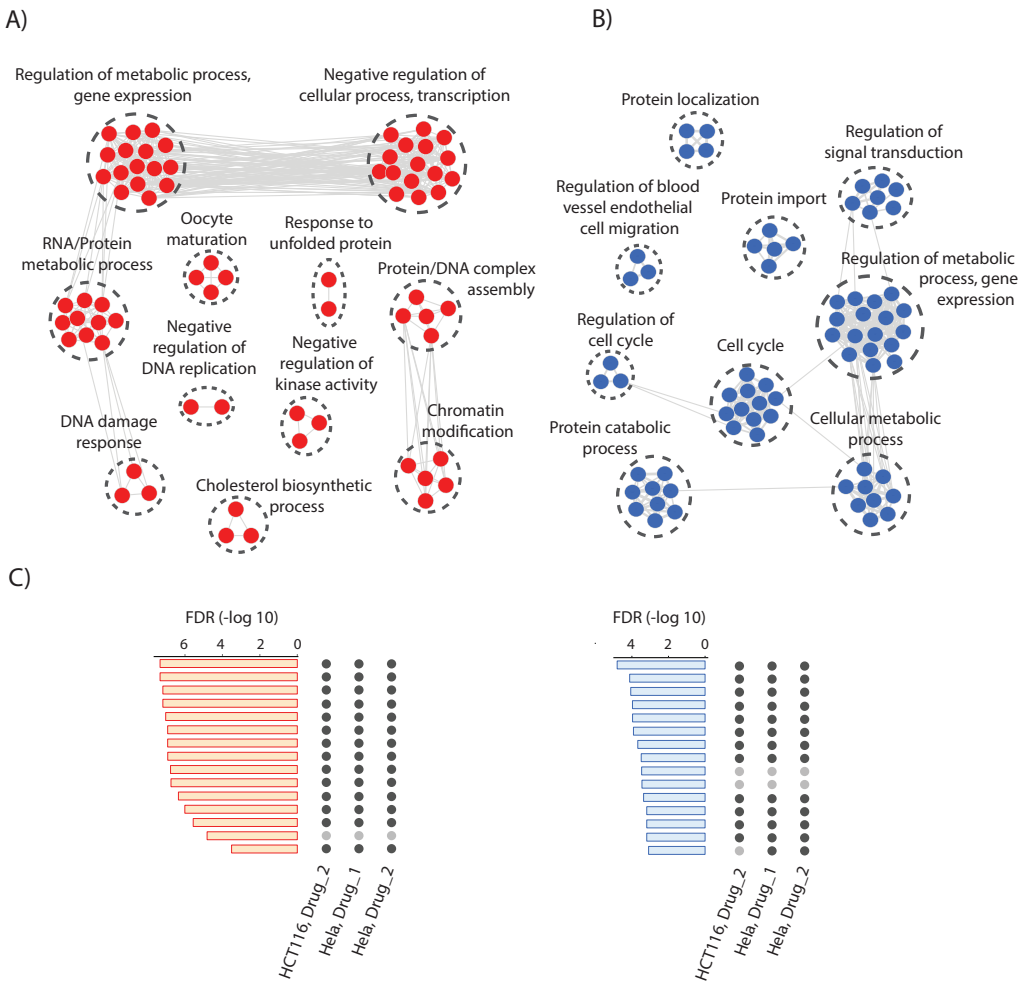

Figure 4.6: GO enrichment analysis for clusters of genes showing similar expression change pattern. **A.** For the cluster of genes with monotonically increasing consensus pattern after inhibiting EIF4A3 with the first inhibitor in HCT116 cell line, we performed GO enrichment analysis using BINGO. Each node represents a GO term enriched in our analysis with false discovery rate ≤ 0.05, and edges show gene sets with common genes present in the input list. GO terms are clustered using ENRICHMENTMAP software. **B.** Similar analysis as in part **A** was carried out for genes in the cluster of monotonically decreasing consensus pattern. **C.** For the top 15 GO terms with lowest FDR, we checked if replicating the analysis with the other compound, or in the other cell line could detect similar GO terms. Bar plots illustrate FDR values; the black circle indicates the same term was also detected in the corresponding data with ≤ 0.05, and the grey circle indicates that the same GO term was not detected.

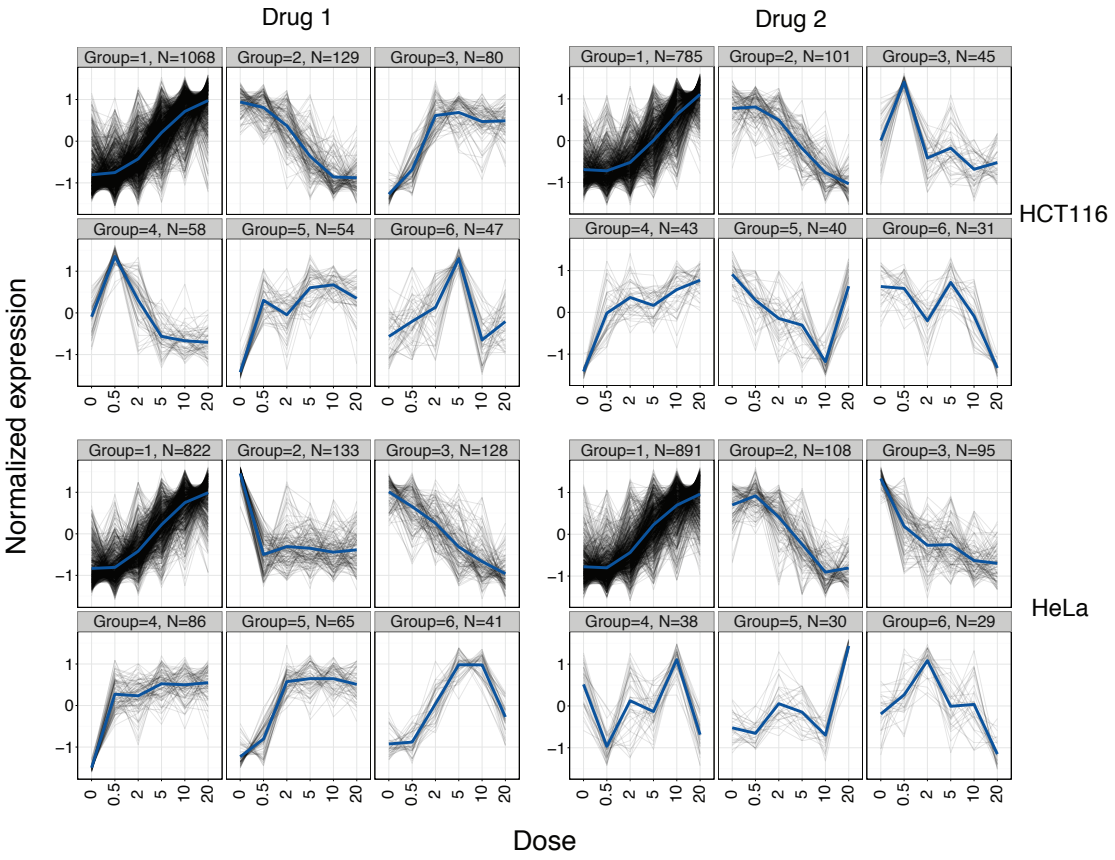

Figure 4.7: Clustering of NMD isoforms response patterns upon inhibiting EIF4A3. Results are presented for distinct compounds in two cell lines. Expression profiles of isoforms known to undergo NMD are clustered using WGCNA [277]. In each case, only the six clusters with the largest number of genes are shown. In contrast to the clustering of gene expression where we observed two dominant clusters (Figure 4.5), here there only exists one dominant cluster constituting genes with monotonically increasing response patterns. The blue line demonstrates the consensus response pattern for each of the clusters (by connecting average values of isoform expression at different drug levels).

4.4 Discussion

In this chapter, we discussed an important goal of molecular Biology research: understanding functions and regulations of genes. We reviewed methods developed to inferring functional and regulatory knowledge from high-throughput sequencing data. Based on the power and limitations of methods discussed, the type of experiment, and the research question, the appropriate method should be employed. With the improvement in technology and the reduction of sequencing costs, data is being generated at a much faster rate. Therefore, the methods should also be adopted to benefit from the amount of extra information available.

Methods performing mechanistic inference reviewed here have been successfully applied to improve our understanding of how an specific response emerges when a condition is modified [257, 267]. Despite being helpful, the methods have some limitations as well. Our knowledge on biological interactions essential for the success of the discussed methods is still incomplete. Additionally, many of the known interactions are indeed context specific without the context being specified in public data bases. Fortunately, the increasing amount of data generated these days seem to make many of such limitations to be only temporary.

We also presented our RNA-seq data consisting of inhibiting proteins at multiple levels. Advancement in therapeutics has made similar data sets much more abundant than before, and consequently, adopting methods to incorporate dose dependent responses in computational analyses is of huge interest. Most of the methods reviewed are only intended to handle situations where two conditions are compared; thus not being optimized to benefit from the extra information provided by inducing various inhibition levels.

Our preliminary analysis showed that increasing inhibition levels of the genes we investigated imposes gradual effects, both at the splicing level and at the expression level. This type of effect can be further investigated to realize primary functions of targets and differentiate them from secondary consequences. When applying a correlation-based clustering methods (WGCNA), the results suggested that the data could be engaged to consistently derive gene functions when distinct compounds and cell lines were used.

Appropriately modifying methods performing mechanistic inference can enhance our

findings using pharmaceutical inhibition data. In order to benefit from these methods, we need to determine which interactions among the known prior interactions are active in a given condition, and accordingly which genes are being regulated by a given candidate gene. An obvious approach is to define interactions based on response correlations, instead of considering direction and the magnitude of changes for the interacting genes in the two conditions case. One issue with using correlation based methods is that they assume linear dependencies among responses which can be violated [234].

Hidden Markov models [279] (HMMs) are also appropriate tools to model the sequence of observed responses in our data. HMMs have been extensively applied to problems where there could be a long range dependencies among a sequence of observations. For instance, the observations can be presented by a series of fold changes and hidden states (which control the generative probabilistic components explaining observations) can take three values: "Up", "Down", and "No change", indicating whether the gene is up-regulated, down-regulated or there is no change compared to the previous inhibitor level. Besides, a probability distribution is assigned to each hidden state from which the observations are derived. Finally, the probability of each possible path (a sequence of "Up", "Down" and "No Change"s) could be calculated to assess the probability of genes showing similar patterns of responses.

In a recent study, Leng *et al* [280] proposed auto-regressive hidden Markov models (AR-HMM) to infer probability of potential paths (a sequences of "up", "Down" and "No change"'s). The model allows to capture the dependence of an observed FPKM value or a read count value in an experiment not only based on the current hidden state (up-regulation, down-regulation or no change), but also on the observed previous FPKM value as well (Figure 4.8). As an extension, separate models could be designed and trained for potential regulation between any two given genes in interaction data bases. Paired observations (transcript abundances of two interacting genes) are derived based on the hidden state of the upstream gene in each model, and the best model describing the observations can define the type of interaction.

In this chapter we have taken the first steps towards developing methods that in future can help to study biological systems, drug effects, and gene functions with the increasing amount of data provided by pharmaceutical agents. We discussed the existing methods,

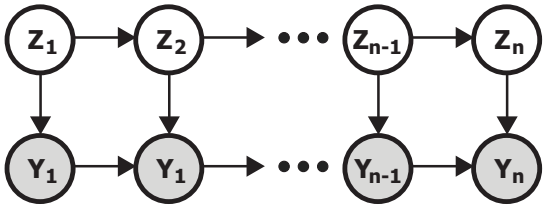

Figure 4.8: An auto-regressive hidden Markov model proposed by Leng *et al* [280] to analyze ordered high-throughput sequencing data. Z_i's show hidden states and can take values from "Up", "Down" and "No change" to represent direction of change between consecutive observations. Shaded circles (Y_i's) represent observations that could report FPKM values or read counts per each gene. Connected nodes enable modelling dependencies among an ordered set of observations.

our type of data, preliminary analysis on their usefulness, and also the way we think the data should be incorporated in the existing pipelines.

Chapter 5

Conclusion

In this thesis, I took a systems biology approach to investigate functions and regulations of alternative splicing. Through AS mechanism, cells expand the capacity of their genomes and orchestrate complex responses. The regulated interplay between components of splicing machinery is essential to maintain normal cellular functions, and consequently, many of the genetic diseases have been associated to impaired splicing. Our approach offers new insight on how AS is regulated and also how it affects related mechanisms. Our study provides additional perspective towards a more comprehensive picture of alternative splicing.

Advancement in high-throughput sequencing technologies and the development of cost-effective methods has brought new opportunities to better understanding of AS mechanism. In all research questions explored here, we benefited from RNA-seq libraries to perform a genome-wide identification of AS events and the corresponding global consequences on transcriptome regulation. Additionally, by taking advantage of replicated experiments, multiple cell lines, and state of the art computational methods, we addressed limitations and uncertainties of RNA-seq data.

In Chapter 2, I presented our findings on tissue specific RNA editing in *Drosophila melanogaster* and its potential role in regulating alternative splicing. We designed a pipeline that utilizes large input data and ADAR's requirement for double-stranded targets to distinguish genuine editing sites from mapping and sequencing errors. We showed that editing events happen 3 times more frequently in exons with multiple acceptor/donor sites

than exons with unique splice site. This finding demonstrates a potential inter-relation between AS and RNA editing. Next, we searched conserved secondary structures in regions where alternative splicing and RNA editing co-occur, and reported conserved structures that may mediate their inter-relation. Our research suggests a tissue specific and gene specific regulation of alternative splicing by RNA editing mediated through formation of RNA structures.

Considering the huge number of editing sites that have been already reported in human, exploring a similar hypothesis in human in future can uncover regions where a similar inter-relation may happen. Additionally, it should be noted that in our study, we used mRNA libraries (poly-A enriched) where most intronic signals were removed. In future studies, using pre-mRNA sequencing data enables investigating editing in more detail, especially in human, where a large number of editing sites have been predicted to happen in intronic regions [104, 113].

In a different prospective, our study identifies RNA structures that form *in vivo*. Although potential RNA structures can be predicted computationally, it is hard to determine whether they actually form *in vivo* in an specific tissue, or at a given time. However, we know ADAR requires double stranded structures which confirms the formation of structures. Once these structures are detected, their potential roles in regulating splicing or their relevance to diseases regardless of RNA editing can be further analyzed. Furthermore, in future studies, mutational experiments will be required in order to validate the importance of these structures in regulating splicing.

In chapter 3, we studied the roles of CDK12 in regulating RNA splicing and transcription. Our RNA-seq data demonstrate that CDK12 expression predominantly influence splicing by regulating the differential usage of alternative last exons. The regulation could be modulated either at the transcription or splicing level. Furthermore, our proteomics data indicates that CDK12 interacts with the components of splicing machinery, especially those associated with splice site selection. We showed that long genes with many exons constitute differentially regulated genes upon knocking down CDK12. We also showed that the regulation of gene expression by CDK12 is tissue specific, however, common pathways are influenced in the two cell lines that we analyzed. DNA damage response genes are one class of common affected genes. We analyzed TCGA data and showed

that the regulation of alternative last exon events that we found in our data could also be observed when comparing samples from CDK12 mutant patients to control patients.

In future studies, our findings on the differential regulation of *ATM* and *DNAJB6* and their potential contribution to the tumorigenicity of breast cancer cells can be further investigated to better understand tumor biology of breast cancer cells harboring genomic alterations in CDK12. Additionally, our study is limited in providing mechanisms that regulate the tissue specific splicing of events such as the one happening in *DNAJB6* which should be taken into account in future studies.

In this study, we only considered splicing events that are already annotated and are present in MISO [73] database. Using methods that enable discovering *de novo* splicing events as discussed in chapter 1 can further increase the number of identified regulated genes, and might help to infer more plausible models to explain functions of *CDK12*. Also, chip-seq experiments can be used to measure the occupancy of RNA polymerase II across the genome for the same cell lines to help distinguish events that are a consequence of disruption in transcription elongation rate and the other ones.

In chapter 4, I presented a review on methods developed to perform mechanistic inference using high-throughput sequencing data, and methods that try to identify a small set of genes and pathways that control the transition between the two given conditions. The methods have been successfully applied to uncover how a response emerges by modifying conditions. Our review provides a guideline to choose appropriate methods based on research questions and available data sets. We also introduced our data sets where genes known to directly or indirectly affecting splicing regulations were progressively inhibited using multiple concentrations of the inhibitor. Using clustering of response patterns and applying gene set enrichment analysis, we showed the data can contribute to exploring functions and regulations of proteins.

By the advancement in therapeutics and decreasing sequencing cost, this type of data will become more accessible. Although we discussed how the reviewed methods could be adopted to incorporate the additional information provided by the introduced data, so far no method has been developed. The next step would be to develop methods (*e.g.* HMM based methods) to benefit more from the additional information provided by the systematic gene inhibition using pharmaceuticals.

Taken together, our systems biology approach in this thesis provides additional insight on regulations and functions of alternative splicing. I hope this study can motivate further investigation of mechanisms discussed and their roles in associated diseases, and eventually lead to the advancement in therapeutics.

Bibliography

[1] Mazloomian, A. & Meyer, I.M., 2015. Genome-wide identification and characterization of tissue-specific RNA editing events in D. melanogaster and their potential role in regulating alternative splicing. *RNA biology* **12**(12): 1391–1401. → pages iv

[2] Tien, J.F., Mazloomian, A., Cheng, S., Hughes, C.S., Chow, C., Canapi, L.T., Oloumi, A., Trigo-Gonzalez, G., Bashashati, A., Xu, J. *et al.*, 2017. CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic acids research* . → pages iv, 77

[3] Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., & Shoemaker, D.D., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**(5653): 2141–2144. → pages 1

[4] Krawczak, M., Reiss, J., & Cooper, D.N., 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Human genetics* **90**(1-2): 41–54. → pages 1

[5] Venables, J.P., 2004. Aberrant and alternative splicing in cancer. *Cancer research* **64**(21): 7647–7654. → pages 1, 81

[6] Marguerat, S. & Bähler, J., 2010. RNA-seq: from technology to biology. *Cellular and molecular life sciences* **67**(4): 569–579. → pages 1

[7] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., & Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**(7): 621–628. → pages 1

[8] Kaletta, T. & Hengartner, M.O., 2006. Finding function in novel targets: *C. elegans* as a model organism. *Nature Reviews Drug Discovery* **5**(5): 387–399. → pages 2

[9] Pandey, U.B. & Nichols, C.D., 2011. Human disease models in *Drosophila melanogaster* and the role of the fly in therapeutic drug discovery. *Pharmacological reviews* **63**(2): 411–436. → pages

[10] Chintapalli, V.R., Wang, J., & Dow, J.A., 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature genetics* **39**(6): 715–720. → pages 2

[11] Smit, A.F., Hubley, R., & Green, P., 1996. RepeatMasker. *Published on the web at http://www repeatmasker org* . → pages 2

[12] Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. *et al.*, 2016. Ensembl 2016. *Nucleic acids research* **44**(D1): D710–D716. → pages 2

[13] Shoemaker, R.H., 2006. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer* **6**(10): 813–823. → pages 2

[14] Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D. *et al.*, 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**(7391): 603–607. → pages 2

[15] Berget, S.M., Moore, C., & Sharp, P.A., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* **74**(8): 3171–3175. → pages 3

[16] Chow, L.T., Gelinas, R.E., Broker, T.R., & Roberts, R.J., 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**(1): 1–8. → pages 3

[17] Han, J., Xiong, J., Wang, D., & Fu, X.D., 2011. Pre-mRNA splicing: where and when in the nucleus. *Trends in cell biology* **21**(6): 336–343. → pages 3

[18] Lim, L. & Burge, C., 2001. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences* **98**(20): 11193–11198. → pages 3

[19] Black, D., 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry* **72**(1): 291–336. → pages 3

[20] Smith, C.W. & Valcárcel, J., 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in biochemical sciences* **25**(8): 381–388. → pages 3

[21] Horowitz, D.S., 2012. The mechanism of the second step of pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA* **3**(3): 331–350. → pages 3, 4

[22] Will, C.L. & Lührmann, R., 2011. Spliceosome structure and function. *Cold Spring Harbor perspectives in biology* **3**(7): a003707. → pages 4

[23] Wahl, M.C., Will, C.L., & Lührmann, R., 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**(4): 701–718. → pages 4

[24] Nilsen, T., 2003. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays* **25**(12): 1147–1149. → pages 4

[25] Nilsen, T.W. & Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**(7280): 457–463. → pages 5

[26] Faustino, N. & Cooper, T., 2003. Pre-mRNA splicing and human disease. *Genes & development* **17**(4): 419–437. → pages 5, 9, 81

[27] Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R. *et al.*, 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**(6114): 1587–1593. → pages 5

[28] Yeo, G., Holste, D., Kreiman, G., & Burge, C.B., 2004. Variation in alternative splicing across human tissues. *Genome biology* **5**(10): 1. → pages 5

[29] Raj, B. & Blencowe, B.J., 2015. Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles. *Neuron* **87**(1): 14–27. → pages 5

[30] Vuong, C.K., Black, D.L., & Zheng, S., 2016. The neurogenetics of alternative splicing. *Nature Reviews Neuroscience* **17**(5): 265–281. → pages 5

[31] Wang, Z. & Burge, C.B., 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**(5): 802–813. → pages 5, 7

[32] Keren, H., Lev-Maor, G., & Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* **11**(5): 345–355. → pages 6

[33] McManus, C.J., Coolon, J.D., Eipper-Mains, J., Wittkopp, P.J., & Graveley, B.R., 2014. Evolution of splicing regulatory networks in *Drosophila*. *Genome research* **24**(5): 786–796. → pages 6

[34] Ast, G., 2004. How did alternative splicing evolve? *Nature Reviews Genetics* **5**(10): 773–782. → pages 6

[35] Lee, Y. & Rio, D.C., 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry* **84**: 291–323. → pages 6

[36] McManus, C.J. & Graveley, B.R., 2011. RNA structure and the mechanisms of alternative splicing. *Current opinion in genetics & development* **21**(4): 373–379. → pages 6

[37] Kim, E., Goren, A., & Ast, G., 2008. Alternative splicing: current perspectives. *Bioessays* **30**(1): 38–47. → pages 6

[38] Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W. *et al.*, 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**(7339): 473–479. → pages 6, 29, 36, 37, 38, 39, 45

[39] Reiter, L.T., Potocki, L., Chien, S., Gribskov, M., & Bier, E., 2001. A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. *Genome research* **11**(6): 1114–1125. → pages 6

[40] McQuilton, P., Pierre, S., Thurmond, J. *et al.*, 2012. FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Research* **40**(D1): D706–D714. → pages 6, 7

[41] Gibilisco, L., Zhou, Q., Mahajan, S., & Bachtrog, D., 2016. The evolution of alternative splicing in *Drosophila*. *bioRxiv* page 054700. → pages 6

[42] Celotto, A.M. & Graveley, B.R., 2001. Alternative splicing of the Drosophila *Dscam* pre-mRNA is both temporally and spatially regulated. *Genetics* **159**(2): 599–608. → pages 7

[43] Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.*, 2003. The UCSC genome browser database. *Nucleic acids research* **31**(1): 51–54. → pages 7, 91

[44] Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S. *et al.*, 2016. The UCSC Genome Browser database: 2016 update. *Nucleic acids research* **44**(D1): D717–D725. → pages 7

[45] Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F. *et al.*, 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**(6012): 1787–1797. → pages 7, 30

[46] Matlin, A.J., Clark, F., & Smith, C.W., 2005. Understanding alternative splicing: towards a cellular code. *Nature reviews Molecular cell biology* **6**(5): 386–398. → pages 7

[47] Maniatis, T. & Tasic, B., 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**(6894): 236–243. → pages 7

[48] Naftelberg, S., Schor, I.E., Ast, G., & Kornblihtt, A.R., 2015. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annual review of biochemistry* **84**: 165–198. → pages 7

[49] Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., & Muñoz, M.J., 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews Molecular cell biology* **14**(3): 153–165. → pages 8, 11

[50] Neugebauer, K.M., 2002. On the importance of being co-transcriptional. *Journal of cell science* **115**(20): 3865–3871. → pages 8

[51] Lai, D., Proctor, J.R., & Meyer, I.M., 2013. On the importance of cotranscriptional RNA structure formation. *RNA* **19**(11): 1461–1473. → pages 8, 22

[52] Baralle, D. & Baralle, M., 2005. Splicing in action: assessing disease causing sequence changes. *Journal of medical genetics* **42**(10): 737–748. → pages 8, 9

[53] Buratti, E. & Baralle, F.E., 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. *Molecular and cellular biology* **24**(24): 10505–10514. → pages 8, 9, 10

[54] Pervouchine, D., Khrameeva, E., Pichugina, M., Nikolaienko, O., Gelfand, M., Rubtsov, P., & Mironov, A., 2012. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA* **18**(1): 1–15. → pages 8

[55] Meyer, I. & Miklós, I., 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic acids research* **33**(19): 6338–6348. → pages 8, 40, 49

[56] Tazi, J., Bakkour, N., & Stamm, S., 2009. Alternative splicing and disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1792**(1): 14–26. → pages 9

[57] Singh, N.N., Androphy, E.J., & Singh, R.N., 2004. An extended inhibitory context causes skipping of exon 7 of *SMN2* in spinal muscular atrophy. *Biochemical and biophysical research communications* **315**(2): 381–388. → pages 9

[58] Garcia-Blanco, M., Baraniak, A., & Lasda, E., 2004. Alternative splicing in disease and therapy. *Nature biotechnology* **22**(5): 535–546. → pages 9

[59] Fu, X.D. & Ares Jr, M., 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* **15**(10): 689–701. → pages 11

[60] Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S., 2013. Function of alternative splicing. *Gene* **514**(1): 1–30. → pages 11

[61] Irimia, M. & Blencowe, B.J., 2012. Alternative splicing: decoding an expansive regulatory layer. *Current opinion in cell biology* **24**(3): 323–332. → pages 11

[62] Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T., & Soreq, H., 2005. Function of alternative splicing. *Gene* **344**: 1–20. → pages 11

[63] Treangen, T.J. & Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**(1): 36–46. → pages 11

[64] Oshlack, A. & Wakefield, M., 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biology direct* **4**(4): 14. → pages 11

[65] Garber, M., Grabherr, M.G., Guttman, M., & Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* **8**(6): 469–477. → pages 11

[66] Ryan, M.C., Cleland, J., Kim, R., Wong, W.C., & Weinstein, J.N., 2012. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics* **28**(18): 2385–2387. → pages 11

[67] Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R., Morrissy, A.S., Morin, R.D., Corbett, R., Tang, M.J., Hou, Y.C., Pugh, T.J. *et al.*, 2010. Alternative expression analysis by RNA sequencing. *Nature methods* **7**(10): 843–847. → pages

[68] Glaus, P., Honkela, A., & Rattray, M., 2012. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**(13): 1721–1728. → pages

[69] Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z.x., Zhou, Q., Carstens, R.P., & Xing, Y., 2012. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic acids research* page gkr1291. → pages

[70] Aschoff, M., Hotz-Wagenblatt, A., Glatting, K.H., Fischer, M., Eils, R., & König, R., 2013. SplicingCompass: differential splicing detection using RNA-Seq data. *Bioinformatics* **29**(9): 1141–1148. → pages

[71] Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., & Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* **31**(1): 46–53. → pages 12

[72] Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., & Kendziorski, C., 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**(8): 1035–1043. → pages 12

[73] Katz, Y., Wang, E.T., Airoldi, E.M., & Burge, C.B., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7**(12): 1009–1015. → pages 12, 54, 57, 92, 93, 106

[74] Anders, S., Reyes, A., & Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. *Genome research* **22**(10): 2008–2017. → pages 12, 35, 45, 46, 135

[75] Wang, W., Qin, Z., Feng, Z., Wang, X., & Zhang, X., 2013. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* **518**(1): 164–170. → pages 12

[76] Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.F., Hammond, S.M., Makowski, L. *et al.*, 2013. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic acids research* **41**(2): e39–e39. → pages 12

[77] Alamancos, G.P., Agirre, E., & Eyras, E., 2014. Methods to study splicing from high-throughput RNA Sequencing data. *Spliceosomal Pre-mRNA Splicing: Methods and Protocols* pages 357–397. → pages 12

[78] Hooper, J.E., 2014. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Human Genomics* **8**(1): 1–6. → pages 13

[79] Gray, M.W., 2012. Evolutionary origin of RNA editing. *Biochemistry* **51**(26): 5235–5242. → pages 13

[80] Benne, R., Van Den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H., & Tromp, M.C., 1986. Major transcript of the frameshifted *coxll* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**(6): 819–826. → pages 13

[81] Scadden, A., 2005. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nature structural & molecular biology* **12**(6): 489–496. → pages 13

[82] Farajollahi, S. & Maas, S., 2010. Molecular diversity through RNA editing: a balancing act. *Trends in Genetics* **26**(5): 221–230. → pages 14

[83] Blow, M., Futreal, P.A., Wooster, R., & Stratton, M.R., 2004. A survey of RNA editing in human brain. *Genome research* **14**(12): 2379–2387. → pages 14, 17, 34

[84] Nishikura, K., 2010. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annual Review of Biochemistry* **79**: 321–349. → pages 14, 16, 20, 34, 35

[85] Barraud, P. & Allain, F., 2012. ADAR Proteins: Double-stranded RNA and Z-DNA Binding Domains. *Current Topics in Microbiology and Immunology* **353**: 35–60. → pages 14, 15, 16, 29

[86] Ramaswami, G. & Li, J.B., 2014. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research* **42**(D1): D109–D113. → pages 15

[87] Paro, S., Li, X., O'Connell, M., & Keegan, L., 2012. Regulation and functions of ADAR in *Drosophila*. *Current topics in microbiology and immunology* **353**: 221–236. → pages 16, 28, 29, 49

[88] Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B. *et al.*, 2010. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**(7339): 473–479. → pages 16

[89] Rodriguez, J., Menet, J.S., & Rosbash, M., 2012. Nascent-seq Indicates Widespread Cotranscriptional RNA Editing in *Drosophila*. *Molecular cell* **47**(1): 27–37. → pages 16, 17, 29, 38, 39, 40, 45, 49, 135

[90] Bass, B.L., 2002. RNA editing by adenosine deaminases that act on RNA. *Annual review of biochemistry* **71**: 817–846. → pages 16, 20

[91] Bass, B.L., 1997. RNA editing and hypermutation by adenosine deamination. *Trends in biochemical sciences* **22**(5): 157–162. → pages 16

[92] Maas, S., Godfried Sie, C., Stoev, I., Dupuis, D., Latona, J., Porman, A., Evans, B., Rekawek, P., Kluempers, V., Mutter, M. *et al.*, 2011. Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing sites. *Biochemical and biophysical research communications* **412**(3): 407–412. → pages 16, 34

[93] Neeman, Y., Levanon, E.Y., Jantsch, M.F., & Eisenberg, E., 2006. RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA* **12**(10): 1802–1809. → pages 16, 19, 34

[94] Hoopengardner, B., Bhalla, T., Staber, C., & Reenan, R., 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science's STKE* **301**(5634): 832–836. → pages 17, 18, 40, 43

[95] Rieder, L.E. & Reenan, R.A., 2012. The intricate relationship between RNA structure, editing, and splicing. In *Seminars in cell & developmental biology*, volume 23, pages 281–288. Elsevier. → pages 17, 20, 26, 28

[96] Athanasiadis, A., Rich, A., & Maas, S., 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS biology* **2**(12): e391. → pages 17

[97] Morse, D.P., Aruscavage, P.J., & Bass, B.L., 2002. RNA hairpins in noncoding regions of human brain and *Caenorhabditis elegans* mRNA are edited by adenosine deaminases that act on RNA. *Proceedings of the National Academy of Sciences* **99**(12): 7906–7911. → pages 17, 34, 35

[98] Danecek, P., Nellåker, C., McIntyre, R.E., Buendia-Buendia, J.E., Bumpstead, S., Ponting, C.P., Flint, J., Durbin, R., Keane, T.M., & Adams, D.J., 2012. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biology* **13**(4): 1–12. → pages 17, 20, 22, 29, 30, 34, 35

[99] Wan, Y., Kertesz, M., Spitale, R., Segal, E., & Chang, H., 2011. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics* **12**(9): 641–655. → pages 17

[100] Yang, Y., Sun, F., Wang, X., Yue, Y., Wang, W., Zhang, W., Zhan, L., Tian, N., Jin, Y. *et al.*, 2012. Conservation and regulation of alternative splicing by dynamic inter-and intra-intron base pairings in Lepidoptera 14-3-3z pre-mRNAs. *RNA biology* **9**(5): 691–700. → pages 17

[101] Daniel, C., Venø, M.T., Ekdahl, Y., Kjems, J., & Öhman, M., 2012. A distant cis acting intronic element induces site-selective RNA editing. *Nucleic Acids Research* **40**(19): 9876–9886. → pages 17, 34

[102] Levanon, E., Hallegger, M., Kinar, Y., Shemesh, R., Djinovic-Carugo, K., Rechavi, G., Jantsch, M., & Eisenberg, E., 2005. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic acids research* **33**(4): 1162–1168. → pages 17, 33

[103] Stark, A., Lin, M., Kheradpour, P., Pedersen, J., Parts, L., Carlson, J., Crosby, M., Rasmussen, M., Roy, S., Deoras, A. *et al.*, 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**(7167): 219–232. → pages 18

[104] Peng, Z., Cheng, Y., Tan, B.C.M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X. *et al.*, 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology* **30**(3): 253–260. → pages 18, 20, 29, 38, 105

[105] Gu, T., Buaas, F.W., Simons, A.K., Ackert-Bicknell, C.L., Braun, R.E., & Hibbs, M.A., 2012. Canonical A-to-I and C-to-U RNA Editing Is Enriched at 3UTRs and microRNA Target Sites in Multiple Mouse Tissues. *PLoS ONE* **7**(3): e33720. doi:{10.1371/journal.pone.0033720}. → pages 18

[106] Palladino, M.J., Keegan, L.P., O'connell, M.A., & Reenan, R.A., 2000. A-to-I pre-mRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity. *Cell* **102**(4): 437–449. → pages 18, 43

[107] Wang, Q., Miyakoda, M., Yang, W., Khillan, J., Stachura, D.L., Weiss, M.J., & Nishikura, K., 2004. Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. *Journal of Biological Chemistry* **279**(6): 4952–4961. → pages 18

[108] Maas, S., Kawahara, Y., Tamburro, K., & Nishikura, K., 2006. A-to-I RNA editing and human disease. *RNA biology* **3**(1): 1–9. → pages 18

[109] Nishikura, K., 2006. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nature Reviews Molecular Cell Biology* **7**(12): 919–931. → pages 18, 19

[110] Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F.J., Rechavi, G., Li, J.B., Eisenberg, E. *et al.*, 2014. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Research* **24**(3): 365–376. → pages 19

[111] Ramaswami, G., Lin, W., Piskol, R., Tan, M.H., Davis, C., & Li, J.B., 2012. Accurate identification of human Alu and non-Alu RNA editing sites. *Nature methods* **9**(6): 579–581. → pages 19, 20, 29, 31

[112] Eggington, J., Greene, T., & Bass, B., 2011. Predicting sites of ADAR editing in double-stranded RNA. *Nature communications* **2**: 319. → pages 19

[113] Bahn, J., Lee, J., Li, G., Greer, C., Peng, G., & Xiao, X., 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome research* **22**(1): 142–150. → pages 19, 20, 21, 29, 31, 105

[114] St Laurent, G., Tackett, M.R., Nechkin, S., Shtokalo, D., Antonets, D., Savva, Y.A., Maloney, R., Kapranov, P., Lawrence, C.E., & Reenan, R.A., 2013. Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in *Drosophila*. *Nature structural & molecular biology* **20**(11): 1333–1339. → pages 19, 20, 21, 28, 36, 37, 38, 39, 40, 45, 49

[115] Tariq, A., Garncarz, W., Handl, C., Balik, A., Pusch, O., & Jantsch, M.F., 2013. RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation. *Nucleic acids research* **41**(4): 2581–2593. → pages 19

[116] Wahlstedt, H., Daniel, C., Enstërö, M., & Öhman, M., 2009. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome research* **19**(6): 978–986. → pages 19, 41

[117] Solomon, O., Oren, S., Safran, M., Deshet-Unger, N., Akiva, P., Jacob-Hirsch, J., Cesarkas, K., Kabesa, R., Amariglio, N., Unger, R. *et al.*, 2013. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *RNA* **19**(5): 591–604. → pages 20, 28, 49, 50

[118] Giuliany, R.S., 2012. *A Novel Statistical Framework for the Accurate Identification of RNA-edits with Application to Human Cancers*. Ph.D. thesis, University of British Columbia. → pages 20, 21

[119] Zhang, Q. & Xiao, X., 2015. Genome sequence-independent identification of RNA editing sites. *Nature methods* **12**(4): 347–350. → pages 20, 21

[120] Li, M., Wang, I., Li, Y., Bruzel, A., Richards, A., Toung, J., & Cheung, V., 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**(6038): 53–58. → pages 20

[121] Pickrell, J.K., Gilad, Y., & Pritchard, J.K., 2012. Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome". *Science* **335**(6074): 1302. → pages 21

[122] Kleinman, C.L. & Majewski, J., 2012. Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome". *Science* **335**(6074): 1302. → pages 21

[123] Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M.A., Condon, A. *et al.*, 2012. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* **28**(2): 167–175. → pages 21

[124] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., & Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly* **125**(2): 167–188. → pages 22

[125] Zuker, M. & Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research* **9**(1): 133–148. → pages 22, 34, 135

[126] Ding, Y. & Lawrence, C.E., 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic acids research* **31**(24): 7280–7301. → pages 22

[127] Ding, Y., Chan, C.Y., & Lawrence, C.E., 2004. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic acids research* **32**(suppl 2): W135–W141. → pages 22

[128] Wiebe, N.J. & Meyer, I.M., 2010. Transat–a method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLoS Comput Biol* **6**(6): e1000823. → pages 23, 46, 136

[129] Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., & Haussler, D., 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**(4): e33. → pages 23

[130] Pedersen, J.S., Meyer, I.M., Forsberg, R., Simmonds, P., & Hein, J., 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic acids research* **32**(16): 4925–4936. → pages 23

[131] Pedersen, J.S., Forsberg, R., Meyer, I.M., & Hein, J., 2004. An evolutionary model for protein-coding regions with conserved RNA structure. *Molecular biology and evolution* **21**(10): 1913–1922. → pages 23

[132] Knudsen, B. & Hein, J., 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research* **31**(13): 3423–3428. → pages 23, 46

[133] Walsh, C.T., Garneau-Tsodikova, S., & Gatto, G.J., 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie International Edition* **44**(45): 7342–7372. → pages 23

[134] Capra, M., Nuciforo, P.G., Confalonieri, S., Quarto, M., Bianchi, M., Nebuloni, M., Boldorini, R., Pallotti, F., Viale, G., Gishizky, M.L. *et al.*, 2006. Frequent alterations in the expression of serine/threonine kinases in human cancers. *Cancer research* **66**(16): 8147–8154. → pages 23, 52, 79

[135] Zhang, J., Yang, P.L., & Gray, N.S., 2009. Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer* **9**(1): 28–39. → pages 24

[136] Harper, J. & Adams, P., 2001. Cyclin-dependent kinases. *Chemical Reviews* **101**(8): 2511–2526. → pages 24

[137] Malumbres, M., 2014. Cyclin-dependent kinases. *Genome Biology* **15**(6): 1–10. → pages 24, 51

[138] Loyer, P., Trembley, J.H., Katona, R., Kidd, V.J., & Lahti, J.M., 2005. Role of CDK/cyclin complexes in transcription and RNA splicing. *Cellular signalling* **17**(9): 1033–1051. → pages 51

[139] Even, Y., Durieux, S., Escande, M.L., Lozano, J.C., Peaucellier, G., Weil, D., & Geneviève, A.M., 2006. CDC2L5, a Cdk-like kinase with RS domain, interacts with the ASF/SF2-associated protein p32 and affects splicing in vivo. *Journal of cellular biochemistry* **99**(3): 890–904. → pages 24

[140] Cheng, S.W.G., Kuzyk, M.A., Moradian, A., Ichu, T.A., Chang, V.C.D., Tien, J.F., Vollett, S.E., Griffith, M., Marra, M.A., & Morin, G.B., 2012. Interaction of cyclin-dependent kinase 12/CrkRS with cyclin K1 is required for the phosphorylation of the C-terminal domain of RNA polymerase II. *Molecular and cellular biology* **32**(22): 4691–4704. → pages 24

[141] Bartkowiak, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., & Greenleaf, A.L., 2010. CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes & development* **24**(20): 2303–2316. → pages 24

[142] Dixon-Clarke, S., Elkins, J., Cheng, S., Morin, G., & Bullock, A., 2014. Structures of the CDK12/CycK complex with AMP-PNP reveal a flexible C-terminal kinase extension important for ATP binding. *Scientific reports* **5**: 17122–17122. → pages 24

[143] Ko, T.K., Kelly, E., & Pines, J., 2001. CrkRS: a novel conserved Cdc2-related protein kinase that colocalises with SC35 speckles. *Journal of cell science* **114**(14): 2591–2603. → pages 24

[144] Taglialatela, A. *CDK12 is a novel oncogene with clinical and pathogenetic relevance in breast cancer*. Ph.D. thesis. → pages 24, 25, 80

[145] Liang, K., Gao, X., Gilmore, J.M., Florens, L., Washburn, M.P., Smith, E., & Shilatifard, A., 2015. Characterization of human cyclin-dependent kinase 12 (CDK12) and CDK13 complexes in C-terminal domain phosphorylation, gene transcription, and RNA processing. *Molecular and cellular biology* **35**(6): 928–938. → pages 24, 53, 55, 58, 61, 63, 68, 70, 75, 77, 79

[146] Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., & Peterlin, B.M., 2011. The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes & development* **25**(20): 2158–2172. → pages 24, 52, 58, 70, 73, 77, 78, 79

[147] Chen, M. & Manley, J.L., 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews Molecular cell biology* **10**(11): 741–754. → pages 25

[148] Bartkowiak, B. & Greenleaf, A.L., 2015. Expression, purification, and identification of associated proteins of the full-length hCDK12/CyclinK complex. *Journal of Biological Chemistry* **290**(3): 1786–1795. → pages 25, 63

[149] Chen, H.H., Wang, Y.C., & Fann, M.J., 2006. Identification and characterization of the CDK12/cyclin L1 complex involved in alternative splicing regulation. *Molecular and cellular biology* **26**(7): 2736–2745. → pages 25

[150] Rodrigues, F., Thuma, L., & Klämbt, C., 2012. The regulation of glial-specific splicing of *Neurexin IV* requires how and cdk12 activity. *Development* **139**(10): 1765–1776. → pages 25

[151] Bajrami, I., Frankum, J.R., Konde, A., Miller, R.E., Rehman, F.L., Brough, R., Campbell, J., Sims, D., Rafiq, R., Hooper, S. *et al.*, 2014. Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer research* **74**(1): 287–297. → pages 25, 52, 70, 79

[152] Ekumi, K.M., Paculova, H., Lenasi, T., Pospichalova, V., Bösken, C.A., Rybarikova, J., Bryja, V., Geyer, M., Blazek, D., & Barboric, M., 2015. Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. *Nucleic Acids Research* **43**(5): 2575–2589. → pages 25, 52, 65, 70, 79

[153] Network, C.G.A.R. *et al.*, 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**(7353): 609–615. → pages 27, 51, 65, 67

[154] Dawson, T.R., Sansam, C.L., & Emeson, R.B., 2004. Structure and sequence determinants required for the RNA editing of ADAR2 substrates. *Journal of Biological Chemistry* **279**(6): 4941–4951. → pages 28

[155] Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M. *et al.*, 2009. Unlocking the secrets of the genome. *Nature* **459**(7249): 927–930. → pages 29, 30

[156] Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., & Salzberg, S.L., 1999. Alignment of whole genomes. *Nucleic Acids Research* **27**(11): 2369–2376. → pages 31, 134

[157] Delcher, A.L., Phillippy, A., Carlton, J., & Salzberg, S.L., 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic acids research* **30**(11): 2478–2483. → pages 31, 134

[158] Needleman, S.B. & Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**(3): 443–453. → pages 31, 134

[159] Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. *et al.*, 2013. Ensembl 2014. *Nucleic acids research* pages D749–D755. → pages 31, 53, 91

[160] Pachter, L., 2012. A closer look at RNA editing. *Nature biotechnology* **30**(3): 246–247. → pages 31, 38

[161] Goya, R., Sun, M.G., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., Senz, J., Crisan, A., Marra, M.A., Hirst, M. *et al.*, 2010. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**(6): 730–736. → pages 32, 33

[162] Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(21): 2987–2993. → pages 32, 34, 53, 91, 134

[163] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**(4): 1–13. → pages 31, 53

[164] Higuchi, M., Single, F.N., Köhler, M., Sommer, B., Sprengel, R., & Seeburg, P.H., 1993. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell* **75**(7): 1361–1370. → pages 34

[165] Bernhart, S.H., Hofacker, I.L., & Stadler, P.F., 2006. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22**(5): 614–615. → pages 35, 40, 135

[166] Ramaswami, G., Zhang, R., Piskol, R., Keegan, L.P., Deng, P., O'Connell, M.A., & Li, J.B., 2013. Identifying RNA editing sites using RNA sequencing data alone. *Nature Methods* **10**(2): 128–132. → pages 35, 38, 39, 45, 49, 135

[167] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., & Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**(3): 562–578. → pages 36, 42, 91, 92, 136

[168] Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., & Stadler, P.F., 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**(1): 1–13. → pages 39, 46, 48

[169] Chawla, G. & Sokol, N.S., 2014. ADAR mediates differential expression of polycistronic microRNAs. *Nucleic Acids Research* **42**(8): 5245–5255. → pages 40

[170] Vesely, C., Tauber, S., Sedlazeck, F.J., von Haeseler, A., & Jantsch, M.F., 2012. Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. *Genome research* **22**(8): 1468–1476. → pages

[171] de Hoon, M., Taft, R., Hashimoto, T., Kanamori-Katayama, M., Kawaji, H., Kawano, M., Kishima, M., Lassmann, T., Faulkner, G., Mattick, J. *et al.*, 2010. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome research* **20**(2): 257–264. → pages 40

[172] Da Wei Huang, B.T.S. & Lempicki, R.A., 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**(1): 44–57. → pages 43

[173] Hood, J.L. & Emeson, R.B., 2012. Editing of neurotransmitter receptor and ion channel RNAs in the nervous system. In *Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing*, pages 61–90. Springer. → pages 43

[174] Jepson, J., Savva, Y., Yokose, C., Sugden, A., Sahin, A., & Reenan, R., 2011. Engineered alterations in RNA editing modulate complex behavior in *Drosophila*: regulatory diversity of adenosine deaminase acting on RNA (ADAR) targets. *The Journal of biological chemistry* **286**(10): 8325–8337. → pages 43

[175] Lai, D., Proctor, J.R., Zhu, J.Y.A., & Meyer, I.M., 2012. R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic acids research* page gks241. → pages 47

[176] Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A. *et al.*, 2011. The UCSC Genome Browser database: update 2011. *Nucleic acids research* **39**(suppl_1): D876–D882. → pages 46

[177] Meyer, I.M. & Miklós, I., 2007. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS computational biology* **3**(8): e149. → pages 46

[178] Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H.A., Marr, M.T., & Rosbash, M., 2011. Nascent-seq Indicates Widespread Cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes & development* **25**(23): 2502–2512. → pages 49

[179] Romano, G. & Giordano, A., 2008. Role of the cyclin-dependent kinase 9-related pathway in mammalian gene expression and human diseases. *Cell Cycle* **7**(23): 3664–3668. → pages 51

[180] Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. *et al.*, 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**(5): 401–404. → pages 51, 79

[181] Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A. *et al.*, 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502**(7471): 333–339. → pages 65

[182] Network, C.G.A. *et al.*, 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418): 61–70. → pages 51, 79

[183] Joshi, P.M., Sutor, S.L., Huntoon, C.J., & Karnitz, L.M., 2014. Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly (ADP-ribose) polymerase inhibitors. *Journal of Biological Chemistry* **289**(13): 9247–9253. → pages 51, 52, 65, 70, 79

[184] Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A. *et al.*, 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**(5): 413–421. → pages 51

[185] Natrajan, R., Wilkerson, P.M., Marchiò, C., Piscuoglio, S., Ng, C.K., Wai, P., Lambros, M.B., Samartzis, E.P., Dedes, K.J., Frankum, J. *et al.*, 2014. Characterization of the genomic features and expressed fusion genes in micropapillary carcinomas of the breast. *The Journal of pathology* **232**(5): 553–565. → pages 52, 57, 65, 70, 79

[186] Kauraniemi, P., Kuukasjärvi, T., Sauter, G., & Kallioniemi, A., 2003. Amplification of a 280-kilobase core region at the *ERBB2* locus leads to activation of two hypothetical proteins in breast cancer. *The American journal of pathology* **163**(5): 1979–1984. → pages 52, 80

[187] Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringnér, M., Sauter, G., Monni, O., Elkahloun, A. *et al.*, 2002. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer research* **62**(21): 6240–6245. → pages

[188] Kao, J. & Pollack, J.R., 2006. RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes. *Genes, Chromosomes and Cancer* **45**(8): 761–769. → pages

[189] Kauraniemi, P., Bärlund, M., Monni, O., & Kallioniemi, A., 2001. New amplified and highly expressed genes discovered in the *ERBB2* amplicon in breast cancer by cdna microarrays. *Cancer research* **61**(22): 8235–8240. → pages

[190] Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F. *et al.*, 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell* **10**(6): 515–527. → pages

[191] Pollack, J.R., Sørlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Børresen-Dale, A.L., & Brown, P.O., 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* **99**(20): 12963–12968. → pages

[192] Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y. *et al.*, 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**(7403): 346–352. → pages

[193] Lawrence, R.T., Perez, E.M., Hernández, D., Miller, C.P., Haas, K.M., Irie, H.Y., Lee, S.I., Blau, C.A., & Villén, J., 2015. The proteomic landscape of triple-negative breast cancer. *Cell reports* **11**(4): 630–644. → pages

[194] Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C. *et al.*, 2015. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**(2): 506–519. → pages 52, 65, 80

[195] Zang, Z.J., Ong, C.K., Cutcutache, I., Yu, W., Zhang, S.L., Huang, D., Ler, L.D., Dykema, K., Gan, A., Tao, J. *et al.*, 2011. Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing. *Cancer research* **71**(1): 29–39. → pages 52, 65

[196] Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B. *et al.*, 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research* **41**(D1): D64–D69. → pages 53

[197] Wu, T.D. & Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7): 873–881. → pages 53, 91, 92

[198] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. *et al.*, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16): 2078–2079. → pages 53, 91, 134

[199] Love, M.I., Huber, W., & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**(12): 550. → pages 53, 68

[200] Anders, S., Pyl, P.T., & Huber, W., 2015. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**(2): 166–169. → pages 53

[201] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., & Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**(5): 511–515. → pages 53

[202] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.*, 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43): 15545–15550. → pages 54, 69, 85, 89

[203] Merico, D., Isserlin, R., Stueker, O., Emili, A., & Bader, G.D., 2010. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one* **5**(11): e13984. → pages 54, 92

[204] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**(11): 2498–2504. → pages 54, 92

[205] Retelska, D., Iseli, C., Bucher, P., Jongeneel, C.V., & Naef, F., 2006. Similarities and differences of polyadenylation signals in human and fly. *BMC genomics* **7**(1): 1. → pages 56

[206] Busch, A. & Hertel, K.J., 2012. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdisciplinary Reviews: RNA* **3**(1): 1–12. → pages 63

[207] Eifler, T.T., Shao, W., Bartholomeeusen, K., Fujinaga, K., Jäger, S., Johnson, J.R., Luo, Z., Krogan, N.J., & Peterlin, B.M., 2015. Cyclin-dependent kinase 12 increases 3 end processing of growth factor-induced c-FOS transcripts. *Molecular and cellular biology* **35**(2): 468–478. → pages 63

[208] Ingham, R.J., Colwill, K., Howard, C., Dettwiler, S., Lim, C.S., Yu, J., Hersi, K., Raaijmakers, J., Gish, G., Mbamalu, G. *et al.*, 2005. WW domains provide a platform for the assembly of multiprotein networks. *Molecular and cellular biology* **25**(16): 7092–7106. → pages 63

[209] Jung, S.Y., Malovannaya, A., Wei, J., OMalley, B.W., & Qin, J., 2005. Proteomic analysis of steady-state nuclear hormone receptor coactivator complexes. *Molecular endocrinology* **19**(10): 2451–2465. → pages 63

[210] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J. *et al.*, 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic acids research* **33**(suppl 1): D284–D288. → pages 64

[211] Elkon, R., Ugalde, A.P., & Agami, R., 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics* **14**(7): 496–506. → pages 63

[212] Anders, S. & Huber, W., 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**: R106–R106. → pages 68

[213] Juan, H., Lin, Y., Chen, H., & Fann, M., 2016. Cdk12 is essential for embryonic development and the maintenance of genomic stability. *Cell Death & Differentiation* **23**(6): 1038–1048. → pages 70

[214] Iorns, E., Martens-de Kemp, S.R., Lord, C.J., & Ashworth, A., 2009. CRK7 modifies the MAPK pathway and influences the response to endocrine therapy. *Carcinogenesis* **30**(10): 1696–1701. → pages 70

[215] Guleria, A. & Chandna, S., 2016. ATM kinase: Much more than a DNA damage responsive protein. *DNA repair* **39**: 1–20. → pages 73

[216] Chuang, J.Z., Zhou, H., Zhu, M., Li, S.H., Li, X.J., & Sung, C.H., 2002. Characterization of a brain-enriched chaperone, MRJ, that inhibits Huntingtin aggregation and toxicity independently. *Journal of Biological Chemistry* **277**(22): 19831–19838. → pages 75

[217] Fayazi, Z., Ghosh, S., Marion, S., Bao, X., Shero, M., & Kazemi-Esfarjani, P., 2006. A *Drosophila* ortholog of the human mrj modulates polyglutamine toxicity and aggregation. *Neurobiology of disease* **24**(2): 226–244. → pages 75

[218] Mitra, A., Fillmore, R.A., Metge, B.J., Rajesh, M., Xi, Y., King, J., Ju, J., Pannell, L., Shevde, L.A., & Samant, R.S., 2008. Large isoform of MRJ (DNAJB6) reduces malignant activity of breast cancer. *Breast Cancer Research* **10**(2): R22. → pages 75, 80

[219] Yu, V.Z., Wong, V.C.L., Dai, W., Ko, J.M.Y., Lam, A.K.Y., Chan, K.W., Samant, R.S., Lung, H.L., Shuen, W.H., Law, S. *et al.*, 2015. Nuclear localization of DNAJB6 is associated with survival of patients with esophageal cancer and reduces Akt signaling and proliferation of cancer cells. *Gastroenterology* **149**(7): 1825–1836. → pages 75

[220] Li, X., Chatterjee, N., Spirohn, K., Boutros, M., & Bohmann, D., 2016. Cdk12 Is A Gene-Selective RNA Polymerase II Kinase That Regulates a Subset of the Transcriptome, Including Nrf2 Target Genes. *Scientific reports* **6**: 21455. → pages 77

[221] Moasser, M.M., 2007. The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene* **26**(45): 6469–6487. → pages 78

[222] Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J., & Helleday, T., 2005. Specific killing of *BRCA2*-deficient tumours with inhibitors of poly (adp-ribose) polymerase. *Nature* **434**(7035): 913–917. → pages 79

[223] Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C. *et al.*, 2005. Targeting the DNA repair defect in *BRCA* mutant cells as a therapeutic strategy. *Nature* **434**(7035): 917–921. → pages

[224] Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M.J. *et al.*, 2009. Inhibition of poly (adp-ribose) polymerase in tumors from *BRCA* mutation carriers. *New England Journal of Medicine* **361**(2): 123–134. → pages 79

[225] Thompson, E.W. & Price, J.T., 2002. Mechanisms of tumour invasion and metastasis: emerging targets for therapy. *Expert opinion on therapeutic targets* **6**(2): 217–233. → pages 80

[226] Modrek, B., Lee, C. *et al.*, 2002. A genomic view of alternative splicing. *Nature genetics* **30**(1): 13–19. → pages 81

[227] Zhang, J. & Manley, J.L., 2013. Misregulation of pre-mRNA alternative splicing in cancer. *Cancer discovery* **3**(11): 1228–1237. → pages 81

[228] Danan-Gotthold, M., Golan-Gerstl, R., Eisenberg, E., Meir, K., Karni, R., & Levanon, E., 2015. Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic acids research* **43**(10): 5130–5144. → pages 81

[229] Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R., & Skotheim, R., 2016. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* **35**(19): 2413–2427. → pages 81

[230] Chen, J. & Weiss, W., 2015. Alternative splicing in cancer: implications for biology and therapy. *Oncogene* **34**(1): 1–14. → pages 81

[231] Kim, E., Goren, A., & Ast, G., 2008. Insights into the connection between cancer and alternative splicing. *Trends in Genetics* **24**(1): 7–10. → pages 82

[232] Tavares, R., Scherer, N.M., Ferreira, C.G., Costa, F.F., & Passetti, F., 2015. Splice variants in the proteome: a promising and challenging field to targeted drug discovery. *Drug discovery today* **20**(3): 353–360. → pages 82

[233] Anczuków, O. & Krainer, A.R., 2015. The spliceosome, a potential Achilles heel of MYC-driven tumors. *Genome Medicine* **1**(7): 1–4. → pages 82

[234] Nelander, S., Wang, W., Nilsson, B., She, Q.B., Pratilas, C., Rosen, N., Gennemark, P., & Sander, C., 2008. Models from experiments: combinatorial drug perturbations of cancer cells. *Molecular systems biology* **4**(1): 216. → pages 82, 102

[235] Liberali, P., Snijder, B., & Pelkmans, L., 2015. Single-cell and multivariate approaches in genetic perturbation screens. *Nature Reviews Genetics* **16**(1): 18–32. → pages

[236] Lehár, J., Zimmermann, G.R., Krueger, A.S., Molnar, R.A., Ledell, J.T., Heilbut, A.M., Short, G.F., Giusti, L.C., Nolan, G.P., Magid, O.A. *et al.*, 2007. Chemical combination effects predict connectivity in biological systems. *Molecular systems biology* **3**(1): 80. → pages 82

[237] Ideker, T., Dutkowski, J., & Hood, L., 2011. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144**(6): 860–863. → pages 83

[238] Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D., & Sander, C., 2011. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research* **39**(suppl 1): D685–D690. → pages 84

[239] Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**(1): 27–30. → pages

[240] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C. *et al.*, 2013. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**(D1): D808–D815. → pages

[241] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P. *et al.*, 2014. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* pages D447–D452. → pages

[242] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M., 2006. BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**(suppl 1): D535–D539. → pages 84

[243] Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L. *et al.*, 2015. The BioGRID interaction database: 2015 update. *Nucleic acids research* **43**(D1): D470–D478. → pages 84

[244] Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R. *et al.*, 2014. A proteome-scale map of the human interactome network. *Cell* **159**(5): 1212–1226. → pages 84

[245] Huang, D.W., Sherman, B.T., & Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**(1): 1–13. → pages 85

[246] Hung, J.H., Yang, T.H., Hu, Z., Weng, Z., & DeLisi, C., 2012. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics* **13**(3): 281–291. → pages 85

[247] Khatri, P., Sirota, M., & Butte, A., 2011. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **8**(2): e1002375–e1002375. → pages 85

[248] Al-Shahrour, F., Díaz-Uriarte, R., & Dopazo, J., 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**(4): 578–580. → pages 85

[249] Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S. *et al.*, 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* **4**(4): R28. → pages

[250] Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z., & Galon, J., 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**(8): 1091–1093. → pages

[251] Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z., 2010. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research* **38**(suppl 2): W64–W70. → pages

[252] Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L. *et al.*, 2006. WEGO: a web tool for plotting GO annotations. *Nucleic acids research* **34**(suppl 2): W293–W297. → pages 85

[253] Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., & Park, P.J., 2005. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America* **102**(38): 13544–13549. → pages 85

[254] Goeman, J.J., Van De Geer, S.A., De Kort, F., & Van Houwelingen, H.C., 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**(1): 93–99. → pages 85

[255] Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., & Romero, R., 2007. A systems biology approach for pathway level analysis. *Genome research* **17**(10): 1537–1545. → pages 85

[256] Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.s., Kim, C.J., Kusanovic, J.P., & Romero, R., 2009. A novel signaling pathway impact analysis. *Bioinformatics* **25**(1): 75–82. → pages 85

[257] Pham, L., Christadore, L., Schaus, S., & Kolaczyk, E.D., 2011. Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proceedings of the National Academy of Sciences* **108**(32): 13347–13352. → pages 85, 86, 101

[258] Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Martínez, M.R., López, G., Mattioli, M., Realubit, R. *et al.*, 2015. Elucidating compound mechanism of action by network perturbation analysis. *Cell* **162**(2): 441–451. → pages 86, 87

[259] Chindelevitch, L., Ziemek, D., Enayetallah, A., Randhawa, R., Sidders, B., Brockel, C., & Huang, E.S., 2012. Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics* **28**(8): 1114–1121. → pages 88

[260] Chindelevitch, L., Loh, P.R., Enayetallah, A., Berger, B., & Ziemek, D., 2012. Assessing statistical significance in causal graphs. *BMC bioinformatics* **13**(1): 35. → pages 89

[261] Krämer, A., Green, J., Pollard Jr, J., & Tugendreich, S., 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**(4): 523–530. → pages 88, 89

[262] Zarringhalam, K., Enayetallah, A., Gutteridge, A., Sidders, B., & Ziemek, D., 2013. Molecular causes of transcriptional response: a Bayesian prior knowledge approach. *Bioinformatics* **29**(24): 3167–3173. → pages 88

[263] Jaeger, S., Min, J., Nigsch, F., Camargo, M., Hutz, J., Cornett, A., Cleaver, S., Buckler, A., & Jenkins, J.L., 2014. Causal network models for predicting compound targets and driving pathways in cancer. *Journal of biomolecular screening* **19**(5): 791–802. → pages 88

[264] Martin, F., Thomson, T.M., Sewer, A., Drubin, D.A., Mathis, C., Weisensee, D., Pratt, D., Hoeng, J., & Peitsch, M.C., 2012. Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Systems Biology* **6**: 54. → pages

[265] Vasilyev, D.M., Thomson, T.M., Frushour, B.P., Martin, F., & Sewer, A., 2014. An algorithm for score aggregation over causal biological networks based on random walk sampling. *BMC Research Notes* **7**(1): 516. → pages

[266] Laenen, G., Ardeshirdavani, A., Moreau, Y., & Thorrez, L., 2015. Galahad: a web server for drug effect analysis from gene expression. *Nucleic Acids Research* **43**(W1): W208–W212. → pages 88

[267] Lefebvre, C., Rajbhandari, P., Alvarez, M.J., Bandaru, P., Lim, W.K., Sato, M., Wang, K., Sumazin, P., Kustagi, M., Bisikirska, B.C. *et al.*, 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology* **6**: 377. → pages 89, 101

[268] Lachmann, A. & Ma'ayan, A., 2009. KEA: kinase enrichment analysis. *Bioinformatics* **25**(5): 684–686. → pages 89

[269] Koschmann, J., Bhar, A., Stegmaier, P., Kel, A.E., & Wingender, E., 2015. Upstream Analysis: An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data. *Microarrays* **4**(2): 270–286. → pages 89

[270] Chen, E.Y., Xu, H., Gordonov, S., Lim, M.P., Perkins, M.H., & Ma'ayan, A., 2012. Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics* **28**(1): 105–111. → pages 89

[271] Funnell, T., Tasaki, S., Oloumi, A., Araki, S., Kong, E., Yap, D., Nakayama, Y., Hughes, C.S., Cheng, S.W.G., Tozaki, H. *et al.*, 2017. CLK-dependent exon recognition and conjoined gene formation revealed with a novel small molecule inhibitor. *Nature Communications* **8**(1): 7. → pages 90, 91

[272] Chazal, P.E., Daguenet, E., Wendling, C., Ulryck, N., Tomasetto, C., Sargueil, B., & Le Hir, H., 2013. EJC core component MLN51 interacts with eIF3 and activates translation. *Proceedings of the National Academy of Sciences* **110**(15): 5903–5908. → pages 91

[273] Haremaki, T. & Weinstein, D.C., 2012. Eif4a3 is required for accurate splicing of the *Xenopus laevis ryanodine receptor* pre-mRNA. *Developmental biology* **372**(1): 103–110. → pages 91

[274] Ngo, J.C.K., Chakrabarti, S., Ding, J.H., Velazquez-Dones, A., Nolen, B., Aubol, B.E., Adams, J.A., Fu, X.D., & Ghosh, G., 2005. Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Molecular cell* **20**(1): 77–89. → pages 91

[275] Ito, M., Iwatani, M., Kamada, Y., Sogabe, S., Nakao, S., Tanaka, T., Kawamoto, T., Aparicio, S., Nakanishi, A., & Imaeda, Y., 2017. Discovery of selective ATP-competitive eIF4A3 inhibitors. *Bioorganic & Medicinal Chemistry* **25**(7): 2200–2209. → pages 91

[276] Iwatani-Yoshihara, M., Ito, M., Ishibashi, Y., Oki, H., Tanaka, T., Morishita, D., Ito, T., Kimura, H., Imaeda, Y., Aparicio, S.A. *et al.*, 2017. Discovery and characterization of a eukaryotic initiation factor 4A-3-selective inhibitor that suppresses nonsense-mediated mRNA decay. *ACS Chemical Biology* . → pages 91

[277] Langfelder, P. & Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559. → pages 91, 95, 97, 100

[278] Maere, S., Heymans, K., & Kuiper, M., 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**(16): 3448–3449. → pages 91, 95

[279] Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2): 257–286. → pages 102

[280] Leng, N., Li, Y., McIntosh, B.E., Nguyen, B.K., Duffin, B., Tian, S., Thomson, J.A., Dewey, C.N., Stewart, R., & Kendziorski, C., 2015. EBSeq-HMM: a Bayesian approach

for identifying gene-expression changes in ordered RNA-seq experiments. *Bioinformatics* **31**(16): 2614–2622. → pages 102, 103

[281] Lange, S.J., Maticzka, D., Möhl, M., Gagnon, J.N., Brown, C.M., & Backofen, R., 2012. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic acids research* **40**(12): 5215–5226. → pages 135

[282] Edgar, R., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5): 1792–1797. → pages 136

Appendix A

Supporting Materials for Chapter 2

A.1 Details of the proposed pipeline

In our analysis, we used dm3 reference genome (fasta file) from UCSC (http://genome.ucsc.edu). *Drosophila melanogaster* annotation file (*BDGP5.74_ensembl.gtf*) was downloaded from the Ensembl web page (http://www.ensembl.org). In making the corresponding annotation file for the *OregonR* genome we employed MUMMER [156, 157] version "3.23" and NEEDLEMAN-WUNSCH [158] program version "0.3.5". To align short reads to the *OregonR* genome we executed: "tophat2 -F 0 -i 40 -g 40 –library-type fr-secondstrand -r 200 –mate-std-dev 20 –segment-length 16 –read-mismatches 5 –read-edit-dist 5" using TOPHAT2 version "v2.0.10". Information such as library type and mate standard deviation were chosen based on the information provided on http://www.modencode.org/.

For each candidate position, we require at least 2 and 5 reads in the flexible and stringent threshold sets, accordingly. We employed SAMTOOLS [162, 198] mpileup to extract reads covering each position. Sites that contain stars in SAMTOOLS mpileup tracks are also discarded (They present evidence for small insertions and deletions near a candidate site).

Additionally, at least one of the observed nucleotides from each variant should be from a high quality read (phred score of at least 20) and more than 5 nucleotides distant from the read ends. This filter can improve the results in two ways: first, random hexamer priming

can cause errors in the 5' starting positions of reads [166]; and second, read ends at splice junctions are prone to being misaligned [89]. We also filter sites where two or more alleles are observed other than the reference allele.

To filter known variations, we use Ensembl fly variant file http://uswest.ensembl.org/info/data/ftp/index.html (Ensembl release 74). Because variations reported in the file only contains variations of chromosomes X, 2 and 3, we ignored all predictions from other regions.

We filter candidates with log likelihood score smaller than 3. Additionally, we require editing ratio to be between 0.03 and 0.97, in order to lower the chance of including homozygous sites in our predictions [166], since sequencing and mapping errors are inevitable. These thresholds are equal in both sets threshold values.

The thresholds for all four of the SAMTOOLS/BCFTOOLS tests are set to 0.15 in flexible thresholding and 0.02 for the stringent thresholding. Our results were generated using SAMTOOLS version "0.1.19".

We employ RNAFOLD [125] with default parameters and RNAPLFOLD [165] with "-W 200 -L 150 -u 1" as suggested [281]; and for each site we calculate the average of pairing probabilities for a local region of length 5 (candidate position extended by two nucleotides from each side). A candidate site passes the structural filter if it is in a highly structured region (based on RNAFOLD [125] energy) or it shows evidence for being a part of a stem (based on RNAPLFOLD [165] energy). We set RNAFOLD [125] thresholds to -10 and -50 for the flexible and stringent threshold sets and we set RNAPLFOLD [165] thresholds to 0.2 and 0.7, accordingly. The analysis in the paper was carried out using RNAFOLD version "2.0.4" and RNAPLFOLD version "2.0.7".

For finding alternatively used exons, we applied DEXSEQ [74] version "1.8.0". In cases that there are transcripts with overlapping exons with different boundaries, DEXSEQ cuts the exons into multiple parts (see [74] for more details) and analyses their usage separately. Each of these exonic parts are considered as an exon in our analysis when we investigate the potential inter-relation between editing and splicing, however, we only report the ones that are longer than 10 nucleotides. Additionally, when we compare two tissues, we only consider genes that are predicted to have FPKM (fragments per kilobase of transcript per million fragments mapped) expression values greater than 2. Expression

values were computed by employing CUFFLINKS [167] package version "2.2.1".

In our analysis, we classify exonic regions into two groups: for each gene, we put all the exons in all the transcripts together; then we find the union of these exonic regions. Next, for each region, if the region constitutes multiple exons that are not identical, we call the region an exonic region with multiple acceptor/donor sites. The other group contains all the other exonic regions.

When we searched for structural features using TRANSAT [128], we only considered those helices that contain at least 8 base-pairs. The 15 fly species alignment was downloaded from UCSC http://genome.ucsc.edu for regions of interest. We added OregonR genome to the alignments and realigned the 16 sequences in each region by employing MUSCLE [282] (version 3.8.31).

Micro-RNA target sites were downloaded from http://microrna.org (August 2010 release), and miRNA sites were downloaded from: http://www.mirbase.org/ (miRBase v19).

A.2 Editing events within or in close vicinity of alternatively spliced exonic regions

The following table presents alternatively spliced exonic parts for which we found editing events within or in close vicinity (+/- 150 nt) of them.

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chr2L	+	2752952	2753088	2753235	2753371	FBgn0031453	CG9894
chr2L	+	2756391	2757424	2756664	2757703	FBgn0031453	CG9894
chr2L	+	2964499	2968083	2964852	2968434	FBgn0010263	Rbp9
chr2L	+	4307994	4308077	4308776	4308859	FBgn0010473	tutl
chr2L	+	4313423	4313836	4314202	4314615	FBgn0010473	tutl
chr2L	+	4313837	4315272	4314616	4316050	FBgn0010473	tutl
chr2L	+	5126158	5138115	5127058	5139015	FBgn0261836	Msp-300
chr2L	+	5157555	5164499	5158455	5165399	FBgn0261836	Msp-300
chr2L	+	5164500	5185064	5165400	5185964	FBgn0261836	Msp-300
chr2L	+	5205305	5205461	5206205	5206361	FBgn0065104	snmRNA:158
chr2L	+	5205462	5206102	5206362	5207002	FBgn0261836	Msp-300
chr2L	+	6497626	6498803	6498642	6499819	FBgn0051637	CG31637
chr2L	+	7050813	7052418	7051760	7053365	FBgn0262872	milt
chr2L	+	8109514	8115564	8110481	8116531	FBgn0261822	Bsg
chr2L	+	8176785	8177403	8177779	8178397	FBgn0031993	Piezo
chr2L	+	9255522	9255655	9256563	9256696	FBgn0028433, FBgn0263984	Ggamma30A, CG43733
chr2L	+	9255656	9255773	9256697	9256814	FBgn0028433	Ggamma30A
chr2L	+	12723188	12723334	12724472	12724618	FBgn0032456	MRP
chr2L	+	14082195	14082304	14083458	14083567	FBgn0028875	nAcRalpha-34E
chr2L	+	16172868	16173007	16174057	16174196	FBgn0001991	Ca-alpha1D
chr2L	+	16748691	16750760	16749944	16752013	FBgn0032600	CG17912
chr2L	+	16750761	16752451	16752014	16753704	FBgn0032600	CG17912
chr2L	+	16778450	16778567	16779728	16779845	FBgn0264695	Mhc
chr2L	+	21124662	21124877	21126159	21126374	FBgn0040297	Nhe2
chr2L	+	22443026	22447535	22444518	22449027	FBgn0040010	CG17493
chr2L	+	22735693	22736072	22737200	22737579	FBgn0041004	CG17715
chr2L	-	227580	228164	227548	228132	FBgn0086902	kis
chr2L	-	1008285	1011311	1008378	1011417	FBgn0031294	IA-2
chr2L	-	2782981	2785243	2783275	2785538	FBgn0004242	Syt1
chr2L	-	3461785	3462873	3462213	3463289	FBgn0005616	msl-2
chr2L	-	3503294	3503904	3503689	3504299	FBgn0014396	tim
chr2L	-	6631823	6633573	6632798	6634548	FBgn0051635	CG31635

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chr2L	-	6792277	6792422	6793228	6793373	FBgn0015777	nrv2
chr2L	-	6792423	6792453	6793374	6793404	FBgn0015777	nrv2
chr2L	-	7227712	7228024	7228700	7229012	FBgn0259111	Ndae1
chr2L	-	7802122	7802654	7803044	7803576	FBgn0031952	cdc14
chr2L	-	9788515	9788572	9789675	9789732	FBgn0042174	CR18854
chr2L	-	9789454	9789583	9790616	9790745	FBgn0042174	CR18854
chr2L	-	9808062	9808242	9809222	9809402	FBgn0032151	nAcRalpha-30D
chr2L	-	9934485	9936150	9935570	9937237	FBgn0051712	CG31712
chr2L	-	9958179	9959380	9959248	9960449	FBgn0032172	CG5850
chr2L	-	11158599	11158963	11159880	11160244	FBgn0259822	Ca-beta
chr2L	-	11828938	11829492	11830277	11830831	FBgn0259225	Pde1c
chr2L	-	17369521	17370498	17370843	17371820	FBgn0032633	Lrch
chr2L	-	21662309	21664103	21663806	21665595	FBgn0032957	CG2225
chr2R	+	2704047	2705453	2704298	2705704	FBgn0033107	koi
chr2R	+	4790682	4792342	4790636	4792296	FBgn0004921	Ggamma1
chr2R	+	5996022	5997233	5995980	5997192	FBgn0004907	14-3-3zeta
chr2R	+	6166227	6172274	6166251	6172306	FBgn0033504	CAP
chr2R	+	6499104	6499267	6499021	6499184	FBgn0263102	psq
chr2R	+	9704073	9704297	9704260	9704484	FBgn0261041	stj
chr2R	+	10164604	10166977	10164786	10167147	FBgn0263397	Ih
chr2R	+	10185231	10185547	10185402	10185718	FBgn0263397	Ih
chr2R	+	12002296	12002709	12002445	12002858	FBgn0034075	Asph
chr2R	+	13249308	13252422	13249777	13252891	FBgn0261642	mbl
chr2R	+	13260853	13266410	13261323	13266881	FBgn0261642	mbl
chr2R	+	13458532	13458825	13458978	13459271	FBgn0040294	POSH
chr2R	+	13458887	13459381	13459333	13459827	FBgn0040294	POSH
chr2R	+	14708584	14708731	14708995	14709142	FBgn0010551	l(2)03709
chr2R	+	15110431	15111243	15110745	15111557	FBgn0263395	hppy
chr2R	+	15112756	15112977	15113070	15113291	FBgn0263395	hppy
chr2R	+	16893090	16894180	16894147	16895245	FBgn0034570	CG10543
chr2R	+	17030976	17030998	17032011	17032033	FBgn0021872	Xbp1
chr2R	+	20770449	20772868	20772049	20774468	FBgn0085442	NKAIN
chr2R	+	20796894	20797221	20798487	20798814	FBgn0085434	NaCP60E
chr2R	+	20797929	20798103	20799522	20799696	FBgn0085434	NaCP60E
chr2R	+	20801661	20801761	20803254	20803354	FBgn0085434	NaCP60E
chr2R	-	674709	675916	674643	675850	FBgn0250830	CG12547
chr2R	-	2814780	2815538	2814895	2815653	FBgn0053558	mim
chr2R	-	2818453	2819232	2818568	2819347	FBgn0053558	mim
chr2R	-	5121984	5123114	5121950	5123080	FBgn0010114	hig
chr2R	-	5172488	5172577	5172464	5172553	FBgn0020621	Pkn

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chr2R	-	5610974	5611177	5611045	5611248	FBgn0259678	sqa
chr2R	-	5771494	5771588	5771508	5771589	FBgn0033463	CG1513
chr2R	-	5911719	5911875	5911708	5911864	FBgn0022382	Pka-R2
chr2R	-	9400439	9405293	9400637	9405491	FBgn0260964	Vmat
chr2R	-	9771446	9781481	9771600	9781634	FBgn0013733	shot
chr2R	-	9954582	9955019	9954752	9955189	FBgn0040752	Prosap
chr2R	-	11150036	11151989	11150309	11152262	FBgn0083959	trpm
chr2R	-	11652369	11652498	11652628	11652757	FBgn0083919	Zasp52
chr2R	-	11661984	11662733	11662263	11663012	FBgn0083919	Zasp52
chr2R	-	19049472	19051556	19050535	19052619	FBgn0085400	CG34371
chr3L	+	893317	895105	893521	895313	FBgn0052479	CG32479
chr3L	+	1620402	1620503	1620817	1620918	FBgn0035244	ABCB7
chr3L	+	2923450	2924848	2924078	2925476	FBgn0262593	Shab
chr3L	+	3085267	3085738	3085865	3086336	FBgn0035397	CG11486
chr3L	+	4429108	4429621	4429716	4430229	FBgn0000038	nAcRbeta-64B
chr3L	+	9068900	9069378	9070078	9070556	FBgn0023479	Tequila
chr3L	+	9824730	9826147	9826028	9827445	FBgn0264489	CG43897
chr3L	+	20368371	20368907	20372412	20372948	FBgn0036980	RhoBTB
chr3L	+	20761984	20762135	20766064	20766215	FBgn0016696	Pitslre
chr3L	+	20762136	20762813	20766216	20766893	FBgn0016696	Pitslre
chr3L	+	21202134	21202496	21206229	21206591	FBgn0037060	CG10508
chr3L	+	21915634	21921636	21919800	21925802	FBgn0262737	mub
chr3L	+	23273132	23276049	23277312	23280229	FBgn0037212	nAcRalpha-80B
chr3L	+	24531314	24532438	24535510	24536634	FBgn0044510	mRpS5
chr3L	-	2039221	2040112	2039681	2040572	FBgn0086906	sls
chr3L	-	2561442	2562353	2561932	2562843	FBgn0010909	msn
chr3L	-	4096290	4096364	4096904	4096978	FBgn0035497	CG14995
chr3L	-	4322064	4322921	4322672	4323529	FBgn0035533	Cip4
chr3L	-	4367031	4368514	4367649	4369133	FBgn0035538	DopEcR
chr3L	-	4824734	4825093	4825328	4825687	FBgn0261797	Dhc64C
chr3L	-	5148383	5150468	5148921	5151007	FBgn0052423	shep
chr3L	-	6944092	6946063	6944904	6946872	FBgn0035720	CG10077
chr3L	-	7172281	7172729	7173238	7173686	FBgn0263218	Dscam2
chr3L	-	7822499	7823929	7823487	7824910	FBgn0016694	Pdp1
chr3L	-	7920585	7922201	7921569	7923185	FBgn0024187	syd
chr3L	-	11549184	11549651	11551372	11551839	FBgn0259481	Mob2
chr3L	-	12199603	12203376	12201959	12205752	FBgn0260941	app
chr3L	-	13424958	13425445	13427842	13428329	FBgn0036360	CG10713
chr3L	-	14497699	14500007	14500714	14503019	FBgn0087007	bbg

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chr3L	-	17048362	17054741	17052007	17058366	FBgn0260943	Rbp6
chr3L	-	17959290	17961085	17962744	17964539	FBgn0000568	Eip75B
chr3L	-	18047937	18049283	18051337	18052698	FBgn0000568	Eip75B
chr3L	-	19132438	19136013	19135998	19139573	FBgn0016797	fz2
chr3L	-	19880748	19884381	19884505	19888138	FBgn0014037	Su(Tpl)
chr3L	-	20145870	20146412	20149780	20150322	FBgn0261556	CG42674
chr3L	-	21186766	21187109	21190860	21191203	FBgn0053054	CG33054
chr3L	-	21187110	21187176	21191204	21191270	FBgn0053054	CG33054
chr3R	+	121421	122682	121423	122684	FBgn0041605	cpx
chr3R	+	528121	530970	528134	530983	FBgn0263346	CG43427
chr3R	+	3019037	3019078	3018693	3018734	FBgn0086372	lap
chr3R	+	3829261	3829801	3828942	3829482	FBgn0037536	CG2698
chr3R	+	5274501	5275314	5274377	5275190	FBgn0261552	ps
chr3R	+	6021459	6021790	6021438	6021769	FBgn0004575	Syn
chr3R	+	6067512	6073989	6067519	6073996	FBgn0261928	CG42795
chr3R	+	7217000	7217144	7217251	7217395	FBgn0004595	pros
chr3R	+	9489835	9490352	9490657	9491174	FBgn0004587	B52
chr3R	+	9516531	9518800	9517357	9519629	FBgn0024555	flfl
chr3R	+	10615974	10618161	10616975	10619164	FBgn0263929	jvl
chr3R	+	11237203	11238136	11238301	11239234	FBgn0041188	Atx2
chr3R	+	11780587	11780904	11781712	11782029	FBgn0013334	Sap47
chr3R	+	12127035	12127383	12128095	12128443	FBgn0250823	gish
chr3R	+	13744178	13745815	13745554	13747191	FBgn0263995	cpo
chr3R	+	13835530	13836146	13836819	13837435	FBgn0263995	cpo
chr3R	+	13998906	13999410	14000262	14000766	FBgn0042693	PP2A-B'
chr3R	+	14002949	14004109	14004328	14005496	FBgn0042693	PP2A-B'
chr3R	+	14794533	14794750	14796353	14796570	FBgn0261262, FBgn0263983	CG42613, CG43732
chr3R	+	14795191	14795317	14797012	14797138	FBgn0261262, FBgn0263983	CG42613, CG43732
chr3R	+	15589530	15589636	15591367	15591473	FBgn0024963	GluClalpha
chr3R	+	16145871	16147064	16147671	16148864	FBgn0261550	CG42668
chr3R	+	16834442	16834471	16836308	16836337	FBgn0013995	Calx
chr3R	+	17036498	17038080	17038410	17039991	FBgn0264357	SNF4Agamma
chr3R	+	18426603	18427004	18428784	18429185	FBgn0051158	Efa6
chr3R	+	20529756	20529892	20532259	20532395	FBgn0003429	slo
chr3R	+	20530492	20531282	20532995	20533784	FBgn0003429	slo
chr3R	+	20531283	20531615	20533785	20534111	FBgn0003429	slo
chr3R	+	20531616	20534411	20534112	20536911	FBgn0003429	slo
chr3R	+	21425417	21431422	21428144	21434145	FBgn0011666	msi
chr3R	+	23530288	23531240	23533099	23534051	FBgn0039544	CG12877

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chr3R	+	24737936	24738410	24740662	24741136	FBgn0259220	Doa
chr3R	+	27659530	27659581	27662904	27662955	FBgn0039883	RhoGAP100F
chr3R	-	622768	623397	622790	623419	FBgn0260794	ctrip
chr3R	-	1108329	1109043	1108387	1109101	FBgn0013576	mtd
chr3R	-	1827447	1827676	1827555	1827784	FBgn0003261	Rm62
chr3R	-	1828380	1828724	1828476	1828820	FBgn0003261	Rm62
chr3R	-	1828725	1829298	1828821	1829394	FBgn0003261	Rm62
chr3R	-	3687162	3687752	3686861	3687451	FBgn0037525	CG17816
chr3R	-	4661546	4663665	4661313	4663432	FBgn0262614	pyd
chr3R	-	5845458	5847136	5845427	5847105	FBgn0053208	Mical
chr3R	-	7590636	7591909	7590844	7592117	FBgn0086910	l(3)neo38
chr3R	-	7629238	7630103	7629438	7630309	FBgn0051116	ClC-a
chr3R	-	7772790	7773053	7773103	7773362	FBgn0037963	Cad87A
chr3R	-	10638940	10642622	10639971	10643653	FBgn0053555	btsz
chr3R	-	11921737	11922221	11922858	11923342	FBgn0026059	Mhcl
chr3R	-	12163638	12164821	12164665	12165848	FBgn0040284	SF2
chr3R	-	13598188	13600676	13599554	13602042	FBgn0262562	CG43102
chr3R	-	13706530	13708449	13707906	13709825	FBgn0053547	Rim
chr3R	-	13708450	13709028	13709826	13710404	FBgn0053547	Rim
chr3R	-	14016821	14019193	14018216	14020580	FBgn0011481	Ssdp
chr3R	-	14963038	14965628	14964776	14967365	FBgn0261285	Ppcs
chr3R	-	19926699	19926891	19929101	19929293	FBgn0013343	Syx1A
chr3R	-	19926892	19929063	19929294	19931465	FBgn0013343	Syx1A
chr3R	-	21179732	21180606	21182356	21183230	FBgn0004509	Fur1
chr3R	-	27424594	27425436	27427953	27428795	FBgn0039858	CycG
chrX	+	936417	936786	936466	936835	FBgn0003638	su(w[a])
chrX	+	936787	936917	936836	936966	FBgn0003638	su(w[a])
chrX	+	1541203	1543216	1541398	1543411	FBgn0000210	br
chrX	+	1677514	1677859	1677689	1678034	FBgn0026086	Adar
chrX	+	1677860	1681927	1678035	1682100	FBgn0026086	Adar
chrX	+	2005630	2007591	2005737	2007698	FBgn0000382	csw
chrX	+	2561982	2562647	2562199	2562864	FBgn0003371	sgg
chrX	+	2568738	2569418	2568955	2569635	FBgn0003371	sgg
chrX	+	2569419	2569437	2569636	2569654	FBgn0003371	sgg
chrX	+	2570917	2571662	2571134	2571879	FBgn0003371	sgg
chrX	+	3232900	3237343	3233361	3237800	FBgn0000479	dnc
chrX	+	3846095	3847427	3846800	3848138	FBgn0029687	Vap-33-1
chrX	+	5132003	5132968	5133459	5134415	FBgn0086911	rg
chrX	+	5293695	5295921	5295157	5297372	FBgn0029761	SK
chrX	+	8124249	8124726	8126462	8126939	FBgn0261873	sdt
chrX	+	8131254	8132287	8133475	8134507	FBgn0261873	sdt

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chrX	+	9007936	9009533	9010179	9011772	FBgn0030089	AP-1gamma
chrX	+	9066982	9081106	9069224	9083342	FBgn0026206	mei-P26
chrX	+	10742995	10743629	10745738	10746360	FBgn0030240	CG2202
chrX	+	11384395	11384582	11387231	11387418	FBgn0052666	Drak
chrX	+	11594708	11595593	11597542	11598427	FBgn0011754	PhKgamma
chrX	+	11691173	11692768	11694026	11695620	FBgn0000259	CkIIbeta
chrX	+	11726864	11728598	11729760	11731493	FBgn0262684	CG43154
chrX	+	12539706	12540359	12542817	12543464	FBgn0030412	tomosyn
chrX	+	12661863	12663735	12665012	12666884	FBgn0030421	CG3812
chrX	+	13604153	13605164	13607457	13608468	FBgn0052627	NnaD
chrX	+	14823230	14823799	14827047	14827616	FBgn0264078	Flo-2
chrX	+	14888251	14889359	14892167	14893275	FBgn0000535	eag
chrX	+	14890535	14890950	14894446	14894861	FBgn0000535	eag
chrX	+	14892240	14892527	14896148	14896435	FBgn0000535	eag
chrX	+	15793632	15795071	15797823	15799262	FBgn0003392	shi
chrX	+	15795072	15795334	15799263	15799525	FBgn0003392	shi
chrX	+	16228492	16230171	16232805	16234476	FBgn0011764	Dsp1
chrX	+	16326218	16326300	16330538	16330620	FBgn0026575	hang
chrX	+	16823620	16824025	16828071	16828476	FBgn0027556	CG4928
chrX	+	18767196	18769775	18772088	18774669	FBgn0085430	CG34401
chrX	+	19395304	19397722	19400519	19402937	FBgn0027621	Pfrx
chrX	+	21249448	21250319	21255317	21256188	FBgn0003423	slgA
chrX	-	6681872	6683314	6683745	6685194	FBgn0259228	C3G
chrX	-	6977733	6977754	6979803	6979824	FBgn0263563	mir-4956
chrX	-	6977755	6977765	6979825	6979835	FBgn0263563	mir-4956
chrX	-	6977766	6977787	6979836	6979857	FBgn0263563	mir-4956
chrX	-	6977918	6977941	6979988	6980011	FBgn0264270	Sxl
chrX	-	7940386	7941253	7942563	7943432	FBgn0004656	fs(1)h
chrX	-	9170062	9170707	9172285	9172930	FBgn0040236	c11.1
chrX	-	9949290	9949353	9951699	9951762	FBgn0030174	CG15312
chrX	-	10217262	10217284	10219824	10219846	FBgn0259170	alpha-Man-I
chrX	-	11859737	11860300	11862684	11863247	FBgn0263111	cac
chrX	-	11913895	11916122	11916861	11919090	FBgn0030366	Usp7
chrX	-	13093610	13093927	13096773	13097090	FBgn0005410	sno
chrX	-	13094040	13094586	13097203	13097742	FBgn0005410	sno
chrX	-	13161461	13161511	13164676	13164726	FBgn0041210	HDAC4
chrX	-	14677766	14679544	14681511	14683283	FBgn0003301	rut
chrX	-	14681551	14682656	14685281	14686386	FBgn0003301	rut
chrX	-	15818777	15818930	15822964	15823117	FBgn0053180	Ranbp16
chrX	-	15879107	15879164	15883302	15883359	FBgn0030719	eIF5
chrX	-	15967717	15968843	15971920	15973046	FBgn0028397	Tob

Chrom	Strand	OregonR start	OregonR end	Reference start	Reference end	Ensembl ID	Gene name
chrX	-	16451371	16454003	16455710	16458346	FBgn0030758	CanA-14F
chrX	-	17819913	17820103	17824619	17824809	FBgn0003380	Sh
chrX	-	17855688	17856565	17860381	17861258	FBgn0003380	Sh
chrX	-	19462271	19462409	19467498	19467636	FBgn0031030	Tao
chrX	-	19462410	19462649	19467637	19467876	FBgn0031030	Tao
chrX	-	19462650	19462663	19467877	19467890	FBgn0031030	Tao
chrX	-	20626419	20630935	20632098	20636616	FBgn0085387	shakB
chrX	-	20630936	20632981	20636617	20638654	FBgn0085387	shakB
chrX	-	21072832	21074690	21078702	21080560	FBgn0052521	CG32521
chrX	-	21074691	21075035	21080561	21080905	FBgn0052521	CG32521
chrX	-	21493913	21494001	21499782	21499870	FBgn0024807	DIP1
chrX	-	21494951	21495131	21500820	21501000	FBgn0024807	DIP1

Table A.1: Alternatively spliced exonic parts for which we found editing events in close vicinity

A.3 Genomic regions with evidence for the inter-relation of RNA editing and alternative splicing

The following table contains the list of genomic regions for which we found evidence that RNA editing may regulate alternative splicing.

Chrom	Strand	OregonR start position	OregonR end position	Reference start position	Reference end position	Ensembl gene ID	Gene name	RNAalifoldEnergy
chr2R	+	17030848	17031148	17031883	17032183	FBgn0021872	Xbp1	-47.47
chr2R	+	17030826	17031126	17031861	17032161	FBgn0021872	Xbp1	-43.97
chr3R	+	6021309	6021609	6021288	6021588	FBgn0004575	Syn	-35.74
chr2R	+	15112606	15112906	15112920	15113220	FBgn0263395	hppy	-35.32
chr3R	+	20530342	20530642	20532845	20533145	FBgn0003429	slo	-34.33
chr2L	+	4307844	4308144	4308626	4308926	FBgn0010473	tutl	-32.99
chr3R	-	1827526	1827826	1827634	1827934	FBgn0003261	Rm62	-32.71
chrX	-	6977637	6977937	6979707	6980007	FBgn0263563	mir-4956	-30.22
chr2L	+	4313686	4313986	4314465	4314765	FBgn0010473	tutl	-28.02
chr2R	+	15112827	15113127	15113141	15113441	FBgn0263395	hppy	-28.01
chr2L	-	3503144	3503444	3503539	3503839	FBgn0014396	tim	-27.77
chr2R	+	20796744	20797044	20798337	20798637	FBgn0085434	NaCP60E	-27.50
chr3R	+	20529742	20530042	20532245	20532545	FBgn0003429	slo	-27.01
chr3L	+	1620353	1620653	1620768	1621068	FBgn0035244	ABCB7	-26.77
chr3L	+	21201984	21202284	21206079	21206379	FBgn0037060	CG10508	-24.89
chr3R	+	20529606	20529906	20532109	20532409	FBgn0003429	slo	-24.83
chr2L	-	9808092	9808392	9809252	9809552	FBgn0032151	nAcRalpha-30D	-24.60
chr3R	+	14795167	14795467	14796988	14797292	FBgn0261262, FBgn0263983	CG42613, CG43732	-24.09
chr2R	-	2819082	2819382	2819197	2819497	FBgn0053558	mim	-23.82
chrX	+	14889209	14889509	14893125	14893424	FBgn0000535	eag	-21.86
chr2R	-	2814630	2814930	2814745	2815045	FBgn0053558	mim	-21.81
chr2L	-	9807912	9808212	9809070	9809372	FBgn0032151	nAcRalpha-30D	-21.19
chr2L	+	5205311	5205611	5206211	5206511	FBgn0065104	snmRNA:158	-20.95
chr2L	+	9255372	9255672	9256413	9256713	FBgn0028433, FBgn0263984	Ggamma30A, CG43733	-20.94
chrX	+	16823875	16824175	16828326	16828626	FBgn0027556	CG4928	-20.77
chr2L	-	9789304	9789604	9790464	9790766	FBgn0042174	CR18854	-20.15
chr3L	+	1620252	1620552	1620667	1620967	FBgn0035244	ABCB7	-19.52
chr2L	-	7227562	7227862	7228550	7228850	FBgn0259111	Ndae1	-18.93
chr3L	-	4322771	4323071	4323379	4323679	FBgn0035533	Cip4	-18.87
chr3L	-	7172579	7172879	7173536	7173836	FBgn0263218	Dscam2	-18.84

Chrom	Strand	OregonR start position	OregonR end position	Reference start position	Reference end position	Ensembl gene ID	Gene name	RNAalifoldEnergy
chr3L	+	3085588	3085888	3086186	3086486	FBgn0035397	CG11486	-18.05
chr3R	+	5274351	5274651	5274219	5274527	FBgn0261552	ps	-17.27
chr3R	+	3018887	3019187	3018543	3018843	FBgn0086372	lap	-17.03
chr2L	+	16778417	16778717	16779695	16779995	FBgn0264695	Mhc	-16.91
chr2L	+	5205155	5205455	5206055	5206355	FBgn0065104	snmRNA:158	-16.51
chr2L	+	9255505	9255805	9256546	9256846	FBgn0028433, FBgn0263984	Ggamma30A, CG43733	-16.25
chr3R	+	27659431	27659731	27662805	27663105	FBgn0039883	RhoGAP100F	-16.14
chr3R	+	24738260	24738560	24740986	24741286	FBgn0259220	Doa	-16.02
chr2L	+	21124727	21125027	21126224	21126524	FBgn0040297	Nhe2	-15.11
chr2R	+	13458737	13459037	13459183	13459483	FBgn0040294	POSH	-14.58
chr3R	+	9490202	9490502	9491024	9491324	FBgn0004587	B52	-14.17
chr3R	+	27659380	27659680	27662754	27663054	FBgn0039883	RhoGAP100F	-13.76
chr2L	+	12723038	12723338	12724322	12724622	FBgn0032456	MRP	-13.49
chr3R	-	27424444	27424744	27427803	27428103	FBgn0039858	CycG	-11.81
chrX	-	21493851	21494151	21499720	21500020	FBgn0024807	DIP1	-11.79
chr2L	-	11829342	11829642	11830681	11830979	FBgn0259225	Pde1c	-11.54
chr2R	+	9704147	9704447	9704334	9704634	FBgn0261041	stj	-11.38
chrX	-	13161361	13161661	13164576	13164874	FBgn0041210	HDAC4	-11.27
chr3R	+	14795041	14795341	14796861	14797162	FBgn0261262, FBgn0263983	CG42613, CG43732	-10.94
chr2L	+	12723184	12723484	12724468	12724768	FBgn0032456	MRP	-10.76
chr3R	-	11921587	11921887	11922708	11923008	FBgn0026059	Mhcl	-10.73
chr3L	+	21202346	21202646	21206441	21206741	FBgn0037060	CG10508	-10.50
chr3L	-	19135863	19136163	19139423	19139723	FBgn0016797	fz2	-9.94
chr2R	+	20797779	20798079	20799372	20799672	FBgn0085434	NaCP60E	-9.58
chr2R	+	6498954	6499254	6498871	6499171	FBgn0263102	psq	-9.31
chr3R	+	7216850	7217150	7217101	7217401	FBgn0004595	pros	-9.15
chrX	-	9949140	9949440	9951549	9951849	FBgn0030174	CG15312	-8.74
chr3L	-	4368364	4368664	4368982	4369283	FBgn0035538	DopEcR	-8.32
chrX	+	16326068	16326368	16330388	16330688	FBgn0026575	hang	-7.85
chr3R	+	122532	122832	122534	122834	FBgn0041605	cpx	-7.50
chr3L	-	4824584	4824884	4825178	4825478	FBgn0261797	Dhc64C	-7.43
chr2R	+	14708581	14708881	14708992	14709292	FBgn0010551	l(2)03709	-7.37
chr3L	-	21186959	21187259	21191053	21191353	FBgn0053054	CG33054	-7.05
chr2R	-	5610824	5611124	5610895	5611195	FBgn0259678	sqa	-6.95
chr3L	+	24531164	24531464	24535360	24535660	FBgn0044510	mRpS5	-6.70
chrX	+	10742845	10743145	10745588	10745888	FBgn0030240	CG2202	-5.77
chrX	-	13093777	13094077	13096940	13097240	FBgn0005410	sno	-5.31
chr3R	-	1827297	1827597	1827405	1827705	FBgn0003261	Rm62	-4.48

Chrom	Strand	OregonR start position	OregonR end position	Reference start position	Reference end position	Ensembl gene ID	Gene name	RNAalifoldEnergy
chrX	-	13161311	13161611	13164526	13164824	FBgn0041210	HDAC4	-4.06
chr2R	-	5911569	5911869	5911558	5911858	FBgn0022382	Pka-R2	-3.63
chr3L	-	13425295	13425595	13428179	13428479	FBgn0036360	CG10713	-3.60
chrX	-	15879014	15879314	15883210	15883509	FBgn0030719	eIF5	-3.57
chr3R	-	7590486	7590786	7590694	7590994	FBgn0086910	l(3)neo38	-2.95
chr2R	+	13458675	13458975	13459121	13459421	FBgn0040294	POSH	-2.66
chr3R	+	12126885	12127185	12127945	12128245	FBgn0250823	gish	-2.31
chr2L	+	22735922	22736222	22737429	22737729	FBgn0041004	CG17715	-2.03
chr3R	+	18426453	18426753	18428640	18428934	FBgn0051158	Efa6	-1.89
chr3R	+	23531090	23531390	23533901	23534201	FBgn0039544	CG12877	-1.70
chrX	+	11384432	11384732	11387268	11387568	FBgn0052666	Drak	-1.62
chr2L	-	6792303	6792603	6793254	6793554	FBgn0015777	nrv2	-1.14
chr2R	+	16894030	16894330	16895092	16895395	FBgn0034570	CG10543	-1.14
chr3L	-	14497549	14497849	14500564	14500864	FBgn0087007	bbg	-0.79
chr3L	-	11549034	11549334	11551222	11551522	FBgn0259481	Mob2	-0.60
chr2R	-	5911725	5912025	5911714	5912014	FBgn0022382	Pka-R2	-0.20
chrX	+	15795184	15795484	15799375	15799675	FBgn0003392	shi	-0.02
chr2L	-	1011161	1011461	1011267	1011570	FBgn0031294	IA-2	0.22
chrX	-	14679394	14679694	14683133	14683433	FBgn0003301	rut	0.96
chr3R	+	12127233	12127533	12128293	12128593	FBgn0250823	gish	1.12
chrX	-	17855538	17855838	17860231	17860531	FBgn0003380	Sh	1.56
chrX	-	13093890	13094190	13097053	13097353	FBgn0005410	sno	1.80
chrX	+	1677709	1678009	1677884	1678184	FBgn0026086	Adar	1.98
chr3L	-	21187026	21187326	21191120	21191420	FBgn0053054	CG33054	2.08
chrX	+	15794921	15795221	15799112	15799412	FBgn0003392	shi	3.13
chrX	+	2569268	2569568	2569485	2569785	FBgn0003371	sgg	3.31
chrX	+	2569287	2569587	2569504	2569804	FBgn0003371	sgg	3.43
chrX	+	14823080	14823380	14826897	14827197	FBgn0264078	Flo-2	4.06

Table A.2: Genomic regions with evidence for the inter-relation of RNA editing and alternative splicing

Appendix B

Supporting Materials for Chapter 3

B.1 Selected TCGA overian serous cystadenocarcinoma samples

Sample ID	Alteration	Amino acid change	Type of mutation
TCGA-13-0891-01	Mutation	T10114_Q1016del	In frame deletion
TCGA-13-1495-01	Mutation	Q602*	Nonsense
TCGA-20-0987-01	Mutation	E928Gfs*27	Frame shift insertion
TCGA-25-1322-01	Mutation	Y901C	Missense
TCGA-25-2392-01	Mutation	R882L	Missense
TCGA-31-1953-01	Mutation	W719*	Nonsense
TCGA-59-2351-01	Mutation	K975E	Missense
TCGA-04-1332-01	Deletion	-	-
TCGA-23-1030-01	Deletion	-	-
TCGA-61-2003-01	Deletion	-	-
TCGA-10-0934-01	Amplification	-	-
TCGA-24-1431-01	Amplification	-	-
TCGA-61-2002-01	Amplification	-	-
TCGA-61-2092-01	Amplification	-	-
TCGA-04-1343-01	Control	-	-
TCGA-04-1348-01	Control	-	-
TCGA-04-1361-01	Control	-	-
TCGA-04-1517-01	Control	-	-
TCGA-09-1662-01	Control	-	-
TCGA-09-2053-01	Control	-	-

Sample ID	Alteration	Amino acid change	Type of mutation
TCGA-10-0927-01	Control	-	-
TCGA-10-0933-01	Control	-	-
TCGA-13-0720-01	Control	-	-
TCGA-13-0724-01	Control	-	-
TCGA-13-0800-01	Control	-	-
TCGA-13-0884-01	Control	-	-
TCGA-13-0897-01	Control	-	-
TCGA-13-0905-01	Control	-	-
TCGA-13-1407-01	Control	-	-
TCGA-13-1483-01	Control	-	-
TCGA-13-1492-01	Control	-	-
TCGA-13-1505-01	Control	-	-
TCGA-13-1506-01	Control	-	-
TCGA-13-1507-01	Control	-	-
TCGA-20-0991-01	Control	-	-
TCGA-23-1032-01	Control	-	-
TCGA-23-1116-01	Control	-	-
TCGA-24-0966-01	Control	-	-
TCGA-24-1419-01	Control	-	-
TCGA-24-1422-01	Control	-	-
TCGA-24-1436-01	Control	-	-
TCGA-24-1471-01	Control	-	-
TCGA-24-1545-01	Control	-	-
TCGA-24-1552-01	Control	-	-
TCGA-24-1553-01	Control	-	-
TCGA-24-1558-01	Control	-	-
TCGA-24-1563-01	Control	-	-
TCGA-24-1564-01	Control	-	-
TCGA-24-1565-01	Control	-	-
TCGA-24-1567-01	Control	-	-
TCGA-24-1603-01	Control	-	-
TCGA-24-1604-01	Control	-	-
TCGA-24-2038-01	Control	-	-
TCGA-24-2261-01	Control	-	-
TCGA-24-2290-01	Control	-	-
TCGA-25-1320-01	Control	-	-
TCGA-25-1321-01	Control	-	-
TCGA-25-2396-01	Control	-	-
TCGA-25-2399-01	Control	-	-
TCGA-30-1862-01	Control	-	-
TCGA-36-1570-01	Control	-	-
TCGA-36-1574-01	Control	-	-

Sample ID	Alteration	Amino acid change	Type of mutation
TCGA-59-2355-01	Control	-	-
TCGA-61-1728-01	Control	-	-
TCGA-61-1919-01	Control	-	-
TCGA-61-2009-01	Control	-	-
TCGA-61-2016-01	Control	-	-
TCGA-61-2095-01	Control	-	-
TCGA-61-2104-01	Control	-	-
TCGA-61-2111-01	Control	-	-

Table B.1: Ovarian serous cystadenocarcinoma samples selected from TCGA

B.2 qRT-PCR validation of identified ALE splicing events

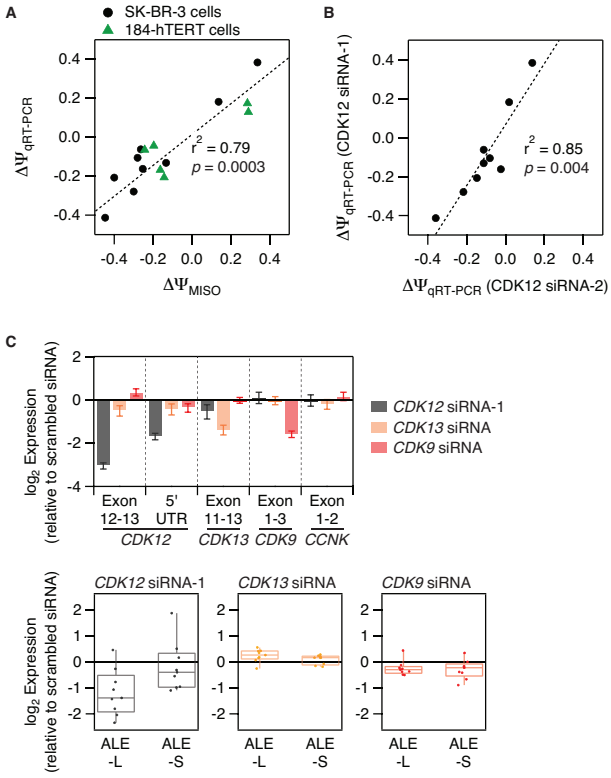

Figure B.1: qRT-PCR analysis of identified alternative splicing events. **A.** Correlation of $\Delta\Psi$ values for a panel of ALE events regulated by *CDK12* in SK-BR-3 and 184-hTERT cells, as determined by RNA-seq (MISO) versus qRT-PCR for genes: *NFX1*, *RIBF1*, *DNAJB6*, *BRCA2*, *DPP9*, *THADA*, *ZFYVE26*, *PADI2* and *ATM*. **B.** Correlation of $\Delta\Psi$ values for SK-BR-3 cells treated with two different *CDK12* siRNA constructs. **C.** Depletion of *CDK13* or *CDK9* did not phenocopy the effect of CDK12 depletion on ALE splicing. For nine genes with ALEs regulated by CDK12, qRT-PCR was used to measure changes in the expression of long (ALE-L) and short (ALE-S) mRNA isoforms after depletion of *CDK12*, *CDK13*, or *CDK9* in SK-BR-3 cells. Also, the depletion of the CDKs did not affect expression of *CCNK*.

B.3 Proteomics analysis of SK-BR-3 after CDK12 depletion

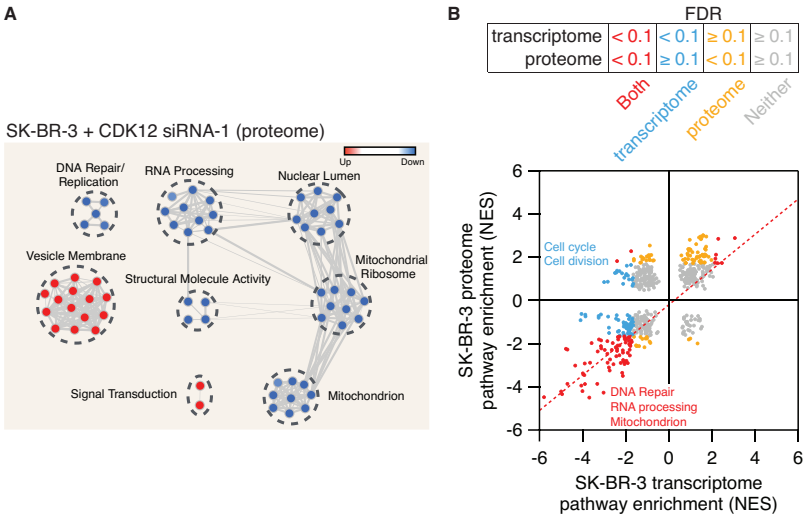

Figure B.2: Proteomic analysis of SK-BR-3 after CDK12 depletion confirms trends observed by differential gene expression analysis. **A.** Enrichment map from global proteome analysis in SK-BR-3 cells by GSEA. **B.** For each pathway, GSEA pre-ranked analysis assigned a normalized enrichment score (NES) representing the extent of over-representation of genes of a pathway at the top or bottom of a ranked list. Positive and negative NES values represent up- and down- regulated pathways, respectively. The dotted red line shows the general trend for NES values significant in both proteomics and transcriptomics datasets (FDR <0.1).

B.4 Up-regulation of cell proliferation pathways in MDA-MD-231 cells by *CDK12*

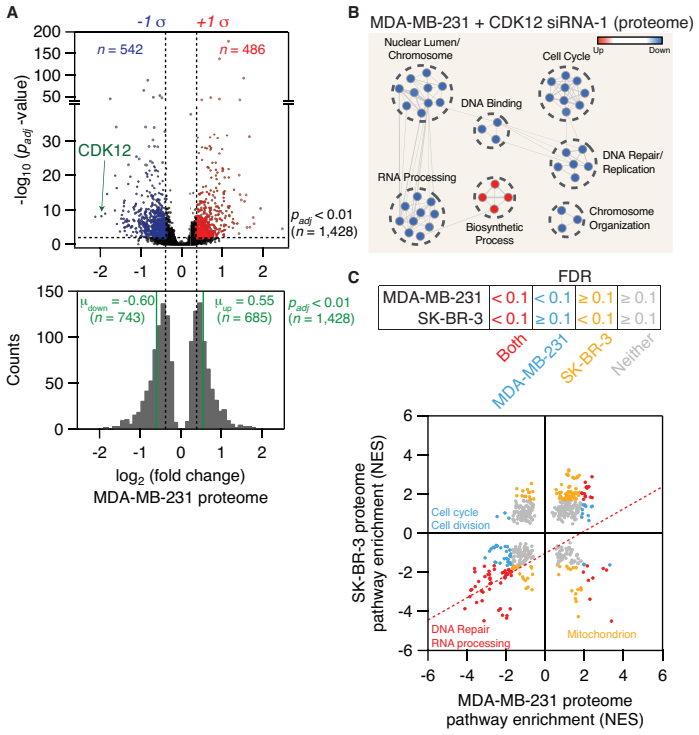

Figure B.3: CDK12 up-regulates cell proliferation pathways in MDA-MB-231 triple-negative breast cancer cells. **A.** Top: volcano plot of the global proteome analysis in MDA-MB-231 cells. Bottom: distribution of fold change values for all differential protein expression events with $padj < 0.01$. **B.** Enrichment map from global proteome analysis in MDA-MB-231 cells by GSEA. **C.** For each pathway, GSEA pre-ranked analysis assigned a normalized enrichment score (NES) representing the extent of over-representation of genes of a pathway at the top or bottom of a ranked list. Positive and negative NES values represent up- and down- regulated pathways, respectively. For each pathway, NES values in the MDA-MB-231 and SK-BR-3 proteome are shown. Red markers represent NES values significant in both cell lines (FDR < 0.1). The dotted red line shows the general trend of these points.