Large-scale study of RNA processing alterations in multiple cancers

Babita Singh

TESI DOCTORAL UPF / 2017

DIRECTOR DE LA TESI

Dr. Eduardo Eyras

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES (DCEXS-UPF)

UNIVERSITAT POMPEU FABRA

मम्मी, ये मेरी पहली किताब तुम्हारे लिए|

Acknowledgements

First of all, I would like to thank my supervisor and mentor Eduardo Eyras. Eduardo, this thesis wouldn't have been possible without you. Thank you for inviting me to this wonderful city Barcelona (which I must confess I had no idea about), and giving me this opportunity to do my PhD in your group. I have learnt a lot from you, your scientific understanding is at par and the way your brain works, I have always envied that. Your silent dedication and passion towards your work has constantly motivated us and sometimes made me question my own career choices. I also admire the effort you have put on to mentor me and the patience you've shown often times (sometimes explaining things over and over again). It was reassuring to know I can always reach you without any inhibitions, even for the smallest doubts. Thank you for being supportive, especially during the last year of my PhD. I was struggling with work and my own personal issues. It would have been so easy to give up, if I didn't find constant support, encouragement and patience from you. I will forever be grateful.

I would also like to thank this research park (PRBB) and people I met here. I cannot emphasize more, how tremendous it has been for being able to work here. There is so much knowledge crammed at this place and the scientific community is so strong. One of the hubs of bioinformatics and right next to the beach! Its open & inclusive atmosphere, interval courses, seminars, pizza talks, volleyball tournaments, beer sessions and various activities made my PhD time fun. Specially interval courses that gave me opportunities to satiate my creative cravings.

To Konrad and his lab for welcoming me in their group for my three months research stay in Wurzberg, Germany. Konrad, thank you for teaching me Python right from the basics. In those three months I was able to develop my own Python based tool, only because of you.

To the GRIB- IT crowd Alfons and Miguel, for troubleshooting and s/w installations, thank you for taking care of this, it definitely made life easier.

To all the wonderful people that came in the lab 486 while my time here. I have learnt a bit from all of you and spent a gala time in your company. Especially thanks to the group I started with, Juanra, Gael, Isaac, Endre,

Amadis, Jose Luis, Jorge and Janet, it took me no time to adjust in this new place and environment because of you all. Emma, Caterina, Hector, Juanlu, Will, Pau and JC, thanks for always keeping the group lively and fun. Our lunch table conversations were definitely a podcast material that sadly nobody will ever know. I am definitely missing so many names here, but you know who you are. Thank you for being around. You all became family and have made me grow as a person many folds. I never felt I am in a different country, despite not being able to speak the language. Spending my time in this group and living in this country gave me such wide perspective and expanded my knowledge, compared to what I came here with.

To my friends in India, my apologies to isolate myself and not being around. To my childhood best friend khushboo, despite being together since we were 2 years old, I missed your wedding and then your baby's birthday. It is unforgivable, I know. My Tinkers gang, Chiya, Ashish, Farheen, Azi, Tarun, why did we ever have to grow up and move on with lives? Iram di for always being around no matter what (even when I distanced myself from everyone else). Anupam ma'm for all that wonderful programming (and life) lessons. You made me believe in magic and to keep on seeking more. Mohit, for pushing me to do Master's degree, when I was kind of settling down with what I had, a job.

Finally, to my family. My brothers, Vijay and Aditya. Thanks Vj, for teaching your honesty, hardwork and perseverance and Adi, for teaching your joyfulness, creativity and curiosity. I feel so proud when I see how we have supported each other all this while and how well you both are doing. A heartfelt thank you to my mom and dad for all the love and warmth we received from you and for all your struggles and sacrifices. For making us believe right from the beginning that knowledge is one true power. Thank you for always letting me choose my own way, I don't know how you both managed to be so different than the rest of the society we saw around. Thank you for not conditioning our minds and letting us learn to form our own understanding and opinions. Even when we had our moments of struggles, my mind was free to dream and imagine, it was all because of you both. Thank you.

Abstract

RNA processing and their alterations are determinant to understand normal and disease cell phenotypes. In particular, specific alterations in the RNA processing of genes has been linked to widely accepted cancer hallmarks. With the availability of large-scale genomic and transcriptomic data for multiple cancer types, it is now possible to address ambitious questions such as obtaining a global view of alterations in RNA processing specific to each cancer type as well as in common across all types. The first objective of this thesis is to obtain a global view of RNA processing alterations across different tumor types along with alterations with respect to RNA binding proteins (*trans*-component), their tumor-type specificity, differential expression, mutations, copy number variation and whether these alterations result in differential splicing. Using data for more than 4000 patients from 11 tumor types, we provide the link between alterations of RNA binding proteins and splicing changes across multiple tumor types. Second objective moves one step further and explores in detail the RNA-processing alterations with respect to mutations on RNA regulatory sequences (*cis*-components). Using whole genome sequencing data for more than 1000 cancer patients, we thoroughly study the sequence of entire genes and report significantly mutated short regions in coding and non-coding parts of genes that are moreover enriched in RNA putative RNA regulatory sites, including regions deep into the introns. The recurrence of some of the mutations in non-coding regions is comparable to some of already known driver genes in coding regions. We further analyze the impact of these mutations at the RNA level by using RNA sequencing from the same samples. This work proposes a novel and powerful strategy to study mutations in cancer to identify novel oncogenic mechanisms. In addition, we share the immense amount of data generated in these analyses so that other researchers can study them in detail and validate them experimentally.

Resumen

El procesamiento del ARN y sus alteraciones son determinantes para entender el fenotipo de las células en condiciones normales y de enfermedad. En particular, alteraciones en el procesamiento de ARN de determinados genes se han vinculado a características distintivas del cáncer ampliamente aceptadas. Con la disponibilidad de datos genómicos y transcriptómicos a gran escala paramúltiples tipos de cáncer, es posible abordar cuestiones ambiciosas como la obtención de una visión global de las alteraciones en el procesamiento de ARN que son específicas para cada tipo de cáncer, así como de aquellas las comunes a varios tipos. El primer objetivo de esta tesis es obtener una visión global de las alteraciones del procesamiento de ARN en diferentes tipos de tumores, así como de las alteraciones en las proteínas de unión a ARN (componente *trans*), y si dichas alteraciones resultan en un procesamiento diferencial del RNA. Utilizando datos de más de 4000 pacientes para 11 tipos de tumores, establecemos la relación entre las alteraciones de las proteínas de unión a ARN y cambios de *splicing* en múltiples tipos de tumores. El segundo objetivo va un paso más allá y explora en detalle las alteraciones del procesamiento de ARN con respecto a mutaciones en las secuencias reguladoras del ARN (componente *cis*). Utilizando datos de genomas completos para más de 1000 pacientes, estudiamos a fondo la secuencia de genes para identificar regiones cortas significativamente mutadas en partes codificantes y no codificantes por proteína, y que además están enriquecidas en posibles sitios reguladores del ARN, incluyendo regiones intrónicas profundas. La recurrencia de las mutaciones en algunas regiones no codificantes es comparable a la de algunos genes *drivers* de cáncer conocidos. Además, analizamos el impacto de estas mutaciones a nivel del ARN mediante el uso de datos de secuenciación de ARN de las mismas muestras. Este trabajo propone una estrategia novedosa y potente para estudiar las mutaciones en cáncer con el fin de identificar nuevos mecanismos oncogénicos. Además, compartimos la inmensa cantidad de datos generados en estos análisis para que otros investigadores los puedan estudiar en detalle y validarlos experimentalmente.

Preface

Before introducing the work in this thesis, perhaps I should talk a bit about my own journey in science so far. Sixteen years ago, in August 2000, I happen to pick up a science magazine at a city bus station in Gorakhpur, India. An initial draft of human genome was just released in June 2000 and this magazine covered the news with an attractive allegory and catchy headline about genes and their 'secrets'. The article started with ‘what is a gene?’ I must say the definition was pretty simple back then. As I turned pages after pages, I read in amusement about the Human Genome Project (HGP) and the amount of data this project was generating. Like a dramatic science fiction storyline, it also mentioned the race between the private (led by J. Craig Venter) and public (led by Francis Collins) endeavors and how all diseases will now be curable. I was enchanted. Two years later, in 2003, when the full human genome got published, I was graduating from school. I submitted my final grade biology report on the Human Genome Project using the same allegory I found on that magazine’s cover page (a bit like the cover page of this thesis). Unfortunately, my biology teacher had no idea about the HGP and during a *viva voce*, I was asked questions from my text book and not from my project. I wasn’t thrilled. He later advised me to pick some ‘soft’ fields for further studies. “Science is a long way till PhD and for girls it is better to choose fields like commerce or arts’, he added.

Mathematics was never my strong subject, so I was already indecisive, but now I wanted to study science. I pursued bachelors in life sciences for three years but didn’t quite enjoy it. The course was on zoology and botany and I wanted to study modern science like genes and DNA. Luckily, when I finished my bachelor’s degree, I found an advertisement about a diploma course in bioinformatics. It was the year 2006, and bioinformatics was still less known in India, compared to thousands of engineering institutions and courses. There were no usual study materials or designed curriculum but I

connected with this field more. Soon, I started learning about NCBI and EMBL, about genomic sequences, gene predictions, sequence alignments, 3D visualizations and programming languages. I felt so powerful that I could ask questions and write a code to get results, everything on my fingertips! This was so rewarding for my instant-gratification seeking curious mind. However, after I finished the course, I learnt that to apply for a PhD, I needed a Master's degree and not diploma, even though I spent the same two years as I would have studied in Masters. At this point, I gave up and joined a research group as a project assistant instead. It was a wet-lab group researching on cancer, but my bioinformatics skills were limited to excel sheet calculations. Anyone who has ever worked in bioinformatics would cringe inside, reading that sentence. After spending a year in that lab, finally I enrolled for a Master's Degree in Bioinformatics in 2009. Three years later, at the time when ENCODE (a massive extension of HGP) got published in September, 2012, I was in talks with my supervisor Eduardo about this PhD position and I felt so delighted to join the place I had just read about in the newspaper! Since then, the four years that I have spent in this PhD I have learnt so much, both at professional and personal level. In this group, I got the opportunity to learn extensively about various aspects of genome informatics, analyze genome sequencing data of thousands of patients. I've worked and collaborated with such smart brains that I often questioned, where do I stand? I got opportunities to meet and learn from pioneering scientists in this field, attend conferences and visit genomics hubs in Europe that I had read about in my text books. Several such moments made me feel both humble and proud.

Finally, sixteen years later, the moment when I felt this journey has come to full circle now, was during the last year of my PhD retreat in 2016, when I attended an hour long debate session held between two renowned experts in the field of genomics, Dr. Mar Alba and Dr. Roderic Guigo, debating about the same question it all started with, '*What is a gene?*'

DNA is considered as the key component of life. In order to study genes, we study DNA sequences, though DNA is not the one directly responsible for diseases. There are alterations in DNA which then lead to altered RNA and proteins responsible for diseases. Thus, RNA is often called the first phenotype of a cell. It is still unclear at which point of time in the 4 billion years old Earth's history, DNA emerged as genetic material and a key component of life. Many independent evidences have established that an RNA world existed prior to a DNA world. This indicates that the biochemical material available was self-sufficient for life processes until some stability factor came along. The fact that many crucial components of cells are composed of RNA or processed through an RNA-based machinery, including the splicing process, establishes the importance of RNA: RNA molecules provide for many genes an intermediate stage between DNA and protein molecules (Cech 2012), there are currently many more genes predicted to be non-coding RNAs than protein-coding, and the transcriptome, the set of all RNAs in the cell, is often considered to be the first phenotype in eukaryotic cells. The various regulatory steps from DNA transcription, through RNA processing, and up until mRNA translation in some cases, decide the fate of cell function. Therefore research in RNA biology has become a cornerstone for modern day research of diseases and therapeutics. This development came along with new technologies to detect and measure RNA molecules, such as RNA sequencing (RNA-seq). In order to understand how things work inside cells and what goes wrong in diseases, much of current research relies on the measurement of RNA using high-throughput methods.

It is perhaps remarkable, that a single experimental approach, like RNA-seq, across multiple samples and conditions, provides enough information to infer multiple mechanisms of disease at molecular level. This thesis describes the development and implementation of tools to study RNA alterations in cancer, hence pushing the limits of knowledge extraction from high-throughput sequencing methods such as data generated by The Cancer Genome Atlas (TCGA). TCGA is a large consortium that started in 2005 and so far has created 2.5 petabytes worth of tumor data from almost

11,000 patients. As this thesis describes later, cancer is a complex multi-genic disease with diverse set of genetic and epigenetic alterations. The different types of cancers sing the same tune but with different lyrics. We can recognize the common tunes now but even a tiny variation in lyrics might change the whole meaning of the song, such as in rare cancer types (occurring < 6 cases in 100,000 patients). As we are aiming towards more inclusive cures such as gene therapies and precision medicines, we need much broader perspective to understand these variations. This means that more genomics data is required and so are the researchers that can study these data. Perhaps a future next step would be the creation of a world databank with the sole purpose of providing multi-level cancer data to researches. Researchers from all over the world could then write grant proposals to work with these data.

Another bottleneck that I often wonder about is whether there will be enough researchers in the field to take on this massive task of data generation and analysis. With funding size getting relatively smaller and limited new job creations, would enough researchers still be interested in pursuing or staying in the field? On the other hand, why data analysis should be limited to researchers only? While tech giants like Bill Gates and Mark Zuckerberg are urging young people to learn programming, influential scientists might urge the same: for people to understand basic genome architecture and the basic analysis pipelines of genomics data. Now with genome sequencing technologies like Nanopore MinIONs that can be plugged into a laptop, with starting kits from $1000, genome data analysis can be the next programming language for people around the world. Would there be ethical concerns involved? This is a matter of perspective. If proper education and considerable background is provided, it is less likely that one will still believe in the era of Jurassic park science fiction. And if they do, then it might be suggesting that we as a scientific community are failing to bridge the gap between the real world and us by creating a niche for ourselves. As Carl Sagan once said, "*We live in a society absolutely dependent on science and technology and yet have cleverly arranged things so that almost no one understands science and technology*."

Table of Contents

I. Introduction

1. Gene, Gene Expression and Splicing

What is a gene? A gene can be defined as the unit of heredity that is transferred from parents to offsprings and which determines certain traits in the individuals. Genes are encoded in the DNA of each cell, and each gene is currently considered to be defined as a stretch of nucleotide sequences in the double-stranded chromosomes. During the process called transcription, a precursor RNA, often called pre-messenger RNA or pre-mRNA, is created by copying the nucleotides from one of the two DNA strands. A gene is made up of introns and exons. Introns are removed, or spliced out, mostly during transcription, and exons are joined together to form mature RNA molecules. The transcription and post-transcriptional maturation of the pre-mRNA gives rise to mature RNA transcripts that are either translated into. However, many RNAs are not translated and exert their function as non-coding RNAs. Each of the steps during the transcription and maturation of the RNA molecules is complex and highly regulated process, involving multiple molecular complexes that work in precise coordination. Explaining all of these steps would be out of the scope of thesis and therefore I will focus only on a small part called ‘alternative splicing’.

1.1 Splicing

In eukaryotes genes are composed of exons and introns, where exons are considered to carry the functional information of the mature RNA molecule: protein-coding regions in the case of protein-

coding genes, structured regions in the case of structural RNAs (e.g. snoRNAs and tRNAs), or some other possible functional regions yet to be characterized for many of the so far measured long non-coding RNAs. These exons are intervened by long sequences of introns, which need to be removed through highly dynamic ribonucleoprotein complex known as spliceosome (Lee and Rio 2015). Splicing is the process whereby introns are removed from the precursor RNA to create mature RNA sequences. This process can occur in different ways for the same gene, giving rise to alternative splicing (AS). Among all the processing steps during gene expression, AS provides perhaps the largest potential for molecular diversity and controlled regulation in the cell. Almost 95% of human genes undergoes alternative splicing (Wang et al. 2008).

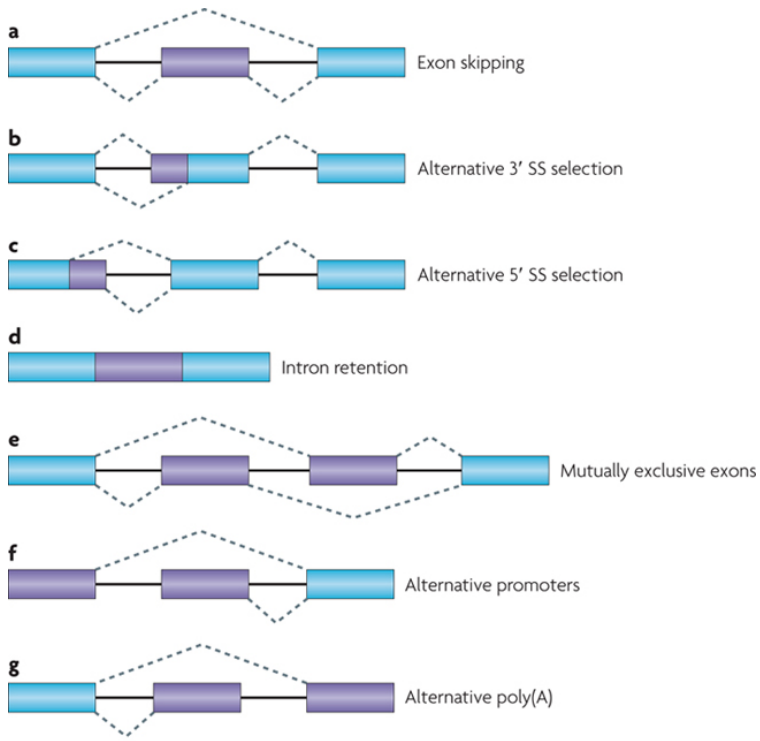


Figure 1. Different types of alternative splicing (Keren et al. 2010).

Alternative splicing is usually described in terms of local variations of the exon-intron structures, or events. There are seven main types of alternative splicing events (Figure 1). a) Exon skipping (SE), also known as cassette exon, is the most common type of splicing event. In SE, exons are included or spliced out from the transcripts. b) In Mutually Exclusive exons (MX) either of the two exons is selected. This type of event is usually related to exons that rarely or never co-exist in a transcript. Another type is alternative 5' splice-site (A5), where two different 5' splice sites (donor sites) may be selected. Likewise, in the alternative 3' splice-site (A3) different 3' splice sites (acceptor sites) may be included in the transcripts. In Intron retention (RI) a complete intron is included in the transcript instead of being spliced out during the splicing process. Another type is Alternative First (AF) exons, where alternate starting exons are selected, and is often coupled with transcriptional regulation. Finally, there is the Alternative Last (AL) exons, which describe two possible 3' terminal exons in the transcripts. These are the simplest and most common local transcript variations, although more complex events can take place.

1.2 Splicing Regulation

1.2.1 Core splicing machinery

The process of splicing is catalyzed by a multi-megadalton ribonucleoprotein (RNP) complex, which is dynamically assembled together to create a highly accurate RNA processing machinery, the spliceosome. The spliceosome works as a film editor, cutting out the irrelevant material (introns) from the show reel, to keep the important information (exons) for the main movie. Introns are generally defined by specific sequence signals (Figure 2).

Splicing signals - core components:

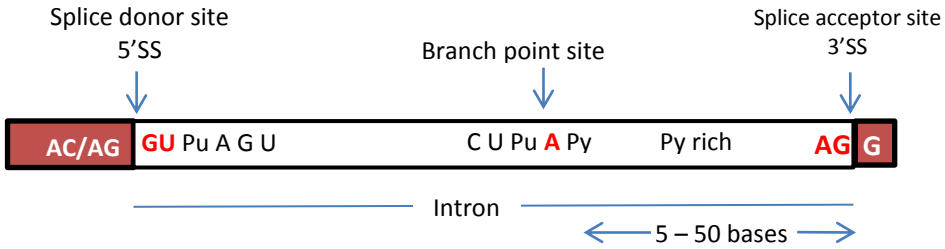


Figure 2. A typical conserved nucleotides for splicing in introns and exons Pu = A/G, Py=C/U

The core components of these signals include the motifs at the 5' and 3' splice-sites, which correspond to multiple positions around the exon-intron boundaries. There is also the polypyrimidine tract (PPT), which is a pyrimidine-rich stretch of nucleotides located 5-50 bases upstream from the 3' splice site, and which recruits the U2 small nuclear RNA auxiliary factors 2 (U2AF2), also called U2AF65. This factor generally binds together with the U2AF1 (U2AF35), which interacts with the 3' splice site. The branch-point site is made of a conserved adenosine and corresponds frequently with the motif CURAY (Corvelo et al. 2010). The BP is initially recognized by the splicing factor 1 (SF1) and also reflects, except for the BP A, it reflects the base-pairing with the U2 snRNA in U2 introns (Figure 2).

In eukaryotes, there are two types of spliceosome machineries: the U2 and the U12. The U2 spliceosome catalyzes U2-type introns, which are the majority of introns in eukaryotes, whereas the U12 spliceosome catalyzes U12-type introns, which are a small subset of the introns and typically occur at the frequency of 1 in every 5,000 to 10,000 introns (Sharp and Burge 1997). The U2 spliceosome

machinery is composed of five small nuclear ribonucleoproteins (snRNPs) named U1, U2, U3, U4, U5 and U6. In addition to the snRNPs, few other proteins are required for the spliceosome assembly, including the U2 small nuclear RNA auxiliary factors 1 (U2AF1) and 2 (U2AF2) and SF1. U12-dependent introns are spliced-out by their specific snRNPs, the U11 and U12 snRNPs, which generally act together as a single complex (Will and Lührmann 2011). U2 and U12 dependent introns have distinct sequence features (Figure 2). Most of the U2 introns have the consensus GT-AG at the splice-site positions. On the other hand, although many U12 introns also have this consensus, a considerable proportion have AT-AC (Turunen et al. 2013).

U2 and U12 splicing signals

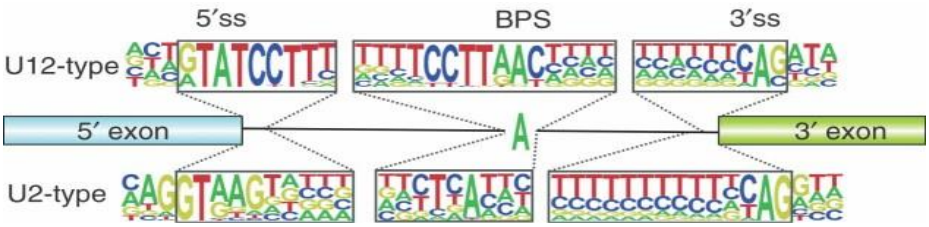


Figure 3. Sequence features of U2 and U12 introns

In U2 introns, the 5’ signal corresponds to the complementarity to the U1 snRNA, whereas in U12 introns, the 5’ss consensus is rather CCUURAY, and corresponds to the base pairing with the U11 snRNA. U12 introns generally show a stronger branch-point consensus and a shorter and almost non-existent PPT as well as a shorter distance between the BP and the 3'ss. In contrast, U2 BPs have a weaker consensus and, although the distance to the 3'ss can be longer, most of the U2 BPs lie between 5nt and 50nt upstream of the 3'ss.

Steps in canonical splicing

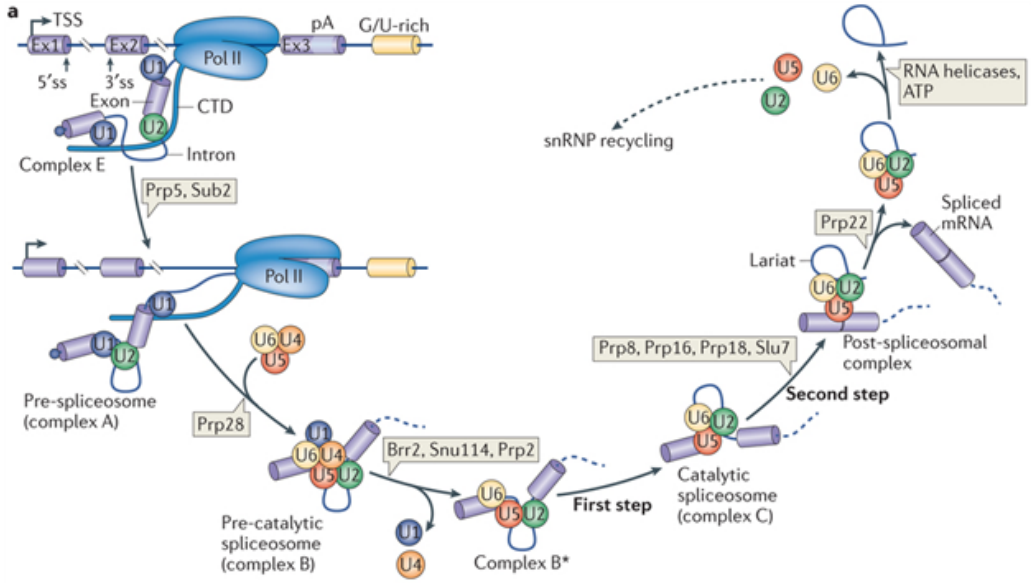


Figure 4. Step wise processes in canonical splicing as explained by (Matera and Wang 2014).

Over several steps of splicing process, spliceosomal proteins form different complexes to perform *canonical* splicing. Step 1: Small nuclear ribonucleoproteins (snRNPs) U1 and U2 recognizes the 5' and 3' splice sites of exon-intron boundaries, mediated by the carboxy-terminal domain of RNAPIIpol-II (Complex E). Step 2: Interaction of snRNPs U1 and U2 with each other results into the formation of pre-spliceosome complex (Complex A) in the presence of helicase pre-mRNA-processing 5 (Prp5) helicase. Step3: Next three other pre-assembled snRNPs U4-U6-U5 are recruited at the site to form complex B. This is catalyzed by helicase pre-mRNA-processing 28 (Prp28). Step4: Complex B then forms catalytic active complex B (complex B*) by undergoing series of rearrangements

which results into the release of snRNPs U4 and U1. This process is mediated by RNA helicases Prp2, Brr2 and snu114. Step5: Complex B generates a catalytic spliceosome complex C that contains free exon and corresponding exon-intron lariat, this is the intermediate state. Step6: Complex C again goes for rearrangements to create post-spliceosomal complex. This complex performs a second catalytic process that contains the lariat intron and spliced exons. Additional helicases are utilized at this step. Step7: At this step, the three snRNPs U2, U5 and U6 are released, mediated by helicase Prp22, and recycled again to be reused for splicing. Finally, the intron lariat undergoes degradation (Figure 4). Most of the mechanisms related to gene expression take place in a coordinated way that couples transcription with pre-mRNA processing. Co-transcriptional splicing seems to be quite prevalent and advantageous for the efficiency of splicing (Naftelberg et al. 2015). There is also plenty of evidence showing that splicing regulation depends on the coupling with the dynamics of RNA polymerase II (RNAPII). During transcription, the spliceosome machinery assembles to perform the process of splicing.

1.2.2 Additional components, (silencers and enhancers)

In contrast to constitutive splicing, which refers to exons and splice-sites that are always processed in the same way, alternative splicing is related to a competitive regulation between exons or splice-sites, resulting in multiple RNA molecules from the same gene locus. Several components participate to decide the fate for skipping or inclusion of exons in alternative splicing, and these will work differently depending on multiple factors, like tissue type, cell-type and disease state. Although alternative splicing often occurs in exons with weaker splice sites the strength of the splice site cannot explain

completely the inclusion level of exons in different conditions (Barash et al. 2010). Indeed, splice sites are not the only signals governing the recognition of exons and the complexity of alternative splicing regulatory process goes much deeper than the core spliceosomal machinery. Alternative splicing is generally controlled or triggered by multiple RNA binding proteins (RBPs), known as splicing regulators.

Splicing Silencers and Enhancers

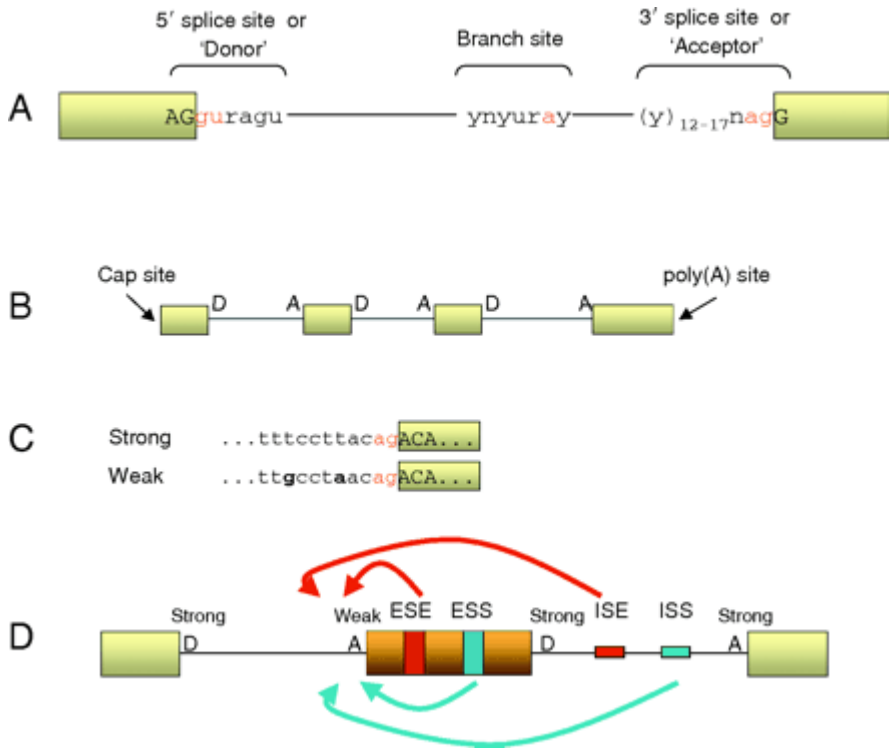


Figure 5. Cis regulatory sites controlling alternative splicing. (Srebrow and Kornblihtt 2006)

The interactions of splicing regulators with their cognate RNA sites generally control and regulate the splicing decision (Witten and Ule 2011). These cis-regulatory sites are found on exons, as well as typically 200-300 nucleotides up or downstream of exons. These regulatory elements are denominated splicing enhancers or silencers, as they can function as activators and repressors of the splicing mechanism, respectively (Fairbrother et al. 2002; Wang et al. 2004). These can occur in exons as Exonic Splicing Enhancers (ESEs) or Silencers (ESSs), and in introns as Intronic Splicing Enhancers (ISEs) or Silencers (ISSs) (Figure 5) (Wang et al. 2012). A large amount of these regulators have been identified using experimental and computational methods (Yeo et al. 2004; Stadler et al. 2006), and they can have a changing role depending on their position along the exon or the intron (Goren et al. 2006). These results highlight the variety of sequences that can function as splicing *cis*-regulatory elements, and their position-specific effects. The determination of the binding affinities of multiple RBPs have helped in the recognition of many of these splicing enhancers and silencers as specific binding sites of RBPs that work as splicing regulators (Ray et al. 2013; Lambert et al. 2014; Ule et al. 2003).

Figure 6 describes various modes of mechanisms in the regulation of alternative splicing through splicing silencers and enhancers.

Mechanism of regulation

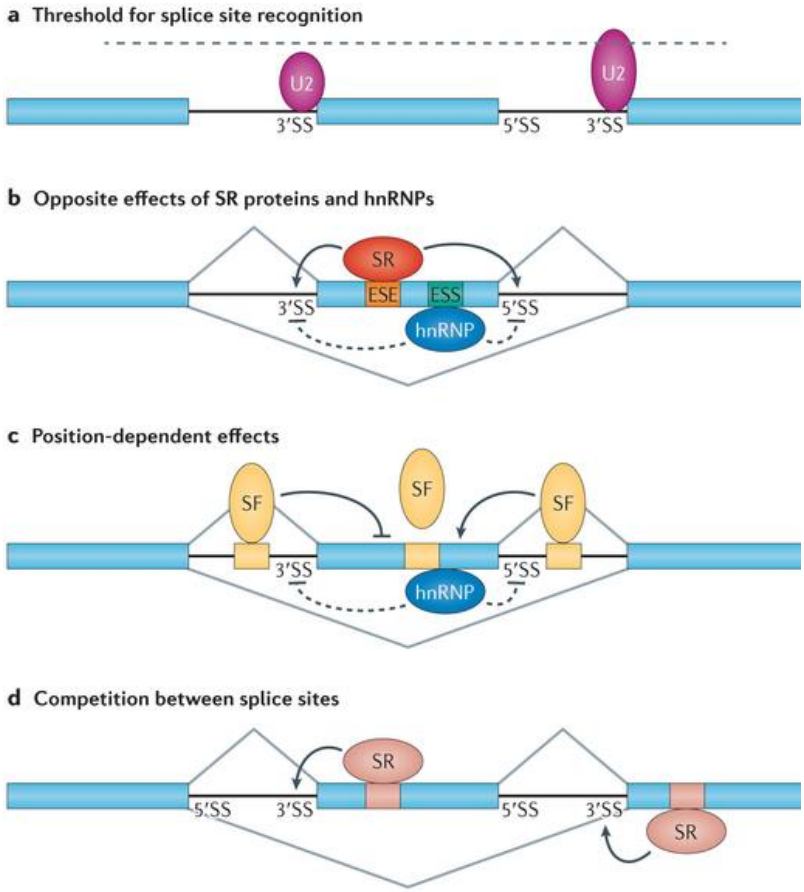


Figure 6. Rules for context-dependent and position-sensitive regulation of alternative splicing.(Fu and Ares 2014)

1.3 Computational Methods to study RNAseq data and splicing

1.3.1 RNA Sequencing

Thanks to the developments in the technologies for the sequencing of ribonucleotides, starting with Sanger's chain termination technique in 1977 (Sanger et al. 1977) and later with the start of the

sequencing of the human genome, genomics has come a long way. Importantly, advancement in computational biology has taken place in parallel. If human landing on the moon was done using computers less powerful than our smartphones today, one could argue the same when comparing current computational genomics research with the methods and technologies available 13 years ago. The human genome project took a total budget of $3 billion and 13 years to accomplish, including the time of genome mapping (Archive), which means $1 per base sequenced every 0.14 second. Some argue that this could now be achieved with **$1000** within a 3-day run with 30X coverage (Illumina HiSeq X Ten) i.e. $0.00000033 dollar per base for every 0.000086 sec. Even though we have not included here the cost of analyzing the data, one may argue that this can be done much faster and cheaper nowadays. The technological revolution in sequencing combined with the advancements in computational biology has led to the appearance of a myriad of methods and algorithm to analyze data. In particular, there are many methods to study splicing from RNA sequencing data (Alamancos et al. 2014).

1.3.2 Computational Methods to study splicing

High-throughput RNA sequencing (RNA-seq) allows the measurement of the set of RNA molecules, the transcriptome, in a sample. RNA-seq data is mostly used to study differential expression analysis of genes, but it allows measuring other patterns in the transcriptome, like differential splicing. Differential splicing is related to a change in the relative abundances of the gene-isoforms (transcripts), which may occur without an observable expression change for the given gene, and hence it provides orthogonal regulatory evidence. If the genome of an organism is available, a

typical method to study alternative splicing starts with mapping RNA-seq reads on the genome to identify the locus of the transcripts they originate from. Otherwise, de novo reads may be assembled into contigs to build a candidate transcriptome. The different methodologies used for splicing analysis from RNA-seq are reviewed in detail in (Alamancos et al. 2014) and briefly described here:

Spliced mapping of reads: First step in the analysis pipeline is to use splice-aware tools to map sequencing reads back to their gene locus of origin. Splice aware mapping tools usually follow two methods; *exon-first* methods use first an unspliced approach to map the reads to genome and create read clusters. Unmapped reads are then used to connect these read-clusters. These methods tend to be faster but usually require high coverage on exons. TopHat (Kim et al. 2013), and STAR (Dobin et al. 2013) are tools that use this approach. *Seed-and-extend* is another method of reads mapping that maps reads as k-mers or substrings and locate the splice-sites by extending matched reads at both directions. This method is comparatively slower although recovers more novel splice-sites and some aligners use both approaches together (see (Alamancos et al. 2014) for a complete set of references).

Quantification of the splicing variants: After mapping reads on their respective transcript coordinates the next step is to estimate the expression levels of transcripts and exon inclusion or exclusions. There are two ways to study splicing quantification 1) *events based:* Here the splicing events are created by using the annotation of exon-intron junctions or directly from the mapped reads, and then subsequently the inclusion levels of each event is estimated. Tools

like SUPPA (Alamancos et al. 2015), MISO (Katz et al. 2010), rMATS (Shen et al. 2014) or MAJIQ (Vaquero-Garcia et al. 2016) provide event quantification. PSI of an event is generally calculated by estimating the ratio between expression of the transcript(s) with a particular event compared to expressions of all transcripts that include or exclude the event (Klinck et al. 2008), which the various methods estimate in different ways from sequencing reads. 2) *isoform based* : Another method of quantification is to assign reads on different isoforms of the gene and estimate the expression for each isoform using, for instance, Cufflinks, Salmon, Kallisto or Sailfish (Trapnell et al. 2010; Patro et al. 2015; Bray et al. 2016; Patro et al. 2014). The isoform expression values are reported in RPKM (Reads per Kilobase per million of mapped reads) or TPM (Transcripts per million) or using relative expression values such as PSI (percent/proportion spliced index), which is calculated as the relative abundance of the transcripts over the total gene abundance.

Alignment free quantification: A new generation of tools appeared recently that can provide transcript quantification without relying on genome alignments. Tools like Sailfish, Salmon and Kallisto (Patro et al. 2014, 2015; Bray et al. 2016) are alignment independent quantification tools that take the sequences of all the transcripts one wants to quantify and outputs the estimated read counts and TPM values for each transcript. This method of transcript quantification is very fast but is limited to annotated transcripts.

Differential splicing: For *events based* methods, the change in splicing is calculated by estimating the changes in the relative abundance of events or exons between two conditions. Methods like

MISO, rMATS, MAJIQ, and SUPPA calculate the difference in PSI (delta-PSI) between conditions. For example SUPPA (Alamancos et al. 2015) provides the delta-PSI (between -1 to +1) for all type of splicing events (SE,A3,A5,RI,MX,AF,AL) and creates a unique identifier for each event (*<gene_id>;<event_type>:<seqname>:<coordinates_of_the_event>:strand*) as shown in Figure 7 (Entizne et al. 2016). Another way to calculate differential splicing is through *isoform based* methods. Tools like Iso-kTSP (Sebestyén et al. 2015), SwitchSeq (Gonzàlez-Porta et al. 2013), DrimSeq (Nowicka and Robinson 2016) and SUPPA also identify transcripts that significantly change in relative abundance between conditions.

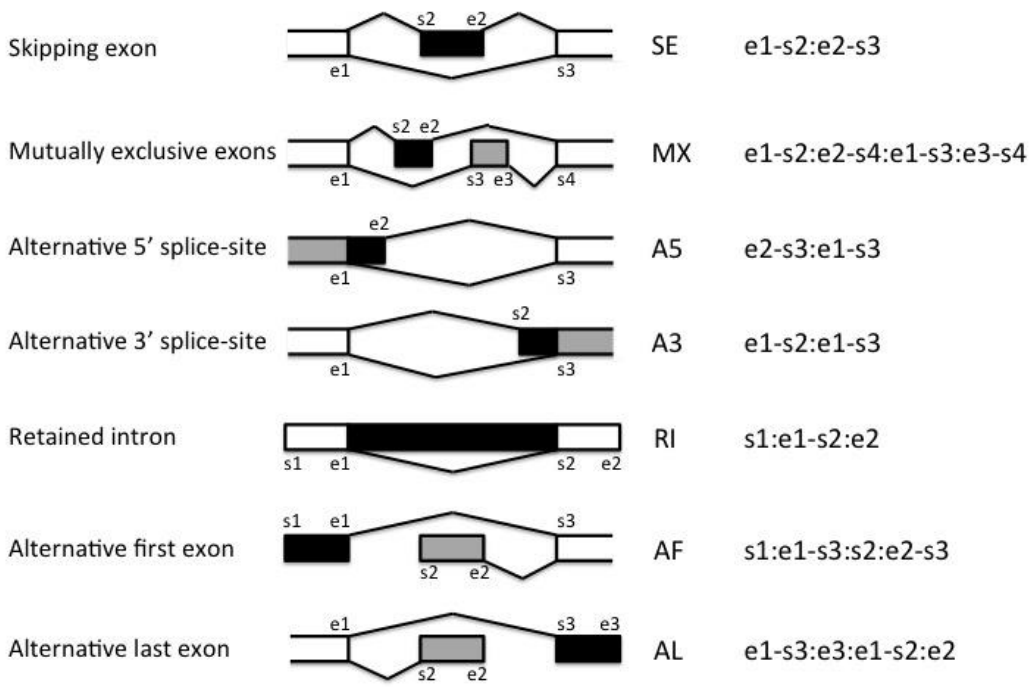


Figure 7. SUPPA event ids to create unique identifiers for alternative splicing events (Alamancos et al. 2015).

1.3.3 METHODS TO STUDY SPLICING REGULATION

Two main features are generally studied in the context of the regulation of splicing and disease: the differences in splicing between normal and disease conditions, and the changes in expression of regulatory RNA binding proteins and splicing factors (Trans component) that might be responsible for the splicing changes through their binding to RNA.

a. Trans-component: Multiple approaches are available to determine the expression change between different conditions computationally. Read counts per gene are normalized using tools like TMM (Robinson and Oshlack 2010), DESeq (Anders and Huber 2010), PoissonSeq (Li et al. 2012), which take into account read depth and transcript length and ignore highly variable transcripts that might skew the results. Tools like DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010) requires already normalized data to perform differential expression analysis. Cuffdiff (Trapnell et al. 2012) is another method to study expression change at the gene and transcript level. The genes thus found changing expression between conditions can further be confirmed using qPCR and Western-blot experiments. Further, the association between the differential splicing observed and the possible regulators can be validated using knockdown or overexpression experiments. For example if a gene shows a positive correlation between its expression change and a change in splicing patterns between conditions, a knockdown of this gene should reverse the splicing patterns (see e.g. (Sebestyén et al. 2016)). This way one could map the impact of RNA binding proteins on splicing.

b. Cis-component: To establish whether the changes in splicing patterns between conditions could be due to the direct interaction of RBPs, one needs to identify the actual binding sites in the sequence of a gene. RNA binding proteins recognize one or more small stretches of nucleotides (4-7nt) on the pre-mRNA (Daubner et al. 2013; Ray et al. 2013). If RNAseq

is the method to study expression and splicing in genes, crosslinking and immunoprecipitation (CLIP) followed by deep-sequencing (CLIP-Seq) has become the standard protocol to identify binding sites for RNA binding proteins (Ule et al. 2003; Macias et al. 2012). This method captures cross-linked RNA-protein complexes, i.e. RNA sequences bound to a specific protein. The fragments of RNA are then sequenced and statistical analysis is then performed to identify the likely binding regions. Methods like CIMS, Piranha or Pyicoclip (Uren et al. 2012; Althammer et al. 2011) are some of the methods developed for this task. See (Bottini et al. 2017) for a recent benchmarking analysis. Using the significant CLIP regions, a sequence enrichment analysis can be performed, by comparing with control regions, to identify a consensus binding motif. The simplest description of a motif consists in a position weight matrix (PWM), which represents the frequency of the four nucleotides A,T,G,C, at each position of the motif. Once a motif is identified, tools such as MEME-suite (Bailey et al. 2009) could be used to scan sequences to locate the positions of the motif in introns or exons of a gene. Tools and methods to study motifs are described in detailed further in the thesis. Alternatively, one can perform motif enrichment analyses in the differentially spliced events, compared to controls, and identify the possible regulators from a set of previously identified binding affinities (Ray et al. 2013; Lambert et al. 2014).

2. Cancer

2.1 Cancer brief description

First known records of cancer date back to 1600 BC in Egyptian literature, and included the description as well as the removal procedure of breast tumors. It was then concluded that this was a disease with no cure (2014). Unlike many other diseases that since their discovery have been successfully treated or eradicated, cancer has become a large jigsaw puzzle. Scientists and doctors are trying to fit in pieces together, but the puzzle remains incomplete. Since 1971 over 200 billion dollars have been invested on basic and clinical research in cancer in the US alone (Begley 2008). This has made possible to reduce the cancer death rate by a total of 25% since its peak in 1991 (Siegel et al. 2017). Cancer research and treatment has come a long way from the initial interventions, including surgically removing the tumors in live patients body (without anesthesia) during the 1800s. Many scientific and surgical landmarks have determined the approach to cancer. For instance, radiation as treatment started to be used in 1903, five years after Curie's discovery of radium. More than hundred years later, it is still an essential component in cancer treatment. The development and implantation of Pap-test screenings (1950s), as well as the programs to increase colorectal and mammography screenings (late 1970s) have contributed to the decrease of deaths related to cancer. The development of chemotherapy, combination of chemotherapy (1950s), vaccines against hepatitis B (1981), as well as the identification of carcinogens such as smoking (1960s), asbestos (1970s), benzene (1980s), radium, etc, have also contributed to the advancement in the prevention and treatment of the disease. However, cancer research has been largely dominated by the strategies of trial and error. It is only recently that specific molecular mechanisms have been identified and specific drugs to target these mechanisms have been developed (early 1990 onwards). For instance, the first ever drug for targeted therapy was Rituximab, developed

against B-cell non-Hodgkin lymphoma (1997); followed by Herceptin for women with early-stage HER2-positive breast cancers (1998). Gleevec was developed for rare Leukemia and imatinib for rare abdominal tumor called gastrointestinal stromal tumor (2001). We could conclude that within a hundred years cancer became a partially curable, preventable disease to some extent. Remarkably, all these achievements took place before the human genome was known.

The developments in genome sequencing have shifted the focus of cancer towards the genetic characterization of the tumors and the development of targeted therapies directed to specific genetic alterations. This has also highlighted the fact that cancer is not a single disease, but many diseases, almost as many as patients. According to recently released annual statistics report from the American Cancer Society, more than 200,000 new cases of *rare* cancers are expected to be reported in 2017 in the US alone (Siegel et al. 2017). Rare cancer cases are those with 6 or fewer cases per 100,000 people per year. In other words, 1 in 5 people diagnosed with cancer in Europe have a rare cancer type. Sometimes this could go as low as 1 case in 100,000 people. If in the pre-genomic era the efforts were focused on identifying and curing common cancer types, in the post-genomic era the attention has shifted towards the search of cures for each individual cancer, exploiting the information of genomics data and working towards patients-specific therapeutic strategies.

2.2 Genomics and cancer

The study of genomics in cancer started with the observation in 1981 by Shih et al. that DNA sequences obtained from carcinoma cell lines were able to transform phenotypically healthy NIH3T3 cells into cancer cells (Shih and Weinberg 1982). A follow up for the exact causes for this observation started revealing several point-mutations such as a single G > T substitution

mutation in HRAS (Reddy et al. 1982). Such observations motivated the search for more such genes and established the importance of genomic sequences in cancer research. Today, only fifteen years after the first draft of the human genome sequence was published, an exhaustive genomic catalogue of mutations linked to multiple cancer types have been put in place by researchers (Lander 2011). We now have vast information on genome sequences, transcription data, methylation patterns, chromatin structures, etc, related to many types of cancer.

Most of the recent research in cancer genomics has been devoted in the identification of the DNA alterations associated to the tumors from multiple individuals. Patients that have already been diagnosed with a given tumor type according to the multiple clinical tests have their tumors resected and the DNA and/or RNA is extracted and sequenced. Compared to normal cells, tumor cells shows differences in growth, morphology, structure, cellular interactions and gene expression. Somatic alterations such as mutations, copy number variations, translocations, insertions and inversions can cause changes in gene expression, which directly affects protein expression Somatic alterations appear spontaneously and could occur anywhere within the 3 billion bases of the human genome. However, not all are linked to a phenotypic change in the cell. . In cancer, proteins that are crucial for normal cell growth and survival are altered. This is why usually the research is focused on alterations falling on coding regions, which only comprise ~2-3% of entire genome. As described earlier and further in the thesis, studying only coding regions does not provide a comprehensive view of the alterations in tumors. Non-coding regions (regions that do not code for protein) contain many regulatory elements that are essential for the control of gene expression and in particular, RNA processing, hence mutations in them can also have a functional impact in the cell.

2.2.1 HALLMARKS OF CANCER

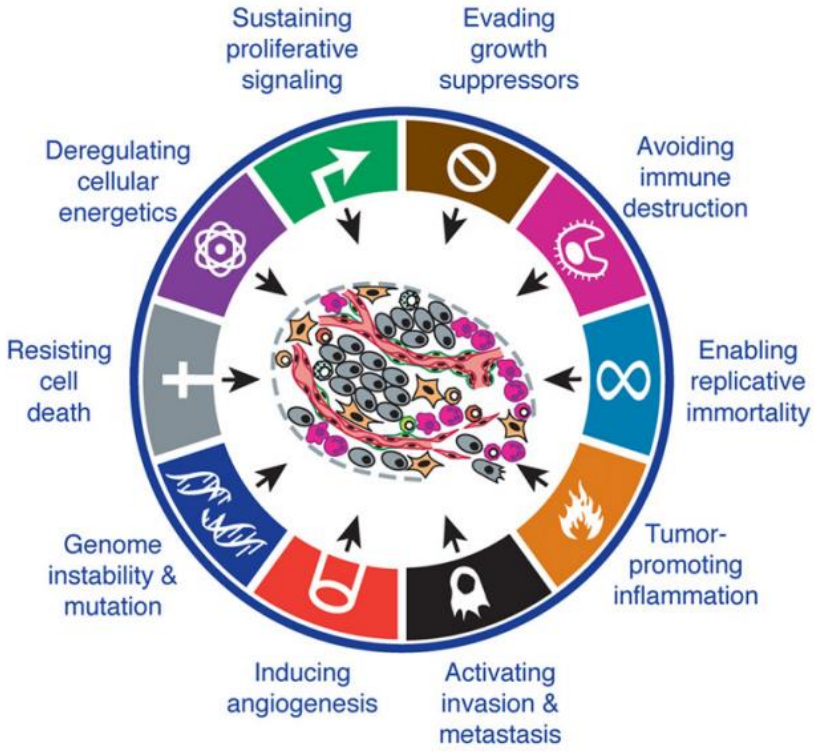


Figure 8. Hallmarks in cancer as proposed by Hanahan and Weinberg, 2011

Tumors are usually divided into benign and malignant. A benign tumor is a mass of cells considered harmless and curable, as they do not spread to other parts of the body. Malignant tumors are the cancerous cells that found a way to multiply uncontrollably and metastasize. Malignant tumors share a minimum set of properties. Hanahan and Weinberg published in the year 2000 the first set of features encompassing the general properties of cancers (Hanahan and Weinberg 2000). They proposed that all cancers 1) sustain proliferative signalling, 2) evade growth suppressors, 3) show resistance to cell death, 4) enable replicative immortality, 5) induce

angiogenesis and 6) activate invasion and metastasis. This provided a common definition for cancer (Hanahan and Weinberg 2000) (Figure 8). However, as research on cancer progressed, along with the availability of more genomics data, it was observed that cancer cells share other consistent properties apart from these six. Following the new developments, Hanahan & Weinberg revisited the cancer hallmarks in another review in 2011 (Hanahan and Weinberg 2011) to include four new hallmarks: 7) avoidance of immune destruction, 8) tumor promoted inflammation, 9) deregulation in cellular metabolism, and 10) genome instability. These ten hallmarks, or a combination of them, are observed in all cancers, and yet the list is not believed to be comprehensive. Several additional hallmarks have been proposed by researchers (Ladomery and Ladomery 2013) in order to address cancer heterogeneity and complexity shared through similar properties.

2.2.2 Alterations in cancer genome

It has been shown through different experiments that cancer follows an evolutionary process at cellular level (Jones et al. 2008). When cells grow abnormally and form a mass outside of the tissue structure, this is known as neoplasm. Neoplasms are classified into four main categories: benign, in-situ, malignant and neoplasms of uncertain behavior. Cancer cells are part of the malignant neoplasms. They are mutant cells competing for space and resources inside a microenvironment. They struggle against elimination by the immune system and disperse throughout the body often to colonize new organs, improving their chances of survival and giving rise to metastasis (Merlo et al. 2006). These neoplasms are

genetically and epigenetically a diverse population of cells altered through somatic mutations and then selected through Darwinian evolution. Somatic mutations providing selective advantage are expected to be maintained in the population and increase their frequency with time, whereas those detrimental to tumor cells are expected to occur at low frequency or disappear from the population. This has motivated the use of recurrence of mutations across different patient samples to identify mutations that may be essential for the survival and growth of the tumor.

Genes that provide advantages to tumor growth are frequently altered to either silence or enhance their function. These genes are divided into two main categories: oncogenes and tumor-suppressors. Genes that participate in cellular-growth pathways, such as *RAS*, are oncogenes and are often associated with gain of function mutations. On the other hand, genes such as *P16*, which function to regulate cell proliferation, act as cell cycle checkpoints against DNA damage, promote apoptosis and DNA repair, etc., are usually tumor suppressor genes. Such genes show loss of function in tumors through point mutations or deletions. Alterations in these genes through mutations, chromosomal rearrangements or gene duplication provides excessive growth promoting signals and uncontrolled division of the cells, hence they are defined as drivers of cancer (Vogelstein et al. 2013).

2.2.3 DRIVERS AND PASSENGER MUTATIONS

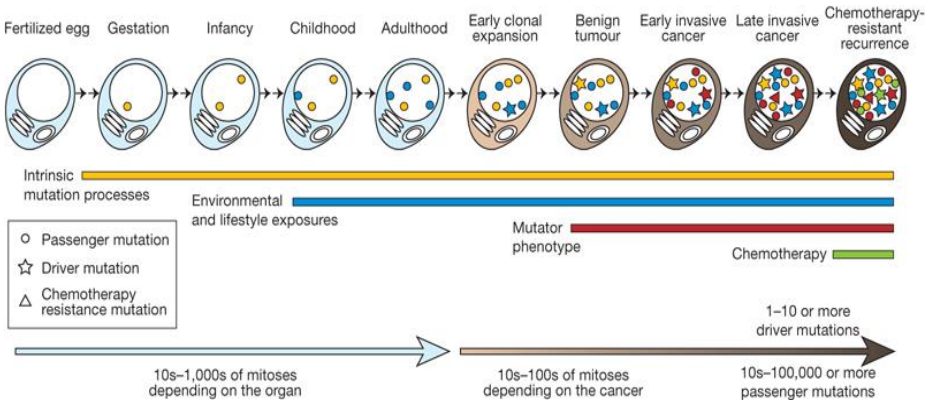


Figure 9. Timing of the somatic mutations acquired by cancer cell as the growth progresses (Stratton et al. 2009).

Cancer is a continuous evolutionary process where cells are 'selected' based on Darwinian evolution. Mutations occur at random on genes and get accumulated, until a specific mutation that gives selective advantage to tumor cells for uncontrolled growth occurs. When this mutation happens it leads to multiple clonal copies in the cell population, often becoming a dominating clone. These mutations will be observed at higher frequency in the DNA sequencing from the bulk tumor sample from the patient. Moreover, if this specific mutation is frequently observed in other patients, it is usually described as a driver mutation. However, as tumors are characterized by increased overall mutation patterns due to genome instabilities and the deregulation of DNA repair (Alexandrov et al. 2013), recurrence is not enough to identify drivers. Besides correction for the mutational background, the functional impact of the mutation is usually evaluated, and hence the analysis has been mostly focused on the protein coding regions of genes (Watson et al.

2013; Martincorena and Campbell 2015). Driver mutations are often observed in oncogenes and tumor suppressors in locations related to their over-activation or inactivation, respectively. Lately, driver mutations were also reported on transcription start sites and other non-coding regions (Huang et al. 2013; Weinhold et al. 2014a). The cells with driver mutations are able to produce more clonal copies and therefore same mutation at same position can be consistently observed across patients. In contrast to driver mutations, passenger mutations are due to the random background pattern of mutations. They are often considered neutral for tumor growth providing no selective advantage, perhaps because they are inconsistent across patients, and therefore difficult to detect (Vogelstein et al. 2013). However, new studies are proposing that passenger mutations hitchhike along with driver mutations and assist driver mutations for specific tumor phenotypes (McFarland et al. 2014).

2.3 Computational method to study cancer mutations

2.3.1 Large Scale Genome Sequencing Studies

Sequencing data to study cancer is generated using either whole exome sequencing (WES) or whole genome sequencing (WGS). In WES, exonic regions, mostly those coding for protein are previously enriched from DNA samples and sequenced. WES is cost effective to get good read depth for ~2% of the genome, and therefore appropriate to study alterations such as protein-coding driver mutations. However, current analyses indicate that we are plateauing in the number of patients that can be explained by protein-affecting

mutations (PAMs) (Garraway and Lander 2013). Despite current efforts, not all regulatory processes have been explored. The reduction of sequencing costs has facilitated the systematic study of other genomic regions in cancer using WGS data (Meienberg et al. 2016). WGS produces full genome sequence, which allows detailed studies for all type of DNA alterations genome wide, including the 98% that corresponds to non-coding regions. The analyses of whole genome sequencing (WGS) from multiple tumors have highlighted the relevance of mutations in regulatory regions, like the TERT promoter or the binding sites for CTCF (Horn et al. 2013; Huang et al. 2013; Katainen et al. 2015) to explain part of the observed tumor phenotypes. WGS data analyses have also highlighted specific mutational signatures linked to tumors and oncogenic mechanisms (Alexandrov et al. 2016; Alexandrov and Stratton 2014; Alexandrov et al. 2013), which may eventually help improving the clinical prognosis and therapy selection for individual tumors. WGS has thus become an essential method nowadays to study cancer data.

The large heterogeneity observed in tumors has continuously raised demands to produce genome-wide sequencing data with high quality cancer cohorts. In fact, sequencing throughput is no longer a problem, but the assembly of well-phenotyped samples is, and this has become the new currency in cancer genomics. Larger cohorts of patients with well-curated clinical data help researchers for in-depth analysis of cancer alterations in relation to prognosis. This has motivated the formation of large international consortia, like ICGC (http://icgc.org/) or TCGA (https://gdc.cancer.gov/) to carry out collaborative efforts to study in depth multiple cancer types using sequencing data from patient samples. TCGA started as a pilot project in the year 2005 in coordination with multiple centers to

provide high quality genomic data for three cancer cohorts: glioblastoma, lung and ovarian cancer. In the second phase, which started in 2009 and ended in 2016, TCGA has provided multi-layer data for 33 different cancer types including whole genome sequencing data for more than 500 patients. The data includes RNA sequencing, CNV profiling, SNP genotyping, DNA methylation profiling, microRNA profiling and whole genome and exome sequencing. For a number of patients, paired normal samples were extracted from the tumor resections and RNA was sequenced as well. TCGA data for 16 different cancer types were used for the work presented in this thesis (Figure 10).

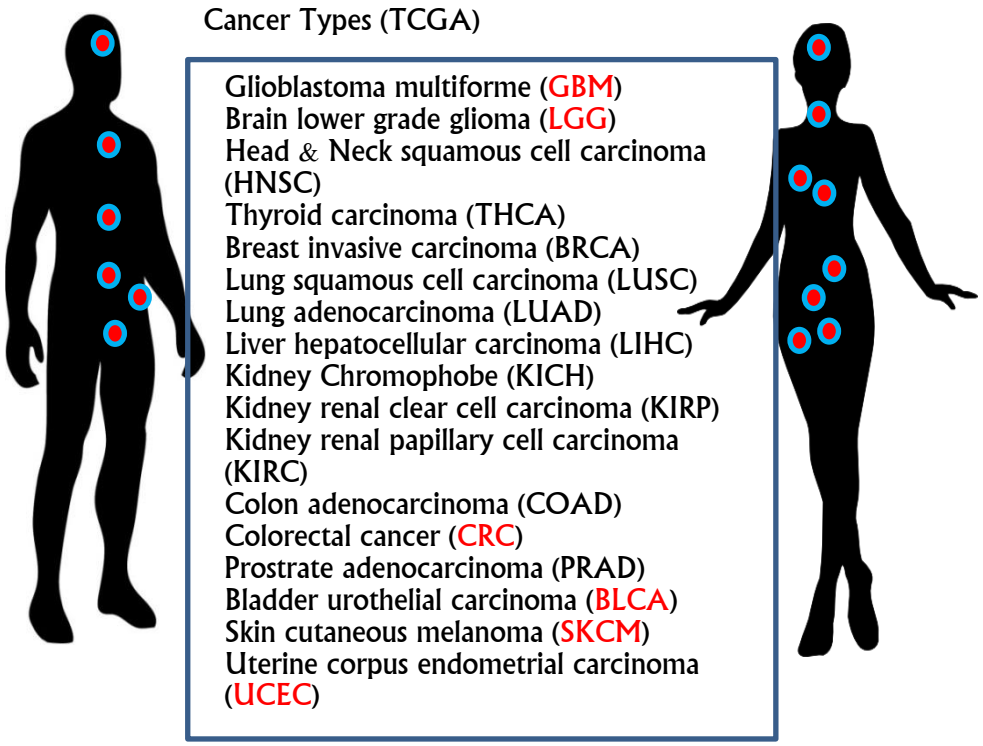


Figure 10. TCGA cancer types used in this thesis. Cancer types in red color were later added to the analysis for the work described in the second chapter.

2.3.2 SOMATIC MUTATION ANALYSIS

After sequence alignment of DNA reads from tumor and paired normal samples, typical pipelines for somatic mutation analysis identify single nucleotide variants (SNVs), copy number variants (CNVs), gene fusions and structural variants. SNVs, including substitutions and indels, are generally detected using tools such as VarScan (Koboldt et al. 2012) and GATK (McKenna et al. 2010). Once identified, these variations are annotated and then studied for either genetic alterations at individual level or at population level (Ding et al. 2014). Significant mutations and driver genes are identified by comparing mutation frequencies against background mutation rates. Besides this, mutation analysis must be corrected for gene length, expression level and replication timing (Lawrence et al. 2013). Gene expression levels correlate with mutation rate, while replication timing shows anti-correlation. Late replicating genes are also found to be prone to more mutations due to unavailability of nucleotides (Lawrence et al. 2013). Tools such as MuSiC (Dees et al. 2012) are utilized to identify significantly mutated genes considering these factors as background. One of the main objectives of mutation analysis is to separate between driver and passenger mutations.

Additional analyses involved the discrimination between oncogenes and tumor suppressors (Schroeder et al. 2014). Oncogenes usually have a clustered pattern of mutations and these are related to activation, whereas tumor suppressors usually have a widespread pattern of mutations along the coding region and these are generally related to protein truncations and/or loss of function. Although

activation of protooncogenes and inactivation of tumor suppressors are associated with mutations, there are also genes with alterations linked to the activation of oncogenic properties or inactivation of tumor-suppressing properties despite not being mutated. These alterations are often related to amplification and overexpression for oncogenes or deletions or downregulation for tumor suppressors (Vogelstein et al. 2013).

3. Splicing and Cancer

3.1.1 Splicing as a hallmark of Cancer

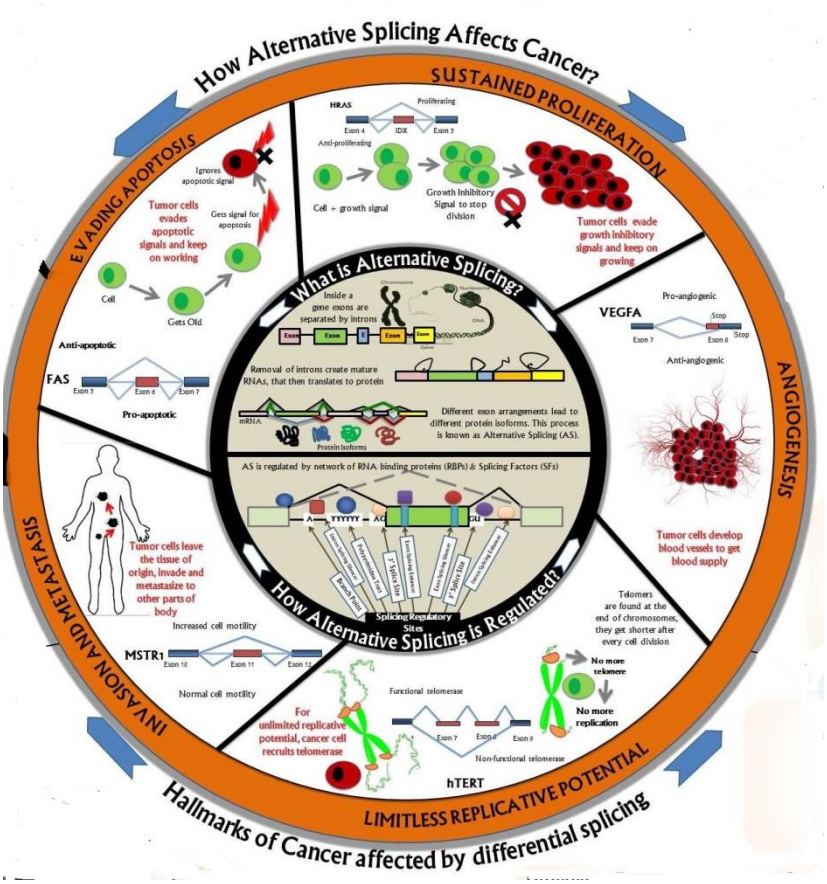


Figure 11. Hallmarks of cancer modified to show splicing association at each step.

Sometimes changes other than mutations or expression alterations of genes can be linked to specific tumorigenic properties. Genes may undergo differential splicing, often producing a switch between a normal protein and a disease specific variant. Differential splicing is relevant for tissue differentiation (Han et al. 2013) and is often

reverted in disease conditions (Irimia et al. 2014; Sebestyén et al. 2016). Differential splicing of genes can disrupt the gene function or protein-protein interactions and ultimately affect downstream pathways. A switch in one or more splicing events has been observed in relation to almost all hallmarks of cancer. For example the protooncogene *HRAS*, important for cellular proliferation, shows mutations on exon 2 that leads to an oncogenic splicing variant (Hartung et al. 2016). Similarly, the alternative splicing of *FAS* produces a soluble form that permits tumor cells to evade the immune system and apoptosis (Cascino et al. 1995; David and Manley 2010). Another proto-oncogene, *MST1R*, may undergo exon 11 skipping, producing a protein variant that provides increased cell motility and migration to tumor cells for invasion and metastasis (Ghigna et al. 2005). Similarly, alternative splicing of *VEGFA* is commonly observed in tumors and it promotes the formation of new blood vessels (David and Manley 2010). Likewise, deregulation of telomerase splicing (hTERT) provides an advantage to tumor cells for limitless replicative potential (Wong et al. 2014). Other examples include an isoform switch in *CTNND1* in relation to cell invasion (Yanagisawa et al. 2008) and the alternative splicing of *BIN1* in relation to apoptosis (Anczuków et al. 2012). In summary, multiple AS changes have been described that essentially recapitulate cancer-associated phenotypes (Figure 11).

3.1.2 ORIGIN OF THE SPLICING ALTERATIONS IN CANCER

Cis and *Trans* alterations

A large body of work has been devoted to determine the different alterations that lead to these AS splicing changes observed in cancer. Trans-acting splicing factors bind to small sequence motifs to promote or repress splicing. Many other RBPs such as CELF proteins, MBNL proteins, QKI, TIA1 and NOVA proteins are also known to regulate splicing in cancer. These proteins show differential expression, somatic mutations and copy number variation in different tumor types, which can have an impact in the RNA processing, and in particular splicing, of multiple genes (Alsafadi et al. 2016; Brooks et al. 2014; Darman et al. 2015; Sebestyén et al. 2016). On the other hand, the cis-component of splicing, i.e. the sequences where these proteins and splicing factors bind are also important to study. Mutations in splicing regulatory sequences are often associated with cryptic splice-site formation and frameshifts, which in turn leads to premature termination of transcripts (Srebrow and Kornblihtt 2006).Somatic mutations have been associated with aberrant splicing of genes in cancer (Jung et al. 2015; Supek et al. 2014). However, these studies focused on splice sites and exonic regions, and show limited evidence on RNA sequencing from the same samples and other more sophisticated analyses are necessary to show the functional impact of splicing in cancer. We describe below the possible alterations in tumors that can lead to alternative splicing changes.

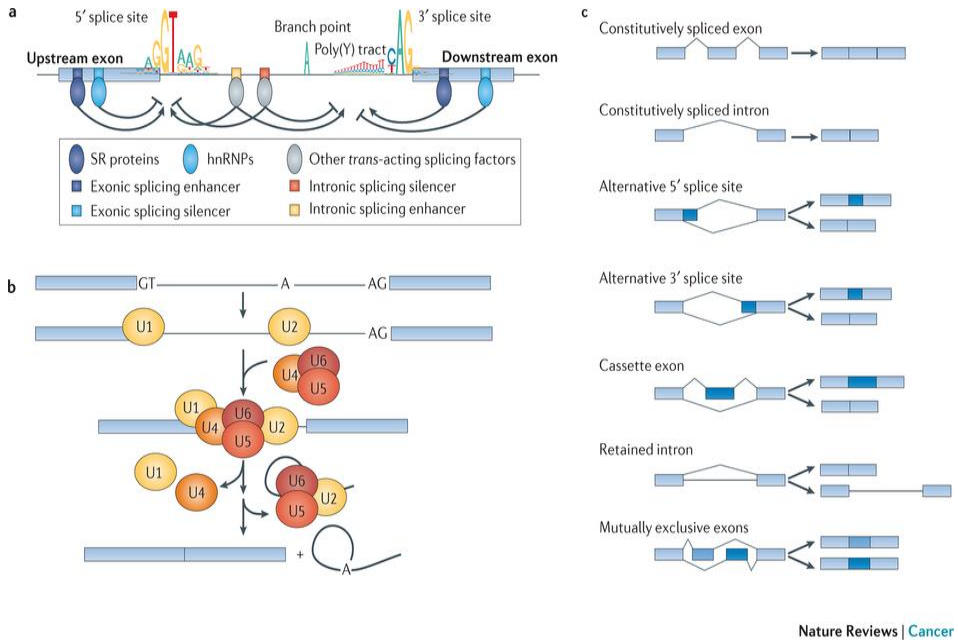


Figure 12. RNA splicing factors as oncogene and tumor suppressors. (Dvinge et al. 2016)

Expression change of splicing factors

Multiple splicing regulatory factors have been observed to trigger tumorigenic properties in cells when overexpressed or downregulated, and have been characterized as oncogenes or tumor-suppressors, respectively, through the changes they induce in alternative splicing (Grosso et al. 2008; Anczuków and Krainer 2016; Dvinge et al. 2016) (Figure 12). Some factors recapitulate this role across multiple tumor types, whereas others show a context dependent expression pattern that may reflect the tissue of origin (Sebestyén et al. 2016). The expression alteration of splicing regulators may have different origins, like copy number alterations (Sebestyén et al. 2016) or through changes in post-transcriptional modifications that are under the control of cell signaling pathways, which are frequently deregulated in tumors. Additionally, splicing factors are transcriptionally controlled by the

oncogene MYC, which is frequently overexpressed in tumors and leads to multiple oncogenic splicing changes through the upregulation of splicing factors (Das et al. 2012; Anczuków and Krainer 2016). The expression changes in splicing factors is also linked to the metabolic transformations associated to tumors, often triggered by specific cellular microenvironments, which leads to alternative splicing chances in genes involved in metabolic processes (David et al. 2010). The link between MYC, splicing and cancer has been further emphasized recently. Components of the spliceosome appear to be essential for the activity of MYC as oncogene, which underscores the central role of splicing in cancer (Hsu et al. 2015; Koh et al. 2015).

It has been further observed that gene expression alterations in cancer appear to recapitulate partially or extensively physiological pathways. For instance, breast tumors show a pattern in the expression splicing factors and splicing events that resemble that of undifferentiated cells, including the down regulation of MBNL1 and a splicing change in NUMB (Sebestyén et al. 2016). Similarly, alternative splicing analysis during metastatic colonization (Lu et al. 2015) show extensive overlap with the changes that occur during epithelial-to-mesenchymal transition (EMT) (Shapiro et al. 2011). However, it is not yet clear whether such cellular programs are fully recapitulated or whether they coexist with other alterations that appear in tumors, thereby providing tumor cells with a variety of molecular repertoires.

Mutations

Access to the genome sequence from multiple tumors has uncovered recurrent mutations in core and auxiliary components of the spliceosome in various tumor types. They occur predominantly in hematological malignancies and often involve the factors SF3B1,

U2AF1, SRSF2 and ZRSR2 (reviewed in (Dvinge et al. 2016)). Although generally at lower rate, splicing factors also appear mutated in solid tumors, including SF3B1 in breast cancer and melanoma (Darman et al. 2015; Furney et al. 2013), U2AF1 and RBM10 in non-small cell lung tumors (Brooks et al. 2014), and HNRNPL in colon tumors (Sebestyén et al. 2016) An analysis of genes coding for known and putative RNA binding proteins has shown that mutations in known and putative regulators of splicing is mostly limited to these cases in solid tumors (Sebestyén et al. 2016). Additionally, expression changes in splicing factors appears to produce more splicing changes in the events, compared to those related with mutations in splicing factors (Sebestyén et al. 2016) or regulatory regions (et al. 2015; Jung et al. 2015), and both types of alterations do not seem to produce the same splicing changes. For instance, modulating the expression of SF3B1 in cells does not recapitulate the changes observed when SF3B1 is mutated (Alsafadi et al. 2016). The identification of the splicing changes related to mutations in splicing factors is instrumental to understand their relevance for cancer development and therapy and is currently an active area of research (Kim et al. 2015; Lee and Abdel-Wahab 2016; Darman et al. 2015; Alsafadi et al. 2016).

Mutations on splicing regulatory sequences

Somatic mutations that disrupt splicing regulatory motifs can also be a source of splicing changes in cancer. For instance, mutations at the exon-intron boundaries have been associated with intron retention in tumor suppressors such as TP53, ARID1A, PTEN, CHD1, MLL2 and PTCH1 (Jung et al. 2015). Similarly, mutations on synonymous sites on coding exons appear enriched in oncogenes and have been proposed to disrupt the splicing of cancer drivers such as ITK, ALK, IDH1 and BCL6 (Supek et al. 2014). Since splicing regulatory

sequences on exons span 4 to 6 nucleotides, hence possibly covering multiple codons, it is likely that mutations on non-synonymous sites also lead to splicing changes in cancer drivers (Sterne-Weiler and Sanford 2014). Intronic mutations also appear to play a crucial role in cancer such as therapy resistance. For instance, a point mutation 51nt upstream of the 3' splice-site of intron 8 of BRAF promotes a splice variant that confers resistance to Vemurafenib treatment (Salton et al. 2015). However, in contrast to exonic mutations, not many recurrent intronic mutations have been described so far beyond the exon-intron boundaries, despite the fact that a significant fraction of the splicing regulation is controlled by intronic regulatory sequences, either through the branch-point and poly-pyrymidine tract sequences, or through intronic splicing enhancers and silencers (Diederichs et al. 2016). This could be due to the fact that intronic regulatory motifs often present positional variability with respect to the exon-intron boundaries and are therefore less straightforward to identify. Although deep intronic mutations may be harder to characterize, they could also affect splicing. For instance, a considerable number of introns harbor distant branch-points located further than 50nt upstream of the 3' splice-site (Corvelo et al. 2010), and the structure of the RNA plays a role in its processing and may bring together distant regions (Lovci et al. 2013). By harnessing the power of characterizing the relevant intronic regulatory regions, we will be able to gain further insights into the disruption of splicing in cancer.

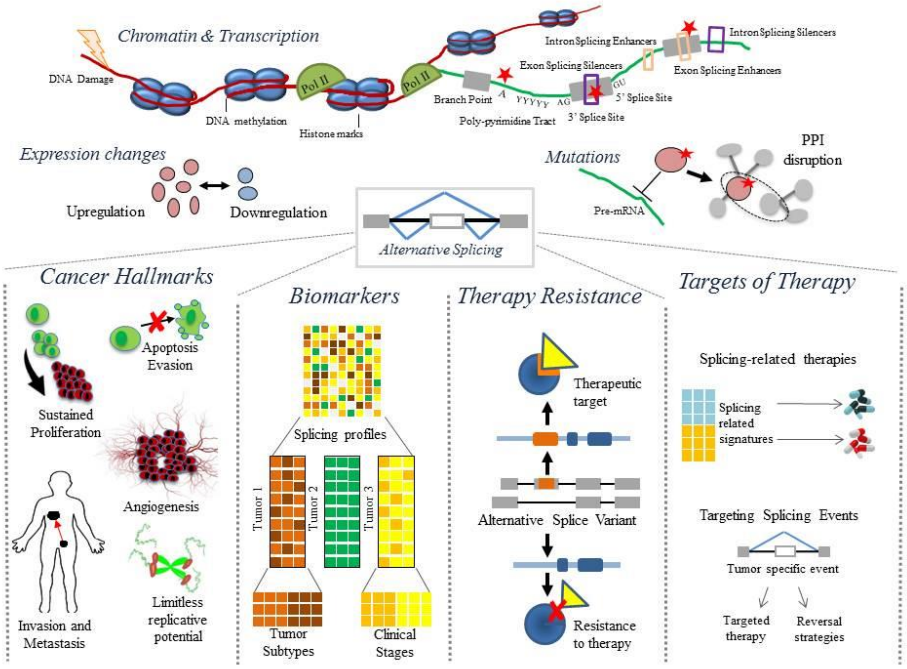


Figure 13. Multiple facets of alternative splicing, providing portal to next generation cancer genomics study and therapeutic targets.

3.2 Functional Impact of Splicing

The analyses of transcriptomes from multiple patient tumor samples have highlighted frequent splicing changes during tumor progression and metastasis transformation (Trincado et al. 2016; Lu et al. 2015) as well as in association to somatic alterations (Darman et al. 2015; Kim et al. 2015; Alsafadi et al. 2016). However, the functional impact of these alternative splicing changes and their significance in cancer is only starting to be elucidated (Figure 13).

3.2.1 ALTERNATIVE SPLICING RECAPITULATES HALLMARKS OF CANCER

Several alternative splicing events have been shown to recapitulate cancer-associated phenotypes. For instance, an exon inclusion change in *NUMB* has been shown to promote cell proliferation (Bechara et al. 2013). Similarly, an exon skipping event in MST1R has been related to the acquisition of cell motility during cancer cell invasion (Ghigna et al. 2005). Moreover, the modulation of these events can recapitulate the tumor phenotype or revert to a normal phenotype (Bechara et al. 2013; Ghigna et al. 2005). Therefore, understanding the general functional effects of alternative splicing potentially leads to the discovery of novel oncogenic mechanisms and therapeutic targets.

Alternative splicing changes have been proposed to remodel the network of protein–protein interactions in a tissue-specific manner (Buljan et al. 2012; Ellis et al. 2012). It is therefore possible that splicing changes in cancer also impact the network of protein-protein interactions, but in a disruptive, non-regulated way. In this direction, a recent study shows that an alternative splicing change in *NFE2L2* that occurs in various tumor types leads to the loss of a protein interaction with its negative regulator KEAP1, thereby providing an alternative way to activate the Nrf2 pathway (Goldstein et al. 2016). This may in fact be a general mechanism whereby splicing alterations disrupt protein-protein interactions of cancer drivers and related pathways, providing other means to impact cell function that are equivalent to classical somatic mutations in drivers. Additionally, alternative splicing may also induce degradation of the transcripts through non-sense mediated decay (Green et al. 2003), a mechanism that was associated to somatic mutations on the splice-sites that induce intron-retention in tumor suppressors(Jung et al. 2015).

3.2.2 Biomarkers

Despite the abundance of splicing changes observed in tumors, only few cases have been characterized for their functional impact. It is possible that the majority of the splicing changes in tumors are passengers, merely reflecting upstream genetic mechanisms and the deregulation of splicing fidelity mechanisms. Yet, they may provide tale-tell signs of specific tumor characteristics. In this context, splicing changes have been shown to separate tumor types and subtypes (Sebestyén et al. 2015) and have been related to tumor stage and patient survival (Shen et al. 2016; Trincado et al. 2016), so they have the potential to be used as biomarkers for specific clinical conditions. This could be relevant for cases for which a known prognostic marker is either not present in the sample or does not exist, as for pediatric tumors (Parsons et al. 2011).

3.2.3 Therapy resistance

Alterations in alternative splicing also appear essential for understanding drug resistance(Lee and Abdel-Wahab 2016). For instance, a considerable proportion of patients that do not respond to targeted treatment against BRAF mutations express a BRAF isoform lacking exons 4–8, which encompass the RAS binding domain (Poulikakos et al. 2011). Interestingly, small-molecule modulators of pre-mRNA splicing are capable of restoring the original BRAF splicing and reduce growth of therapy-resistant cells(Salton et al. 2015). Similarly, alternative splicing also impacts immunotherapy in leukemia due to the disrupted activity of the splicing factor SRSF3 (Sotillo et al. 2015). These results highlight the importance of characterizing the

transcriptome for therapy and suggest that specific splicing alterations may provide a selective advantage to tumors.

3.2.4 Targeted therapy

There is a growing interest to search for splicing-related alterations for which specific therapies could be developed. One of the strategies being tested at the moment consists in the synthetic design of antisense oligonucleotides (AONs) that target specific splicing events. AONs are able to revert alternative splicing events to restore normal cellular phenotypes (Bechara et al. 2013; Ghigna et al. 2005), and have reached already clinical trial stage for some splicing-related disorders(Havens and Hastings 2016). Another promising strategy for cancer therapeutics is the use of small molecule compounds that modulate the activity of splicing factors (Salton and Misteli 2016; Lee and Abdel-Wahab 2016). These therapies have a wide range of effects depending on the tumor type or the mutational status of the targeted splicing factor. Thus, it becomes essential to know which patients may benefit from splicing-related therapies. One such possible class includes patients with overexpressed MYC in tumors, which are more dependent on the activity of the spliceosome (Koh et al. 2015; Hsu et al. 2015).

Alternative splicing events are also emerging as direct actionable alterations for targeted therapies. This is the case of the skipping of MET exon 14 observed in some lung cancer patients, resulting in a deletion of the protein region that inhibits its kinase catalytic activity (Kong-Beltran et al. 2006). Importantly, the skipping of this exon is sufficient for MET activation and tumors that harbor the event respond to MET-targeted therapies (Frampton et al. 2015; Paik et al. 2015). Although this splicing change in MET has been explained so far as a

result of somatic mutations on exon 14 or on its splice-sites, it is conceivable that the same splicing change could occur due to other mechanisms yet to be discovered. These results raise the interesting possibility that an alternative splicing event could be used as direct target of therapy. Thus, either as direct targets or as a means to characterize the tumor, the splicing properties may become fundamental to identify therapeutic vulnerabilities and potential resistance. This may be particularly relevant for tumors lacking somatic mutations in genes with known targeted therapy, as these patients cannot benefit from currently available therapies.

3.2.5 Future towards personalized medicine & Role of splicing

Alternative splicing changes that characterize and contribute to the pathophysiology of cancer are triggered by alterations in a complex network of different mechanisms. These combinatorial effects have some interesting implications. Different alterations in tumors may in turn impact RNA processing and splicing in similar ways. For instance, mutations in RBM10 or downregulation of QKI lead to the same splicing change in NUMB that promotes cell proliferation (Zong et al. 2014; Bechara et al. 2013). This suggests that the splicing alterations observed in tumors may be indicative of a phenotypic advantage, and some may even phenocopy somatic mutations in cancer drivers to induce similar functional impacts. Accordingly, a subset of the splicing changes in cancer may play an important role in the neoplastic process independently of or in conjunction with the already characterized genetic alterations.

It is not clear yet whether a single splicing change may be sufficient to induce an oncogenic transformation in a normal tissue context, or even

whether splicing events can be considered cancer drivers. It is possible that the splicing-related effects are additive, contributing to and maintaining specific properties or favoring certain cellular environments that modulate the oncogenic impact of somatic mutations. Consistent with this, there is a relation between specific tumor microenvironments and alternative splicing (Brosseau et al. 2014). Additionally, somatic mutations in splicing factors are generally heterozygous, and appear to require a normal functional splicing machinery to exert their oncogenic function (Fei et al. 2016; Lee and Abdel-Wahab 2016). For example, the ratio of both mutant and wild-type U2AF1 splicing factor influences the splice-site selection in lung adenocarcinomas, questioning the functional significance of the mutant U2AF1 cells(Fei et al. 2016). This suggests a context-dependent effect, by which somatic alterations may become relevant in the presence of certain splicing-related signatures. This is further supported by recent findings showing that tumors with overexpressed MYC are highly dependent on the splicing machinery for survival and may be more sensitive to splicing-related therapies (Hsu et al. 2015; Koh et al. 2015).

In conclusion, as selection on the tumor clones is exerted on the phenotype rather than on the genotype, we propose that the splicing patterns may define relevant molecular phenotypes in tumors, despite their genetic heterogeneity. The characterization of tumor transcriptomes - with respect to splicing - thus becomes essential to understand their clinical properties and to select appropriate therapeutic strategies.

II. Objectives

This thesis is mainly divided into two parts to fulfill following objectives:

1 Regulation of splicing through *Trans* Components

The analysis of splicing changes associated to the mutation and expression alterations in the RBP genes in 11 different cancer types.

2. Regulation of splicing through *cis* Components

An exhaustive search for mutations on RBP motifs inside genes using whole genome sequencing data to study the impact of these mutations on mRNA expression and splicing.

III. Results

Results: Objective 1

Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. 2016;26(6):732–44. DOI: 10.1101/gr.199935.115

Keywords: alternative splicing, RNA binding proteins, splicing networks, cancer

Results: Objective 2

Novel mutational patterns that impact RNA-processing in human tumors

Babita Singh[1], Juan L. Trincado[1], P.J. Tatlow[2], Stephen R. Piccolo[2,3], Eduardo Eyras[1,4,*]

[1]Pompeu Fabra University (UPF), E08003 Barcelona, Spain.
[2]Brigham Young University, Provo, Utah, USA
[3]University of Utah, Salt Lake City, Utah, USA
[4]Catalan Institution for Research and Advanced Studies (ICREA). E08010 Barcelona, Spain

*correspondence to: eduardo.eyras@upf.edu

Abstract

A major challenge in cancer research is to determine the significance of nucleotide variants in regions that do not code for protein. To address this question, we have performed a comprehensive study of mutations along genes from whole genome sequencing data for more than 1000 tumor samples to identify significantly mutated regions (SMRs). Systematic sequence analysis reveals recurrent patterns of mutations in motifs associated to specific RNA binding proteins in introns as well as 5' and 3' untraslated regions (UTRs) in protein coding genes, and in exons and introns of non-coding RNAs. Analysis of RNA sequencing from the same samples identifies alterations in

RNA-processing and expression associated to these mutations, revealing novel oncogenic mechanisms. One of these alterations is a recurrent mutation in a in CT-rich motif in the 5'UTR for multiple genes, including the cancer gene driver SPOP, which shows a significant alteration in RNA processing. Our study describes the first genome-wide map of somatic mutations that potentially impact protein-RNA interactions in cancer and their impact on RNA processing. Furthermore, our integrative analysis of analogous regions enriched with mutations allows the interpretation of non-coding variants in tumors.

Introduction

Cancer arises from genetic and epigenetic alterations that interfere with essential mechanisms of the normal life cycle of cells such as DNA repair, replication control and cell death (Hanahan and Weinberg 2011). The search for cancer driver mutations, which confer a selective advantage to cancer cells, has been traditionally studied in terms of how they directly affect protein sequences (Vogelstein et al. 2013). However, systematic studies of cancer genomes have highlighted a large number of mutations and mutational processes across the tumor genomes (Alexandrov et al. 2013; Weinhold et al. 2014). Moreover, specific non-coding mutations have been identified as tumorigenic, like those found in the TERT promoter (Horn et al. 2013; Huang et al. 2013) or at CTCF binding sites (Katainen et al. 2015). Currently, a major challenge in cancer genomics research is to determine the significance and potential pathogenic involvement of somatic variants in regions that do not code for proteins (Piraino and Furney 2016).

Current methods to detect potential driver mutations in non-coding regions have been mostly based in the recurrence of mutations in

specific regulatory regions in combination with the measurement of the potential functional impacts (Melton et al. 2015; Fredriksson et al. 2014; Mularoni et al. 2016; Weinhold et al. 2014), recurrence combined with sequence conservation or polymorphism data (Khurana et al. 2013; Piraino and Furney 2017) or using the enrichment with respect to specific mutational backgrounds (Lochovsky et al. 2015; Lanzós et al. 2017). Most of the methods have been restricted to specific genomic regions, like potential regulatory regions, and only few of them have measured the impact of mutations on the RNA and this has been mainly focused on overall gene expression. On the other hand, deep-intronic mutations or mutations that affect RNA processing have not been thoroughly studied.

Transcribing and mature RNA molecules are bound by multiple RNA binding proteins (RBPs), which have specific roles at different steps during RNA processing, including RNA translation, RNA stability and RNA localization, and are critical for the proper control of gene expression (Fu and Ares 2014; Rissland 2017). RBPs can act as auxiliary and sometimes necessary factors to regulate splicing and RNA processing and often antagonize each other in normal cellular programs and disease states (Eperon et al. 2000; Zhu et al. 2001; David et al. 2010; Bonomi et al. 2013). Importantly, the same RBP may participate in a wide-rage of RNA processing activities besides splicing regulation. For instance, SR proteins can control splicing, mRNA nuclear export, nonsense-mediated mRNA decay and mRNA translation (Long and Caceres 2009; Maslon et al. 2014). Recent high throughput studies have uncovered many new proteins with potential RNA binding capabilities (Baltz et al. 2012; Castello et al. 2012; Conrad et al. 2016), highlighting the relevance of RBPs in gene expression and providing perhaps the largest potential for dynamic gene regulation. Specific RBPs and RBP families are linked to specific

functions and cellular pathways. They control the processing of the RNA from multiple gene loci, and therefore can have major influence in specific cellular mechanisms. For instance, SRSF10 controls the alternative splicing of multiple genes involved in DNA damage repair (Shkreta et al. 2016), and proteins from the RBFOX family are involved in neuronal differentiation program (Kim et al. 2013), neuronal function (Gehman et al. 2011) and the maintenance of a mesenchymal splicing program in normal and tumor cells (Venables et al. 2013a). Similarly, the ESRP proteins maintain an epithelial phenotype and their downregulation induce alternative splicing changes that trigger cell motility (Warzecha et al. 2010), and MBNL proteins maintain a differentiated cellular state that is reversed in some tumor tissues (Han et al. 2013; Sebestyén et al. 2016).

Multiple experimental approaches have established that RBPs generally interact with RNAs through short motifs of 4-7 nucleotides (Ule et al. 2003; Lambert et al. 2014; Ray et al. 2013; Oberstrass et al. 2005). These motifs may occur anywhere along the pre-mRNA, including introns and exons coding regions (CDS) and untraslated 5' and 3' regions (5UTR/3UTR), as well as in short and long non-coding RNAs (Sterne-Weiler and Sanford 2014; Haerty and Ponting 2015; Michlewski et al. 2008). Mutations in RNA regulatory motifs, including binding motifs for RBPs, have been linked before to RNA processing alterations in disease (Cartegni et al. 2002; Anczuków et al. 2015). The phenotypes triggered by the alterations in RNA-processing regulators involve changes in multiple genes, but these do not occur in the same way in all cancer patients (Brooks et al. 2014; Sebestyén et al. 2016; Alsafadi et al. 2016; Darman et al. 2015). Since RBP binding motifs are wide-spread along gene loci, and somatic mutations may occur anywhere along the genome, it is possible that somatic mutations in non-coding, as well as coding regions, will impact the processing of

RNA to produce a phenotypic impact in tumor cells similar to that produced by other alterations.

Studies carried out so far on mutations affecting RNA processing and alternative splicing have mainly focused on a fraction of the motifs associated with the core splicing machinery (Jung et al. 2015) or in protein coding regions only (Supek et al. 2014). For instance, mutations on the exon-intron boundaries have been associated with intron retention in tumor suppressors such as *TP53, ARID1A, PTEN, CHD1, MLL2* and *PTCH1* (Jung et al. 2015). Similarly, mutations on synonymous sites on coding exons appear enriched in oncogenes and have been proposed to disrupt the splicing of cancer drivers such as *ITK, ALK, IDH1 and BCL6* (Supek et al. 2014). Exonic splicing regulatory sequences span multiple nucleotides and cover multiple codons, hence it is likely that mutations that change the amino-acid can also affect some regulatory sequences and induce splicing changes in cancer (Sterne-Weiler and Sanford 2014). In this direction, the systematic analysis of sequence variants on an exon has revealed that more than 50% of nucleotide substitutions can induce splicing changes (Ke et al. 2011; Julien et al. 2016), with similar effects on synonymous and non-synonymous sites (Julien et al. 2016). Accordingly, a large proportion of the somatic mutations on gene loci can be expected to impair RNA processing, which in turn will have a functional impact. In that respect, characterizing tumor transcriptomes is extremely relevant to identify and characterize the functional impact of somatic alterations (Singh and Eyras 2016). Tumor alterations that impact RNA processing and, in particular, alternative splicing, have raised much recent interest since they uncover novel oncogenic mechanisms and open up new therapeutic opportunities (Lee and Abdel-Wahab 2016). However, genome-wide studies on mutations

affecting all possible sequence elements bound by RBPs, including intronic ones, and measuring the impact on RNA, are lacking.

To understand the alterations of RNA processing in cancer at a global level, we have carried out a comprehensive study of the mutation patterns in exons and introns from coding and non-coding genes using mutation data from whole-genome sequencing for more than 1000 samples from multiple tumor tissues. In our approach, we detect significantly mutated regions (SRMs) along genes taking into account regional variability and nucleotide biases. To increase the power to describe novel oncogenic mechanisms involving non-coding regions, we studied the enrichment of known and putative motifs for RNA binding proteins with respect to control regions. This novel approach allows us to related SMRs with each other by the presence of recurrent sequence motifs, inferring potentially common regulatory roles. Additionally, unlike previous methods, we do measure the impact on RNA processing using RNA sequencing from the same samples testing changes in transcript expression, alternative splicing and aberrant splicing. This unprecedented study reveals a new layer of somatic alterations in cancer that may be relevant to explain transcriptome alterations in cancer. Our approach provides a way to evaluate the relevance of many non-coding variants of unknown function and gives a new interpretation to some of the variants in coding regions.

Results

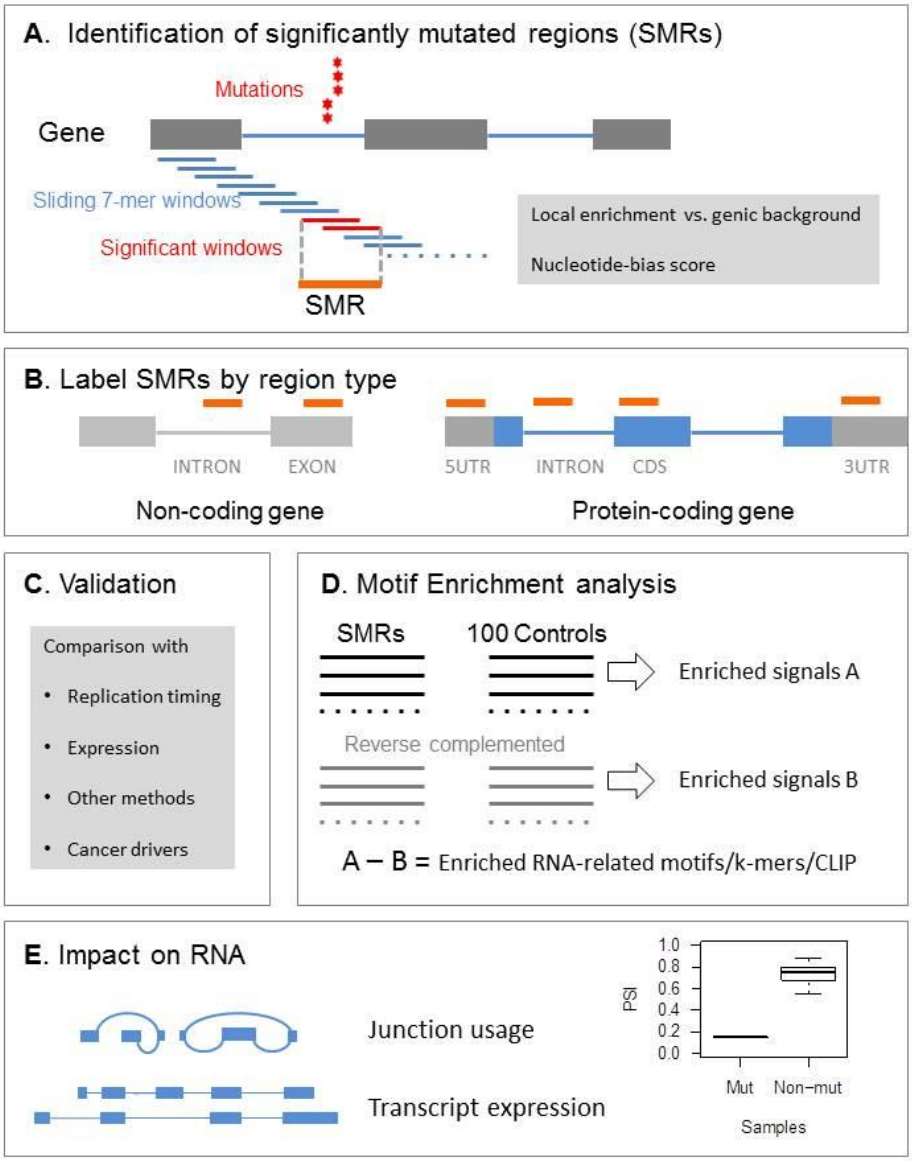


Figure 1. (A) Illustration of the calculation of significantly mutated regions (SMRs). Short k-mer windows (k=7 in our study) along genes are tested for the enrichment in mutations with respect to the gene mutation rate and with respect to the local nucleotide biases (Methods). Significant windows are clustered by region type, producing the SMRs.

Unbiased search of significantly mutated regions (SMRs) along gene loci

RNA binding proteins (RBPs) generally interact with pre-mRNAs through short motifs of 4-7 nucleotides, which may occur anywhere along the pre-mRNA. We thus performed an exhaustive search for mutation enrichment using overlapping genomic windows of length 7 along each gene locus to determine the patterns of somatic mutations in potential RBP motifs, (Fig. 1) (Methods). We considered two datasets of somatic mutations from whole-genome sequencing (WGS) in multiple tumor types: 505 samples from 14 tumor types (Fredriksson et al. 2014) (PAN505) (Supp. Table S1) and 507 samples from 10 tumor types (Alexandrov et al. 2013) (PAN507) (Supp. Table S2), which were analyzed independently. We used a double test to account for the local variations and nucleotide biases in mutation rates to assess the possible enrichment of mutations at every 7-mer window along each gene locus. To account for the local variations in mutational processes, each 7-mer window was tested for enrichment by comparing with the mutation rate in the same gene locus (Methods). The majority of 7-mer windows follow an expected uniform distribution (Supp. Fig 1A) and we selected those windows with p-value < 0.05 after correcting for multiple testing.

To account for the nucleotide biases we performed an additional test by comparing the mutation count in the 7-mer windows with the expected mutation count calculated from the window nucleotide sequence and the mutation rate per nucleotide of the gene locus. This provided a nucleotide bias (NB) score per window defined as the log2-likelihood of the observed versus the expected counts (Methods). To determine the cut-off for the NB score, we further compared the distribution of scores in windows with 3 or more mutations with the

distribution for windows with 1 mutation, which were considered to be more likely to reflect the background mutations. This comparison shows that NB-score = 6 provides a good separation between windows with 3 or more mutations and windows with 1 mutation (Supp. Fig. 1B). After these filters (corrected p-value < 0.05 and NB score > 6), our exhaustive analysis produced a total of 74771 and 8893 significant 7-mer window falling in 8159 and 1247 genes for the PAN505 and PAN507 datasets, respectively, indicating possible mutational hotspots.

The functional impact of somatic mutations as well as the selection processes they are subjected to may vary depending which genic region they fall in. We thus classified 7-mer windows according to whether they fall in a 5' or 3' untranslated region (5UTR/3UTR), a coding sequence (CDS), an exon in a long non-coding RNA (EXON), or in an intron (INTRON) (Fig. 1) (Methods). These windows were then clustered into significantly mutated regions (SMRs), producing a total of 18458 SMRs for PAN505, containing a total of 73728 substitutions; and 1609 SMRs for PAN507 containing 5247 substitutions (Supp. Figs. 1C and 1D). The discrepancy between the two sets is due to the different starting number of mutations (Supp. Tables S1 and S2). The majority of SMRs are in introns (Table 1). We also see a large representation on the exons in non-coding RNAs (EXON) (Table 1). Most of the predicted SMRs are between 7 and 15 nucleotide long (Supp. Fig. 2).

Table1. Number of significantly mutated regions (SMRs) per cohort and per region type.

SMR	PAN505	PAN507
CDS	225	21
5UTR	120	9
3UTR	298	8
INTRON	17474	1544
EXON	341	24

There is a known correlation between observed somatic mutations in cancers and DNA replication timing that can be a source of artifacts in mutational driver predictions (Lawrence et al. 2013; Liu et al. 2013). To validate our SMRs, we thus calculated the replication timing in the regions where SMRs fall and observed no association with mutation count (Supp. Fig. 3). Another potential source of artifacts is the known correlation between gene expression and the rate of somatic mutations in cancer (Lawrence et al. 2013). Using RNA-seq from the same samples in the PAN505 cohort and measuring the expression of transcripts containing the SMRs (Methods), we observed no association of the mutation count in SMRs with expression (Supp. Fig. 4).

To further validate our SMRs we used the tool LARVA (Lochovsky et al. 2015) to assess their significance using a model that accounts for the over dispersion of the mutation rate and the replication timing (Methods). We observed overall a good correspondence between the

significance provided by our method and the significance given by LARVA (Supp. Fig. 5). In particular, we found a good agreement for intronic SMRs, providing support for our intronic SMRs.

Prediction of novel and known significant SMRs in coding and non-coding regions

For the PAN505 dataset, our method found SMRs in 415 cancer drivers from a set of 889 collected from the literature (Sebestyén et al. 2016). We found 38 in CDS SMRs, including the drivers *BRAF, KRAS*, *NRAS*, *HRAS, TP53, CTNNB1, PIK3CA, PIK3R1, IDH1 and SF3B1* (Fig. 2) (Supp. Table S3). A recent approach based on the measure of the functional impact also recovered significantly mutated CDS regions in *BRAF*, *IDH1*, *KRAS*, *PIK3CA* and *PIK3R1* using the same data (Mularoni et al. 2016). Significantly mutated coding regions in *SF3B1*, *CTNNB1*, *TP53* and *KRAS* were also recovered before using a genome-wide search based on mutation enrichment and evolutionary conservation (Piraino and Furney 2017). In total, we found CDS SMRs in 13 of the 41 genes identified before with the PAN505 data (Mularoni et al. 2016). The discrepancy could stem from the fact that we predicted SMRs only using all samples from PAN505. Our method also found CDS SMRs in drivers that were not found by previous methods, including *NRAS*, *EP300* and *ATM.*

Additionally, we recovered 5UTR SMRs in 17 of the 44 different genes identified previously (Mularoni et al. 2016), including *C16orf59*, *TAF11* and *TBC1D12.* Similarly, we recover 3UTR SMRs in 3 of the 12 different genes identified before (Mularoni et al. 2016), including one 3UTR SMR in *CYP4F31P* (Fig. 2) (Supp. Table S3). We also found novel cases, like a 5UTR SMR in the DEAH-box helicase *DHX16* and a 3UTR SMR in the TP53 inducible gene 3 *TP53TG3D* (Supp. Fig. 6).

In total, we found 8 5UTR SMRs and 40 3UTR SMRs in cancer drivers. Among those, we found a novel 5UTR SMR in the cancer driver *SPOP.*

We also found many SMRs in the exons of non-coding transcripts (EXON SMRs). For instance, we found an SMR in an exon of an annotated non-coding transcript in the DEAD-Box Helicase 17 gene *DDX17* with 10 substitutions, and an SMR in an exon at the 3' end of the long non-coding RNA *CTD-3148I10.15* (Supp. Fig. 6), both of which have not been described before by any other method. Comparing with a list of 46 lncRNAs related to cancer (Lanzós et al. 2017), we found EXON SMRs only in one of them, *TCL6*. On the other hand, for 11 of these 46 lncRNAs we found INTRON SMRs.

All genic regions that could not be matched to an exon (CDS, UTR or EXON) were classified as intronic (Fig. 2). We found INTRON SMRs in 7 of the 13 genes reported in (Mularoni et al. 2016), including the cancer gene drivers *TP53*, *NF1*. As our analysis is exhaustive along gene loci and all positions along entire introns were tested, we recovered many more intronic SMRs than in previous reports. In particular, we found INTRON SMRs in 381 different cancer driver genes, including the Tyrosine-Protein Kinase Receptor *EPHB1,* with 14 substitutions (Supp. Fig. 6)

From the PAN507 dataset we found SMRs in 91 cancer drivers, 9 in CDS, including *FGFR1*, *TP53* and *KRAS*, 2 in 5UTR, in *BCL2* and *TBC1D12*, one 3UTR SMR in the cancer driver *NBPF1*. Finally, we found 82 INTRON SMRs in cancer drivers, including one in *EPHB1.* We found only 26 overlapping SMRs between PAN505 and PAN507. One of them is the 5UTR SMR in *TBC1D12*, already described before

(Mularoni et al. 2016). Moreover, there were in common 5 CDS SMRs, including *TP53* and *KRAS.* Finally, there were 15 intronic SMRs overlapping between the two sets PAN505 and PAN507. One of them corresponds to 3 mutations deep in the intron of the *RAD51* Paralog B gene. *RAD51B is* involved in DNA repair, which have not been described before. The low overlap suggests very different mutational processes and functional alterations between the two tested cohorts.

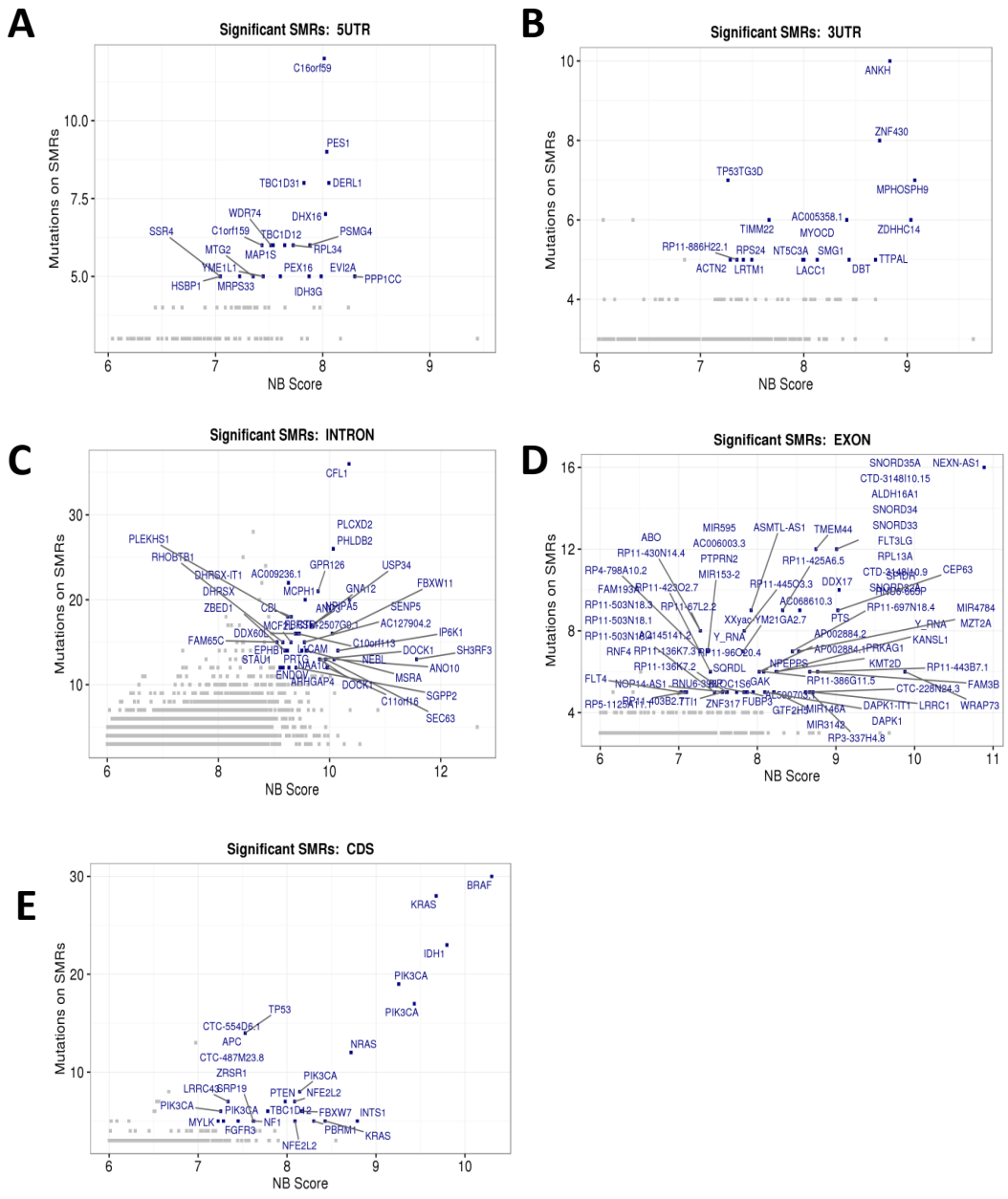


Figure 2. Significantly mutated regions (SMRs). We show the SMRs directed in 5' UTRs **(A),** 3'UTRs **(B),** introns **(C),** exons of non-coding RNAs (EXON) **(D)** and coding sequences (CDS) **(E).** We show the gene name for the SMRs with NB-score 8 or greater and with 5 or more mutations, except for the INTRON SMRs, where we highlight the cases with 14 or more mutations.

SMRs are enriched for putative RBP binding motifs

Mutations and expression alterations in RBPs have an impact in specific cellular programs in cancer (Brooks et al. 2014; Sebestyén et al. 2016; Alsafadi et al. 2016; Darman et al. 2015). However, little is known about whether mutations in the binding sites for RBPs are frequent in cancer and could be related to specific oncogenic mechanisms. To determine whether the identified SMRs are related to a general disruption of specific RNA processing mechanisms, we studied the sequence motifs and biochemical signals related to RBP binding present in the SMRs. First, using sequence motifs for 331 different RBPs (Methods), we tested RBP-motif enrichment in the SMRs compared with control sets of the same region type, adjusting for length and G+C content (Methods). To obtain enriched motifs that are likely associated to RNA rather than DNA, we reverse-complemented all SMRs and controls and repeated the enrichment analysis (Methods). All those motifs that appeared enriched in both calculations for the same region-type were then eliminated. This yielded a total of 10, 9, 5, 4 and 10 enriched RBP-motifs (z-score > 1.96, motif count > 5) in CDS, 5UTR, 3UTR, INTRON and EXON SMRs, respectively for the PAN505 cohort (Supp. Table S4). A total of 3299 SMRs (81 CDS, 85 5UTR, 45 3UTR, 2956 INTRON and 132 EXON) harbor at least one of the enriched motifs.

Among the enriched motifs, we found QKI, RBMS1 and PCBP2 motifs in CDS SMRs (Fig. 3). QKI motif mutations only appear in 2 CDS SMRs in the genes *APC* and *ATM* with mutations from colorectal cancer (CRC). In 5UTRs we found motifs for PTBP1, SRSF11 and RBM5 (Fig. 3). PTBP1 motif mutations in 5UTR SMRs are

predominant in melanoma (SKCM). Interestingly, many mRNAs contain the so-called 5' terminal oligo-pyrimidine tract (5'TOP) motif that is relevant for translational regulation (Meyuhas 2000; Pichon et al. 2012) and PTBP1 may bind these TOP motifs to regulate translation (Sawicka et al. 2008; Pichon et al. 2012). We found PTBP1 motifs mutated in 59 5UTR SMRs, including one in the cancer gene driver *SPOP*, which was found to be frequently mutated in prostate tumors (Cancer Genome Atlas Research Network 2015). Here we found a 5UTR SMR in *SPOP* with 3 mutations in uterine carcinoma (UCEC), suggesting new *SPOP* alterations with relevance in other tumors (Supp. Fig. 6).

In INTRON SMRs, we found a large overrepresentation of binding motifs of the PABPC5, ESRP2 and the ring finger protein ZFP36 (Fig. 3). PABPC5 binds to the poly-A stretches in mRNAs in the cytoplasm. The enrichment found could indicate an alternative function related to the processing of pre-mRNAs. Mutations in ZFP36 motifs are the most abundant in introns and appear widespread across all tumor types. ZFP36 binds to AU-rich motifs in mRNAs to promote their degradation (Lai et al. 2002) and plays a key role in the post-transcriptional regulation of the tumor necrosis factor TNF (Resch et al. 2014). Here we found a mutated ZFP36 motif enriched in mutations in the intron of a member of the TNF Receptor Superfamily *TNFRSF11A*. It is possible that intronic ZFP36 motifs enhance the degradation of intronic RNAs, and the frequent mutation on these sites leads to the aberrant expression of intronic sequences and unprocessed transcripts as observed before in multiple tumors (Dvinge et al. 2015). SMRs in exons of non-coding RNAs are enriched in TIA1, RC3H1, and CELF2 motifs, among others. Mutations in RC3H1, which appear in EXON, INTRON and 3UTR SMRs, are quite abundant and present in multiple

tumor types. TIA1 and CELF2 motifs in EXON SMRs are also frequent across all tumor types.

For all the SMRs harboring any of the enriched motifs, we calculated averaged phastCons scores from multiple alignments of primates and mammals (Methods). For each of the enriched motifs, we observed some of the motif-containing SMRs that are highly conserved (Supp. Fig. 7A). In the CDS SMRs we observe a subset highly conserved. A fraction of the PTBP1 motifs in 5UTR SMRs show high conservation indicating that these sites are functional. Some of the intronic SMRs with ESRP2 and ZFP36 motifs also show high conservation. In general, EXON and INTRON SMRs show lower conservation. Interestingly, a number of EXON SMRs with TIA1 motifs show high conservation, indicating a potential relevant role of TIA1-like motifs on non-coding RNAs. We further performed an unbiased k-mer enrichment analysis, using k=7, in the SMRs. As before, we tested the enrichment of k-mers compared to controls of the same type and discarded those cases that appeared also enriched after reverse complementing SMRs and controls (Supp. Table S5) (Methods). This k-mer enrichment analysis recovered motifs previously found and uncovered new ones (Supp. Fig. 7B). For instance, in 5UTR regions we found multiple CT-rich k-mers, highlighting other possible mutated TOP-sites. In EXON SMRs we found T-rich motifs, further supporting the enrichment of motifs for TIA1, and suggesting other T-rich motifs. Interestingly, in CDS SMRs we found GA-rich motifs (Supp. Fig. 7B). Furthermore, a subset of found enriched k-mers show high conservation across species (Supp. Fig. 7C).

To gain further evidence of frequent mutations in specific RBP binding sites we performed an enrichment analysis of binding sites from 91

cross-linking immunoprecipitation (CLIP) experiments for 68 different RBPs (Supp. Table S6) (Methods). As before, we repeated the enrichment analysis after reversing the strand of all SMRs and controls to ensure that the significant associations are related to RNA. Most of the enrichments found were due to a small number of overlaps between SMRs and CLIP regions (between 2 and 5) (Supp. Table S7), except for IGF2BP1 and TARDBP CLIP and intronic SMRs, which had 8 and 42 overlaps, respectively. Although we had motifs for both proteins in our analysis above, these did not appear enriched, indicating a discrepancy between the motifs used and the CLIP signals. The CLIP enrichment in INTRON SMRs suggests a possible nuclear regulatory role of IGF2BP1 and TARDBP that is potentially impaired in cancer.

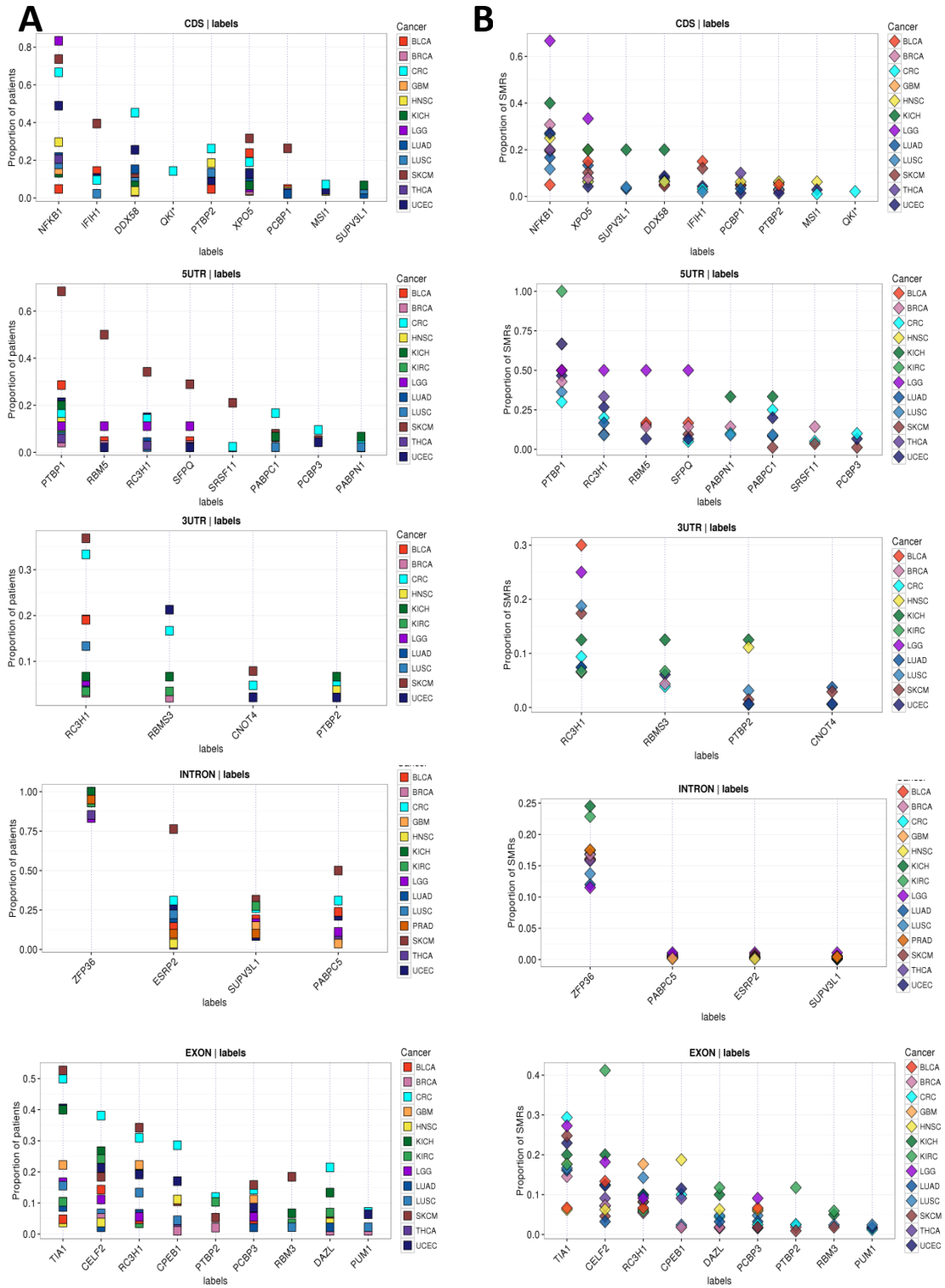


Figure 3. Enriched RBP motifs in SMRs. For each of the region types (CDS, 5UTR, 3UTR, INTRON, EXON), and for each RBP motif (x axes) found enriched in that region, we give the proportion of patients that have at least one SMR containing the motif **(A)**, as well as the proportion of SMRs that contain the RBP motif **(B)**. These proportions are divided by tumor type, which are indicated by color.

Somatic mutations show positional biases on consensus RBP motifs

We next decided to determine whether there are particular positions in the found enriched motifs that could be more frequently mutated than others. To this end, for each of the enriched motifs we grouped the conserved SMRs containing this motif and extracted the piece of sequence corresponding to the instance of the motif. These instances were then used to form a multiple sequence alignment (MSA) to determine equivalent positions of the binding motifs across motif instances in the different SMRs. Additionally, over the consensus motif built from the MSA, we calculated the density of somatic mutations and germline mutations per position (Methods). This analysis showed that PTBP1 motifs in 5UTR SMRs are frequently mutated in two positions that show mostly C in the consensus (Fig. 4A). However, these two positions show T>C and C>T mutations, which in theory would leave the motif CT-rich (Fig. 4A). These mutations occur in a large number of genes, including the cancer driver *SPOP*, and they seem predominant in melanoma (SKCM) (Supp. Fig. 8). We also found frequent mutations of a motif for ESRP2 binding in introns, with a strong enrichment for G>A mutations at position 4 that would disrupt the motif (Fig. 4B). Interestingly, this position does not show any overlap with germline mutations. Mutations in intronic ESRP2 motifs occur in multiple genes, including the Estrogen receptor *ESR1*, which contains an SMR with mutations in ESRP2 in BRCA (breast cancer) and UCEC (uterine cancer) samples (Supp. Fig. 8). Among the enriched motifs in CDS SMRs we had found one putative for QKI (Fig. 4C). Although not many mutations fall in this motif, all of them fall in the same position and would potentially disrupt the motif. All these mutations occur in colorectal tumors (CRC) and fall in the cancer drivers *APC* and *ATM*

(Fig. 4C). Interestingly, in EXON SMRs we find a number of mutations on the motif for PUM1, which shows high conservation (Fig. 4D). The majority of those mutations are G>A in a position that would disrupt the motif. Here we see it is mutated in transcripts of the non-coding RNA genes RP11-136K7.3 and LINC00473 and a non-coding transcript in the protein-coding gene SEC61G (Fig. 4D). Although PUM1 has been observed to bind to 3'UTR regions (Li et al. 2010), our results suggest also a role in non-coding RNAs.

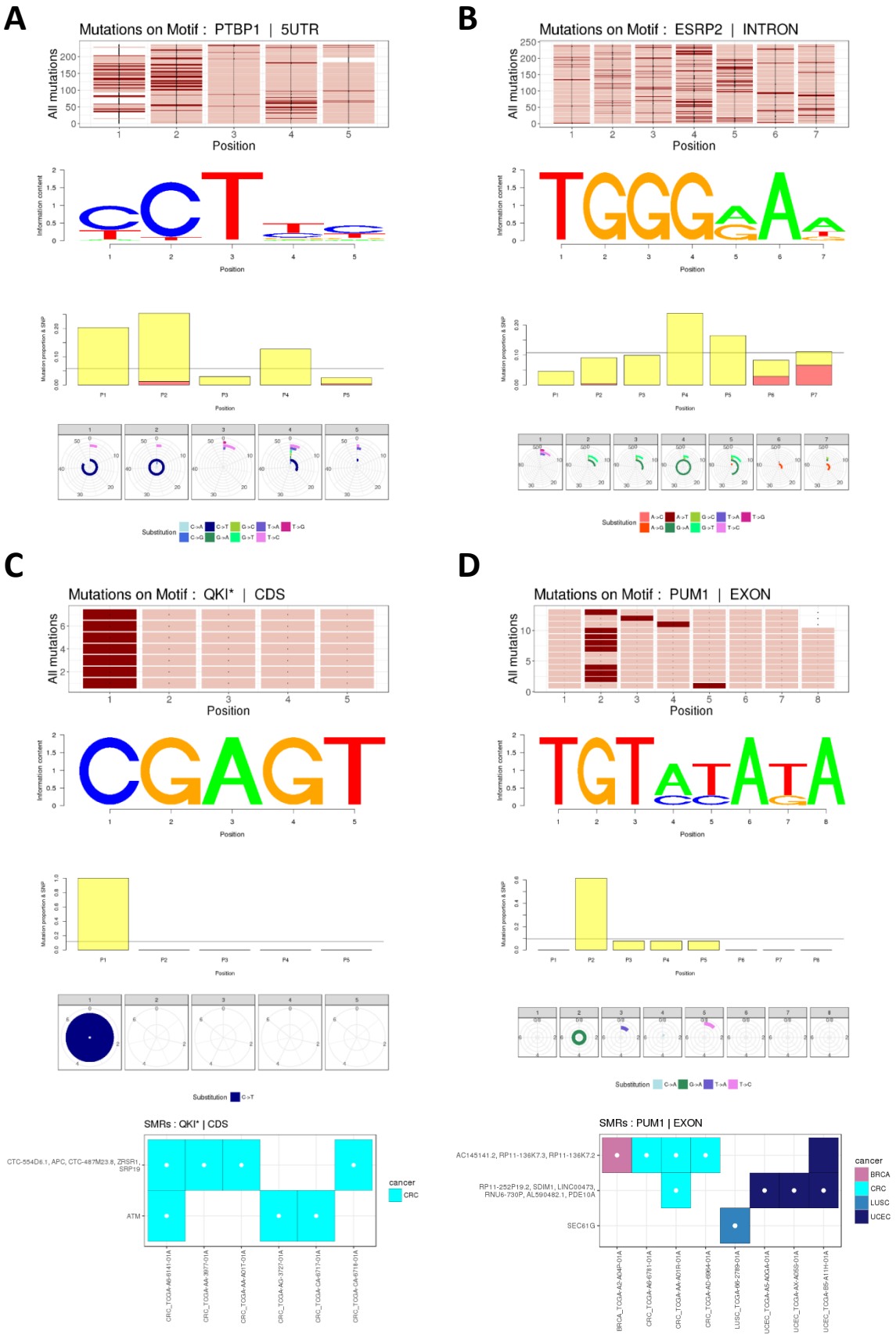

A
Mutations on Motif : PTBP1 | 5UTR
All mutations
Position
Information content
Mutation proportion & SNP
Substitution
B
Mutations on Motif : ESRP2 | INTRON
All mutations
Position
Information content
Mutation proportion & SNP
Substitution
C
Mutations on Motif : QKI* | CDS
All mutations
Position
Information content
Mutation proportion & SNP
Substitution
SMRs : QKI* | CDS
CTC-554D6.1, APC, CTC-487M23.8, ZRSR1, SRP19
ATM
cancer
CRC
D
Mutations on Motif : PUM1 | EXON
All mutations
Position
Information content
Mutation proportion & SNP
Substitution
SMRs : PUM1 | EXON
AC145141.2, RP11-136K7.3, RP11-136K7.2
RP11-252P19.2, SDIM1, LINC00473, RNU6-730P, AL590482.1, PDE10A
SEC61G
cancer
BRCA
CRC
LUSC
UCEC

Figure 4. Positional biases of cancer mutations in RBP motifs. Mutation distribution of somatic mutations in PTBP1 motifs in 5UTR SMRs **(A),** in ESRP2 motifs in INTRON SMRs **(B)**, in QKI motifs in CDS SMRs **(C)**, and in PUM1 motifs in EXON SMRs **(D)**. In each plot, the upper panel shows the multiple sequence alignment (MSA) built with the SMRs containing the motif. Light red lines indicate that this position is covered by the SMR. A dark red line indicates that this position is mutated. White spaces indicate positions not covered by any SMR. If the mutation falls outside the motif, the SMR is still aligned and shown. In the second panel we show the logo built from the MSA. In the third panel we show the density distribution of mutations per position, calculated as the fraction of all mutations considered in the MSA that fall in each position. In orange we indicate the proportion of those mutations that coincide with a germline SNP in the same position of the SMR. The lower panel shows the number of substitutions observed at each position color-coded by type of substitution. Substitutions are indicated according to the strand of the gene hosting the SMR. In **(C)** and **(D)** we indicate in a matrix plot below the genes (y axis) with SMRs containing the QKI and PUM1 motifs. The tumor types are color-coded and patients are given in the x-axis. A white dot inside the color square indicates that the mutation falls inside the motif for that patient.

Somatic mutations in SMRs show an impact in RNA processing and expression

To determine the impact on RNA of the somatic mutations observed on enriched RBP binding motifs, we analyzed RNA-seq data from the same samples of the PAN505 dataset. We first estimated the impact of the mutations on the expression of transcripts. Transcript abundances in TPM units were estimated per sample (Methods). For each patient with a mutation in an SMR, we considered each transcript overlapping the SMR and compared its abundance with the distribution of

abundances of the same transcript in patients from the same tumor type that did not harbor any mutation in the same SMR (|z-score|>1.96 and $|\log_{10}$ fold change|>0.5) (Methods). Most of the significant changes detected correspond to mutations in 5UTR, INTRON and EXON SMRs. The motif in 5UTR SMRs that most frequently associates to transcript expression changes is PTBP1 (Fig. 5A). Interestingly, mutations in the DDX58 motif in the CDS are associated to expression changes in *KRAS*, and mutations in QKI in the CDS are associated to expression changes in the cancer driver *ATM* (Fig. 5B) (Supp. Table S8). In INTRON SMRs, the majority of significant changes in expression occur in for ZFP36 (Fig. 5C), whereas in EXON SMRs the most common motif with effects on transcript expression was TIA1 (Figs. 5D). Both motifs harbor frequent T>C and C>T mutations (Figs. 5E and 5F).

To determine the impact on RNA-processing, we analyzed all possible exon-exon junctions defined from spliced reads mapped to the genome. All junctions appearing in any given patient were clustered to define homogeneous clusters across patients but were quantified per patient (Methods). For each patient with an observed mutation in an SMR, we considered each junction overlapping that SMR and compared the junction inclusion level with the distribution of inclusion levels of the same junction in patients from the same tumor type that did not have any mutation in the same SMR (Methods). Using a minimum |delta-PSI| = 0.1, the majority of changes are observed in mutations in intronic SMRs (Fig. 6) (Supp. Table S9). Interestingly, the 5UTR SMR *SPOP* is associated with the significant inclusion of a new junction (Fig. 6) (Supp. Table S9), which indicates that the novel SMR found in *SPOP* could impact its processing.

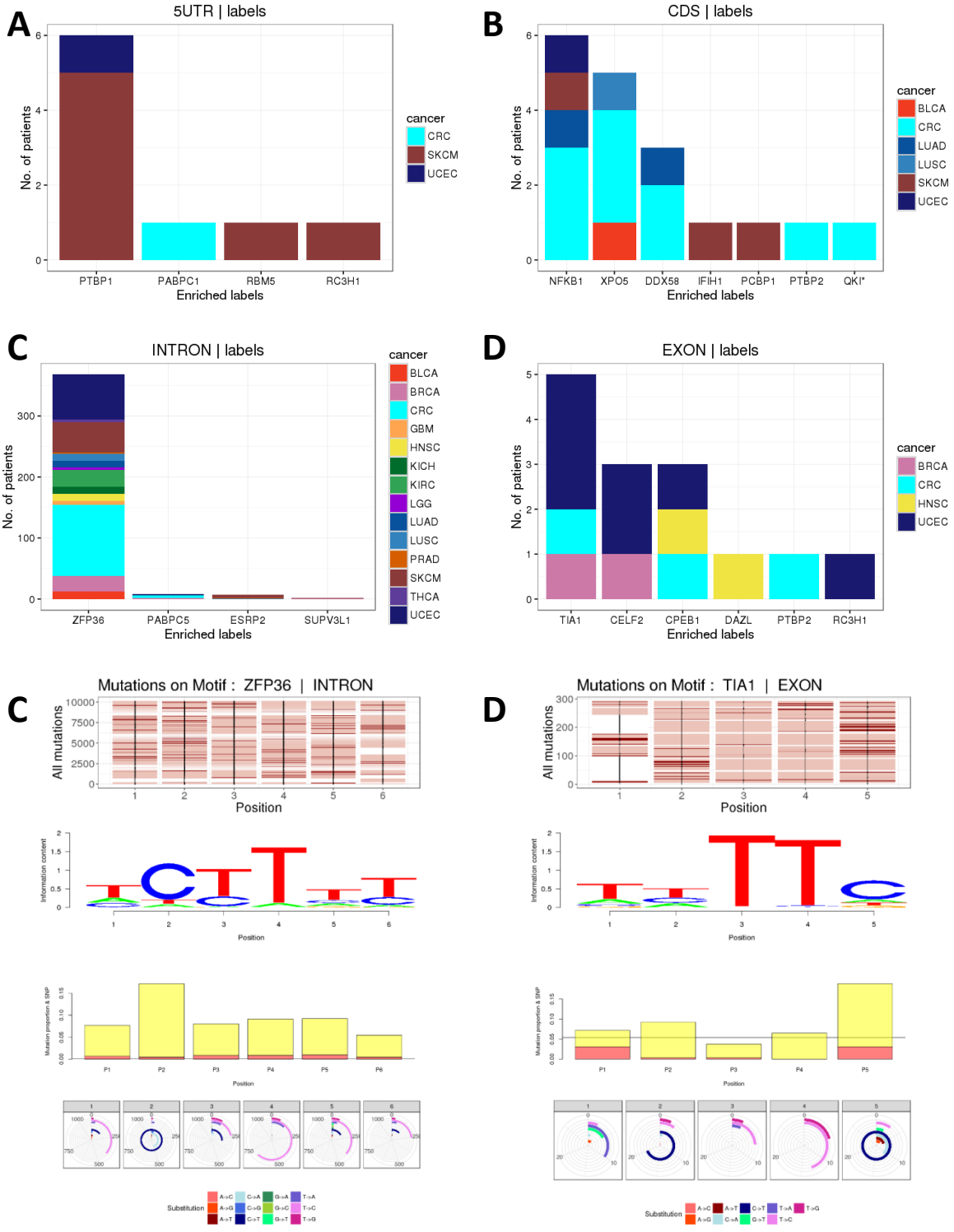

A
5UTR | labels
No. of patients
Enriched labels
PTBP1
PABPC1
RBM5
RC3H1
cancer
CRC
SKCM
UCEC
B
CDS | labels
No. of patients
Enriched labels
NFKB1
XPO5
DDX58
IFIH1
PCBP1
PTBP2
QKI*
cancer
BLCA
CRC
LUAD
LUSC
SKCM
UCEC
C
INTRON | labels
No. of patients
Enriched labels
ZFP36
PABPC5
ESRP2
SUPV3L1
cancer
BLCA
BRCA
CRC
GBM
HNSC
KICH
KIRC
LGG
LUAD
LUSC
PRAD
SKCM
THCA
UCEC
D
EXON | labels
No. of patients
Enriched labels
TIA1
CELF2
CPEB1
DAZL
PTBP2
RC3H1
cancer
BRCA
CRC
HNSC
UCEC
C
Mutations on Motif : ZFP36 | INTRON
All mutations
Position
Substitution
D
Mutations on Motif : TIA1 | EXON
All mutations
Position
Substitution

Figure 5. Transcript expression changes associated to mutations in RBP motifs. For each RBP motif enriched in 5UTR SMRs **(A)**, CDS SMRs **(B)**, INTRON SRMs **(C)**, or EXON SMRs **(D)** we show the number of patients (y axis) for which a mutation in the motif is associated with a significant change in transcript expression. **(E)** and **(F):** distribution of somatic mutations in the ZFP36 motif in INTRON SMRs, and in the TIA1 motif in EXON SMRs, respectively. The upper panel shows the multiple sequence alignment (MSA) built with the SMRs containing the motif. Light red lines indicate that this position is covered by the SMR. A dark red line indicates that this position is mutated. White spaces indicate positions not covered by the SMR. If the mutation falls outside the motif, the SMR is still aligned and shown. In the second panel we show the logo built from the MSA. In the third panel we show the density distribution of mutations per position, calculated as the fraction of all mutations in considered in the MSA that fall in a given position. In orange we indicate the proportion of those mutations that coincide with a germline SNP in the same position of the SMR. The lower panel shows the number of substitutions observed at each position color-coded by type of substitutions. Substitutions are indicated according to the strand of the gene hosting the SMR.

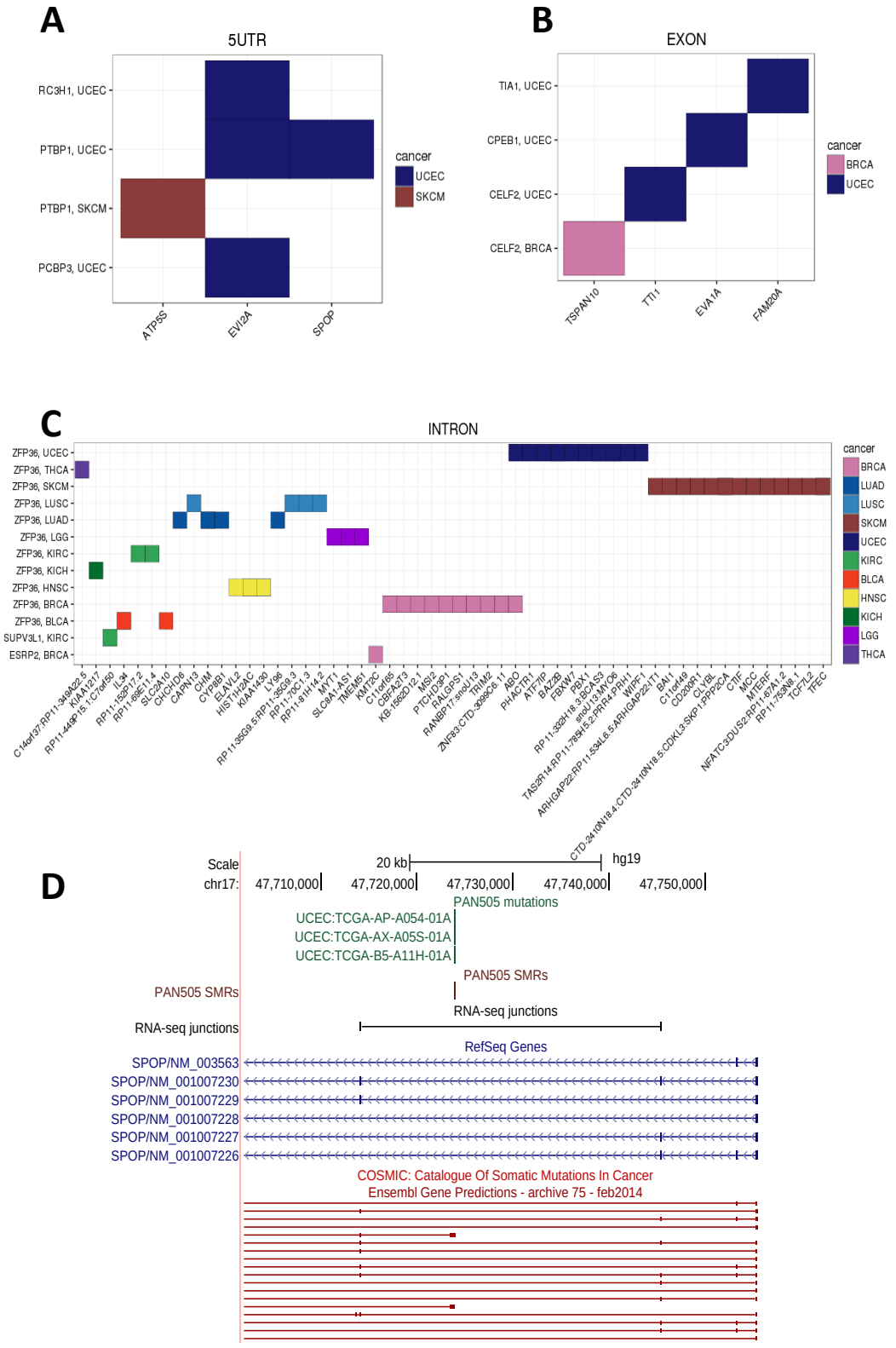

A
5UTR
RC3H1, UCEC
PTBP1, UCEC
PTBP1, SKCM
PCBP3, UCEC
ATP5S
EVI2A
SPOP
cancer
UCEC
SKCM
B
EXON
TIA1, UCEC
CPEB1, UCEC
CELF2, UCEC
CELF2, BRCA
TSPAN10
TTI1
EVA1A
FAM20A
cancer
BRCA
UCEC
C
INTRON
ZFP36, UCEC
ZFP36, THCA
ZFP36, SKCM
ZFP36, LUSC
ZFP36, LUAD
ZFP36, LGG
ZFP36, KIRC
ZFP36, KICH
ZFP36, HNSC
ZFP36, BRCA
ZFP36, BLCA
SUPV3L1, KIRC
ESRP2, BRCA
cancer
BRCA
LUAD
LUSC
SKCM
UCEC
KIRC
BLCA
HNSC
KICH
LGG
THCA
D
Scale
chr17:
20 kb
hg19
47,710,000
47,720,000
47,730,000
47,740,000
47,750,000
PAN505 mutations
UCEC:TCGA-AP-A054-01A
UCEC:TCGA-AX-A05S-01A
UCEC:TCGA-B5-A11H-01A
PAN505 SMRs
PAN505 SMRs
RNA-seq junctions
RNA-seq junctions
RefSeq Genes
SPOP/NM_003563
SPOP/NM_001007230
SPOP/NM_001007229
SPOP/NM_001007228
SPOP/NM_001007227
SPOP/NM_001007226
COSMIC: Catalogue Of Somatic Mutations In Cancer
Ensembl Gene Predictions - archive 75 - feb2014

Figure 6. Changes in junction usage associated to mutations in RBP motifs. We show the genes (x-axis) and the RBP motifs and tumor types (y-axis) in which a mutation on the RBP motif in 5UTR **(A)**, EXON **(B)** and INTRON **(C)** SMRs has been found to be associated with a significant change in the usage of a splicing junction (|z-score|>1.96, |delta-PSI|>0.1). **(D)** We show the junction in the cancer driver SPOP that increases PSI in association with a mutation in a TC-rich motif in a 5'UTR exon. This exon is present in the Ensembl annotation but not in the UCSC/RefSeq annotation.

Discussion

We have described a novel method to identify and characterize relevant cancer mutations in coding and non-coding regions exploiting their potential to be involved in protein-RNA binding. We performed an unbiased identification of significantly mutated regions (SMRs) based on their enrichment of mutations according to local mutation rates and nucleotide mutation biases. Although nucleotide mutation signatures have been established at genome-scale using tri-nucleotides (Alexandrov et al. 2013), at the scale of a gene-locus at which our method operates there are not enough mutation counts to establish similar frequency models. In fact, the single-nucleotide signature used by our method will generally penalize observed mutations more often than higher order models, hence making our approach conservative. Additionally, global mutational biases (Mularoni et al. 2016) might not accurately represent the local mutational biases at gene scale. We showed that our local approach did not present biases in relation to replication timing and gene expression values, and had a good correspondence with other previous methods that take these features into account to assess mutational enrichment (Lochovsky et al. 2015).

Our approach recovers known SMRs in coding regions and UTRs, and predicts new cases, some also in non-coding RNAs and introns. Our method present several advantages with respect to previous methods. Since we test all positions along genes, we can detect potentially relevant deep intronic mutations. This is an advantage over previous methods that have only tested positions on exons or in adjacent intronic regions (Lanzós et al. 2017; Mularoni et al. 2016). Additionally, unlike previous methods (Lanzós et al. 2017; Mularoni et al. 2016; Piraino and Furney 2017), we provide precise location of the mutations on the candidate regulatory region and describe the commonalities between them. This is crucial to enable the interpretation of the non-coding mutations to identify regulatory mechanisms that may be altered in cancer. At the initial stage of analysis we do not assume any specific functional impact, like secondary structure or conservation. Other features may determine the processing, stability and function of a pre-mRNA or mature RNA molecules, and here we assumed that these are mediated through their ability to interact with proteins. Accordingly, we uncovered many new significantly mutated non-coding regions. Also, these interactions are not necessarily conserved across species or may have some sequence redundancy; hence some of the regions we described will be missed by methods based on species or population conservation. Additionally, unlike previous approaches, we have tested the impact on RNA of the found significant mutations. We directly measured the relevance to RNA processing by measuring the stranded enrichment of binding motifs for RNA binding proteins (RBPs). Previous methods that tested the impact on RNA did not perform an analysis in relation to mutations on specific RBP binding motifs or using RNA sequencing from the same samples (Supek et al. 2014). Recently, a method was developed that restricted the search to conserved positions (Piraino and Furney 2017). Although conservation

can be used as a proxy for functional relevance, we observed that not all instances of the regulatory motifs corresponding to a common regulatory factor are equally conserved. In our approach, we tested their functional relevance by relating the sequence to other sites in the genome to identify recurrent motifs and by directly measuring the impact on RNA from the same samples.

We found a large number of intronic SMRs with an impact in the expression and splicing of genes, including non-coding RNAs. Other methods assumed that the function of lncRNAs or of non-coding regions in the RNA is determined through its structure or that only mutations on exons can affect their function. However, lncRNAs are not as much structured as initially thought (Rivas et al. 2017). Also, exonic as well as intronic regions are relevant for RNA processing; hence mutations in any of those regions could impact the integrity of the lncRNA and therefore its function. Accordingly, one could argue that it is not correct to use intronic regions as mutational background. Although we found multiple SMRs in introns and exons of non-coding RNAs that are conserved and share a regulatory motif, we did not find any in the cancer related lncRNA gene *MALAT1*. *MALAT1* has multiple somatic mutations in the studied cohorts but these are quite scattered through the gene locus. As our method was based on an unbiased search of short windows enriched for mutations, these did not pass the filters in *MALAT1*. Previous methods (Mularoni et al. 2016, Lanzos et al. 2017) found *MALAT1* among their predictions. In one case, this was based on the test of the impact of individual mutations on the secondary structure (Mularoni et al. 2016). *MALAT1* is known to be processed into smaller RNA fragments, which are likely processed from secondary structures (Wilusz et al. 2008). However, *MALAT1* function seems to be exerted as a long RNA rather than a processed one (Gutschner et al. 2013). Thus it is not clear yet whether and how

the mutations disrupt *MALAT1* function. In another method, *MALAT1* was also identified using a method that compares with the mutation count on exons with a background composed of introns and flanking regions (Lanzos et al.2017). However, *MALAT1* has only two introns that are extremely short, which may introduce some biases in the analysis. *MALAT1* involvement in cancer has been related to its expression (Gutschner et al. 2013), so it remains to be elucidated whether the mutations in *MALAT1* have any relevance or are an artifact of its extremely high expression (Lawrence et al. 2013).

Our approach presents various limitations. The analysis of non-coding mutations may be still underpowered due to the small number of patients analyzed in each cohort. In fact, the PAN507 presented very few SMRs and no significantly enriched RBP motifs. We also observed a low overlap between the SMRs calculated from each cohort, so we are still far from saturation. Another limitation is that to be able to relate functionally analogous regions we had to work with specific representations of RNA binding motifs. The analysis is thus limited by how accurate this representation is. Short fixed nucleotide strings hold sufficient information to understand protein-RNA interactions (Daubner et al. 2013) but do not generalize well to genome-wide searches. Moreover, despite the development of new methodologies to describe the sequence specificities of RBPs (Alipanahi et al. 2015), their precision at genome scale (the fraction of real sites from all predictions genome-wide) remain uncertain. Additionally, different RBPs bind very similar or even the same sequences; hence the identification of protein-RNA interactions remain challenging. The availability of high-throughput methods to identify regions of protein-RNA interactions at genome scale, like CLIP-seq, has the potential to help in this direction (Sundararaman et al. 2016), but we detected very little overlap with the

sequence motif approach. We conclude that there is still a limitation in the capacity to accurately identify the regions for protein-RNA binding.

We found frequent mutations associated to CT-rich and T-rich motifs in introns, UTRs and exons from non-coding RNAs. Although we linked those to specific factors, including ZFP36, PTBP1, PTBP2 and TIA1, they could be binding sites for other RBPs with similar affinities. This mutational pattern in these motifs frequently observed melanoma (SKCM). It is possible that this mutation is partly associated to the enriched mutational processes in tumors caused by specific mutagens, like ultraviolet light (Viros et al. 2014; Pleasance et al. 2010a) or tobacco smoke (Govindan et al. 2012; Pleasance et al. 2010b). In fact, it is likely that the vast majority of non-coding mutations identified are passengers and merely reflect the mutational processes of the tumor (Pleasance et al. 2010b; Alexandrov et al. 2016). It is possible that except for a handful of cases, most of the non-coding mutations with a functional impact tend to occur in few patients. This motivates the extension of the assumption of recurrence to analogous sites at different genomic positions, which allows describing similar phenotypes arising from different alterations.

Methods

Detection of significantly mutated regions (SMRs)

We aimed to identify significantly mutated regions (SMRs) in both coding and non-coding regions of genes taking into account regional and sequence mutational biases. We used data for somatic mutations from whole genome sequencing for 505 tumor samples from 14 tumor types (Fredriksson et al. 2014) (cohort PAN505), as well as data for

507 tumor samples from (Alexandrov et al. 2013) (cohort PAN507). For the PAN505 cohort we used all substitutions except those with precise allelic match to a known germline variant in dbSNP138. For the PAN507 we considered all substitutions provided in (Alexandrov et al. 2013). Our method to identify significantly mutated regions (SMRs) works as follows. We only considered substitutions falling in the genomic extensions spanned by genes (Gencode annotation version 19). We used genes annotated as protein-coding as well as non-coding. For each gene, the method works by first finding short windows that show an enrichment of mutations according to the mutation rate in the same gene locus and according to the mutation nucleotide biases. As RNA binding proteins (RBPs) interact with pre-mRNAs through short nucleotide stretches, we considered windows of size 7 (Fig. 1). We used a sliding window approach, whereby along a gene locus all overlapping windows of length 7 and harboring at least one mutation were tested. Using shorter windows increases the number of computations but the results are similar. Using larger windows we would lose positional resolution.

For each 7-mer window we performed a double test to determine the enrichment to account for the local variations and nucleotide biases in mutation rates. Given a window with n mutations in a gene of length L and N mutations, we performed a binomial test using an expected local mutation rate of N/L. All tested windows in a gene were corrected for multiple testing using the Benjamini-Hochberg method. Additionally, to account for the nucleotide biases we compared the mutation count in a window with the expected count according to the distribution of mutations at each nucleotide in the same gene locus. For each base a we calculated the rate of mutations falling in that base along a gene $R(a) = m(a)/n(a)$, where $n(a)$ is the number of a bases in the gene and $m(a)$ is the number of those bases that are mutated. The expected

mutation count is then calculated using the nucleotide counts in the window and the mutation rate per nucleotide. For instance, for the 7-mer window AACTGCAG, the expected count was calculated as: $E = 3R(A) + 2R(C) + 2R(G) + R(T)$. This was compared to the number of mutations in the window n to define a nucleotide bias (NB) score: $NB\text{-}score = log2(n/E)$ for each 7-mer window. We filtered out 7-mer windows corresponding to single-nucleotide repeats (e.g. AAAAAAA). Further, we compared the NB-scores of windows with only 1 or 2 mutations with windows with >= 3 mutations and set the NB-score to be > 6. Further, we kept only 7-mer windows with adjusted p-value < 0.05. Although our p-values are various orders of magnitude lower than the expected values, the cases that we do not consider significant show a trend similar to the expected values. Also, the extra constraint on the NB-score will reduce the number of possible false cases.

Significant 7-mer windows were classified according to the genic region in the same strand in which they fall: 5' or 3' untranslated regions (5UTR/3UTR), coding sequence (CDS), exon in non-coding RNA (EXON), or intron (INTRON). To unambiguously identify each window to a region type using the precedence CDS > 5UTR/3UTR > EXON > INTRON. That is, if a window overlapped a CDS, it was classified as CDS; else, if it overlapped an UTR, it was labeled as UTR; else, if it mapped an exon in a non-coding RNA, it was labeled as EXON. Remaining windows were labeled as intronic. Windows that lie across region boundaries were split into two subwindows, each entirely within a region type. Windows and sub-windows were then clustered according to genomic overlap into significantly mutated regions (SMRs) of a single type, keeping only those of length 7 or longer, which ensures that each SMR contains at least one full significant 7-mer window. To each SMR, we assigned the highest

score and the lowest p-value of the windows comprising the cluster. Code for this analysis is available at https://github.com/comprna/mira

Comparison to expression, replication timing analyses and LARVA

Data for replication time was obtained from (Locovsky et al. 2015). As this data does not cover the entire genome, only SMRs for which replication time was available were analyzed. Expression data was calculated for the same samples from the PAN505 cohort. For each SMR from the PAN505 cohort, we considered those annotated transcripts overlapping with the SMR. Using RNA-seq from the patients in whom the SMR appeared mutated, we calculated the total TPM for the overlapping transcripts per patient and averaged these values across patients. For each SMR we compared the average expression of the SMR-containing transcripts in the mutated samples with the number of mutations. Although some SMRs appeared associated with low or no expression, as we cannot assess this for all of the cohorts analyzed, hence we did not remove these from further analysis. For comparison, all SMRs obtained in each cohort were analyzed with LARVA (Lochovsky et al. 2015) with the same mutation data. Specifically, we compared the significance of our SMRs with the significance from LARVA with the model with a beta-binomial distribution with the replication timing correction (p-bbd-cor), which accounts for the over dispersion of the mutation rates and regional biases.

Control regions for SMR comparison

For each region type (CDS, 5UTR, 3UTR, INTRON, EXON) we generated control regions by sampling non-overlapping regions of the

same type from the entire Gencode annotation controlling for G+C content and length. For each SMR, we sampled randomly 100 control regions of the same length and the same type from the annotation allowing for a maximum variation of G+C content of 5%. Each of these 100 control regions was collated into a different control set to create 100 control sets each with one region matching each of the SMRs.

Motif and CLIP enrichment analysis

We studied the enrichment of potential RNA-binding protein (RBP) motifs on the SMRs with potential relevance in the studied tumors. We considered k-mer motifs (k variable length) for 330 RBPs present in ATtRACT DB (Giudice et al. 2016) and for SRMM4 (Raj et al. 2014). We located these k-mers in the identified SMRs. For each RBP name we calculated the number of SMRs in which any of its associated k-mers appears. Similarly, we calculated these counts for the 100 control sets; and calculated a mean-based z-score comparing the observed counts in the SMRs to the distribution in control regions separately for each region type. The unbiased analysis for k-mers (k=7) was done in a similar way, but considering each k-mer separately. For each k-mer we counted the number of times it appeared and a z-score was computed for each individual k-mer comparing the distribution of counts in the control regions. The motif enrichment analysis was repeated reversing the strand of all observed SMRs and the control regions. RBP names or k-mers, that appeared significantly enriched in the direct and reversed analyses for the same region type were discarded. We considered significantly enriched the RBP names and k-mers with z-score > 1.96 and with > 5 counts. We further only kept k-mers that appeared in > 2 SMRs.

We gathered binding sites from 91 CLIP experiments for 68 different RBPs from multiple sources (Sundararaman et al. 2016; Shao et al. 2014; Rodor et al. 2016; Yang et al. 2015; Raj et al. 2014; Bechara et al. 2013; Best et al. 2014) and selected the available significant CLIP clusters from each experiment. For datasets with two replicates we selected the intersecting regions of the significant CLIP clusters: the genomic ranges covered by both replicates. For each CLIP dataset, we then performed a Fisher's exact test with the number of SMRs and control regions that show overlap or not with the CLIP regions. The association test was repeated by reversing the strand of all SMRs and control regions and CLIP experiments that appeared significantly enriched in both analyses were discarded.

Conservation analysis

PhastCons scores were obtained for the 7-way and 20-way alignments for hg38 from UCSC and transformed to hg19 coordinates using the liftOver tool (Tyner et al. 2017). The conservation score was calculated per SMR by averaging the PhastCons scores across the positions of the SMR.

Significant mutations per position of a motif

Given an enriched motif RBP name, we considered all the SMRs where the k-mers associated to that RBP appeared. We performed a multiple sequence alignment (MSA) of those k-mers using MUSCLE (Edgar 2004), including the possible multiple instances of a k-mer. The sequences of the SMRs were then aligned according to the relative position of the k-mers in the MSA. Sequence logos were built from this alignment of SMR sequences. Somatic mutations were counted per position relative to the MSA to obtain the fraction of mutated SMRs per

position in the built motif. As a control, we shuffled the same mutations on the positions aligned in each SMR to produce an average expected number of mutations per position associated to that motif. Germline mutations from the 1000 genomes project (1000 Genomes Project Consortium et al. 2010) were also counted per position in the SMR MSA to estimate the number of mutated motifs per position in the germline.

RNA-seq data analysis

RNA-seq was obtained for the same samples from PAN505 from TCGA (https://gdc-portal.nci.nih.gov/). RefSeq transcript abundances in TPM units were estimated using Salmon (Patro et al. 2015) and inclusion levels (PSI values) for alternative splicing events were estimated using SUPPA (Alamancos et al. 2015). RNA-seq reads were also mapped to the human genome (hg19) with STAR (Dobin et al. 2013). From the BAM files we obtained all possible exon-exon junctions defined by spliced reads that appear in any of the samples. All defined junctions were then grouped into junction-clusters. Two junctions were clustered together if they shared at least one splice-site. Clusters were built considering all junctions present in all patients, but junction read-counts were assigned per patient. Then, for each sample, the read-count per junction was normalized by the total read count of all junctions in that cluster to define a junction inclusion level. If a junction was not expressed in a given sample, it was given zero inclusion level. The software for the junction analysis can be found here https://github.com/comprna/Junckey.

For the positions in the motif that were significantly more mutated compared to the randomized control, we calculated whether mutations were associated to a change in RNA splicing and expression. For the

splicing analysis we calculated the inclusion levels (PSI) of events and junctions per sample. For each selected motif, and for each SMR containing the motif, we tested the association of the mutation with a change in RNA-splicing using outlier statistics. We compared each patient with a mutation in the selected position of the SMR with all the patients for the same tumor type that did not have any mutations in the same SMR. The significance of splicing change was measured in terms of a z-score derived from the junction PSI of the mutated patient and the distribution of junction PSIs in the other non-mutated patients. This was performed in the same way for alternative splicing events. To test the association with expression changes we performed a similar analysis but using the transcript abundances in TPM units for the transcripts including the SMR. For each SMR and each transcript containing the SMR, we compare the transcript $\log_{10}$(TPM) for the patient having a mutation in the SMR with the distribution of $\log_{10}$(TPM) values for the same transcript in the patients with no mutation in that SMR. We considered a change significant for |z-score|>1.96 and the difference between the observed $\log_{10}$(TPM) and the mean of $\log_{10}$(TPM) in patients without mutations in the SMR greater than 0.5 or lower than -0.5.

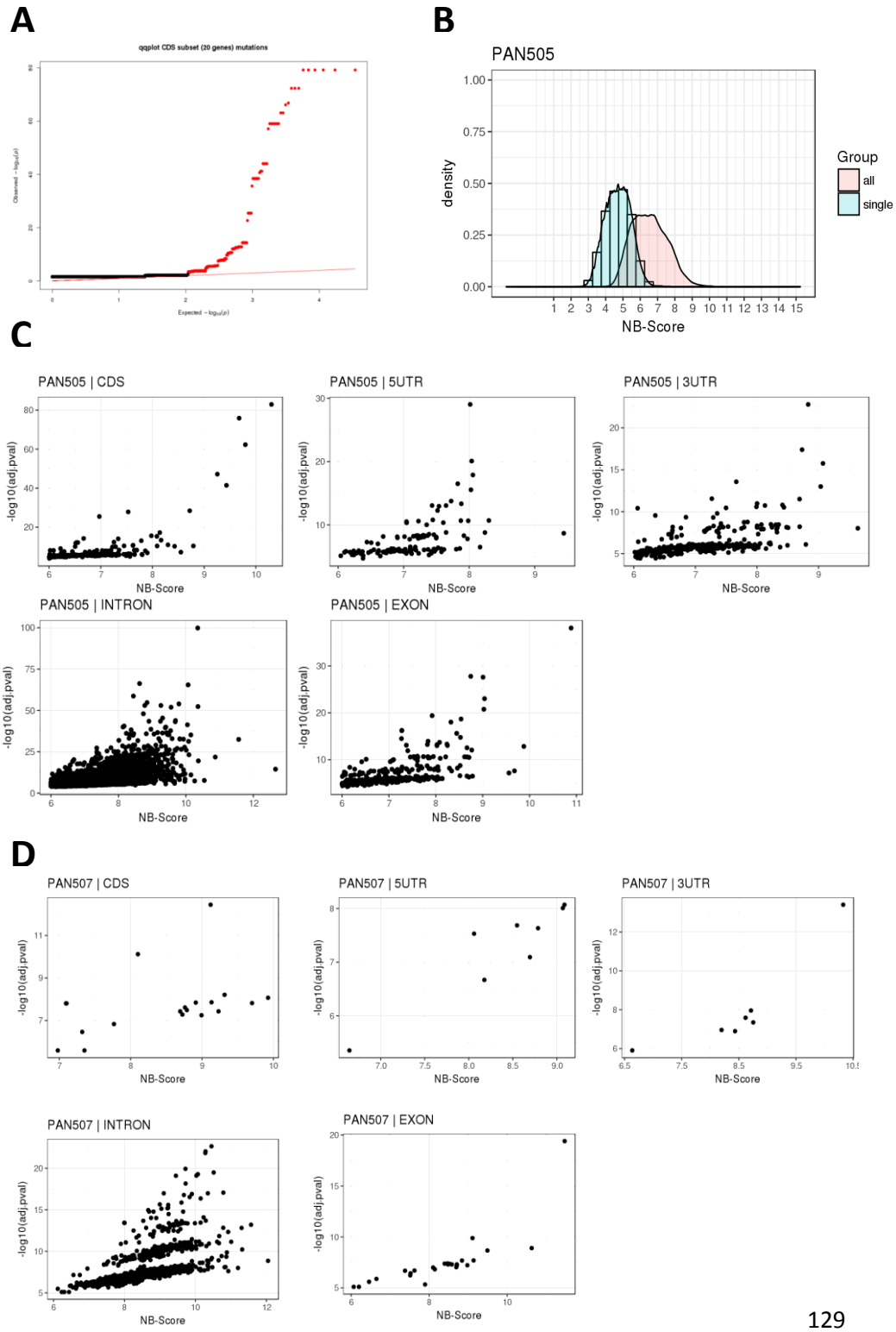
Supplementary Data and Software

Supplementary Data for this manuscript is available at:
http://comprna.upf.edu/Data/MutationsRBPMotifs/
Code use in this manuscript is available at:
https://github.com/comprna/MIRA
All plots in high resolution for the cases shown in the manuscript and for other examples are available at
http://comprna.upf.edu/Data/MutationsRBPMotifs/

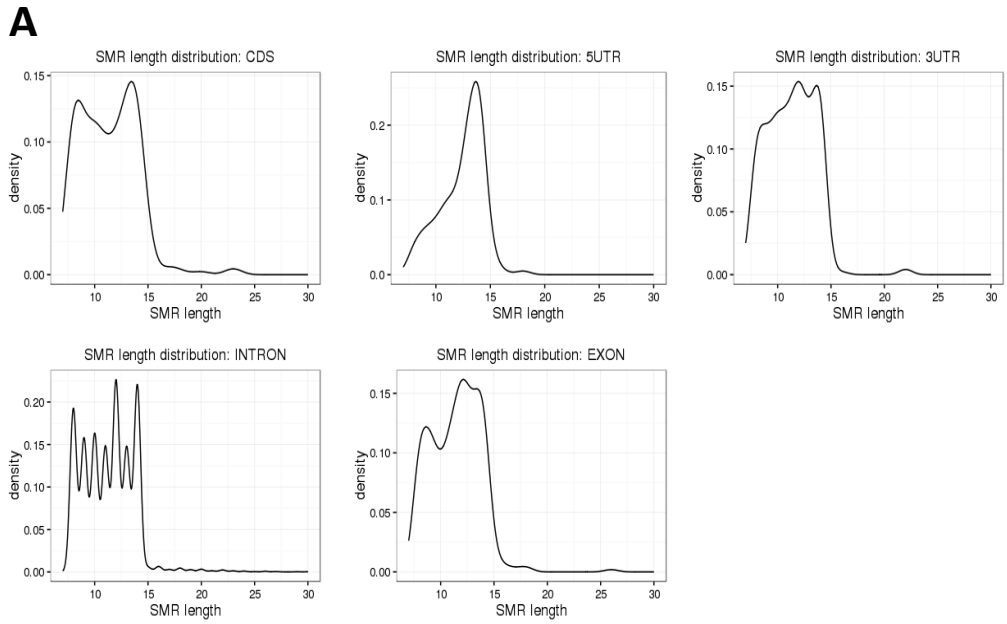
Acknowledgements

We would like to thank R. Jonhson, N. Lopez-Bigas, M. Taylor, and A. Lanzós for useful discussions. This work was supported by the MINECO and FEDER (BIO2014–52566-R) and AGAUR (SGR2014–1121). BS was funded by an FPI grant from the Spanish Government with reference BES-2012-052683.

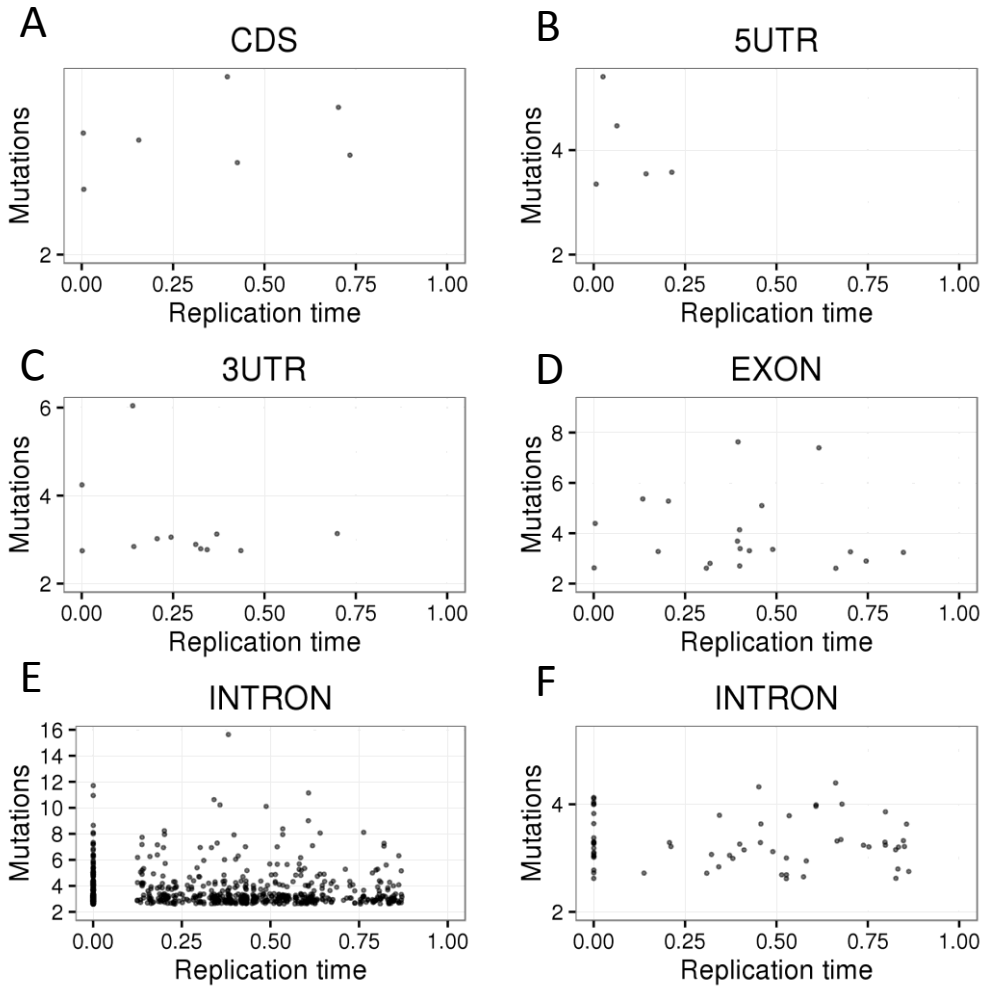
Supplementary Figures



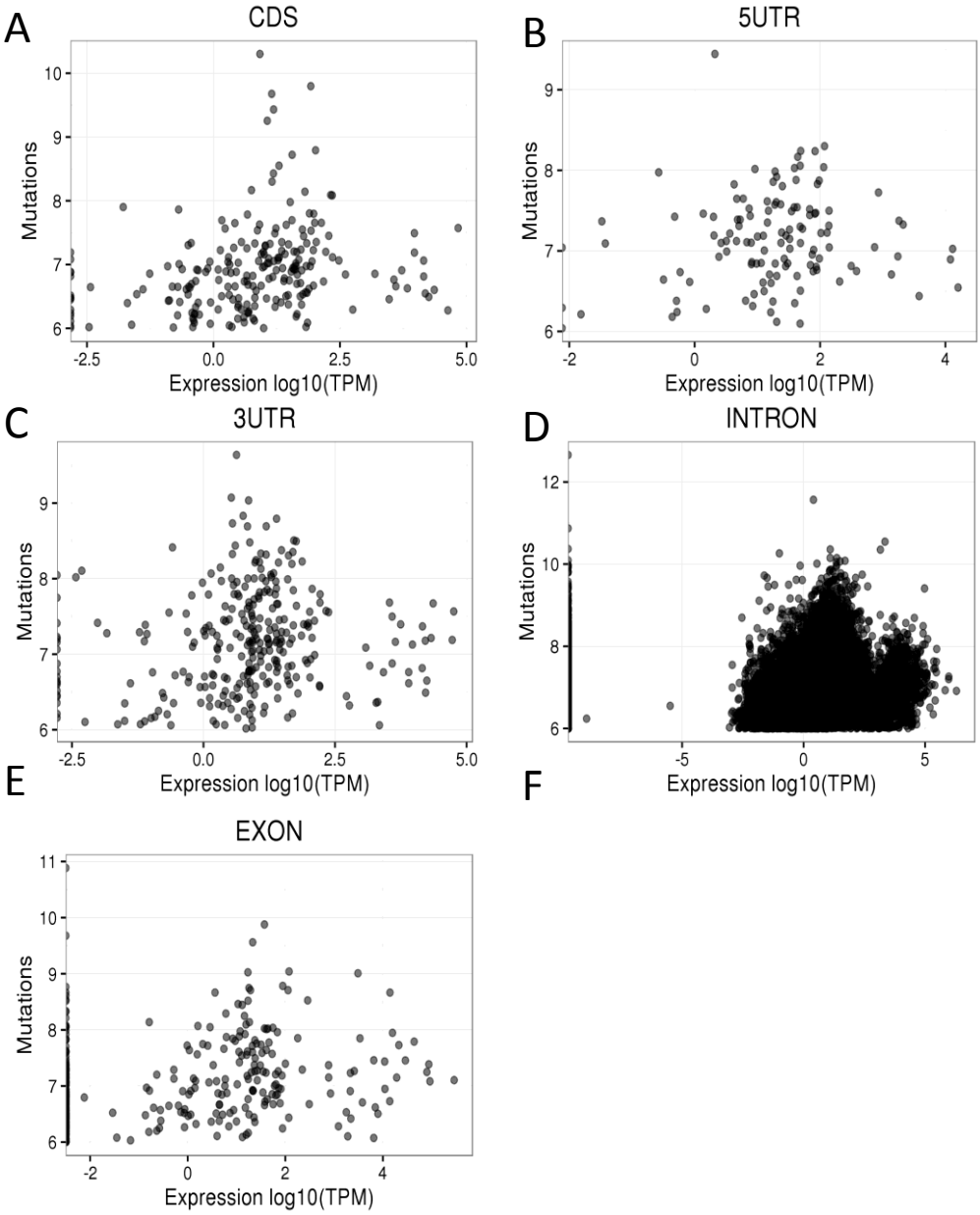
Supplementary Figure 1. Identification of significantly mutated regions (SMRs) (A) QQ-plot comparing the distribution of p-values in our calculated 7-mer windows, using all windows with 1 or more mutations, with the uniform distribution. In red we indicate those that we are taking as significant: 3 or mutations and corrected p-value < 0.05. **(B)** Comparison of the distributions of nucleotide bias (NB) scores in 7-mer windows with 1 mutation (blue) and in 7-mer windows with 3 or more mutations. We selected 7-mer windows with 3 or more mutations and with NB-score >= 6. **(C)** Distribution of the NB-scores and p-values (in $-\log_{10}$ scale) of the identified SMRs from the PAN505 dataset separated by region type. **(D)** Distribution of the NB-scores and p-values (in $-\log_{10}$ scale) of the identified SMRs from the PAN507 dataset separated by region type.



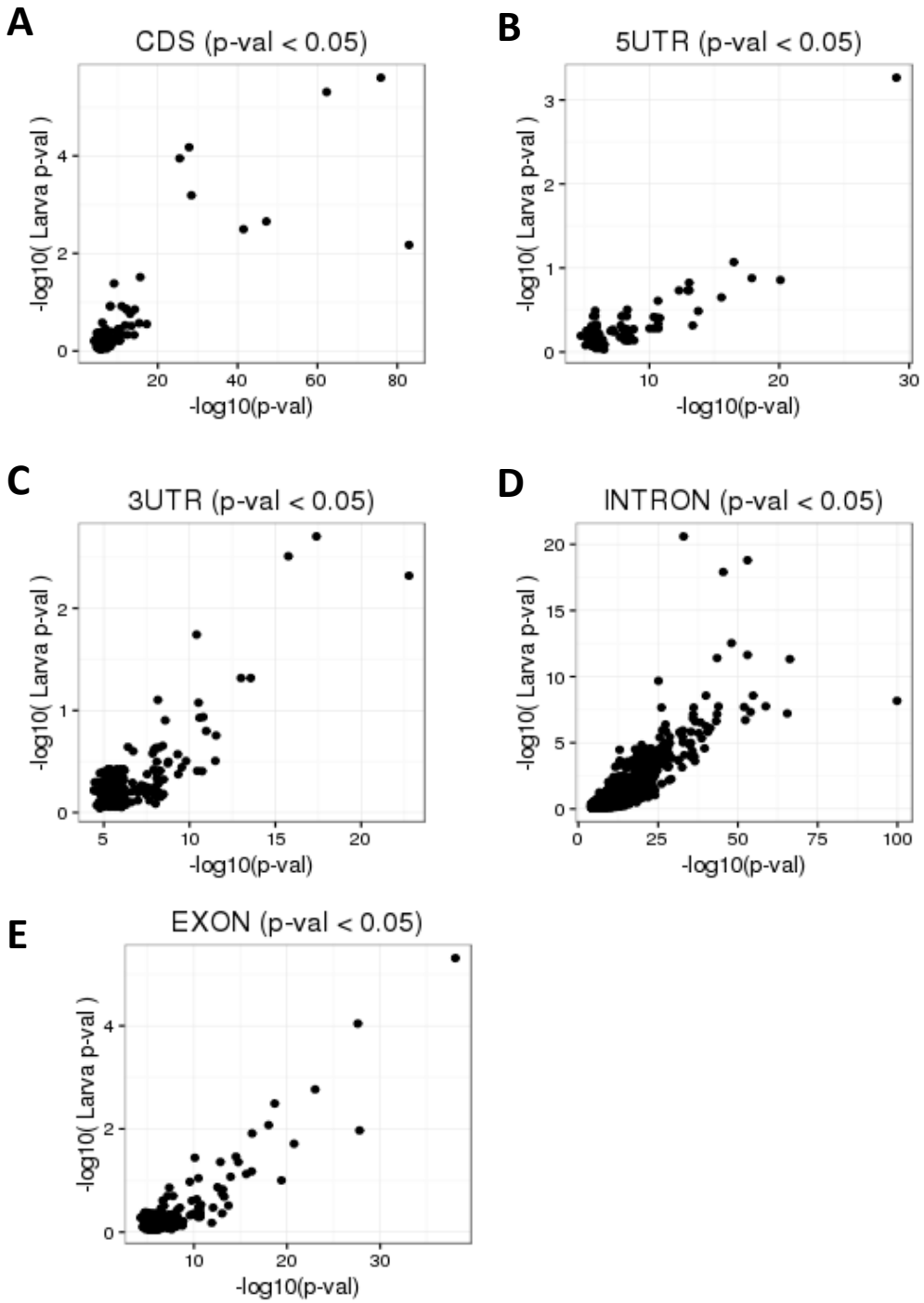
Supplementary Figure 2. Length distribution of SMRs. (A) Length distribution of the identified significantly mutated regions (SMRs) for the PAN505 dataset separately for each region type: protein-coding sequences (CDS), 5' and 3' untraslated regions (5UTR/3UTR), introns (INTRON) and exons of non-coding genes (EXON).



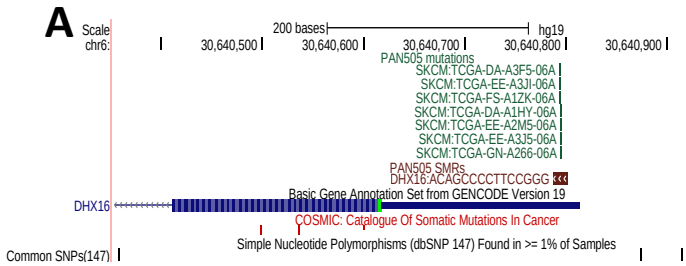
Supplementary Figure 3. (A-E) Relation between replication timing (x-axis) and number of mutations (y-axis) in the SMRs from the PAN505 cohort falling in regions with replication timing data from (Lochovsky et al. 2015). **(F)** For the PAN507 cohort only intronic SMRs fell in genomic regions with replication timing information.

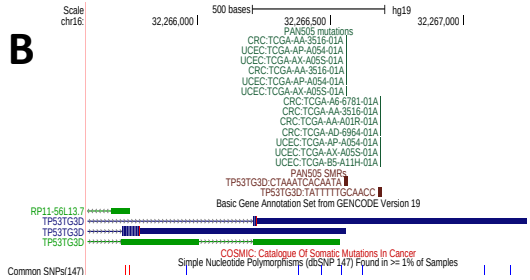


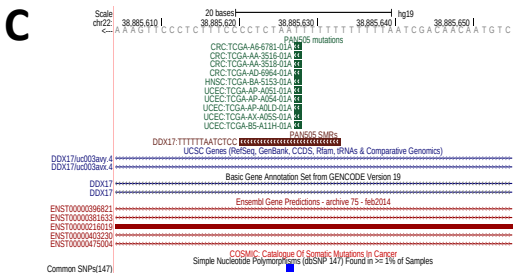
Supplementary Figure 4 (A-E) Comparison between expression and mutation count. For every SMR in the PAN505 cohort, we compared the mutation count (y axis) with the average expression of the transcripts, in $\log_{10}$(TPM) units (x axis), in the same patients harboring the mutations in the SMR. For each transcript including an SMR, we considered the expression of the transcript in the same patients that harbor the mutation. These expression values were averaged over all patients.

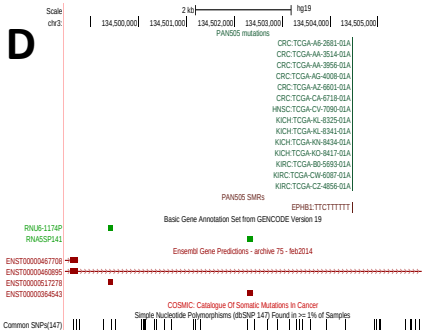


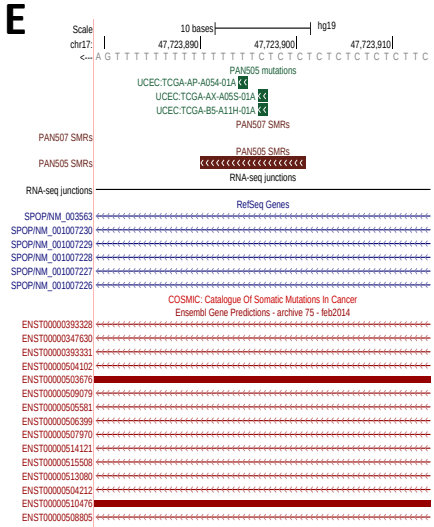
Supplementary Figure 5. Comparison between our SMRs and LARVA (Lochovsky et al. 2015) for the CDS (A), 5UTR (B), 3UTR (C), INTRON (D) and EXON (E) SMRs. For each SMR, we plot the p-value provided by our method in –log10 scale (x axis) and the p-value given by LARVA in –log10 (y axis). LARVA was used with the beta-binomial distribution (bbd) and the correction for replication timing (cor).

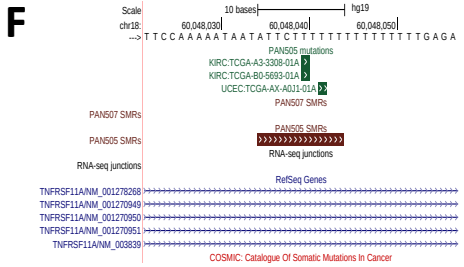

A
Scale
chr6:
200 bases
hg19
30,640,500
30,640,600
30,640,700
30,640,800
30,640,900
PAN505 mutations
SKCM:TCGA-DA-A3F5-06A
SKCM:TCGA-EE-A3JI-06A
SKCM:TCGA-FS-A1ZK-06A
SKCM:TCGA-DA-A1HY-06A
SKCM:TCGA-EE-A2M5-06A
SKCM:TCGA-EE-A3J5-06A
SKCM:TCGA-GN-A266-06A
PAN505 SMRs
DHX16:ACAGCCCCTTCCGGG
Basic Gene Annotation Set from GENCODE Version 19
DHX16
COSMIC: Catalogue Of Somatic Mutations In Cancer
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)


B
Scale
chr16:
500 bases
hg19
32,266,000
32,266,500
32,267,000
PAN505 mutations
CRC:TCGA-AA-3516-01A
UCEC:TCGA-AP-A054-01A
UCEC:TCGA-AX-A05S-01A
CRC:TCGA-AA-3516-01A
UCEC:TCGA-AP-A054-01A
UCEC:TCGA-AX-A05S-01A
CRC:TCGA-A6-6781-01A
CRC:TCGA-AA-3516-01A
CRC:TCGA-AA-A01R-01A
CRC:TCGA-AD-6964-01A
UCEC:TCGA-AP-A054-01A
UCEC:TCGA-AX-A05S-01A
UCEC:TCGA-B5-A11H-01A
PAN505 SMRs
TP53TG3D:CTAAATCACAATA
TP53TG3D:TATTTTTGCAACC
Basic Gene Annotation Set from GENCODE Version 19
RP11-56L13.7
TP53TG3D
TP53TG3D
TP53TG3D
COSMIC: Catalogue Of Somatic Mutations In Cancer
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)

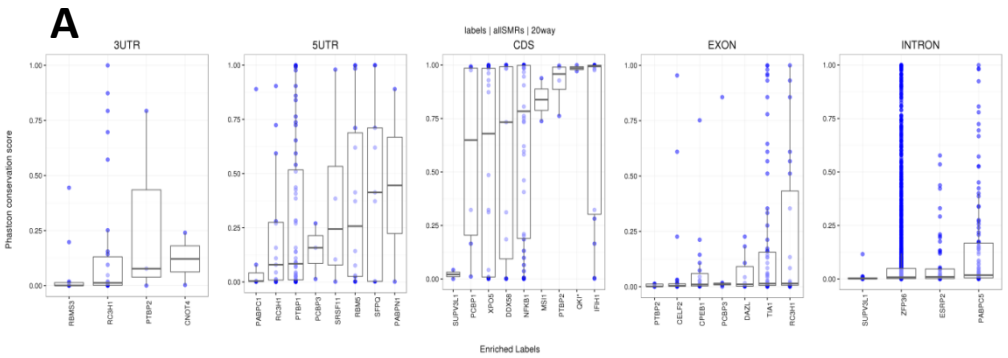

C
Scale
chr22:
20 bases
hg19
38,885,610
38,885,620
38,885,630
38,885,640
38,885,650
PAN505 mutations
CRC:TCGA-A6-6781-01A
CRC:TCGA-AA-3516-01A
CRC:TCGA-AA-3518-01A
CRC:TCGA-AD-6964-01A
HNSC:TCGA-BA-5153-01A
UCEC:TCGA-AP-A051-01A
UCEC:TCGA-AP-A054-01A
UCEC:TCGA-AP-A0LD-01A
UCEC:TCGA-AX-A05S-01A
UCEC:TCGA-B5-A11H-01A
PAN505 SMRs
DDX17:TTTTTTTAATCTCC
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
DDX17/uc003anv.4
DDX17/uc003avk.4
Basic Gene Annotation Set from GENCODE Version 19
DDX17
DDX17
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000396821
ENST00000381633
ENST00000216019
ENST00000403230
ENST00000475004
COSMIC: Catalogue Of Somatic Mutations In Cancer
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)

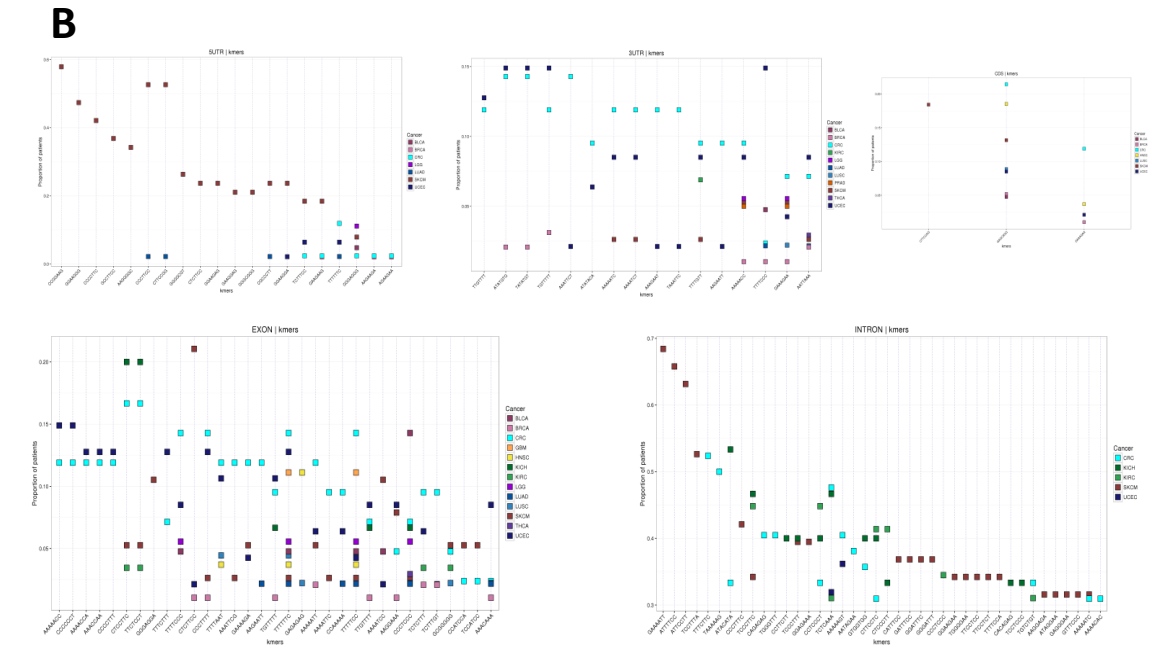

D
Scale
chr3:
2 kb
hg19
134,500,000
134,501,000
134,502,000
134,503,000
134,504,000
134,505,000
PAN505 mutations
CRC:TCGA-A6-2681-01A
CRC:TCGA-AA-3514-01A
CRC:TCGA-AA-3956-01A
CRC:TCGA-AG-4008-01A
CRC:TCGA-AZ-6601-01A
CRC:TCGA-CA-6718-01A
HNSC:TCGA-CV-7099-01A
KICH:TCGA-KL-8325-01A
KICH:TCGA-KL-8341-01A
KICH:TCGA-KN-8434-01A
KICH:TCGA-KO-8417-01A
KIRC:TCGA-B0-5693-01A
KIRC:TCGA-CW-6087-01A
KIRC:TCGA-CZ-4856-01A
PAN505 SMRs
EPHB1:TTCTTTTTT
Basic Gene Annotation Set from GENCODE Version 19
RNU6-117P
RNASSP141
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000467706
ENST00000460895
ENST00000517278
ENST00000364543
COSMIC: Catalogue Of Somatic Mutations In Cancer
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)

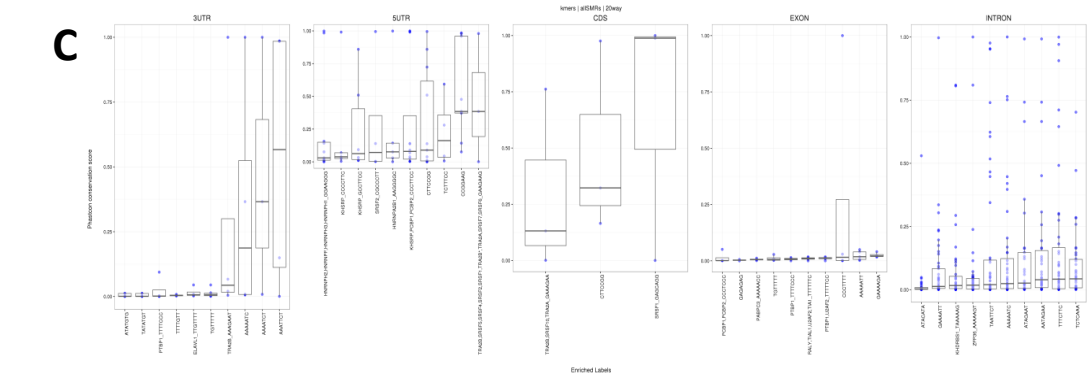

E
Scale
chr17:
10 bases
hg19
47,723,890
47,723,900
47,723,910
PAN505 mutations
UCEC:TCGA-AP-A054-01A
UCEC:TCGA-AX-A05S-01A
UCEC:TCGA-B5-A11H-01A
PAN507 SMRs
PAN507 SMRs
PAN505 SMRs
PAN505 SMRs
RNA-seq junctions
RNA-seq junctions
RefSeq Genes
SPOP/NM_003563
SPOP/NM_001007230
SPOP/NM_001007229
SPOP/NM_001007228
SPOP/NM_001007227
SPOP/NM_001007226
COSMIC: Catalogue Of Somatic Mutations In Cancer
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000393328
ENST00000347630
ENST00000393331
ENST00000504102
ENST00000503676
ENST00000509079
ENST00000505581
ENST00000506399
ENST00000507970
ENST00000514121
ENST00000515508
ENST00000513080
ENST00000504212
ENST00000510476
ENST00000508805


F
Scale
chr18:
10 bases
hg19
60,048,030
60,048,040
60,048,050
PAN505 mutations
KIRC:TCGA-A3-3308-01A
KIRC:TCGA-B0-5693-01A
UCEC:TCGA-AX-A0J1-01A
PAN507 SMRs
PAN507 SMRs
PAN505 SMRs
PAN505 SMRs
RNA-seq junctions
RefSeq Genes
TNFRSF11A/NM_001278268
TNFRSF11A/NM_001270949
TNFRSF11A/NM_001270950
TNFRSF11A/NM_001270951
TNFRSF11A/NM_003839
COSMIC: Catalogue Of Somatic Mutations In Cancer

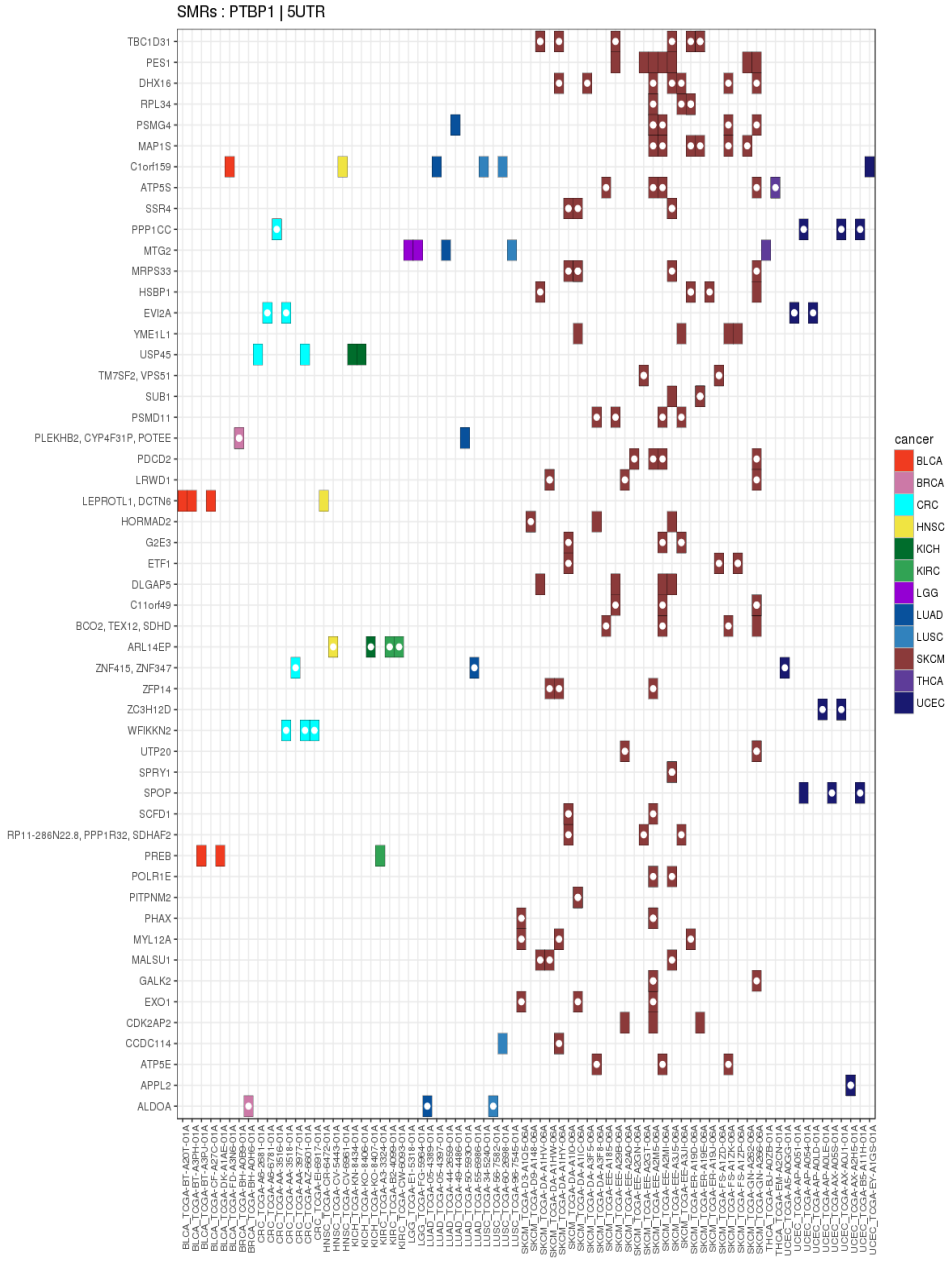
Supplementary Figure 6 Novel significantly mutated regions. (A) Novel 5UTR SMR identified in *DHX16*. **(B)** Example of the novel 3UTR SMR found in *TP53TG3D*. **(C)** Novel EXON SMR found in a non-coding transcript of the gene *DDX17.* **(D)** Novel INTRON SMR found in *EPHB1* from the PAN505 dataset. **(E)** Novel 5UTR mutations in *SPOP* falling on a putative PTP1 binding site. The SMR is in a 5' untraslated region that is annotated only in Gencode/Ensembl. **(F)** Novel INTRON SMR in the gene *TNFRSF11A*, with mutations falling in a putative ZFP36 motif.


A
labels | allSMRs | 20way
3UTR
5UTR
CDS
EXON
INTRON
Phastcon conservation score
Enriched Labels

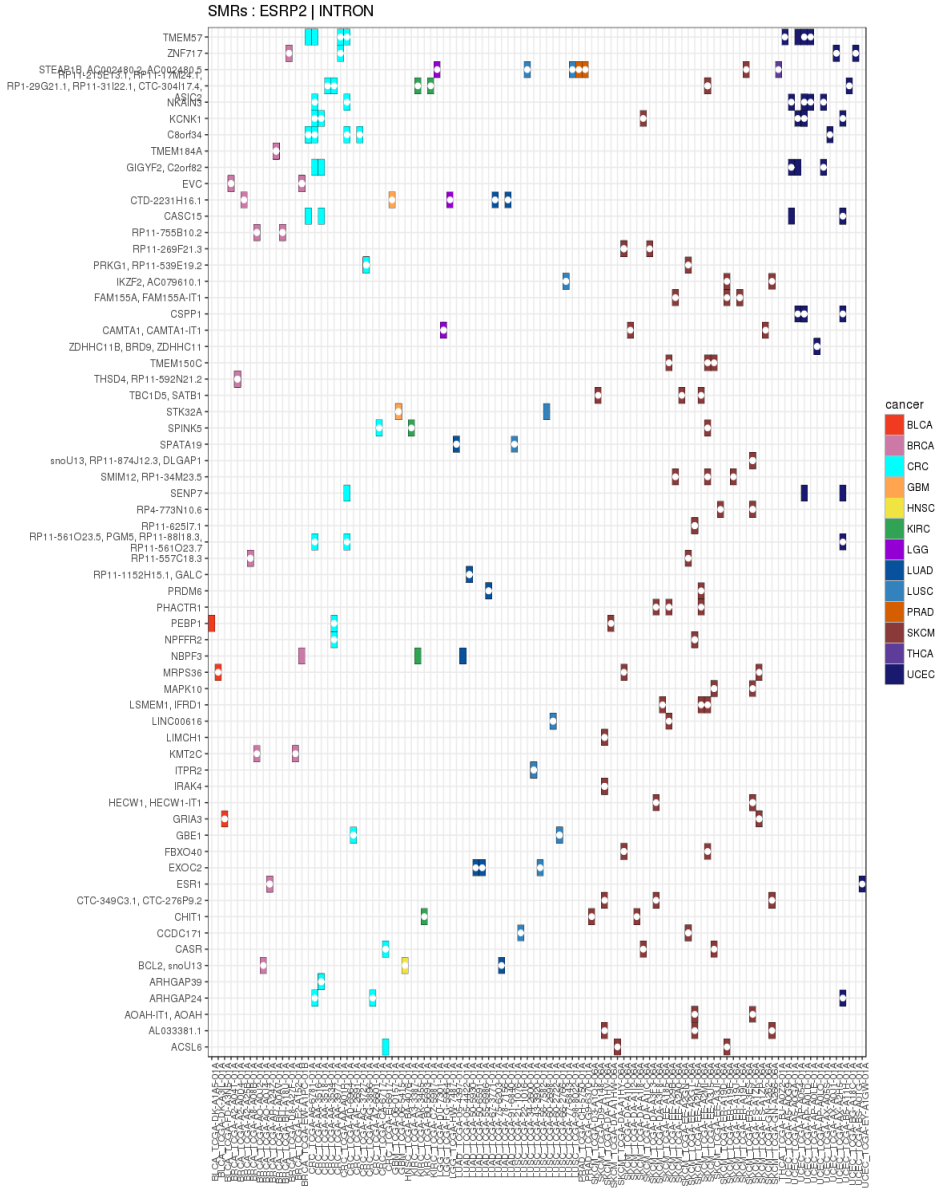

B
5UTR | kmers
3UTR | kmers
CDS | kmers
EXON | kmers
INTRON | kmers
Proportion of patients
kmers
Cancer


C
kmers | allSMRs | 20way
3UTR
5UTR
CDS
EXON
INTRON
Phastcon conservation score
Enriched Labels

Supplementary Figure 7. (A) Phastcons conservation score distributions (y axis) for the SMRs harboring enriched motifs RBP motifs (x axis). SMRs and motifs are separated by region type. The conservation is calculated from the 20-way multiple species alignment from UCSC. **(B)** Top 20 (or all if they are less than 20) enriched k-mers (z>1.96 and k-mer count > 2). For each region type, we give the proportion of patients (y-axis) that show SMRs enriched in each k-mer (x-axis). **(C)** Phastcons conservation score distributions (y axis) for the SMRs harboring enriched k-mer motifs (x axis). SMRs and k-mers are separated by region type. The conservation is calculated from the 20-way multiple species alignment from UCSC.



Supplementary Figure 8. (A)



Supplementary Figure 8. Genes and patients with SMRs that contain enriched motifs. We show matrix plots with the genes (y axis) with SMRs containing the PTBP1 **(A)** and ESRP2 **(B)** motifs. The tumor types are color-coded and patients are given in the x-axis. A white dot inside the color square indicates that the mutation falls inside the motif for that patient.

IV. Discussion

With the opportunity provided by large-scale genomics data for multiple cancer types that became available to the research community, this study started with the ambitious motivation to find patterns of common and specific RNA processing alterations going on in different cancer types. Indeed we observed multi-layer RNA processing alterations pervasive in cancer that could categorize specific tumor types as well as define a common set of alterations such as in the hallmarks of cancer. The RNA binding proteins that are involved in RNA processing were frequently deregulated in all of the 11 tumors studied, more often due to copy number variations and expression changes than mutations. We further studied association between RBPs deregulation with respect to splicing change. Many such splicing changes were observed in cancer driver genes and were also related to various cancer hallmarks. Conversely, many other splicing changes were observed specific to each tumor type and subtype, such as breast tumor of subtype Luminal A and B. This finding suggests that alternative splicing patterns is an important factor to consider for diagnosis as well as for therapeutic purposes along with somatic mutations and epigenetic alterations. We also reported several RBPs that were deregulated specific to each tumor type with an enriched motif binding sites on differential spliced events. It shows their possible specificity for splicing alterations in those tissue types, such as TRA2B, which appears amplified and overexpressed in lung squamous carcinoma. Interestingly, we observed splicing alterations in targets of TRA2B, like the gene CHEK1. One of the key results that arose from this study was the source of origin for the deregulation of RBPs as not all of them could be described through DNA alterations in the gene locus. Further to expand our study and validate this claim we studied one particular RBP, MBNL1, which was frequently deregulated in tumors with an enrichment of its binding sites on differentially alternative spliced events including the mitotic gene NUMA1. We

observed splicing changes related to MBNL1 recapitulated splicing patterns of undifferentiated cells such as in the gene NUMA1. We showed NUMA1 alternative splicing leads to higher proliferation and increased centrosome amplification in normal cells. This study provides multiple distinct as well as common splicing regulatory networks that could be explored further to work out the detail of tumor-specific key modulators for differential splicing.

Our next objective was to explore the open-ended questions raised by the previous work such as the origin of deregulation of RNA binding proteins and previously unexplored components such as whether mutations on the binding sites for these RBPs might be leading to differential splicing or RNA processing alterations. As the binding motifs were found to be enriched both in intron and exons of differentially alternative spliced events, to study the *cis*-components, we decided to study mutations for 13 different cancer patients obtained from whole genome sequencing data. We designed a novel method to categorized relevant cancer mutations in both coding as well non-coding regions. Our goal was to create an unbiased search in entire gene regions as the regulatory sites known for RNA processing binds both in coding as well as non-coding regions. We recovered several known and novel mutation hotspots (SMRs) across genes such as 3UTRs, 5UTRs and introns, which are not often studied. Our method provides advantage in studying potentially relevant deep intronic mutations that were not described previously. We further observed these SMRs show enrichment in RNA binding motifs mostly in intronic and UTR regions and often certain positions of these motifs are more mutated than others. Further, many of the transcripts with these SMRs were found to be changing significantly in expression in the patients where the mutations on the motifs were observed. This suggests there

is still an unexplored territory with respect to somatic mutations and RNA processing alterations in cancer. This could be due to scarcity of whole genome sequencing data that provides limitations to extract statistically relevant cases. It is still unclear whether these mutations, especially in non-coding regions provide selective growth advantage to cells, such as in cancer driver genes. Most likely they are passenger mutations providing sustainability to tumor cells. However, this study shows the importance of expanding the search for mutations to non-coding regions for their potential to explain relevant phenotypes in multiple cancers.

V. Conclusions

The main contributions from the work presented in this thesis could be summarized as follows:

RNA processing alterations through trans components

- RNA binding proteins (RBPs) are frequently deregulated in different cancers and often their expression patterns categorize tumor types.
- Much of the deregulation of RBPs is due to copy number alterations and expression changes.
- Deregulation in RBPs due to mutations or differential expression are associated with abnormal splicing patterns in cancer.
- Differential splicing shows common and specific patterns in different cancer types.
- Differential splicing events have enrichment of deregulated RBP motifs in multiple cancers.
- There are specific splicing regulatory modules (networks) altered in different cancers.
- We validated that MBNL1 protein contributes to cell proliferation and genome instability through the splicing regulation of NUMA1.

RNA processing alterations through cis components

- There are significantly mutated regions (SMRs) at gene loci consistent across several cancer types.
- These SMRs are highly prevalent in non-coding regions such as introns and non-coding exons.
- These SMRs are enriched for multiple putative binding sites for RNA binding proteins (RBPs), hence they probably affect RNA processing.
- Some positions of the RBP motifs are mutated in cancer more frequently than expected..
- Significantly enriched RBPs motifs on SMRs show impact on RNA processing and expression using RNA sequencing data from the same patients.

VI. Other scientific work

1) I performed the analyses of RNA-seq data in two cell lines to study the relation between AGO1-related chromatin states and alternative splicing:
Alló M, Agirre E, Bessonov S, Bertucci P, Gómez Acuña L, Buggiano V, Bellora N, Singh B, Petrillo E, Blaustein M, Miñana B, Dujardin G, Pozzi B, Pelisch F, Bechara E, Agafonov DE, Srebrow A, Lührmann R, Valcárcel J, Eyras E, Kornblihtt AR. (2014) Argonaute-1 binds transcriptional enhancers and controls constitutive and alternative splicing in human cells. PNAS 4;111(44):15622-9. PMID: 25313066

2) I performed the analyses of RNA-seq data in two cell lines to study the relation between of different chromatin states and alternative splicing:

González-Vallinas J, Pagès A, Singh B, Eyras E. (2015) A semi-supervised approach uncovers thousands of intragenic enhancers differentially activated in human cells. BMC Genomics. 16:523. doi: 10.1186/s12864-015-1704-0. PMID: 26169177

3) In a collaboration with W. Ritchie's group in the development of a method to detect intron retention (IRFinder), I used motif enrichment tool MoSEA (https://github.com/comprna/MoSEA) to identify the possible RBP factors that control intron retention:

Middleton R, Gao D, Thomas A, Singh A, Au A, Wong JJL, Bomane A, Cosson B, Eyras E, Rasko JEJ, Ritchie R. IRFinder: Assessing the impact of intron retention on mammalian gene expression. Genome Biology, in press.

4) I used MoSEA (https://github.com/comprna/MoSEA) to identify the possible RBP factors that control the differential splicing of events during differentiation of iPS cells into bipolar neurons. This analysis confirmed that CELF, RBFOX and SRRM proteins are

important for neuronal differentiation. Additionally, ESRP proteins also may play a role in the maintainance of the splicing pattern in the differentiated state:

Juan C Entizne, JL Trincado, G Hysenaj, B Singh, M Skalic, DJ Elliott, E Eyras Fast and accurate differential splicing analysis across multiple conditions with replicates (2016). Biorxiv doi: https://doi.org/10.1101/086876

5) I wrote a review about the role of alternative splicing in cancer together with my supervisor Eduardo Eyras:

Singh B, Eyras E. (2016) The role of alternative splicing in cancer. Transcription. 22:e1268245. doi: 10.1080/21541264.2016.1268245. PMID: 28005460.

VII. References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73.

Alamancos GP, Agirre E, Eyras E. 2014. Methods to Study Splicing from High-Throughput RNA Sequencing Data. pp. 357–397.

Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. 2015. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21**: 1521–31. http://www.ncbi.nlm.nih.gov/pubmed/26179515 (Accessed February 13, 2017).

Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, et al. 2016b. Mutational signatures associated with tobacco smoking in human cancer. *Science (80-)* **354**: 618–622.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013b. Signatures of mutational processes in human cancer. *Nature* **500**: 415–421.

Alexandrov LB, Stratton MR. 2014. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24**: 52–60. http://www.ncbi.nlm.nih.gov/pubmed/24657537 (Accessed February 13, 2017).

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the

sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–8. http://www.ncbi.nlm.nih.gov/pubmed/26213851.

Allo M, Agirre E, Bessonov S, Bertucci P, Gomez Acuna L, Buggiano V, Bellora N, Singh B, Petrillo E, Blaustein M, et al. 2014. Argonaute-1 binds transcriptional enhancers and controls constitutive and alternative splicing in human cells. *Proc Natl Acad Sci U S A* **111**: 15622–15629.

Alsafadi S, Houy A, Battistella A, Popova T, Wassef M, Henry E, Tirode F, Constantinou A, Piperno-Neumann S, Roman-Roman S, et al. 2016. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun* **7**: 10615. http://www.ncbi.nlm.nih.gov/pubmed/26842708.

Althammer S, González-Vallinas J, Ballaré C, Beato M, Eyras E. 2011. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics* **27**: 3333–40. http://www.ncbi.nlm.nih.gov/pubmed/21994224 (Accessed February 13, 2017).

Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, Raimer A, Sun S, Jensen MA, Hua Y, et al. 2015. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol Cell* **60**: 105–17. http://www.ncbi.nlm.nih.gov/pubmed/26431027.

Anczuków O, Krainer AR. 2016. Splicing-factor alterations in cancers. *RNA* **22**: 1285–301. http://www.ncbi.nlm.nih.gov/pubmed/27530828.

Anczuków O, Rosenberg AZ, Akerman M, Das S, Zhan L, Karni R, Muthuswamy SK, Krainer AR. 2012. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary

epithelial cell transformation. *Nat Struct Mol Biol* **19**: 220–228. http://www.ncbi.nlm.nih.gov/pubmed/22245967 (Accessed February 13, 2017).

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106 (Accessed February 13, 2017).

Archive HGI. *About the Human Genome Project*. U.S. Department of Energy & Human Genome Project program http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml (Accessed February 13, 2017).

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–W208. http://www.ncbi.nlm.nih.gov/pubmed/19458158 (Accessed February 13, 2017).

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Second Int Conf Intell Syst Mol Biol* **2**: 28–36.

Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, et al. 2012. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol Cell* **46**: 674–90.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**: 53–59. http://www.ncbi.nlm.nih.gov/pubmed/20445623 (Accessed February 12, 2017).

Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. 2013.

RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Mol Cell* **52**.

Begley S. 2008. Rethinking the War on Cancer. http://europe.newsweek.com/rethinking-war-cancer-88941?rm=eu (Accessed February 13, 2017).

Best A, James K, Dalgliesh C, Hong E, Kheirolahi-Kouhestani M, Curk T, Xu Y, Danilenko M, Hussain R, Keavney B, et al. 2014. Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. *Nat Commun* **5**: 4760. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4175592&tool=pmcentrez&rendertype=abstract.

Bonomi S, di Matteo A, Buratti E, Cabianca DS, Baralle FE, Ghigna C, Biamonti G. 2013. HnRNP A1 controls a splicing regulatory circuit promoting mesenchymal-to-epithelial transition. *Nucleic Acids Res* **41**: 8665–79. http://www.ncbi.nlm.nih.gov/pubmed/23863836.

Bordonaro M. 2013. Crosstalk between Wnt signaling and RNA processing in colorectal cancer. *J Cancer* **4**: 96–103.

Bottini S, Hamouda-Tekaya N, Tanasa B, Zaragosi L-E, Grandjean V, Repetto E, Trabucchi M. 2017. From benchmarking HITS-CLIP peak detection programs to a new method for identification of miRNA-binding sites from Ago2-CLIP data. *Nucleic Acids Res* gkx007.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Brooks AN, Choi PS, de Waal L, Sharifnia T, Imielinski M, Saksena G, Pedamallu CS, Sivachenko A, Rosenberg M, Chmielecki J, et al. 2014. A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered

splicing events. *PLoS One* **9**: e87361. http://www.ncbi.nlm.nih.gov/pubmed/24498085.

Brosseau J-P, Lucier J-F, Nwilati H, Thibault P, Garneau D, Gendron D, Durand M, Couture S, Lapointe E, Prinos P, et al. 2014. Tumor microenvironment-associated modifications of alternative splicing. *RNA* **20**: 189–201. http://www.ncbi.nlm.nih.gov/pubmed/24335142.

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Mol Cell* **46**: 871–883.

Cancer Genome Atlas Research Network. 2015. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**: 1011–25.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**: 285–98. http://www.ncbi.nlm.nih.gov/pubmed/11967553.

Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. 2006. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet* **38**: 1043–1048.

Cascino I, Fiucci G, Papoff G, Ruberti G. 1995. Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. *J Immunol* **154**.

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, et al. 2012. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149**: 1393–1406.

Clauset A, Newman M, Moore C. 2004. Finding community structure in very large networks. *Phys Rev E* **70**: 66111.

Conrad T, Albrecht A-S, de Melo Costa VR, Sauer S, Meierhofer D, Ørom UA. 2016. Serial interactome capture of the human cell nucleus. *Nat Commun* **7**: 11212. http://www.ncbi.nlm.nih.gov/pubmed/27040163.

Corvelo A, Hallegger M, Smith CWJ, Eyras E. 2010. Genome-Wide Association between Branch Point Properties and Alternative Splicing ed. I.M. Meyer. *PLoS Comput Biol* **6**: e1001016.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2015. *Nucleic Acids Res* **43**: 662–669.

Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, Bailey SL, Bhavsar EB, Chan B, Colla S, et al. 2015. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep* **13**: 1033–45. http://www.ncbi.nlm.nih.gov/pubmed/26565915.

Das S, Anczuków O, Akerman M, Krainer AR. 2012. Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. *Cell Rep* **1**: 110–7. http://www.ncbi.nlm.nih.gov/pubmed/22545246.

Daubner GM, Cléry A, Allain FH-T. 2013. RRM-RNA recognition: NMR or crystallography…and new findings. *Curr Opin Struct Biol* **23**: 100–8. http://www.ncbi.nlm.nih.gov/pubmed/23253355.

David CJ, Chen M, Assanah M, Canoll P, Manley JL. 2010. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature* **463**: 364–8. http://www.ncbi.nlm.nih.gov/pubmed/20010808.

David CJ, Manley JL. 2010. Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged. *Genes Dev* **24**: 2343–2364.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**: 1589–98. http://www.ncbi.nlm.nih.gov/pubmed/22759861 (Accessed February 13, 2017).

Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, et al. 2016. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol Med* **8**: 442–57. http://www.ncbi.nlm.nih.gov/pubmed/26992833.

Ding L, Wendl MC, McMichael JF, Raphael BJ. 2014. Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* **15**: 556–70. http://www.ncbi.nlm.nih.gov/pubmed/25001846 (Accessed February 13, 2017).

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dorman SN, Viner C, Rogan PK. 2014. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci Rep* **4**: 7063.

Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433–3434.

Dvinge H, Bradley RK, Kim E, Abdel-Wahab O, Bradley RK. 2015. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med* **7**: 45. http://www.ncbi.nlm.nih.gov/pubmed/26113877.

Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. 2016. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer* **16**: 413–30.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–7.

Ellis JD, Barrios-Rodiles M, ??olak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, et al. 2012a. Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Mol Cell* **46**: 884–892.

Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC, et al. 2012b. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**: 353–360.

Entizne JC, Trincado JL, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. 2016. Fast and accurate differential splicing analysis across multiple conditions with replicates.

Eperon IC, Makarova O V, Mayeda A, Munroe SH, Cáceres JF, Hayward DG, Krainer AR. 2000. Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol* **20**: 8303–18. http://www.ncbi.nlm.nih.gov/pubmed/11046128.

Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**: 1007–13.

Fei DL, Motowski H, Chatrikhi R, Prasad S, Yu J, Gao S, Kielkopf CL, Bradley RK, Varmus H. 2016. Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival. *PLoS Genet* **12**: e1006384. http://www.ncbi.nlm.nih.gov/pubmed/27776121.

Frampton GM, Ali SM, Rosenzweig M, Chmielecki J, Lu X, Bauer TM, Akimov M, Bufill JA, Lee C, Jentz D, et al. 2015. Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors. *Cancer Discov* **5**: 850–860.

Fredriksson NJ, Ny L, Nilsson JA, Larsson E. 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**: 1–7. http://dx.doi.org/10.1038/ng.3141.

Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**: 432–41.

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

Furney SJ, Pedersen M, Gentien D, Dumont AG, Rapinat A, Desjardins L, Turajlic S, Piperno-Neumann S, de la Grange P, Roman-Roman S, et al. 2013. SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov* **3**: 1122–9. http://www.ncbi.nlm.nih.gov/pubmed/23861464.

Garraway LA, Lander ES. 2013. Lessons from the Cancer Genome. *Cell* **153**: 17–37. http://www.ncbi.nlm.nih.gov/pubmed/23540688 (Accessed February 13, 2017).

Gehman LT, Stoilov P, Maguire J, Damianov A, Lin C-H, Shiue L, Ares M, Mody I, Black DL. 2011. The splicing regulator Rbfox1

(A2BP1) controls neuronal excitation in the mammalian brain. *Nat Genet* **43**: 706–11.

Ghigna C, Giordano S, Shen H, Benvenuto F, Castiglioni F, Comoglio PM, Green MR, Riva S, Biamonti G. 2005. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol Cell* **20**: 881–90. http://www.ncbi.nlm.nih.gov/pubmed/16364913.

Giudice G, Sánchez-Cabo F, Torroja C, Lara-Pezzi E. 2016. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**.

Goehe RW, Shultz JC, Murudkar C, Usanovic S, Lamour NF, Massey DH, Zhang L, Camidge DR, Shay JW, Minna JD, et al. 2010. hnRNP L regulates the tumorigenic capacity of lung cancer xenografts in mice via caspase-9 pre-mRNA processing. *J Clin Invest* **120**: 3923–3939.

Golan-Gerstl R, Cohen M, Shilo A, Suh SS, Bakàcs A, Coppola L, Karni R. 2011. Splicing factor hnRNP A2/B1 regulates tumor suppressor gene splicing and is an oncogenic driver in glioblastoma. *Cancer Res* **71**: 4464–4472.

Goldstein LD, Lee J, Gnad F, Klijn C, Schaub A, Reeder J, Daemen A, Bakalarski CE, Holcomb T, Shames DS, et al. 2016. Recurrent Loss of NFE2L2 Exon 2 Is a Mechanism for Nrf2 Pathway Activation in Human Cancers. *Cell Rep*.

Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* **14**: R70. http://genomebiology.com/2013/14/7/R70.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G.

2006. Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. *Mol Cell* **22**: 769–781. http://www.ncbi.nlm.nih.gov/pubmed/16793546 (Accessed February 12, 2017).

Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, Maher CA, Fulton R, Fulton L, Wallis J, et al. 2012. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**: 1121–34. http://www.ncbi.nlm.nih.gov/pubmed/22980976.

Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC, Brenner SE. 2003. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* **19 Suppl 1**: i118-21.

Grosso AR, Carmo-Fonseca M. 2014. The Potential of Targeting Splicing for Cancer Therapy. In *Nuclear Signaling Pathways and Targeting Transcription in Cancer* (ed. R. Kumar), pp. 313–336.

Grosso AR, Martins S, Carmo-Fonseca M. 2008. The emerging role of splicing factors in cancer. *EMBO Rep* **9**: 1087–93. http://www.ncbi.nlm.nih.gov/pubmed/18846105.

Gutschner T, Hämmerle M, Diederichs S. 2013. MALAT1 -- a paradigm for long noncoding RNA function in cancer. *J Mol Med (Berl)* **91**: 791–801. http://www.ncbi.nlm.nih.gov/pubmed/23529762.

Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**: 333–46. http://www.ncbi.nlm.nih.gov/pubmed/25589248.

Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**: 241–245.

Hanahan D, Weinberg RA. 2011. Hallmarks of Cancer: The Next Generation. *Cell* **144**: 646–674.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100**: 57–70.

Hartung A-M, Swensen J, Uriz IE, Lapin M, Kristjansdottir K, Petersen USS, Bang JM V., Guerra B, Andersen HS, Dobrowolski SF, et al. 2016. The Splicing Efficiency of Activating HRAS Mutations Can Determine Costello Syndrome Phenotype and Frequency in Cancer ed. A. Goriely. *PLOS Genet* **12**: e1006039. http://www.ncbi.nlm.nih.gov/pubmed/27195699 (Accessed February 13, 2017).

Havens MA, Hastings ML. 2016. Splice-switching antisense oligonucleotides as therapeutic drugs. *Nucleic Acids Res* **44**: 6549–63.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, et al. 2014. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* **158**: 929–944.

Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, et al. 2013b. TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science (80-)* **339**: 959–961.

Hsu TY-T, Simon LM, Neill NJ, Marcotte R, Sayad A, Bland CS,

Echeverria G V, Sun T, Kurley SJ, Tyagi S, et al. 2015. The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature* **525**: 384–8. http://www.nature.com/doifinder/10.1038/nature14985%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26331541.

Huang C-S, Shen C-Y, Wang H-W, Wu P-E, Cheng C-W. 2007. Increased expression of SRp40 affecting CD44 splicing is associated with the clinical outcome of lymph node metastasis in human breast cancer. *Clin Chim Acta* **384**: 69–74.

Huang FW, Hodis E, Xu MJ, Kryukov G V., Chin L, Garraway LA. 2013. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**: 957–9. http://www.ncbi.nlm.nih.gov/pubmed/23348506.

Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, et al. 2012. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**: 1107–1120.

Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallières M, Tapial J, Raj B, O'Hanlon D, et al. 2014. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell* **159**: 1511–1523. http://www.ncbi.nlm.nih.gov/pubmed/25525873 (Accessed February 13, 2017).

Jangi M, Sharp PA. 2014. Building Robust Transcriptomes with Master Splicing Factors. *Cell* **159**: 487–498.

Jensen MA, Wilkinson JE, Krainer AR. 2014. Splicing factor SRSF6 promotes hyperplasia of sensitized skin. *Nat Struct Mol Biol* **21**: 189–197.

Jia R, Li C, McCoy JP, Deng CX, Zheng ZM. 2010. SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. *Int J Biol Sci* **6**: 806–826.

Jones S, Chen W -d., Parmigiani G, Diehl F, Beerenwinkel N, Antal T, Traulsen A, Nowak MA, Siegel C, Velculescu VE, et al. 2008. Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci* **105**: 4283–4288. http://www.ncbi.nlm.nih.gov/pubmed/18337506 (Accessed February 13, 2017).

Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. 2016. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat Commun* **7**: 11558.

Jung H, Lee D, Lee J, Park D, Kim YJ, Park W-Y, Hong D, Park PJ, Lee E. 2015. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47**: 1242–1248. http://dx.doi.org/10.1038/ng.3414.

Kammerer S, Roth RB, Hoyal CR, Reneland R, Marnellos G, Kiechle M, Schwarz-Boeger U, Griffiths LR, Ebner F, Rehbock J, et al. 2005. Association of the NuMA region on chromosome 11q13 with breast cancer susceptibility. *Proc Natl Acad Sci U S A* **102**: 2004–2009.

Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol* **14**: 185–93. http://www.ncbi.nlm.nih.gov/pubmed/17310252.

Katainen R, Dave K, Pitkänen E, Palin K, Kivioja T, Välimäki N, Gylfe AE, Ristolainen H, Hänninen UA, Cajuso T, et al. 2015b. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**: 818–821.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–15. http://www.ncbi.nlm.nih.gov/pubmed/21057496.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* **21**: 1360–74.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**: 345–355. http://www.nature.com/nrg/journal/v11/n5/box/nrg2776_BX1.html (Accessed February 12, 2017).

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**: 1235587. http://www.ncbi.nlm.nih.gov/pubmed/24092746.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013a. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36. http://www.ncbi.nlm.nih.gov/pubmed/23618408.

Kim E, Ilagan JO, Liang Y, Daubner GM, Lee SC-W, Ramakrishnan A, Li Y, Chung YR, Micol J-B, Murphy ME, et al. 2015. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell* **27**: 617–30. http://www.ncbi.nlm.nih.gov/pubmed/25965569.

Kim KK, Nam J, Mukouyama Y-S, Kawamoto S. 2013b. Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development. *J Cell Biol* **200**: 443–58.

Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Gervais-Bird J, Madden R, Paquet ??ric R., Koh C, Venables JP, Prinos P, et al. 2008. Multiple alternative splicing markers for ovarian cancer. *Cancer Res* **68**: 657–663.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**: 568–576. http://www.ncbi.nlm.nih.gov/pubmed/22300766 (Accessed February 13, 2017).

Koh CM, Bezzi M, Low DHP, Ang WX, Teo SX, Gay FPH, Al-Haddawi M, Tan SY, Osato M, Sabò A, et al. 2015. MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature* **523**: 96–100. http://www.ncbi.nlm.nih.gov/pubmed/25970242.

Kong-Beltran M, Seshagiri S, Zha J, Zhu W, Bhawe K, Mendoza N, Holcomb T, Pujara K, Stinson J, Fu L, et al. 2006. Somatic mutations lead to an oncogenic deletion of Met in lung cancer. *Cancer Res* **66**: 283–289.

Kotak S, Busso C, Gönczy P. 2013. NuMA phosphorylation by CDK1 couples mitotic progression with cortical dynein function. *EMBO J* **32**: 2517–2529.

Kwon SC, Yi H, Eichelbaum K, Föhr S, Fischer B, You KT, Castello A, Krijgsveld J, Hentze MW, Kim VN. 2013. The RNA-binding protein repertoire of embryonic stem cells. *Nat Struct Mol Biol* **20**: 1122–30.

Ladomery M, Ladomery M. 2013. Aberrant Alternative Splicing Is Another Hallmark of Cancer. *Int J Cell Biol* **2013**: 1–6. http://www.hindawi.com/journals/ijcb/2013/463786/.

Lai WS, Kennington EA, Blackshear PJ. 2002. Interactions of CCCH zinc finger proteins with mRNA: non-binding tristetraprolin mutants exert an inhibitory effect on degradation of AU-rich element-containing mRNAs. *J Biol Chem* **277**: 9606–13.

Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. 2014. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* **54**: 887–900. http://www.ncbi.nlm.nih.gov/pubmed/24837674.

Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* **470**: 187–197. http://www.nature.com/doifinder/10.1038/nature09792 (Accessed February 13, 2017).

Lanzós A, Carlevaro-Fita J, Mularoni L, Reverter F, Palumbo E, Guigó R, Johnson R. 2017. Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Sci Rep* **7**: 41544. http://www.ncbi.nlm.nih.gov/pubmed/28128360.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**: R29.

Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013b. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214–218.

Lee SC-W, Abdel-Wahab O. 2016. Therapeutic targeting of splicing in cancer. *Nat Med* **22**: 976–86. http://www.ncbi.nlm.nih.gov/pubmed/27603132.

Lee Y, Rio DC. 2015. Mechanisms and Regulation of Alternative Pre-

mRNA Splicing. *Annu Rev Biochem* **84**: 291–323. http://www.annualreviews.org/doi/10.1146/annurev-biochem-060614-034316 (Accessed February 12, 2017).

Li J, Witten DM, Johnstone IM, Tibshirani R. 2012. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**: 523–538. http://www.ncbi.nlm.nih.gov/pubmed/22003245 (Accessed February 13, 2017).

Li X, Quon G, Lipshitz HD, Morris Q. 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA* **16**: 1096–107. http://www.ncbi.nlm.nih.gov/pubmed/20418358.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**: 417–425.

Liu L, De S, Michor F. 2013. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun* **4**: 1502. http://www.ncbi.nlm.nih.gov/pubmed/23422670.

Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. 2015. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* **43**: 8123–34. http://www.ncbi.nlm.nih.gov/pubmed/26304545.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15–27. http://www.ncbi.nlm.nih.gov/pubmed/19061484.

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al. 2013. Rbfox proteins

regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**: 1434–42.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. http://www.ncbi.nlm.nih.gov/pubmed/25516281 (Accessed February 13, 2017).

Lu Z, Huang Q, Park JW, Shen S, Lin L, Tokheim CJ, Henry MD, Xing Y. 2015. Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. *Mol Cancer Res* **13**: 305–18. http://www.ncbi.nlm.nih.gov/pubmed/25274489.

Macias S, Plass M, Stajuda A, Michlewski G, Eyras E, Cáceres JF. 2012. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nat Struct Mol Biol* **19**: 760–766. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3442229&tool=pmcentrez&rendertype=abstract.

Maguire SL, Leonidou A, Wai P, Marchiò C, Ng CK, Sapino A, Salomon A-V, Reis-Filho JS, Weigelt B, Natrajan RC. 2015. SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J Pathol* **235**: 571–580.

Martincorena I, Campbell PJ. 2015. Somatic mutation in cancer and normal cells. *Science (80-)* **349**: 1483–1489. http://www.ncbi.nlm.nih.gov/pubmed/26404825 (Accessed February 13, 2017).

Maslon MM, Heras SR, Bellora N, Eyras E, Cáceres JF. 2014. The translational landscape of the splicing factor SRSF1 and its role in mitosis. *Elife* **2014**.

Matera AG, Wang Z. 2014. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* **15**: 108–121.

http://www.nature.com/doifinder/10.1038/nrm3742 (Accessed February 12, 2017).

McFarland CD, Mirny LA, Korolev KS. 2014. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc Natl Acad Sci U S A* **111**: 15138–43. http://www.ncbi.nlm.nih.gov/pubmed/25277973 (Accessed February 13, 2017).

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303. http://www.ncbi.nlm.nih.gov/pubmed/20644199 (Accessed February 13, 2017).

Meienberg J, Bruggmann R, Oexle K, Matyas G. 2016. Clinical sequencing: is WGS the better WES? *Hum Genet* **135**: 359–62. http://www.ncbi.nlm.nih.gov/pubmed/26742503 (Accessed February 13, 2017).

Melton C, Reuter JA, Spacek D V, Snyder M. 2015. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet* **47**: 710–6. http://www.ncbi.nlm.nih.gov/pubmed/26053494.

Merlo LMF, Pepper JW, Reid BJ, Maley CC. 2006. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**: 924–935. http://www.nature.com/doifinder/10.1038/nrc2013 (Accessed February 13, 2017).

Meyuhas O. 2000. Synthesis of the translational apparatus is regulated at the translational level. *Eur J Biochem* **267**: 6321–30.

Michlewski G, Guil S, Semple CA, Cáceres JF. 2008. Posttranscriptional regulation of miRNAs harboring conserved

terminal loops. *Mol Cell* **32**: 383–93. http://www.ncbi.nlm.nih.gov/pubmed/18995836.

Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, Blencowe BJ. 2011. Global Profiling and Molecular Characterization of Alternative Splicing Events Misregulated in Lung Cancer. *Mol Cell Biol* **31**: 138–150. http://mcb.asm.org/cgi/doi/10.1128/MCB.00709-10.

Moon H, Cho S, Loh TJ, Oh HK, Jang HN, Zhou J, Kwon Y-S, Liao DJ, Jun Y, Eom S, et al. 2014. SRSF2 promotes splicing and transcription of exon 11 included isoform in Ron proto-oncogene. *Biochim Biophys Acta* **1839**: 1132–1140.

Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. 2016. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**: 128. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0994-0.

Naftelberg S, Schor IE, Ast G, Kornblihtt AR. 2015. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**: 165–98. http://www.ncbi.nlm.nih.gov/pubmed/26034889.

Nowicka M, Robinson M. 2016. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*.

Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL, et al. 2005. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309**: 2054–7. http://www.ncbi.nlm.nih.gov/pubmed/16179478.

Oltean S, Bates DO. 2013. Hallmarks of alternative splicing in cancer. *Oncogene* 1–8.

Paik PK, Drilon A, Fan P-D, Yu H, Rekhtman N, Ginsberg MS, Borsu L, Schultz N, Berger MF, Rudin CM, et al. 2015. Response to MET Inhibitors in Patients with Stage IV Lung Adenocarcinomas Harboring MET Mutations Causing Exon 14 Skipping. *Cancer Discov* **5**: 842–849.

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC-H, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. 2011. The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**: 435–9.

Patro R, Duggal G, Kingsford C. 2015a. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*.

Patro R, Mount SM, Kingsford C. 2014b. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32**: 462–464.

Pichon X, Wilson LA, Stoneley M, Bastide A, King HA, Somers J, Willis AEE. 2012. RNA binding protein/RNA element interactions and the control of translation. *Curr Protein Pept Sci* **13**: 294–304.

Piraino SW, Furney SJ. 2016. Beyond the exome: the role of non-coding somatic mutations in cancer. *Ann Oncol Off J Eur Soc Med Oncol* **27**: 240–8. http://www.ncbi.nlm.nih.gov/pubmed/26598542.

Piraino SW, Furney SJ. 2017. Identification of coding and non-coding mutational hotspots in cancer genomes. *BMC Genomics* **18**: 17. http://www.ncbi.nlm.nih.gov/pubmed/28056774.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray

SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, et al. 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**: 191–6. http://www.ncbi.nlm.nih.gov/pubmed/20016485.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C, et al. 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**: 184–90. http://www.ncbi.nlm.nih.gov/pubmed/20016488.

Poulikakos PI, Persaud Y, Janakiraman M, Kong X, Ng C, Moriceau G, Shi H, Atefi M, Titz B, Gabay MT, et al. 2011. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature* **480**: 387–90.

Raj B, Irimia M, Braunschweig U, Sterne-Weiler T, O'Hanlon D, Lin ZY, Chen GI, Easton LE, Ule J, Gingras AC, et al. 2014. A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol Cell* **56**: 90–103.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–7. http://www.ncbi.nlm.nih.gov/pubmed/23846655.

Reddy EP, Reynolds RK, Santos E, Barbacid M. 1982. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**: 149–152. http://www.nature.com/doifinder/10.1038/300149a0 (Accessed February 13, 2017).

Resch U, Cuapio A, Sturtzel C, Hofer E, de Martin R, Holper-Schichl

YM. 2014. Polyubiquitinated tristetraprolin protects from TNF-induced, caspase-mediated apoptosis. *J Biol Chem* **289**: 25088–100.

Rissland OS. 2017. The organization and regulation of mRNA-protein complexes. *Wiley Interdiscip Rev RNA* **8**. http://www.ncbi.nlm.nih.gov/pubmed/27324829.

Rivas E, Clements J, Eddy SR. 2017. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods* **14**: 45–48. http://www.ncbi.nlm.nih.gov/pubmed/27819659.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140. http://www.ncbi.nlm.nih.gov/pubmed/19910308 (Accessed February 13, 2017).

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25 (Accessed February 13, 2017).

Rodor J, Pan Q, Blencowe BJ, Eyras E, Cáceres JF. 2016. The RNA-binding profile of Acinus, a peripheral component of the exon junction complex, reveals its role in splicing regulation. *RNA* **22**: 1411–26.

Salton M, Kasprzak WK, Voss T, Shapiro BA, Poulikakos PI, Misteli T. 2015. Inhibition of vemurafenib-resistant melanoma by interference with pre-mRNA splicing. *Nat Commun* **6**: 7103.

Salton M, Misteli T. 2016. Small Molecule Modulators of Pre-mRNA

Splicing in Cancer Therapy. *Trends Mol Med* **22**: 28–37. http://www.ncbi.nlm.nih.gov/pubmed/26700537.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463–7.

Sawicka K, Bushell M, Spriggs KA, Willis AE. 2008. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* **36**: 641–7.

Schroeder MP, Rubio-Perez C, Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2014. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics* **30**: i549–i555. http://www.ncbi.nlm.nih.gov/pubmed/25161246 (Accessed February 13, 2017).

Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, Valcárcel J, Eyras E. 2016. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res* **26**: 732–44.

Sebestyén E, Zawisza M, Eyras E. 2015. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res* **43**: 1345–1356.

Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, Zhou J, Qiu J, Jiang L, Li H, et al. 2014. Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol* **21**: 997–1005.

Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, Burge CB, Gertler FB. 2011. An emt-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* **7**.

Sharp PA, Burge CB. 1997. Classification of Introns: U2-Type or U12-Type. *Cell* **91**: 875–879.

Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**: E5593-601. http://www.ncbi.nlm.nih.gov/pubmed/25480548.

Shen S, Wang Y, Wang C, Wu YN, Xing Y. 2016. SURVIV for survival analysis of mRNA isoform variation. *Nat Commun* **7**: 11548. http://www.ncbi.nlm.nih.gov/pubmed/27279334.

Shih C, Weinberg RA. 1982. Isolation of a Transforming Sequence from a Human Bladder Carcinoma Cell Line. *Cell* **29**: 161–69.

Shkreta L, Toutant J, Durand M, Manley JL, Chabot B. 2016. SRSF10 Connects DNA Damage to the Alternative Splicing of Transcripts Encoding Apoptosis, Cell-Cycle Control, and DNA Repair Factors. *Cell Rep* **17**: 1990–2003.

Siegel RL, Miller KD, Jemal A. 2017b. Cancer statistics, 2017. *CA Cancer J Clin* **67**: 7–30.

Singh B, Eyras E. 2016. The role of alternative splicing in cancer. *Transcription* e1268245. http://www.ncbi.nlm.nih.gov/pubmed/28005460.

Smyth GK. 2005. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397–420, Springer New York.

Sotillo E, Barrett DM, Black KL, Bagashev A, Oldridge D, Wu G, Sussman R, Lanauze C, Ruella M, Gazzara MR, et al. 2015. Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. *Cancer*

Discov **5**: 1282–1295.

Srebrow A, Kornblihtt AR. 2006. The connection between splicing and cancer. *J Cell Sci* **119**.

Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. 2006. Inference of Splicing Regulatory Activities by Sequence Neighborhood Analysis. *PLoS Genet* **2**: e191. http://www.ncbi.nlm.nih.gov/pubmed/17121466 (Accessed February 12, 2017).

Sterne-Weiler T, Sanford JR. 2014. Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome Biol* **15**: 201. http://www.ncbi.nlm.nih.gov/pubmed/24456648.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* **458**: 719–724.

Sundararaman B, Zhan L, Blue SM, Stanton R, Elkins K, Olson S, Wei X, Van Nostrand EL, Pratt GA, Huelga SC, et al. 2016. Resources for the Comprehensive Discovery of Functional RNA Elements. *Mol Cell* **61**: 903–13. http://www.ncbi.nlm.nih.gov/pubmed/26990993.

Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**: 1324–35. http://www.ncbi.nlm.nih.gov/pubmed/24630730.

Suvorova ES, Croken M, Kratzer S, Ting LM, de Felipe MC, Balu B, Markillie ML, Weiss LM, Kim K, White MW. 2013. Discovery of a Splicing Regulator Required for Cell Cycle Progression. *PLoS Genet* **9**.

Suzuki K. 2008. The multi-functional serpin, protein C inhibitor: Beyond thrombosis and hemostasis. *J Thromb Haemost* **6**: 2017–2026.

The Cancer Genome Atlas Network. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**: 576–582.

The Cancer Genome Atlas Network. 2012a. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**: 519–525.

The Cancer Genome Atlas Network. 2012b. Comprehensive molecular portraits of human breast tumours. *Nature* **490**: 61–70.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46–53.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515. http://www.ncbi.nlm.nih.gov/pubmed/20436464 (Accessed February 13, 2017).

Trincado JL, Sebestyén E, Pagés A, Eyras E. 2016. The prognostic potential of alternative transcript isoforms across human tumors. *Genome Med* **8**: 85. http://www.ncbi.nlm.nih.gov/pubmed/27535130.

Turunen JJ, Niemelä EH, Verma B, Frilander MJ. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* **4**: 61–76. http://www.ncbi.nlm.nih.gov/pubmed/23074130 (Accessed February 12, 2017).

Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, et al. 2017. The UCSC Genome Browser database: 2017 update. *Nucleic*

Acids Res **45**: D626–D634.

Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**: 1212–5. http://www.ncbi.nlm.nih.gov/pubmed/14615540.

Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov F V, Hodges E, Hannon GJ, Sanford JR, Penalva LOF, Smith AD. 2012. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**: 3013–20. http://www.ncbi.nlm.nih.gov/pubmed/23024010 (Accessed February 13, 2017).

Vanharanta S, Marney CB, Shu W, Valiente M, Zou Y, Mele A, Darnell RB, Massagué J. 2014. Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. *Elife*.

Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**: e11752. http://www.ncbi.nlm.nih.gov/pubmed/26829591.

Venables JP, Brosseau J-P, Gadea G, Klinck R, Prinos P, Beaulieu J-F, Lapointe E, Durand M, Thibault P, Tremblay K, et al. 2013a. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Mol Cell Biol* **33**: 396–405. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3554129&tool=pmcentrez&rendertype=abstract.

Venables JP, Lapasset L, Gadea G, Fort P, Klinck R, Irimia M, Vignal E, Thibault P, Prinos P, Chabot B, et al. 2013b. MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in

pluripotent stem cell differentiation. *Nat Commun* **4**: 2480.

Viros A, Sanchez-Laorden B, Pedersen M, Furney SJ, Rae J, Hogan K, Ejiama S, Girotti MR, Cook M, Dhomen N, et al. 2014. Ultraviolet radiation accelerates BRAF-driven melanomagenesis by targeting TP53. *Nature* **511**: 478–82. http://www.ncbi.nlm.nih.gov/pubmed/24919155.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW, Kinzler KW. 2013a. Cancer genome landscapes. *Science* **339**: 1546–58.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6. http://www.ncbi.nlm.nih.gov/pubmed/18978772.

Wang Y, Chen D, Qian H, Tsai Y, Shao S, Liu Q, Dominguez D, Wang Z. 2014. The Splicing Factor RBM4 Controls Apoptosis, Proliferation, and Migration to Suppress Tumor Progression. *Cancer Cell* **26**: 374–389. http://dx.doi.org/10.1016/j.ccr.2014.07.010.

Wang Y, Ma M, Xiao X, Wang Z. 2012. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol* **19**: 1044–1052. http://www.nature.com.online.uchc.edu/nsmb/journal/v19/n10/full/nsmb.2377.html%5Cnhttp://www.nature.com/doifinder/10.1038/nsmb.2377.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**: 831–45.

Warzecha CC, Jiang P, Amirikian K, Dittmar K a, Lu H, Shen S, Guo

W, Xing Y, Carstens RP. 2010. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J* **29**: 3286–3300. http://dx.doi.org/10.1038/emboj.2010.195.

Watermann DO, Tang Y, Hausen A Zur, Jäger M, Stamm S, Stickeler E. 2006. Splicing factor Tra2-β1 is specifically induced in breast cancer and regulates alternative splicing of the CD44 gene. *Cancer Res* **66**: 4774–4780.

Watson IR, Takahashi K, Futreal PA, Chin L. 2013. Emerging patterns of somatic mutations in cancer. *Nat Rev Genet* **14**: 703–718. http://www.ncbi.nlm.nih.gov/pubmed/24022702 (Accessed February 13, 2017).

Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. 2014b. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**: 1160–1165.

Will CL, Lührmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3**: a003707. http://www.ncbi.nlm.nih.gov/pubmed/21441581 (Accessed February 12, 2017).

Wilusz JE, Freier SM, Spector DL. 2008. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135**: 919–32. http://www.ncbi.nlm.nih.gov/pubmed/19041754.

Witten JT, Ule J. 2011. Understanding splicing regulation through RNA splicing maps. *Trends Genet* **27**: 89–97. http://www.ncbi.nlm.nih.gov/pubmed/21232811 (Accessed February 12, 2017).

Wong MS, Wright WE, Shay JW. 2014. Alternative splicing regulation

of telomerase: a new paradigm? *Trends Genet* **30**: 430–8. http://www.ncbi.nlm.nih.gov/pubmed/25172021 (Accessed February 13, 2017).

Xiao R, Sun Y, Ding J-H, Lin S, Rose DW, Rosenfeld MG, Fu X-D, Li X. 2007. Splicing regulator SC35 is essential for genomic stability and cell proliferation during mammalian organogenesis. *Mol Cell Biol* **27**: 5393–5402.

Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, Jin C, Zhou Y, Wen L, Ren J. 2011. GPS 2.1: Enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng Des Sel* **24**: 255–260.

Yanagisawa M, Huveldt D, Kreinest P, Lohse CM, Cheville JC, Parker AS, Copland JA, Anastasiadis PZ. 2008. A p120 catenin isoform switch affects rho activity, induces tumor cell invasion, and predicts metastatic disease. *J Biol Chem* **283**: 18344–18354. http://www.ncbi.nlm.nih.gov/pubmed/18407999.

Yang M, Sun T, Wang L, Yu D, Zhang X, Miao X, Liu J, Zhao D, Li H, Tan W, et al. 2008. Functional variants in cell death pathway genes and risk of pancreatic cancer. *Clin Cancer Res* **14**: 3230–3236.

Yang Y-CT, Di C, Hu B, Zhou M, Liu Y, Song N, Li Y, Umetsu J, Lu ZJ. 2015. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* **16**: 51.

Yeo G, Holste D, Kreiman G, Burge CB. 2004. Variation in alternative splicing across human tissues. *Genome Biol* **5**: R74. http://www.ncbi.nlm.nih.gov/pubmed/15461793 (Accessed February 12, 2017).

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R,

Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**: 64–69.

Zheng Z, Zhu H, Wan Q, Liu J, Xiao Z, Siderovski DP, Du Q. 2010. LGN regulates mitotic spindle orientation during epithelial morphogenesis. *J Cell Biol* **189**: 275–288.

Zhu J, Mayeda A, Krainer AR. 2001. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell* **8**: 1351–61. http://www.ncbi.nlm.nih.gov/pubmed/11779509.

Zong FY, Fu X, Wei WJ, Luo YG, Heiner M, Cao LJ, Fang Z, Fang R, Lu D, Ji H, et al. 2014. The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *PLoS Genet* **10**.

2014. The History of Cancer.

I thank FPI grant from the Spanish Government (BES-2012-052683)
for funding of this thesis and GRIB-UPF for their support.