Characterization of clinically relevant RNA alterations for personalized cancer medicine

Juan Luis Trincado Alonso

TESI DOCTORAL UPF / 2018

THESIS SUPERVISOR

Dr. Eduardo Eyras

DEPARTMENT

Computational RNA Biology Group
Research Programme on Biomedical Informatics (GRIB)
Department of Experimental and Health Sciences (DCEXS)
Universitat Pompeu Fabra

Acknowledgements

Someone told me that the “acknowledgments” should be the first thing to write on a thesis. Now that I’m writing it the last, I understand why. There are so many people who have been part of these 4 years of PhD, so many people who have taught me so much that I do not have time to list them all. For sure I will forget to mention many, so sorry (and thanks) in advance.

First of all, thanks to the supervisor of this thesis, Eduardo Eyras. Thank you for believing in me and giving me the chance to immerse myself in science. Thank you for allowing me to work on something exciting. Thanks for all your explanations. I can’t understand how you manage to remember all the papers you read hehe. Thanks for all the conversations and for these 4 years of scientific adventure. If migrating to Oceania happens in my head, you will be the first one I will call.

Life in research is frantic. There are so many people that come and go that is incredible the amount of people you meet in 4 years. I am really grateful for all the people I met in the lab. I've learned a lot from everyone and I'm glad for having shared work and life with you: Babita, Hector, Amadis, Jorge, Will, Pau, JC, Endre, Isaac, Adria, Marina, Emma, Caterina and more that for sure I’m forgetting. Thanks for all the good and hard moments of work. For all the lunch chats, all the beer sessions, all the scientific debates, all the Vollywood matches… Special mention to our captain Pau for accepting me, even if coming from the capital of the kingdom. We did it man!

Special thanks also to Robert Castelo and Mar Albá for sharing all their knowledge and giving their advice on the group meetings. And also thanks to all Nuria’s lab. You all helped me a lot.

Thanks to all the other people in the PRBB. I know that this is the best place I’ve ever worked at and probably the best place where I will ever work. Not only because of the beach and Beltran’s sandwiches, but for the people. This place creates the perfect environment for making science but also for meeting people. And I have met extraordinary people. Thanks to all the GRIB people: Juanma, Janet, Adriano, Miha. Don’t let CRG or EMBL people eat us. Special thanks to Alfons and Miguel, the IT guys. Sorry for filling the cluster with tons of jobs.

Many thanks to my alternative volley team: Antonio, Julia, Fer, Estepi, Laura, Vicki. Next year we will win something for sure.

Thanks a lot for all my chorists. It has been a great pleasure filling the auditorium with our voices. Special thanks to Oscar, our director, who suffer our not perfect voices weekly with the biggest smile. And also to Imma, Maria, Erica and Aleix for all your extra work. You're the best.

Thanks to Dr. Eric Wang for allowing me to visit his lab. It was an enriching experience. And special thanks to all the people from Conesa's lab for adopting me. You made me talk less English than in Spain..
Thanks also to Dr. Vazquez-Poleti for showing me the path of bioinformatics.

Como no, gracias a toda la gente que hace Sigeco. Aunque sois todos unos juanines, el que pasen los años y ver que seguís ahí me hace darme cuenta de la suerte que tengo de no haber elegido a los Conquis. Desde que os conocí habéis estado siempre presentes, en los momentos malos y en los buenos, día a día. Y aunque no escriba un solo mail, saber que siempre estáis ahí, aunque estemos desperdigados por el mundo, pase lo que pase, es algo muy grande que creo no mucha gente tiene. Gracias de verdad amigos. Y gracias especiales al Dr. Francesc. Si estoy terminando esta tesis es en gran parte gracias a ti. Tú me motivaste a venir a Barcelona e hiciste que no me sintiera extraño en esta ciudad. Gracias por todo lo que hemos vivido.

Gracias Edu, compañero de máster y de vida, por meterte conmigo en esta locura. Tambien tú tienes gran culpa de que esté terminando esta tesis. A Gonsalito por sus locuras. A Naranjo y Anarjo por sus bbqs en la azotea. A Josep Baulida por ser tan buen casero. A los Juanlu Paga, por recordarme siempre que Juanlu va a pagar.

Y por último, gracias a las personas más importantes en mi vida. A ti Laura. Gracias por todo tu apoyo siempre. Gracias por hacerme querer cambiar las cosas. Gracias por tu entusiasmo infinito y por la pasión que pones en todo. Gracias por estos 3 maravillosos años de aventura y por ser cómo eres.

Y a mi familia. A mis hermanos, que son dos de los ejemplos de persona en las que algún día querría convertirme. Y a mis padres, por haberme criado de la mejor manera y por haberme apoyado en las decisiones que he tomado. Soy lo que soy gracias a vosotros y doy gracias por vosotros cada día.

Y gracias a mis abuelas, porque si no las incluyo aquí me desheredarán jeje. Porque con vosotras empezó todo.

Esta tesis es tan mia como vuestra.
Gracias a todos!!Thanks a lot!!Merci a tothom!!

Abstract

Alternative splicing is a major contributor to the variability of eukaryotic transcriptomes. Several works have shown evidence on how aberrant splicing is linked to various diseases, like cancer. Tumors originate from genetic alterations, but their progression involves phenotypic advantages driven by changes in the transcriptome. The study of these alterations allows researchers to elucidate the mechanisms occurring during cancer progression. More important, there is a growing interest in these alterations have resulted as new targets of cancer therapy.
This thesis addresses two key questions: the development of computational methods for the study of RNA splicing alterations and their application in cancer samples for personalized medicine. We aimed to develop methods and perform analyses that could provide a new perspective on cancer analysis with the potential of opening new strategies for treatment.

Resumen

El splicing alternativo es uno de los mayores contribuidores a la variabilidad del transcriptoma. Varios trabajos han mostrado cómo cambios aberrantes de splicing se producen en diversas enfermedades, como el cáncer. Los tumores se producen por alteraciones genéticas, pero su progresión implica ventajas fenotípicas producidas por cambios del transcriptoma. El estudio de estas alteraciones permite a los investigadores dilucidar los mecanismos que ocurren durante la progresión del cáncer. Es importante remarcar que dichas alteraciones han ganado un gran interés como nuevas dianas de terapia dirigida en cáncer.
Esta tesis aborda dos cuestiones fundamentales: el desarrollo de métodos computacionales para el estudio de las alteraciones del splicing en el ARN y su aplicación en muestras de cáncer para la medicina personalizada. Hemos aspirado a desarrollar métodos y llevar a cabo análisis que puedan ofrecer una nueva perspectiva sobre el análisis de cáncer con el potencial de abrir nuevas estrategias de tratamiento.

Preface

I love science. For a long time, I was devoted to do something related with biology. However, I realized that my passion was computer science. I like everything about computers and technology and I decided to study a degree on it. I felt great with my decision. Until one day I learnt about bioinformatics, thanks to my teacher Jose Luis Vazquez-Poletti. He told me about this discipline and I felt in love, again. It mixed my two passions: computers and biology. It was computer "science" for real. And now I'm finishing a thesis about this topic. I have enjoyed a lot these 4 years of PhD and I hope that the reader could feel also this same enthusiasm about all the work done. I hope that this would be just my first step in this fascinating world.

If I had to define the work of a bioinformatician, or computational biologist, I would say that it is based on two keystones: the analysis and extraction of meaningful results from omics-data and the development of methods for this task. This thesis is a bioinformatics approach to the study of splicing in cancer for personalized medicine. We hope that the tools developed could be useful for the scientific community and the analysis performed contribute to generate more knowledge in this field.

Juan Luis Trincado Alonso
Barcelona, September 2018

List of publications

The list of publications is presented in reverse chronological order. Publications 2) and 3) compose chapters 2.1 and 2.2 of the results of the thesis:

1. Singh, B., **Trincado, J. L.**, Tatlow, P., Piccolo, S. R., & Eyras, E. (2018). Genome Sequencing and RNA-motif Analysis Reveal Novel Damaging Non-coding Mutations in Human Tumors. Molecular Cancer Research, 6175, molcanres.0601.2017. https://doi.org/10.1158/1541-7786.MCR-17-0601
2. **Trincado, J. L.**, Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biology, 19(1). https://doi.org/10.1186/s13059-018-1417-1
3. **Trincado, J. L.**, Sebestyén, E., Pagés, A., & Eyras, E. (2016). The prognostic potential of alternative transcript isoforms across human tumors. Genome Medicine, 8(1). https://doi.org/10.1186/s13073-016-0339-3
4. Alamancos, G. P., Pagès, A., **Trincado, J. L.**, Bellora, N., & Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA, 21(9). https://doi.org/10.1261/rna.051557.115

List of contents

1 INTRODUCTION

1.1. From DNA to RNA

The transcriptome is the set of all RNA molecules in a cell. These molecules are defined as a product of the transcription from DNA. They are the connection between the genetic code and the functional molecules that run cells, the proteins. In the past, scientists have put the spotlight on the study of protein products derived from RNA. However, only focusing on protein activity is not telling the cell's whole story. There is a huge variability at transcriptome level that could not be readily detected on proteins and that has important phenotypic effects. Therefore, quantification and characterization of the transcriptome is essential to understand the activity of genes and their regulation.

Year after year, the number of total protein coding genes has been decreasing. In the last human GENCODE reference release (November 2017, GRCh38) there were a total of 58.381 genes, from which 19.901 are protein coding. A recent work suggests that the current number of protein coding genes is still an overestimate and should in fact be around 18.000 (Abascal et al., 2018). Nevertheless, the number of annotated proteins is much higher (around 90.000 (The UniProt Consortium, 2015)). This difference could be explained because of the large number of protein-coding annotated transcripts (82.335 according to GENCODE). There is a series of tightly regulated and conserved mechanisms that allow the synthesis of multiple transcripts from a single gene: alternative transcript initiation, alternative splicing and alternative polyadenylation. We are going to review briefly these processes.

1.1.1. mRNA processing

Here we give a general overview for each of the processing steps from the transcription of DNA to the maturation of the messenger RNA (mRNA) in eukaryotes:

- Synthesis of pre-mRNA: the RNA polymerase II (RNApol II) is the unit in charge of performing the synthesis of the complementary RNA. The place where the RNApol II binds is the promoter, where RNApol II interacts with other proteins called transcription factors (TFs). The presence of different transcription factors could influence the selection different promoters by the RNA pol II for cleaving and starting the transcription at different transcription start sites (TSSs). This could produce transcripts with different first exons (alternative promoter usage).
- 5' capping: the newly synthetized pre-mRNA is capped with a modified nucleotide. This protects the mRNA molecule from ribonuclease degradation and enables ribosome recruitment after the exportation to the cytoplasm.
- Splicing: it is the process by which regions of the pre-mRNA are excised out (introns) or included (exons) in the final processed mRNA. Combinations of different sets of exons will give rise to different transcripts and therefore, could translate into different proteins (Figure 1). This process is known as Alternative Splicing (AS). We will review this in more depth below.
- Polyadenylation: once the whole mRNA molecule has been synthetized, a poly-A tail is added to the 3' end. This tail plays a major role on nuclear transportation and stability. Similarly to the selection of the promoter region, there are different sites where the poly-A tail could be added, i.e. alternative polyadenylation sites.

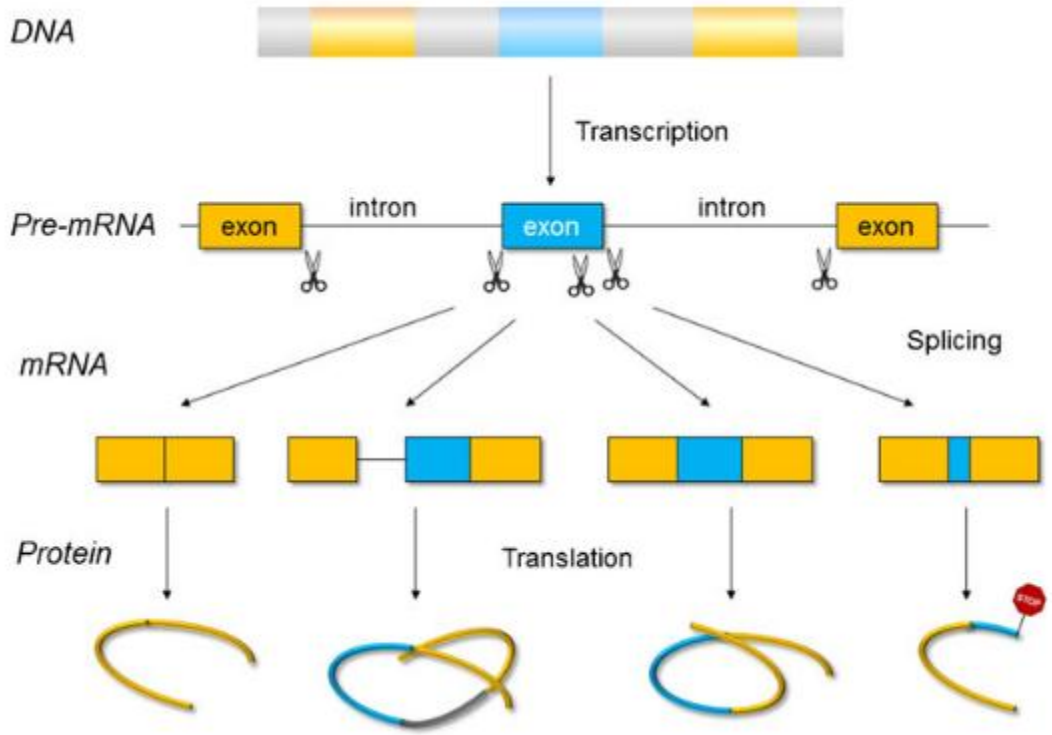


Figure 1: Alternative Splicing (AS) allows the synthesis of multiple transcripts from the same gene locus, and potentially give rise to different proteins by combining different sets of exons (Lara-Pezzi, Gómez-Salinero, Gatto, & García-Pavía, 2013)

For a long time it was thought that the process of transcription and RNA processing took place independently of each other, and that splicing would not take place until the polymerase had not released the pre-mRNA molecule. However, it has been shown that this post-processing starts before the RNA molecule has been completely synthetized. This was first observed in 1988 by Beyer and Osheim in Drosophila (Beyer & Osheim, 1988), and there is now evidence that splicing occurs mostly co-transcriptionally (Carrillo Oesterreich et al., 2016) and that transcription regulation can also influence splicing and alter the final product (Schor, Gómez Acuña, & Kornblihtt, 2013).

The variations in the RNA processing steps mentioned above (alternative promoter usage, alternative splicing and alternative polyadenylation) are the major contributors to the variability of the transcriptome. This evidences the importance of these processes and motivates the study of how they contribute to molecular diversity.

1.2. Splicing

Splicing is the process of intron removal and exon ligation. The exon and intron terms were coined by the Nobel laureate Walter Gilbert in 1978 (Gilbert, 1978), referring as expressed and non-expressed intragenic regions respectively. Some exons carry the genetic information that will code for a peptide sequence. Intronic sequences, as well as a subset of exons, do not encode protein products, but they play an important role on gene regulation.

Alternative Splicing (AS) consists on the multiple ways in which introns may be excised from the nascent transcript, and allows the generation of multiple transcripts from a single gene. Several studies have shown that genes may express alternative transcripts in a cell, tissue or pathway specific way (Baralle & Giudice, 2017). This suggests a complex regulatory mechanism underneath.

1.2.1. Splicing regulation

Which exons will be spliced in is determined by many factors. Three of the major players on this decision are the spliceosome, cis-regulatory sequences and trans-acting factors.

1.2.1.1. The spliceosome

Splicing is catalyzed by a dynamic ribonucleoprotein complex called spliceosome, composed by five small nuclear RNAs (U1, U2, U4, U5 and U6), which form complexes with more than 200 proteins (Will & Lührmann, 2011). The spliceosome recognizes 4 main sequences in the pre-mRNA:

- 5’ splice site (5’ ss): present at the beginning of the intron. It is composed of a highly conserved di-nucleotide, mostly GT, surrounded by less conserved positions, extending to a motif of approximately 6-9 nt corresponding to the base-pairing with the U1 snRNA.
- 3’ splice site (3’ ss): present at the end of the intron. Similarly to the 5’ss, it is composed by a highly conserved dinucleotide, mostly AG, preceded often by a T or C (Padgett, 2012).

- Branch point (BP) adenosine: located ~15-50 nucleotides upstream of the 3' splice site. It is surrounded by a weaker motif that base pairs with the U2 snRNA.
- Polypirimidine tract (PPT): usually 15-20 bases long, located downstream of the branch point and close to the 3' end of the intron. This region is generally bound by the U2AF heterodimer (Sickmier et al., 2006).

There are two types of spliceosomes: the major and the minor spliceosomes. Each one recognizes and processes a different type of introns: U2 and U12 introns, respectively. U2 introns are the most abundant ones (around 99.5%). U12 introns have different sequence elements and they normally co-occur together in genes with U2 introns (C. F. Lin, Mount, Jarmoowski, & Makaowski, 2010). In spite of the low amount of U12 introns, they have important functions at limiting expression of their host genes (Turunen, Niemelä, Verma, & Frilander, 2013). The splice sites of the U2 introns have a canonical GT-AG consensus, whereas the U12 could also be GC-AG or AT-AC. Additionally, U12 introns present a more conserved BP motif and a much shorter PPT (Turunen et al. 2013).

Here we give a brief overview of how the major spliceosome works, a more detailed explanation of the process can be found in (Matera & Wang, 2014). U1 recognizes and binds the 5' ss. U2 interacts with the branch point. U2 auxiliary factors (U2AF heterodimer) recognize the polypyrimidine tract and the 3'ss. With the recruitment of the U4/U5/U6 tri-snRNP the 5' ss is cleaved and joined to the branch point forming a lariat. The 3' end of the intron is next cleaved at the 3' ss and the two exons are ligated together (Figure 2).

The decision as to which exon is removed or included often involves RNA cis-regulatory sequence elements and trans-acting factors known as splicing factors (SFs).

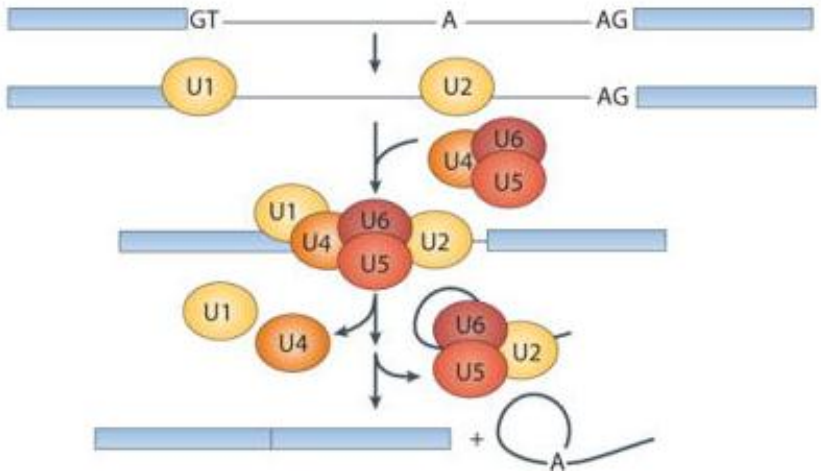


Figure 2: Simplified illustration of the intron excision and ligation of two adjacent exons. The steps shown are: recognition of the 5' and 3' splice sites by the U1 and U2 small nuclear ribonucleoprotein complexes (snRNPs), assembly of the snRNPs into the active spliceosome, the excision of the intron lariat and the ligation of the two exons (Dvinge, Kim, Abdel-Wahab, & Bradley, 2016)

1.2.1.2. Cis-regulatory motifs and trans-acting factors

The spliceosome alone does not process all introns. SFs influence the process of splice site choice. These proteins recognize specific cis-regulatory motifs that depending on the effect they produce, and the location, are classified as:

- Exonic Splicing Enhancers (ESE) or Inhibitors (ESI): enhance or inhibit the inclusion of the exon they are in.
- Intronic Splicing Enhancers (ISE) or Inhibitors (ISI): enhance or inhibit the inclusion of an adjacent exon.

Most studies have looked for motifs nearby splice sites. However, recent studies have shown that distant motifs could be as important as those closer to splice sites (Lovci et al., 2013). Two of the most important splicing factors are serine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). SR proteins mostly bind exonic splicing enhancers, promoting the inclusion of the cis-associated exons. On the other hand, hnRNPs usually bind splicing inhibitors, repressing inclusion. There are other SFs that depending on the region they bind they could have antagonistic effects (Goren et al., 2006).

Many of these regulatory motifs were initially identified using high-throughput computational methods (Barash et al., 2010; Yeo, Holste, Kreiman, & Burge, 2004) as well as experimental screening approaches (Stadler et al., 2006).

Currently, one the most effective ways to obtain an unbiased view of the binding specificities of splicing factors, as well as RBPs, is crosslinking and immunoprecipitation (CLIP) method, in any of its different flavors (F. C. Y. Lee & Ule, 2018). CLIP detects direct RNA-protein interactions by inducing crosslinking with ultraviolet radiation. The use of this technology has been crucial for characterizing the relevant motifs associated to many of the splicing regulatory proteins. RNAcompete is a different method that allows the identification of preferred binding sequences for RBPs using synthetic sequences (Ray et al., 2017). These results have been gathered into databases like ATtRACT (Giudice, Sánchez-Cabo, Torroja, & Lara-Pezzi, 2016; Ray et al., 2009), which contains a large compendium of RNA motifs recognized by splicing factors, obtained from multiple experimental techniques. Additionally, high-throughput experimental methods have been developed to expand the catalogue of RBPs, and hence describe potential new RNA processing regulators (Hentze, Castello, Schwarzl, & Preiss, 2018). There are currently around 1400 proteins estimated to be RBPs, with more than 170 estimated as splicing factors at present (Sebestyén, Singh, Miï¿½ana, et al., 2016), many of which still lack a description their binding sites (X. D. Fu & Ares, 2014). The compendium of RBPs with new biological roles is likely to grow in the near future. It is quite challenging to characterize the binding specificities of RBPs and SFs, since the same genomic sequence could be recognized by different factors. A recent study demonstrated that, in spite of the large amount of identified RNA binding proteins (RBPs), the diversity of motifs is much lower that expected (Dominguez et al., 2018). In addition, some proteins could function in complexes with other proteins, influencing their binding affinities.

1.2.1.3. Other regulators of splicing

There are other agents that affect the regulation of the splicing machinery:

- RNA Pol II processing: as splicing occurs co-transcriptionally, the speed of transcription can play a role in the recognition of splice sites, and in the recruitment of RBPs recruited, often by the RNA Poll II, to the pre-mRNA (Schor et al., 2013).

- Antisense RNAs (asRNAs): asRNAs are single stranded RNAs transcribed from the opposite strand of a gene locus that can hybridize to the pre-mRNA or mRNA blocking its processing. These molecules have been shown to be potential regulators of AS. For instance, the splicing of MSH6 gene, involved in DNA-mismatch repair, was shown to be correlated with expression of antisense FBXO11 (Morrissy, Griffith, & Marra, 2011).
- DNA methylation: It has been shown that alternatively spliced exons show lower levels of methylation than constitutively spliced exons (Lev Maor, Yearim, & Ast, 2015). For instance, CD45 exon 5 inclusion is inhibited by DNA methylation (Shukla et al., 2011).
- Histone modifications: the chromatin state, defined by the biochemical modifications of nucleosome histone tails also influence the splicing of alternative exons. For instance, elevated levels of trimethylation of H3K9me3 have been related to the repression of alternative exons of several genes like CD44 (Saint-André, Batsché, Rachez, & Muchardt, 2011)
- Long Noncoding RNA (lncRNA): These molecules have been shown to influence AS (Romero-Barrios, Legascue, Benhamed, Ariel, & Crespi, 2018). For instance, lncRNA MALAT1 blocks the recruitment of splicing factors to pre-mRNAs. In non-small cell lung cancer, MALAT1 is often depleted increasing splicing factor levels and producing a perturbed splicing pattern (Tripathi et al., 2010).
- RNA editing: Deamination of adenosine to inosine in exons and/or introns has been shown to be related to AS. For instance, cassette exons were found to be significantly enriched with adenosine-to-inosine RNA editing sites by ADAR enzymes compared to constitutive exons (Solomon et al., 2013).
- RNA modification: N6-methyladenosine (m6A) has been recently identified as another modification that influences AS; m6A can alter the structure in mRNAs by enhancing the binding of HNRNPC (Bartosovic et al., 2017).

All these mechanisms regulate and impact splicing, thereby possibly giving rise to AS. We will describe next how to study AS from high-throughput RNA sequencing data.

1.3. Alternative Splicing quantification

Alternative splicing analysis requires the quantification of the transcriptome. Next generation sequencing technologies have facilitated this task. Their ability to provide massively parallel analysis from several samples at much reduced cost has led to the explosion of RNA bioinformatics. Whole transcriptome shotgun sequencing, also known as RNA-seq, is the standard approach for transcriptome profiling nowadays. On the other hand, third-generation sequencers from PacBio or Oxford Nanopore Technologies use advances in nanotechnology to sequence full-length RNA molecules (ref). These technologies are emerging as new promising techniques for transcriptome analysis (de Jong et al., 2017; Tardaguila et al., 2017). Their longer reads could allow the direct resolution of isoform structure. On the other hand, the higher error rates, price, and low throughput still make these technologies not the standard in the field. Since short-read RNA-seq is the standard technology, we are going to focus on methods that use this type of data.

Prior to quantifying AS, the first step is read alignment. This could be done mapping to the genome or to a transcriptome reference. Using an existing transcriptome has the drawback that possible unannotated transcripts will be excluded from the quantification. There is a plethora of methods to map reads. STAR (Dobin et al., 2013) or TopHat (D. Kim et al., 2013) are two of the most widely used mappers. Reads mapped to a genome be assembled into longer contigs to reconstruct potential transcripts and estimate their abundance (Pertea et al., 2015; Trapnell et al., 2010). Reads mapped to a transcriptome must be re-distributed into transcripts using optimization methods to estimate transcript abundances (ref). Recently, a new generation of methods has appeared that provide a pseudo-alignment or quasimapping to the transcriptome, like Salmon (Patro, Duggal, Love, Irizarry, & Kingsford, 2017) or Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016), which have improved significantly the speed at quantifying transcript abundances. Other methods like Trinity (Haas et al., 2013) allows performing de novo transcriptome assembly and quantification, which is especially relevant in experiments on poorly annotated species.

The task of quantifying AS could be addressed from different perspectives. One option would be to calculate the independent usage of each exon, or each independent exon piece, like in DEXSeq (Anders, Reyes, & Huber, 2012). Using this approach, it is possible to detect changes between conditions associated to specific exonic regions. However, this approach could mask

splicing changes that would not be visible unless neighboring exons are considered. In Figure 3 we illustrate this with an example. Accordingly, methods that study the splicing variation in the context of the neighboring splice sites produce more reliable results.

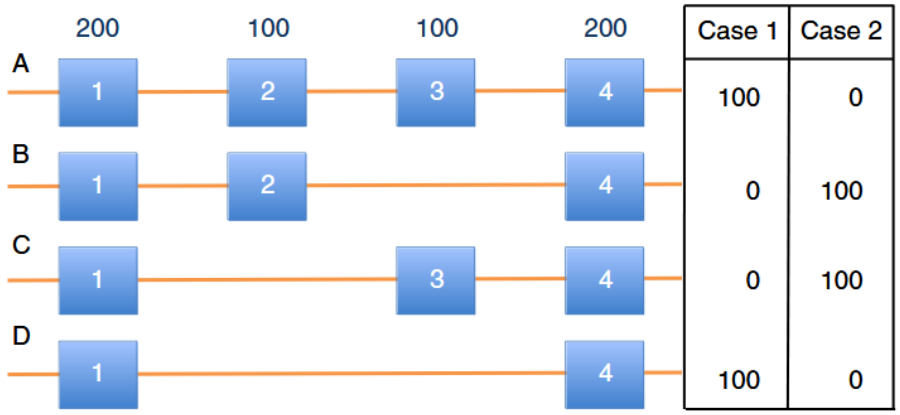


Figure 3: Only looking at exon usage could be masking splicing changes. The image shows a gene with four transcripts, with different sets of exons. In individual 1, only isoforms A and D are found with abundance of 100 copies, whereas in individual 2, only B and C are found, with the same abundance. If we focus our analysis on exon usage only, we will not see any change between the two individuals on their exons abundances, whereas they do show differential splicing. Adapted from (Monlong, Calvo, Ferreira, & Guigó, 2014).

One of the most common ways of studying splicing is through alternative splicing events, which involve binary choices of exons or splice sites. In Figure 4 we describe the most commonly studied alternative splicing patterns. Alternative first and alternative last exons are not pure AS events, since they are generated by the selection of alternative promoter or polyadenylation sites, respectively. As they often involve the interplay of splicing regulators, they are generally included as alternative splicing patterns.

There are several methods for quantifying events using RNA-seq data. The main differences between them are their definition of alternative splicing events and the method of obtaining the inclusion level of the event or percent/proportion spliced in (PSI), which represents the proportion of transcripts that include an exon or splice site. In this thesis we present SUPPA2, a lightweight method that exploits normalized transcript abundance values for obtaining differential splicing on binary events. SUPPA2, like rMATS (Shen et al., 2014), obtains differential AS changes relying on the previous definition of events from the annotation. Other methods, like MAJIQ choices (Vaquero-Garcia et al., 2016), obtain differential splicing changes based on local splicing variations, which are not necessarily binary.

Another possibility for quantifying differential splicing is to study the change in relative abundances of transcript isoforms. One would assign reads to isoforms and measure the percentage of expression explained by each isoform at different conditions. Cuffdiff, a utility from Cufflinks calculates the significance of differential splicing by comparing isoform expression estimations between conditions (Trapnell et al., 2010). MISO is another popular method that identifies differentially regulated isoforms using Bayesian inference (Katz, Wang, Airoldi, & Burge, 2010). Iso-kTSP identifies significant "switches" on isoform changes between conditions (Sebestyén, Zawisza, & Eyras, 2015). DRIMseq is another method to study differential transcript usage by modeling transcript counts with the Dirichlet-multinomial distribution (Nowicka & Robinson, 2016).

Both perspectives, event and isoform quantification, are complementary to each other and provide distinct information. Events provide a local description of the splicing change, which is useful to perform validations by RT-PCR or to give a mechanistic hypothesis based on the sequence content by looking at motifs affected by mutations or enrichment of RBPs. On the other hand, isoforms are the actual molecules carrying function. The whole sequence of the transcript allows obtaining the ORF and inferring the possible changes at protein level, and consequently study the functional domains affected. So, both approaches are valid and useful for AS analysis.

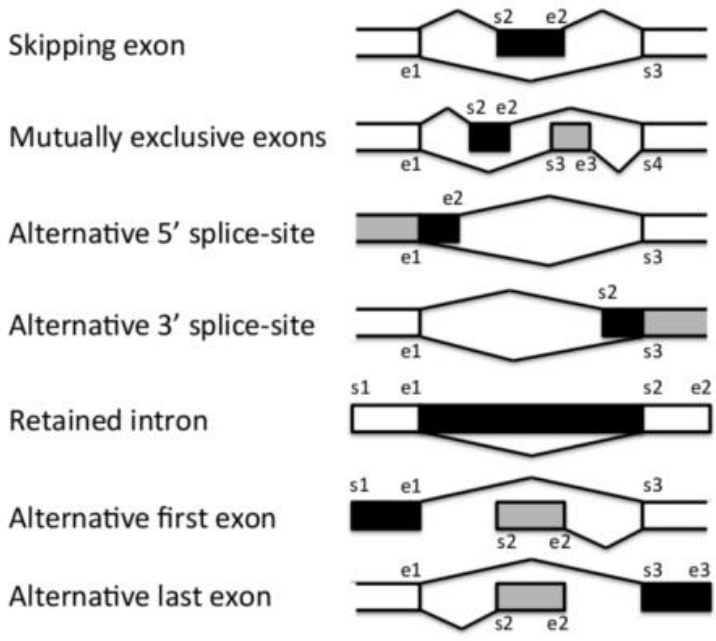


Figure 4: Most studied splicing variations (G.P. Alamancos, Pagès, Trincado, Bellora, & Eyras, 2015)

A different approach to study differential splicing consists of the study of associations between genotypes and splicing variation. Splicing quantitative trait locus (sQTL) analysis is a useful approach to discover genetic alterations connected with AS. The PSI values are treated as traits and tested for association with sequence variants, in the same way as in genome-wide association studies (GWAS). Some of the methods available are GLiMMPS, which identifies sQTLs at event level (Zhao, Lu, Park, Zhou, & Xing, 2013) or sQTLseekeR, which obtains the same at transcript level (Monlong et al., 2014). Notably, these methods allow the analysis for the discovery of any DNA alteration in *cis* or *trans*.

It is important to remark that there is no single "perfect" pipeline. Each method has pros and cons and depending on the biological question at hand and the context of the experiment, a method may be better suited than others. In addition, the results provided using different techniques could be complementary to each other and shed light on a problem. Alternatively, using the overlapping results between different tools has been proven to be an effective way to reduce the number of false positives (Z. H. Zhang et al., 2014).

With the increasing popularity of open source version control systems like Github or Bitbucket, users can check if a given tool has been modified after paper publication and whether it is actively maintained. In fact, this has been shown to be connected with higher number of citations and a wider community of users (Russell, Johnson, Ananthan, Harnke, & Carlson, 2018). In addition, it gives the community an opportunity to add contributions to the code and improve it. The reviews Alamancos, Agirre, & Eyras, 2014 and Conesa et al., 2016 provide more information and interesting summaries of the most widely used methods.

1.4. Functional impact of alternative splicing

1.4.1. Evidence of functional relevance

Thanks to the recent advances in transcriptional profiling, it has been extendedly proved that alternative splicing is one of the major forces driving transcriptome diversity (Harrow et al, 2012, Sanchez-Pla et al, 2012). It has been shown that 95% of human genes produce more than one isoform (Gerstein et al., 2014). What functions all these different transcripts have is still a question of heated debate.

Most genes are considered to have a principal or "canonical" transcript. In Ensembl, this isoform is defined according to APPRIS, which is a database for the annotation of the principal isoforms per gene based on structure, function and conservation (Rodriguez et al., 2013). When there is no information about a gene in APPRIS, Ensembl labels the longest isoform as principal. On the other hand, analysis of RNA-seq from many tissues and individuals has shown that the most highly expressed isoform in a gene may change between normal tissues (Gonzalez-Porta, Frankish, Rung, Harrow, & Brazma, 2013) and between normal and tumor samples (Sebestyén et al., 2015). This change in the principal isoform could thus change the open reading frame (sequence of nucleotide triplets that are read as codons specifying amino acids) resulting in the generation of different proteins. This could result in acquisition of new functions or even the complete loss of function. In this way, AS provides functional diversity at the level of enzymatic activities, subcellular localizations, protein-protein, protein-DNA, and protein-ligand physical interactions (Kelemen et al., 2013). This is particularly illustrated by the fact that interaction partners specific to alternative isoforms tend to be expressed in a highly tissue-specific manner (Yang et al., 2016).

AS regulation has been observed to have a great prominence in brain. Almost 400 splicing events change in cerebral cortex between embryo and adult mice, with more than 30% of the associated genes not showing any change at expression level (Dillman et al., 2013). AS is particularly relevant to neuronal development in mammals; and there is a strong evolutionary conservation of "microexons", generally defined as exons of length <28nt, which modulate the activity of protein interaction domains in neurogenesis (Dergai et al., 2010), and are deregulated in Autism (Irimia et al., 2014). Several RBPs like PTBP1 or RBFOX has been shown to regulate brain development (X. Zhang et al., 2016).

AS is also crucial in striated muscle (heart and skeletal). During heart development, there are important expression changes in CELF1, MBNL1, RBFOX1 and 2 and RBM24, leading to several splicing changes (T. A. Cooper & Giudice, 2014) TTN (titin) is the gene with the largest number of exons in mammals. In neonates, N2BA isoform of this gene is expressed whereas in adults a shorter version (N2B) is the most abundant (Bang et al., 2001), and the relative abundance of the two isoforms is controlled by RBM20 (S. Li, Guo, Dewey, & Greaser, 2013). In skeletal muscle differentiation, specific splicing networks have been shown to operate orchestrated by RBPs like PTBP, QKI or RBFOX2 (R. K. Singh et al., 2014).

Also important is the regulation of splicing during spermatogenesis. Mammalian testis is among the tissues with higher transcriptome variability (Soumillon et al., 2013). Numerous splicing events are differentially expressed during spermatogenesis, showing enrichment of several RBPs like PTBP1, PTBP2, TRA2B and STAR family proteins (Schmid et al., 2013).

AS is involved in other biological processes like gender differentiation. A skipping exon event in the DSX gene has been described to be essential for gender differentiation in Drosophila. In males, exons 1,2,3,5 and 6 are joined to form an mRNA that encodes a transcriptional regulatory protein required for male development. In females, exons 1,2,3, and 4 are joined together instead, and a polyA signal in exon 4 causes cleavage of the mRNA at that site (Lynch & Maniatis, 1996). This influence in sexual differentiation has actually been used for pest control (G. Fu et al., 2007).

There is also a link between T cell activation and splicing. T cells are in charge of triggering an immune response against antigens. Global splicing networks are coordinated during T cell activation, together with nucleosome occupancy and RBP activity (Gaudreau, Heyd, Bastien, Wilhelm, & Moroy, 2012); and it has been proposed that intron retention might regulate T cell activation (Ni et al., 2016).

AS has also been seen changing under general physiological conditions of the body. Stress on students has been linked to the skipping of an exon in the SMG-1 kinase, having a downstream effects on p53 pathway (Kurokawa et al., 2010). As illustrated, AS has a strong relevance for life.

In spite of all these findings, thousands of transcript isoforms still lack a functional characterization. There is still a long way until we fully understand the functional roles of developmental splicing networks.

1.4.2. One gene, multiple proteins; or one gene, one protein?

A still largely debated point is how much is the diversity of splicing transcript variations reflects at proteomic level. In Tress et al. 2008 the authors confirmed the presence of multiple alternative gene products in two different proteomic studies on Drosophila. Nine years later, these same authors supported the opposite view that the fraction of splice isoforms that produce stably folded proteins is scarce (Tress, Abascal, & Valencia, 2017a). In Abascal et al., 2015, the authors revisited eight large-scale mass spec experiments on human and detected splice events for only 246 genes. In 2014, two large-scale proteomics analyses appeared that contradicted these views (M.-S. Kim et al., 2014; Wilhelm et al., 2014). In these studies, 22-37% of the genes with multiple protein isoforms appeared to have peptide evidence for more than one isoform. These studies were contested with the argument that there was not sufficient quality in the detected peptides, hence leading to an overestimation in the number of protein coding genes found (Ezkurdia, Vázquez, Valencia, & Tress, 2014)

Disputing this skeptic view, in Blencowe, 2017, and in other previous works (Bensimon, Heck, & Aebersold, 2012; Blakeley, Siepen, Lawless, & Hubbard, 2010), it was argued that mass spectrometry does not have enough sensitivity yet for detecting splicing variations. In addition, the fact that the methods for peptide identification rely on the proteins having specific ion signatures limits the detection of sufficiently diverse proteins. Additionally, some works have pointed out also the difficulty of detecting short proteins (Slavoff et al., 2013) as well as short-lived proteins that are rapidly degraded in purification procedures (Fälth et al., 2006). Other techniques like ribosome profiling (Ribo-seq) are supporting the idea that a major proportion of the AS variants detected are engaged and possibly translated by ribosomes (Weatheritt, Sterne-Weiler, & Blencowe, 2016). On the other hand, as it was commented in Tress, Abascal, & Valencia, 2017b, detection of transcripts bound by the ribosome does not ensure that a stable protein will be produced. There are some machine-learning methods that make use of different

structural, evolutionary, regulatory and network features, which predict that around 30% of annotated protein isoforms are stably folded (Hao et al., 2015).

The fact that there seems to be a lack of proteomic evidence for the splicing variants does not diminish its biological value in physiological and disease contexts. In fact, AS could generate alternative functional transcripts that do not all necessarily code for proteins. For instance, intron retention events often introduce a premature stop codon that triggers nonsense-mediated decay (NMD) (Green et al., 2003). NMD is a mechanism that degrades mRNAs harboring premature stop codons limiting the translation of abnormal proteins. During translation the ribosome removes the exon junction complexes formed after joining the exons. If a termination codon is found 50nt upstream of one of these complexes, the transcript will be subjected to NMD, thereby regulating the expression of genes (Zheng, 2016).

Another example of the functionality of alternative splicing is the modification of the cellular localizations of transcripts. Splicing changes could appear in untranslated regions that would not affect the coding part but the transportation of the transcript. Transcripts retained in the nucleus would not be translated, hence affecting gene expression (S. Sun, Zhang, Sinha, Karni, & Krainer, 2010). Finally, there is an increasing number of splicing events in non-coding RNAs (Kiegle, Garden, Lacchini, & Kater, 2018).

For all of this, the importance of AS goes beyond the encoding of alternative proteins. Moreover, the presence of AS in almost all biological functions, makes particularly relevant the study of its deregulation, which could give rise to various disease types.

1.5. Alternative Splicing in disease

Understanding the role of splicing in disease is fundamental to uncover new disease mechanisms and to investigate new ways of therapeutic intervention. The exon-intron architecture of eukaryotic genes provide some evolutionary advantages (Sharp, 1994), but the correct processing of pre-mRNAs by the splicing machinery is susceptible to be affected by different genetic alterations. Since many of the mutations falling on exonic or intronic regions could disrupt splicing, it is predicted that between 15 and 50% of disease-causing mutations affect splicing (Scotti & Swanson, 2016). The most common type of alterations that disrupts splicing are cis-acting, meaning that they are affecting core regulatory sequences (5' ss, 3' ss or branch point) or RBP binding motifs (Vaz-Drago, Custódio, & Carmo-Fonseca, 2017). One of the first discovered cases was a mutation in the 3' ss in the gene HBB, which encodes β-globin, which lowered the levels of this protein causing anemia (Maquat et al., 1980).

Disease-causing splicing mutations are often linked to the production of aberrant protein products. Mutations in the LMNA gene have been related with expression of different aberrant transcripts, which are associated with several diseases called "laminopathies" (Luo, Mastaglia, & Wilton, 2014). One of these is the Progeria syndrome, which is caused by a single mutation in exon 11 of LMNA and activates a cryptic splice site, giving rise to a truncated protein with a 50 amino acid deletion (Pendás et al., 2002). Similarly, Familial Dysautonomia is a recessive disease caused by a mutation in the intronic region of IKBKAP gene. Most of the affected patients show a T to C transition in the sixth base of the intron. This mutation leads to exon 20 skipping, causing a shift in the reading frame and degrading the transcript by NMD (Slaugenhaupt et al., 2001). A mutation falling on any region has thus the potential for triggering a disease phenotype.

Recently, two splicing related diseases have attracted wide media coverage due to the development of new treatments. One is Duchenne muscular dystrophy, which is caused by mutations in the donor splice sites of exons 16 and 45 in the gene DMD, leading to the production of an aberrant protein (Fletcher et al., 2013). On the other hand, mutations or deletions of the SMN1 gene leads to Spinal Muscular atrophy in patients, as the homologue SMN2 does fully splice in the same way as SMN1, hence it does not produce a fully functional protein at the required level (Lorson & Androphy, 2000). Antisense oligonucleotides (AONs)

directed against the SMN2 pre-mRNA can recover the correct splicing outcome and lead to the necessary production of functional protein to stop motor neuron degeneration (Hua et al., 2011; Wein et al., 2014).

Other type of alterations affecting RNA processing in disease have been described. For instance, mutations in PRPF6, a core spliceosome component cause global splicing dysregulation in patients in the retina leading to retinitis pigmentosa (Tanackovic et al., 2011); and mutations in RBM20 induces defects in heart development leading to cardiac disease (ref). The relevance of AS in the neuronal system development also reflects in roles in neurodevelopmental disorders. The protein nSR100 is a key regulator of microexon inclusion (length < 28 nt) in neurogenesis (Raj et al., 2014) and its deregulation has been linked with Autism Spectrum Disorder (ASD) (Quesnel-Vallières et al., 2016). In myotonic dystrophy, a CTG expansion in the 3′ untranslated region of the DMPK gene causes a sequestration of two splicing regulatory proteins: CUGBP1 and MBNL1. As a result, a network of alternative splicing events is changed, which causes abnormalities in heart and skeletal muscle development (Ranum & Cooper, 2006). Splicing alterations are not only linked to inheritable genetic disease, it is also observed associated to the somatic alterations that take place in tumors.

1.6. Cancer

The systematic sequencing of the genomes from multiple cancers has highlighted cancer as a genetic disease (Martinocorena & Campbell, 2015). A normal cell becomes a tumor cell driven by changes in the DNA. These alterations range from single point mutations to copy number variations, insertions, deletions, translocations, inversions, etc. (Stratton, Campbell, & Futreal, 2009). Cancer could in fact be seen as an evolutionary process in which cells undergo multiple mutations, not all necessarily producing a phenotypic effect. However, mutations conferring some selective advantage, like a more proliferative phenotype or evasion of apoptosis, are able to survive and keep spreading (Figure 5).

The genetic analysis of tumor cells allows the elucidation of which mutations are the ones that gave rise to the cancer. But even though cancer is a disease of the genome, the impact of these somatic alterations are reflected through the transcriptome, which represents a first read-out of the cell phenotype. Additionally, not all tumoral processes can be explained through somatic mutations. For instance, invasive phenotypes may be influenced by the gene expression patterns of the microenvironment(Calon et al., 2012). As natural selection acts on the phenotype rather than on the genotype, it is therefore crucial to study tumor transcriptomes to understand the mechanisms of cell transformation. As AS is the main driver of transcriptome diversity, it is thus equally important to study the role of AS in cancer.

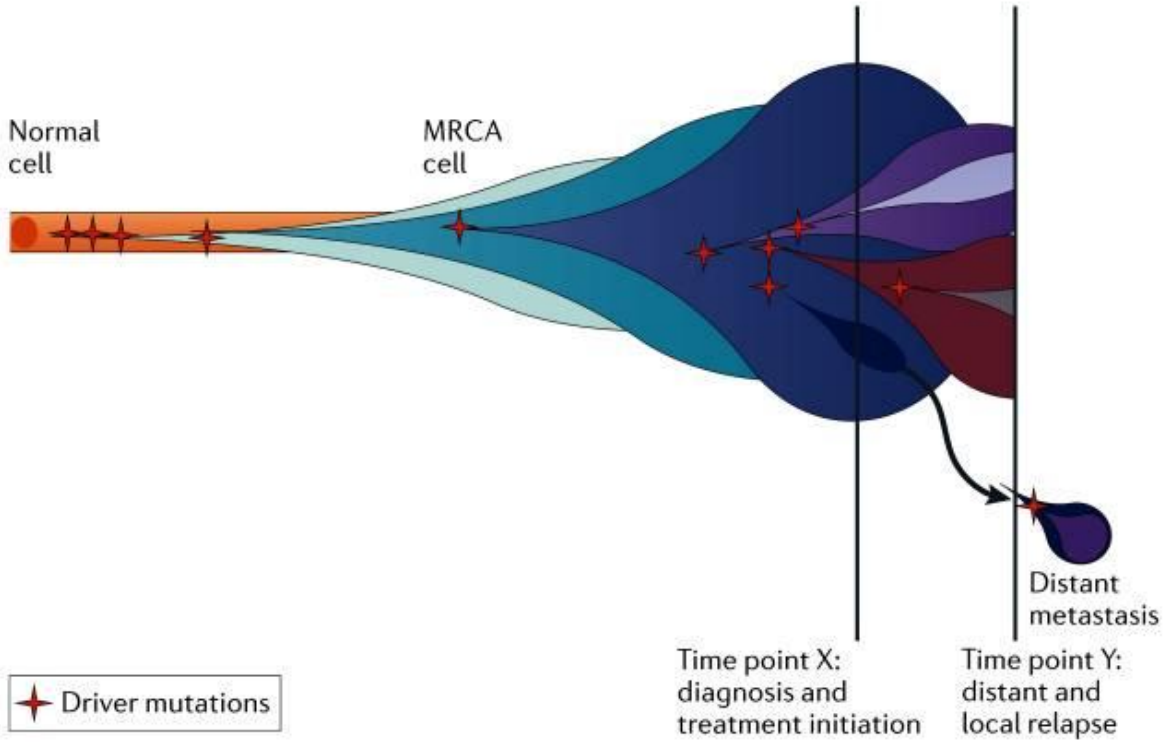


Figure 5: Accumulation of driver mutations gives rise to new subpopulations with different phenotypes. At certain point, simultaneous subclones could give rise to relapse and metastasis. Therapy can act as an additional selection pressure by killing some cells but allowing resistant clones to prevail (Yates & Campbell, 2012)

1.6.1. Alternative Splicing in Cancer

Cancer is not a single disease but rather a set of diseases, each built as a mix of different phenotypes. Hanahan and Weinberg defined the general properties of a malignant tumor as hallmarks (Hanahan & Weinberg, 2011). Later on, AS has been shown to contribute to these cancer hallmarks (Oltean & Bates, 2014). Several genes have been described having isoforms with different exon compositions carrying antagonistic functions with a role in cancer (Figure 6). An example is the gene FAS and its role in apoptosis. The isoform including exon 6 produces membrane-bound FAS protein, which promotes apoptosis, whereas skipping of the same exon produces a soluble protein that inhibits apoptosis (Cheng et al., 1994). Another example is VEGFA where extension of exon 8 in the transcript promotes the formation of new blood vessels (David & Manley, 2010). Splicing of MST1R exon 11 increases cell motility and promotes metastasis (Ghigna et al., 2005). Recently, new cases on emerging cancer hallmarks such as inflammation and avoidance of immune detection has been discovered. This is the case

of the CD19 antigen, which is recognized and killed by CAR-T cells. However, exon 2 skipping fails to trigger this defense mechanism (Sotillo et al., 2015).

1.6.1.1. Mutations on regulatory sequences

Direct cis-splicing changes could be induced if mutations fall within splice sites or regulatory elements. Somatic mutations at exon-intron boundaries frequently lead to intron retention events (Jung et al., 2015). Also, mutations that are "silent", i.e. not altering the peptide sequence, could also have important tumorigenic consequences. TP53 has the highest number of recurrent inactivating mutations of this type compared to other cancer genes (Supek, Miñana, Valcárcel, Gabaldón, & Lehner, 2014). Although harder to characterize, mutations in intronic regions far from the exon-intron boundaries have been also related to AS misregulation in cancer by disrupting e.g. branch-points, polypyrimidine tracts or intronic splicing motifs (Diederichs et al., 2016).

1.6.1.2. Mutations in splicing factors

Mutations affecting components of the spliceosome have been identified across multiple cancer types. Alterations in the splicing core factor SF3B1 are associated with 3' cryptic splice site recognition and altered branchpoint selection (Darman et al., 2015). SF3B1, U2AF1 and SRSF2 are among the most commonly mutated factors in tumors, and occur most prominently in hematological malignancies and uveal melanoma (Dvinge et al., 2016) . Importantly, these alterations are mutually exclusive with one another (S. C. W. Lee et al., 2018). In solid tumors, the most prevalent are SF3B1 in breast cancer and melanoma and U2AF1 in non-small cell lung tumors. Alterations in auxiliary splicing factors also promotes splicing deregulation, like RBM10 in non-small cell lung cancer or HNRNPL in colon cancer (Bechara, Sebestyén, Bernardis, Eyras, & Valcárcel, 2013; Sebestyén, Singh, Miï¿½ana, et al., 2016) .

1.6.1.3. Expression changes in splicing factors

Overexpression or downregulation of SFs have been observed to trigger also tumorigenic properties in cells. Indeed, it has been shown that these expression changes produce more splicing changes than mutations (Sebestyén, Singh, Miï¿½ana, et al., 2016). These changes

could come from alterations in transcription regulators, copy number alterations (Sebestyén, Singh, Miï¿½ana, et al., 2016) or post-transcriptional modifications (Feinberg, Koldobskiy, & Göndör, 2016). A commonly overexpressed factor in tumors is SRSF1, which regulates events participating in a wide range of processes like cell proliferation, apoptosis or signaling pathways (Das & Krainer, 2014). In contrast, factors like QKI are commonly downregulated cancer. QKI depletion in non-small cell lung cancer affects the splicing of NUMB who has a key role in cell proliferation (Zong et al., 2014). Especially relevant is the role of oncogene MYC at controlling multiple SFs. MYC overexpression is frequent in tumors, leading to upregulation of these SFs (Anczukow & Krainer, 2016). Several other cases have been reported (Dvinge et al., 2016).

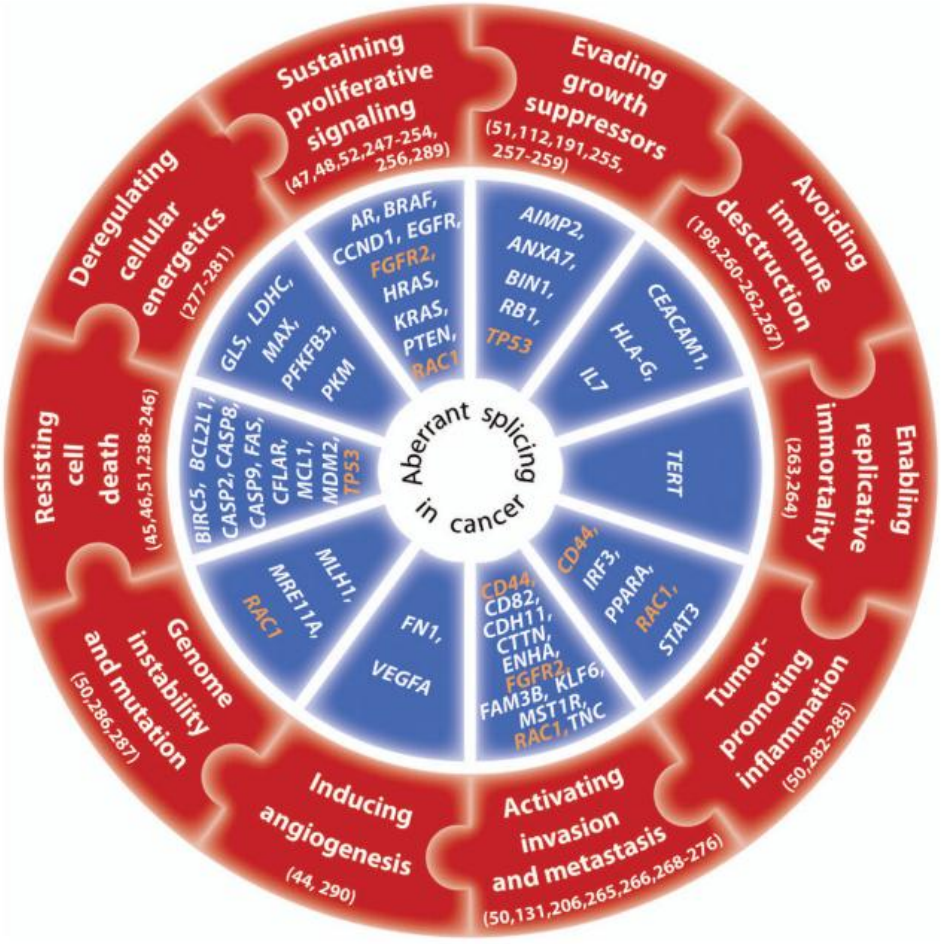


Figure 6: Alternative splicing of several genes have been directly implicated with each of the cancer hallmarks (Sveen, Kilpinen, Ruusulehto, Lothe, & Skotheim, 2016)

1.7. Therapeutic possibilities of Alternative Splicing

We described briefly above some of the diseases that aberrant splicing could lead to. These alterations can in fact be turned into therapeutic opportunities. AS has been extensively studied as a therapeutic option. Additionally, AS has great implications at therapeutic level as informative signature of therapeutic vulnerability and resistance.

1.7.1. Alternative Splicing and therapeutic response

Alternative splicing has shown to be crucial to understand therapeutic response in cancer. A proportion of lung adenocarcinoma patients harbor mutations on and around exon 14 of MET gene, a known proto-oncogene, resulting in exon skipping of that exon. Patients having this alteration and no other oncogenic mutation responded to two drugs approved for targeted therapy in lung cancer, crizotinib and cabozantinib (Paik et al., 2015). Alterations in AS are also important for drug resistance. In B cell acute lymphoblastic leukemia, CD19 gene, expressed on this cancer, can be targeted with CART-19. Nevertheless, in 30% of the cases the patients are resistant to the treatment. In a recent study, it was observed that SRSF3 expression alteration promotes exclusion of exon 2 in CD19, preventing the targeting by CART-19 and failing to kill the cancer cells (Sotillo et al., 2015). Similarly, a proportion of melanoma patients with mutated BRAF are resistant to BRAF inhibitor vemurafenib, which was explained by a skipping of exons 4 to 8 that leads to the loss of the RAS-binding domain (Poulikakos et al., 2011). These examples evidence how certain splicing alterations may confer a selective advantage to tumors and allow the identification of patients that may or may not benefit from specific therapies.

1.7.2. Therapeutic targeting of alternative splicing

AS alterations are possible targets for certain therapies. Here we give a brief overview on the most important strategies.

1.7.2.1. Antisense oligonucleotides (AONs)

One of the most promising molecular tools for splicing therapy is AONs. These molecules are designed to be antisense to an RNA and through base pairing to modulate the splicing of the target RNA. In this way, if a specific transcript is recognized as pathogenic, the use of AONs could prevent the expression of the disease-related transcript, or activate the expression of a fully normal transcript. Some studies with AONs are showing promising results in muscular maladies. In myotonic dystrophy Type 1, targeting 3' UTR region of DMPK gene in mice show a significant reduction of aberrant mRNA in this gene and an improvement on body weight and muscle strength without overt toxicity (Jauvin et al., 2017). Erythropoietic protoporphyria is another important case, in which 90% of people with this disease inherit a biallelic polymorphism in intron 3 of the FECH gene. This creates a cryptic splice site enhancing inclusion of exon 4, which produces a premature stop codon resulting in a decrease of the associated protein. Applying an AON reduces the production of the aberrant transcript and the increase production of FECH wild type (Oustric et al., 2014). Nusinersen, the first AON based therapy, and the first and most promising treating for spinal muscular atrophy (SMA), was approved in December 2016 by the Food and Drug Administration (FDA). This compound induces full inclusion of exon 7 in the SMN2 mRNA, otherwise only partial, by blocking a downstream intronic splicing silencer. The drug has lead to a dramatic improvement in motor neuron function in clinical trials (Claborn, Stevens, Walker, & Gildon, 2018).

In cancer, there are several ongoing trials testing AONs. Unfortunately, the results so far do not show a significant reduction of cancer growth. One of the major issues is how to efficiently deliver the drugs into the tumors (Moreno & PÃªgo, 2014). In colorectal cancer, a study has shown that the combination of AONs and chemotherapy on patients effectively reduces mRNA levels of eIF4E, which is deregulated in this cancer (Duffy et al., 2016). Nevertheless, they did not observe a reduction on eIF4E protein expression, so it is not yet clear the applicability of this therapy. In a study on different cancer types, it was identified a signature of 27 miRNA consistently up or down regulated involved in cell growth and apoptosis (Volinia et al., 2006). The use of anti miRNA oligonucletoides (AMOs) on all these miRNA resulted in increased deaths of cancer cells (Z. Wang, 2010). These works show promising results but there is still a long way until AONs or AMOs will be accepted as a standard therapy in cancer.

1.7.2.2. Small molecule compounds

Another promising strategy for cancer therapeutics is the use of small molecule compounds that modulate the spliceosomal core components and splicing regulators (S. C. W. Lee & Abdel-Wahab, 2016). Several drugs have been proposed to target SF3B1 and destabilize the interaction of U2 snRNP with the pre-mRNA (Kaida et al., 2007; Kotake et al., 2007; SAKAI, ASAI, OKUDA, KAWAMURA, & MIZUI, 2004). These molecules show promising results at efficiently disrupting the U2 snRNP and arresting cell cycle. In particular, H3B-8800 has shown promising results and is now entering phase I clinical trials (Seiler et al., 2018). Similarly, there is a series of small molecules (Cpd-1,2 and 3) that inhibit phosphorylation of SR proteins via kinase blocking. Administration of these compounds reduces SR phosphorylation and induces splicing alterations and protein depletion for genes involved in growth and survival (Araki et al., 2015).

New potential therapies exploit the vulnerabilities of the tumor. Recently, it was shown recently that tumors with MYC overexpression are highly dependent on the splicing machinery and may be more sensitive to splicing-therapies (Hsu et al., 2015). For instance, knockdown of PRMT5 gene, which methylates Sm protein components of the U2 snRNP, stop proliferation in MYC-driven lymphomas (Koh et al., 2015). The combination of AONs with splicing-targeting small compounds seems to be quite promising as a therapeutic strategy, since it encapsulates the advantages of both approaches: being specific and having a potent effect with a small dose to revert oncogenic AS events (Makowski, Vigevani, Albericio, Valcarcel, & Alvarez, 2017).

1.7.2.3. Cancer immunotherapy

In the last few years, cancer immunotherapy has been developing rapidly, chaining a number of remarkable successes in new treatments for tumors with dismal prognosis. Cancer cells are able to avoid immune system recognition either by constitutively expressing immune inhibitors, or by altering the expression of mediators of immune response (Vinay et al., 2015). Immunotherapies exploits the fact that these cells often produce aberrant sequences (neoantigens) that are presented on the cell surface (neoepitopes) and can be detected by T cells, which would then destroy the tumor cells. These therapies attempt to “help” the immune

system by inhibiting the brakes (immune checkpoints) to recognize tumor neoepitopes and kill the cancer cells.

Some studies on mice have shown that mutated peptides predicted to bind the MHC complex, which interacts with T cells to trigger the immune response, are effective at blocking tumor growth (Castle et al., 2012; Kreiter et al., 2015). Clinical trials are proving that these treatments are effective on patients with melanoma (Carreno et al., 2015; Sahin et al., 2017) and small cell lung cancer (ref).

Recent studies are showing that splicing deregulation in cancer could hold a great potential as a source of neoepitopes. Aberrant splicing could potentially change open reading frames (ORFs) and generate new peptides, some of which may be presented by the MHC-I or MHC-II complexes, thereby representing potential new triggers of immune response that can be awaken with the use of immune checkpoint inhibitors (Jayasinghe, Cao, Gao, Wendl, Vo, Reynolds, Zhao, Ding, et al., 2018; Smart et al., 2018). In this thesis, we have investigated this question on samples from small-cell lung cancer, one of the deadliest cancers. We have developed a pipeline for the exhaustive identification of neoepitopes from all non-annotated splicing changes, which we have validated using mass spectrometry for MHC-I associated proteins. We describe how splicing alterations lead not only to the generation of neo-epitopes, but also and to a higher extent, to the deletion of native epitopes, hence providing a new mechanism of immune escape. This and the other examples presented evidence how fundamental is to investigate the molecular mechanisms of splicing regulation at designing highly specific therapeutic tools.

1.8. Molecular signatures for staging and prognosis in cancer

Cancer staging is a systematic approach for determining the severity and prospect of a tumor at the time of detection. In 1977, the American Joint Committee on Cancer (AJCC) defined the popularized TNM cancer staging system in the 1st edition of "Manual for staging of cancer". This manual guides clinicians using the size of the tumor (T), the spread to the lymph nodes (N) and the presence or absence of metastasis (M) for determining cancer's anatomy stage. The aim is that according to this staging an estimation of the patient prognosis could be given. There is a list of tests carried out by pathologists in order to determine the staging, but these tests

could be sometimes incomplete or inconclusive. Therefore, it would be desirable to have an unbiased method for determining staging from a molecular perspective, complementing those tests or even better, replacing them.

Breast cancer is one of the most studied cases for staging and several classification signatures have been proposed. One of the most extended ones is Oncotype DX (Partin & Mamounas, 2011) which consists of a PCR assay of 21 genes. This method quantifies risk of distant recurrence and predicts chemotherapy benefit in estrogen receptor positive patients. In fact, in 2017 the AJCC adopted OncotypeDX as the first molecular signature for staging (Giuliano, Connolly, Edge, & Mittendorf, 2017). In other cases markers are based on microarray analysis of gene expression, like MammaPrint, which analyzes the activity of 70 genes in early-stage breast cancer for predicting the use of adjuvant chemotherapy (van de Vijver et al., 2002). Recently, another new signature based on five-lncRNA has been proposed for predicting disease free survival (J. Li et al., 2018). There are other established biomarkers for breast cancer stratification, like estrogen and progesterone receptor, HER2 or Ki67. Along with these, other new prognostic markers have been proposed: measuring expression levels of proteins cyclin E, B-Myb, Twist and DMP1β is informative to predict poor survival and likelihood of metastasis (Inoue & Fry, 2016).

Other cancers are also studied for the identification of prognostic signatures. In colorectal cancer, several have been proposed, from miRNA signatures (J.-X. Zhang et al., 2013) to lncRNA (Xue et al., 2017) or gene expression markers coupled with methylation levels (Liu et al., 2017). In endometrial cancer, a 12-gene expression signature was develop for predicting the risk of lymph node metastasis (Kang et al., 2018). Other molecular markers for different cancer types are described in (Nair, Sandhu, & Sharma, 2018). As expected, there is a remarkable overlap of biomarker genes across various cancer types, including well-known oncogenes and tumor suppressors, like TP53, FAS or PTEN.

Many of these markers are based on measuring gene expression changes. On the other hand, there are works proving that using isoform-level changes is more informative for biological classification tasks than gene-level (Johnson, Dhroso, Hughes, & Korkin, 2018; Z. F. Zhang, Pal, Bi, Tchou, & Davuluri, 2013). In this thesis we present an analysis on exploiting transcript isoform changes for the task of predicting clinical staging and prognosis. We show that, when

optimizing Machine Learning classifiers, transcripts provide more informative biomarkers than genes.

Other proposed markers are based on mutational signatures. These signatures are related with the age of patients, mutagenic exposures (like tobacco smoking) or defects in DNA repair (Alexandrov et al., 2013). Regarding cancer immunotherapy, some studies have found a strong correlation between mutational burden and clinical benefit (Van Allen et al., 2015) while some other did not find such a relation (Hugo et al., 2016). Recently, a pan-cancer analysis highlighted the relation between mutational signatures and neo-epitope burden as relevant markers of immune therapy response, beyond mutational burden (Miao et al., 2018).

A big issue with all these markers is the lack of enough confidence and limited reproducibility on new samples. As a false prediction could have tragic consequences for patients, constant reevaluation of these predictors on new samples is thus necessary in order to obtain the most accurate tools for the clinicians. Nevertheless, encouraging examples like OncoptypeDX evidence how bioinformatics has a key role to play in cancer clinical decision-making.

2 RESULTS

2.1 SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions

This section presents SUPPA2, a method for the calculation of differential splicing across multiple conditions taking into account biological variability. This method was applied on different sets of data, yielding results as good as, and sometimes better than, other methods in much less time. The method alongside an extended explanation of its usage is available online: https://github.com/comprna/SUPPA. The data analyzed in this manuscript is available in the following link: https://github.com/comprna/SUPPA_supplementary_data

Manuscript presented in this section:

> Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). **SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions**. Genome Biology, 19(1). https://doi.org/10.1186/s13059-018-1417-1

2.2 The prognostic potential of alternative transcript isoforms across human tumors

This section presents the development of models trained with splicing changes in tumor samples for prognosis. We investigated whether it is possible to infer signatures for clinical staging and survival in cancer patients by using alternative splicing variations between different clinical groups. All the details of the models built and signatures obtained are detailed in the manuscript and as additional files in the online version.

Manuscript presented in this section:

Trincado, J. L., Sebestyén, E., Pagés, A., & Eyras, E. (2016). The prognostic potential of alternative transcript isoforms across human tumors. Genome Medicine, 8(1). https://doi.org/10.1186/s13073-016-0339-3

2.3 The immunogenic impacts of splicing alterations in small cell lung cancer

This section presents a new method for the identification of non-annotated splicing junctions and their evaluation at immunogenic level. We test this methods on cell lines and small cell lung cancer patients. All the details of the pipeline as well as figures and supplementary tables are available in the following link: http://github.com/comprna/ePydoor

Manuscript presented in this section:

> Trincado, J. L., Reixachs, M., Yokota, J., & Eyras, E. (2018). The immunogenic impacts of splicing alterations in small cell lung cancer. *In preparation.*

2.3.1. Abstract

We describe a novel approach for the exhaustive identification of neo-epitopes from tumor-specific splicing alterations, including aberrant spliced junctions, retained introns and exonizations. Using mass spectrometry for MHC-I associated proteins, we show that splicing derived neo-epitopes are processed and presented by MHC-I complexes. We applied this method to a cohort of 123 small cell lung cancer patients (SCLC) and found that tumor-specific splicing alterations more frequently eliminate than create epitopes, hence uncovering a new mechanism of immune escape in SCLC.

2.3.2. Main

Identification of cancer neo-antigens arising from tumor specific mutations provide an effective way to develop anticancer vaccines (Sahin & Türeci, 2018) and determine the efficacy of immune checkpoint inhibitors (Van Allen et al., 2015). Deregulation of splicing in cancer has been shown to represent an additional source of tumor neo-epitopes from the aberrant selection of splice sites (Jayasinghe, Cao, Gao, Wendl, Vo, Reynolds, Zhao, Climente-González, et al., 2018; Kahles et al., 2018) or through intron retention (Andersen et al., 2013; Smart et al., 2018). Splicing alterations in cancer thus provide a general mechanism to elicit tumor-specific immune responses. However, it is not known yet whether splicing alterations could also provide a strategy for tumors to evade immune-mediated elimination.

We present here a new approach to exhaustively identify the immunogenic impacts from all types of tumor-specific splicing alterations, including aberrant splice sites and intron retentions, as well as new exon skipping and exonization events. Our method identifies tumor-specific splicing alterations from tumor bulk RNA sequencing (RNA-seq) and tests against multiple potential confounding factors (Fig. 1a) (Supp. Fig. 1a) (Methods). Protein changes induced by the splicing alterations are predicted using a reference proteome, and peptides with binding affinity to the major histocompatibility complex class I (MHC-I) or II (MHC-II) are predicted using the corresponding human leukocyte antigen (HLA) types for each patient. The method further calculates candidate MHC-I and MHC-II binders that are gained or lost in the tumor as a consequence of the splicing alterations.

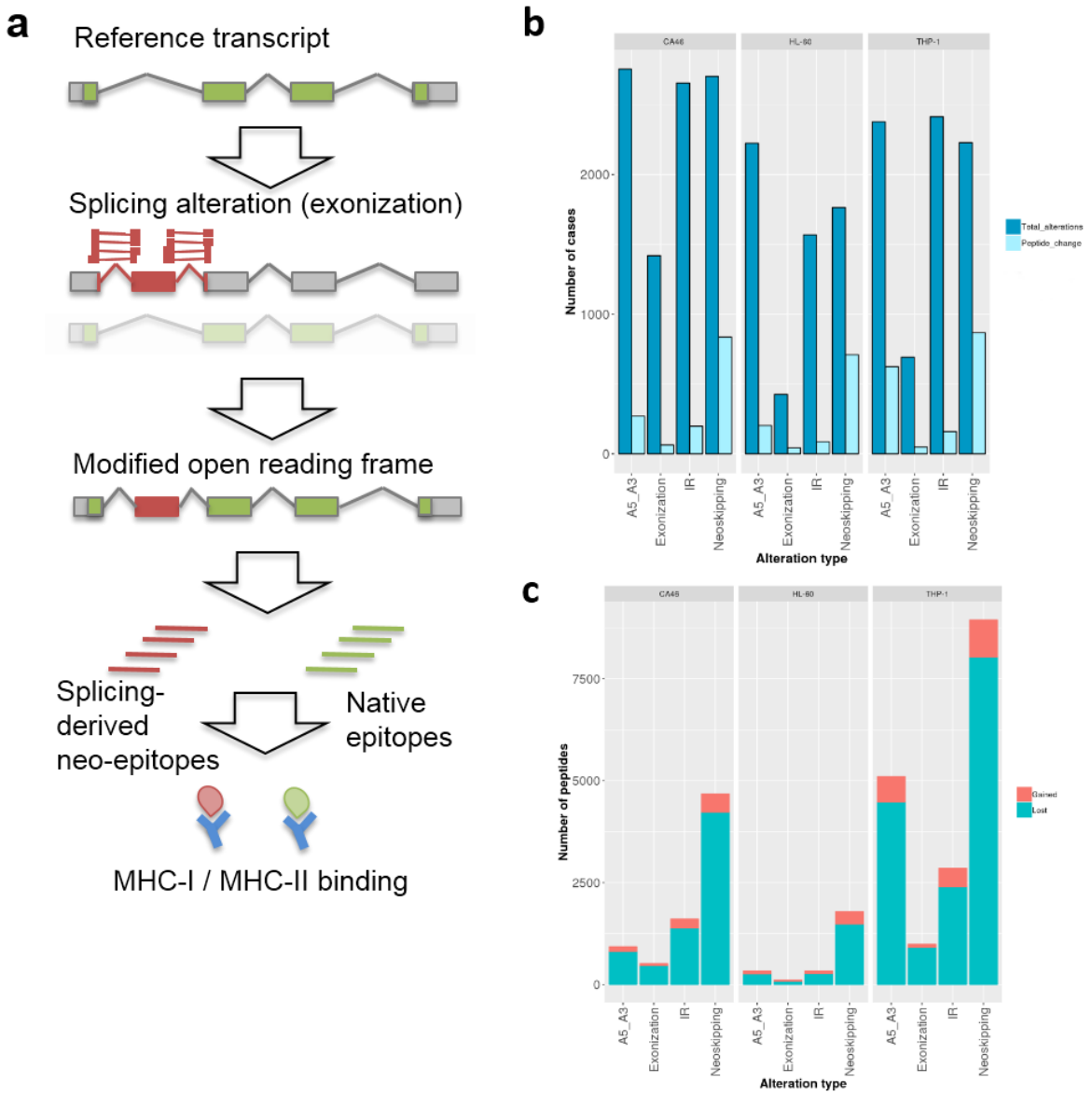


Figure 1. Immunogenic impact of splicing alterations. (a) Illustration of our method to identify splicing-related neo-epitopes. Splicing alterations are defined as variations with respect to the annotation and classified as aberrant splice sites, new exon skippings, new exonizations and intron retention. The figure illustrates an exonization. The reference open reading frame is modified according to the splicing alteration and the reference and modified proteins are tested for affinity with the MHC-I and MHC-II complexes (see Methods for details). (b) Validation of our method using the cell lines CA46, HL-60 and THP-1. We show the number of splicing alterations in each cell line and the fraction producing a change in the encoded open reading frame. (c) Number of candidate splicing-related neo-epitopes detected (gained) and native epitopes that are depleted as a consequence of the splicing alterations (lost) for each of the splicing alterations in each cell line.

To validate our method we analyzed RNA-seq data and MHC-I associated proteomics data for the cancer cell lines CA46, HL-60 and THP-1 (Barretina et al., 2012; Ritz et al., 2016). Aberrant splice sites and new exon skipping events are the most common type of alterations,

followed by intron retentions and exonizations (Fig. 1b). The majority of splicing alterations either does not affect the encoded protein or are rejected for being potentially degraded by non-sense mediated decay (NMD). From the remaining cases, exon skippings and intron retention events potentially produce the largest number of protein changes compared to the other alterations (Fig. 1b). HLA-types for the cell lines were predicted from the RNA-seq data and matched those reported previously (Ritz et al., 2016). We predicted potential MHC-I binders (binding affinity ≤ 500nM) derived from the splicing alterations, keeping only those peptides that did not appear in the cell line reference proteome (Methods). This produced 830 (CA46), 461 (HL-60) and 2072 (THP-1) candidate neo-epitopes (Supp. Table 1) (Fig. 1c). Using MHC-I associated mass-spectrometry (MS) data available for the same cell lines (Ritz et al. 2016), we were able to validate 2 peptides from THP-1 in the genes *PCNP* (FAIGSQTTK) and *RPS10* (LLFKEGVMV) (Supp. Table 2), both appearing from aberrant splice site selection. These results show the potential by any splicing alteration for producing neo-epitopes. To further understand the impacts of splicing alterations in these cell lines, we also calculated the epitopes present in a reference proteome and potentially lost as a consequence of the splicing alterations. We found 1384 (CA46), 270 (HL-60) and 2401 (THP-1) candidate MHC-I binders that would be depleted as a consequence of the splicing alterations. Furthermore, we were able to validate a higher number of these candidate (194) using the MHC-I associated MS data (169) (Supp. Table 2). This suggests that splicing alterations could also deplete self-antigens in a cell.

We applied our methodology to a cohort of 123 small cell lung cancer (SCLC) patients (George et al., 2015; Iwakawa et al., 2015; Peifer et al., 2012; Rudin et al., 2012) (Supp. Table 3) (Supp. Fig. 1b). SCLC is the most aggressive type of lung cancer, with a very early relapse after chemotherapy treatment and an average survival of 5% after 5 years of diagnosis (S. Cooper & Spiro, 2006). (S. Cooper & Spiro, 2006). SCLC is one of the cancer types with largest mutation burden (Fig. 2a), which has been found to correlate with the efficacy of immune therapy in SCLC (Hellmann et al., 2018). Additionally, SCLC presents a significantly higher density of mutations in introns compared to exons (Fig. 2a), which may be indicative of a widespread impact on RNA-processing. Accordingly, SCLC represents a relevant case to investigate how splicing alterations may contribute to neo-epitope burden and impact the antigenicity of tumor cells.

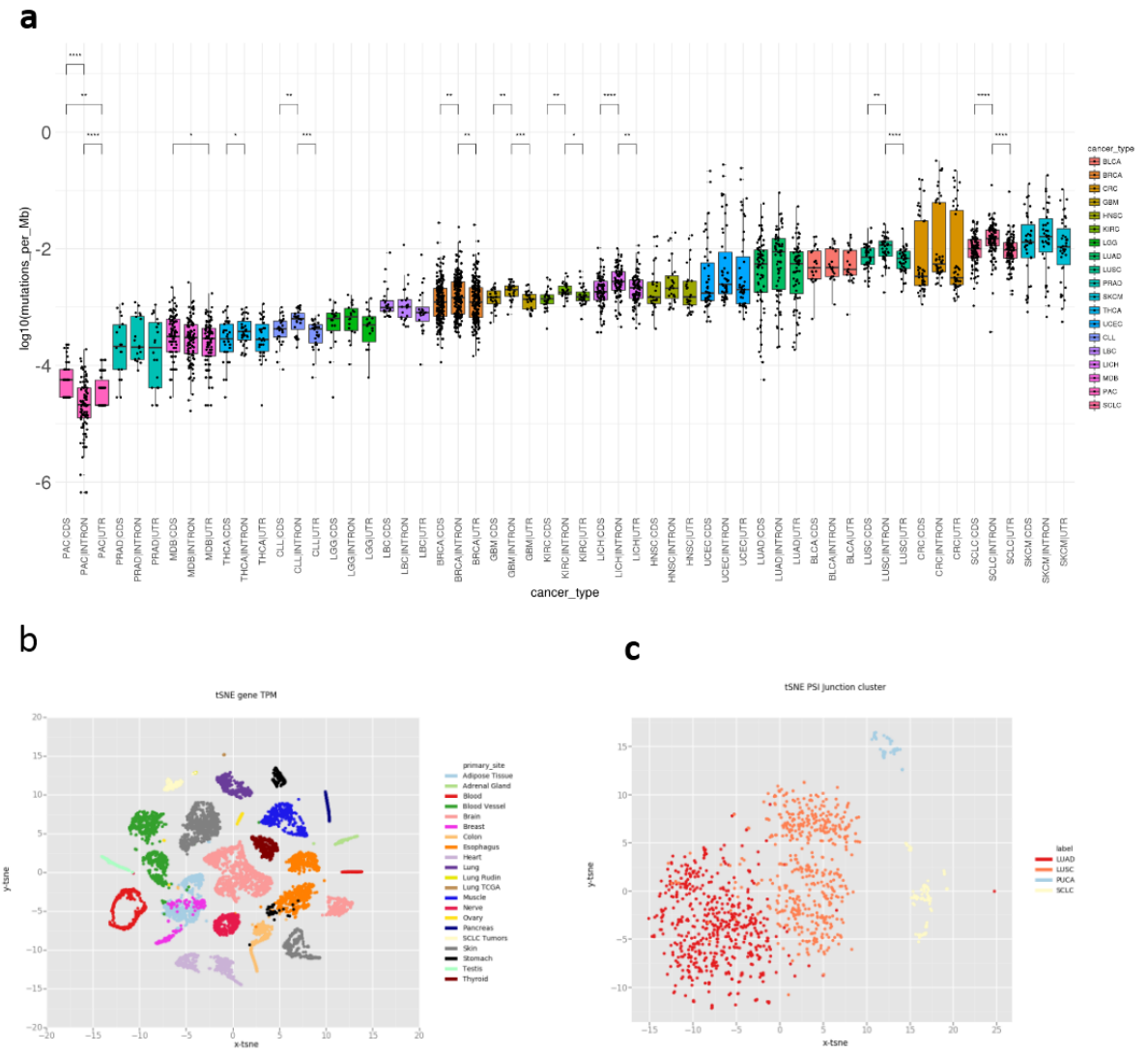


Figure 2. SCLC specific splicing alterations. (a) Mutation burden calculated separately for introns, coding exons and non-coding exons from whole genome sequencing (WGS) data for a number of tumor types, including SCLC. (b) tSNE plot based on gene expression of the SCLC and GTEX samples. (c) PCA/tSNE of tumor-specific junctions in SCLC, together with lung adenocarcinoma (LUAD) samples, lung squamous cell carcinoma (LUSC), and pulmonary carcinoids (PUCA) samples.

An exhaustive compendium of splicing alterations was calculated per sample as described before. To select those alterations that were specific to the SCLC tumors we had to use the appropriate matching controls in terms of tissue identity and cellular content (Aran et al., 2017; Sebestyén et al., 2015). Gene expression provides a signal of tissue identity (Saha et al., 2017) and can help identifying the right controls for the study of splicing alterations associated to cancer (Black et al., 2017). Following this principle, we compared the gene expression patterns of the SCLC samples with the expression from 18 different normal tissues from 7859 individuals (GTEx Consortium, 2015) (Fig. 2b). Normal tissues separate from each other, except for stomach and oesophagus, and for breast and adipose tissue. SCLC patients form a

clearly separated group, with no clear resemblance to other tissues, although the lie close to normal lung samples. This suggests that a comparison with a general compendium of normal splicing profiles would be an appropriate way to extract SCLC specific alterations.

We filtered out all splicing alterations that appeared in the annotations and splicing junctions from a large set of normal samples (GTEx Consortium, 2015; Nellore et al., 2016; Pertea et al., 2018; Rudin et al., 2012) (Methods). We found a total 14643 aberrant splice sites, 7039 intron retentions, 1311 aberrant exon skippings, and 290 new exonizations that are tumor-specific in SCLC, affecting 2955, 149, 620, and 169 genes, respectively (Supp. Fig. 2). These alterations distribute homogeneously across all samples and show no association to tumor mutation burden or overexpression of MYC genes (Fig. 2d). However, the level of detection of aberrant splice sites, new skippings and exonizations depend on the number of spliced reads in a sample (Supp. Fig. 2).

We used this approach to compare the new splicing junctions in our SCLC cohort, with 515 lung adenocarcinomas (LUAD) samples (Cancer Genome Atlas Research Network, 2014), 496 lung squamous cell carcinomas (LUSC) samples (The Cancer Genome Atlas Network, 2012), as well as 69 pulmonary carcinoids (PUCA) (Fernandez-Cuesta et al., 2014), which are lowly proliferative lung malignancies characterized by the expression of neuroendocrine differentiation markers, and potentially similar to SCLC (Bunn et al., 2016; Fernandez-Cuesta et al., 2014). Low dimensional projection of all the tumor-specific junctions from these four cohorts show differences between SCLC samples and the rest of tumors (Fig. 2c), indicating possible SCLC specific biomarkers.

From the set of detected tumor-specific splicing alterations, 3890 (27%) of the aberrant junctions, 85 (29%) of the new exonizations, 753 (10%) of the intron retentions and 804 (61%) of the aberrant exon skippings occurred within the open reading frame (ORF) of the host gene and therefore could potentially impact the encoded protein (Supp. Table 3-6). Using a reference proteome for the SCLC samples we calculated the altered ORF potentially induced by the splicing alterations (Methods). 61% of the new skipping events, 29% of new exonizations, 26% of aberrant splice sites, and 10% of intron retentions, produce a protein change.

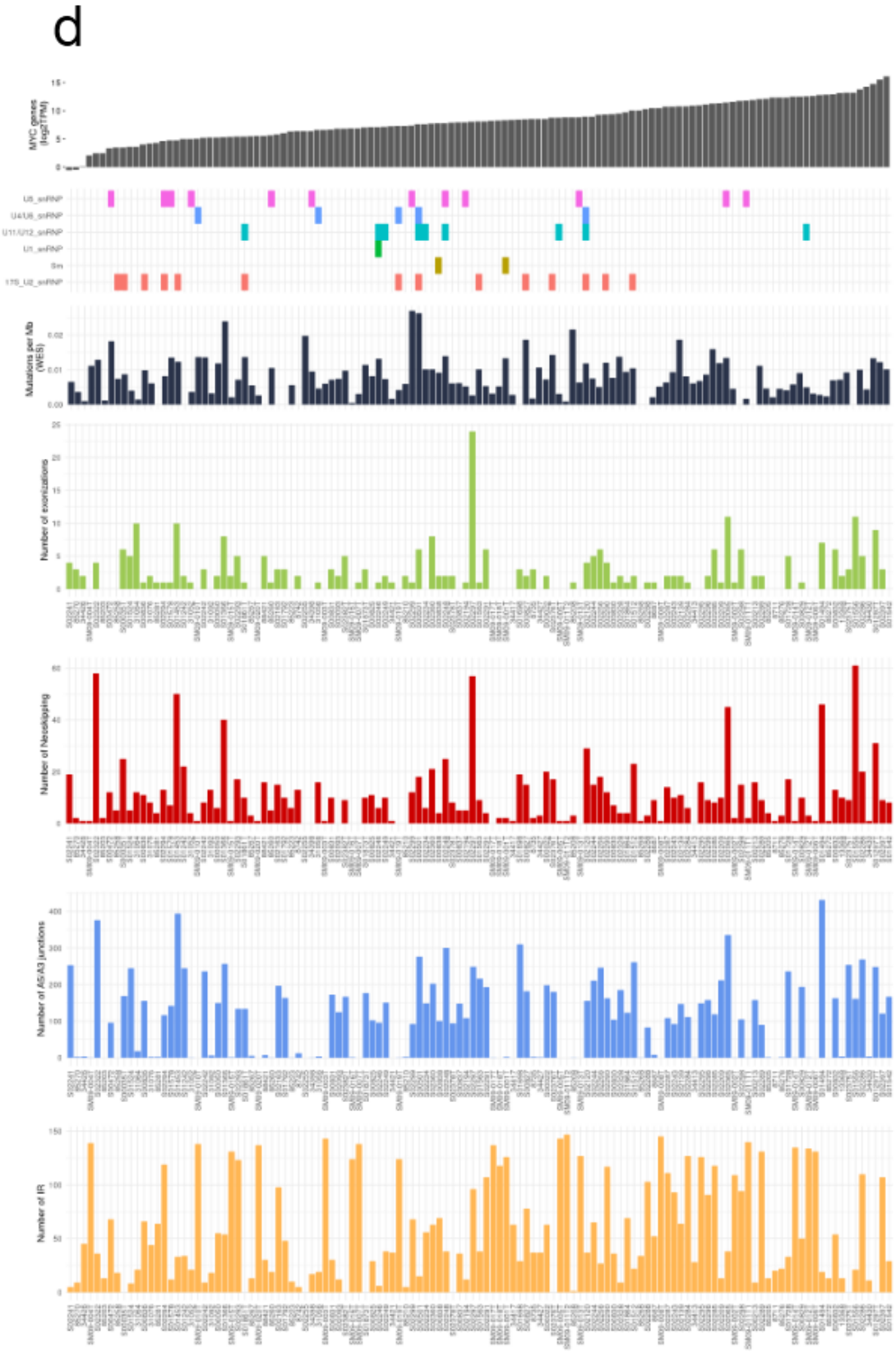


Figure 2. SCLC specific splicing alterations. (d) SCLC specific splicing alterations detected in each patient. Upper panels show the expression of the three MYC genes (MYC+ MYCL + MYCN), the presence of mutations in components of the spliceosomal complexes, and the mutation burden. Mutation data is only given in those samples for which WGS or WES is available

To evaluate the immunogenic impacts induced by the splicing alterations, we predicted HLA-I and HLA-II types from the RNA-seq for the SCLC samples. To validate the identification of HLA-types directly from the tumor sample, we compared the HLA-type predictions from RNA-seq reads from tumor and matched normal samples for 24 SCLC cases (Rudin et al., 2012) (Fig. 3a). Overall, there was 80-90% agreement between the predictions with the tumor and the normal RNA-seq. Next, using the altered and the reference ORFs, we searched for candidate MHC-I binders (binding affinity ≤ 500nM), and those specific of the altered ORFs were considered candidate MHC-I neo-epitopes. We identified a total of 12.422 candidate MHC-I neo-epitopes, with the majority (63%) associated to aberrant splice junctions (Figs. 3b and 3c) (Supp. Fig. 3). Using mass-spectrometry data for MHC-I associated proteins in lymphoblastoid cell lines (Lanoix et al., 2018a) we were able to validate 1458 (11.7%) of the predicted MHC-I neo-epitopes (Supp. Tables 7 and 8). As MHC-II neo-epitopes may be relevant for immunotherapy (Sun et al. 2017; Lu et al. 2017), we also measured the binding affinities with the sample-specific HLA class II alleles. We identified a total of 6618 (7.9%) candidate MHC-II neo-epitopes, with the majority (82%) associated to aberrant splice sites (Supp. Tables 9 and 10). We did not observe any significant correlation of the expression of T cell markers or immune checkpoint genes with the presence of predicted MHC-I or MHC-II neo-epitopes (Supp. Fig. 3). Thus, although splicing-related neo-epitopes may indicate vulnerability for immune therapy in some tumor types (Jayasinghe et al. 2018; Kahles et al. 2018), this might not be the case for SCLC.

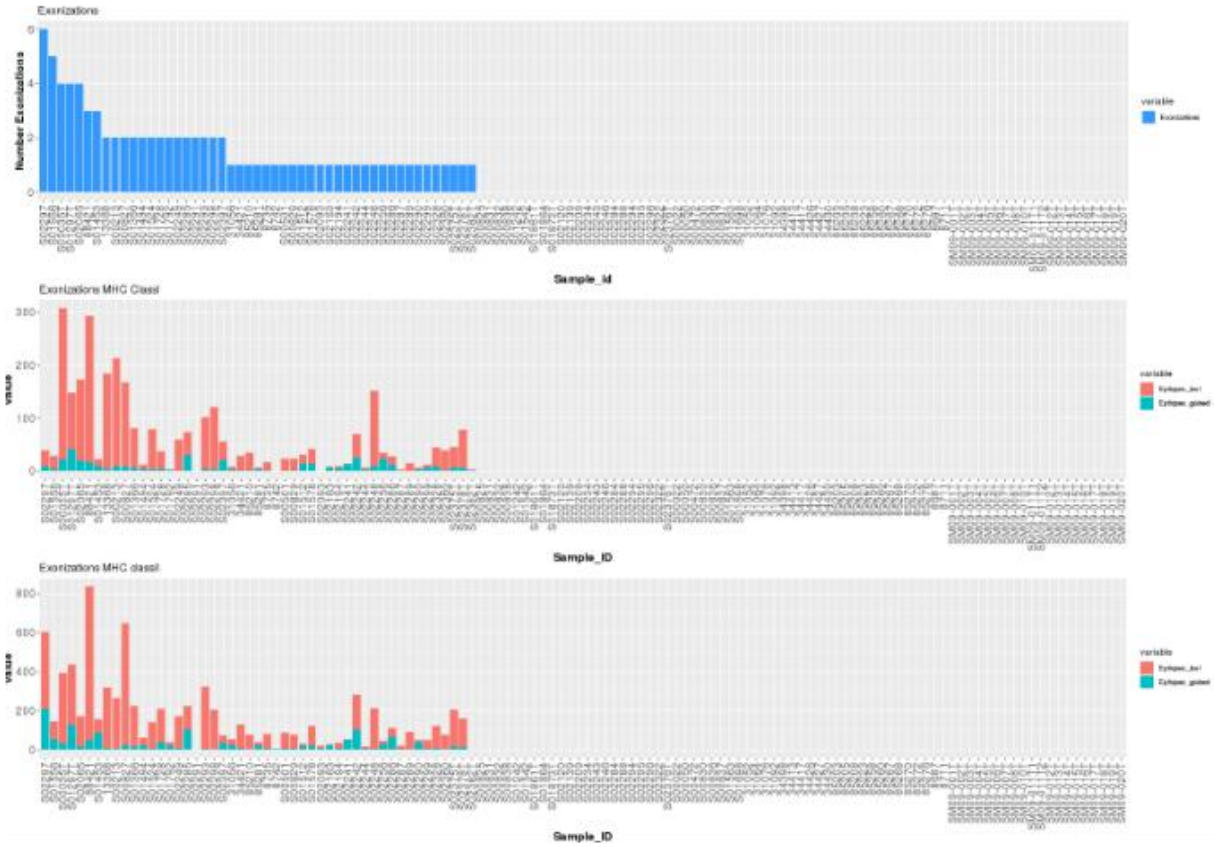
Studies so far have focused on the creation of neo-epitopes through splicing alterations (Jayasinghe et al. 2018; Kahles et al. 2018; Smart et al. 2018). However, splicing alterations in cancer frequently remove protein coding regions (Climente-González, Porta-Pardo, Godzik, & Eyras, 2017); hence, they could lead to the depletion of peptides with immunogenic potential. To test this hypothesis, we calculated whether the splicing alterations would delete MHC-I or MHC-II epitopes that are present in the reference proteome, i.e. they are would not appear in the altered ORFs (Methods). Interestingly, we observed that splicing alterations in SCLC patients more frequently eliminate epitopes than create them. In total, we observed around 7 times more epitopes removed than gained by splicing alterations (Figs. 3b and 3c)(Supp. Fig. 3). This imbalance towards the elimination of epitopes occurred at the level of the number of predicted immunogenic peptides, the number of events producing immunogenic peptides, and the number of genes involved. Moreover, this effect is not specific of any type of splicing

alteration. Finally, mass-spectrometry data for MHC-I associated proteins in lymphoblastoid cell lines validated many more depleted epitopes than neo-epitopes. This suggests that splicing alterations in SCLC could provide a general mechanism for the cancer cell to evade potential immune responses.

2.3.3. Discussion

We have described a new tool to exhaustively determine how splicing alterations may impact the antigenicity of tumor cells, and showed that splicing alterations can contribute with new epitopes as well as eliminate native ones. Our method presents a number of novelties and advantages with respect to previous approaches. It is exhaustive in the type of alterations tested, including new spliced junctions, retained introns and exonizations, making possible an assessment at unprecedented scale of the splicing-derived neo-epitope burden. As the method only requires RNA-seq data from a tumor sample, it is applicable in the absence of DNA sequencing from the patient. Unlike previous studies (Jayasinghe, Cao, Gao, Wendl, Vo, Reynolds, Zhao, Climente-González, et al., 2018; Kahles et al., 2018; Smart et al., 2018), our analysis describes tumor-specific alterations by comparing to a large compendium of normal samples, and testing for potential MHC-II neo-epitopes, which are also relevant for immunotherapy (Lu et al., 2017; Z. Sun, Chen, Meng, Wei, & Liu, 2017).

b


Exonizations
Number Exonizations
Sample_Id
Exonizations MHC ClassI
value
Sample_ID
Exonizations MHC classI
value
Sample_ID
variable

c

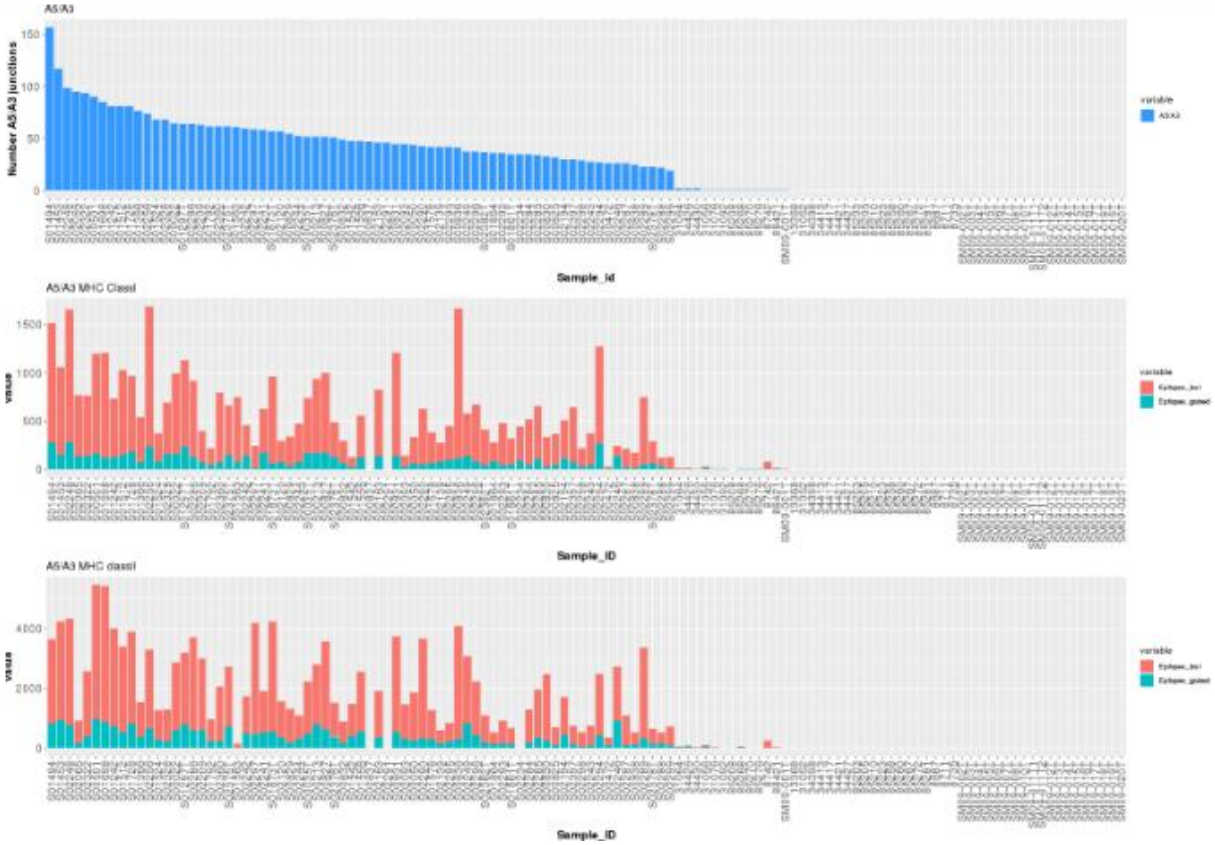

A5/A3
Number A5/A3 junctions
Sample_Id
A5/A3 MHC ClassI
value
Sample_ID
A5/A3 MHC classI
value
Sample_ID
variable

Figure 3. Epitope production and depletion in SCLC patients. (a) For each MHC Class I (HLA-A, HLA-B, HLA-C) and II (HLA-DQA, HLA-DQB, HLA-DRB) prediction from PHLAT (red) or SeqHLA (blue), we show the proportion of RNA-seq samples (over a total of 24) in which the prediction on the tumor sample coincides with the prediction on the matched normal sample. (b) Upper panel: Number of exonizations per SCLC sample that impact the open reading frame (upper panel). Middle panel: Number of candidate MHC-I binders per sample that are created (blue), i.e. neo-epitopes, or depleted (red) through exonizations. Lower panel: Number of candidate MHC-II binders per sample that are created (blue), i.e. neo-epitopes, or depleted (red) through exonizations. (c) Same as in (b) but for aberrant splice sites.

We found that tumor-specific splicing alterations in SCLC more frequently eliminate epitopes than generate new ones, providing a new mechanism of immune escape. Unlike non-small cell lung cancers (NSCLC), SCLC genomes lack known actionable alterations (Bunn et al., 2016; George et al., 2015). Immune checkpoint inhibitors are effective against SCLC (Antonia et al., 2016), and biomarkers of response to therapy are being developed (Hellmann et al., 2018). Our findings provide a new signature of the immunogenic status of SCLC, which could prove relevant for future clinical decision-making.

Tumor specific splicing alterations in SCLC may appear through multiple mechanisms. We only observed a small fraction stemming from somatic mutations in cis or in trans. However, other mechanisms are possible. MYC genes are frequently amplified and/or overexpressed in SCLC, and there is a established link between overexpression of MYC and splicing alterations in cancer through different mechanisms (Hsu et al., 2015; Koh et al., 2015). There are moreover multiple splicing alterations in cancer in the absence of mutations on splicing factors, and these appear lineage specific (Dvinge & Bradley, 2015; Sebestyén, Singh, Miñana, et al., 2016). Thus the SCLC-specific splicing alterations described could be related to lineage specific characteristics.

As the ability of the immune system to identify malignant cells relies on the tumor cells maintaining sufficient antigenicity, it is thus essential to exhaustive explore all potential immunogenic impacts through the variety of splicing alterations that may rise specifically in tumors. We have shown here that tumor-specific splicing alterations contribute not only to the generation of neo-epitopes, but also and in greater proportion to their depletion, hence uncovering a new mechanism to evade recognition by immune cells.

2.3.4. Methods

Datasets

We gathered whole genome sequencing (WGS) data for 100 SCLC patients (George et al., 2015), and whole exome sequencing data (WES) for 103 (Rudin et al., 2012) and 19 (Iwakawa et al., 2015) SCLC patients. Additionally, we gathered RNA sequencing (RNA-seq) data from 123 patients (George et al., 2015; Iwakawa et al., 2015; Peifer et al., 2012; Rudin et al., 2012), 72 of which with WGS data available and 111 with WES data available. We estimated the tumor purity of the SCLC samples from the gene expression data using ESTIMATE R package (v.1.0.13) (Yoshihara et al., 2013). All t-SNE plots have been produced using the package sklearn (Pedregosa et al., 2012). Samples with more than 30% of junctions with missing values were filtered. Remaining missing values were mean-imputed. We used the 100 first principal components for the tSNE generation. For visualization, we used a learning rate of 300 and perplexity of 30.

Identification of tumor-specific splicing alterations

All RNA-seq samples were mapped to the genome (hg19) using STAR (Dobin et al., 2013). As described before (B. Singh, Trincado, Tatlow, Piccolo, & Eyras, 2018), mapped spliced reads with at least a common splice site across two o more samples were clustered using LeafCutter (Y. I. Li et al., 2018), with a minimum of 30 reads per cluster and a minimum fraction of reads of 0.01 in a cluster supporting a junction. Read-counts per junction were normalized over the total of reads in a cluster. Junction clusters were defined across all patients, but normalized read counts were calculated per patient. Junctions were classified as novel if either or both of the splice-sites were not present in the annotation (Gencode v19) (Harrow et al., 2012), they had at least 10 supporting reads it in at least one tumor sample, and did not appear in normal samples (GTEx Consortium, 2015; Nellore et al., 2016; Rudin et al., 2012). For the cancer cell line data we proceeded in the same way, defining junction clusters independently for each cell line. To define exonizations, we considered all pairs of junctions not present in normal samples (GTEx Consortium, 2015; Nellore et al., 2016; Rudin et al., 2012) that would define a potential new internal exon no longer than 500nt, with flanking canonical splice site motifs (AG-GT) on the same strand of the host gene. We kept only cases with more than 5 reads validating each splice site. Tumor specific new exon skippings, we

considered those junctions selected before that moreover were skipping known exons and defining new connections between adjacent exons not present in the normal samples. To define retained introns we used KMA (Pimentel, Conboy, & Pachter, 2015) to extend the Gencode (v19) transcriptome with new potential retained introns (RIs), which we quantified for each RNA-seq sample with Kallisto (Bray et al., 2016). To filtered out RIs that are not tumor specific, we calculated RI events with SUPPA (Trincado et al., 2018) from the Gencode (Harrow et al., 2012) and the CHESS 2.0 (Pertea et al., 2018) annotations, and removed predicted RIs that appeared in these annotations. We also removed RIs that we predicted in normal lung (Rudin et al., 2012). To control for confounding signals due to overall lack of transcript processing, for each splicing alteration, for all types, we compared the expression of the alteration with 100 randomly cases from the same gene, and compared the expression the observed event with the control distribution using an (Empirical Cumulative Distribution Function) ECDF test. Junctions were compared with other junctions, exonizations were compared with genic regions of similar length, and retained introns were compared with other introns. For the cell line data we proceeded in a similar way, but without removing the alterations in normal samples, as those tests were focused on the presentation of splicing-derived neo-epitopes.

Association with somatic mutations

Somatic mutations were filtered out if they overlapped with frequent (>1% allele frequency) SNPs (dbSNP 144). The association of splicing changes to cis mutations was tested by comparing the inclusion value (for the junction), normalized expression (for the RI), or exon coverage (for novel exon) in the mutated sample against the distribution of values in samples without mutations in the same region to obtain a z-score. From the z-score a p-value was obtained, which was corrected for multiple testing using the Benjamini-Hochberg method. We considered mutations falling in the region covering the exon, junction or RI, plus 200nt on either side. This test was carried out for all splicing alterations with mutations nearby, and then the intersection with tumor-specific events was considered.

Protein impact of the splicing alterations

We built a reference transcriptome using the most abundant isoform for each gene, measured as the transcript isoform with the largest and greater than 1 mean expression, in transcripts per million (TPM) units, across all patients. Transcript abundance was calculated using Salmon

(Patro et al., 2017). A reference proteome was defined using the proteins encoded by these reference transcripts. For each splicing alteration, a modified transcript was then built using as scaffold the reference transcript. Unless the splicing alteration only affected the untranslated region (UTR), an altered protein was calculated from the longest open reading frame (ORF) (start to stop) predicted on the modified transcript. If no ORF was predicted, the splicing alteration was not considered. If the splicing alteration deleted the region of the start codon, the closes downstream start codon was used. Further, if the stop codon in the altered ORF was located further than 50nt to a downstream splice site, the case was discarded as potential NMD target. This method is available at http://github.com/comprna/ePydoor

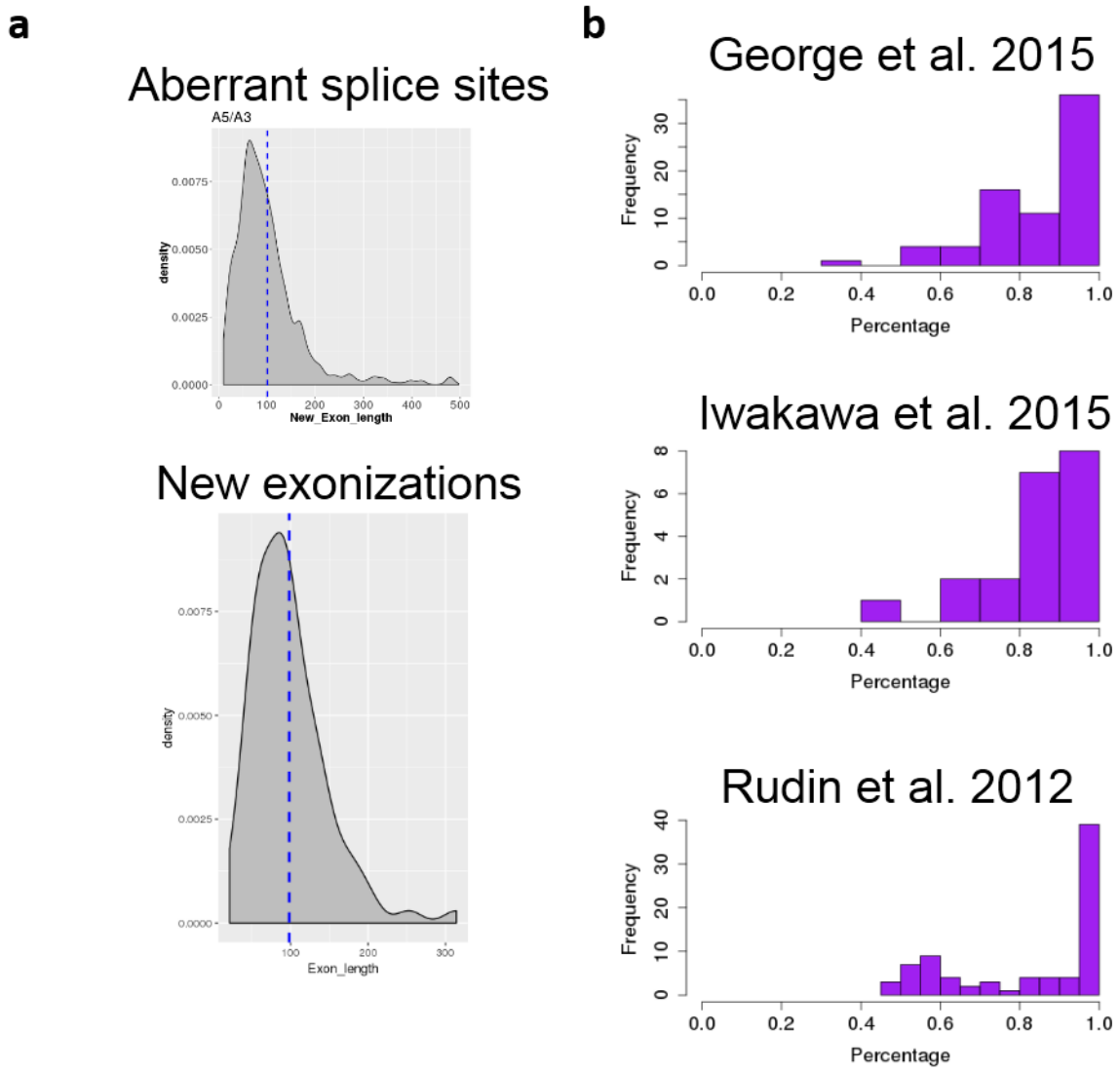
Prediction of neo-epitopes from splicing alterations

We inferred the HLA-type from the tumor RNA-seq for each patient, or from the cell RNA-seq for tested cell lines, using PHLAT (Bai, Ni, Cooper, Wei, & Fury, 2014). We predicted potential MHC-I binders with NetMHC-4.0 (Andreatta & Nielsen, 2016), and with NetMHCpan-4.0 (Jurtz et al., 2017) for the classes missing in NetMHC-4.0.0. Those peptides in common between the reference and the altered protein were discarded. Peptides in the altered protein with binding affinity $\leq$ 500nM, but not present in the reference protein, were considered candidate neo-epitopes; whereas peptides in the reference protein with binding affinity $\leq$ 500nM, but not present in the altered protein with binding affinity $\leq$ 500nM, were considered deleted native epitopes. We performed the same analysis for MHC-II binders using predictions from NetMHCII-2.3, and complementing them with the predictions from NetMHCIIpan-3.2 for the missing types (Jensen et al., 2018). Candidate epitopes created or eliminated by splicing alterations are given in the supplementary material.

Validation of neo-epitope prediction

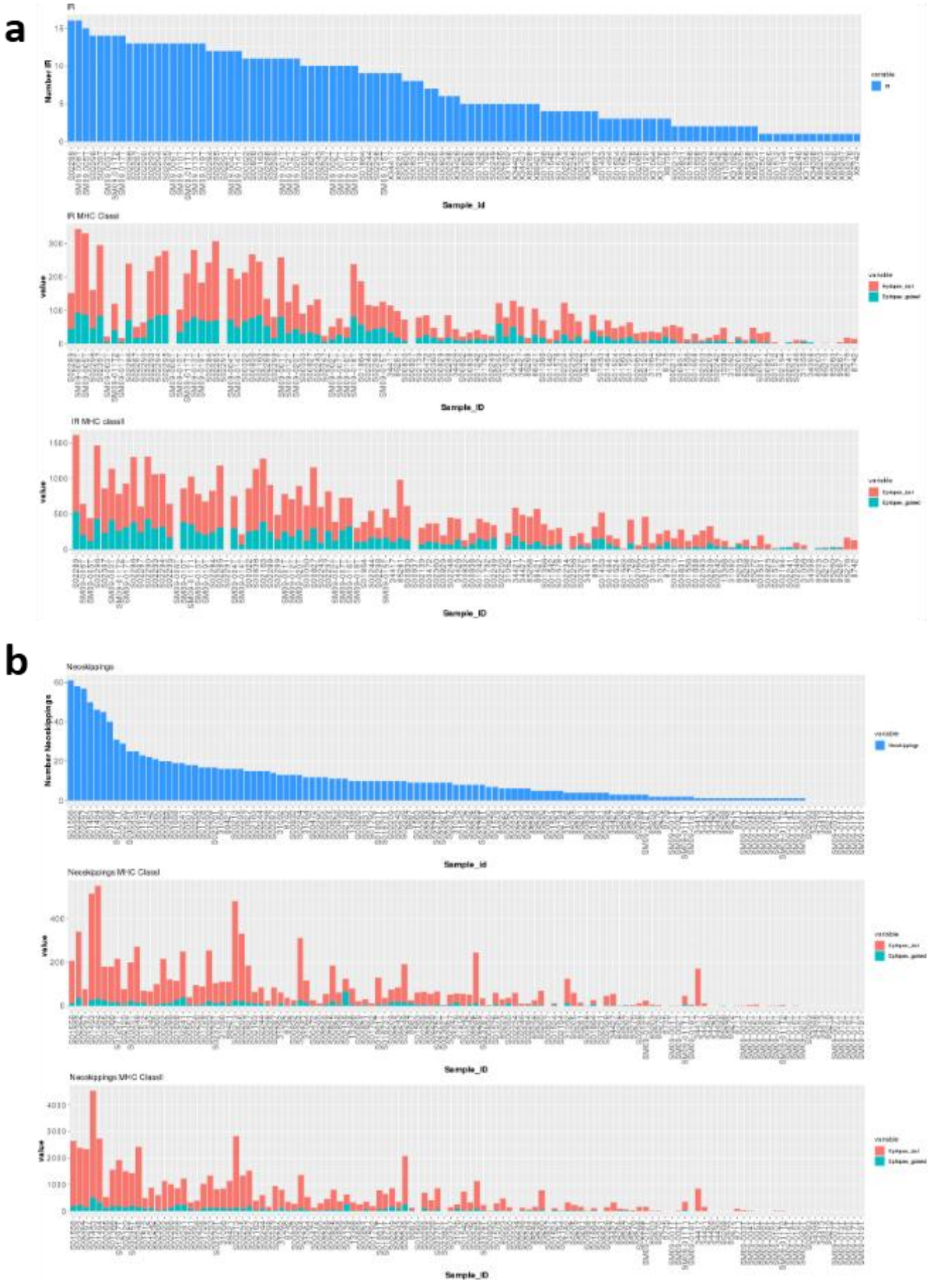
We used RNA-seq data and MHC-I associated mass-spectrometry (MS) data for the cell lines CA46, HL-60 and THP-1. Peptides from the MS experiments were compared against the candidate MHC-I binders derived from the splicing alterations in each cell line. A match was considered if the MHC-I binder was exactly equal to of was completely included in the MS peptide. For the SCLC samples we used MHC-I associated MS from lymphoblastoid cell lines (Lanoix et al., 2018a). Candidate MHC-I binders, either generated or deleted by the splicing alterations, were validated in a similar way.

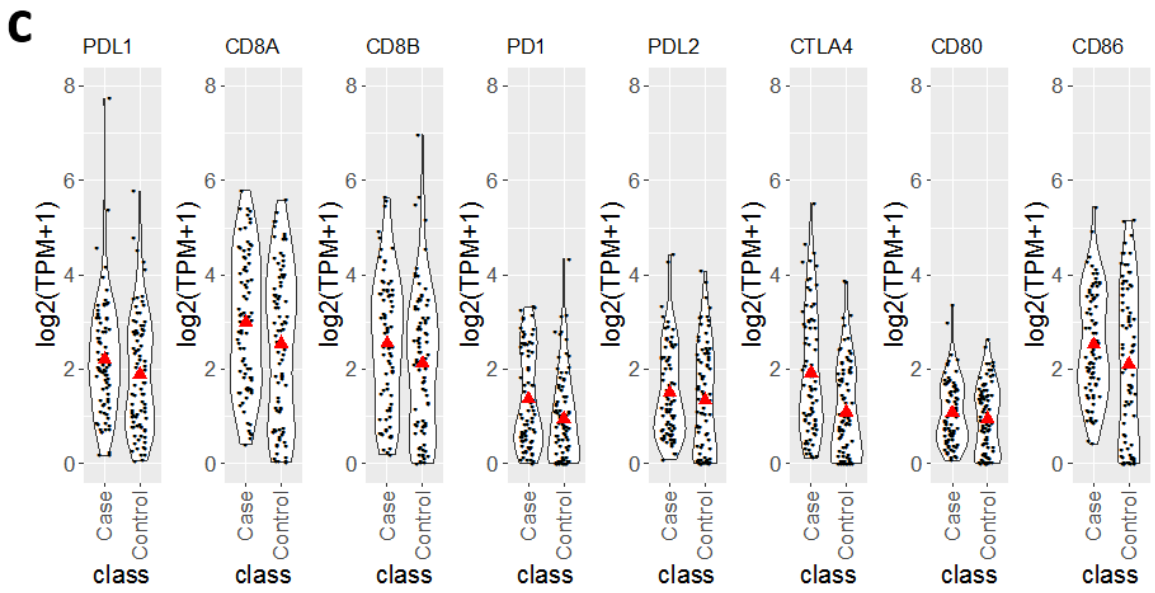
2.3.5. Supplementary figures



Supp. Figure 1. **(a)** Length distributions of the new exons produced as a consequence of aberrant splice sites (upper panel) or new exonizations (lower panel). The lengths follow a extreme value distribution with mean value 100, similar to known exons. **(b)** Purity of the SCLC samples. For each one of the three cohorts used for this study, we give the distribution of tumor purity values (between 0 and 1), calculated with ESTIMATE (Yoshihara et al., 2013).

Supp. Fig. 2. SCLC specific splicing alterations that impact encoded proteins. For each SCLC sample we plot the number of SCLC specific alterations that impact the encoded protein of the host gene (after removing the cases leading to NMD), as a function of the number of spliced reads mapped in the sample (top panel) and the total number of reads mapped uniquely (second panel).


a
IR
Number IR
Sample_id
IR MHC Classl
value
Sample_ID
IR MHC classII
value
Sample_ID
b
Neoskippings
Number Neoskippings
Sample_id
Neoskippings MHC Classl
value
Sample_ID
Neoskippings MHC ClassII
value
Sample_ID



Supp. Fig. 3. Epitope production and depletion in SCLC patients. **(a)** Upper panel: Number of intron retentions per SCLC sample that impact the open reading frame (upper panel). Middle panel: Number of candidate MHC-I binders per sample that are created (blue), i.e. neo-epitopes, or depleted (red) through exonizations. Lower panel: Number of candidate MHC-II binders per sample that are created (blue), i.e. neo-epitopes, or depleted (red) through exonizations. **(b)** Same as in (a) but for new exon skippings. **(c)** Expression of various immune cell markers and immune checkpoint genes in samples with (case) and without (control) predicted MHC-I neo-epitopes.

2.3.6. Author contributions

E.E. proposed the study. J.L.T. developed the software and performed the analyses. M.R. perform the analysis on the mutations. E.E. supervised the analyses. E.E. and J.L.T. wrote the manuscript with essential inputs from J.Y.

2.3.7. Acknowledgements

This work was supported by the Spanish Government and FEDER with grant BIO2017-85364-R, and by AGAUR with grants SGR2017-1020.

2.3.8. Additional files

Available online in http://github.com/comprna/ePydoor

Additional file 1: Supplementary Tables (XLSX)

2.3.9. References

1. Andersen RS, Andersen SR, Hjortsø MD, Lyngaa R, Idorn M, Køllgård TM, Met O, Thor Straten P, Hadrup SR. 2013. High frequency of T cells specific for cryptic epitopes in melanoma patients. Oncoimmunology 2: e25374. http://www.ncbi.nlm.nih.gov/pubmed/24073381.
2. Andreatta M, Nielsen M. 2016. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics 32: 511–7. http://www.ncbi.nlm.nih.gov/pubmed/26515819.
3. Antonia SJ, López-Martin JA, Bendell J, Ott PA, Taylor M, Eder JP, Jäger D, Pietanza MC, Le DT, de Braud F, et al. 2016. Nivolumab alone and nivolumab plus ipilimumab in recurrent small-cell lung cancer (CheckMate 032): a multicentre, open-label, phase 1/2 trial. Lancet Oncol 17: 883–895. http://www.ncbi.nlm.nih.gov/pubmed/27269741.
4. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, Goga A, Sirota M, Butte AJ. 2017. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun 8: 1077. http://www.ncbi.nlm.nih.gov/pubmed/29057876.
5. Bai Y, Ni M, Cooper B, Wei Y, Fury W. 2014. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics 15: 325. http://www.ncbi.nlm.nih.gov/pubmed/24884790.
6. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov G V, Sonkin D, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603–7. http://www.ncbi.nlm.nih.gov/pubmed/22460905.
7. Black KL, Naqvi AS, Hayer KE, Yang SY, Gillespie E, Bagashev A, Pillai V, Tasian S, Gazzara MR, Carroll M, et al. 2017. Aberrant splicing in B-cell acute lymphoblastic leukemia. bioRxiv. http://biorxiv.org/content/early/2017/12/01/225136.abstract.
8. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34: 525–7. http://www.ncbi.nlm.nih.gov/pubmed/27043002.
9. Bunn PA, Minna JD, Augustyn A, Gazdar AF, Ouadah Y, Krasnow MA, Berns A, Brambilla E, Rekhtman N, Massion PP, et al. 2016. Small Cell Lung Cancer: Can Recent Advances in Biology

and Molecular Biology Be Translated into Improved Outcomes? J Thorac Oncol 11: 453–74. http://www.ncbi.nlm.nih.gov/pubmed/26829312.

10. Cancer Genome Atlas Research Network. 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511: 543–50. http://www.ncbi.nlm.nih.gov/pubmed/25079552.
11. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. 2017. The Functional Impact of Alternative Splicing in Cancer. Cell Rep 20: 2215–2226. http://www.ncbi.nlm.nih.gov/pubmed/28854369.
12. Cooper S, Spiro SG. 2006. Small cell lung cancer: treatment review. Respirology 11: 241–8. http://www.ncbi.nlm.nih.gov/pubmed/16635081.
13. 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.
14. Dvinge H, Bradley RK. 2015. Widespread intron retention diversifies most cancer transcriptomes. Genome Med 7: 45. http://www.ncbi.nlm.nih.gov/pubmed/26113877.
15. Fernandez-Cuesta L, Peifer M, Lu X, Sun R, Ozretić L, Seidel D, Zander T, Leenders F, George J, Müller C, et al. 2014. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. Nat Commun 5: 3518. http://www.ncbi.nlm.nih.gov/pubmed/24670920.
16. George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, Leenders F, Lu X, Fernández-Cuesta L, Bosco G, et al. 2015. Comprehensive genomic profiles of small cell lung cancer. Nature 524: 47–53. http://www.ncbi.nlm.nih.gov/pubmed/26168399.
17. GTEx Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648–60. http://www.ncbi.nlm.nih.gov/pubmed/25954001.
18. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760–74. http://www.ncbi.nlm.nih.gov/pubmed/22955987.
19. Hellmann MD, Callahan MK, Awad MM, Calvo E, Ascierto PA, Atmaca A, Rizvi NA, Hirsch FR, Selvaggi G, Szustakowski JD, et al. 2018. Tumor Mutational Burden and Efficacy of Nivolumab Monotherapy and in Combination with Ipilimumab in Small-Cell Lung Cancer. Cancer Cell 33: 853–861.e4. http://www.ncbi.nlm.nih.gov/pubmed/29731394.
20. 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.
21. Iwakawa R, Kohno T, Totoki Y, Shibata T, Tsuchihara K, Mimaki S, Tsuta K, Narita Y, Nishikawa R, Noguchi M, et al. 2015. Expression and clinical significance of genes frequently

mutated in small cell lung cancers defined by whole exome/RNA sequencing. Carcinogenesis 36: 616–21. http://www.ncbi.nlm.nih.gov/pubmed/25863124.

22. Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, Zhao Y, Climente-González H, Chai S, Wang F, et al. 2018. Systematic Analysis of Splice-Site-Creating Mutations in Cancer. Cell Rep 23: 270–281.e3. http://www.ncbi.nlm.nih.gov/pubmed/29617666.
23. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, Nielsen M. 2018. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology 154: 394–406. http://www.ncbi.nlm.nih.gov/pubmed/29315598.
24. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol 199: 3360–3368. http://www.ncbi.nlm.nih.gov/pubmed/28978689.
25. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, Cancer Genome Atlas Research Network, et al. 2018. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell 34: 211–224.e6. http://www.ncbi.nlm.nih.gov/pubmed/30078747.
26. 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.
27. Lanoix J, Durette C, Courcelles M, Cossette É, Comtois-Marotte S, Hardy M-P, Côté C, Perreault C, Thibault P. 2018. Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. Proteomics 18: e1700251. http://www.ncbi.nlm.nih.gov/pubmed/29508533.
28. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK. 2018. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet 50: 151–158. http://www.ncbi.nlm.nih.gov/pubmed/29229983.
29. Lu Y-C, Parker LL, Lu T, Zheng Z, Toomey MA, White DE, Yao X, Li YF, Robbins PF, Feldman SA, et al. 2017. Treatment of Patients With Metastatic Cancer Using a Major Histocompatibility Complex Class II-Restricted T-Cell Receptor Targeting the Cancer Germline Antigen MAGE-A3. J Clin Oncol 35: 3322–3329. http://www.ncbi.nlm.nih.gov/pubmed/28809608.
30. Nellore A, Jaffe AE, Fortin J-P, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips Iii RA, Karbhari N, Hansen KD, Langmead B, et al. 2016. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol 17: 266. http://www.ncbi.nlm.nih.gov/pubmed/28038678.
31. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. http://www.ncbi.nlm.nih.gov/pubmed/28263959.

32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2012. Scikit-learn: Machine Learning in Python. J Mach Learn Res 12: 2825–2830. http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490.
33. Peifer M, Fernández-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, Plenker D, Leenders F, Sun R, Zander T, et al. 2012. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet 44: 1104–1110.
34. Pertea M, Shumate A, Pertea G, Varabyou A, Chang Y-C, Madugundu AK, Pandey A, Salzberg S. 2018. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. bioRxiv. http://biorxiv.org/content/early/2018/05/29/332825.abstract.
35. Pimentel H, Conboy JG, Pachter L. 2015. Keep Me Around: Intron Retention Detection and Analysis.
36. Ritz D, Gloger A, Weide B, Garbe C, Neri D, Fugmann T. 2016. High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. Proteomics 16: 1570–80. http://www.ncbi.nlm.nih.gov/pubmed/26992070.
37. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, Bergbower EA, Guan Y, Shin J, Guillory J, et al. 2012. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet 44: 1111–6. http://www.ncbi.nlm.nih.gov/pubmed/22941189.
38. Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, GTEx Consortium, Engelhardt BE, Battle A. 2017. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res 27: 1843–1858. http://www.ncbi.nlm.nih.gov/pubmed/29021288.
39. Sahin U, Türeci Ö. 2018. Personalized vaccines for cancer immunotherapy. Science 359: 1355–1360. http://www.ncbi.nlm.nih.gov/pubmed/29567706.
40. 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. http://www.ncbi.nlm.nih.gov/pubmed/27197215.
41. 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: 1345–56. http://www.ncbi.nlm.nih.gov/pubmed/25578962 (Accessed January 4, 2018).
42. Singh B, Trincado JL, Tatlow PJ, Piccolo SR, Eyras E. 2018. Genome sequencing and RNA-motif analysis reveal novel damaging noncoding mutations in human tumors. Mol Cancer Res 16.
43. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, Fugmann T, Wong K-K, Van Allen EM. 2018. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol. http://www.ncbi.nlm.nih.gov/pubmed/30114007.

44. Sun Z, Chen F, Meng F, Wei J, Liu B. 2017. MHC class II restricted neoantigen: A promising target in tumor immunotherapy. Cancer Lett 392: 17–25. http://www.ncbi.nlm.nih.gov/pubmed/28104443.
45. The Cancer Genome Atlas Network. 2012. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489: 519–525.
46. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol 19: 40. http://www.ncbi.nlm.nih.gov/pubmed/29571299.
47. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Foppen MHG, Goldinger SM, et al. 2015. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science 350: 207–211. http://www.ncbi.nlm.nih.gov/pubmed/26359337.
48. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine D a, et al. 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 4: 2612. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3826632&tool=pmcentrez&rendertype=abstract.

3 DISCUSSION

This thesis could be summarized as producing two main outputs: The elaboration of methods for the study of alternative splicing changes and their application in real data for the extraction of RNA alterations from cancer patients that could be relevant for clinical therapeutics. Chapter 2.1 presented a method for the study of differential splicing across multiple conditions: SUPPA2. We showed that this method allows for rapid and accurate quantification of splicing changes and offer useful functionalities not provided by previous existing methods. Chapter 2.2 provided evidence about how we can use these splicing changes to train models for the prediction of clinical staging and survival in cancer patients. Importantly, these models highlight the higher granularity of transcriptomic changes compared with gene expression changes and how they could be useful for obtaining new biomarkers. Finally, chapter 2.3 expanded the study of alternative splicing to all non-annotated changes to detect aberrant spliced products that could be relevant for cancer immunotherapy. All these points are thoroughly analyzed in the discussion of the manuscripts. The following discussion is focused on analyzing the impact of the results for the scientific community, reviewing the limitations of the analyses and commenting future perspectives.

3.1. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions

3.1.1 Main findings

In a previous version of this method, we developed a software oriented to exploit the high speed of state-of-the-art transcriptome quantification algorithms, like Salmon or Kallisto (Bray et al., 2016; Patro et al., 2017) to quantify alternative splicing events. We proved that SUPPA was able to analyze splicing with comparable or higher accuracy than other methods but at an unforeseen speed (G.P. Alamancos et al., 2015). In this new version, SUPPA2, we incorporated new necessary functionalities, including differential splicing calculation and event clustering. Additionally, we showed again the speed improvement respect other methods without compromising accuracy, and demonstrated experimentally the veracity of the values predicted. The method is open source and available to all scientific community.

3.1.2. Impact of the presented research

There is a great variety of methods for splicing quantification (subsection 1.3). To our knowledge, SUPPA2 is the first one that makes use of transcript quantification to infer differential splicing. This make this method incredibly fast compared with others. In addition, the tool is agnostic of the transcript quantification method making it easier to integrate in other pipelines.

SUPPA2 provides a fast way of quantifying splicing events and isoforms. Nevertheless, the fact of finding significant PSI changes does not necessarily mean they are biologically meaningful. It is necessary to perform downstream analysis to put in context this information, e.g. by finding possible mechanisms that would explain these changes. In order to fulfill this aspect, we incorporated the ability to perform event clustering. This is an interesting feature that could be useful for identifying splicing networks of events sharing common mechanisms or functions. Since not all the significant events are generally driven by the same mechanisms, this clusters could be helpful for guiding downstream RBP motif discovery (Carazo, Romero, & Rubio, 2018).

Since we released SUPPA2, we have noticed a great feedback from many users. Several labs are using it for their analysis and they have helped us to keep on improving the method by detecting possible issues or suggesting new features. This makes us think that this tool is appreciated and useful for the scientific community.

3.1.3. Limitations

SUPPA2 relies on a genome annotation to define events. For species poorly annotated, this could be a problem since a lot of the transcripts are not annotated, hence cannot be measured. Depending on the species, a method that is not dependent on the annotation could be more suitable.

Another limitation, already commented in the introduction (subsection 1.3), is that there is no clear agreement between different methods. In spite of this, the results from each method made sense biologically. This highlights how crucial it is to try different approaches and avoid relying on a single method. The differences in the definition of what a splicing change is or how to test for differential splicing probably leads to each method detecting a different portion of the biological variability. Since this an intrinsic limitation for any method, we encourage users not to rely on a single method.

SUPPA2 is coded in python. This is currently one of the most used languages in bioinformatics (Russell et al., 2018). The growing community of developers behind python tools for bioinformatics is one of their strengths. Nevertheless, there are faster programming languages, like C, that could improve even more SUPPA2 performance. In addition, we have noticed an excess of RAM usage when dealing with big datasets (>500 samples) that could be improved in future implementations.

Finally, the definition of the splicing events measured by SUPPA2 is based on the most common binary variations (subsection 1.3, figure 4). Nevertheless, other complex splicing changes may occur that may be left undetected. Other methods, like MAJIQ (Vaquero-Garcia et al., 2016), are not restricted to a set of predefined events an allow the discovery of these complex events (Figure 7).

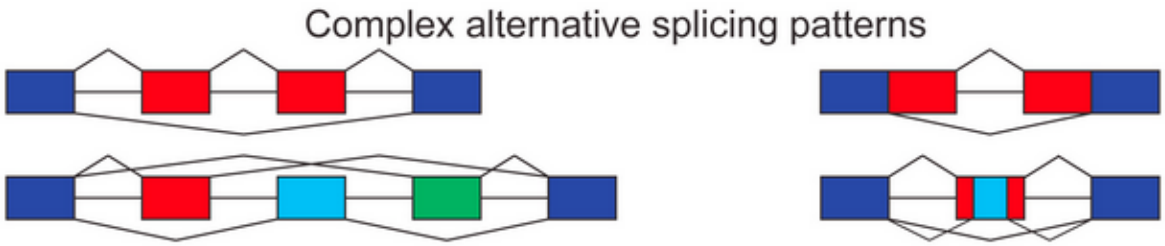


Figure 7: Complex splicing patterns involving non-binary choices of splice site or exons *(Park, Pan, Zhang, Lin, & Xing, 2018)*

3.1.4. Future perspectives

There are some improvements to be made on SUPPA2. As we have mentioned, downstream analysis is necessary for explaining the possible PSI changes (Subsection 3.1.2.). In our lab, we developed a tool for exploiting the PSI changes from SUPPA2 to perform a motif scan and enrichment analysis of RBPs: MoSEA (B. Singh et al., 2018). This method is available online: https://github.com/comprna/MoSEA. In the same way, we want to extend the functionalities of SUPPA2 with some other methods for RBP/SF discovery.

The method has been tested on Illumina bulk RNA-seq data. This is still the standard technology for splicing analysis. Nevertheless, the emergence of single cell technologies are opening new horizons in research, especially in cancer (Liang & Fu, 2017). There are some methods propose for analyzing splicing in single cell RNA-seq experiments (Huang & Sanguinetti, 2017; Song et al., 2017). A prospective work would be to test how well performs SUPPA2 on single cell data. Since SUPPA2 relies on transcript quantification, if these values are adapted for their proper calculation with this technology, our method could work accurately. Kallisto has already an option for this (https://pachterlab.github.io/kallisto/singlecell.html) and Salmon authors have developed a new method for this purpose (Srivastava, Smith, Sudbery, & Patro, 2018). Long read technologies also offer new opportunities to interrogate the transcriptome in multiple contexts (Garalde et al., 2018). Provided a normalized molecular count from long reads, SUPPA2 can already be used to study differential splicing and differential transcript usage with a reference annotation. The adaptation to new transcriptomes derived from the long-read sequencing will require further developments.

3.2. The prognostic potential of alternative transcript isoforms across human tumors

3.2.1 Main findings

This chapter presented a study of the predictive power that splicing changes hold regarding staging and survival of cancer patients. Transcript and event signatures were derived from RNA-seq samples from 12 cancer types from TCGA. These signatures yielded good accuracies to separate between early and late stage patients, especially at predicting metastasis and lymph node invasion. Blind prediction on non-labelled samples was validated by finding a significant separation according to survival. The same methodology was applied for extracting signatures in tumor subtypes in breast cancer and melanoma.

3.2.2. Impact of the presented research

The presented study is a proof of concept on the relation between splicing and clinical staging. Previous works have investigated the possibility of finding cancer biomarkers for stratification of patients using gene expression (Subsection 1.8). In Z. F. Zhang et al., 2013 the authors show how isoform expression changes discriminate better between cancer and non-cancer cell lines than gene expression changes. Our study is the first one that explores the possibility of using inclusion levels of isoforms and events for patient classification.

The obtained signatures are available as supplementary data. The importance of this study lies not on the specific signatures, but on the fact that relative isoform expression describes a more detailed picture of the underlying mechanisms of cancer progression than gene expression. We did not observe a conserved signature across cancer types, indicating lineage specific mechanisms. However, we saw a consistent enrichment of genes involved in DNA repair, MYC targets and mTORC signaling for all cancer types, evidencing a phenotypic convergence related to the RNA processing alterations. Additionally, we observed splicing changes previously described in the literature related with cancer progression.

We performed this analysis by comparing samples classified as early and late according to TNM staging. Clinicians normally perform this classification without prior molecular

characterization. Recently, albeit only for certain cancer types like breast cancer, molecular characterization has started to be used to assess patient staging (Giuliano et al., 2017). Importantly, we also applied this methodology for ER- subtypes in breast cancer patients, identifying a transcript signature that separates between early and late patients with a significant association with survival. This is especially interesting since gene markers for this subtypes are harder to determine (Taherian-Fard, Srihari, & Ragan, 2014). The results of this work support the use of splicing changes as possible molecular markers for cancer staging. Our results also indicate that these models could be particularly useful in cases where it is not possible to determine if there is metastasis or lymph node invasion.

3.2.3. Limitations and future work

The described workflow extracts features (isoforms or events) that hold a predictive value individually and obtains the minimum set of features that explains the greater variance of the dataset. We used subsampling and randomization through the process and assessed the performance of the models via cross-validation to reduce as much as possible potential biases. Nevertheless, we are conscious that there could be still potential confounding factors. These are more likely to have an impact in the tumor types with a low number of patients available.

A possible limitation is the fact that we used univariate feature selection, i.e. we tested the contribution of each feature (isoform) individually. However, several works have evidenced how splicing factors intervene in cellular programs by remodeling entire splicing networks in cancer (Germann, Gratadou, Dutertre, & Auboeuf, 2012). It would be interesting to improved the analysis by combining the contribution of potentially related features at discriminating across stages. In this context, multivariate feature selection techniques would be effective, although they would imply a higher computational burden.

We built our predictive models using logistic model tress. The election of this or another model in this and similar studies is usually based on the accuracy observed in the cross-fold validation, but can be considered to be an arbitrary choice, hence introducing potential biases. A comparison with other classifiers (random forests, neural networks) would be helpful to assess the objectivity of the predictive models.

Each cancer type presents very different cellular and molecular characteristics; and the same stage in different cancer types could have different implications for the patient. Moreover, within each cancer stage there may be additional molecular subgroups and substages that could be critical for the patient outcome. Therefore, some of the signatures obtained here could be too coarse-grained and therefore less relevant for prognosis. On the other hand, we have shown that our methodology is applicable also for each substage. For some cancer types, it could be more interesting to focus on specific subtypes that are still difficult to characterize, like ER- breast cancer.

If clinicians have access to the tumor, TNM staging is generally simple to perform and robust. Therefore, molecular markers might not be necessary. However, there are some cancer types for which access to the affected organ and extraction of the tumor mass is difficult or impossible, like liver or biliary cancer. For these cancers biopsies are difficult to perform. Alternatives like liquid biopsies are a promising way to follow cancer progression and to perform regular screenings. Recent works are obtaining splicing changes using this technology (Bao et al., 2018). Since we have performed this analysis on solid tumors from TCGA, an interesting improvement would be to apply the same methodology on liquid biopsies for determining the stage of a tumor.

As we mentioned in the introduction, all these markers need a constant evaluation on new sets of patients. Since we finished this work, more databases of cancer samples have become available, e.g. through the International Cancer Genome Consortium (ICGC). Prospective work would include the reevaluation of the signatures on new sets of patients.

3.3. The immunogenic impacts of splicing alterations in small cell lung cancer

3.3.1. Main findings

This chapter presented an in silico strategy for the study of epitopes produced by splicing alterations. We developed a pipeline that allows for the identification of aberrant splice sites, exonizations, new exon skippings and intron retentions that are tumor specific and measure their individual immunogenic impact. We tested this method on cell lines and tumor samples from small cell lung cancer (SCLC) obtaining epitopes produced by these alterations and validating them by proteomics data.

3.3.2. Impact of the presented research

During the last few years, cancer immunotherapy has become a topic of great interest. Recent works have investigated to what extent splice site creating mutations or intron retention events are responsible for the origin of neo-epitopes. The method developed in this work allows for genome-wide inspection of all possible non-annotated splicing alterations, including those previously tested (aberrant splice site and intron retention) and new ones (exon skipping and exonization).

In addition, this approach takes also into account the possible epitopes that could be lost due to the splicing alteration. We observed that these alterations tended to delete more epitopes than create new ones, suggesting a mechanism of immune escape.

This is the first analysis that explores the immunogenic impact of splicing in SCLC. Since this cancer is one of the most mutated and no targeted alterations has been found yet, this analysis opens a new perspective for SCLC clinical management that should be explored further.

3.3.3. Limitations and future work

For the validation with the cell lines we used proteomics data generated with the same cell lines (Ritz et al, 2016). For the SCLC epitopes we did not have mass-spectrometry data available from the same samples. Because of this, we decided to use a set of peptides from Lanoix et al., 2018. A potential future development will be to obtain peptides binding MHC complexes and T-cell receptor repertoires from the same tumor samples studied. This would provide a better validation of the putative epitopes.

There are several questions for the SCLC samples that require further investigation at the light of the results. Although we observed many SCLC specific splicing alterations, very few were associated with cis-mutations. Future work would involve looking for trans-mutation on spliceosome components or splicing factors and disrupted motifs to explain the observed splicing alterations. Also important will be to investigate the functional impact that all these alterations cause at the level of protein domains or protein-protein interactions. We observed some degree of mutual exclusion between the expression of MYC genes and mutations in the spliceosome. Finally, prospective work would be to investigate if these behaviors occurs in other cancer types. SCLC tumors present very different etiological and molecular characteristics from the lung tumors. Understanding further the molecular similarities with pulmonary carcinoids, a possibly related condition, may help improving our understanding of SCLC.

4 CONCLUSIONS

1. We developed SUPPA2, a tool for fast calculation of differential splicing across multiple conditions taking into account biological variability.

2. We implemented a clustering method that detects groups of events with similar splicing patterns. We proved that this method obtains meaningful groups previously reported in the literature with the potential of finding new regulatory features.

3. Using machine learning techniques, we exploited splicing changes across patients for deriving signatures for cancer staging and prognosis. We showed that these signatures yielded good accuracies at discriminating between patients at early and late cancer stage, and especially at predicting metastasis and lymph node invasion.

4. We developed a pipeline for the identification of all type of tumor-specific splicing changes and the subsequent evaluation of their immunogenic impact.

5. We have applied this method to SCLC samples and described a widespread elimination of native epitopes, in contrast to the birth of neo-epitopes, in relation to splicing. This could suggest a mechanism for the cancer cell to evade immune responses.

5 REFERENCES

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M., del Pozo, A., Vázquez, J., … Tress, M. L. (2015). Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004325

Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J. M., … Tress, M. L. (2018). Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky587

Alamancos, G. P., Agirre, E., & Eyras, E. (2014). Methods to study splicing from high-throughput RNA sequencing data. *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-62703-980-2_26

Alamancos, G. P., Pagés, A., Trincado, J. L., Bellora, N., & Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA (New York, N.Y.), 21*(9), 1521–1531. https://doi.org/10.1261/rna.051557.115

Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N., & Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, *21*(9). https://doi.org/10.1261/rna.051557.115

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Anczukow, O., & Krainer, A. R. (2016). Splicing-factor alterations in cancers. *RNA*. https://doi.org/10.1261/rna.057919.116

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*. https://doi.org/10.1101/gr.133744.111

Andersen, R. S., Andersen, S. R., Hjortsø, M. D., Lyngaa, R., Idorn, M., Køllgård, T. M., … Hadrup, S. R. (2013). High frequency of T cells specific for cryptic epitopes in

melanoma patients. *Oncoimmunology*, *2*(7), e25374. https://doi.org/10.4161/onci.25374

Andreatta, M., & Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics (Oxford, England)*, *32*(4), 511–517. https://doi.org/10.1093/bioinformatics/btv639

Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *In ACM Sigmod Record*, *28*(2), 49–60.

Antonia, S. J., López-Martin, J. A., Bendell, J., Ott, P. A., Taylor, M., Eder, J. P., … Calvo, E. (2016). Nivolumab alone and nivolumab plus ipilimumab in recurrent small-cell lung cancer (CheckMate 032): a multicentre, open-label, phase 1/2 trial. *The Lancet. Oncology*, *17*(7), 883–895. https://doi.org/10.1016/S1470-2045(16)30098-5

Araki, S., Dairiki, R., Nakayama, Y., Murai, A., Miyashita, R., Iwatani, M., … Nakanishi, O. (2015). Inhibitors of CLK protein kinases suppress cell growth and induce apoptosis by modulating pre-mRNA splicing. *PLoS ONE*, *10*(1), 1–18. https://doi.org/10.1371/journal.pone.0116929

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., … Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, *8*(1), 1077. https://doi.org/10.1038/s41467-017-01027-z

Bai, Y., Ni, M., Cooper, B., Wei, Y., & Fury, W. (2014). Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*, *15*, 325. https://doi.org/10.1186/1471-2164-15-325

Bang, M. L., Centner, T., Fornoff, F., Geach, A. J., Gotthardt, M., McNabb, M., … Labeit, S. (2001). The complete gene sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circulation Research*. https://doi.org/10.1161/hh2301.100981

Bao, Q., Gong, L., Wang, J., Wen, J., Shen, Y., & Zhang, W. (2018). Extracellular Vesicle RNA Sequencing Reveals Dramatic Transcriptomic Alterations Between Metastatic and Primary Osteosarcoma in a Liquid Biopsy Approach. *Annals of Surgical Oncology*. https://doi.org/10.1245/s10434-018-6642-z

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm.2017.27

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., … Frey, B. J. (2010). Deciphering the splicing code. *Nature*. https://doi.org/10.1038/nature09000

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., …

Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607. https://doi.org/10.1038/nature11003

Bartosovic, M., Molares, H. C., Gregorova, P., Hrossova, D., Kudla, G., & Vanacova, S. (2017). N6-methyladenosine demethylase FTO targets pre-mRNAs and regulates alternative splicing and 3′-end processing. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx778

Bechara, E. G., Sebestyén, E., Bernardis, I., Eyras, E., & Valcárcel, J. (2013). RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2013.11.010

Bensimon, A., Heck, A. J. R., & Aebersold, R. (2012). Mass Spectrometry–Based Proteomics and Network Biology. *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev-biochem-072909-100424

Best, A., James, K., Dalgliesh, C., Hong, E., Kheirolahi-Kouhestani, M., Curk, T., … Elliott, D. J. (2014). Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. *Nature Communications*, *5*, 4760. https://doi.org/10.1038/ncomms5760

Beyer, A. L., & Osheim, Y. N. (1988). Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes & Development*, *2*(6), 754–765. https://doi.org/10.1101/gad.2.6.754

Black, K. L., Naqvi, A. S., Hayer, K. E., Yang, S. Y., Gillespie, E., Bagashev, A., … Thomas-Tikhonenko, A. (2017). Aberrant splicing in B-cell acute lymphoblastic leukemia. *BioRxiv*.

Blakeley, P., Siepen, J. A., Lawless, C., & Hubbard, S. J. (2010). Investigating protein isoforms via proteomics: A feasibility study. *Proteomics*. https://doi.org/10.1002/pmic.200900445

Blencowe, B. J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2017.04.001

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*. https://doi.org/10.1038/nbt.3519

Brown, J. W. S., Calixto, C. P. G., & Zhang, R. (2016). High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. *The New Phytologist*. https://doi.org/10.1111/nph.14208

Bunn, P. A., Minna, J. D., Augustyn, A., Gazdar, A. F., Ouadah, Y., Krasnow, M. A., …

Hirsch, F. R. (2016). Small Cell Lung Cancer: Can Recent Advances in Biology and Molecular Biology Be Translated into Improved Outcomes? *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer*, *11*(4), 453–474. https://doi.org/10.1016/j.jtho.2016.01.012

Busskamp, V., Lewis, N. E., Guye, P., Ng, A. H. M., Shipman, S. L., Byrne, S. M., … Church, G. M. (2014). Rapid neurogenesis through transcriptional activation in human stem cells. *Molecular Systems Biology*, *10*, 760.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V. F., Iglesias, M., Céspedes, M. V., … Batlle, E. (2012). Dependency of Colorectal Cancer on a TGF-β-Driven Program in Stromal Cells for Metastasis Initiation. *Cancer Cell*. https://doi.org/10.1016/j.ccr.2012.08.013

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, *511*(7511), 543–550. https://doi.org/10.1038/nature13385

Carazo, F., Romero, J. P., & Rubio, A. (2018). Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors. *Briefings in Bioinformatics*, (May), 1–18. https://doi.org/10.1093/bib/bby005

Carreno, B. M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A. A., … Linette, G. P. (2015). A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science*. https://doi.org/10.1126/science.aaa3828

Carrillo Oesterreich, F., Herzel, L., Straube, K., Hujer, K., Howard, J., & Neugebauer, K. M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell*. https://doi.org/10.1016/j.cell.2016.02.045

Castle, J. C., Kreiter, S., Diekmann, J., Löwer, M., Van De Roemer, N., De Graaf, J., … Sahin, U. (2012). Exploiting the mutanome for tumor vaccination. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-11-3722

Cheng, J., Zhou, T., Liu, C., Shapiro, J., Brauer, M., Kiefer, M., … Mountz, J. (1994). Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science*. https://doi.org/10.1126/science.7510905

Claborn, M. K., Stevens, D. L., Walker, C. K., & Gildon, B. L. (2018). Nusinersen: A Treatment for Spinal Muscular Atrophy. *Annals of Pharmacotherapy*. https://doi.org/10.1177/1060028018789956

Climente-González, H., Porta-Pardo, E., Godzik, A., & Eyras, E. (2017). The Functional Impact of Alternative Splicing in Cancer. *Cell Reports*, *20*(9), 2215–2226. https://doi.org/10.1016/j.celrep.2017.08.012

Cole, B. S., Tapescu, I., Allon, S. J., Mallory, M. J., Qiu, J., Lake, R. J., … Lynch, K. W. (2015). Global analysis of physical and functional RNA targets of hnRNP L reveals distinct sequence and epigenetic features of repressed and enhanced exons. *RNA (New York, N.Y.)*, *21*(12), 2053–2066. https://doi.org/10.1261/rna.052969.115

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*. https://doi.org/10.1186/s13059-016-0881-8

Cooper, S., & Spiro, S. G. (2006). Small cell lung cancer: treatment review. *Respirology (Carlton, Vic.)*, *11*(3), 241–248. https://doi.org/10.1111/j.1440-1843.2006.00850.x

Cooper, T. A., & Giudice, J. (2014). RNA-binding proteins in heart development. *Advances in Experimental Medicine and Biology*.

Darman, R. B., Seiler, M., Agrawal, A. A., Lim, K. H., Peng, S., Aird, D., … Buonamici, S. (2015). Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Reports*. https://doi.org/10.1016/j.celrep.2015.09.053

Das, S., & Krainer, A. R. (2014). Emerging Functions of SRSF1, Splicing Factor and Oncoprotein, in RNA Metabolism and Cancer. *Molecular Cancer Research*. https://doi.org/10.1158/1541-7786.MCR-14-0131

David, C. J., & Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged. *Genes and Development*. https://doi.org/10.1101/gad.1973010

de Jong, L. C., Cree, S., Lattimore, V., Wiggins, G. A. R., Spurdle, A. B., Miller, A., … Walker, L. C. (2017). Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. *Breast Cancer Research*. https://doi.org/10.1186/s13058-017-0919-1

Dergai, M., Tsyba, L., Dergai, O., Zlatskii, I., Skrypkina, I., Kovalenko, V., & Rynditch, A. (2010). Microexon-based regulation of ITSN1 and Src SH3 domains specificity relies on introduction of charged amino acids into the interaction interface. *Biochemical and Biophysical Research Communications*. https://doi.org/10.1016/j.bbrc.2010.07.080

Diederichs, S., Bartsch, L., Berkmann, J. C., Fröse, K., Heitmann, J., Hoppe, C., … Wullenkord, R. (2016). The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Molecular Medicine*, *8*(5), 442–457. https://doi.org/10.15252/emmm.201506055

Dillman, A. A., Hauser, D. N., Gibbs, J. R., Nalls, M. A., McCoy, M. K., Rudenko, I. N., … Cookson, M. R. (2013). MRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nature Neuroscience*. https://doi.org/10.1038/nn.3332

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts635

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., … Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2018.05.001

Duffy, A. G., Makarova-Rusher, O. V., Ulahannan, S. V., Rahma, O. E., Fioravanti, S., Walker, M., … Greten, T. F. (2016). Modulation of tumor eIF4E by antisense inhibition: A phase I/II translational clinical trial of ISIS 183750—an antisense oligonucleotide against eIF4E—in combination with irinotecan in solid tumors and irinotecan-refractory colorectal cancer. *International Journal of Cancer*, *139*(7), 1648–1657. https://doi.org/10.1002/ijc.30199

Dvinge, H., & Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine*, *7*(1), 45. https://doi.org/10.1186/s13073-015-0168-9

Dvinge, H., Kim, E., Abdel-Wahab, O., & Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc.2016.51

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). https://doi.org/10.1.1.71.1980

Ezkurdia, I., Vázquez, J., Valencia, A., & Tress, M. (2014). Analyzing the first drafts of the human proteome. *Journal of Proteome Research*. https://doi.org/10.1021/pr500572z

Fälth, M., Sköld, K., Norrman, M., Svensson, M., Fenyö, D., & Andren, P. E. (2006). SwePep, a Database Designed for Endogenous Peptides and Mass Spectrometry. *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.M500401-MCP200

Feinberg, A. P., Koldobskiy, M. A., & Göndör, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2016.13

Fernandez-Cuesta, L., Peifer, M., Lu, X., Sun, R., Ozretić, L., Seidel, D., … Thomas, R. K. (2014). Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids.

Nature Communications, *5*, 3518. https://doi.org/10.1038/ncomms4518

Fletcher, S., Meloni, P. L., Johnsen, R. D., Wong, B. L., Muntoni, F., & Wilton, S. D. (2013). Antisense suppression of donor splice site mutations in the dystrophin gene transcript. *Molecular Genetics & Genomic Medicine*. https://doi.org/10.1002/mgg3.19

Froussios, K., Mourao, K., Barton, G. J., & Schurch, N. J. (2017). Identifying differential isoform abundance with RATs: a universal tool and a warning. *BioRxiv*. https://doi.org/10.1101/132761

Fu, G., Condon, K. C., Epton, M. J., Gong, P., Jin, L., Condon, G. C., … Alphey, L. (2007). Female-specific insect lethality engineered using alternative splicing. *Nature Biotechnology*. https://doi.org/10.1038/nbt1283

Fu, X. D., & Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3778

Garalde, D. R., Snell, E. A., Jachimowicz, D., Heron, A. J., Bruce, M., Lloyd, J., … Turner, D. J. (2016). Highly parallel direct RNA sequencing on an array of nanopores. *BioRxiv*, 068809. https://doi.org/10.1101/068809

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., … Turner, D. J. (2018). Highly parallel direct RN A sequencing on an array of nanopores. *Nature Methods*, *15*(3), 201–206. https://doi.org/10.1038/nmeth.4577

Gaudreau, M.-C., Heyd, F., Bastien, R., Wilhelm, B., & Moroy, T. (2012). Alternative Splicing Controlled by Heterogeneous Nuclear Ribonucleoprotein L Regulates Development, Proliferation, and Migration of Thymic Pre-T Cells. *The Journal of Immunology*. https://doi.org/10.4049/jimmunol.1103142

George, J., Lim, J. S., Jang, S. J., Cun, Y., Ozretić, L., Kong, G., … Thomas, R. K. (2015). Comprehensive genomic profiles of small cell lung cancer. *Nature*, *524*(7563), 47–53. https://doi.org/10.1038/nature14664

Germann, S., Gratadou, L., Dutertre, M., & Auboeuf, D. (2012). Splicing programs and cancer. *Journal of Nucleic Acids*. https://doi.org/10.1155/2012/269570

Gerstein, M. B., Rozowsky, J., Yan, K. K., Wang, D., Cheng, C., Brown, J. B., … Waterston, R. (2014). Comparative analysis of the transcriptome across distant species. *Nature*. https://doi.org/10.1038/nature13424

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P. M., … Biamonti, G. (2005). Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Molecular Cell*, *20*(6), 881–890. https://doi.org/10.1016/j.molcel.2005.10.026

Gilbert, W. (1978). Why genes in pieces? *Nature*. https://doi.org/10.1038/271501a0

Giudice, G., Sánchez-Cabo, F., Torroja, C., & Lara-Pezzi, E. (2016). ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*. https://doi.org/10.1093/database/baw035

Giuliano, A. E., Connolly, J. L., Edge, S. B., & Mittendorf, E. A. (2017). Breast Cancer — Major Changes in the American Joint Committee on Cancer Eighth Edition Cancer Staging Manual. *A Cancer Journals for Clinicians*. https://doi.org/10.3322/caac.21393.

Gonzalez-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*(7), R70. https://doi.org/doi:10.1186/gb-2013-14-7-r70

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., … Ast, G. (2006). Comparative Analysis Identifies Exonic Splicing Regulatory Sequences-The Complex Definition of Enhancers and Silencers. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2006.05.008

Green, R. E., Lewis, B. P., Hillman, R. T., Blanchette, M., Lareau, L. F., Garnett, A. T., … Brenner, S. E. (2003). Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. In *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btg1015

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*. https://doi.org/10.1038/nprot.2013.084

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*. https://doi.org/10.1016/j.cell.2011.02.013

Hao, Y., Colak, R., Teyra, J., Corbi-Verge, C., Ignatchenko, A., Hahne, H., … Kim, P. M. (2015). Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins. *Cell Reports*. https://doi.org/10.1016/j.celrep.2015.06.031

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774.

https://doi.org/10.1101/gr.135350.111

Hellmann, M. D., Callahan, M. K., Awad, M. M., Calvo, E., Ascierto, P. A., Atmaca, A., … Antonia, S. J. (2018). Tumor Mutational Burden and Efficacy of Nivolumab Monotherapy and in Combination with Ipilimumab in Small-Cell Lung Cancer. *Cancer Cell*, *33*(5), 853–861.e4. https://doi.org/10.1016/j.ccell.2018.04.001

Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, *19*(5), 327–341. https://doi.org/10.1038/nrm.2017.130

Hsu, T. Y. T., Simon, L. M., Neill, N. J., Marcotte, R., Sayad, A., Bland, C. S., … Westbrook, T. F. (2015). The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature*, *525*(7569), 384–388. https://doi.org/10.1038/nature14985

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., … Liu, J. (2013). DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Research*, *41*(2). https://doi.org/10.1093/nar/gks1026

Hua, Y., Sahashi, K., Rigo, F., Hung, G., Horev, G., Bennett, C. F., & Krainer, A. R. (2011). Peripheral SMN restoration is essential for long-term rescue of a severe spinal muscular atrophy mouse model. *Nature*. https://doi.org/10.1038/nature10485

Huang, Y., & Sanguinetti, G. (2017). BRIE: Transcriptome-wide splicing quantification in single cells. *Genome Biology*. https://doi.org/10.1186/s13059-017-1248-5

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., … Lo, R. S. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*. https://doi.org/10.1016/j.cell.2016.02.065

Inoue, K., & Fry, E. A. (2016). Novel Molecular Markers for Breast Cancer. *Biomarkers in Cancer*, *8*, BIC.S38394. https://doi.org/10.4137/BIC.S38394

Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., … Blencowe, B. J. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*. https://doi.org/10.1016/j.cell.2014.11.035

Iwakawa, R., Kohno, T., Totoki, Y., Shibata, T., Tsuchihara, K., Mimaki, S., … Yokota, J. (2015). Expression and clinical significance of genes frequently mutated in small cell lung cancers defined by whole exome/RNA sequencing. *Carcinogenesis*, *36*(6), 616–621. https://doi.org/10.1093/carcin/bgv026

Jauvin, D., Chrétien, J., Pandey, S. K., Martineau, L., Revillod, L., Bassez, G., … Puymirat, J. (2017). Targeting DMPK with Antisense Oligonucleotide Improves Muscle Strength

in Myotonic Dystrophy Type 1 Mice. *Molecular Therapy - Nucleic Acids*. https://doi.org/10.1016/j.omtn.2017.05.007

Jayasinghe, R. G., Cao, S., Gao, Q., Wendl, M. C., Vo, N. S., Reynolds, S. M., … Ding, L. (2018). Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Reports*, *23*(1), 270–281.e3. https://doi.org/10.1016/j.celrep.2018.03.052

Jayasinghe, R. G., Cao, S., Gao, Q., Wendl, M. C., Vo, N. S., Reynolds, S. M., … Ding, L. (2018). Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Reports*. https://doi.org/10.1016/j.celrep.2018.03.052

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., … Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, *154*(3), 394–406. https://doi.org/10.1111/imm.12889

Johnson, N. T., Dhroso, A., Hughes, K. J., & Korkin, D. (2018). Biological classification with RNA-Seq data: Can alternatively spliced transcript expression enhance machine learning classifier? *RNA (New York, N.Y.)*. https://doi.org/10.1261/rna.062802.117

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y. J., Park, W. Y., … Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nature Genetics*. https://doi.org/10.1038/ng.3414

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., & Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of Immunology (Baltimore, Md. : 1950)*, *199*(9), 3360–3368. https://doi.org/10.4049/jimmunol.1700893

Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., … Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, *34*(2), 211–224.e6. https://doi.org/10.1016/j.ccell.2018.07.001

Kaida, D., Motoyoshi, H., Tashiro, E., Nojima, T., Hagiwara, M., Ishigami, K., … Yoshida, M. (2007). Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. *Nature Chemical Biology*. https://doi.org/10.1038/nchembio.2007.18

Kang, S., Thompson, Z., Mcclung, E. C., Abdallah, R., Lee, J. K., Gonzalez-Bosquet, J., … Chon, H. S. (2018). Gene expression signature-based prediction of lymph node metastasis in patients with endometrioid endometrial cancer. *International Journal of Gynecological Cancer*, *28*(2), 260–266. https://doi.org/10.1097/IGC.0000000000001152

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*.

https://doi.org/10.1038/nmeth.1528

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene*. https://doi.org/10.1016/j.gene.2012.07.083

Kiegle, E. A., Garden, A., Lacchini, E., & Kater, M. M. (2018). A Genomic View of Alternative Splicing of Long Non-coding RNAs during Rice Seed Development Reveals Extensive Splicing and lncRNA Gene Families. *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2018.00115

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, K. K., Nam, J., Mukouyama, Y.-S., & Kawamoto, S. (2013). Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development. *The Journal of Cell Biology*, *200*(4), 443–458. https://doi.org/10.1083/jcb.201206146

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., … Pandey, A. (2014). A draft map of the human proteome. *Nature*. https://doi.org/10.1038/nature13302

Koh, C. M., Bezzi, M., Low, D. H. P., Ang, W. X., Teo, S. X., Gay, F. P. H., … Guccione, E. (2015). MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature*. https://doi.org/10.1038/nature14351

Kotake, Y., Sagane, K., Owa, T., Mimori-Kiyosue, Y., Shimizu, H., Uesugi, M., … Mizui, Y. (2007). Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature Chemical Biology*. https://doi.org/10.1038/nchembio.2007.16

Kreiter, S., Vormehr, M., Van De Roemer, N., Diken, M., Löwer, M., Diekmann, J., … Sahin, U. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature*. https://doi.org/10.1038/nature14426

Kurokawa, K., Kuwano, Y., Tominaga, K., Kawai, T., Katsuura, S., Yamagishi, N., … Rokutan, K. (2010). Brief naturalistic stress induces an alternative splice variant of SMG-1 lacking exon 63 in peripheral leukocytes. *Neuroscience Letters*. https://doi.org/10.1016/j.neulet.2010.08.031

Lahat, A., & Grellscheid, S. N. (2016). Differential mRNA Alternative Splicing. In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing* (pp. 105–119). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-31350-4_5

Lanoix, J., Durette, C., Courcelles, M., Cossette, É., Comtois-Marotte, S., Hardy, M.-P., … Thibault, P. (2018a). Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. *Proteomics*, *18*(12), e1700251. https://doi.org/10.1002/pmic.201700251

Lanoix, J., Durette, C., Courcelles, M., Cossette, É., Comtois-Marotte, S., Hardy, M. P., … Thibault, P. (2018b). Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. *Proteomics*. https://doi.org/10.1002/pmic.201700251

Lara-Pezzi, E., Gómez-Salinero, J., Gatto, A., & García-Pavía, P. (2013). The alternative heart: Impact of alternative splicing in heart disease. *Journal of Cardiovascular Translational Research*, *6*(6), 945–955. https://doi.org/10.1007/s12265-013-9482-z

Lee, F. C. Y., & Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2018.01.005

Lee, S. C. W., & Abdel-Wahab, O. (2016). Therapeutic targeting of splicing in cancer. *Nature Medicine*. https://doi.org/10.1038/nm.4165

Lee, S. C. W., North, K., Kim, E., Jang, E., Obeng, E., Lu, S. X., … Abdel-Wahab, O. (2018). Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations. *Cancer Cell*. https://doi.org/10.1016/j.ccell.2018.07.003

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, *84*(1), 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316

Lev Maor, G., Yearim, A., & Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2015.03.002

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Li, J., Wang, W., Xia, P., Wan, L., Zhang, L., Yu, L., … Xu, C. (2018). Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in patients with breast cancer. *International Journal of Cancer*, 1–21. https://doi.org/10.1002/ijc.31573

Li, S., Guo, W., Dewey, C. N., & Greaser, M. L. (2013). Rbm20 regulates titin alternative splicing as a splicing repressor. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1362

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using

LeafCutter. *Nature Genetics*, *50*(1), 151–158. https://doi.org/10.1038/s41588-017-0004-9

Li, Y. I., Sanchez-Pulido, L., Haerty, W., & Ponting, C. P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Research*, *25*(1), 1–13. https://doi.org/10.1101/gr.181990.114

Li, Y., McGrail, D. J., Xu, J., Mills, G. B., Sahni, N., & Yi, S. (2018). Gene Regulatory Network Perturbation by Genetic and Epigenetic Variation. *Trends in Biochemical Sciences*, *xx*, 1–17. https://doi.org/10.1016/j.tibs.2018.05.002

Liang, S. B., & Fu, L. W. (2017). Application of single-cell technology in cancer research. *Biotechnology Advances*, *35*(4), 443–449. https://doi.org/10.1016/j.biotechadv.2017.04.001

Lin, C. F., Mount, S. M., Jarmoowski, A., & Makaowski, W. (2010). Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evolutionary Biology*, *10*(1). https://doi.org/10.1186/1471-2148-10-47

Lin, E., Li, L., Guan, Y., Soriano, R., Rivers, C. S., Mohan, S., … Modrusan, Z. (2009). Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers. *Molecular Cancer Research : MCR*, *7*(9), 1466–1476. https://doi.org/10.1158/1541-7786.MCR-08-0522

Liu, J., Li, H., Sun, L., Wang, Z., Xing, C., & Yuan, Y. (2017). Aberrantly methylated-differentially expressed genes and pathways in colorectal cancer. *Cancer Cell International*. https://doi.org/10.1186/s12935-017-0444-4

Lorson, C. L., & Androphy, E. J. (2000). An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/9.2.259

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., … Yeo, G. W. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2699

Lowery, L. A., Rubin, J., & Sive, H. (2007). Whitesnake/sfpq is required for cell survival and neuronal development in the zebrafish. *Developmental Dynamics : An Official Publication of the American Association of Anatomists*, *236*(5), 1347–1357. https://doi.org/10.1002/dvdy.21132

Lu, Y.-C., Parker, L. L., Lu, T., Zheng, Z., Toomey, M. A., White, D. E., … Rosenberg, S. A. (2017). Treatment of Patients With Metastatic Cancer Using a Major Histocompatibility

Complex Class II-Restricted T-Cell Receptor Targeting the Cancer Germline Antigen MAGE-A3. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *35*(29), 3322–3329. https://doi.org/10.1200/JCO.2017.74.5463

Luo, Y. B., Mastaglia, F. L., & Wilton, S. D. (2014). Normal and aberrant splicing of LMNA. *Journal of Medical Genetics*, *51*(4), 215–223. https://doi.org/10.1136/jmedgenet-2013-102119

Lynch, K. W., & Maniatis, T. (1996). Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer. *Genes and Development*. https://doi.org/10.1101/gad.10.16.2089

Makowski, K., Vigevani, L., Albericio, F., Valcarcel, J., & Alvarez, M. (2017). Sudemycin k: A synthetic antitumor splicing inhibitor variant with improved activity and versatile chemistry. *ACS Chemical Biology*. https://doi.org/10.1021/acschembio.6b00562

Maquat, L. E., Kinniburgh, A. J., Beach, L. R., Honig, G. R., Lazerson, J., Ershler, W. B., & Ross, J. (1980). Processing of human beta-globin mRNA precursor to mRNA is defective in three patients with beta+-thalassemia. *Proceedings of the National Academy of Sciences*, *77*(7), 4287–4291. https://doi.org/10.1073/pnas.77.7.4287

Martinocorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*. https://doi.org/10.1126/science.aab4082

Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3742

Miao, D., Margolis, C. A., Vokes, N. I., Liu, D., Taylor-Weiner, A., Wankowicz, S. M., … Van Allen, E. M. (2018). Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nature Genetics*. https://doi.org/10.1038/s41588-018-0200-2

Monlong, J., Calvo, M., Ferreira, P. G., & Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, *5*(May). https://doi.org/10.1038/ncomms5698

Moreno, P. M. D., & PÃªgo, A. P. (2014). Therapeutic antisense oligonucleotides against cancer: hurdling to the clinic. *Frontiers in Chemistry*. https://doi.org/10.3389/fchem.2014.00087

Morrissy, A. S., Griffith, M., & Marra, M. A. (2011). Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Research*. https://doi.org/10.1101/gr.113431.110

Nair, M., Sandhu, S. S., & Sharma, A. K. (2018). Cancer molecular markers: A guide to

cancer detection and management. *Seminars in Cancer Biology*, (February), 1–17. https://doi.org/10.1016/j.semcancer.2018.02.002

Nellore, A., Jaffe, A. E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., … Leek, J. T. (2016). Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology*, *17*(1), 266. https://doi.org/10.1186/s13059-016-1118-6

Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., … Zhu, J. (2016). Global intron retention mediated gene regulation during CD4+ T cell activation. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw591

Norris, A. D., Gao, S., Norris, M. L., Ray, D., Ramani, A. K., Fraser, A. G., … Calarco, J. A. (2014). A pair of RNA-binding proteins controls networks of splicing events contributing to specialization of neural cell types. *Molecular Cell*, *54*(6), 946–959. https://doi.org/10.1016/j.molcel.2014.05.004

Nowicka, M., & Robinson, M. D. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*. https://doi.org/10.12688/f1000research.8900.2

Oltean, S., & Bates, D. O. (2014). Hallmarks of alternative splicing in cancer. *Oncogene*. https://doi.org/10.1038/onc.2013.533

Oustric, V., Manceau, H., Ducamp, S., Soaid, R., Karim, Z., Schmitt, C., … Gouya, L. (2014). Antisense oligonucleotide-based therapy in human erythropoietic protoporphyria. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2014.02.010

Padgett, R. A. (2012). New connections between splicing and human disease. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2012.01.001

Paik, P. K., Drilon, A., Fan, P. D., Yu, H., Rekhtman, N., Ginsberg, M. S., … Ladanyi, M. (2015). Response to MET inhibitors in patients with stage IV lung adenocarcinomas harboring met mutations causing exon 14 skipping. *Cancer Discovery*. https://doi.org/10.1158/2159-8290.CD-14-1467

Park, E., Pan, Z., Zhang, Z., Lin, L., & Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2017.11.002

Partin, J. F., & Mamounas, E. P. (2011). Impact of the 21-gene recurrence score assay compared with standard clinicopathologic guidelines in adjuvant therapy selection for node-negative, estrogen receptor-positive breast cancer. *Annals of Surgical Oncology*.

https://doi.org/10.1245/s10434-011-1698-z

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. https://doi.org/10.1038/nmeth.4197

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*(5), 462–464. https://doi.org/10.1038/nbt.2862

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Peifer, M., Fernández-Cuesta, L., Sos, M. L., George, J., Seidel, D., Kasper, L. H., … Thomas, R. K. (2012). Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nature Genetics*, *44*(10), 1104–1110. https://doi.org/10.1038/ng.2396

Pendás, A. M., Zhou, Z., Cadiñanos, J., Freije, J. M. P., Wang, J., Hultenby, K., … López-Otín, C. (2002). Defective prelamin A processing and muscular and adipocyte alterations in Zmpste24 metalloproteinase-deficient mice. *Nature Genetics*. https://doi.org/10.1038/ng871

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*. https://doi.org/10.1038/nbt.3122

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A. K., … Salzberg, S. (2018). Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *BioRxiv*.

Pimentel, H., Conboy, J. G., & Pachter, L. (2015). Keep Me Around: Intron Retention Detection and Analysis.

Pimentel, H., Parra, M., Gee, S., Ghanem, D., An, X., Li, J., … Conboy, J. G. (2014). A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research*, *42*(6), 4031–4042. https://doi.org/10.1093/nar/gkt1388

Pimentel, H., Parra, M., Gee, S. L., Mohandas, N., Pachter, L., & Conboy, J. G. (2016). A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research*, *44*(2), 838–851. https://doi.org/10.1093/nar/gkv1168

Poulikakos, P. I., Persaud, Y., Janakiraman, M., Kong, X., Ng, C., Moriceau, G., … Solit, D. B. (2011). RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature*, *480*(7377), 387–390. https://doi.org/10.1038/nature10662

Quesnel-Vallières, M., Dargaei, Z., Irimia, M., Gonatopoulos-Pournatzis, T., Ip, J. Y., Wu, M., … Cordes, S. P. (2016). Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2016.11.033

Raj, B., Irimia, M., Braunschweig, U., Sterne-Weiler, T., O'Hanlon, D., Lin, Z. Y., … Blencowe, B. J. (2014). A global regulatory mechanism for activating an exon network required for neurogenesis. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2014.08.011

Ranum, L. P. W., & Cooper, T. A. (2006). RNA-MEDIATED NEUROMUSCULAR DISORDERS. *Annual Review of Neuroscience*. https://doi.org/10.1146/annurev.neuro.29.051605.113014

Ray, D., Ha, K. C. H., Nie, K., Zheng, H., Hughes, T. R., & Morris, Q. D. (2017). RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. *Methods*. https://doi.org/10.1016/j.ymeth.2016.12.003

Ray, D., Kazan, H., Chan, E. T., Castillo, L. P., Chaudhry, S., Talukder, S., … Hughes, T. R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*. https://doi.org/10.1038/nbt.1550

Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., & Fugmann, T. (2016). High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics*, *16*(10), 1570–1580. https://doi.org/10.1002/pmic.201500445

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G., … Tress, M. L. (2013). APPRIS: Annotation of principal and alternative splice isoforms. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1058

Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F., & Crespi, M. (2018). Splicing regulation by long noncoding RNAs. *Nucleic Acids Research*, (February), 1–16. https://doi.org/10.1093/nar/gky095

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rudin, C. M., Durinck, S., Stawiski, E. W., Poirier, J. T., Modrusan, Z., Shames, D. S., …

Seshagiri, S. (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nature Genetics*, *44*(10), 1111–1116. https://doi.org/10.1038/ng.2405

Russell, P. H., Johnson, R. L., Ananthan, S., Harnke, B., & Carlson, N. E. (2018). A large-scale analysis of bioinformatics code on GitHub. *BioRxiv*, 321919. https://doi.org/10.1101/321919

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., … Battle. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research*, *27*(11), 1843–1858. https://doi.org/10.1101/gr.216721.116

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B. P., Simon, P., Löwer, M., … Türeci, Ö. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*. https://doi.org/10.1038/nature23003

Sahin, U., & Türeci, Ö. (2018). Personalized vaccines for cancer immunotherapy. *Science (New York, N.Y.)*, *359*(6382), 1355–1360. https://doi.org/10.1126/science.aar7112

Saint-André, V., Batsché, E., Rachez, C., & Muchardt, C. (2011). Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.1995

SAKAI, T., ASAI, N., OKUDA, A., KAWAMURA, N., & MIZUI, Y. (2004). Pladienolides, New Substances from Culture of Streptomyces platensis Mer-11107 II. Physico-chemical Properties and Structure Elucidation. *The Journal of Antibiotics*. https://doi.org/10.7164/antibiotics.57.180

Schmid, R., Grellscheid, S. N., Ehrmann, I., Dalgliesh, C., Danilenko, M., Paronetto, M. P., … Elliott, D. J. (2013). The splicing landscape is globally reprogrammed during male meiosis. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt811

Schor, I. E., Gómez Acuña, L. I., & Kornblihtt, A. R. (2013). Coupling between transcription and alternative splicing. *Cancer Treatment and Research*. https://doi.org/10.1007/978-3-642-31659-3_1

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews Genetics*, *17*(1), 19–32. https://doi.org/10.1038/nrg.2015.3

Sebestyén, E., Singh, B., Miï¿½ana, B., Pagï¿½s, A., Mateo, F., Pujana, M. A., … Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Research*, *26*(6), 732–744. https://doi.org/10.1101/gr.199935.115

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M. A., … Eyras, E.

(2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Research*, *26*(6), 732–744. https://doi.org/10.1101/gr.199935.115

Sebestyen, E., Zawisza, M., & Eyras, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Research*, *43*(3), 1345–1356. https://doi.org/10.1093/nar/gku1392

Sebestyén, E., Zawisza, M., & Eyras, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1392

Seiler, M., Yoshimi, A., Darman, R., Chan, B., Keaney, G., Thomas, M., … Buonamici, S. (2018). H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nature Medicine*. https://doi.org/10.1038/nm.4493

Sharp, P. A. (1994). Split Genes and RNA Splicing (Nobel Lecture). *Angewandte Chemie International Edition in English*. https://doi.org/10.1002/anie.199412291

Shen, S., Park, J. W., Lu, Z., Lin, L., Henry, M. D., Wu, Y. N., … Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1419161111

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., … Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*. https://doi.org/10.1038/nature10442

Sickmier, E. A., Frato, K. E., Shen, H., Paranawithana, S. R., Green, M. R., & Kielkopf, C. L. (2006). Structural Basis for Polypyrimidine Tract Recognition by the Essential Pre-mRNA Splicing Factor U2AF65. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2006.05.025

Singh, B., Trincado, J. L., Tatlow, P. J., Piccolo, S. R., & Eyras, E. (2018). Genome sequencing and RNA-motif analysis reveal novel damaging noncoding mutations in human tumors. *Molecular Cancer Research, 16*(7). https://doi.org/10.1158/1541-7786.MCR-17-0601

Singh, R. K., Xia, Z., Bland, C. S., Kalsotra, A., Scavuzzo, M. A., Curk, T., … Cooper, T. A. (2014). Rbfox2-coordinated alternative splicing of Mef2d and Rock2 controls myoblast fusion during myogenesis. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2014.06.035

Slaugenhaupt, S. A., Blumenfeld, A., Gill, S. P., Leyne, M., Mull, J., Cuajungco, M. P., …

Gusella, J. F. (2001). Tissue-Specific Expression of a Splicing Mutation in the IKBKAP Gene Causes Familial Dysautonomia. *The American Journal of Human Genetics*. https://doi.org/10.1086/318810

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., … Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology*. https://doi.org/10.1038/nchembio.1120

Smart, A. C., Margolis, C. A., Pimentel, H., He, M. X., Miao, D., Adeegbe, D., … Van Allen, E. M. (2018). Intron retention is a source of neoepitopes in cancer. *Nature Biotechnology*, (October 2017). https://doi.org/10.1038/nbt.4239

Solomon, O., Oren, S., Safran, M., Deshet-Unger, N., Akiva, P., Jacob-Hirsch, J., … Eyal, E. (2013). Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *RNA*. https://doi.org/10.1261/rna.038042.112

Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., & Yeo, G. W. (2017). Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2017.06.003

Sotillo, E., Barrett, D. M., Black, K. L., Bagashev, A., Oldridge, D., Wu, G., … Thomas-Tikhonenko, A. (2015). Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. *Cancer Discovery*. https://doi.org/10.1158/2159-8290.CD-15-1020

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., … Kaessmann, H. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*. https://doi.org/10.1016/j.celrep.2013.05.031

Srivastava, A., Smith, T. S., Sudbery, I., & Patro, R. (2018). Alevin : An integrated method for dscRNA-seq quantification, 1–20.

Stadler, M. B., Shomron, N., Yeo, G. W., Schneider, A., Xiao, X., & Burge, C. B. (2006). Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.0020191

Stratton, M., Campbell, P., & Futreal, P. (2009). The cancer genome. *Nature*, *458*(7239), 719–724. https://doi.org/10.1038/nature07943.The

Sun, S., Zhang, Z., Sinha, R., Karni, R., & Krainer, A. R. (2010). SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nature Structural and Molecular Biology*, *17*(3), 306–312. https://doi.org/10.1038/nsmb.1750

Sun, Z., Chen, F., Meng, F., Wei, J., & Liu, B. (2017). MHC class II restricted neoantigen: A promising target in tumor immunotherapy. *Cancer Letters*, *392*, 17–25. https://doi.org/10.1016/j.canlet.2016.12.039

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., & Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*. https://doi.org/10.1016/j.cell.2014.01.051

Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., & Skotheim, R. I. (2016). Aberrant RNA splicing in cancer; Expression changes and driver mutations of splicing factor genes. *Oncogene*, *35*(19), 2413–2427. https://doi.org/10.1038/onc.2015.318

Taherian-Fard, A., Srihari, S., & Ragan, M. A. (2014). Breast cancer classification: Linking molecular mechanisms to disease prognosis. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbu020

Tanackovic, G., Ransijn, A., Thibault, P., Elela, S. A., Klinck, R., Berson, E. L., … Rivolta, C. (2011). PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddr094

Tardaguila, M., Fuente, L. de la, Marti, C., Pereira, C., Pardo-Palacios, F. J., Risco, H. del, … Conesa, A. (2017). SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *BioRxiv*, 118083. https://doi.org/10.1101/118083

The Cancer Genome Atlas Network. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, *489*(7417), 519–525. https://doi.org/10.1038/nature11404

The UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku989

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2450

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. https://doi.org/10.1038/nbt.1621

Tress, M. L., Abascal, F., & Valencia, A. (2017a). Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*.

https://doi.org/10.1016/j.tibs.2016.08.008

Tress, M. L., Abascal, F., & Valencia, A. (2017b). Most Alternative Isoforms Are Not Functionally Important. *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2017.04.002

Tress, M. L., Bodenmiller, B., Aebersold, R., & Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology*. https://doi.org/10.1186/gb-2008-9-11-r162

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, *19*(1), 40. https://doi.org/10.1186/s13059-018-1417-1

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., … Prasanth, K. V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2010.08.011

Turunen, J. J., Niemelä, E. H., Verma, B., & Frilander, M. J. (2013). The significant other: Splicing by the minor spliceosome. *Wiley Interdisciplinary Reviews: RNA*. https://doi.org/10.1002/wrna.1141

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., … Garraway, L. A. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. https://doi.org/10.1126/science.aad0095

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. a M., Voskuil, D. W., … Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa021967

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., … Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *ELife*. https://doi.org/10.7554/eLife.11752

Vaz-Drago, R., Custódio, N., & Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human Genetics*. https://doi.org/10.1007/s00439-017-1809-4

Venables, J. P., Brosseau, J.-P., Gadea, G., Klinck, R., Prinos, P., Beaulieu, J.-F., … Chabot, B. (2013). RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Molecular and Cellular Biology*, *33*(2), 396–405. https://doi.org/10.1128/MCB.01174-12

Venables, J. P., Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Koh, C., … Elela, S. A. (2008). Identification of alternative splicing markers for breast cancer. *Cancer Research*, *68*(22), 9525–9531. https://doi.org/10.1158/0008-5472.CAN-08-1769

Vinay, D. S., Ryan, E. P., Pawelec, G., Talib, W. H., Stagg, J., Elkord, E., … Kwon, B. S. (2015). Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Seminars in Cancer Biology*. https://doi.org/10.1016/j.semcancer.2015.03.004

Volinia, S., Calin, G. A., Liu, C.-G., Ambs, S., Cimmino, A., Petrocca, F., … Croce, C. M. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(7), 2257–2261. https://doi.org/10.1073/pnas.0510565103

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., … Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*. https://doi.org/10.1038/nature07509

Wang, Z. (2010). The Concept of Multiple-Target Anti-miRNA Antisense Oligonucleotide Technology. *Journal of Biological Chemistry*, *261*(26), 12282–12289. https://doi.org/10.1007/978-1-60761-863-8

Weatheritt, R. J., Sterne-Weiler, T., & Blencowe, B. J. (2016). The ribosome-engaged landscape of alternative splicing. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.3317

Wein, N., Vulin, A., Falzarano, M. S., Szigyarto, C. A. K., Maiti, B., Findlay, A., … Flanigan, K. M. (2014). Translation from a DMD exon 5 IRES results in a functional dystrophin isoform that attenuates dystrophinopathy in humans and mice. *Nature Medicine*. https://doi.org/10.1038/nm.3628

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., … Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*. https://doi.org/10.1038/nature13319

Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*. https://doi.org/10.1101/cshperspect.a003707

Xue, W., Li, J., Wang, F., Han, P., Liu, Y., & Cui, B. (2017). A long non-coding RNA expression signature to predict survival of patients with colon adenocarcinoma. *Oncotarget*, *8*(60), 101298–101308. https://doi.org/10.18632/oncotarget.21064

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., … Vidal, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*. https://doi.org/10.1016/j.cell.2016.01.029

Yates, L. R., & Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3317

Yeo, G., Holste, D., Kreiman, G., & Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome Biology*. https://doi.org/10.1186/gb-2004-5-10-r74

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., … Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, *4*, 2612. https://doi.org/10.1038/ncomms3612

Zhang, J.-X., Song, W., Chen, Z.-H., Wei, J.-H., Liao, Y.-J., Lei, J., … Luo, J.-H. (2013). Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *The Lancet Oncology*. https://doi.org/10.1016/S1470-2045(13)70491-1

Zhang, R., Calixto, C. P. G., Marquez, Y., Venhuizen, P., Tzioutziou, N. A., Guo, W., … Brown, J. W. S. (2017). A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Research*, *45*(9), 5061–5073. https://doi.org/10.1093/nar/gkx267

Zhang, R., Calixto, C. P. G., Tzioutziou, N. A., James, A. B., Simpson, C. G., Guo, W., … Brown, J. W. S. (2015). AtRTD - a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana. *New Phytologist*, *208*(1). https://doi.org/10.1111/nph.13545

Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E., & Hogenesch, J. B. (2014). A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(45), 16219–16224. https://doi.org/10.1073/pnas.1408886111

Zhang, X., Chen, M. H., Wu, X., Kodani, A., Fan, J., Doan, R., … Walsh, C. A. (2016). Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. *Cell*. https://doi.org/10.1016/j.cell.2016.07.025

Zhang, Z. F., Pal, S., Bi, Y., Tchou, J., & Davuluri, R. V. (2013). Isoform level expression profiles provide better cancer signatures than gene level expression profiles. *Genome Medicine*. https://doi.org/10.1186/gm437

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., … Zhao, Q. Y. (2014). A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0103207

Zhao, K., Lu, Z., Park, J., Zhou, Q., & Xing, Y. (2013). GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biology*, *14*(7), R74. https://doi.org/10.1186/gb-2013-14-7-r74

Zheng, S. (2016). Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression. *International Journal of Developmental Neuroscience*. https://doi.org/10.1016/j.ijdevneu.2016.03.003

Zong, F. Y., Fu, X., Wei, W. J., Luo, Y. G., Heiner, M., Cao, L. J., … Hui, J. (2014). The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1004289