OXFORD

Briefings in Bioinformatics, 20(4), 2019, 1358–1375

doi: 10.1093/bib/bby005
Advance Access Publication Date: 29 January 2018
Paper

Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors

Fernando Carazo, Juan P. Romero and Angel Rubio

Corresponding author. Angel Rubio, Group of Bioinformatics, TECNUN, University of Navarra, Paseo Manuel Lardizábal 15, 20018 San Sebastián, Spain. Tel.: +34 943 21 98 77; E-mail: arubio@tecnun.es

Abstract

Alternative splicing (AS) has shown to play a pivotal role in the development of diseases, including cancer. Specifically, all the hallmarks of cancer (angiogenesis, cell immortality, avoiding immune system response, etc.) are found to have a counterpart in aberrant splicing of key genes. Identifying the context-specific regulators of splicing provides valuable information to find new biomarkers, as well as to define alternative therapeutic strategies. The computational models to identify these regulators are not trivial and require three conceptual steps: the detection of AS events, the identification of splicing factors that potentially regulate these events and the contextualization of these pieces of information for a specific experiment. In this work, we review the different algorithmic methodologies developed for each of these tasks. Main weaknesses and strengths of the different steps of the pipeline are discussed. Finally, a case study is detailed to help the reader be aware of the potential and limitations of this computational approach.

Key words: splicing factors; alternative splicing; RNA-binding proteins; bioinformatics

Introduction

Alternative splicing (AS) is the mechanism by which a single pre-mRNA molecule can lead to different mature mRNA molecules, called isoforms or transcripts. In this process, exons can be either included or excluded, shortened or lengthened and skipped or retained. The transcriptome is the complete set of mRNA isoforms in an organism. The phenomenon of AS was first described by Berget *et al.* [1], where it was shown that one adenovirus produced several transcripts during its infectious cycle.

The number of discovered isoforms increases as the study of an organism improves. In humans, around 95% of multi-exonic genes present AS events in diverse conditions [2, 3]. The paradigm 'one gene-one protein' has switched to the present situation in which most genes encode several proteins because of AS [4].

The functions altered by AS can be different: biomass generation, induction of angiogenesis, loss of genomic stability or deterioration of the immune system among others [5]. Besides, the influence of AS on neoplasms and other diseases is well known [6, 7]. In fact, studies suggest that approximately one-third of all disease-causing mutations modify splicing [8]. The regulation of AS has become a therapeutic strategy, and it is also revealing new therapeutic targets [6]. It has also been

Fernando Carazo is a PhD student at the University of Navarra (Spain). He holds an MSc degree in Industrial Engineering. He is interested in deciphering the splicing rules in different diseases and in analyzing high-throughput data in translational medicine.
Juan P. Romero received his PhD in Bioinformatics at Tecnun, University of Navarra (Spain) in 2017 and now works at CIMA in the Onco-Hematology Department. His research interests include the development of algorithms to identify alternative splicing events and machine learning techniques to study the transcriptome.
Angel Rubio is an Industrial Engineer and holds a PhD by the University of Navarra (1999). His research has focused on alternative splicing analysis, analysis of copy number alterations and integration of disparate biological data sources (miRNA and mRNA expression, number of copies and expression, etc.). He has been the Director of the Biomedical Engineering Department of the University of Navarra. At present, he is a Professor of Biostatistics, Next Generation Sequencing and Genomics and Proteomics at the University of Navarra.
Submitted: 24 October 2017; **Received (in revised form):** 14 December 2017

shown that all the cancer paradigms [9] have their counterpart in aberrant splicing [5].

The mechanism of splicing involves a complex biological machinery with several elements, such as RNA-binding proteins (RBPs), other *trans*-acting factors or specific sequences in the mRNA, which are the target signals of RBPs. Even epigenetics has been shown to play a role in AS [10–12]. The detection of AS itself is not trivial and requires specific algorithms and software to identify and label AS events.

Multiple approaches have been developed to understand the link between AS events and splicing regulatory elements (SREs) in different diseases. Several works analyzed brain-specific splicing factors (SFs) such as *NOVA1* and *NOVA2* [13–15]. Besides, one of the first global analysis for the identification of cancer-associated AS events and regulators was performed by Danan-Gotthold *et al.* in 2015 [16]. As these influential works, numerous strategies have emerged to decipher the splicing mechanism associated with diverse pathologies.

The scope of this review is the description of the computational approaches to detect splicing events and to predict their regulatory elements, i.e. upstream analysis of AS. We will not discuss the functional effects of AS but only its detection and potential regulation. Finally, a case study is detailed to help the reader be aware of the potential and limitations of these computational approaches.

Overview of splicing

Splicing is a posttranscriptional process in which nucleotide sequences, called introns, are removed from the pre-mRNA. The resulting product is a mature mRNA molecule that includes 5′ and 3′ untranslated regions (5′/3′ UTRs) and coding regions (exons) joined together to form a single mRNA strand. The splicing process is orchestrated by the spliceosome, a complex machinery made up of different subunits known as small nuclear ribonucleoproteins (snRNPs) and other protein complexes. snRNPs are noncoding and non-polyadenylated RNA–protein complexes that carry out their functions in the nucleoplasm [17]. A deeper explanation of the splicing mechanism has been included in the Supplementary Material (Section 1: splicing mechanism).

Splicing events in eukaryote cells can be classified into two main groups: constitutive splicing events, which always occur and give rise to the same isoforms independently of the tissue or pathological situation; and AS events, which lead to different isoforms. AS events have been divided into several canonical classes, as shown in Supplementary Figure S1.

Regulatory elements of AS

The mechanisms that control and regulate AS are still subject to active research [3, 11, 18–20]. Here, we focus on two key elements in the regulation of AS: *cis*-acting RNA elements and *trans*-acting factors. A scheme of the elements that take part in the AS process is shown in Figure 1A.

Cis-acting RNA elements

Cis-elements are RNA sequences (or motifs) in the pre-mRNA that allow the recognition of specific exonic/intronic regions by the spliceosome. Mainly, they comprise the canonical splicing signals and the SREs.

Splicing signals are essential sequences (the 5′ splice site, the 3′ splice site and the adenine branch point) for recognition by the spliceosome. SREs are divided into exonic splicing enhancers, exonic splicing silencers (ESSs), intronic splicing enhancers and intronic splicing silencers. The activity of SREs depends on the recruitment of molecules, which impact positive or negatively in the splicing reaction steps [21]. One common example is the polypyrimidine tract-binding protein (PTB), which causes exon skipping after binding to an ESS by avoiding the formation of the exon definition complex [22]. The effect of *cis*-acting RNA elements can be altered because of factors such as decoy splice-sites [17] or the surrounding context [19].

Trans-acting splicing regulators

Trans-acting splicing regulators are analytes—usually proteins—that, by interacting with the mRNA, modulate its AS. Most of these *trans*-acting factors are RBPs.

RBPs are proteins that bind to single- or double-stranded RNA and play key roles in posttranscriptional gene regulation, such as regulation of AS, mRNA stabilization, mRNA location, polyadenylation or translation [23]. They usually have modular designs and consist of various repeats of just a few basic RNA-binding domains, which have, in turn, different strategies to RNA binding. The ability to selectively recognize and bind mRNAs is crucial for the correct functionality of RBPs [3, 11, 18–20]. Most frequent RNA-binding domains, namely, RNA-recognition motif, heterogeneous nuclear ribonucleoprotein K-homology, double-stranded RNA binding domain and Zinc fingers—are described in the Section 1 of the Supplementary Material. We refer the reader to [19, 20, 24–31] for further details.

These domains are the main players of RBPs-mRNA interactions. In turn, these interactions have been extensively studied and compiled in several databases [32–36], but many of them are still unknown. The ATtRACT database [37], which contains curated and validated data from the main databases of RBPs-mRNA interactions (CisBP-RNA [32], SpliceAid-F [38] and RBPDB [39]), collects the binding information of 370 RBPs, which represent about 30% of the ∼1300 RBPs that have currently been discovered [31, 40].

RBPs that participate in the AS regulation are called SFs. These RBPs mainly include serine- or arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). The binding profile and functions of a number of SFs have been previously studied [41–44] and reviewed [19]. In general, SR proteins—such as *SRSF1* or *SRSF2*—are considered positive splicing regulators, as they promote exon inclusion [45–47]. In contrast, hnRNPs—such as *hnRNP A1*, *hnRNP A2* or *hnRNP B1*—seem to have the opposite effect, as they avoid the formation of the splicing machinery.

In addition to the spliceosome, there are other processes that play an important role in the regulation of the AS, such as DNA methylation [12], chromatin status [10], histone modifications [11], phosphorylation of the corresponding RBPs [48] or the secondary structure of the pre-mRNA [49]. In many cases, changes in either of these processes impact on AS.

In the next sections, we discuss the computational approaches to identify AS events and predict their context-dependent regulators. Although some individual parts of these tasks have already been covered by other reviews (Table 1), this work tries to provide a broader view of algorithms developed to unveil the complex regulation of AS.

Computational approaches to identify splicing and its regulatory elements

The algorithms will be presented in a conceptual sequential order: first, we discuss the algorithms to detect AS from

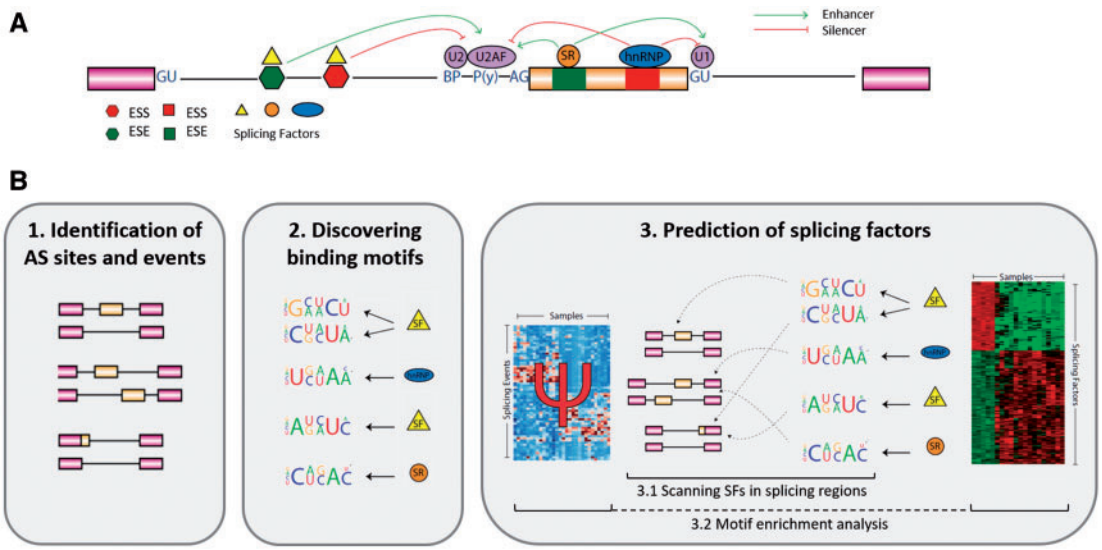


Figure 1. (**A**) Overview of AS process. An example of a cassette exon with its regulating elements is shown. The branch point (BP), and the polypyrimidine tract (PY) are also represented. (**B**) General pipeline to detect SFs: (1) identification of AS events in a specific condition; (2) identification of binding motifs of RBPs; and (3) prediction of splicing regulatory factors. This task is, in turn, divided into: (3.1) scanning SF's motifs in splicing regions and (3.2) motif enrichment analysis using PSI, expression levels of RBPs and other sources of information (driver mutations, CNVs, coexpression networks, etc.). Each of the boxes corresponds to a section in the main text.

Table 1. Description of previous reviews of experimental and computational methods related to the upstream analysis of the splicing process

Review	Description	Algorithms reviewed	Reference
A survey of software for genome-wide discovery of differential splicing in RNA-seq data	A review of the software available for analysis of RNA-seq data for differential splicing	**Identification of AS sites and events** Cuffdiff 2, MISO, DEXSeq, DSGseq, MATS, DiffSplice, Splicing compass, AltAnalyze	Hooper [50]
Advances in the characterization of RNA-binding proteins	Experimental and computational methods for detection of protein–RNA interactions	**Experimental methods for detection of protein–RNA interactions** (Protein-centric) RIP, HiTS-CLIP, PAR-CLIP, iCLIP, eCLIP, RNA-compete, SEQRS, RBNS, RNA-MaP, HiTS-RAP, MITOMI. (RNA-centric) TRAP/RAT, RaPID, RiboTrap, RNA-assisted chromatography, protein microarray, MS2-BioTRAP, ChIRP, CHART, RAP-MS, Interactome capture **Motif discovery** MEME, RBPmap, SeAMotE, RNAcontext	Marchese *et al.* [34]
High-throughput characterization of protein–RNA interactions	Review of (a) experimental characterization of RBP-RNA interactions, (b) algorithms to predict RNA secondary structure and (c) motif finding tools	**Motif discovery** MEME, PhyloGibbs, cERMIT, DRIMUST, StructuRED, TEISER, RNAcontext, GraphProt, CMfinder, RNApromo, MEMERIS	Cook *et al.* [24]
Evaluating tools for TFBS prediction	Review and performance comparison of (a) *de novo* motif discovery tools and (b) transcription factors binding sites prediction tools	**Motif discovery** rGADEM, HOMER, ChIP- Munk, MEME-ChIP **Scanning motifs** Baycis, Cister, MCast, Comet, ClusterBuster, Matrix-Scan, Clover, FIMO, Patser, PossumSearch	Jayaram *et al.* [51]
Finding the target sitesof RNA-binding proteins	Comprehensive review of resources and methods to detect protein–RNA interactions. It focuses on the importance of the secondary structure of RNA	**Resources for RBP binding sites** ARESITE, CisBP-RNA, CLIPz (no longer available), doRiNA, RBPDB, Rfam, UTRSite **Motif discovery** MatrixREDUCE, MEME, MEMERIS, REFINE, AMADEUS, Aptamotif, CMfinder, cERMIT, COVE, FIRE, RNAalifold, RNAcontext, RNApromo	Li *et al.* [52]

high-throughput transcriptomic data [RNA sequencing (RNA-seq) and microarrays]. Next, we show different methods to discover the binding motifs of RBPs and, finally, we describe how to combine this information with experiment-specific data (expression of RBPs and relative usage of the exons in an event among others) to predict context-dependent regulators of AS. A summary of the pipeline followed can be found in Figure 1B.

Identification of as sites and events

An AS event is a local alteration of the splicing pattern on a gene, that in turn originates different isoforms. Some of these alterations occur more frequently and are called 'canonical' events. These canonical events are described in Supplementary Figure S1. In the case of canonical events, each event has two alternative configurations—the exon is either included or excluded, the 3′ extension can be included or excluded and so on. There are other events that can involve more complex patterns of AS for the same locus in the mRNA—e.g. three exons that can be skipped and are mutually exclusive.

Detecting the AS events is necessarily the first step in the identification of potential regulators: once the events with differential usage in different conditions are identified, using other computational methods it is possible to predict context-dependent regulators.

The task of identifying AS events has already been studied. Hooper [50] reviewed some tools that detect AS with RNA-seq data, but in the past 3 years, there has been a huge development of this family of algorithms. This section includes 33 methods to identify splicing events using either microarrays or RNA-seq data. Table 2 summarizes the reviewed algorithms. In the following paragraphs, we explain the criteria to include a method (rows of the Table 2) and their key characteristics (groups and columns of Table 2).

Criteria to include a method in the table. We focus on tools that detect AS events. We do not consider pipelines that quantify the expression of (novel) transcripts, such as Cufflinks [86, 87], MISO [88], SpliceGrapher [89] or Stringtie [90]. We do not either consider Nanopore or PacBio, as they are not suitable to pinpoint splicing events but the whole sequences of transcripts [91]. We do include methods that, taking as input isoform concentrations and structure, predicted by the previous or other algorithms, detect the presence of splicing events.

The reason why we only examine algorithms that detect AS events is that the transcriptome reconstruction is a problem much more difficult to solve. In fact, it was shown to be an NP-hard problem [92]. Different heuristics have been proposed, but they are far from perfect. Steijger *et al.* [93] stated that the best-performing methods have precision and recall around 40–50% at the transcript level for simulated data. This means that <50% of the predicted transcripts are correct and that <50% of the transcripts are recovered. These results worsen for complex genes with many transcripts. The same reference states that precision and recall rise to 80–90% if the analysis is performed at the exon level. Once the exons and the junctions that link them are known (i.e. given the splicing graph), the identification of AS events is straightforward. These facts make it more sensitive and reliable to focus on events than on transcripts to identify the potential regulators of AS. On the other hand, as the SFs bind to specific regions of the pre-mRNA, even if the isoforms were used as input to the algorithms, it would be necessary to perform the analysis at the event level (as many algorithms do).

We include some methods that use arrays. Our experience in the detection of events using microarrays and RNA-seq is that top results using both technologies show strong coherence between them [94]. RNA-seq, of course, has an edge on its ability to detect novel events, but the required computing resources using microarrays are much smaller. Therefore, we have decided to include microarray's methods that can be applied to junction arrays. This filter also helps to remove methods that are no longer maintained.

Key characteristics. Table 2 is split into three groups: methods based on RNA-seq that discover novel events, methods based on RNA-seq that do not discover novel events and methods based on arrays. The boundary between RNA-seq methods that discover novel events and methods that do not is blurred. Most algorithms that detect annotated events use as input the transcript structure (GTF file) and the estimated transcript expression. If this information is generated by an isoform deconvolution software such as Cufflinks, they can also be used to detect novel events, as Cufflinks (and other methods) predict novel transcripts given the RNA-seq data. In this case, these methods would be unraveling an non-deterministic polynomial-time hard combinatorial problem (isoform deconvolution) to solve a much easier one (event detection and quantification).

The main methodologies proposed to quantify AS events are the percent spliced-in (PSI or Ψ) and the splicing index (SI). PSI [95] is an estimate of relative usage of each alternative path (specific configuration of exons and/or junctions) of an AS event. Estimates of PSI can be validated using a third technology such as PCR (either quantitative or standard). On the other hand, the SI states the relative signal/coverage of an exon or a junction compared with the whole gene.

The SI has two drawbacks. AS every exon or junction has its own SI, the coherence of the SI change of the different exons and junctions involved in an event is not taken into account. For example, for a cassette event in which the exon is skipped in a tumoral condition, the SI of the junction that skips the cassette exon will be positive, the SI of the junctions of the cassette exon and the cassette exon itself will be negative. As the SI is not summarized for the whole event, it corresponds to the researcher to state the coherence between these signals. On the other hand, SI is difficult to be validated using PCR because it would require to run a PCR for every exon and junction to measure the average value. In contrast, the PSI value can easily be validated using PCR. Finally, SI may show spurious changes even for constitutive exons. Algorithms that return the PSI are therefore preferred. In Figure 2 it can be seen both PSI and SI calculations for a cassette event.

Some of RNA-seq methods use only the exon or only the junction reads to quantify the splicing events. Either of them are theoretically inferior to integrating both sources of information. Junction reads tend to be more scarce and more difficult to map than exon reads [96]. Methods based on junction reads are more sensitive to the characteristics of the aligner than methods that integrate them with exon reads [96]. On the other hand, exon reads alone can miss changes in isoforms that correspond to the less expressed path. The coverage of the junction that skips a cassette exon would be especially informative to state the change if its isoform is weakly expressed. Methods that exploit both sources of information are preferred.

Discussion. Mats—predecessor of rMats [64]—and DEXseq [58] were the first algorithms developed with this purpose. Both of them are actively maintained and several improvements, and new functionalities have been included in them. For example, the ability to detect novel events was included in rMats in late 2016, and DEXseq has recently improved its underlying statistical analysis. The statistics related with these two methods are briefly described in the additional material. rMats is based on the PSI, and DEXseq performs a statistical analysis indirectly based on SI.

Some published methods show a comparison with other algorithms. For example, Spladder [67], rMats [64], SpliceGrapher [89] and JuncBase [15] were compared by the developers of Spladder. Using simulated data, the number of detected events using JuncBase or SpliceGrapher is larger than using rMats. On the

Table 2. Algorithms for the identification of as events

Algorithm family	General aspects	Algorithm	Operating system	E.	S.	V.	PSI	Information used for quantification	References
RNA-seq novel events	(+) Detect nonannotated events	**AltAnalyze**	All	✓	✓	✓	×	Exons and junctions	[53]
		ASpli*	All	✓	✓	×	✓	Only Junctions	[54]
	(–) Time-consuming and complexity of the algorithms	**CASH**	All	✓	✓	×	~	Exons and junctions	[55]
		DEXseq	All	×	✓	×	×	Only Exons	[56–58]
		DiffSplice	Linux	ASM	✓	✓	✓	Exons and junctions	[59]
		EventPointer	All	✓	✓	✓	✓	Exons and junctions	[60]
		Gess	All	CE	✓	×	✓	Only exons	[61]
		JuncBASE	All	✓	✓	×	×	Only junctions	[15]
		Leafcutter	All	×	✓	✓	✓	Only junctions	[62]
		MAJIQ+VOILA	Linux	✓	✓	✓	✓	Junction reads	[63]
		rMATS	Linux	✓	✓	×	✓	Exons and junctions	[64]
		SGSeq	All	✓	×	✓	✓	Exons and junctions	[65]
		SPLADDER	All	✓	✓	×	✓	Exons and junctions	[66, 67]
		SplicePie	Linux	~	✓	×	✓	Exons and junctions	[68]
		SplicingTypesAnno	All	✓	✓	✓	✓	Exons and junctions	[69]
RNA-seq known events	(+) Better adapted to compare disparate experiments (+) Faster (–) Non-novel events	**ASATP***	All	✓	✓	✓	×	Expression of isoforms involved in event	[70]
		ASprofile	Linux	✓	×	×	×	Expression of isoforms involved in event	[71]
		DSGseq	All	×	✓	×	×	Exons	[72]
		IMAS*	All	×	✓	✓	✓	Exons	[73]
		SpliceR	All	✓	×	×	×	Expression of isoforms involved in event	[74]
		SpliceSEQ	All	✓	✓	✓	×	Exons and junctions	[75]
		SpliceTrap	Linux	✓	✓	×	✓	Exons	[76]
		SplicingCompass	All	×	✓	✓	×	Exons and junctions	[77]
		SplicingExpress	Linux	✓	✓	✓	×	Expression of isoforms involved in event	[78]
		SUPPA	All	✓	✓	✓	✓	Expression of isoforms involved in event	[79]
		Vast-Tools	Linux	✓	✓	✓	✓	Exons and junctions	[80]
Arrays	(+) Good performance (–) Non-novel events	**AltAnalyze**	All	✓	✓	✓	×	Only Exons	[53]
		EventPointer	All	✓	✓	✓	✓	Exons and junctions	[60]
		ExonPointer	All	CE	✓	✓	×	Exons and junctions	[81]
		IGems	All	×	✓	✓	×	Only Exons	[82]
		MADS+	All	✓	✓	✓	×	Exons and junctions	[83]
		RASA	NA	×	✓	×	×	Exons and junctions	[84]
		TAC 4.0	Windows	✓	✓	✓	×	Exons and junctions	[85]

Notes: *There is not a peer-reviewed reference for this algorithm. E: event classification; S: this method provides statistics; V: visualization; PSI: whether the PSI is returned; CE: cassette exon; ASM: alternative splicing module (any type of event without labeling the canonical ones). It is divided into three groups: algorithms that use RNA-seq to discover novel and non-novel events, and microarray-based algorithms. Other characteristics, such as algorithm's input data and some comments of each algorithm can be found in the Supplementary Table S1.

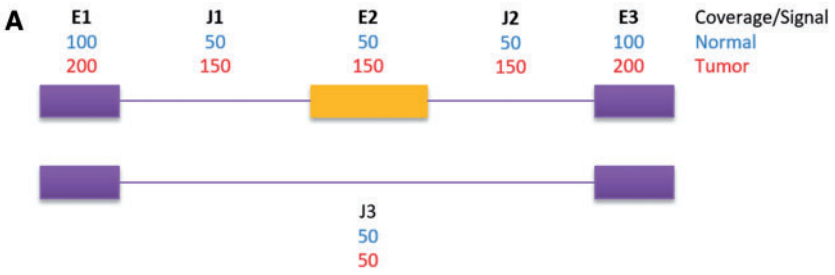
contrary, rMats shows better true discovery rates than JuncBase when top-ranked events were selected. Spladder outperforms all the others in any respect in this comparison.

SUPPA developers compared it against MISO and MATS. In this study, MATS (an event-based method) was shown to outperform MISO (a transcript deconvolution method) and FIMMO to be the algorithm with best performance [79]. SGSeq developers compare it against methods based on the deconvolution of the transcriptome. The conclusion of both references is that PSI estimation is more reliable when using event-based methods than when using methods based on transcriptome deconvolution. These conclusions must be taken with prudence, as there can be confirmation bias.

Some algorithms do not discover novel events. Despite this disadvantage, these algorithms can be advisable in some situations. First, they can be better adapted to compare disparate experiments as long as the same reference transcriptome is used across such experiments. Second, these methods can provide results in much shorter time, using a fast isoform-quantification algorithm such as Kallisto [97] or Salmon [98]. It is up to the user to decide if the discovery of novel events is worth the additional burden of time and storage or the difficulties in performing meta-analyses.

Table 2 can be used as a guide to select the proper AS detector. If the analysis requires the detection of novel events, only the algorithms of the firs group can be used. Among them, those that use information of exons and junctions and provide the PSI and event classifications should be preferred (i.e. rMats, EventPointer-SGSeq, SPLADDER or SplicingTypesAnno).

If novel events are not required, using algorithms in the second group—probably after quantifying the isoforms using Kalisto or Salmon—is preferred. Among them, several algorithms return the PSI and classify the corresponding events [i.e. SpliceTrap (69), SUPPA (29) or Vast-Tools(687)].



B

	E1	J1	E2	J2	E3	J3
SI (normal)	0.58	-0.42	-0.42	-0.42	0.58	-0.42
SI (tumor)	0.42	0.00	0.00	0.00	0.42	-1.58
ΔSI	-0.17	0.42	0.42	0.42	-0.17	-1.17

PSI (normal)	0.50
PSI (tumor)	0.75
ΔPSI	0.25

$$Ave.Coverege = \frac{E1 + E2 + E3 + J1 + J2 + J3}{6}$$

$$SI_{E1} = \log_2\left(\frac{E1}{Ave.Coverege}\right)$$

$$PSI = \frac{(J1 + E2 + J2)/3}{J3 + (J1 + E2 + J2)/3}$$

Figure 2. (**A**) Toy example of an exon cassette with differential splicing across two conditions (normal and tumor). Coverage of exons and junctions in both conditions are included. (**B**) Computation of SI and PSI for a toy example. SI is computed using the log ratio of the coverage of each exon/junction with the average coverage of the whole gene-hear simplistically consider as the average of the coverage of the exons and junctions of the gene. On the other hand, PSI considers the ratio of the mean coverages of the exons and junctions that include the cassette exons (J1, J2 and E2) and the sum of the coverages of both isoforms.

Table 3. Computational methods aimed at discovering motifs

Algorithm subtype	Main algorithms	Algorithm with best performance (reference)
Based on RNA primary structure	MEME [101], cERMIT [102], phyloGibbs [102], GLAM2 [103], HOMER [104], ChIP- Munk [105], DREME [106], rGADEM [107], MEME-ChIP [108], DRIMUST [109], RBPmap [110], SeAMotE [111]	**rGADEM** (Jayaram *et al.* [51])
Based on RNA secondary structure	MEMERIS [112], RNApromo [113], StructRED [114], RNAcontext [115], CMfinder [115], TEISER [116], mCarts [117], GraphProt [118]	**RNAcontext*** and **MatrixREDUCE** (Kazan *et al.* [115])

Notes: *Note that the authors of RNAcontext algorithm are also the authors of its corresponding comparative review. These methods extract binding motifs using CLIP-, RIP- or CHIP-seq experiments.

Finally, the number of algorithms using arrays is smaller. EventPointer, AltAnalyze and IGEMs have been recently deployed and showed their performance in real data. Only EventPointer returns the PSI.

Discovering binding motifs of SFs

There are two main approaches to identify pairs of proteins–RNAs that have affinity to bind together: using the output data of biological experiments that characterize protein–RNA interactions [such as cross-linking and immunoprecipitation sequencing (CLIP-seq), photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), RNA immunoprecipitation sequencing (RIP-seq)] and analyzing protein and RNA structures to predict potential binding sites. The first method combines experimental data with computational algorithms (Table 3), whereas the second is purely computational.

Only two references lie on the second family of methods [99, 100]. The computational burden of these two methods make them non-suitable to be applied genome-wide but to check the interaction of specific pairs of RBP-RNAs. These methods are not used in any of the computational approaches to find context-dependent SFs (Table 5).

Discovering binding motifs using protein-centric experiments

Experimental methods that characterize protein–RNA interactions can be divided into protein-centric methods, which identify binding RNAs for a particular protein, and RNA-centric methods, which discover the proteins that interact with a specific RNA region. RNA-centric methods are not strictly

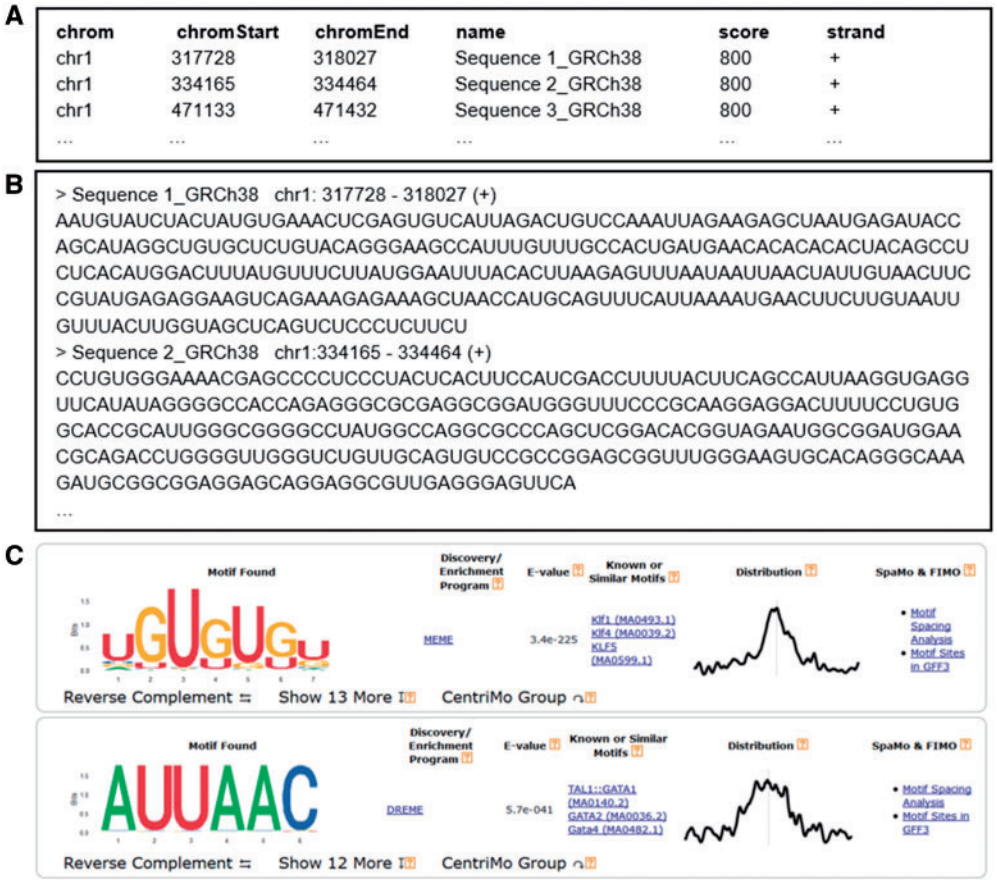


Figure 3. (**A**) Common output data (BED file) of RBP-RNA interaction experiments such as CLIP-seq. It consists of the genomic ranges in which RBP bind to, the sequence name, a binding score and the strand. (**B**) Input data of most motif discovery algorithms: a FASTA file of sequences of RNA. (**C**) Output of MEME-ChIP algorithm: logos of PWMs, the algorithm used to find each motif, E-values of the discovered PWM, similar known motifs, centered distribution of the motif in input sequences and other options to perform additional analyses.

high-throughput, as only a few proteins bind to a specific locus in the genome. Therefore, we focus the discussion on the protein-centric ones (see Section 2 of Supplementary Material for details).

Most protein-centric experimental methods rely on RNA immunoprecipitation, namely, a protein antigen is precipitated using a protein-specific antibody, followed by RNA identification using either microarrays or RNA-seq. RNA immunoprecipitation (RIP) and CLIP are the main methodologies [14, 119]. They differ on the protocol for immunoprecipitation.

In RIP techniques, a protein antigen is precipitated using a protein-specific antibody, and RNAs are identified using either microarray (RIP-chip) or RNA-seq (RIP-seq) [119]. The main limitation of RIP relies on the low resolution and background noise that causes the detection of nonspecific interactions.

Protein-centered methods were improved with ultraviolet cross-linking and denaturing techniques (CLIP) and can be measured by RNA-seq (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HiTS-CLIP) [14]). Figure 3A shows an example of the output data of a CLIP-seq experiment (BED file). More details about these experimental methods can be found in the review of Marchese *et al.* [34].

Protein-centric techniques usually return a collection of RNA sequences attached or close to the binding sites of RBPs. The corresponding genomic regions for these sequences are long (50–500 nt) compared with the loci where the RBPs bind, typically a few nucleotides. There are several databases that compile information of these experiments, such as DoRiNA [120], CLIPdb [121] or POSTAR [122] (Table 1).

Identifying short recurring motifs makes it possible to predict RBP-mRNA potential binding sites without the need of a RIP or CLIP experiment. Algorithms to find recurring motifs in multiple, unaligned and long sequences are known as motif discovery algorithms. The input of these programs is a collection of sequences (FASTA files) given by an RIP or a CLIP experiment. The output is a set of motifs that appear recurrently in the given sequences. These motifs are usually represented as position weight matrices (PWMs). An example of the input and output of this family of algorithms is shown in Figure 3.

RNAcompete (with the companion database CisBP) performs a different approach: it uses a pool of nucleotide k-mers randomly generated to determine the preferred RNA sequence of an RBP [123]. Once RNAs bind to a tagged RBP, they are pulled down with a fluorescent label and measured by microarrays. This method outputs k-mers of nucleotides whose affinity to the RBPs is especially high. Using this method, the step of discovering motifs can be skipped, as RNAcompete directly provides the binding motif with high affinity to the RBPs. Only a procedure to merge motifs with high-affinity that are similar is required.

Motif discovery algorithms have been reviewed before [24], so we only briefly describe their main features to provide a wide view of the complete pipeline. In the cited review (Table 1) and

in the Section 3 of the Supplementary Material, the reader can find a deeper description of these algorithms.

Most of these methods have been borrowed from the detection of transcription factor-binding sites (TFBSs). Only later, they were applied to the detection of splicing factor binding sites (SFBSs). Among all the features that differentiate TFBS from SFBS—such as SFBS specific motifs, motif length or preferable location in RNA—SFBS discovering algorithms consider, almost exclusively, the primary and secondary structure of RNA, as stated in [24].

The simplest approach to detect motifs is to use only nucleotide sequences. Within this group, one of the most widely used algorithms is MEME [101]. This tool uses probabilistic models based on the maximum likelihood estimation to look for recurring and fixed-length motifs from unaligned sequences.

The MEME algorithm belongs to a broad set of motif-based tools called MEME-suite [124], which contains several variants of this software. DREME [106] uses other models for discovering motifs, GLAM2 [103] allows finding gapped motifs with arbitrary insertions or deletions and MEME-ChIP [108] is an algorithm that performs a comprehensive motif analysis. The MEME-ChIP algorithm, additionally, incorporates other useful motif-based functions, such as analyzing the similarity of predicted motifs with known motifs (TomTom [125]), automatically grouping predicted motifs by similarity (CentriMo [126]), predicting preferred spacing between pair of motifs (SpaMo [127]) and creating a GFF file for visualizing the predicted motifs in integrative genomics viewer [128] or any genome browser.

Other algorithms, using similar approaches, enable finding ungapped motifs: phyloGibbs (which incorporates phylogeny) [102], SeAMotE [111] and cERMIT [102]; and gapped motifs: HOMER [104], ChIP- Munk [105], rGADEM [107], MatrixREDUCE [129], DRIMUST [109] and RBPmap [110].

The MEMERIS algorithm [112] is an extension of MEME that combines primary and secondary structure to find motifs. It uses the single-strandedness information of sequences as prior knowledge in the MEME's expectation maximization model.

Other algorithms are StructRED [114], which uses mRNA expression levels in addition to the FASTA files, RNAcontext [115], which is available on the RBPmotif Web server [115], GraphProt [118], which uses a graph-based encoding, CMfinder [115], mCarts [117], TEISER [116] and RNApromo [113]. These methods are deeply described in the reviews cited in Table 1.

Several authors have compared motif-discovering algorithms [51, 130]. Jayaram *et al.* [51] evaluated their performance using ChIP-Seq data. In this analysis, they showed that rGADEM was the best-performing tool for discovering motifs.

Prediction of splicing regulatory factors

The final step in the pipeline is the identification of the RBPs that induce differential AS events across the conditions of the study. This section is split into two parts: scanning the SF motifs in the splicing regions and identifying the potential regulators by using some type of enrichment analysis. These tasks are depicted in Figure 4A and B.

Scanning SFs' motifs in splicing regions

Once the SFs' motifs are known, they are scanned across the splicing regions. This approach can potentially save costs, as binding sites can be predicted without having to use protein-centric experiments. On the other hand, the predicted binding sites can be used to make sound hypothesis on the potential regulators to be validated by an ulterior RIP or CLIP experiment.

Algorithms to scan motifs in nucleotide sequences have been deeply studied and reviewed, as they are a key element for unveiling TFBS [51, 132–134]. As it occurs with motif discovery algorithms, these methods were adapted from algorithms developed to scan TFBS. These methods can be divided into methods to find individual occurrences and methods to discover clusters of binding sites (Table 4, Supplementary Figure S2).

FIMO [135] is a software of the MEME-Suite, which allows finding individual occurrences of motifs in DNA, RNA or protein sequences. It computes a log-likelihood ratio for each motif in each position in the given sequences and calculates the associated q-values assuming a model in which sequences are randomly generated. This method was found to outperform others when detecting TFBS [51].

Cis-regulatory modules (CRMs) are sets of RBP motifs locally enriched in the given sequences. CRM discovering algorithms return a single score for each CRM that combines the matches of its RBP motifs. A set of RBP motifs with a global significant score could provide evidence that they are acting together.

The MCAST algorithm [136] yields a list of predicted CRMs ranked by *E*-value. Each CRM represents a group of PWMs that frequently appear together in query sequences. MCAST was found to outperform any other algorithm to find TFBS clusters in [51].

Other algorithms used to discover CRMs such as BayCis [137], Cister [138], Cluster-Buster [139], CisModule [140] or EMCModule [141] were also reviewed in [51].

Motif enrichment analysis and refinement of results

Once the putative binding sites of RBPs in the transcriptome are known, RBPs can be associated to a gene (RxG) or to a splicing event (RxE). Methodologies that predict SFs can be divided into two broad groups depending on these relationships: methods focused on genes (RxG) and methods focused on individual AS events (RxE).

The first group consists of finding in the literature (or using the mapping) the relationships between RBPs and genes (RxG) and comparing the recurrence of RBPs in genes with spliced and non-spliced events.

Using the RxG strategy, de Miguel *et al.* [148] discovered the key role of protein quaking (*QKI*) in the regulation of splicing in non-small cell lung cancer (NSCLC). They identified the events (only cassette exons) using ExonPointer [81] and performed an enrichment analysis of genes with differentially spliced exons in different gene sets of putative regulators. *QKI* was found to be the most significantly enriched gene set. Experimental work showed the functional implications of the depletion of *QKI* in NSCLC cell lines.

A straightforward refinement is the study of individual splicing events and their potential RBP regulators—inferred by a motif scanning algorithm (RxE). RBP motifs hit regions where AS events occur. RBPs whose hits are significantly enriched in differentially spliced events are potential regulators of AS (Figure 4B).

Following this methodology, Danan-Gotthold *et al.* [16] analyzed splicing events with a potential role in solid tumors and predicted putative regulators of the splicing patterns for each tumor type. They developed their own algorithms to perform the motif scanning and the estimation of PSI for exon events. To assess statistical significance, they compared the frequency of occurrences between spliced and non-spliced events using a Fisher's exact test.

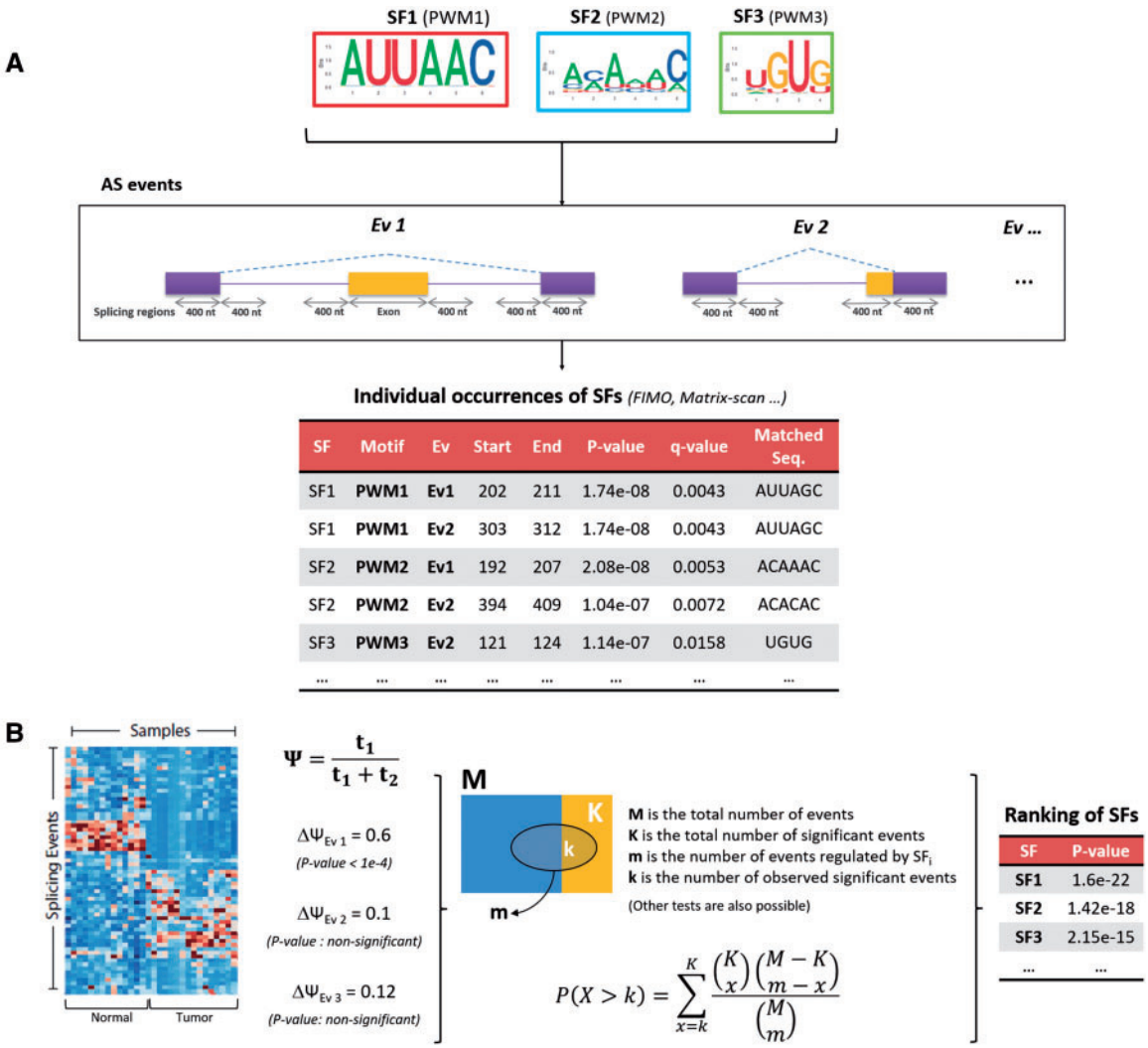


Figure 4. Example of the pipeline for predicting context-dependent SFs (SFs). (**A**) Scanning SFs' motifs in splicing regions (typically 300–400 nt upstream and downstream the AS events [131]). A set of three PWMs associated to three SFs is shown. PWMs are examined in the splicing regions three different AS events. A statistical analysis is performed to get a table of individual occurrences, which contains the hits of each motif against the events. (**B**) Performing a motif enrichment analysis using ΔΨ (PSI) of AS events. An example of the statistics is shown. Other tests are also possible. The main output of the pipeline is a ranking of SFs, which are predicted to regulate the splicing pattern under study.

Table 4. Computational methods aimed at scanning motifs against DNA/RNA regions

Algorithm subtype	Main algorithms	Algorithm with best performance (reference)
Individual occurrences	Clover [142], PossumSearch [143], FIMO [135], Matrix-scan and Patser [144–146]	**FIMO** (Jayaram *et al.* [51])
Cluster of binding sites	Cister [138], Comet [147], MCAST [136], Cluster-Buster [139], CisModule [140], EMCModule [141], BayCis [137]	**MCAST** (Jayaram *et al.* [51])

In a similar approach, Sebestyén *et al.* [149] carried out a study of the alterations of RBPs in cancer and associated splicing changes. They performed a comprehensive analysis of 1300 RBPs in multiple tumors of TCGA. They analyzed mutation, copy number and gene expression patterns combined with AS changes and the binding motif enrichment analysis of spliced events. AS events were identified from a transcript quantification based on a known annotation using SUPPA [79]. They used FIMO [135] to scan motifs in splicing regions considering a hit if the *P*-value was <0.001. They evaluated the differential number of hits between spliced regions and non-spliced regions of the same size controlling for the G + C content. Finally, they measured the possible influence of RBPs by relating their expression with the splicing pattern of each event. With this methodology, they discovered that *MBNL1*—an SF associated with cell differentiation—controls the AS of several genes involved in the cell.

Aghamirzaie *et al.* [150] developed a different method called CoSpliceNet, which is based on coexpression networks of

transcripts and SFs. They found RBPs that are strongly correlated with transcripts. Then, they used MEME to find conserved motifs in intron and exon sequences adjacent to events (i.e. in a cassette exon they differentiated four regions: Intron-3′, Exon-3′, Intron-5′ and Exon-5′) and found motifs in each of them separately. Finally, they identified significantly enriched motifs and constructed a co-splicing network.

We depict the application of these and other similar approaches in Table 5. This table includes the required inputs, the output of the algorithms as well as other characteristics.

Case study

To illustrate the whole pipeline and the difficulties that appear in each step, we include a worked case study. References in Table 5 perform its own analysis making a succession of decisions, such as selecting the AS detecting algorithm, downloading an RBP motif database, choosing a motif scanning algorithm or performing an enrichment test.

This case study is performed on a previous experiment with some collaborators (GSE 76902) [60]. In this experiment, the SF *SRSF1* is knocked down using small interfering RNA (siRNA) on the A549 lung adenocarcinoma cell line. The experiment includes three conditions: cells treated only with the vehicle of the transfection (Lipofectamine 2000, Invitrogen), cells treated with scramble siRNA (i.e. a sequence that will not lead to the specific degradation of any cellular mRNA) and cells transfected with an siRNA that targets *SRSF1*. These three groups are referred to as Control, SCR and KO-*SRSF1*, respectively. Each condition has three biological replicates that, in turn, are hybridized three times (nine hybridizations on HTAv2 microarrays).

As a preliminary step, we compared the expression changes (Figure 5A and B) of 1243 genes that code RBPs between conditions SCR and KO-*SRSF1* to confirm the knock down effect of *SRSF1* and to evaluate the expression changes of other RBPs. Aroma.affymetrix pipeline was performed to summarize the expression values for each gene. Differential expression was performed using LIMMA [153]. *P*-values were adjusted for multiple testing using the Benjamini–Hochberg procedure. Interestingly, not only *SRSF1* significantly changed its expression but also other RBP genes.

We found 134 RBPs (of 1243) with differential changes of expression (adjusted *P*-value < 0.05 and $|\log_2\text{-fold change}| > 0.5$). As expected, *SRSF1* had the best *P*-value (adjusted *P*-value = 9.1e-29) with a $\log_2$FC of -1.5. The differential expression of RBPs occurred in both directions but not with the same proportion (we found 72 and 28% of RBPs downregulated and upregulated, respectively).

For identifying AS events, we compared the splicing pattern of KO-*SRSF1* against SCR cells. EventPointer was used to discover the AS events, as it is the only algorithm that returns the PSI value using arrays. If the experiment had been performed with RNAseq, rMats or Spladder would be the methods of choice as described in the section 'Identification of as sites and events'. We set a filter based on the expression of genes—if the gene is not expressed, there is no point in discussing splicing. All genes whose expression was under quantile 0.25 in all the samples were discarded. Of the theoretical 97 482 events interrogated by the array, 35 963 pass the expression threshold and 3686 showed a *P*-value < 0.001 according to the LogFC test (approximately 4% of the events). The application of this expression filter when identifying the AS events is crucial to ignore irrelevant events, as it will be shown later.

The sequence of the neighborhood of the events (400 nt upstream or downstream, equivalent to Figure 4A: splicing regions) was extracted taking into account their corresponding strand for every splicing event interrogated by the array HTAv2. The size of the neighborhood is somehow arbitrary, but it is in a range according to [131].

To assess the validity of the 400 nt selection, we used two different CLIP data sets that target *SRSF1*: CLIP-seq data of HEK 293 human cell line from Sanford *et al.* [47] and CLIP-seq data of mouse embryo fibroblasts from Pandit *et al.* [46]. We mapped these data sets against the human genome and found that most CLIP hits (∼70%) were located within the selected 400 bp window (Supplementary Figure S3). The mapping between the mouse and the human genomes was performed using the liftover tool of UCSC [154, 155].

Motif enrichment analysis was performed using the PWMs from the ATtRACT database. This database contains the largest number of PWMs collected from different resources.

Many RBPs in ATtRACT include several nearly identical annotated PWMs collected from different studies. We grouped similar PWMs into a single motif using the Kullback Leibler (KL) divergence [156] (Figure 6C). If two motifs of a certain RBP are similar (KL < 0.5; Figure 6B), they are merged into a single one (for convenience, we selected the longest one). Following this criterion, we got 487 PWMs (24% of PWMs were lost).

We used the FIMO algorithm to scan these PWMs against the neighboring regions of AS events (as recommended in [51]). We built the background as a one-order Markov model and set the default threshold of *P*-values to consider a significant hit (*P*-value $<$ 1e-4). We constrained FIMO to search only in the strand of the corresponding gene.

Consensus binding motifs or RBPs are often too short to get statistically significant matches. In our analysis, motifs with six nucleotides or less gave no significant hits (FIMO's *P*-value $<$ 1e-4). In the ATtRACT database, almost 44% of motifs have six nucleotides or less (Figure 6A). Interestingly, the motif length is related to the information content of motifs and to the significance of hits (Figure 6D).

There are motifs whose entropy is high, i.e. they lack well-defined binding sites. For this reason, 10% of motifs with 7 nt or more had no significant hits when scanned against the transcriptome. We finally got significant matches for 445 PWMs that correspond to 125 RBPs.

Once we have the significant hits against the event regions, we studied the significance of RBPs in differentially spliced events by using a Fisher's exact test. In total, 14 of 125 RBPs were significantly enriched (Fisher *P*-value $<$ 1e-3; Table 6). *SRSF1* was one of them (Fisher *P*-value = 8.32E-04). However, 12 of the 14 RBPs were even more significant than *SRSF1*. These findings could be considered false positives, as the only direct interaction was precisely *SRSF1*. Interestingly, 9 of the 14 RBPs are differentially expressed (Table 6) and 13 of the 14 RBPs have strong relationships—direct or indirect—with *SRSF1* according to the STRING database [157] (Figure 7) and [158]. Somehow, these false positives are showing the relationships of these RBPs in the experiment and the tight coupling among the SFs, as the depletion of *SRSF1* provokes significant changes in the expression of other SFs.

It is important to note that, before applying the expression filter described above, the enrichment *P*-values were inaccurate (for example, the *P*-value of *SRSF1* was nonsignificant).

We evaluated whether *SRSF1* promotes exon inclusion or exclusion by comparing PSI values between KO-*SRSF1* and SCR samples (selecting just the cassette exon events). *SRSF1* was found to be positive splicing regulator (*P*-value = 5.82e-06), which is in accordance with the bibliography [45, 47].

Table 5. Main approaches to find context-specific SFs

Reference	Title	Inputs	Ouputs	Platform	Novel events	Software	Type of AS events (algorithm)	Data access (samples)	RBPs	Description and comments
Danan-Gotthold *et al.* [16]	Identification of Recurrent Regulated Alternative Splicing Events across Human Solid Tumors	E, Ψ, RxE	1, 2	RNA-seq	Yes	No	Only cassettes	TCGA	*RBFOX2*, *QKI*, *CELF2*, *MBNL1*, *MBNL2* and *PTBP1*	A large-scale study of AS in human solid tumors
Sebestyán *et al.* [149]	Large-Scale Analysis of Genome and Transcriptome Alterations in Multiple Tumors Unveils Novel Cancer-Relevant Splicing Networks	E, Ψ, RxE, O	1, 2, 3	RNA-seq	No	No	CE, IR, A3, A5, AF, AL, MX (SUPPA)	TCGA	1348 RBPs (104 with motifs, CISbp)	Analysis of widespread alterations in the expression of RBP genes, novel mutations and copy number variations in association with multiple AS changes in cancer drivers and oncogenic pathways
Aghamirzaie *et al.* [150]	CoSpliceNet: a framework for co-splicing network inference from transcriptomics data	E, Ψ*, RxE	1, 2, 3	RNA-seq and Arrays	No	Yes (open-source)	Isoform-specific analysis	GEO: GSE74692	Defined by the user	A tool for co-splicing network inference, which can be used to identify SFs and their candidate targets pre-mRNAs
Zhang *et al.* [151]	MYCN Controls an Alternative RNA Splicing Program in High-Risk Metastatic Neuroblastoma	E, Ψ*, RxE	1, 2	RNA-seq	No	No	Isoform-specific analysis	dbGaP: phs000868	*RBFOX1*, *RBFOX3*, *CELF2*, *CELF6*, *PTBP1* and *HNRNPA1*	Analysis of transcription factors that regulate SF genes by analyzing splicing patterns
de Miguel *et al.* [148]	A Large-Scale Analysis of Alternative Splicing Reveals a Key Role of QKI in Lung Cancer	Ψ, RxG	1	Arrays (HJAY)	No	No	Only cassettes (ExonPointer)	Non-Public	Targets of *QKI*. Manually curated.	Analysis of AS of lung cancer and *QKI*
Correa *et al.* [152]	Functional Genomics Analyses of RNA-Binding Proteins Reveal the Splicing Regulator SNRPB as an Oncogenic Candidate in Glioblastoma	E, Ψ, O	1	RNA-seq and Arrays (Human U219)	No	No	CE, MX, IR, A3, A5, AF, AL	TCGA & SRA	1542 RBPs (CISbp)	A systematic study of RBPs in GBM
Sveen *et al.* [5]	Aberrant RNA Splicing in Cancer; Expression Changes and Driver Mutations of Splicing Factor Genes	E, Ψ, O	1	RNA-seq, Arrays	Yes	No	CE, IR, A3, A5, AF, AL, MX (SpliceSeq)	TCGA, MediSapiens	261 RBPs	Review of aberrant RNA splicing and its regulation in several cancer types
Brooks *et al.* [15]	Conservation of an RNA regulatory map between *Drosophila* and mammals	Ψ, RxE	1, 2	RNA-seq	Yes	No	CE, IR, A3, A5, AF, AL, MX (JuncBASE)	Non-Public	*NOVA1*, *NOVA2*	Analysis of the conservation of RNA regulatory elements of *NOVA1* and *NOVA2*

Notes: The table headers display the following fields: Inputs (E: expression of RBPs; Ψ: percent spliced-in; Ψ*: isoform relative usage; RxG: RBP–gene relationships; RxE: RBP–event relationships; O: other genetic sources of information such as copy number variations, or mutations). Outputs (1: SFs that regulate a specific condition; 2: SFs that regulate a specific event; 3: interaction network of RBPs). Platform: platform for mRNA measuring. Type of events (CE: cassette event; IR: intron retention; A3: alternative 3′; A5: Alternative 5′; AF: alternative first; AL: alternative last; MX: mutual exclusive; C: complex event). RBPs covered in each work.

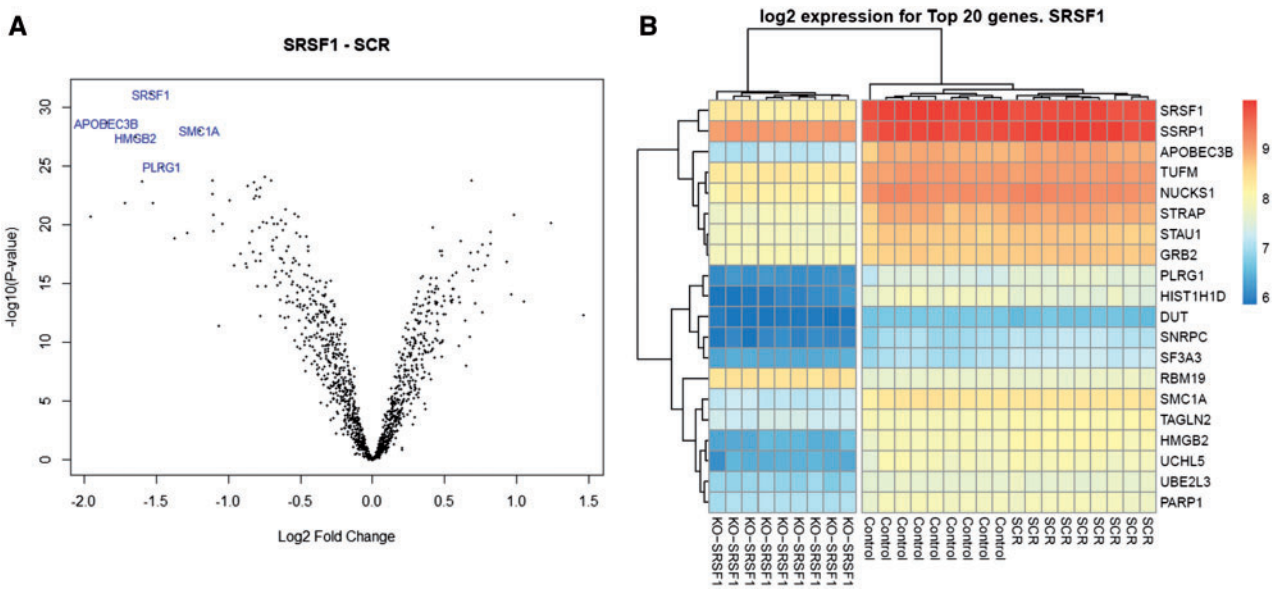


Figure 5. (**A**) Volcano plot of RBP genes corresponding to a LIMMA analysis that compares KO-*SRSF1* versus SCR. Top five genes are highlighted. (**B**) Heatmap of log2 expression of the 20 most enriched RBP genes among three conditions (Control, SCR and KO-*SRSF1*).

Finally, we evaluated the enrichment of AS events in the aforementioned CLIP experiments using also a Fisher test. The enrichment for the union of both experiments was more significant than for any predicted motif (Fisher *P*-value = 1.25E-23). It is interesting to note that the Pandit's CLIP-seq data were strongly significant (more than any other test), despite being data from a different organism.

According to these results, the identification of *SRSF1* as driver of this change is difficult to pinpoint, as other SFs are even more significant than it. However, *SRSF1*'s enrichment *P*-value was strongly significant (in fact, it is in the position 13 of 125 RBPs). In addition, most of the other significantly enriched RBPs were differentially expressed and related to *SRSF1*. Even in the case of RBPs that are not differentially expressed (*TIA1* and *TIAL1*), there is strong evidence of their interaction with *SRSF1* [158] not reflected in the STRING database. Consequently, these RBPs could be participating in the regulation of AS splicing as well. The differential expression of the SFs across the studied conditions helps to filter out some SFs.

Somehow, the described experiment was optimal to 'discover' the SF regulating the differentially spliced events. However, the enrichment analysis alone was not sufficient to infer the key role of *SRSF1*. The combination of the enrichment analysis with differential expression is a must to uncover the key regulators in the experiment. In fact, most of the methods in Table 5 combine both sources of information.

Discussion and conclusion

We have outlined a conceptual computational pipeline to infer AS regulators. The first step is to detect the AS events, the second is to predict RBP-mRNA-binding sites and the last is, using both pieces of information, to predict the context-dependent SFs that regulate splicing in a specific condition.

Regarding the detection of AS events, we have already discussed qualitatively the different algorithms in the corresponding section. It would be desirable a comparison that states their performance also quantitatively. However, this task is not trivial at all: events are difficult to match across algorithms, different outputs (SI and PSI for example) can hardly be compared, a ground truth simulated experiment able to fairly compare the methods should be designed and, of course, the results should be compared in real samples with a proper validation strategy. Nevertheless, the provided comparison is still useful and can be used to guide the researcher to find the algorithms that better suit his/her needs.

Regarding the prediction of RBPs' binding sites, the review includes different algorithms that discover motifs based on RIP and CLIP methods. These motifs—along with other ones obtained from RNAcompete techniques—are included in databases. Using these databases could potentially save costs, as there is no need to perform additional biological experiments to predict candidate binding sites of a RBP in a specific sample. It is important to point out that the broadest database collects motifs of for only around 30% of known RBPs.

Once the motifs are selected, different software packages identify the loci in the transcriptome where there are putative hits of these motifs. All the methods reviewed to scan PWMs are borrowed from the detection of TFBSs. Although the algorithms developed to scan motifs in DNA sequences can also be used with RNA sequences, this is a simplification. As previously stated, the secondary structure of RNA and the specific characteristics of SFBS play a key role in the binding process. A potential improvement of specificity and sensitivity would be achieved by this information, as some methods to extract the PWMs do it. Specifically, MEMERIS and GraphProt use the secondary structure of RNA to check single-stranded regions.

The PWMs for RBPs are usually short and repetitive and, consequently, prone to have too many potential binding sites in the transcriptome. This fact, in turn, makes it difficult to find hits that are statistically significant. As we have pointed out in the case study, only motifs >7 nt achieve statistical significance. Short motifs—6 nt or less—were discarded by the motif scanning algorithm.

The splicing machinery is complex. The interaction networks and synergistic effects of RBPs should also be considered. The process of regulation of the splicing is guided by a group of RBPs acting as a whole and not only by their individual activity

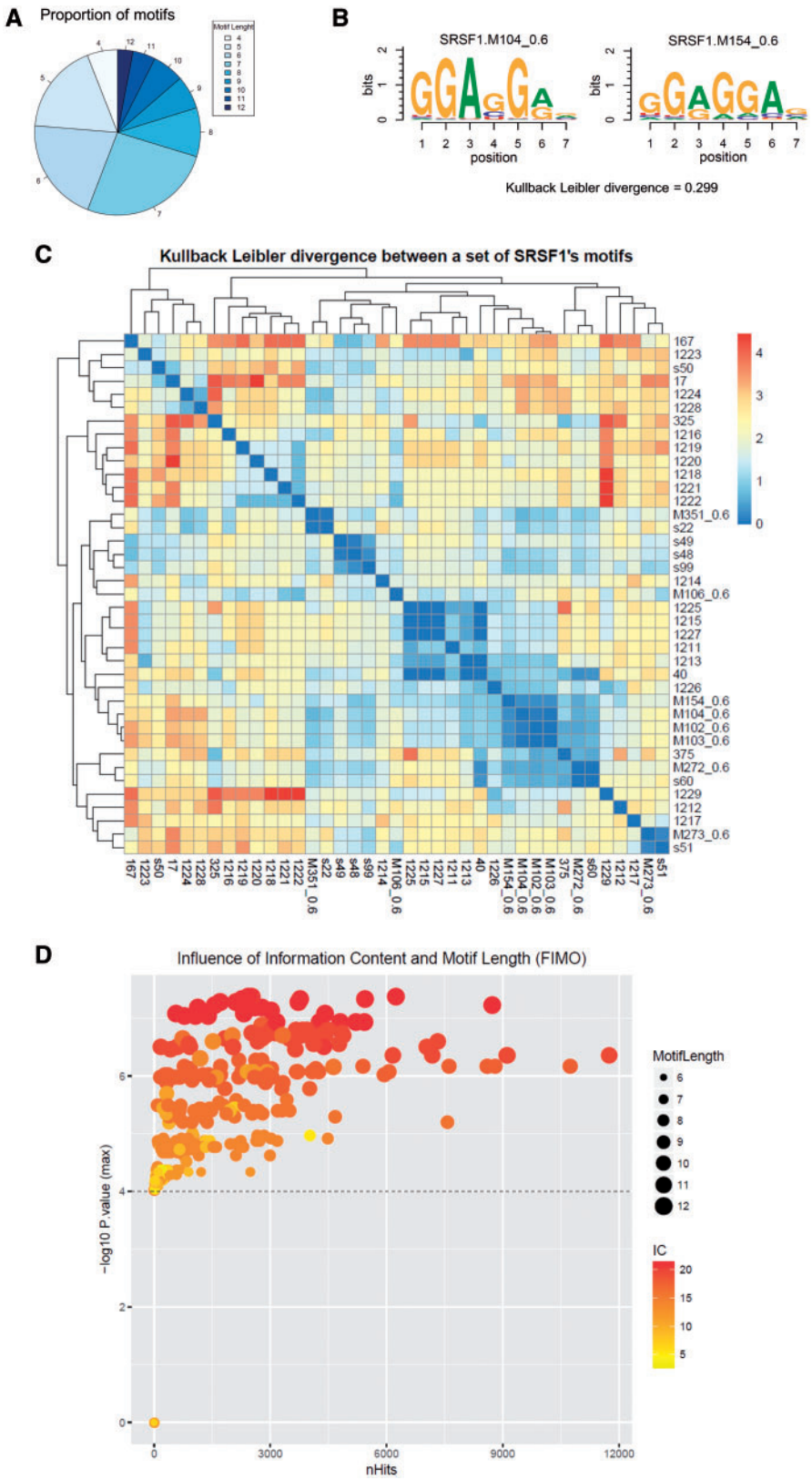


Figure 6. (**A**) Proportion of *Homo Sapiens*' motifs of ATtRACT database according to their length. (**B**)Two similar PWMs of *SRSF1* which were joined together. (**C**) KL divergence between a set of *SRSF1*'s motifs. Every pair of motifs with KL divergence <0.5 were merged. (**D**) Each dot represents a PWM. The influence of information content (IC) and motif length with the number of significant hits (nHits; FIMO's *P*-value < 1e-4) and the significance of hits (−log10 of the best hit using FIMO) is shown. PWMs with no significance hits (FIMO's *P*-value < 0.001) were discarded and not shown in the figure.

Table 6. RBPs Predicted to be splicing regulators (fisher *P*-value < 1e-3)

RBP ranking	RBP/CLIP-seq	PWM	Fisher *P*-value (PSI)	LIMMA adjusted *P*-value (differential expression)
	Clip-seq_ *SRSF1*_union	–	1.25E-23	**9.18E-29**
	Clip-seq_ *SRSF1*_Pandit	–	8.30E-19	**9.18E-29**
1	*ELAVL2*	ELAVL2.1093	7.60E-15	**1.35E-04**
2	*TIAL1*	TIAL1.1287	1.75E-13	2.55E-01
3	*TIA1*	TIA1.1284	3.78E-11	NA in HTAv2
4	*ELAVL4*	ELAVL4.1095	9.17E-11	2.04E-01
	Clip-seq_ *SRSF1*_Sanford	–	2.28E-09	**9.18E-29**
5	*ELAVL1*	ELAVL1.161	1.39E-08	**2.57E-15**
6	*AKAP1*	AKAP1.97	1.15E-06	**5.62E-07**
7	*ELAVL3*	ELAVL3.119	2.82E-05	4.32E-01
8	*SSB*	SSB.58	2.64E-04	1.80E-01
9	*HNRNPH2*	HNRNPH2.925	3.03E-04	4.60E-02
10	*TRA2A*	TRA2A.s77	5.52E-04	2.40E-02
11	*SF1*	SF1.120	5.66E-04	**5.83E-12**
12	*SRSF2*	*SRSF2*.1311	6.83E-04	**4.05E-16**
13	*SRSF1*	*SRSF1*.1223	8.32E-04	**9.18E-29**
14	*PTBP1*	PTBP1.1012	9.13E-04	**5.10E-09**

Notes: CLIP-seq data of Pandit and Sanford are also included. The ranking of RBPs, the RBPs' name, the best PWMs, the Fisher *P*-values of PSI and the adjusted *P*-values of the enrichment analysis are shown (LIMMA adjusted *P*-values < 1e-3 in bold).

(i.e. a given genomic sequence could be differently recognized by the same RBP depending on the expression of the other RBPs). This fact makes the elucidation of the regulators a much harder problem. The case study illustrates the tight control of the expression of different SFs. More than 100 SFs showed strong differential expression across the conditions. This differential expression makes it difficult to pinpoint which of the differentially expressed SFs is the driver of the change.

Finally, we summarized different works that applied these methodologies for deciphering context-dependent SFs in a certain experiment. The procedure focused on RBP–gene interactions (RxG) is the simplest pipeline, as there is no need to predict or scan RBP motifs against a transcriptome. However, using this approach, two SFs that regulate different AS events of the same genes cannot be distinguished from each other. This drawback can be resolved by analyzing RBP–event relationships (RxE).

All these methods sensibly combine information of the expression of the RBPs with overrepresentation of their putative targets in the corresponding experiment. Overrepresentation alone does not seem to be sufficient to accurately identify the drivers of the changes.

One potential reason is that the specificity and sensitivity of the methods to map PWMs are far from perfect. As stated above, one of the problems is the PWMs themselves: many of them are too short to predict the binding sites accurately. The weakest part of the pipeline is the identification of the binding sites for the RBPs: the computational prediction of these sites is prone to errors (both false positives and false negatives).

In fact, in the case study, the overrepresentation of SFBS using CLIP-seq data instead of motif-scanning was much more significant. In the long term, once the RIP- and CLIP- based techniques are settled down and results for most SF readily

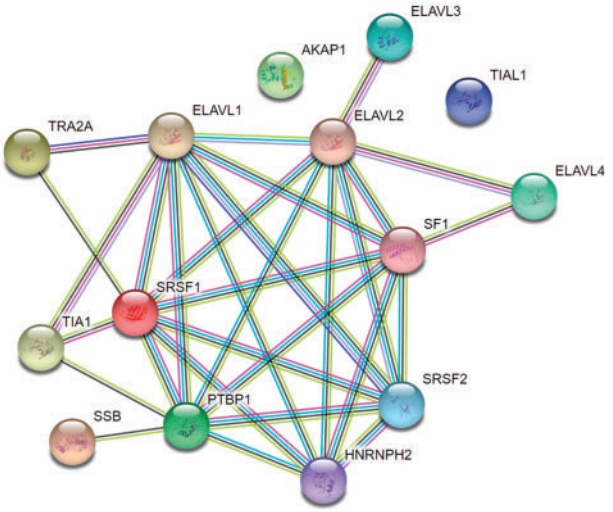


Figure 7. String's interactions network of the 15 significantly enriched RBP genes (Fisher *P*-value < 1e-3; number of nodes: 14; number of edges: 31). *TIAL1* does not appear in the STRING database, but it is also related to *SRSF1* [158].

available, it makes more sense to use this information than scanning the putative motifs (even for different cell lines).

Despite the concerns described in this work, it is possible to predict splicing regulators with acceptable sensitivity and precision. In fact, different functional studies showed that the predictions were indeed correct. The described methodologies do not substitute RIP- and CLIP- based experiments but complement them by providing some candidates to be driver regulators in the condition under study. Besides, this approach could help the scientific community to understand the regulation networks of SFs and infer groups of SFs that cooperate in the regulation of AS.

Key Points

- Deciphering the regulation of AS is conceptually divided into three steps: detection of AS events, estimation of their interactions with SFs and contextualization for a specific experiment.
- There are many methods to detect and quantify AS events. Most of them are recent. Several algorithms that use RNA-seq data detect novel unannotated events.
- The motifs of the SFs tend to be small and repetitive making it difficult to have good precision pinpointing the binding sites.
- The algorithms that include the contextualization of the results have helped to discover novel roles in the regulation of splicing that were experimentally validated.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors are grateful to Francisco J. Planes, Lucía Campuzano and Xabier Cendoya for their comments on the preparation of this manuscript.

Funding

This work was supported by the Basque Government with the grant promoting doctoral theses for young pre-doctoral researcher (grant number PRE_2016_1_0194 to F.C.).

References

1. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 1977;**74**:3171–5.
2. Pan Q, Shai O, Lee LJ, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**:1413–15.
3. Modrek B, Lee C. A genomic view of alternative splicing. *Nat Genet* 2002;**30**(1):13–19.
4. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 2010;**463**:457–63.
5. Sveen A, Kilpinen S, Ruusulehto A, *et al*. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* 2016;**35**:2413–27.
6. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. *Nat Biotechnol* 2004;**22**(5):535–46.
7. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. *Biochim Biophys Acta Mol Basis Dis* 2009;**1792**(1):14–26.
8. Lim KH, Ferraris L, Filloux ME, *et al*. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci USA* 2011;**108**(27):11093–8.
9. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**(5):646–74.
10. Allemand E, Myers MP, Garcia-Bernardo J, *et al*. A broad set of chromatin factors influences splicing. *PLoS Genet* 2016;**12**: e1006318.
11. Luco R, Pan Q, Tominaga K, *et al*. Regulation of alternative splicing by histone modifications. *Science* 2010;**327**(5968): 996–1000.
12. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. *Trends Genet* 2015;**31**(5): 274–80.
13. Ule J, Ule A, Spencer J, *et al*. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* 2005;**37**(8):844–52.
14. Licatalosi DD, Mele A, Fak JJ, *et al*. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008;**456**(7221):464–9.
15. Brooks AN, Yang L, Duff MO, *et al*. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res* 2011;**21**(2):193–202.
16. Danan-Gotthold M, Golan-Gerstl R, Eisenberg E, *et al*. Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Res* 2015; **43**(10):5130–44.
17. Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol cell Biol* 2014;**15**(2):108–21.
18. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* 2009;**10**(11):741–54.
19. Fu XD, Ares M, Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* 2014;**15**: 689–701.
20. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* 2007;**8**: 479–90.
21. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 2008;**14**(5):802–13.
22. Izquierdo JM, Majós N, Bonnal S, *et al*. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell* 2005;**19**:475–84.
23. Glisovic T, Bachorik JL, Yong J, *et al*. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 2008; **582**(14):1977–86.
24. Cook KB, Hughes TR, Morris QD. High-throughput characterization of protein-RNA interactions. *Brief Funct Genomics* 2015;**14**(1):74–89.
25. Maris C, Dominguez C, Allain FHT. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* 2005;**272**(9):2118–31.
26. Braddock DT, Louis JM, Baber JL, *et al*. Structure and dynamics of KH domains from FBP bound to single-stranded DNA. *Nature* 2002;**415**(6875):1051–6.
27. Lewis HA, Musunuru K, Jensen KB, *et al*. Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* 2000;**100**: 323–32.
28. Laver JD, Li X, Ancevicius K, *et al*. Genome-wide analysis of Staufen-associated mRNAs identifies secondary structures that confer target specificity. *Nucleic Acids Res* 2013;**41**(20): 9438–60.
29. Theunissen O, Rudt F, Guddat U, *et al*. RNA and DNA binding zinc fingers in Xenopus TFIIIA. *Cell* 1992;**71**(4):679–90.
30. Hall TMT. Multiple modes of RNA recognition by zinc finger proteins. *Curr Opin Struct Biol* 2005;**15**(3):367–73.
31. Castello A, Fischer B, Eichelbaum K, *et al*. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 2012;**149**(6):1393–406.
32. Ray D, Kazan H, Cook KB, *et al*. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 2013; **499**:172–7.
33. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet* 2014;**15**(12):829–45.
34. Marchese D, de Groot NS, Lorenzo Gotor N, *et al*. Advances in the characterization of RNA-binding proteins. *Wiley Interdiscip Rev RNA* 2016;**7**:793–810.
35. Liu ZP, Liu S, Chen R, *et al*. Structure alignment-based classification of RNA-binding pockets reveals regional RNA recognition motifs on protein surfaces. *BMC Bioinformatics* 2017; **18**:27.
36. Han H, Braunschweig U, Gonatopoulos-Pournatzis T, *et al*. Multilayered control of alternative splicing regulatory networks by transcription factors: molecular cell. *Mol Cell* 2017; **65**(3):539–53.
37. Giudice G, Sánchez-Cabo F, Torroja C, *et al*. ATtRACT—a database of RNA-binding proteins and associated motifs. *Database* 2016;**2016**:baw035.
38. Giulietti M, Piva F, D'Antonio M, *et al*. SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res* 2013;**41**:125–31.
39. Cook KB, Kazan H, Zuberi K, *et al*. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* 2011;**39**:301–8.
40. Aken BL, Achuthan P, Akanni W, *et al*. Ensembl 2017. *Nucleic Acids Res* 2017;**45**(D1):D635–42.
41. Weyn-Vanhentenryck SM, Mele A, Yan Q, *et al*. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* 2014;**6**(6):1139–52.

42. Rossbach O, Hung LH, Khrameeva E, *et al*. Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biol* 2014;**11**(2):146–55.
43. Charizanis K, Lee KY, Batra R, *et al*. Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy. *Neuron* 2012;**75**(3):437–50.
44. Daughters RS, Tuttle DL, Gao W, *et al*. RNA gain-of-function in spinocerebellar ataxia type 8. *PLoS Genet* 2009;**5**(8): e1000600.
45. Änkö ML, Müller-McNicoll M, Brandl H, *et al*. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol* 2012; **13**(3):R17.
46. Pandit S, Zhou Y, Shiue L, *et al*. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell* 2013;**50**(2):223–35.
47. Sanford JR, Wang X, Mort M, *et al*. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* 2009;**19**(3):381–94.
48. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 2010; **11**:75–87.
49. Warf MB, Berglund JA. The role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* 2010;**35**(3):169–78.
50. Hooper JE. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum Genomics* 2014; **8**:3.
51. Jayaram N, Usvyat D, R Martin AC. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* 2016, doi:10.1186/s12859-016-1298-9.
52. Li X, Kazan H, Lipshitz HD, *et al*. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA* 2014;**5**: 111–30.
53. Emig D, Salomonis N, Baumbach J, *et al*. AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res* 2010;**38(Suppl 2)**:W755.
54. Mancini E, Iserte J, Yanovsky M, *et al*. ASpli: analysis of alternative splicing using RNA-Seq. Bioconductor: R package version 1.2.3, 2017.
55. Wu W, Zong J, Wei N, *et al*. CASH: a constructing comprehensive splice site method for detecting alternative splicing events. *Brief Bioinform* 2017, doi:10.1093/bib/bbx034.
56. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
57. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**(12):550.
58. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;**22**(10):2008–17.
59. Hu Y, Huang Y, Du Y, *et al*. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res* 2013;**41**:e39.
60. Romero JP, Muniategui A, De Miguel FJ, *et al*. EventPointer: an effective identification of alternative splicing events using junction arrays. *BMC Genomics* 2016;**17**:467.
61. Ye Z, Chen Z, Lan X, *et al*. Computational analysis reveals a correlation of exon-skipping\nevents with splicing, transcription and epigenetic factors. *Nucleic Acids Res* 2014;**42**(5): 2856–69.
62. Li YI, Knowles DA, Pritchard JK. LeafCutter: annotation-free quantification of RNA splicing. *Nature Genetics* 2017;**50**:151–8.
63. Vaquero-Garcia J, Barrera A, Gazzara MR, *et al*. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 2016;**5**:e11752.
64. Shen S, Park JW, Lu Z, *et al*. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci USA* 2014;**111**(51):E5593–601.
65. Goldstein LD, Cao Y, Pau G, *et al*. Prediction and quantification of splice events from RNA-seq data. *PLoS One* 2016;**11**(5): e0156132.
66. Drewe P, Stegle O, Hartmann L, *et al*. Accurate detection of differential RNA processing. *Nucleic Acids Res* 2013;**41**(10):5189–98.
67. Kahles A, Ong CS, Zhong Y, *et al*. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* 2016;**32**(12):1840–7.
68. Pulyakhina I, Gazzoli I, 't Hoen PA, *et al*. SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Res* 2015;**43**: 11068.
69. Sun X, Zuo F, Ru Y, *et al*. SplicingTypesAnno: annotating and quantifying alternative splicing events for RNA-Seq data. *Comput Methods Programs Biomed* 2015;**119**(1):53–62.
70. Z-G-L Github. Alternative Splicing Analysis Tool Package (ASATP). https://github.com/Z-G-L/ASATP.
71. Florea L, Song L, Salzberg SL. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Res* 2013;**2**:188.
72. Wang W, Qin Z, Feng Z, *et al*. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* 2013;**518**(1):164–70.
73. Han S, Lee Y. IMAS: integrative analysis of multi-omics data for alternative splicing. Bioconductor: R package version 1.0.0, 2017.
74. Vitting-Seerup K, Porse BT, Sandelin A, Waage J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* 2014;**15**:81.
75. Ryan MC, Cleland J, Kim R, *et al*. SpliceSeq: a resource for analysis and visualization of RNA-seq data on alternative splicing and its functional impacts. *Bioinformatics* 2012; **28**(18):2385–7.
76. Wu J, Akerman M, Sun S, *et al*. Splice trap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics* 2011;**27**(21):3010–16.
77. Aschoff M, Hotz-Wagenblatt A, Glatting KH, *et al*. SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* 2013;**29**(9):1141–8.
78. Kroll JE, Kim J, Ohno-Machado L, *et al*. Splicing express: a software suite for alternative splicing analysis using next-generation sequencing data. *PeerJ* 2015;**3**:e1419.
79. Alamancos GP, Pagès A, Trincado JL, *et al*. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* 2015;**21**(9):1521–31.
80. Irimia M, Weatheritt RJ, Ellis JD, *et al*. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 2014;**159**(7):1511–23.
81. De Miguel FJ, Sharma RD, Pajares MJ, *et al*. Identification of alternative splicing events regulated by the oncogenic factor *SRSF1* in lung cancer. *Cancer Res* 2014;**74**(4):1105–15.
82. Sood S, Szkop KJ, Nakhuda A, *et al*. iGEMS: an integrated model for identification of alternative exon usage events. *Nucleic Acids Res* 2016;**44**(11):e109.
83. Shen S, Warzecha CC, Carstens RP, *et al*. MADS+: discovery of differential splicing events from Affymetrix exon junction array data. *Bioinformatics* 2010;**26**(2):268–9.
84. Seok J, Xu W, Davis RW, *et al*. RASA: robust alternative splicing analysis for human transcriptome arrays. *Sci Rep* 2015;**5**: 11917.

85. ThermoFisher. Transcriptome Analysis Console (TAC) software 4.0. 2017. https://www.thermofisher.com/es/es/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/affymetrix-transcriptome-analysis-console-software.html.
86. Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–15.
87. Trapnell C, Hendrickson DG, Sauvageau M, *et al*. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013;**31**:46–53.
88. Katz Y, Wang ET, Airoldi EM, *et al*. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;**7**:1009–15.
89. Rogers MF, Thomas J, Reddy AS, *et al*. SpliceGrapher: detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. *Genome Biol* 2012;**13**(1):R4.
90. Pertea M, Pertea GM, Antonescu CM, *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5.
91. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 2015;**13**(5):278–89.
92. Bernard E, Jacob L, Mairal J, *et al*. Efficient RNA isoform identification and quantification from RNA-seq data with network flows. *Bioinformatics* 2014;**30**(17):2447–55.
93. Steijger T, Abril JF, Engström PG, *et al*. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* 2013;**10**(12):1177–84.
94. Romero JP, Ortiz-Estévez M, Muniategui A, *et al*. Comparison of RNA-seq and Microarray Platforms for Splice Event Detection using a Cross-Platform Algorithm. *bioRxiv* 2017, 197798.
95. Wang ET, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008;**456**(7221):470–6.
96. Baruzzo G, Hayer KE, Kim EJ, *et al*. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods* 2016;**14**:135–9.
97. Bray NL, Pimentel H, Melsted P, *et al*. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**:525–7.
98. Patro R, Duggal G, Love MI, *et al*. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**(4):417–19.
99. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 2011;**12**(1):489.
100. Bellucci M, Agostini F, Masin M, *et al*. Predicting protein associations with long noncoding RNAs. *Nat Methods* 2011;**8**:444–5.
101. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;**2**:28–36.
102. Siddharthan R, Siggia ED, Van Nsmwegea E. PhyloGibbs: a gibbs sampling motif finder that incorporates phylogeny that incorporates phylogeny. *PLoS Comput Biol* 2005;**1**:0534–56.
103. Frith MC, Saunders NF, Kobe B, *et al*. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* 2008;**4**(5):e1000071.
104. Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89.
105. Kulakovskiy IV, Boeva VA, Favorov AV, *et al*. Deep and wide digging for binding motifs in ChIP-seq data. *Bioinformatics* 2010;**26**(20):2622–3.
106. Bailey TL. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011;**27**(12):1653–9.
107. Mercier E, Droit A, Li L, *et al*. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChiP-Seq. *PLoS One* 2011;**6**(2):e16432.
108. Machanick P, Bailey TL. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* 2011;**27**(12):1696–7.
109. Leibovich L, Paz I, Yakhini Z, *et al*. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Res* 2013;**41**:174–9.
110. Paz I, Kosti I, Ares M, *et al*. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res* 2014;**42**:361–7.
111. Agostini F, Cirillo D, Ponti R, *et al*. SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC Genomics* 2014;**15**(1):925.
112. Hiller M, Pudimat R, Busch A, *et al*. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res* 2006;**34**(17):e117.
113. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci USA* 2008;**105**(39):14885–90.
114. Li X, Quon G, Lipshitz H, *et al*. Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure. *RNA* 2010;**16**(6):1096–107.
115. Kazan H, Ray D, Chan ET, *et al*. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol* 2010;**6**:28.
116. Goodarzi H, Najafabadi H, Oikonomou P, *et al*. Systematic discovery of structural elements governing mammalian mRNA stability. *Nature* 2012;**485**(7397):264–8.
117. Zhang C, Lee KY, Swanson MS, *et al*. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res* 2013;**41**(14):6793–807.
118. Maticzka D, Lange SJ, Costa F, *et al*. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol* 2014;**15**(1):R17.
119. Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc* 2006;**1**(1):302–7.
120. Blin K, Dieterich C, Wurmus R, *et al*. DoRiNA 2.0-upgrading the dorina database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* 2015;**43**(D1):D160–7.
121. Yang YC, Di C, Hu B, *et al*. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* 2015;**16**:51.
122. Hu B, Yang YC, Huang Y, *et al*. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* 2016;**45**:(D1):D104–14.
123. Ray D, Kazan H, Chan ET, *et al*. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* 2009;**27**(7):667–70.
124. Bailey TL, Boden M, Buske FA, *et al*. MEME suite: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**:202–8.
125. Gupta S, Stamatoyannopoulos JA, Bailey TL, *et al*. Quantifying similarity between motifs. *Genome Biol* 2007;**8**(2):R24.

126. Bailey TL, MacHanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* 2012;**40**:e128.
127. Whitington T, Frith MC, Johnson J, *et al*. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res* 2011;**39**(15):e98.
128. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al*. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**(1):24–6.
129. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 2006;**22**:141–9.
130. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.
131. Barash Y, Calarco JA, Gao W, *et al*. Deciphering the splicing code. *Nature* 2010;**465**(7294):53–9.
132. Bulyk ML. Computational prediction of transcription-factor binding site locations. *Genome Biol* 2003;**5**(1):201.
133. Hannenhalli S. Eukaryotic transcription factor binding sites—modeling and integrative search methods. *Bioinformatics* 2008;**24**(11):1325–31.
134. Tran NT, Huang CH. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol Direct* 2014;**9**:4.
135. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;**27**(7):1017–18.
136. Bailey TL, Noble WS. Searching for statistically significant regulatory modules. *Bioinformatics* 2003;**19(Suppl 2)**:ii16.
137. Lin TH, Ray P, Sandve GK, *et al*. BayCis: a Bayesian hierarchical HMM for cis-regulatory module decoding in metazoan genomes. In: *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology*. Springer-Verlag, Singapore, 2008, Vol. 4955, 66–81.
138. Frith MC, Hansen U, Weng Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* 2001;**17**(10):878–89.
139. Frith MC, Li MC, Weng Z. Cluster-buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* 2003; **31**(13):3666–8.
140. Zhou Q, Wong WH. CisModule: *de novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci USA* 2004;**101**(33):12114–19.
141. Gupta M, Liu JS. *De novo* cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci USA* 2005;**102**(20): 7079–84.
142. Frith MC, Fu Y, Yu L, *et al*. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 2004;**32**(4):1372–81.
143. Beckstette M, Homann R, Giegerich R, *et al*. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* 2006;**7**:389.
144. Hertz GZ, Hartzell GW, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Bioinformatics* 1990;**6**(2):81–92.
145. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999;**15**(7–8):563–77.
146. Turatsinze JV, Thomas-Chollier M, Defrance M, *et al*. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc* 2008;**3**: 1578–88.
147. Frith MC. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 2002;**30**(14):3214–24.
148. de Miguel FJ, Pajares MJ, Martínez-Terroba E, *et al*. A large-scale analysis of alternative splicing reveals a key role of QKI in lung cancer. *Mol Oncol* 2016;**10**(9):1437–49.
149. Sebestyén E, Singh B, Miñana B, *et al*. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res* 2016;**26**(6):732–44.
150. Aghamirzaie D, Collakova E, Li S, *et al*. CoSpliceNet: a framework for co-splicing network inference from transcriptomics data. *BMC Genomics* 2016;**17**(1):845.
151. Zhang S, Wei JS, Li SQ, *et al*. MYCN controls an alternative RNA splicing program in high-risk metastatic neuroblastoma. *Cancer Lett* 2016;**371**(2):214–24.
152. Correa BR, de Araujo PR, Qiao M, *et al*. Functional genomics analyses of RNA-binding proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma. *Genome Biol* 2016;**17**:125.
153. Smyth GK. Limma: linear models for microarray data. In: Gentlemana R, Carey V, Dudoit S, *et al*. (eds). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer, 2005, 397–420.
154. Kent WJ, Sugnet CW, Furey TS, *et al*. The human genome browser at UCSC. *Genome Res* 2002;**12**(6):96–1006.
155. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 2009;**25**(14):1841–2.
156. Kullback S, Leibler RA. On information and sufficiency. *Inst Math Stat* 1951;**22**:79–86.
157. Szklarczyk D, Morris JH, Cook H, *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017; **45**(D1):D362–8.
158. Delestienne N, Wauquier C, Soin R, *et al*. The splicing factor ASF/SF2 is associated with TIA-1-related/TIA-1- containing ribonucleoproteic complexes and contributes to post-transcriptional repression of gene expression. *FEBS J* 2010; **277**(11):2496–514.