Published in final edited form as:
Trends Cancer. 2021 April ; 7(4): 347–358. doi:10.1016/j.trecan.2020.12.015.

Computing the Role of Alternative Splicing in Cancer

Zhaoqi Liu[1,2,*], **Raul Rabadan**[3,4,*]

[1]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

[2]China National Center for Bioinformation, Beijing 100101, China

[3]Program for Mathematical Genomics, Columbia University, New York, NY 10032, USA

[4]Departments of Systems Biology and Biomedical Informatics, Columbia University, New York, NY 10032, USA

Abstract

The vast majority of human genes undergo alternative splicing, and dysregulation of alternative splicing contributes to tumor initiation and progression. Computational analysis of genomic and transcriptomic data enables the systematic characterization of alternative splicing and its functional role in cancer. In this review, we summarize the latest computational approaches to studying alternative splicing in cancer and the current limitations of the most popular tools in this field. Finally, we describe some of the current computational challenges in the characterization of the role of alternative splicing in cancer.

Keywords

alternative splicing; computational analysis; cancer; spliceosomal mutations

mRNA Splicing and Altered Regulation in Cancer

Pre-mRNA splicing is required for the maturation of almost all mammalian mRNAs. Alternative splicing (AS) refers to the process by which a pre-mRNA can be processed into different mature mRNA molecules where an exon/intron could be differentially included/excluded by the choice of alternative specific splice sites (Box 1). Alternative splicing enables variable transcripts from the same DNA template, and plays an extensive role in generating protein complexity [1]. It has been estimated that in humans around 95% of genes undergo alternative splicing to produce a large variety of transcripts in a cell, tissue type, and condition-specific manner [2, 3], which suggests that most cellular processes are dependent on the splicing machinery.

[*]Correspondence: liuzq@big.ac.cn (Z.Liu), rr2579@cumc.columbia.edu (R.Rabadan).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Accumulated evidence shows that aberrations in the splicing process could contribute to cancer initiation, progression, and treatment failure through switching isoform expression of key proteins involved in apoptosis, metabolism, and cell signaling [4, 5]. For instance, alternate isoforms of pyruvate kinase M (PKM) and epidermal growth factor receptor (EGFR) that are frequently expressed in glioma affect metabolism and promote tumor proliferation [6, 7]. The variant isoform of CD44 is well studied in many cancer types and associated with epithelial to mesenchymal transition [8, 9]. In melanoma, expression of splicing isoforms of BRAF(V600E) lacking the RAS-binding domain confers resistance to RAF inhibitors [10]. Similarly, in prostate cancer, expression of the androgen-receptor isoform encoded by splice variant 7 lacking the ligand-binding domain is associated with resistance to enzalutamide and abiraterone [11].

Cancer-associated AS events can occur by two main mutation mechanisms. *Cis*-acting somatic mutations can hinder splicing of individual introns or generate new splice sites. For instance, distinct splice-altering mutations are found in the p53 tumor suppressor gene (TP53), introducing novel stop codons that truncate the protein [12]. Splicing can also be deregulated by *trans*-acting mutations in splicing regulatory proteins including SR proteins, hnRNPs and other splicing factors (Box 1). Dysfunction of such proteins may have a larger impact on splicing dysregulation and even alter the entire transcription network [13]. Recently, large-scale genomic analysis has revealed the mutational landscape of splicing-related genes in human cancers [14] and provided genetic evidence directly linking RNA splicing regulation to cancer (Box 2).

Given the high prevalence of splicing dysfunction in cancer and its pervasive effect on the transcriptome, significant computational efforts are needed and have been invested for the identification and quantification of AS events on a genome-wide scale. Computational analyses provide a more complete understanding of how splicing dysfunction alter splicing globally in cancer, and become a fundamental step before down-stream experimental investigation in most studies of cancer splicing. The scope of this review is the discussion of the latest development, possible improvement and current challenges of computational studies in characterization the role of alternative splicing in cancer.

Computational Deciphering of Splicing Dysregulation

The increase in read-depth and decrease in cost of high-throughput RNA-sequencing data (RNASeq) has enabled the systematic characterization of alternative splicing in a context-dependent manner (Figure 1). The analysis of these data was enabled by a variety of computational tools developed in the last few years [15–17]. However, the output of these tools varies significantly, sometimes with dramatic differences, leading to conflicting interpretations [16].

These computational tools mainly fall into two methodological categories: AS detection on the 1) whole transcript level or 2) specific event level (Figure 1B). Early studies used transcriptome deconvolution to reconstruct full-length isoforms and quantify the relative expression abundances of each isoform (for instance, Cufflinks [18], DiffSplice [19] and MISO [20]). However, transcriptome reconstruction is overall a challenging problem and is especially complicated in long genes with many transcripts [21]. It is often more convenient to directly focus on each AS event given the specific exon and junction information. For this reason, most of the extensively used and validated tools used today are event-based (for instance, rMATS [22], MAJIQ [23] and JuncBASE [24]). In these tools, local AS events are first identified in each sample using variable exon reads and junction reads (linking exons or cryptic intronic splice sites) between biological conditions or from a background annotation dataset. Next, a value is assigned to quantify the ratio of expression switch on each AS event. The most commonly used measure is called Percent-Spliced-In (PSI), a value in the interval zero to one, which provides the fraction of mRNA reads supporting each AS event. Adjustments of PSI evaluation can be found across different tools, including normalization to junction and read length (rMATS), correcting for GC content (MAJIQ), and batch difference between samples (JUM [21]). After quantification and correction, one can identify significant alternatively spliced events using proper statistical evaluation across experimental conditions.

While many tools limit their detection power to currently well-annotated references, detecting unannotated events with novel splice sites requires different strategies. For instance, *SF3B1* hotspot mutations induce novel 3'ss usages, many of which are not reported in the latest annotation of functional isoforms [25]. One conventional solution is to enlarge the feed-in reference by generating a dataset-specific .gtf file, by using tools (e.g. Cufflinks) to conduct the de novo isoform reconstruction. Most of the leading tools have been updated in recent years to include the feature of novel AS detection, which is more computationally intensive and often requires additional experimental validation.

These computational tools mainly report five common patterns of AS: skipping or inclusion of a cassette exon, alternative 5′ or 3′ splice site choice, intron retention, and mutually exclusive exons (Figure 1B), although certain complex or mixed pattern of AS can occur [21]. Many of these tools (e.g. rMATS, MISO, JuncBASE) preferentially report exon-inclusion/skipping events, which are the most frequent AS pattern in animals. However, intron-related AS events have drawn increasing attention for their role in understanding tumorigenesis [26] and treatment design [27]. Identifying true-positive intron retention events is a difficult task, as it requires manual review of putative events in IGV due to the repetitive nature of intronic sequences (inaccurate read mapping) or un-annotated small/non-coding transcripts from the antisense strand. Recently, an annotation-free tool, JUM, has been specifically designed for quantifying intron retention by requiring approximately uniformly distribution of reads across the entire intronic region to reduce false positive calls [21].

Some studies are not designed with distinct conditions affecting splicing, for instance, investigating any potential effects of splicing in a specific tumor cohort without prior knowledge of any splicing changes. These studies proceed first by the description and characterization of all AS events and then by the identification of potential regulators of these AS events. The most straightforward way is to directly correlate the inclusion level of each AS with different RBP status (e.g. genetic alterations or transcriptomic expression). This approach was used in a *trans*-splicing quantitative trait loci (sQTL) analysis that linked somatic single nucleotide variant (SNV) positions with alternative splicing changes in 8,255 samples [28]. In another example, all known binding motifs of each RBP were screened for a significant enrichment for matching nucleotide sequences in alternative splicing regions [29]. Such analyses are limited, because not all key splicing-related proteins directly bind to RNA, and not all RBPs have been confirmed with high confidence motifs.

A systematic evaluation of differential splicing tools applied to four datasets, using PCR-validated splicing events as the background truth, found that MAJIQ and rMATS out-performed other tools overall [16]. However, it is still highly recommended to use more than one tool due to the relatively large variability of the results reported by the different approaches [16]. Besides direct employment of publicly available tools, uniquely designed/modified algorithms with enhanced sensitivity and specificity will no doubt be more powerful when applied to specific datasets by incorporating prior knowledge of the context-dependent scientific question.

Notably, the computational workflow described above is built on second-generation short-read RNA-seq technique. Natural limitations of short-read sequencing do exist to have an impact on AS detection, such as low unique-mapping rate, especially at complex loci. However, short-read RNA-seq still represents the standard and widely used method in cancer splicing analysis, not merely because the extensive computational efforts (as summarized above) but the low cost to produce high throughput reads. Intriguingly, a few studies estimate the effect of sequencing depth and length of short read RNA-seq on splicing analysis [16, 30, 31]. Overall, these studies suggest a minimum of 50 million reads per sample and length of 100 bp serving as a baseline for accurate splicing quantification.

The increasing use of long-read **Nanopore or PacBio sequencing** (see Glossary) have provided improved reconstruction of full spectrum of isoform profiles and solutions to many of the drawbacks of using short-reads in splicing (for instance, identification of full-length transcripts with retained intron). To date, growing interests and requirements accelerate the fast-pace development of computational tools (nicely archived at “https://long-read-tools.org/”) for long-read sequencing in the last decade [32]. A part of these tools, for instance Iso-Con [33], SQANTI [34] and FLAIR [35], enable the full-length detection of alternative spliced transcripts. Typically, key steps of such detection pipelines include reads error-correction, subgroup clustering, reads collapsing and isoform annotation. Currently, the study of AS analysis using long-read technique is still at its early stage, and continuous efforts are needed to reduce the high false-positive rate of detected isoforms. And high-quality isoform annotation tools and databases are required to keep pace with the novel transcript identification. Meanwhile, accurate quantification of isoform expression is still challenging, due to the relative low reads counts and sequencing coverage biases [32]. Meanwhile, sort-read techniques provide an excellent option to improve these limitations, because it has a larger throughput, lower error rates, and are widely used for many other analyses beyond splicing. Future best practices may involve coupled analysis using both techniques [36].

Computational Refinement of Cancer-associated Aberrant Splicing

The next challenge after a successful characterization of AS events is the functional interpretation of their effects, i.e. how specific events may contribute to the diverse phenotypes expected in cancer cells (Figure 1B). The goal of this part of the workflow is to determine which of AS events are functionally relevant to cancer out of the full list of identified events. The first step is to extract significant changes between conditions, focusing on recurrent and robust/reproducible AS changes. One may apply different thresholds to the output of the computational splicing analysis, including thresholding the statistical *q*-value, the absolute changes in PSI, and the median read counts across replicates. For instance, when identifying cryptic 3'ss induced by *SF3B1* mutations, a minimum PSI change of 0.2 is recommended (Low-abundance isoforms that confer gain-of-function or dominant-negative effect do exist, but are in general rare.) Ideally, we wish no cryptic reads (PSI=0) from wide type samples, under the assumption that an AS event will act as a perfect switch to turn on or off the carcinogenic 3'ss selection. However, after investigating the splicing patterns in more than 10,000 TCGA (The Cancer Genome Atlas) samples, one recent study found that this assumption does not reflect the biological reality. Widespread occurrence of weak cryptic 3′ss usage by many well-known targets of mutant SF3B1 (e.g. *MAP3K7*, *PPP2R5A*) was detected in samples without *SF3B1* lesions and even in normal cells [37]. This result indicates that these cryptic 3′ss are inherently active and very faintly present in normal conditions, but are dramatically elevated in cases with *SF3B1* hotspot mutations.

Another way to determine relevant AS events is by overlapping the identified events from different biological systems including patient data, CRISPR-based cell lines, and transgenic mouse models [38, 39]. Although animal models are becoming the top choice for mechanistic studies, genetic engineering usually requires an extensive amount of experimental effort and time, and the consistency of the splicing pattern between different species must be confirmed before a comparison can be made.

Functional AS events usually cause expression changes on gene or protein level. Most instances of intron retention or **poison exon** result in the introduction of premature stop codons upstream of the normal stop codon. Subsequently, there is **nonsense-mediated decay** (NMD) of the mRNA or production of a truncated protein. Thus, significant alternative splicing changes are expected to alter the expression of target genes. This information could be integrated into the identification of a shorter list of functional AS events (due to the poor overlap between AS targets and differential expressed genes). It is also expected that dysfunction of *trans*-acting splicing factors may alter the global regulatory network as a result of aberrant splicing events in key genes. Recent work [39] showed the impact of mutant SF3B1 on gene-regulatory networks by elucidating the effect of *SF3B1* mutations on post-translational regulation of multiple proteins with well-established roles in tumorigenesis.

Besides regulatory network analysis, previously curated cancer-associated gene sets can also be used to inform the functional effect of splicing events. A routine practice is to directly pool top-ranked AS target genes into functional enrichment analysis. Typically, the top terms involve splicing processes, like 'mRNA splicing', 'mRNA processing', 'translation' on the top of the output list. But there would be very few disease-relevant terms with significant *q*-values, because sometimes only one or two key splicing perturbations would be to enough to change the activity of particular pathways. Thus, how to effectively use pathway-based information in identifying cancer-associated AS events, are need to be better defined. A recent study developed a pathway enrichment-guided study of alternative splicing by correlating transcriptional signatures of cancer driver pathways with the identified AS events and established a role for MYC in regulating RNA splicing by controlling the incorporation of NMD-determinant exons in genes encoding RBP [40]. MYC is frequently altered in cancer cells and has long been recognized to have a genetic dependency on the splicing machinery [6, 41]. Targeting the spliceosome is a therapeutic vulnerability in MYC-driven cancers [42]. Interestingly, one recent study found that besides being a splicing regulator, MYC is also regulated by splicing errors in SF3B1-mutant cells [39].

In summary, identification of cancer-associated mis-splicing effects involves rigorous quality control of the raw AS calls to filter technical artifacts, cross validation using independent datasets or biological systems, and integration of alternative transcriptomic information, such as changes in regulatory network activity and dysregulated signaling pathways (Figure 1B).

Computational challenges in cancer splicing

In the next four sub-sections, we discuss some interesting and challenging topics in cancer splicing, that can be addressed through computational approaches (Figure 2A–D).

Pan-Cancer Splicing Analysis

In the year 2012, TCGA launched a pan-cancer analysis project to compare to examine the similarities and differences among the genomic and cellular alterations across 12 tumor types [43, 44]. Investigating alternative splicing in a pan-cancer cohort is a standard computationally-driven task in disclosing commonly-shared and lineage-independent splicing landscapes (Figure 2A). A different analysis characterized alternative splicing across 32 TCGA cancer types from 8,705 patients and identified increased neojunctions in tumors versus normal tissue, and *trans*-acting variants associated with AS events [28]. Another pan-cancer study reported a high frequency of common somatic alterations in splicing factor genes, suggesting that altered splicing may represent an underappreciated hallmark of tumorigenesis [14]. However, there are still many fundamental questions need to be better elucidated (see Outstanding Questions). For instance, given the low overlap of splicing defects and mutually exclusive pattern of key spliceosomal mutations, what are the convergent effects of such mutations in a single tumor type (e.g. MDS-RARS) or across distinct histological cancer types. Recurrent spliceosomal mutations only happen in some specific tumor types, while very rare in others. So, is there and what is the common process across the diverse tumor types with frequent mutations in splicing factors? Given the high number of proteins and genes involved in the splicing process, why are only a small subset of splicing factors (SF3B1, SRSF2, U2AF1 and ZRSR2) found recurrently mutated in cancer? Some genes, e.g. *SF3B1*, show different hotspot mutations in different cancers, for instance, the K700 amino acid is frequently mutated in chronic lymphocytic leukemia, but the 625 amino acid is frequently mutated in uveal melanoma; why are there cell type specific mutations, and what are their functions?

One benefit of pan-cancer analyses is that they can increase the statistical power to identify very rare mutations associated with specific splicing effects. For instance, one recent study utilized an unbiased pan-cancer analysis to identify mutations in another spliceosomal gene, SURP and G-patch domain containing 1 (*SUGP1*) that recapitulate the usage of cryptic 3′ss known to be found in mutant SF3B1 expressing cells [37]. This also recapitulates previous biochemical studies indicating that the loss of SF3B1 interaction with SUGP1 mimics the effects of *SF3B1* mutations on splicing [45]. This work on *SUGP1* was also supported by a recent study from another research group [46]. Such computational strategies could be applied to many other recurrent spliceosomal mutations in human cancer.

Deep Learning-based Splicing Analysis

By taking advantage of an ever-increasing amount of available genomics data, deep learning techniques have been proposed to enhance the characterization of molecular alterations improving the state-of-the-art performance for many genomics tasks, including alternative splicing analysis [47]. Improvements are quickly coming from new input data and better refinement of the biological questions, making these models increasingly accurate (Figure 2C). Recent work in this area includes deep neural network studies of alternative splicing using *cis*-sequence information [48, 49]. The mRNA expression levels of *trans* RBPs have also been incorporated as useful features to achieve a better characterization of alternative splicing in low expression target genes or when analyzing RNA-seq data with modest coverage [50]. In this study, the RBP expression profiles are from knocking-down experiments by the ENCODE consortium. However, most recurrent spliceosomal mutations in cancer are results in change-of-function, rather than lost-of-function. Therefore, a better fitting of such model to specific cancers, is to enlarge the training set by adding available datasets with change-of-function mutations in splicing factors. Most deep learning application rely on black-box frameworks and involve multiple layers of non-linear combination of raw inputs. This, as in other deep learning applications, hinders interpretability, with little or no information provided on the splicing machinery alterations associated to changes in AS. The development of interpretable deep networks will be paramount in the discovery of causal links between actions and effects in cancer splicing [51].

Modelling the effect of epigenetic features on AS

It has been widely accepted that epigenetic modifications regulate alternative splicing by either influencing the transcription elongation rate of RNA polymerase II (Figure 2B) or direct interactions with proteins that mark exon-intron junction of pre-mRNA [52]. Genome-wide mapping has revealed enrichment of histone modifications (for instance, H3K36me3) on exons relative to introns, which have been implicated in the regulation of alternative splicing [53] (Figure 2B). On the other hand, spliceosomal proteins can likewise influence chromatin structure and histone modifications, which imply a complex feedback loop of regulation [54]. Interestingly, one recent study identified frequent overlap of mutations in *IDH2* and *SRSF2* in human AML that together promote aberrant splicing and increased

DNA methylation of reduced expression of INTS3, which contributes to leukemogenesis [55]. Besides modification of DNA, RNA modifications have also been found to regulate AS. For example, perturbation of the dynamic status of **N6-methyladenosine (m6A) modification** could affect the interaction with SR proteins that may be involved in modulating AS [56]. However, integrating genome-wide epigenetic data with AS modelling to get a regulatory landscape between epigenetics, splicing, and cancer remains a computational challenge. Selecting appropriate datasets and methodologies (for instance, deep learning-based methods discussed above) will provide a means to model the effects of epigenetics on splicing [57].

Calculating alternative splicing derived neoantigens

Finally, after accurate depiction and understanding of alternative splicing in cancer biology, the last important task is to develop pharmacological modulation of splicing for therapeutic strategy (Figure 2D). Direct disruption of splicing efficiency increased sensitivity of cancer cells with spliceosomal mutations *in vivo*, however, some patients unfortunately exhibited unexpected side effects [58]. Immunotherapies have improved objective responses in many tumors with high burden of protein changes [59]. T cell recognition of cancers relies upon presentation of tumor-specific antigens generated by nonsynonymous mutations by MHC molecules [60] (Figure 2D). Interestingly, a recent study suggested that tumor-specific alternative splicing events are far more abundant than somatic single-nucleotide variants [28]. A recent publication presented a computational approach to identifying neoepitopes derived from intron retention events in tumor transcriptomes, which was confirmed by mass spectrometry presented on MHC class I [27]. After the identification of high-confidence AS events, genome annotations were used to extract intronic nucleotide sequences and open reading frame orientation, and sample-specific HLA alleles were computed and examined for putative peptide-MHC I binding affinity (e.g. POLYSOLVER [61]). Expanding such analysis to full types (besides intron retention) of AS events are of great interest in able to be coupled as downstream optional analysis for current available AS detection tools. Such method will be of particular interest in tumors with functional splicing changes (MDS, uveal melanoma, etc.). These observations also suggest a potential approach to activate host anti-tumor immune response by coupling spliceosome inhibition (to increase the immunogenicity) with immunotherapies. Nonetheless, experimental validation of the immunogenicity of such splicing-derived neoantigens will need to be seriously assessed.

Concluding Remarks

RNA splicing is a critical mediator of gene expression and regulator of proteome diversity. Alterations in splicing, including common change-of-function mutations in spliceosome genes, have been suggested to promote tumorigenesis. By utilizing quantitative cancer biology analyses, a number of computational methods have been developed and proven to play an important role in systematically identifying high-confidence AS events in a context-dependent manner. However, more efforts will be required to customize downstream computational analysis to decode the mechanistic consequences of splicing alterations in cancer pathogenesis. We suggest that coupled analysis of the impact of splicing dysfunction on the activity of gene-regulatory networks or cancer signaling pathways may help to discover key functional events and guide further experimental studies.

In this review, we have highlighted several research directions in cancer splicing related to or driven by computational analysis. First, we underscored that advances in cancer genomics projects (e.g. TCGA) have enabled high resolution detection and comparison of AS across a wide-range of tissues in a pan-cancer manner. Second, we suggested the importance of incorporating epigenetic features into AS analyses. The large availability of data is enabling the development and application of *in-silico* approaches in artificial intelligence science to increase the sensitivity and specificity of AS detection. Lastly, we suggested that the characterization of potential splicing-derived neoantigens may be leveraged with recent advances in immunotherapy to open new therapeutic avenues for AS-related tumors.

Acknowledgments

We thank Karen Gomez for her helpful suggestions in writing. Z. Liu is supported by the CAS Hundred Talents Program. R. Rabadan is supported by NIH grants (R01 CA185486, R01 CA179044, U54, CA193313, and U54 209997) and the NSF/SU2C/V-Foundation Ideas Lab Multidisciplinary Team (PHY-1545805). R. Rabadan is a member of the AimedBio SAB and a founder of Genotwin. None of these activities are related to this publication.

Glossary

Branchpoint

An important sequence (adenine) for mRNA splicing, which often located between 18 to 40 nucleotides upstream from the 3′ splice site. During splicing process, the branchpoint attack on the 5′ splice site to form the intron lariat.

N6-methyladenosine (m6A) modification

The addition of a methyl group at position N6 of adenosine, which is the most common mRNA modification in mammalian cells.

Nanopore or PacBio sequencing

Two widely-used long-read sequencing providers. Long-read sequencing, also called third-generation sequencing, have the capability to produce substantially long sequences of DNA. The common sequencing length is between 10,000 and 100,000 base pairs.

Nonsense-mediated decay

An mRNA quality-control mechanism that selectively degrades mRNA transcripts with premature termination codons.

Poison exon

One type of cassette exon, which contains a premature termination codon. When being included, this exon will lead to premature truncation of the transcript.

Polypyrimidine tract

One important *cis*-acting sequence elements in pre-mRNA splicing. Polypyrimidine tract is rich with pyrimidine nucleotides, and is usually 15–20 base pairs long, located about 5–40 base pairs before the 3’ splice site. The polypyrimidine tract can affect 3’ splicing by selecting alternative branchpoint.

Spliceosome

A large and complex molecular which is assembled by small nuclear RNAs and protein complexes. Spliceosome play key functions in the removal of introns from pre-mRNAs.

Reference

1. Nilsen TW and Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463 [PubMed: 20110989]
2. Pan Q, et al. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413–1415 [PubMed: 18978789]
3. Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476 [PubMed: 18978772]
4. Zhang J and Manley JL (2013) Misregulation of pre-mRNA alternative splicing in cancer. Cancer Discov 3, 1228–1237 [PubMed: 24145039]
5. David CJ and Manley JL (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev 24, 2343–2364 [PubMed: 21041405]
6. David CJ, et al. (2010) HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature 463, 364–368 [PubMed: 20010808]
7. Babic I, et al. (2013) EGFR mutation-induced alternative splicing of Max contributes to growth of glycolytic tumors in brain cancer. Cell Metab 17, 1000–1008 [PubMed: 23707073]
8. Vos MC, et al. (2016) MMP-14 and CD44 in Epithelial-to-Mesenchymal Transition (EMT) in ovarian cancer. J Ovarian Res 9, 53 [PubMed: 27590006]
9. Brown RL, et al. (2011) CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. The Journal of clinical investigation 121, 1064–1074 [PubMed: 21393860]
10. Poulikakos PI, et al. (2011) RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). Nature 480, 387–U144 [PubMed: 22113612]
11. Antonarakis ES, et al. (2014) AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. N Engl J Med 371, 1028–1038 [PubMed: 25184630]
12. Supek F, et al. (2014) Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell 156, 1324–1335 [PubMed: 24630730]
13. Anczukow O and Krainer AR (2016) Splicing-factor alterations in cancers. Rna 22, 1285–1301 [PubMed: 27530828]
14. Seiler M, et al. (2018) Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell reports 23, 282–296. e284 [PubMed: 29617667]
15. Wang J, et al. (2015) A survey of computational methods in transcriptome-wide alternative splicing analysis. Biomol Concepts 6, 59–66 [PubMed: 25719337]
16. Mehmood A, et al. (2019) Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform
17. Carazo F, et al. (2019) Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors. Brief Bioinform 20, 1358–1375 [PubMed: 29390045]
18. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515 [PubMed: 20436464]
19. Hu Y, et al. (2013) DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res 41, e39 [PubMed: 23155066]
20. Katz Y, et al. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7, 1009–1015 [PubMed: 21057496]
21. Wang QQ and Rio DC (2018) JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. P Natl Acad Sci USA 115, Eb181–Eb190

22. Shen S, et al. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A 111, E5593–5601 [PubMed: 25480548]
23. Vaquero-Garcia J, et al. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. elife 5, e11752 [PubMed: 26829591]
24. Brooks AN, et al. (2011) Conservation of an RNA regulatory map between Drosophila and mammals. Genome research 21, 193–202 [PubMed: 20921232]
25. DeBoever C, et al. (2015) Transcriptome sequencing reveals potential mechanism of cryptic 3'splice site selection in SF3B1-mutated cancers. PLoS Comput Biol 11, e1004105 [PubMed: 25768983]
26. Zhang D, et al. (2020) Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. Nature Communications 11, 1–19
27. Smart AC, et al. (2018) Intron retention is a source of neoepitopes in cancer. Nature biotechnology 36, 1056–1058
28. Kahles A, et al. (2018) Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell 34, 211–224 e216 [PubMed: 30078747]
29. Danan-Gotthold M, et al. (2015) Identification of recurrent regulated alternative splicing events across human solid tumors. Nucleic Acids Res 43, 5130–5144 [PubMed: 25908786]
30. Liu R, et al. (2014) Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. BMC bioinformatics 15, 364 [PubMed: 25511303]
31. Chhangawala S, et al. (2015) The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome biology 16, 131 [PubMed: 26100517]
32. Amarasinghe SL, et al. (2020) Opportunities and challenges in long-read sequencing data analysis. Genome biology 21, 1–16
33. Sahlin K, et al. (2018) Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. Nature communications 9, 1–12
34. Tardaguila M, et al. (2018) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome research 28, 396–411
35. Tang AD, et al. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nature communications 11, 1–12
36. Au KF, et al. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci U S A 110, E4821–4830 [PubMed: 24282307]
37. Liu ZQ, et al. (2020) Pan-cancer analysis identifies mutations in SUGP1 that recapitulate mutant SF3B1 splicing dysregulation. P Natl Acad Sci USA 117, 10305–10312
38. Liu B, et al. (2020) Mutant SF3B1 promotes AKT and NF-kB driven mammary tumorigenesis. J Clin Invest
39. Liu ZQ, et al. (2020) Mutations in the RNA Splicing Factor SF3B1 Promote Tumorigenesis through MYC Stabilization. Cancer Discovery 10, 806–821 [PubMed: 32188705]
40. Phillips JW, et al. (2020) Pathway-guided analysis identifies Myc-dependent alternative pre-mRNA splicing in aggressive prostate cancers. Proc Natl Acad Sci U S A 117, 5269–5279 [PubMed: 32086391]
41. Koh CM, et al. (2015) MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. Nature 523, 96–+ [PubMed: 25970242]
42. Hsu TYT, et al. (2015) The spliceosome is a therapeutic vulnerability in MYC-driven cancer. Nature 525, 384–+ [PubMed: 26331541]
43. Cancer Genome Atlas Research, N., et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120 [PubMed: 24071849]
44. Liu Z and Zhang S (2014) Toward a systematic understanding of cancers: a survey of the pan-cancer study. Front Genet 5, 194 [PubMed: 25071824]
45. Zhang J, et al. (2019) Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. Mol Cell 76, 82–95 e87 [PubMed: 31474574]

46. Alsafadi S, et al. (2020) Genetic alterations of SUGP1 mimic mutant-SF3B1 splice pattern in lung adenocarcinoma and other cancers. Oncogene, 1–12

47. LeCun Y, et al. (2015) Deep learning. Nature 521, 436–444 [PubMed: 26017442]

48. Jaganathan K, et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535–548 e524 [PubMed: 30661751]

49. Louadi Z, et al. (2019) Deep Splicing Code: Classifying Alternative Splicing Events Using Deep Learning. Genes 10, 587

50. Zhang Z, et al. (2019) Deep-learning augmented RNA-seq analysis of transcript splicing. Nat Methods 16, 307–310 [PubMed: 30923373]

51. Samek W, et al. (2017) Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296

52. Laurent L, et al. (2010) Dynamic changes in the human methylome during differentiation. Genome Res 20, 320–331 [PubMed: 20133333]

53. Spies N, et al. (2009) Biased chromatin signatures around polyadenylation sites and exons. Mol Cell 36, 245–254 [PubMed: 19854133]

54. De Almeida SF and Carmo-Fonseca M (2012) Design principles of interconnections between chromatin and pre-mRNA splicing. Trends in biochemical sciences 37, 248–253 [PubMed: 22398209]

55. Yoshimi A, et al. (2019) Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. Nature 574, 273–277 [PubMed: 31578525]

56. Yang Y, et al. (2015) Dynamic m 6 A modification and its emerging regulatory role in mRNA splicing. Science Bulletin 60, 21–32

57. Pacini C and Koziol MJ (2018) Bioinformatics challenges and perspectives when studying the effect of epigenetic modifications on alternative splicing. Philosophical Transactions of the Royal Society B: Biological Sciences 373, 20170073

58. Jamieson D, et al. (2016) A phase I pharmacokinetic and pharmacodynamic study of the oral mitogen-activated protein kinase kinase (MEK) inhibitor, WX-554, in patients with advanced solid tumours. Eur J Cancer 68, 1–10 [PubMed: 27693888]

59. Rizvi NA, et al. (2015) Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science 348, 124–128 [PubMed: 25765070]

60. Schreiber RD, et al. (2011) Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science 331, 1565–1570 [PubMed: 21436444]

61. Shukla SA, et al. (2015) Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 33, 1152–1158 [PubMed: 26372948]

62. Wahl MC, et al. (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136, 701–718 [PubMed: 19239890]

63. Turunen JJ, et al. (2013) The significant other: splicing by the minor spliceosome. Wiley Interdiscip Rev RNA 4, 61–76 [PubMed: 23074130]

64. Matera AG and Wang Z (2014) A day in the life of the spliceosome. Nat Rev Mol Cell Biol 15, 108–121 [PubMed: 24452469]

65. Wang Z and Burge CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14, 802–813 [PubMed: 18369186]

66. Graveley BR, et al. (1998) A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. The EMBO journal 17, 6747–6756 [PubMed: 9822617]

67. Zhou Z and Fu XD (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma 122, 191–207 [PubMed: 23525660]

68. Krecic AM and Swanson MS (1999) hnRNP complexes: composition, structure, and function. Curr Opin Cell Biol 11, 363–371 [PubMed: 10395553]

69. Hentze MW, et al. (2018) A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol 19, 327–341 [PubMed: 29339797]

70. Zahler AM, et al. (1992) SR proteins: a conserved family of pre-mRNA splicing factors. Genes Dev 6, 837–847 [PubMed: 1577277]

71. Barash Y, et al. (2010) Deciphering the splicing code. Nature 465, 53–59 [PubMed: 20445623]

72. Yoshida K, et al. (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. Nature 478, 64–69 [PubMed: 21909114]

73. Graubert TA, et al. (2012) Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nature Genetics 44, 53–U77

74. Papaemmanuil E, et al. (2011) Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med 365, 1384–1395 [PubMed: 21995386]

75. Quesada V, et al. (2012) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nature Genetics 44, 47–52

76. Harbour JW, et al. (2013) Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. Nature Genetics 45, 133–135 [PubMed: 23313955]

77. Imielinski M, et al. (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 150, 1107–1120 [PubMed: 22980975]

78. Maguire SL, et al. (2015) SF3B1 mutations constitute a novel therapeutic target in breast cancer. J Pathol 235, 571–580 [PubMed: 25424858]

79. Tyner JW, et al. (2018) Functional genomic landscape of acute myeloid leukaemia. Nature 562, 526–531 [PubMed: 30333627]

80. Cherry S and Lynch KW (2020) Alternative splicing and cancer: insights, opportunities, and challenges from an expanding view of the transcriptome. Genes Dev 34, 1005–1016 [PubMed: 32747477]

81. Dvinge H, et al. (2016) RNA splicing factors as oncoproteins and tumour suppressors. Nature Reviews Cancer 16, 413 [PubMed: 27282250]

82. Darman RB, et al. (2015) Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. Cell reports 13, 1033–1045 [PubMed: 26565915]

83. Manley JL, et al. (2020) SF3B1 mutant-induced missplicing of MAP3K7 causes anemia in myelodysplastic syndromes. bioRxiv

84. Inoue D, et al. (2019) Spliceosomal disruption of the non-canonical BAF complex in cancer. Nature 574, 432–436 [PubMed: 31597964]

85. Wang L, et al. (2016) Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. Cancer Cell 30, 750–763 [PubMed: 27818134]

86. Dolatshad H, et al. (2016) Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes. Leukemia 30, 2322–2331 [PubMed: 27211273]

87. Arzalluz-Luque Á and Conesa A (2018) Single-cell RNAseq for the study of isoforms—how is that possible? Genome biology 19, 110 [PubMed: 30097058]

88. Chen G, et al. (2019) Single-cell RNA-seq technologies and related computational data analysis. Frontiers in genetics 10, 317 [PubMed: 31024627]

89. Wen WX, et al. (2020) Technological advances and computational approaches for alternative splicing analysis in single cells. Computational and structural biotechnology journal 18, 332–343 [PubMed: 32099593]

90. Velten L, et al. (2015) Single-cell polyadenylation site mapping reveals 3′ isoform choice variability. Molecular systems biology 11, 812 [PubMed: 26040288]

91. Karlsson K, et al. (2017) Alternative TSSs are co-regulated in single cells in the mouse brain. Molecular systems biology 13, 930 [PubMed: 28495919]

92. Ramsköld D, et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature biotechnology 30, 777–782

93. Welch JD, et al. (2016) Robust detection of alternative splicing in a population of single cells. Nucleic acids research 44, e73–e73 [PubMed: 26740580]

94. Huang Y and Sanguinetti G (2017) BRIE: transcriptome-wide splicing quantification in single cells. Genome biology 18, 1–11 [PubMed: 28077169]

95. Song Y, et al. (2017) Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Molecular cell 67, 148–161. e145 [PubMed: 28673540]

96. Vu TN, et al. (2018) Isoform-level gene expression patterns in single-cell RNA-sequencing data. Bioinformatics 34, 2392–2400 [PubMed: 29490015]

Box 1.

Mechanism of mRNA splicing

Splicing is a co-transcriptional process in which non-coding introns are removed and adjacent exons are joined together to form a single mRNA strand. This process is orchestrated by the large macromolecular complex known as **spliceosome** (see Glossary), which recognizes major introns through specific sequence motifs at the exon-intron boundaries (splice sites), **branchpoint** sites, and the **polypyrimidine tract** upstream of the AG splice site [62] (Figure I). Interestingly, one recent study demonstrated that pyrimidines downstream the AG site may also play a role in the aberrant branchpoint recognition [38]. Small nuclear ribonucleoproteins (snRNPs) are complexes of RNA and proteins that interact with the pre-mRNA, mediating the splicing process [62]. The major U2-type spliceosome consists of five snRNP complexes, U1, U2, U4, U5, and U6, which get dynamically altered composition and structure during the splicing process [62]. In addition to the major spliceosome, there is a functionally analogous ribonucleoprotein complex, the minor spliceosome, that catalyzes the splicing of U12-type introns [63].

The entire splicing process may be simplified as a two-step transesterification reaction. Initially, the U1 snRNP binds the 5′ splice site of the intron. The U2 auxiliary factor (U2AF) complex binds to the 3′ of the intron and recruits the U2 snRNP, which binds to the branchpoint and replaces splicing factor 1 (SF1). After recruitment of U5.U4/U6 tri-snRNP, the first transesterification reaction occurs and forms an intron lariat at the 3′ part of the exon. Next, the second transesterification reaction at the 3′ splice site releases the 3′ exon, which leads to exon ligation and excision of the intron lariat [64].

Besides the spliceosome complex, additional action of *trans*-acting RNA-binding proteins (RBPs) is required to further regulate the splicing process [65]. RBPs bind to sequence motifs located in exonic or intronic regions that can mediate the promotion or repression of particular splicing products [66]. The classic *trans*-acting RBPs are the serine/arginine-rich (SR) proteins, heterogenous nuclear ribonucleoproteins (hnRNPs), and other hnRNP-like proteins [67–69] (Figure I). SR proteins are characterized by containing one or two copies of an RNA recognition motif (RRM) domain as well as a domain rich in serine-arginine (SR) dipeptides that make protein-protein interactions [70]. In contrast, hnRNPs are more structurally diverse, containing different types of RNA-binding domains and relatively unstructured domains that likely contribute to protein-protein interactions [68]. Overall, there are over hundreds of proteins with potential roles in splicing regulation [71], and somatic alternations or dysregulation of these RBPs could potentially be linked to human diseases, such as cancer.

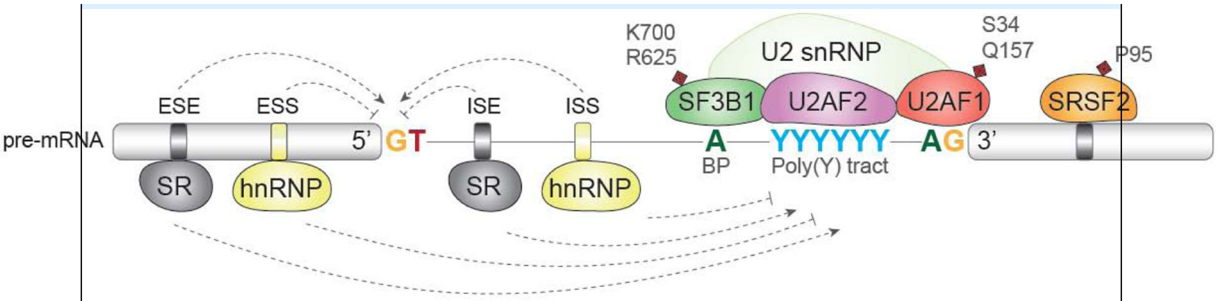

Figure I in Box 1. Splicing regulation mechanism (major introns).

Key sequence features that govern splicing include consensus sequences of the 5′ and 3′ splice sites, branchpoint (BP) sites, and the polypyrimidine tract upstream of the AG splice site. At the beginning of splicing, U2 auxiliary factor (U2AF) complex binds to the 3′ of the intron and recruits the U2 snRNP to interact with the branchpoint. SF3B1 is a key component of the U2 snRNP that makes direct contact with the substrate. Alternative splicing is regulated by trans-acting splicing factors including SR proteins, and hnRNPs. Enhancer auxiliary elements are denoted in black for exonic (ESE) or intronic (ISE) splicing enhancers. Silencer auxiliary elements are denoted in yellow for exonic (ESS) or intronic (ISS) splicing silencers. The most commonly mutated splicing factors in human cancer include *SF3B1* (hotspot: K700 and R625), *U2AF1* (S34 and Q157) and *SRSF2* (P95).

Box 2.

Commonly mutated splicing factors in human cancer

Several splicing factors are recurrently mutated in cancer, including splicing factor 3B, subunit 1 (*SF3B1*), serine/arginine-rich splicing factor 2 (*SRSF2*), U2 small nuclear RNA auxiliary factor 1 (*U2AF1*) and zinc finger, RNA-binding motif and serine/arginine-rich 2 (*ZRSR2*) [72–74]. These spliceosomal mutations have been discovered across a broad range of tumor types, including myelodysplastic syndromes (MDS) [72], chronic lymphocytic leukemia (CLL) [75], uveal melanoma (UVM) [76], lung adenocarcinoma (LUAD) [77], breast invasive carcinoma (BRCA) [78] and others [79]. We briefly recapitulate the latest findings of the most commonly mutated splicing factor, SF3B1, in human cancers (for other splicing factors, see recent comprehensive reviews [80, 81]).

SF3B1 frequently contains heterozygous mutations at very specific residues (known as “hotspots”). These mutations promote the usage of upstream branchpoints during the splicing reaction, resulting in the use of cryptic upstream 3′ splice sites (3′ss) [25, 82]. Recent work has uncovered 3’ splicing patterns specific to *SF3B1* mutational hotspots. *SF3B1*K700E mutations alter splicing of a specific subunit (PPP2R5A) of the PP2A serine/threonine phosphatase complex to confer post-translational MYC and BCL2 activation, which is therapeutically intervenable using an FDA-approved drug [39]. In addition to *PPP2R5A*, other key mis-spliced targets have also been extensively studied and proven to promote tumorigenesis or contribute to disease phenotype. A universally mis-spliced gene in *SF3B1*-mutated tumors, *MAP3K7*, was recently associated with activation of the NF-kB pathway in mammary epithelial tumors with *SF3B1* mutations [38]. *MAP3K7* mis-splicing has also been associated with accelerated erythroid differentiation and apoptosis, potentially explaining the origin of anemia in MDS patients harboring *SF3B1* mutations [83]. Mis-splicing and subsequent loss of *BRD9*, a noncanonical BAF complex subunit, led to enhanced tumor growth and transformation in tumors harboring *SF3B1* mutations [84]. Other interesting mis-spliced target genes of *SF3B1* mutations include *DLV2*, which modulates Notch signaling in CLL [85], and *ABCB7*, which is associated with the increased mitochondrial iron accumulation found in MDS patients [86]. These studies and others demonstrate the role of *SF3B1* as an oncogenic driver implementing tumorigenesis through diverse cellular processes.

Box 3.

Single-cell splicing studies

Alternative splicing events detected from bulk RNASeq are mixed signals averaged over cell populations, that have limited power to delineate the splicing heterogeneity in different tumor clones. In contrast, splicing analysis on single-cell resolution draw huge interests recently by demonstrating the power to unravel isoform expression dynamics in different cellular types [87–89]. Moreover, identification of tumor heterogeneity driven by distinct single-cell splicing events may guide the development of targeted treatment [89]. Limited by the coverage biases on 3'/5' sites of UMI (unique molecular identifier) type sequencing, early single-cell splicing studies focus on the detection of the alternative polyadenylation sites or transcription start sites [90, 91]. More recently, the development of full-length capture techniques, e.g. Smart-seq [92] and single-molecule sequencing, significantly improved reads coverage across entire transcript, showing great advantages in the detection of alternative splicing and isoform usage on single-cell level.

Currently, a few computational approaches for single-cell splicing have been developed, for instance, SingleSplice [93], BRIE [94], and Expedition [95]. Specifically, SingleSplice defines a concept of ‘alternative splicing modules’ and utilizes a statistical model to detect local isoform usage, rather than full-length transcript [93]. BRIE incorporates a Bayesian regression module for differential isoform quantification, by finding a balance between the sequencing depth/quality and an informative prior distribution trained from GENCODE database [94]. In contrast, Expedition suite only uses well aligned-reads to define a custom alternative splicing index, and have advantage in describing a distribution of exon inclusion in a population of single cells [95]. By applying these tools and others, recent studies have successfully uncovered significant isoform switching events on single-cell resolution in human cancers, which is invisible from standard gene expression analysis [92, 96]. However, study of cancer splicing on single-cell level is still in its infancy. Currently, rather than urgent requirement of more elegant methodological design, major barriers of this area of research come from the technical limitations of single-cell library preparation and sequencing protocols, for instance, low reads coverage, high dropout rate, high sequencing errors and inevitable technical noise [87–89].

Outstanding Questions

How can splicing events functionally relevant to cancer be extracted from the long lists of AS generated by standard computational tools?

Given the low overlap of splicing defects and mutually exclusive pattern of key spliceosomal mutations, what are the convergent effects of these mutations in a single tumor type (e.g. MDS-RARS) or across distinct tumor types?

Why do recurrent spliceosomal mutations only occur in particular tumor types (MDS, uveal melanoma, etc). What are the common characteristics of these tumors?

Why are only a small subset of splicing factors (SF3B1, SRSF2, U2AF1 and ZRSR2) mutated among the hundreds of proteins involved in splicing?

What is the reason for the selection of specific mutational hotspots in a specific tumor type (e.g. *SF3B1* K700 in CLL, R625 in UVM), and why are the hotspots different in other tumors?

What is the full spectrum of genetic/transcriptomic changes that could generate the same splicing pattern induced by well-studied spliceosomal mutations?

How to develop interpretable deep learning-based approaches to discover causal links between actions and effects in cancer splicing?

How can genome-wide epigenetic features be integrated with AS modelling to obtain the landscape of epigenetic effects in splicing and cancer? What are the most appropriate datasets and methodologies?

How can the methodology used for bulk RNA-sequencing be extended to single-cell RNA-sequencing (Box 3)?

Highlights

Accumulating evidence indicates that recurrent spliceosomal mutations contribute to the initiation and progression of several cancers through diverse fundamental cellular processes.

A number of computational tools are used to characterize splicing effects in cancer. These tools present limitations that can be overcome by running alternative splicing analysis with multiple tools and integrating the results.

Extracting splicing events functionally relevant to cancer requires rigorous quality control to filter technical artifacts, cross validate the events using independent datasets, and integrate alternative approaches including regulatory network characterization and cancer signaling pathway analyses.

By taking advantage of the increasing amount of genomic data, deep learning-based methods have dramatically improved the state-of-the-art performance of alternative splicing analysis.

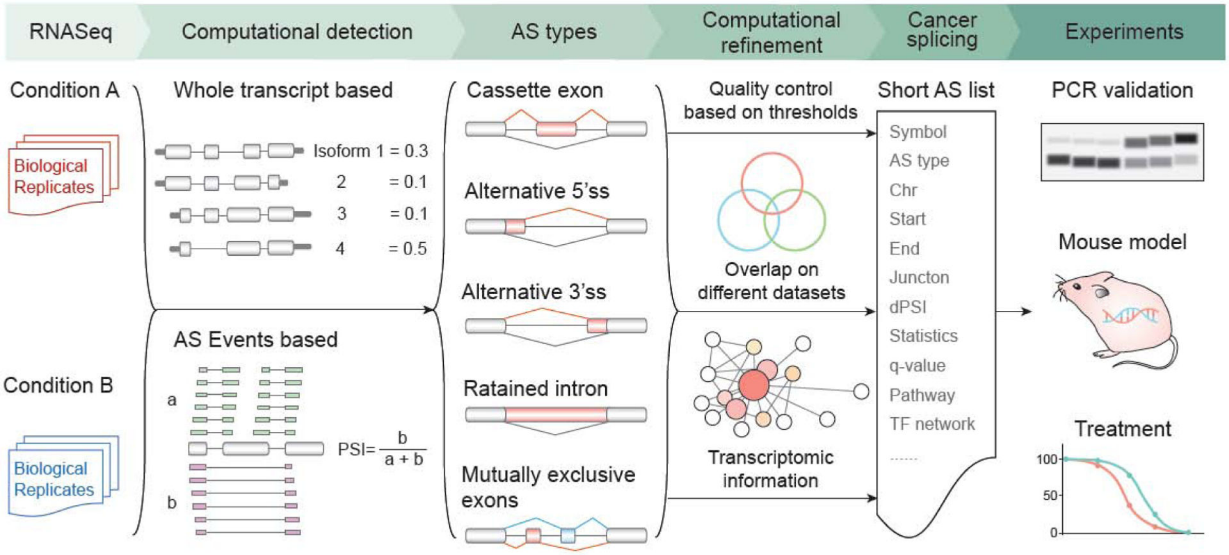

Figure 1. General summary of computational workflow for alternative splicing analysis.

A computational pipeline for the study of alternative splicing events includes: 1) Computational detection of AS events from RNASeq data in a context-dependent manner. There are two main categories of methods: whole transcript-based or events-based. 2) Computational refinement to identify cancer-associated mis-splicing effects for downstream experiments. This step involves rigorous quality control on the raw AS calls to filter technical artifacts, cross validation using independent datasets or biological systems, and integration of alternative transcriptomic information, such as changes in regulatory network activity and dysregulated signaling pathways.

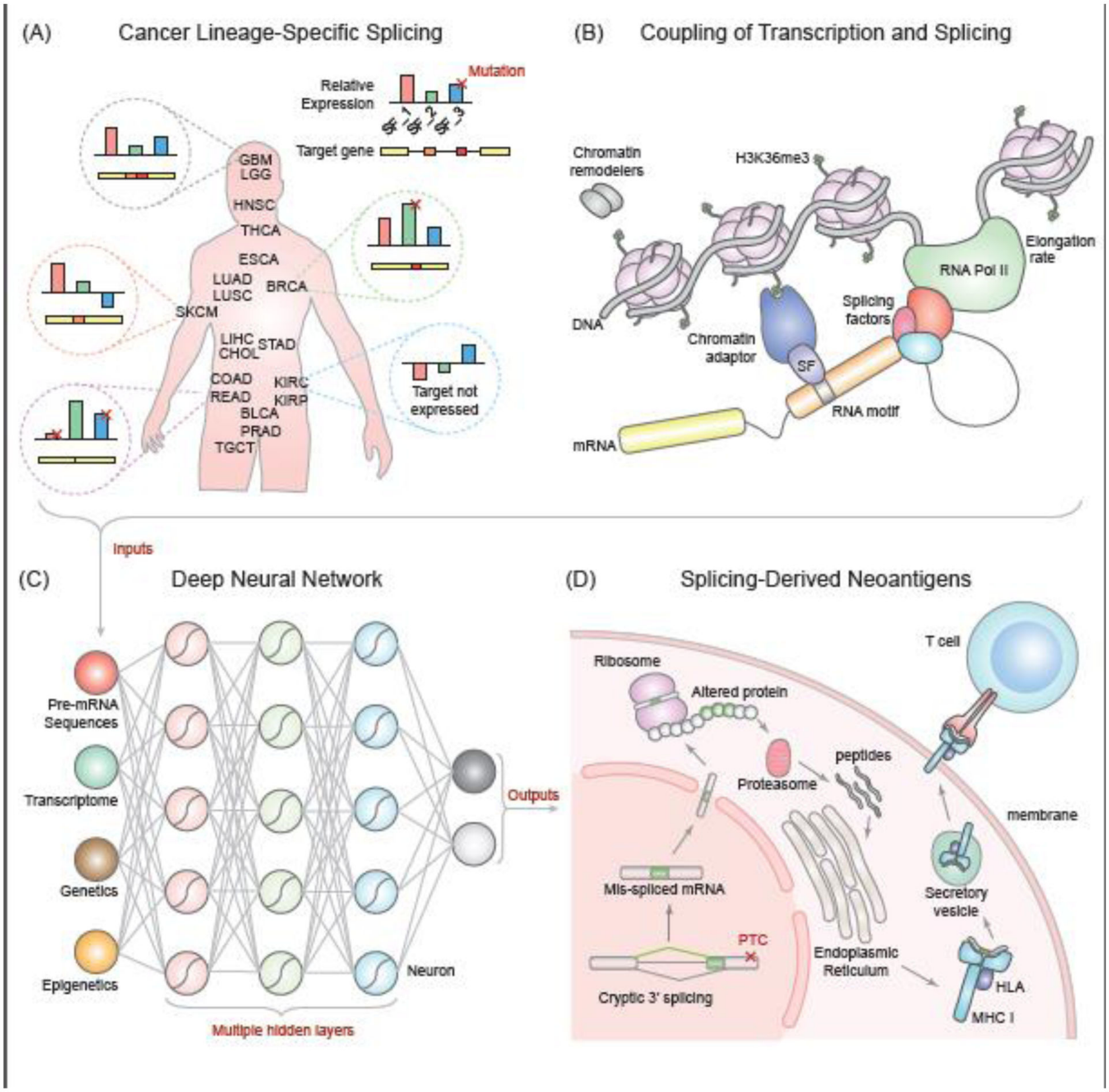

Figure 2. Current computational challenges in the characterization of the role of alternative splicing in cancer.

(A) Pan-cancer analysis of alternative splicing to uncover commonly-shared and lineage-independent splicing landscapes. SF: Splicing factor. (B) Epigenetic modifications regulate alternative splicing by influencing the transcription elongation rate of RNA polymerase II or binding an adaptor protein that reads specific histone marks and in turn recruits splicing factors. Trimethylated histone 3 lysine 36 (H3K36me3) attracts the chromatin-binding factor MRG15 that acts as an adaptor protein and by protein-protein interaction helps to recruit splicing factors. (C) Deep learning techniques have been proposed to improve the state-of-the-art performance for alternative splicing analysis. (D) Characterization of potential splicing-derived neoantigens may be leveraged with recent advances in immunotherapy to open new therapeutic avenues for AS-related tumors. Part of aberrantly spliced transcripts are translated into protein, which is processed into short residue peptides by the proteasome and then shuttled into the endoplasmic reticulum via the transporter associated with antigen processing. Finally, splicing-derived peptides are loaded onto major histocompatibility complex class I (MHC I), and the peptide–MHC complexes can be potentially recognized by T cells. PTC: premature termination codon.