Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing

Wei Jiang, Liang Chen

Quantitative and Computational Biology, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, United States

ABSTRACT

Alternative splicing contributes to the majority of protein diversity in higher eukaryotes by allowing one gene to generate multiple distinct protein isoforms. It adds another regulation layer of gene expression. Up to 95% of human multi-exon genes undergo alternative splicing to encode proteins with different functions. Moreover, around 15% of human hereditary diseases and cancers are associated with alternative splicing. Regulation of alternative splicing is attributed to a set of delicate machineries interacting with each other in aid of important biological processes such as cell development and differentiation. Given the importance of alternative splicing events, their accurate mapping and quantification are paramount for downstream analysis, especially for associating disease with alternative splicing. However, deriving accurate isoform expression from high-throughput RNA-seq data remains a challenging task. In this mini-review, we aim to illustrate I) mechanisms and regulation of alternative splicing, II) alternative splicing associated human disease, III) computational tools for the quantification of isoforms and alternative splicing from RNA-seq.

© 2020 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Upon the completion of the Human Genome Project in 2003, the discrepancy between the number of annotated protein-coding genes and the number of observed human polypeptides reveals the widespread violation of the "one gene–one polypeptide" hypothesis. It is now commonly accepted that in higher eukaryotes, alternative splicing plays a remarkably important role in increasing protein diversity by allowing one gene to generate distinct protein isoforms and increasing the complexity of gene expression regulation [1]. In humans, up to 95% of multi-exon genes undergo alternative splicing to encode proteins with different functions in distinct cellular processes [2]. Furthermore, around 15% of human hereditary diseases and cancers are reported to be associated with alternative splicing [3,4]. The analyses and studies of alternative splicing will fundamentally advance our understanding of mRNA complexity and its regulation, provide valuable insights to understand disease etiology, and assist the development of therapeutic interventions for splicing-related diseases.

2. Mechanisms and regulation of alternative splicing

Constitutive splicing is the process that mRNA is spliced identically producing the same isoforms, while alternative splicing generates different isoforms through using different sets of exons. Typically, there are five major subtypes of alternative splicing [5] (Fig. 1). 1) Exon skipping (also known as cassette exons) is reported to be the most common alternative splicing event in mammalian cells, which results in complete skipping of one or more exons [6,7]. 2) Mutually exclusive exons are a rare subtype where two or more splicing events are no longer independent. They are executed or disabled in a coordinated manner [8]. 3) Alternative 5' splice sites (alternative donors): the usage of an alternative 5' donor site, which changes the 3' boundary of the upstream exon. 4) Alternative 3' splice sites (alternative acceptors): opposite to alternative 5' splice sites, it is the usage of an alternative 3' splice junction site causing the change of the 5' boundary of the downstream exon. 5) Intron retention (IR) is the process that one or more introns remain unspliced in the mRNA. The fate of those intron-retaining mRNA can be different [9,10]. Some of them are degraded by the nonsense-mediated decay pathway, while others could generate new protein isoforms [11]. Often, malfunctional proteins could be produced from IR and lead to diseases. In addition to the five major subtypes, alternative polyadenylation sites and alternative promoters are often discussed under this topic. Although these two also increase the coding potential of genomes, they have very different mechanisms and are not directly related to splicing. We will not discuss them further in this mini-review.

2.1. General mechanisms of pre-mRNA splicing

Pre-mRNA splicing occurs in a large ribonucleoprotein complex (RNP) known as spliceosome [12–14]. The spliceosome is a dynamic complex mainly consisting of five small nuclear ribonucleoproteins (snRNPs) (U1, U2, U4/U6, U5) that recognize and assemble on each intron to ultimately remove introns from a transcribed pre-mRNA (Fig. 2) [15–19]. During the assembly, U1 binds to the 5' splice site with the assistance of the U2 auxiliary factor protein (U2AF) forming base pairing between the U1 snRNA and the splice site. This earliest formed complex (complex E or commitment complex) then binds to U2 to form the complex A (pre spliceosome). Formation of the complex A ensures the intron to be spliced out and the last exon to be retained during the final step. The complex B (precatalytic spliceosome) is formed by the complex A joining U4, U5 and U6. Then a series of intricate reorganization events occur in order to form the complex C (catalytic step 1 spliceosome). Firstly, the U1 interaction at the donor site is replaced by the U6 snRNP, followed by U1 and U4 leaving the complex B. This lastly formed complex C catalyzes two transesterification reactions of the splicing. During the first transesterification, the phosphate at the 5' splice site is attacked by the 2'- hydroxyl group at the branch point which results in the 5' end of the intron being cleaved from the upstream exon, and then joining to the branch point by a phosphodiester bond. In the second transesterification, the phosphate at the 3' splice site of the intron is attacked by the 3'- hydroxyl of the downstream exon. This step finally releases the intron as well ligates the two exons by a phosphodiester bond [6].

2.2. Alternative splicing mechanisms

Alternative splicing is regulated by the interaction between cis-acting regulatory sequences and corresponding trans-acting regulatory proteins. There are two major types of cis-acting elements that either promote (enhancers) or inhibit (silencers) splicing activity of nearby splice sites. Splicing enhancers can be located either in the exon (exonic splicing enhancers, ESEs) or in the intron (intronic splicing enhancers, ISEs). They bind to splicing activator proteins such as serine/arginine-rich family of nuclear phosphoproteins (SR protein family) to increase the chance of an adjacent site being spliced. Splicing silencers include exonic splicing silencers (also known as cassette exons) splicing activity of nearby splice sites. Splicing silencers can be located either in the intron (intronic splicing silencers, ISSs) or in the exon (ESSs) and intronic splicing silencers harbored in the neighboring intron (ISSs). They bind to splicing repressor proteins such as heterogeneous nuclear ribonucleopro-
Fig. 1. Constitutive and five major types of alternative splicing. a: Constitutive splicing; b: exon skipping (cassette exons); c: mutually exclusive exons; d: alternative 5’ splice sites (alternative donors); e: alternative 3’ splice sites (alternative acceptors); f: intron retention.

Fig. 2. Stepwise schematic presentation of general pre-mRNA splicing. Abbreviations: BP: brunch point; SS: splice site.
teins (hnRNPs) to inhibit the splicing of nearby sites or stimulate exon skipping.

Occasionally, the presence or absence of one single regulator is adequate to alternatively splice a pre-mRNA [20–23]. But more often, more than one cis-acting regulatory sequences have to work together with splicing activators or repressors to enhance or inhibit the splicing activity at the splicing site to determine an alternative splicing pathway [10,20,24,25]. Interestingly, hnRNPs proteins do not always act as splicing repressors nor does SR protein always act as splicing activators. HnRNPs is a well-conserved RNA-binding protein family in mammals [26]. In humans, hnRNPA1 usually binds to ESS or ISS to repress exon inclusion by steric actions, or binds to both sides of a cassette exon forming a loop to skip the exon [27,28]. However, during the interaction with Fas pre-mRNA, hnRNPA1 promotes exon 6 inclusion through interrupting 5’ splice donor site selection of exon 5 [29]. Another member of the family hnRNPL also has the ability to both activate or suppress exon 5 of CD45 gene [30]. More importantly, this characteristic of splicing factors often appears to be strongly position-dependent [31]. For example, a splicing factor would serve as a splicing suppressor or an activator in different pre-mRNAs depending on whether it is bound to exons or introns. A well-studied splicing factor, neuron-specific RNA-binding protein Nova-1 is one of such splicing factors. CLIP Analysis has revealed that over 91% of Nova-dependent exon inclusion events occurred near alternative 5’ splice sites or constitutive 3’ splice sites, while 74% of Nova-dependent exon skipping occurs near constitutive 5’ splice sites [32].

Splicing factors are imperative in regulating splicing events. Being able to pinpoint the binding sites between RNA and splicing factors provides us valuable information about splicing regulation. The predominant methods for investigating RNA-protein binding sites are based on cross-linking followed by immunoprecipitation (CLIP) [22,33]. The CLIP method coupling with high-throughput sequencing permits the transcriptome-wide investigation of RNAs that interact with the protein of interest [34]. A variety of CLIP-based methods have been developed. Among them, high-throughput sequencing-CLIP (HITS-CLIP or CLIP-Seq) [34], photoactivatable-ribonucleoside enhanced CLIP (PAR-CLIP) [35], and individual CLIP (iCLIP) [36] are the three major ones. HITS-CLIP was originally applied to identify genome-wide protein-RNA interactions for the Nova protein family [32]. It is a well-established and effective method, though the false negative rate is high as a result of low cross-linking efficiency [34]. Comparing to HITS-CLIP, PAR-CLIP has improved the efficiency due to the incorporation of photoactivatable ribonucleoside analogs. This protocol has effectively increased the resolution and signal-to-noise ratio [35,37]. The potential drawback of PAR-CLIP is that the treatment could be toxic [38,39], but some studies suggest otherwise [35,37]. iCLIP has achieved an even higher resolution and efficiency compared to the first two. However, the experimental set-up and computational analysis are complicated.

The three CLIP-based experiments along with other variations have extended the scope of the genome-wide map of protein-RNA interactions. Choices of CLIP-based methods, experimental design, and the selection of proper bioinformatic pipelines have been reviewed in detail [38,40,41]. Such high-resolution maps of protein-RNA interactions advance the understanding of splicing regulation, and how protein-RNA interactions play a role in human physiological and pathological process. In addition, the detection of RNA-protein interaction can reveal biomarkers, more importantly it could identify potential therapeutic targets [34].

The protein-RNA interaction is not the only player in regulating alternative splicing. RNA structures also play a significant role [42]. They could either promote splicing or inhibit splicing. In general, RNA structures aid splicing by bringing splicing signals close to each other. For example, in the adenoviruses ADML gene, a stem-loop is formed in order to bring the 3’ splice site into the proximity of three possible branch points [43]. Another astonishing example of promoting alternative splicing through RNA structures is through a docking and selector site in the Drosophila melanogaster DSCAM gene. DSCAM encodes over 38,000 distinct mature mRNA isoforms by mutually exclusive splicing of 95 alternative exons [44]. The majority of its diversity is generated by the exon 6 cluster, which contains 48 alternative exons. A conserved selector sequence is complementary to a portion of the docking site. This RNA-RNA base pairing between the docking site and selector sequence ensures that only one of the 48 alternative exons is used to produce mature mRNA [45]. This pattern illustrates that these competing RNA secondary structures are the key element to maintain and assist the formation of protein isoforms. As for RNA structure suppressing splicing, native RNA structures often sequester important sequences required for splicing, thus repress the usage of splice sites. In some other cases, structures close to exons or splice sites may have negative impact on the recruitment of U1 or U2 snRNPs [46]. To conclude, these delicate machineries integrate and interact with each other to regulate alternative splicing in distinct cellular machines.

3. Alternative splicing associated human disease

Considering the important role of pre-mRNA splicing in composing protein diversity and maintaining organism functionality, it is no surprise that disruption of normal splicing patterns can cause gene dysfunction and even disease. There are around 20,000 human protein-coding genes but almost 150,000 transcript isoforms. Thus, on average each human gene has about seven transcript isoforms. Meanwhile, a recent study finds that over 30% of tissue-dependent transcript variations are constituted by local splicing variations [47]. Given that such high level of human genome complexity is exemplified by the flexibility of each gene with alternative splicing, it is natural to consider such flexibility as a risk factor. Indeed, a great number of human diseases have been reported to be linked to defective splicing.

Mutation disrupting either trans-acting regulatory proteins or cis-acting regulatory sequences could lead to aberrant splicing. In general, trans-acting mutations are rarer than the other. It is very likely due to the fact that disruptions in basal factors of the splicing machinery are generally more lethal compared to mutations altering splicing of a single gene in cis [48]. We will discuss some well-characterized diseases caused by mutations in trans-acting splicing factors as well as mutations in cis-acting regulatory sequences leading to different types of aberrant alternative splicing.

3.1. Cerebro-Costo-Mandibular syndrome caused by disruption in the core splicing machinery

Mutations that have impact on spliceosome components or regulatory factors of mRNA processing have been established as the basis of some craniofacial disorders. Cerebro-Costo-Mandibular Syndrome (CCMS) is one of rare craniofacial disorders characterized by the Pierre Robin sequence (severe micrognathia, glossoptosis, soft cleft palate, and upper airway obstruction) and posterior rib defects. Other symptoms including intellectual disability and microcephaly have been reported [49,50]. Although most cases appear to be de novo, familial examples have been reported as well with both autosomal dominant and recessive inheritance [51]. Both Lynch et al. and Bacrot et al. identified the mutations in SNRPB as a cause of CCMS [52,53]. As mentioned above, the spliceosome consists of U1, U2, U4/U6, and U5 snRNPs. Each snRNP protein includes seven cores/Sm proteins. SNRPB-encoded small nuclear
3.2. Mutations in trans-splicing factors resulting in tumorigenesis

Alternative splicing possesses ubiquitous and flexible gene regulation in humans. Therefore, cancer cells often exploit this characteristic to expand and survive. Myelodysplastic syndromes (MDS) are a group of diverse cancers in the bone marrow caused by poorly understood mechanisms that lead to formation of immature blood cells [60]. The splicing factor SF3B1 encoding subunit 1 of the splicing factor 3b protein complex, which is a major component of U2 snRNP, is the most frequently mutated gene in MDS patients [61–64]. The splicing factor 3b together with 3a bind the branchpoint sequence in pre-mRNA. The stable binding is indispensable to recruit and anchor U2 snRNP to the pre-mRNA [65]. SF3B1 also serves as a component of the minor U12-type spliceosome [66] and has a role in the commitment complex [67]. SF3B1 knockdown leads to growth inhibition and deregulation of numerous genes and pathways [61]. Due to the important role of SF3B1 in spliceosome machinery, it is not surprising that RNA-seq analysis of tumor tissues with SF3B1 mutations has revealed global splicing defects caused by the perturbed branch region fidelity [68–72]. Other splicing factors involving 3′-splice site recognition such as U2AF1 and SRSF2 have also been reported to harbor somatic mutations associated with MDS [73,74]. The detailed mechanism and pathways of aberrant splicing in cancer progression have been reported and discussed extensively in those referred reviews [75–79]. More importantly, the identification of mutations in those splicing factors suggests that spliceosome machinery could be a therapeutic target for certain cancers [61,62].

3.3. Spliceostatin A, a potent antitumor compound inhibiting splicing

As we discussed above, defects in spliceosome machinery lead to major dysfunctions and disorders. Interestingly, the inhibition of pre-mRNA splicing in cancer cells using Spliceostatin A (SSA) can suppress their proliferation [80–82]. SSA is a methylated derivative of a natural product FR901464 that inhibits pre-mRNA splicing in vitro and in vivo by binding to SF3b, a protein subcomplex of U2 snRNP which is indispensable for the recognition of pre-mRNA branchpoint. The interaction between SF3b 155-kDa subunit and mRNA is disrupted by SSA, which leads to nonproductive recruitment of U2 snRNP to 5′ branchpoint. Furthermore, down-regulation of genes that are crucial for cell division was observed, explaining the anti-proliferative effects of SSA [82]. SSA is not only a valuable target in cancer treatment, it can also suppress viral replication since splicing is vital for certain class of viruses to infect hosts [82–84].

3.4. Disease associated with changes in ratios of protein isoforms caused by dysregulation in alternative splicing

Tauopathies describe a class of neurodegenerative disorders characterized by neuronal and/or glial inclusions composed of the microtubule associated protein, tau [85]. The tau protein family consists of a group of six highly soluble protein isoforms produced by alternative splicing from a single gene MAPT (microtubule associated protein tau) [86]. The functionality of the tau protein family is mostly in maintaining the stability of microtubules in axons where it binds to microtubules via microtubule repeat regions. One of these microtubule binding regions is encoded by the alternatively spliced exon 10. The longest tau isoform (4R isoform) has four microtubules repeat-regions (R1, R2, R3, R4) at the C-terminal caused by exon 10 inclusion while the shortest isoform (3R isoform) has three repeats (R1, R3 and R4) due to the exon 10 skipping event. These splicing events are spatially and temporally regulated, and literatures documented that mutations in exon 10 alter its normal ratio of inclusion and exclusion. This fraction change has been confirmed to be the cause of frontotemporal dementia with parkinsonism linked to chromosome 17 [87,88]. Growing evidence also shows that several other neuron degenerative diseases such as Alzheimer’s disease, Parkinson’s disease and Huntington’s disease can be associated with microtubule dysfunction caused by imbalanced ratio of tau isoforms [89–94].

The Wilm’s tumor gene WT1 is another example that alteration of its isoform expression causes human disease: Frasier syndrome. WT1 encodes a zinc finger protein that binds to DNA. Its exon 5 and 9 are alternatively spliced leading to the formation of four isoforms [95,96]. An intronic point mutation on WT1 causes the disruption of alternative splicing at the splice donor site of exon 9, which further prevents the synthesis of WT1 + KTS isoform. Studies suggest that the gonadal development may be particularly sensitive to the imbalance of WT1 + KTS and WT1 – KTS which eventually leads to Frasier syndrome [97].

The above examples illustrate that abnormal changes in isoform ratios can cause human diseases. Moreover, in such cases it is imperative to accurately estimate isoform abundance in order to shed light on the relative contribution of each isoform to the different physiologic states. We will discuss isoform quantification in detail in part III.

3.5. Distinct disease severity caused by point mutations in mutually exclusive exons

Timothy syndrome is a rare congenital disorder that primarily affects the heart but can also affect many other tissues including teeth, nervous systems, and immune systems. There are two documented types of Timothy syndrome: classical (type-1) and atypical (type-2). They are both caused by de novo point mutations in CACNA1C, a gene encodes an alpha-1 subunit of a voltage-dependent calcium channel [98]. At least 19 out of 55 exons of CACNA1C are subject to alternative splicing [99–101]. Its transmembrane segment IS6 is encoded by inclusion or exclusion of the mutually exclusive 8 and 8a exons [102]. Point mutations on those exons lead to a mutated channel that shows a much slower voltage-dependent inactivation leading to a larger influx of calcium ions. Since the mutations only occur in one of the mutually exclusive exons, usually the ion channel encoded by the unaffected exon will function normally. Exon 8 is expressed in the smooth muscle, while exon 8a is expressed in the cardiac muscle [100,102]. Thus, mutations on CACNA1C cause different level of severity in patients with Timothy syndrome. The most fatal mutation takes places in the cardiac exon 8a, which often leads to cardiac arrhythmia, while mutations found in the smooth-muscle-related exon results in a less severe outcome [98,103].

3.6. Familial dysautonomia caused by a mutation in the 5′ splice site leading to an exon skipping event

Familial dysautonomia (FD) is a rare, recessive genetic disorder that affects the development and survival of sensory sympathetic
and parasympathetic nerve cells in the autonomic nervous system. The autonomic nervous system controls involuntary actions such as digestion, breathing, and the regulation of blood pressure and body temperature. Therefore, patients with FD show various symptoms including insensitivity to pain, difficulty swallowing, poor growth, pneumonia, labile blood pressure and gastrointestinal dysmotility. FD is the result of loss-of-function of an inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein (IKAP, also known as elongator complex protein 1, ELPI) [104]. Most FD patients have a point mutation in the 5’ splice donor site of exon 20. It is a T > C transition that weakens the intronic part of the 5’ splice site and results in a skipping event of exon 20. Translation of this mRNA produces a truncated IKAP protein, which misses all amino acids encoded from exon 20. Thus, the decreased level of functional IKAP protein expression causes many FD symptoms. In addition, studies also indicate that IKAP deficiency results in down regulation of genes involved in oligodendrocyte differentiation and myelination which could be the cause of demyelination symptoms of FD patients [48,105]. This example not only demonstrates the complexity of mutations in splice sites, it also suggests that such a mutation may have impact on multiple physiological process.

3.7. Disease caused by intron retention

Intron retention (IR) has been the least described subtype of alternative splicing until recently. Intron-retaining transcripts were previously thought to be non-functional since they would be degraded by nonsense-mediated decay, a surveillance pathway that prevents such aberrant mRNA from being translated into potentially harmful abnormal proteins [11,106–109]. However, recent studies have confirmed that IR is not only a conserved form of alternative splicing, but also plays an essential role in controlling and enhancing the complexity of gene expression [9,110]. Here, we will discuss some consequences of abnormal IR and the emergent role they play in diverse diseases.

IR is a widespread event occurring in many diseases. Point mutations occurring at critical splice control points such as donor or acceptor sites can result in partial or full retention of specific introns. Autoimmune polyendocrine syndrome type 1 (APS-1) is a rare and complex recessively inherited disorder of immune-cell dysfunction with multiple autoimmunities [111,112]. Loss-of-function mutations of the autoimmune regulator gene (AIRE) have been reported to cause APS-1. AIRE is a transcription factor expressed in the thymus medulla and lymph nodes. It controls the local transcription of tissue-specific proteins typically expressed in peripheral tissues, thus allowing the negative selection of self-reactive T cells [113,114]. The c.463G > A transition in AIRE generates an aberrant transcript retaining intron 3. This aberrant intron 3-retaining transcript generates a truncated protein with a premature stop codon. The truncated protein contains only the first normal functional 154 amino acids followed by 48 aberrant amino acids [115].

Recent studies also revealed that IR events are very common across cancer patients. Partial retention of intron 6 of GSTP1 has been confirmed as one cause of the head and neck cancer [116]. Truncated cyclin D1b is produced by CCND1 with partially retained intron 4, which leads to the prostate and esophageal cancer [117,118]. Moreover, in a recent large-scale transcriptome profiling study, IR events are observed in solid cancers originating from bladder, breast, colon, colon, head and neck, kidney, liver, lung, prostate, rectum, stomach, thyroid, and endometrium, as well as in acute myeloid leukemia [119].

3.8. Transcriptome-wide alternative splicing analysis in disease

Aberrant splicing greatly contributes to the pathologies of many diseases as mentioned. Indeed, the detection of differential splicing not only serves as predictive, diagnostic, or prognostic biomarkers but also stratifies patients with different conditions. Here we briefly discuss examples of transcriptome-wide sequencing unveiling differential splicing events in diseases. Lin et al. identified 593 differential splicing events between Huntington’s disease (HD) and control brains, and four splicing factors with significantly altered expression [120]. The follow-up study verified the impact of the splicing factor PTBP1 on disease-associated splicing patterns in HD patients [120]. Similar study identified hundreds of aberrant splicing events in Alzheimer’s disease [121]. For cancers, differential alternative splicing is more than a diagnostic biomarker. Increasing evidence indicated that it could elucidate the progression and serve as prognostic biomarkers [122,123].

Many studies employed univariate and multivariate Cox regression analyses to identify survival-associated alternative splicing events and to compute risk scores [124–132]. More recently, a novel statistical model SURVIV (Survival analysis of mRNA isoform variation) has shown its ability to outperforms the conventional Cox regression model [133]. SURVIV uses a survival-measurement-error model to estimate uncertainty of mRNA isoform ratios and assesses the association between isoform ratios and survival time of cancer patients. SURVIV outperforms Cox regression model under many different settings especially with low-depth or moderate-depth RNA-seq data. This makes SURVIV extra practical since many clinical studies have not yet reached deep sequencing depth.

Beyond the significant role of alternative splicing in disease diagnosis and prognosis, it is more than informative to deep dive into disease pathological mechanisms to investigate how dysfunctions of alternative splice leading to distinct diseases and to connect phenotypes with genotypes. Despite the profound genetic understanding for many human genetic diseases, causative genetic variations or functional roles of highly correlated variations of many diseases still consist of a big piece of puzzle. To facilitate the understanding from genetic variants to human pathologies, it is necessary to accurately quantify gene expression at the individual transcript isoform level since isoforms are the ultimate units to execute gene functions. In addition to the expression quantification of annotated genes, the discovery of novel transcript isoforms and novel splicing events is also crucial for deciphering the role of alternative splicing in human physiology and pathologies.

4. Computational tools for isoform quantification from RNA-seq

Over the past few years, massively parallel RNA sequencing (RNA-seq) has become a powerful tool for comprehensive transcriptomic analysis. Furthermore, with the cost of deep sequencing dropping drastically, large scale studies for expression and alternative splicing have become plausible. Such studies will expedite the discovery of more splicing-related diseases and the gained knowledge has the potential to develop preventive and therapeutic interventions for these diseases.

However, precise quantification of alternative splicing is still hindered by the technological limitation, mainly the limited read length. During an RNA-seq experiment, mRNA is extracted from the tissue, then fragmented and reverse transcribed into cDNA, which is further amplified and sequenced by high-throughput, short-read sequencing methods. Ideally transcriptome assembly can be performed to reconstruct genome regions being transcribed. However, considering the typical read length of RNA-seq ranging from 50 bp to 150 bp, and the fact that transcript isoforms of the
same genes are usually difficult to distinguish, the expression quantification of transcript isoforms remain challenging.

Moreover, the discovery of novel transcripts using short reads is one of the most challenging tasks in RNA-seq [134]. Short reads hardly ever spanning across splice junctions complicate the inference of full-length transcripts especially for lowly expressed transcripts. The identification of transcription start and end sites also remains a difficult mission [135]. Recently, another computation difficulty emerges as sequencing depth further increases due to the affordable sequencing. The massive sequence data may not only overwhelm hardware resources, but also confound heuristic algorithms that may not be scalable to that gigantic size of data [136].

Various computational tools have been developed in order to tackle these major complications in the last decade. In general, methods for alternative splicing analysis can be divided into two major categories: isoform-centric approaches and exon-centric approaches. Isoform-centric approaches quantify full-length transcript isoforms, and then the alternative splicing ratio can be estimated based on the ratio between isoforms including and isoforms excluding an alternative exon of interest. Exon-centric approaches directly focus on the quantification of exon inclusion ratios.

Base on whether the reference genome or the reference transcriptome is utilized, transcript quantification can be categorized as reference-based approaches or de novo assembly-based ones [137]. Reference-based approaches first align reads to a reference genome. Then those mapped reads are distributed to known annotated isoforms computationally. Or a splicing graph representing all possible alternative splicing events can be built upon the mapped reads and individual isoforms are finally assembled by traversing the graph. The high sensitivity of reference-based approach permits the discovery of novel transcript as well [138]. For example, several popular methods, such as StringTie [139] and Cufflinks [138] apply genome-aligned reads and take advantage of existing annotations when assembling transcripts including novel ones. Transcript isoform quantification is performed simultaneously or afterwards. Although the majority of reference-based quantification methods align reads to the genome, recently there are methods that developed as alignment-free approaches. Such alignment-free methods pre-index reference transcripts, break sequence reads into k-mers, and then preform fast matches to the pre-indexed transcripts as “pseudo”-alignments.

In general, referenced-based approaches are computationally economic and more suitable for isoform quantification especially when the high-quality reference is available. But for non-model organisms de novo assembly-based method is indispensable. De novo assembly-based approaches usually rely on a De Bruijn graph to assemble isoforms, therefore the required computational resources are enormous. Regarding to the alternative splicing analysis, even though some studies have demonstrated that de novo assembly-based method is capable of detecting novel splicing events, the application on human datasets remain sparse [140,141].

In this review, we will mainly focus on reference-based isoform-centric or exon-centric methods. Representative packages for the detection and quantification of isoforms or splicing events with or without the discovery of novel isoforms or splicing events will be discussed. To facilitate users’ selection, we summarized a collection of benchmark characteristics for these representative tools based on simulated studies, literature reviews [142–144], and software documentation (Table 1). We will briefly mention de novo assembly-based methods afterwards.

4. Isoform-centric analysis tools

4.1. Cufflinks

Although developed nearly a decade ago, Cufflinks is perhaps one of the most popular approach for expression quantification, especially for simultaneous novel transcript discovery and abundance estimation [138]. Cufflinks adopts and extends the ideas from expressed sequence tag (EST) assemblers such as PASA, which collapse alignments to transcripts on the basis of splicing compatibility [145]; as well as Dillworth’s theorem which was originally applied to assemble a parsimonious set of haplotypes from sequencing reads of a mixed virus population [146]. The approach reduces the transcript assembly problem into searching a maximum matching in a weighted bipartite graph that represents compatibilities among fragments. Their validation results suggest that Cufflinks is capable of not only improving transcriptome-based genome annotation but also discovery of novel transcripts.

4.1.2. StringTie

StringTie is another highly efficient transcript quantification tool which has the ability to discover novel transcripts [139]. It uses a novel network flow algorithm combined with an optional de novo assembly step to discover and quantify transcripts concurrently. StringTie has assembled 53% more transcripts than Cufflinks in a benchmark of 90 million reads from human blood. Moreover, it runs faster and consumes less memory. StringTie belongs to the updated Tuxedo protocol (HISAT, StringTie, Ballgown) which achieves faster speed, substantially memory saving, and more accurate overall results than the previous Tuxedo protocol (Tophat, Cufflinks, Cuffdiff) [147]. The output of StringTie is also compatible
with downstream specialized tools such as Cuffdiff, DESeq2, edgeR, and so on.

4.1.3. RSEM

RSEM is also a popular tool for quantifying gene and isoform abundances from single-end or paired-end RNA-Seq data [142,143,148,149]. RSEM takes original FASTQ sequence files as well as BAM/SAM alignment files. It is equipped with Bowtie2, STAR and HISAT aligners internally for read alignment. Or users can choose their preferred alternative aligner by providing aligned BAM/SAM files. RSEM is an Expectation-Maximization (EM) algorithm-based method, which assigns mapped reads to transcript isoforms and estimates the maximum likelihood (ML) of relative abundances of transcript isoforms. However, since the assignment of reads to isoforms are resulted from iterations of the EM method, the performance of RSEM is relatively inefficient [142,149]. Given the recent massive sequence data size, the speed of RSEM has become a major drawback.

4.1.4. WemiQ

Precise transcriptome quantification is also hindered by non-uniform short-read sampling. The severity of overdispersion can be exemplified by non-uniform read distribution along single-isoform genes. The hidden bias is multifactorial with many unknown causes, which varies between sequencing platforms and protocols. To address this fundamental problem and fully harness the power of transcriptomics data, Chen’s group has developed an integrated framework that can tackle the overdispersion issue in bulk RNA-seq and single-cell RNA-seq (scRNA-seq) [150-152]. The RNA-seq bias in these models is estimated by a generalized-Poisson model (GPSeq) in a data-adaptive and assumption-free manner [150]. GPSeq estimates the bias directly from read-count distribution and does not need to specify the (often unknown) bias sources. GPSeq outperforms commonly used bias correction methods and still efficiently removes bias in cases that traditional methods fail [151].

For isoform-level quantification, they developed WemiQ [151]. The heterogeneity of read counts along a multi-isoform gene is caused by both RNA-seq bias and annotation heterogeneity (i.e., exonic positions are shared by different isoforms). The challenge is to separate bias from signals during deconvolution of isoform expression. WemiQ addresses this challenge by assigning different weights to reads at different positions. Weights are derived from bias-correction factors using GPSeq. Then an EM algorithm is used to distribute reads among different isoforms and maximize the weighted loglikelihood. Both simulation and empirical analyses showed that WemiQ significantly improves the accuracy of isoform quantification and estimation of exon inclusion rates.

4.1.5. eXpress

eXpress is a more recent tool that takes advantage of an online-EM inference procedure that permits accurate inference of transcript abundance after a single pass over the read alignments [153]. In eXpress, Roberts and Pachter modified the online EM algorithm that resolves the fragment-assignment problem into one that works directly with estimated counts instead of relative abundances. In the algorithm, each incoming fragment can be mapped to any number of target sequences and assigned to its mapped target based on the previously estimated counts. As fragments being processed, their assignment allows the algorithm to update and improve parameter estimates. This dynamic scheme greatly enhances the convergence speed and improves software performance. However, since eXpress is still an alignment-based method, it requires input alignments in the SAM/BAM format. Thus, although the procedure itself is fast, the alignment step is still inevitable and time consuming.

4.1.6. Sailfish

Alignment-based quantification tools are usually relatively slow, below we will introduce three alignment-free methods developed in the past few years. Sailfish is a k-mer-based approach, which completely avoids the time-consuming step of mapping reads to a reference genome [154]. Expression quantification is performed by extraction of k-mers from reads followed by exact matching of the k-mers using a hash table. This approach provides much faster quantification estimates than other approaches (typically 20 times faster). Although it claimed there was no trade-off of accuracy loss, fragmenting sequence reads into k-mers indeed loses valuable information [142,143]. K-mers may align to more transcripts than the read itself since they are shorter, thereby leading to loss of accuracy.

4.1.7. Kallisto

Kallisto is another alignment-free k-mer-based method developed to tackle the inadequate accuracy mentioned in the previous method [143]. It is based on pseudoalignment of reads and uses fast hashing of k-mers together with the transcriptome de Bruijn graph, which has been proven to be crucial for DNA and RNA assembly [155]. Clustering of pseudoalignments originated from the same transcripts into equivalence classes permits simpler likelihood function and more efficient algorithm convergence. The accuracy of Kallisto is similar to those of alignment-based RNA-seq quantification approaches since its pseudoalignments explicitly maintain the information provided by k-mers across reads, yet the speed of Kallisto is two orders of magnitude faster.

4.1.8. Salmon

Salmon is introduced by the same developers of Sailfish; it is also an ultra-fast alignment-free method [156]. Interestingly the developers no longer utilize k-mer-based algorithm. Salmon applies a two-phase parallel inference procedure consisting of a reduced data representation, and a novel lightweight read alignment algorithm. The first online phase uses a variant of stochastic, collapsed variational Bayesian inference (SCVB0) to estimate initial expression levels, model parameters, and to construct equivalence classes over the input fragments. During the second offline phase, a variational Bayesian EM algorithm is applied over a reduced representation of the data to refine the initial transcript abundance estimates until a data-dependent convergence criterion is achieved. This method achieves both significantly faster speed and state-of-art accuracy, and it is now widely applied in RNA-seq data analysis because of those merits. One convenience Salmon provides is that, unlike other two alignment-free methods, it also takes SAM/ BAM files as input. Therefore, for analyses requiring better quality control, Salmon may be the choice.

4.2. Exon-centric analysis tools

4.2.1. MISO

As mentioned above, event-based approaches are the emerging direction for alternative splicing analysis. MISO (Mixture-of-isoforms) is a statistical model that quantifies expression of alternatively spliced exons and isoforms and identifies differentially regulated ones across samples [157]. The MISO model uses probabilistic method (Bayesian inference) to compute the probability of the reads originated from a certain isoform. MISO enables both exon-centric analysis and isoform-centric analysis with improved accuracy for quantification and differential detection across samples.

4.2.2. SUPPA

SUPPA is another powerful tool to investigate alternative splicing events [158]. SUPPA generates alternative splicing events from
an input annotation file. It calculates the percentage or “proportion spliced-in” (PSI) to describe relative abundance of splicing events or transcript isoforms, exploiting the fast transcript quantification. Differences of these relative abundances (APS) across different samples are also reported to quantify differential splicing. SUPPA achieves a much faster computational speed, while its accuracy is on a par with most standard methods based on analysis of experimentally validated events. SUPPA would be an effective and affordable choice for splicing analysis of big datasets. While SUPPA itself is restricted to annotated splicing events, coupling with novel transcript reconstruction methods such as StringTie makes it a powerful tool to identify novel splicing events.

4.2.3. SplAdder

While most of event-based methods do not support the identification of novel splicing events, SplAdder is a powerful tool to tackle the task. To avoid the computational burden posed by the identification of complete transcripts, SplAdder takes an innovative approach which treats individual splicing events as proxy for transcriptome characteristics [159]. SplAdder takes a given annotation and summarizes all transcripts of a gene into a splicing graph. The splicing graph is further augmented with new information extracted from RNA-seq (i.e. new introns and exon segments detected in the alignment). Alternative splicing events can be identified through the augmented annotation graph, and eventually be quantified with RNA-seq data. SplAdder is able to detect all main types of alternative splicing as well as multiple (coordinated) exon skips. In terms of accuracy, SplAdder has achieved an overall high accuracy in the benchmark comparing to various other state-of-the-art methods, while remaining computationally efficient.

4.3. De novo assembly-based and reference-free tools

Discovery of novel transcripts has been challenging ever since the question was posed. Although tools such as Cufflinks [138], StringTie [139], SLIDE [160,158], and IsoLasso [161] take advantage of different algorithms (EM, Lasso, network flow algorithm, and so on) and incorporate existing annotations to perform transcript discovery, the results of accurate transcript reconstruction exhibit certain disagreements and remain unsatisfactory [135]. More importantly, those genome-guided novel transcript detection may be biased by the process itself [134,162]. One way to mitigate the bias from the reference genome is to apply de novo transcript assembly through packages such as Trinity [163], Trans-AbySS [164], and Oases [165]. Trinity is by far the most widely applied de novo assembler with over 10,000 citations. It is named Trinity as three software modules were involved. Inchworm first assembles the RNA-seq data into linear contigs; Chrysalis then groups related contigs and constructs de-Brujin graphs; finally Butterfly examines the reads and reports full-length transcripts through dynamic programming. In splicing analysis, de novo assembly-based approaches usually first reconstruct and quantify isoforms, and then quantify alternative splicing events. One exception, Kiss-Saplice [166], de novo assembles alternative splicing events directly, and achieves a higher accuracy than other de novo full-length transcript assemblers.

4.4. Comparison of different analysis tools

Most of the isoform quantification with known transcripts exploit either EM algorithms or Bayesian inference to perform the read-count deconvolution. The difference mainly lies in the algorithm convergence speed [142]. One exception is WemiQ which also handles the bias correction in RNA-seq to improve the quantification accuracy. Generally, alignment-free methods, such as Salmon and Kallisto are extremely fast. One should take advantage of those tools when processing tremendously large-scale data. Alignment-based methods sometimes have slightly better performance in accuracy than alignment-free methods according to previous evaluation [142].

The accuracy of novel transcript or splicing event discovery still remains questionable, due to the lack of agreement of different methods, one should consider combine two or more approaches to validate with each other if high accuracy is desired. As mentioned before, the quantification of novel isoforms suffers from bias when deploying methods incorporating existing annotation. Using de novo assembly-based methods such as Trinity followed by downstream quantification packages such as RESM, eXpress, or Salmon may mitigate the issue. A recent study shows that in terms of conducting alternative splicing annotation and differential analysis, the results of de novo assembly-based methods and reference-based methods overlap by only 70% with noticeable differences [141]. Since assembly-based methods are capable of detecting more novel events and reference-based methods perform better for lowly expressed transcripts, Benoit-Pilven et al. suggest that the combination of those would be a better alternative [141].

Isoform-centric methods and exon-centric methods complement each other. The choice should be decided by the raised biological questions. Each of the different approaches could shed light on different discovery. While isoform-centric methods rely on quantification of full-length transcripts, exon-centric approaches have bypassed this step by directly exploiting the quantification of alternative splicing events. Some studies concluded that exon-centric methods are superior to isoform-centric methods [167–169]. Another study showed that some exon-centric methods may not handle multiple replicates well [170]. We also argue that although exon-centric methods are more sensitive with known transcripts, isoform-centric ones provide more flexibility in novel and complicated alternative splicing events. Current isoform-centric methods are mainly hindered by the computational accuracy issue resulting from short reads. Recent advances in third-generation sequencing (long-read technology) such as single molecule real time sequencing (SMRT) from Pacific Biosciences, or the Nanopore sequencing from Oxford can produce substantially longer reads. With reads longer than a typical full-length transcript, in the near future, the accuracy of isoform quantification will allow the direct quantification of splicing events.

5. Conclusion

It has been more than four decades since Walter Gilbert first proposed the concept of alternative splicing: special combination of exons being spliced together to make special differentiation products [171]. Our understanding about alternative splicing is advancing rapidly with the help of molecular research, high-throughput sequencing and bioinformatic tools, yet there are still more to explore how alternative splicing functions at cellular level. As discussed above, quantification of full-length transcripts and lowly expressed genes are the two major stumbling blocks for alternative splicing studies based on RNA-seq. We envision advances in two types of technologies may resolve those issues in the near future, 1) Long-read sequencing, SMRT and Nanopore sequencing technologies, has the potential to sequence the entire transcript bypassing the reconstruction issue. Those technologies have been proven applicable in revealing full-length transcripts and alternative splicing events [172–176]. 2) Single-cell RNA-seq (scRNA-seq) examines gene expression on a cellular resolution with optimized next-generation sequence technologies. With proper protocols and bioinformatic pipeline, scRNA-seq has the potential to better unravel lowly expressed transcripts and splicing
events [177–180]. No doubt there are still limitations in both technologies. Long-read sequencing suffers from high error rates, thus de novo transcript detection and quantification is still unsatisfactory [181]; scRNA-seq usually yields fewer expressed genes caused by limited sequence depth and suffers from little distinguishment between technical and biological noise [180]. But with the combination of those newer technologies with traditional ones, more and more powerful computing tools, complete understanding of alternative splicing will be fulfilled in the near future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been supported by the National Institutes of Health [R01GM137428].

References

[1] Graveley BR. Alternative splicing: increasing diversity in the proteomic world. Trends Genet 2001;17(2):100–7.

[2] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40(12):1413–5.

[3] Marquez Y, Brown JW, Simpson C, Bara A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 2012;22(6):1184–95.

[4] Cui Y, Cai M, Stanley HE. Comparative analysis and classification of cassette exons and constitutive exons. Biomed Res Int 2017;2017.

[5] Sammeth M, Foisssac S, Guigo R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol 2008;4(8):e1000147.

[6] Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 2003;72(1):291–336.

[7] Sugnet CW, Kent WJ, Ares M, Haussler D. Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput 2004;66–77.

[8] Sammeth M. Complete alternative splicing events are bubbles in splicing graphs. J Comput Biol 2009;16(8):1117–40.

[9] Vanschick DA, Schmitz U, Wong J, Rasko JE, editors. Challenges in defining Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40(12):1413–5.

[10] Grigorieff N. Three-dimensional structure of C

[11] Chebli K, Gattoni R, Schmitt P, Hildwein G, Stevenin J. The 216-nucleotide Bacrot S, Doyard M, Huber C, Alibeu O, Feldhahn N, Lehalle D, et al. Mutations Marquez Y, Brown JW, Simpson C, Bara A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 2012;22(6):1184–95.

[12] Wang T, Xiao G, Chu Y, Zhang MQ, Corey DR, Xie Y. Design and bioinformatics analysis of genome-wide splicing CLIP experiments. Nucleic Acids Res 2015;43(11):5263–74.

[13] Spitzen J, Hafner M, Landthaler M, Ascano M, Farazi T, Wardle G, et al. PAR-CLIP (Photolabile Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. Methods in enzymology. 539: Elsevier; 2014. p. 113–61.

[14] Wang T, Xiao G, Chu Y, Zhang MQ, Corey DR, Xie Y. Design and bioinformatics analysis of genome-wide splicing CLIP experiments. Nucleic Acids Res 2015;43(11):5263–74.

[15] Grigorieff N. Three-dimensional structure of C

[16] Wang Z, Xiao X, Van Nostrand E, Burge CB. General and specific functions of splicing regulation: from a parts list of regulatory ensembles to an integrated splicing code. Trends Genet 2013;29(1):59–66.

[17] Uhl M, Houwaart T, Corrado G, Wright PR, Backofen R. Computational enzymology. 539: Elsevier; 2014. p. 113-61.

[18] Spitzen J, Hafner M, Landthaler M, Ascano M, Farazi T, Wardle G, et al. PAR-CLIP (Photolabile Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. Methods in enzymology. 539: Elsevier; 2014. p. 113–61.

[19] Wang T, Xiao G, Chu Y, Zhang MQ, Corey DR, Xie Y. Design and bioinformatics analysis of genome-wide splicing CLIP experiments. Nucleic Acids Res 2015;43(11):5263–74.

[20] Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila genome-wide insights into brain alternative RNA processing. Nature 2009;461(7266):1050–5.

[21] Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila genome-wide insights into brain alternative RNA processing. Nature 2009;461(7266):1050–5.

[22] Wu W, Jiang and L. Chen Computational and Structural Biotechnology Journal 19 (2021) 183–195
Reduced fidelity of branch point recognition and alternative splicing induced by the anti-tumor drug spliceostatin A. Genes Dev 2011;25(5):445–59.

Wang L, Brooks AN, Fan J, Wan Y, Gambe R, Li S, et al. Targeting viral proteostasis limits influenza virus, HIV, and dengue virus infection. Immunity 2016;44(1):46–58.

Yang N, Gibbs JS, Hickman HS, Reynoso GV, Ghosh AK, Bennink JR, et al. Development of a viral-derived translation initiation event that generates a common peptide from influenza A Virus M2 and M1 mRNAs. J Immunol (Baltimore, Md 1950) 2016;196(9):3608.

Irwin DJ. Tauopathies as clinicopathological entities. Parkinsonism Related Disorders 2016;22:529–33.

Sergeant N, Delacourte A, Buele L. Tau protein as a differential biomarker of tauopathies. Biochim Biophys Acta (BBA)-Mol Basis Dis 2005;1739(2–3):179–97.

Wolfe MS. The role of tau in neurodegenerative diseases and its potential as a therapeutic target. Scientifica 2012;2012.

Andreas A. Misregulation of tau alternative splicing in neurodegeneration and dementia. Alternative Splicing and Disease. Springer, 2006. p. 89–107.

Spallini MG, Goertmer M. Tau protein pathology in neurodegenerative diseases. Trends Neurosci 2009;21(10):428–33.

Mudher A, Lovestone S. Alzheimer’s disease–do taus and tangles finally collide? Trends Biochem Sci 2011;36(12):655–65.

Fernández-Nogales M, Lucas J. Altered levels and isoforms of tau and nuclear membrane invaginations in Huntington’s disease. Front Cell Neurosci 2019;13.

Corder K, Kong Ó, Bruns G. The genomic organization and expression of the human pre-mRNA. Nat Chem Biol 2007;3(9):576–83.

Das R, Zhou Z, Reed R. Functional association of U2 snRNP with the ATP-dependent spliceosomal complex E. Mol Cell 2000;5(5):779–87.

U2-dependent spliceosome. RNA 2004;10(6):929–41.

Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5 homologue and an SF3b DEAD-box protein. EMBO J 2002;21(18):4978–88.

Will CL, Schneider C, Hoxlbsch M, Urlaub H, Raubt M, Elbasser S, et al. The human 18S U11/U12 snRNPs contains a set of novel proteins not found in the U2-dependent spliceosome. RNA 2004;10(6):929–41.

Das R, Zhou Z, Reed R. Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. Mol Cell 2000;5(5):779–87.

DeBoever C, Ghia EM, Shepard RJ, Jepsen K, et al. Transcriptome sequencing reveals potential mechanism of cryptic 3′ splice site selection in SF3B1-mutated cancers. PLoS Comput Biol 2011;7(13):e1002255.

Alsaﬁdi S, Hooj A, Battistella A, Popovata T, Wassef M, Henry E, et al. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. Nat Commun 2016;7(1):1–12.

Darmen RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, et al. Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. Cell Rep 2015;13(5):1033–45.

Wang L, Brooks AN, Fan J, Yan W, Gamble R, Li S, et al. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. Cancer Cell 2016;30(5):750–63.

Kesarwani AK, Ramirez O, Gupta AK, Yang X, Murthy T, Minella AC, et al. Cancer-associated SF3B1 mutations recognize otherwise inaccessible cryptic 3′ splice sites within RNA secondary structures. Oncogene 2017;36(8):1123–33.

Je EM, Yoo NJ, Kim YJ, Kim MS, Lee SH. Mutational analysis of splicing machinery genes SF3B1, U2AF1 and SF3B2 in myeloproliferative neoplasms. Blood 2004;103(1):2439–46.

Srivastava S, Dey O, Maitra A, Paul D, et al. SF3B1 gene mutations in myeloid disorders. J Hematol Oncol 2011;4:1.

and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myeloproliferative syndromes. Blood, J Am Soc Hematol 2012;118(24):6239–46.

Characterization of SF3B1 mutations in myelodysplastic syndromes and myeloproliferative neoplasms. Blood, J Am Soc Hematol 2011;121(10):1384–95.

Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5 homologue and an SF3b DEAD-box protein. EMBO J 2002;21(18):4978–88.

Will CL, Schneider C, Hoxlbsch M, Urlaub H, Raubt M, Elbasser S, et al. The human 18S U11/U12 snRNPs contains a set of novel proteins not found in the U2-dependent spliceosome. RNA 2004;10(6):929–41.

Das R, Zhou Z, Reed R. Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. Mol Cell 2000;5(5):779–87.

DeBoever C, Ghia EM, Shepard RJ, Jepsen K, et al. Transcriptome sequencing reveals potential mechanism of cryptic 3′ splice site selection in SF3B1-mutated cancers. PLoS Comput Biol 2011;7(13):e1002255.

Alsaﬁdi S, Hooj A, Battistella A, Popovata T, Wassef M, Henry E, et al. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. Nat Commun 2016;7(1):1–12.

Darmen RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, et al. Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. Cell Rep 2015;13(5):1033–45.

Wang L, Brooks AN, Fan J, Yan W, Gamble R, Li S, et al. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. Cancer Cell 2016;30(5):750–63.

Kesarwani AK, Ramirez O, Gupta AK, Yang X, Murthy T, Minella AC, et al. Cancer-associated SF3B1 mutations recognize otherwise inaccessible cryptic 3′ splice sites within RNA secondary structures. Oncogene 2017;36(8):1123–33.

Je EM, Yoo NJ, Kim YJ, Kim MS, Lee SH. Mutational analysis of splicing machinery genes SF3B1, U2AF1 and SF3B2 in myeloproliferative neoplasms. Blood 2004;103(1):2439–46.

Hafner-Tsai C, Nagaia Y, Grosman V, Okuno Y, Bacher U, Nagae G, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia 2014;28(2):241–7.

David C, Manley JL. Alternative pre-mRNA splicing regulation in cancer. Pharmacol Ther 2006;107(2–3):243–64.

Venables JP. Aberrant and alternative splicing in cancer. Cancer Res 2004;64(21):7647–54.

Obha S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene. 2014;33(46):5311-8.

Kalina Z, Zayakin P, Silina K, Line A. Alterations of pre-mRNA splicing in cancer. Genes, Chromosomes and Cancer. 2005;42(4):342-57.

Srivastava S, Dey O, Maitra A, et al. SF3B1 gene mutations in myeloid disorders. J Hematol Oncol 2011;4:1.

Kesarwani AK, Ramirez O, Gupta AK, Yang X, Murthy T, Minella AC, et al. Cancer-associated SF3B1 mutations recognize otherwise inaccessible cryptic 3′ splice sites within RNA secondary structures. Oncogene 2017;36(8):1123–33.

Srivastava S, Dey O, Maitra A, et al. SF3B1 gene mutations in myeloid disorders. J Hematol Oncol 2011;4:1.
Ma J, Xiang Y, Lin Z, Xue Y, Mao M, et al. SMRT sequencing analysis reveals the full-length transcripts and alternative splicing patterns in Ananas comosus var. bracteatus. PeerJ 2019;7. e7062.

Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 2013;31(11):1009–14.

Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. Am J Hum Genet 2018;102 (1):11–26.

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun 2017;8(1):1–11.

de Jong LC, Cree S, Lattimore V, Wiggins GA, Spurdle AB, Miller A, et al. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. Breast Cancer Res 2017;19 (1):1–9.

Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50(8):1–14.

Sekula M, Gaskins J, Datta S. Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects. Biometrics 2019;75(4):1051–62.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods 2013;10(11):1096–8.

Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 2012;30(8):777–82.

Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. PLoS One 2012;7(10):e46679.