

## Chapter 2

# Global Approaches to Alternative Splicing and Its Regulation—Recent Advances and Open Questions

Yun-Hua Esther Hsiao, Ashley A. Cass, Jae Hoon Bahn, Xianzhi Lin and Xinshu Xiao

**Abstract** Pre-mRNA splicing is an essential RNA processing step in eukaryotes. Alternative splicing generates distinct spliced isoforms of the same gene, thereby dramatically increasing transcriptome diversity. Since most human genes undergo alternative splicing, this process contributes to a wide spectrum of biological functions in healthy and disease states. Splicing is closely regulated by various *cis*-regulatory elements and *trans*-factors. With the advent of high-throughput experimental technologies and bioinformatic algorithms, we now have powerful means to study alternative splicing globally and uncover its functional impact and regulatory mechanisms. As more RNA sequencing (RNA-Seq) data from normal and disease conditions are becoming available, many studies are underway to dissect global misregulation of splicing in diseases and develop novel splicing-targeted therapeutics. In this chapter, we first discuss the experimental and bioinformatic approaches for identification of alternative splicing, followed by a comprehensive review on the state-of-the-art methodologies to study splicing regulation. In addition, we discuss the current challenges and open questions in the RNA splicing field

---

Y.-H.E. Hsiao · X. Xiao (✉)

Department of Bioengineering, University of California Los Angeles, Los Angeles, CA 90095, USA

e-mail: gxxiao@ucla.edu

A.A. Cass · X. Xiao

Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA 90095, USA

J.H. Bahn · X. Lin · X. Xiao

Department of Integrative Biology and Physiology, University of California Los Angeles, Los Angeles, CA 90095, USA

X. Xiao

Molecular Biology Institute, University of California Los Angeles, Los Angeles, CA 90095, USA

X. Xiao

UCLA, Terasaki Life Sciences Building 2000A, 610 Charles E. Young Drive, Los Angeles, CA 90095-1570, USA

including gene expression kinetics, co-transcriptional splicing, and therapeutic approaches targeting splicing.

**Keywords** Alternative splicing • RNA • RNA-Seq • Gene regulation

## 2.1 Introduction

First discovered nearly 40 years ago [1, 2], pre-mRNA splicing consists of a series of biochemical reactions that function to remove introns and ligate flanking exons. Exon–intron boundaries are defined by highly conserved consensus sequences including the 5' splice site (5'ss, or donor site), 3' splice site (3'ss, or acceptor site), and branch point sequences (BPSs) (Fig. 2.1). These sequences are recognized by the spliceosome, a dynamic multi-ribonucleoprotein complex composed of small nuclear ribonucleoproteins (snRNPs) (refer to [3] for detailed reviews). The spliceosome is the basic machinery that carries out splicing reactions.

In recent years, it was estimated that more than 90 % of human genes are processed through alternative splicing where multiple spliced isoforms are generated from a single gene, thus significantly increasing transcriptome diversity [4–6]. The most extreme case of alternative splicing is the *Drosophila* Down Syndrome cell adhesion molecule gene (*Dscam*) which includes 48 exons and can theoretically produce 38,016 alternative transcripts from a single gene [7]. Different types of alternative splicing exist with the most common ones being exon skipping, alternative 5'ss usage, alternative 3'ss usage, mutually exclusive exons, and intron retention [8].

It is now well established that alternative splicing contributes to a wide spectrum of cellular functions [9]. Disruption of normal splicing was reported for a large number of human diseases, which has been reviewed extensively [10–12]. As a functionally critical process, alternative splicing is regulated by a myriad of *cis*-elements and *trans*-acting factors (Fig. 2.1). Splicing regulatory elements (SREs) reside in exons or introns and function to either enhance or silence splicing. These *cis*-elements are thus named accordingly as: exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs), and intronic splicing silencers (ISSs). These *cis*-elements interact with many *trans*-acting factors (i.e., splicing factors), including serine/arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) [13]. RNA secondary structures also affect alternative splicing, likely by facilitating or blocking accessibility of splicing factors to their cognate RNA [14].

Understanding the regulatory mechanisms of alternative splicing in health and disease is an essential topic of gene regulation. Recent advances in high-throughput technologies and related bioinformatic methodologies are enabling exciting discoveries in this area. Here, we first focus on global approaches for splicing identification, followed by an in-depth review of methodologies to study splicing regulatory mechanisms.

## 2.2 Identification and Validation of Alternative Splicing Events

### 2.2.1 Identification of Alternative Splicing Events

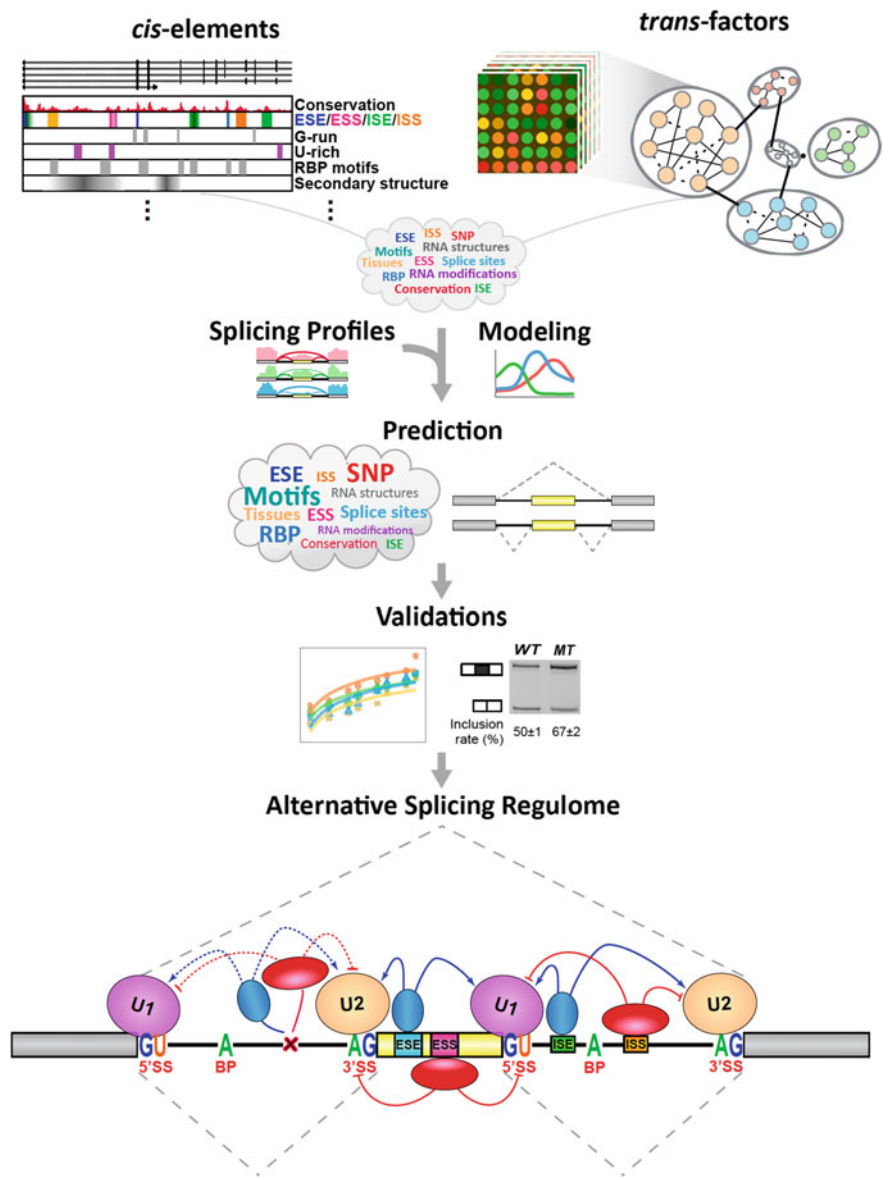
#### 2.2.1.1 High-Throughput Experimental Approaches

The first high-throughput method developed to detect and quantify alternative splicing events was customized microarrays [15–17]. An initial study by Hu et al. [18] used multi-probe design of Affymetrix arrays to detect splicing variants, demonstrating the utility of microarrays for splicing analyses. Later studies [19, 20] developed different techniques to improve the microarray probe design and successfully profiled alternative splicing events and their expression on the genome-wide scale. Johnson et al. [19] used splice junction arrays to probe around 10,000 human multi-exon genes across 52 tissues. Besides the known alternative splicing events, they were also able to discover novel spliced isoforms of many genes. Pan et al. [20] took the focused probe design approach (see review [16]) with three exon body probes and three spliced junction probes for each known alternative splicing event to achieve more sensitive expression quantification. In this study, they were able to globally determine the tissue specificity of alternative splicing events in mouse tissues. Many recent studies adopted different probe designs and microarray platforms to investigate splicing profiles and splicing levels in healthy and disease samples (reviewed in [15–17]).

Since the advent of next-generation sequencing (NGS), RNA-Seq became an essential technology for global studies of alternative splicing (Fig. 2.2a). It provides a means to directly or indirectly sequence the RNA molecules in a high-throughput manner. At present, often-used RNA-Seq methods first convert the RNA sample of interest into cDNAs, which are then made into a sequencing library that consists of short DNA fragments (corresponding to the RNA of interest) flanked by pre-designed adapter oligos. The DNA library is then sequenced from one end (single-end sequencing, or SE) or both ends (paired-end sequencing, or PE) to yield final RNA-Seq reads [21]. The resulting RNA-Seq reads correspond to a snapshot of RNA expression in the respective cellular sample.

RNA-Seq is advantageous in several ways. First, it can detect novel isoforms and alternative splicing events that are not yet annotated [22, 23]. Second, RNA-Seq is not affected by the cross-hybridization problem that confounds many microarray-based studies [21]. Third, RNA-Seq data can provide relatively accurate quantification of levels of gene expression and splicing [21, 24]. Lastly, RNA-Seq provides single-nucleotide information that enables studies of genetic variants [25, 26] and RNA editing sites [27–30], in addition to gene or exon expression. Using RNA-Seq, a large number of alternative splicing events were identified in human and mouse tissues [4, 5].

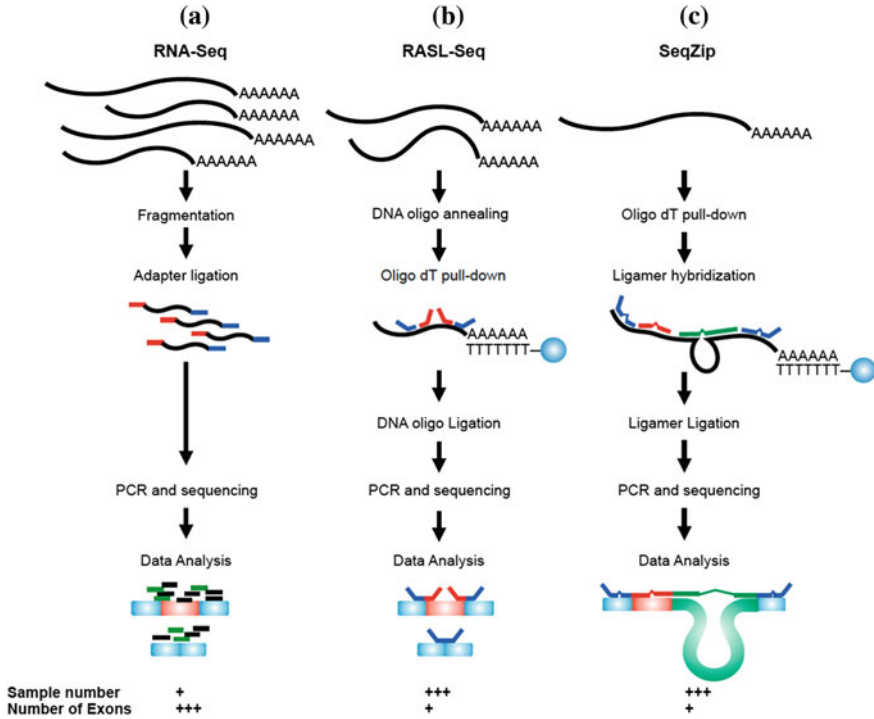
Although RNA-Seq has dramatically improved our knowledge on alternative splicing, there are still remaining challenges to be addressed [21, 31]. RNA-Seq



◀ **Fig. 2.1** Overview of previous studies in alternative splicing regulation. *Cis*-regulatory elements and *trans*-acting factors are key components in the splicing regulatory networks (alternative splicing regulome), which have been actively examined. Combined with global profiles of alternative splicing patterns, bioinformatic models were developed to predict the relative impacts of different regulators and splicing outcome of a given exon. Experimental validations are critical steps to evaluate the accuracy of the predicted splicing regulation. The *bottom* diagram illustrates well-known components of splicing regulation. The *yellow box* represents an alternatively skipped exon, which has ESE and ESS motifs that can be recognized by splicing factors. The flanking introns of this exon harbor ISE and ISS motifs. Interactions between the splicing factors and the core splicing machinery (U1, U2 snRNPs, etc.) are illustrated. Splicing enhancers (ESEs, ISEs) normally promote exon inclusion, which is represented by the *arcs with arrowheads*, whereas splicing silencers (ESSs, ISSs) repress exon inclusion, which is represented by *flat-headed arcs*. Genetic variants may disrupt splicing motifs and alter the binding strength of splicing factors (illustrated by the *x*). Other mechanisms such as RNA modifications or RNA secondary structures may also affect alternative splicing, which are not illustrated in this diagram. ESE: exonic splicing enhancer; ESS: exonic splicing silencer; ISE: intronic splicing enhancer; ISS: intronic splicing silencer; 5'ss: 5' splice site; 3'ss: 3' splice site and BP: branch point

library construction is the first critical step. Different library preparation protocols were developed to study various biological questions and thus have their own merits and limitations [32, 33]. In addition, RNA-Seq library generation protocols often need optimization for specific RNA samples based on sample quality, concentration, and other variables. In large-scale experiments, batch effects in RNA-Seq data may be a critical problem to consider [34], which may mislead study conclusions if not properly accounted for. Finally, RNA-Seq experiments are still costly, especially for studies of alternative splicing. In such applications, reads covering spliced junctions are examined closely to guide the identification and quantification of alternative splicing. Thus, it is highly desirable to have a relatively large number of spliced reads. Often-used settings of RNA-Seq in splicing studies favor PE reads, long read length (e.g., >75 bp), and high sequencing depth (≥100 million PE reads for human samples) [35, 36].

Alternative approaches were developed to address some of the above challenges in RNA-Seq. For example, RNA-mediated oligonucleotide annealing, selection, and ligation with next-generation sequencing (RASL-Seq) allows for RNA-Seq of a limited set of exons in hundreds or thousands of biological samples [37] (Fig. 2.2b). Thus, it is ideal for large-scale analysis of up to 500 exons in complex networks or pathways [37]. The main difference between RNA-Seq and RASL-Seq is the use of oligonucleotides that recognize a specific spliced junction in the latter method. After ligating the pairs of oligos, these specific RNAs are then isolated with biotinylated oligo-dTs and pulled down with streptavidin-coated magnetic beads. A unique barcode for each sample is incorporated during PCR, allowing for pooled sequencing of >1500 samples per lane [37]. Analyzing expression of a limited number of genes in many samples has clinical applications, such as screening for drugs that inhibit splicing events implicated in cancer [38]. One factor of consideration in RASL-Seq is the efficiency and specificity of ligation; Rnl2 was shown to have higher efficiency than T4 ligase [39].



**Fig. 2.2** High-throughput experimental approaches for splicing detection. **a** RNA-Seq, the most popular method for splicing analysis, begins with creating cDNA libraries of fragmented RNA. Then, sequencing adapters are added to make a sequencing library, followed by PCR amplification and sequencing. In data analysis, reads that span spliced exon junctions and those that are located within exon bodies are identified bioinformatically to detect and quantify alternative splicing. This method can provide data for many expressed exons (++++) in the sample of interest. The cost of RNA-Seq is relatively high, which may limit the number of samples (+) that can be analyzed in a specific study. **b** RASL-Seq requires a pair of pre-designed oligonucleotides that recognize specific splice junctions of intact (i.e., unfragmented) mRNA. Biotinylated oligo-dTs with streptavidin-coated magnetic beads are then used to pull down the RNA. Barcode incorporation during PCR allows for pooled sequencing of ~1500 samples per sequencing lane. Compared to RNA-seq, RASL-Seq is ideal for few (up to 500) exons (+) in hundreds or thousands of samples (+++). **c** SeqZip uses DNA “ligamers” to directly sequence long transcript isoforms, causing intermediate regions to loop out. Compared to RNA-Seq and RASL-Seq, SeqZip is specialized for targeting long transcripts

A limitation common to all sequencing-based methods is the sequencing read length, which is typically much shorter than the full-length isoform of long transcripts. Full-length isoforms are thus reconstructed computationally using overlapping reads, though there is always a degree of uncertainty [40, 41]. To overcome this limitation, a new method SeqZip was recently developed [42] (Fig. 2.2c). It uses ~40–60nt DNA “ligamers” that recognize the 5' and 3' ends of single or multiple alternatively spliced exons that may be thousands of nucleotides apart,

causing the intermediate sequence to loop out [42]. Multiple ligamers hybridized to the same transcript are then ligated together, thereby connecting distant exons in the same transcript. Assessing the length and sequence of the DNA ligamers allows for deduction of the full-length isoform. Thus, SeqZip greatly improves the ability to sequence long transcripts.

Aside from the above-mentioned RNA-based approaches, protein-based approaches may be used to identify changes in protein expression resulting from alternative splicing events. Mass spectrometry has been used to identify alternative splicing events in breast cancer [43]. Still, RNA-based approaches are far more commonly used for alternative splicing identification. The choice of experimental method depends on the experimental goal. As sequencing technology improves, so will the ability to identify alternative splicing events.

### **2.2.1.2 Bioinformatic Algorithms for Analyzing Alternative Splicing Using RNA-Seq**

Current bioinformatic methods for analyzing alternative splicing in RNA-Seq can be largely classified into two categories: exon-centric and isoform-centric. Exon-centric approaches directly estimate the splicing level of each exon typically by calculating its percent spliced-in (PSI) [4], a measure of the frequency of exon inclusion among all mature mRNA molecules of the gene (also see reviews [44, 45]). In contrast, isoform-centric methods aim to quantify the abundance of each alternative isoform of the gene, which can be followed by further comparisons to determine differential splicing [46–48].

The benefit of using exon-centric splicing detection is that the type and PSI of each alternatively spliced exon are directly interrogated. Such single-exon information is useful in designing experiments to validate and further examine these events [36, 49]. PSI can be calculated in different ways. First, abundance of reads aligned directly to alternative exon junctions is used, with the exon body reads optionally included [36, 49]. However, it is difficult to precisely estimate the PSI value in cases of complex alternative splicing. To overcome this problem, other tools, such as SplAdder and DiffSplice [50, 51], adopt a splicing graph strategy to capture the complexity of alternative splicing by building a graph of spliced isoforms where nodes represent exons and edges represent spliced introns. Input RNA-Seq data are used to update the alternative path in the graph. The challenge in these approaches is that the splicing graph can be complicated by poorly supported events, so post-filtering is necessary to reduce false positives. In general, exon-centric methods alone do not support identification of novel alternative splicing events due to their requirement of gene annotation.

Instead of focusing on specific splicing events, isoform-centric methods use RNA-Seq to construct isoforms and estimate their expression levels [52–55]. Most tools also utilize the reference genome to guide isoform reconstruction, but others perform *de novo* transcriptome assembly without relying on the reference genome. The latter type is particularly helpful for alternative splicing analyses in species

with poorly annotated genomes. Early isoform-centric methods were developed under the assumption that the read distribution is uniform, though this is rarely the case. New methods are now available to account for RNA-Seq read non-uniformity [56, 57]. Another recent development for isoform-centric analysis is the alignment-free approach, which bypasses the time-consuming alignment step by building a hash index from the reference transcripts using sequence k-mers as keys and applying an expectation maximization algorithm to estimate isoform abundance [46, 47]. This approach speeds up the computational time considerably while maintaining prediction accuracy. However, it remains to be evaluated whether such methods perform well in the presence of sample-specific genetic variants.

Once alternative splicing is identified, both classes of methods provide a means to detect differentially spliced events. The outcome from exon-centric analyses is a list of differentially spliced events that can be directly used for further analysis (e.g., experimental validation, functional interpretation, and regulatory studies). On the other hand, isoform-centric analysis captures the splicing complexity of a series of related events within the same isoform, but further steps are often needed to pinpoint individual splicing events of interest. In Table 2.1, we summarize often-used tools for splicing analysis.

### 2.2.2 *Validation of Alternative Splicing Events*

In silico tools that detect alternative splicing events based on RNA-Seq data usually generate a large number of candidates. A subset of these events should be experimentally validated in vivo or in vitro. Verification experiments for alternative splicing events are readily carried out by reverse transcription followed by PCR (RT-PCR) using primers that target flanking constitutive exons [58]. This strategy works well for alternative splicing events in genes with intermediate or high expression levels. In order to verify lowly expressed events, in vitro minigene expression analysis by RT-PCR can be utilized [59–61]. Compared with in vivo assays, the minigene system is able to validate events regardless of their endogenous expression level. However, since only a limited region flanking the exon of interest can be cloned into the minigene vector, this in vitro approach may not faithfully reproduce in vivo splicing patterns. It should be noted that both types of experiments are considered low-throughput and labor intensive, thus only validating a relatively small number of events.

High-throughput methods for validation of alternative splicing events are in great demand and several such approaches are on the horizon. For example, RT-PCR experiments may be scaled up when used in conjunction with microfluidic devices [62]. In addition, recent methods, such as the “designer exons” approach [63], may be further developed for this purpose. With the rapid technology development in synthetic biology and genome editing, it is likely that high-throughput splicing validation will soon become a reality.



**Table 2.1** Tools for analysis of alternative splicing using RNA-Seq data

Function	Category	Program	URL	Input	Gene annotation
AS prediction	Isoform-centric	Cufflinks	<a href="http://cole-trapnell-lab.github.io/cufflinks/cufflinks">http://cole-trapnell-lab.github.io/cufflinks/cufflinks</a>	SAM or BAM files	No
AS prediction	Isoform-centric	eXpress	<a href="http://bio.math.berkeley.edu/eXpress">http://bio.math.berkeley.edu/eXpress</a>	SAM or BAM files	No
AS prediction	Isoform-centric	Trinity	<a href="http://trinityrnaseq.github.io">http://trinityrnaseq.github.io</a>	FASTQ files	Yes
AS prediction	Isoform-centric	Trans-ABYSS	<a href="http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss">http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss</a>	FASTQ files	Yes
AS prediction	Isoform-centric	Scripture	<a href="http://www.broadinstitute.org/software/scripture">http://www.broadinstitute.org/software/scripture</a>	SAM or BAM files	No
AS prediction	Isoform-centric	RSEM	<a href="http://deweylab.biostat.wisc.edu/rsem">http://deweylab.biostat.wisc.edu/rsem</a>	FASTA or FASTQ files, transcript expression	Yes
AS prediction	Isoform-centric	PennSeq	<a href="http://sourceforge.net/projects/pennseq">http://sourceforge.net/projects/pennseq</a>	UCSC genome browser gene annotation, SAM files	Yes
AS prediction	Isoform-centric	RNA-Skim	<a href="http://www.csbio.unc.edu/rs">http://www.csbio.unc.edu/rs</a>	Specialized transcriptome FASTA, RNA-Seq FASTQ files	Yes
AS prediction	Isoform-centric	Sailfish	<a href="http://www.cs.cmu.edu/~ckingsf/software/sailfish">http://www.cs.cmu.edu/~ckingsf/software/sailfish</a>	<i>k</i> -mer size, RNA-Seq in FASTA or FASTQ, transcriptome in GTF	Yes
AS prediction	Isoform-centric	Sequrio	<a href="http://fafner.meb.ki.se/biostatwiki/sequrio">http://fafner.meb.ki.se/biostatwiki/sequrio</a>	BAM files	Yes
AS prediction	Exon-centric	SplAdder	<a href="https://github.com/ratschlab/spladder">https://github.com/ratschlab/spladder</a>	GFF annotation, BAM files	Yes
AS prediction	Exon-centric	SpliceTrap	<a href="http://nulai.cshl.edu/splicetrap/doc/help.html">http://nulai.cshl.edu/splicetrap/doc/help.html</a>	Gene annotation in BED or GTF format, TXdb exon isoform database, FASTA or FASTQ files	Yes
AS prediction	Exon-centric	ESFinder	<a href="http://mlg.hit.edu.cn/ybai/ES/ESFinder.html">http://mlg.hit.edu.cn/ybai/ES/ESFinder.html</a>	GTF annotation, UCSC genome browser AS events, BAM files	Yes
AS prediction	Exon-centric	SplicePie	<a href="https://github.com/pulyakhina/splicing_analysis_pipeline">https://github.com/pulyakhina/splicing_analysis_pipeline</a>	GTF annotation, BAM and BAM index files	Yes
AS prediction	Both	MISO	<a href="https://miso.readthedocs.org/en/fastmiso">https://miso.readthedocs.org/en/fastmiso</a>	GFF annotation, BAM files	Yes

(continued)

Table 2.1 (continued)

Function	Category	Program	URL	Input	Gene annotation
DAS	Isoform-centric	CuffDiff 2	<a href="http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html">http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html</a>	GFF/GTF annotation, SAM files	Yes
DAS	Isoform-centric	IUTA	<a href="http://www.niehs.nih.gov/research/resources/software/biostatistics/iuta/index.cfm">http://www.niehs.nih.gov/research/resources/software/biostatistics/iuta/index.cfm</a>	GFF annotation, BAM files	Yes
DAS	Isoform-centric	SplicingCompass	<a href="http://www.ichip.de/software/SplicingCompass.html">http://www.ichip.de/software/SplicingCompass.html</a>	Read coverage in GFF, BED files	Yes
DAS	Isoform-centric	rSeqDiff	<a href="http://www-personal.umich.edu/~jianghui/rseqdiff">http://www-personal.umich.edu/~jianghui/rseqdiff</a>	“sampling_rates” files from rSeq [167], gene expression files	Yes
DAS	Isoform-centric	FDM	<a href="http://csbio-linix001.cs.unc.edu/nextgen/software/FDM">http://csbio-linix001.cs.unc.edu/nextgen/software/FDM</a>	GTF annotation, SAM files	Yes
DAS	Isoform-centric	rDiff	<a href="http://bioweb.me/rdiff">http://bioweb.me/rdiff</a>	GFF annotation, BAM files	Yes
DAS	Exon-centric	DEXSeq	<a href="http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html">http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html</a>	GFF/GTF annotation, SAM files	Yes
DAS	Exon-centric	DSGSeq	<a href="http://bioinfo.au.tsinghua.edu.cn/software/DSGseq">http://bioinfo.au.tsinghua.edu.cn/software/DSGseq</a>	BAM and BED files	Yes
DAS	Exon-centric	DiffSplice	<a href="http://www.netlab.uky.edu/p/bioinfo/DiffSplice">http://www.netlab.uky.edu/p/bioinfo/DiffSplice</a>	Parsed SAM files, program configuration files	No
DAS	Exon-centric	dSpliceType	<a href="http://dsplice.type.sourceforge.net">http://dsplice.type.sourceforge.net</a>	GFF annotation, read coverage in bedgraph format, junction files in BED format	Yes
DAS	Exon-centric	MATS/rMATS	<a href="http://maseq-mats.sourceforge.net">http://maseq-mats.sourceforge.net</a>	GFF annotation, FASTQ or BAM files, Bowtie index	Yes
DAS	Exon-centric	rSeqNP	<a href="http://www-personal.umich.edu/~jianghui/rseqnp">http://www-personal.umich.edu/~jianghui/rseqnp</a>	Transcript expression files	No
DAS	Both	MISO	<a href="https://miso.readthedocs.org/en/fastmiso">https://miso.readthedocs.org/en/fastmiso</a>	GFF annotation, BAM files	Yes

(continued)

Table 2.1 (continued)

Function	Category	Program	URL	Input	Gene annotation
DAS	Both	SUPPA	<a href="https://bitbucket.org/regulatorygenomicsupf/suppa">https://bitbucket.org/regulatorygenomicsupf/suppa</a>	GTF annotation, transcript expression files	Yes
Spliced mapper	NA	HMMSplicer	<a href="http://derisilab.ucsf.edu/software/hmmsplicer">http://derisilab.ucsf.edu/software/hmmsplicer</a>	FASTQ files, Bowtie index	Yes
Spliced mapper	NA	PASSion	<a href="https://trac.nbic.nl/passion">https://trac.nbic.nl/passion</a>	FASTQ files, SMALT index	Yes
Spliced mapper	NA	PASTA	<a href="http://www.biotech.ufl.edu/cores/bioinformatics/dibig/dibig-software/pasta">http://www.biotech.ufl.edu/cores/bioinformatics/dibig/dibig-software/pasta</a>	FASTQ files, Bowtie index	Yes
Spliced mapper	NA	OLego	<a href="http://zhanglab.c2b2.columbia.edu/index.php/OLego">http://zhanglab.c2b2.columbia.edu/index.php/OLego</a>	FASTA or FASTQ files, BWT index	Yes
Spliced mapper	NA	TrueSight	<a href="http://bioen-compbio.bioen.illinois.edu/TrueSight/">http://bioen-compbio.bioen.illinois.edu/TrueSight/</a>	FASTQ files, Bowtie index	Yes
Spliced mapper	NA	UnSplicer	<a href="http://opal.biology.gatech.edu/paul/unsplicer/index.htm">http://opal.biology.gatech.edu/paul/unsplicer/index.htm</a>	FASTQ files, Bowtie index	Yes
Spliced mapper	NA	JAGuaR	<a href="http://www.bcgsc.ca/platform/bioinfo/software/jaguar">http://www.bcgsc.ca/platform/bioinfo/software/jaguar</a>	FASTQ files, BWA index	Yes
Spliced mapper	NA	Rail-RNA	<a href="https://github.com/nellore/rail">https://github.com/nellore/rail</a>	FASTQ files, Bowtie index	Yes

AS: alternative splicing; DAS: differential alternative splicing; Spliced mapper: read aligners specialized in mapping junction reads

## 2.3 Methodologies for Studies of Splicing Regulation

Pre-mRNA splicing is regulated by a large number of *cis*-elements and *trans*-acting factors. In this section, we will review the bioinformatic and experimental approaches for the identification and analysis of splicing regulatory mechanisms.

### 2.3.1 *Cis-Regulation of Alternative Splicing*

#### 2.3.1.1 Splice Site Consensus Sequence

Splice site sequences are among the best-characterized *cis*-elements in splicing regulation, owing to the simplicity of their identification. Each internal exon is flanked by a 5'ss and a 3'ss. Thus, splice site sequences can be easily collected based on gene annotation. The majority of human exons are flanked by the GU-AG canonical sequences. However, the splice site signals normally involve a much longer sequence motif, which confers specificity and a dynamic range of splice site strength. Using known splice site sequences as training data, many algorithms were developed to predict splice site strength (see reviews [64, 65]). The most intuitive model is the position weight matrix (PWM), which is straightforward to implement but fails to consider the positional dependency between nucleotides in the splice site [66]. Other algorithms adopt more sophisticated probabilistic models such as neural networks or maximum entropy to more accurately estimate the splice site scores [67, 68].

#### 2.3.1.2 Branch Point Sequences (BPSs)

The prediction of BPS is challenging because its location in the intron can be highly variable. For example, a BPS may be close to the 3'ss (~40nt upstream) or 100–400nt upstream of the 3'ss in the AG exclusion zone (AGEZ) [69]. Additionally, the BPS motif is highly degenerate [70] and multiple potentially functional BPSs may exist in a particular intron. A number of bioinformatic methods were developed to identify BPS and evaluate their strength. Human Splice Finder [66] uses PWMs and the algorithm proposed by Gooding et al. [69] to search for BPS candidates in a limited region. Another predictive approach makes use of sequence conservation and partial sequence complementarity of U2 snRNA to the BPS [71, 72]. A recent study showed that using machine learning methods such as support vector machines together with polypyrimidine and other sequence information could increase accuracy in BPS prediction [73]. Pastuszak et al. took advantage of the fact that Splicing Factor 1 (SF1) recognizes BPSs and restricted their motif analysis to sites with high SF1 binding affinity to predict BPS with relatively high accuracy [74].

Recently, a few studies used the NGS technology to identify BPS globally. In the RNA-Seq data, a minority of reads may derive from the junction of the 5'ss-branch point of the intron lariat. A search for such reads has led to successful identification of hundreds of BPS in human RNA-Seq data sets [75, 76]. The advantage of these approaches is that they do not require prior knowledge about the BPS locations or sequences. However, one drawback is that lariat reads are very rare among those generated from standard RNA-Seq libraries. Thus, very deeply sequenced data sets are needed to obtain adequate lariat read coverage. Another NGS-based method, called CaptureSeq [77], was applied recently to identify BPS [78]. In this method, tiling arrays were designed that contain oligonucleotide probes to target the 5'ss-branch point junctions [78]. cDNAs from the RNA samples of interest were then hybridized, eluted, and sequenced. As a complementary approach, RNase R digestion was applied to enrich for reads containing BPS without requiring pre-designed arrays. This study identified >50,000 human BPS in >10,000 genes, which enabled further investigation of global features of this class of splicing regulatory signal [78].

### 2.3.1.3 Splicing Regulatory Elements

Besides the core splicing signals, a large number of motifs in the exons or introns can also regulate splicing (Fig. 2.1) [8]. Identification and characterization of these SREs are instrumental to the understanding of splicing regulatory mechanisms. In general, genome-wide experimental or bioinformatic screens have been designed to identify SREs. Wang et al. developed the first large-scale screen of ESSs using splicing reporter assays in cultured cells [59]. This effort successfully identified hundreds of ESS sequences and shed light on the global properties of these elements. Later, a number of other experimental screens were carried out to identify different types of SREs [79–82]. These studies greatly expanded the catalog of known or predicted SREs without the associated *trans*-factors necessarily identified. Other experimental methods that pinpoint SREs for known splicing factors will be discussed later.

In addition to the experimental approaches, bioinformatic methods are also essential to SRE studies. Fairbrother et al. developed a motif comparison approach, RESCUE-ESE, to identify ESEs by evaluating motif enrichment correlated with different features of splicing [83]. Similar principles were applied later to identify other types of SREs [84, 85]. A myriad of other bioinformatic methods were also developed for this purpose, such as those based on comparative genomics [86], PWMs [87], or machine learning techniques [88–91].

With the increasing number of SREs, a great deal of effort was dedicated to understand the functional interaction among different elements and their context-dependent roles in splicing regulation. For example, Bayesian networks were used to study coevolutionary relationships of SREs in eukaryotes that reflect functional interaction [60]. Bioinformatic and statistical methods, combined with experimental approaches, were used to infer combinatorial function of different

types of SREs [92–94]. The function of individual motifs (corresponding to one splicing factor) was studied in detail via bioinformatic modeling and analysis to reveal their context-dependent function globally [61, 95–97] (refer to [98] for a detailed review of this topic).

### 2.3.2 Genetic Variants Associated with Splicing

Genetic variants [such as mutations or single-nucleotide polymorphisms (SNPs)] play important roles in gene regulation because they can potentially alter *cis*-regulatory motifs. Previous studies estimated that 15–60 % of point mutations that result in human genetic diseases disrupt splicing [10, 99–102]. In recent years, exciting progress has been made in analyzing the involvement of genetic variants in modulating alternative splicing, which is reviewed in this section.

#### 2.3.2.1 Splicing QTLs

Splicing quantitative trait loci (sQTL) analysis is an often-used method to identify SNPs associated with splicing phenotypes. In this method, the correlation between SNP genotypes and exon inclusion levels is examined using different means, ranging from simple linear correlation to model-based analysis [103–105]. Early sQTL studies used microarrays to detect isoform or exon expression levels, which is rapidly replaced by RNA-Seq-based analysis. However, this method requires a large number of samples to achieve adequate statistical power. In addition, sQTL analyses only deduce correlative relationships, without the capability of pinpointing the causal SNP for splicing alteration.

#### 2.3.2.2 Machine Learning-Based Methods

In contrast to sQTLs, methods based on machine learning principles aim to predict the functional (causal) SNP that modulates alternative splicing. Different types of machine learning or statistical methods were adopted for this purpose [106–108]. One study used a random forest-based strategy and predicted exonic splicing-altering variants [106]. Another study developed a splicing code where “code quality” was optimized using information theory on a large number of features [109]. This splicing code was applied to predict genetic variants that may alter splicing [108–110]. One common challenge to such approaches is the limited availability of training data sets that should include experimentally validated SNPs with confirmed function in splicing and those that are known to have no influence on splicing. To overcome this problem, previous studies used disease-causing exonic mutations from existing databases as positive training data set and common SNPs in the general population as negative data set (assuming they do not affect

splicing) [106, 107]. In contrast, the splicing code-based studies used human RNA-Seq data of different tissues to derive the code, without the need of direct model training using splicing-related variants [108–110].

### 2.3.2.3 Allele-Specific Alternative Splicing

To infer genetic regulation of alternative splicing, another powerful approach is built upon allele-specific expression (ASE) of genetic variants. ASE refers to the biased expression of the two alleles of a variant in diploid cells. RNA-Seq data provide single-nucleotide information that is appropriate for ASE studies. One advantage of ASE analysis is that the two alleles of a variant serve as within-sample controls of each other, which naturally eliminates the environmental and *trans*-acting effects that might alter splicing patterns or introduce variance in the data [111]. Nevertheless, one challenge in using RNA-Seq for ASE analysis lies in the step of read mapping. It is now clear that standard mapping methods induce a mapping bias that favors the reference allele of the genetic variant because the reference genome is utilized in mapping [112, 113]. Various strategies were developed to reduce this type of bias [27, 28]. Once ASE patterns are identified, they can be further analyzed to detect allele-specific alternative splicing events, as proposed in [25]. While sQTL studies and machine learning methods necessitate many data points for correlative analysis or model training, the ASE-based approach can predict splicing-associated genetic variants using RNA-Seq data of a single individual. Thus, it is both cost-effective and computationally inexpensive.

## 2.3.3 *Trans-acting Regulators of Alternative Splicing*

### 2.3.3.1 Methods for Identification of Splicing Factors

Recently, an increasing number of RNA-binding proteins (RBPs) have been identified as regulators of splicing [98]. However, the associated splicing factors are not yet known for a large number of SREs identified using the experimental or bioinformatic methods described above. To this end, a modified RNA affinity purification method was used to identify *trans*-factors for known SREs [81, 82, 114]. In addition, in vivo siRNA screens targeting known splicing factors were also used to reveal the *trans*-factor for specific SREs [79, 115–117].

Previous efforts were also dedicated to predict or validate proteins with splicing regulatory activity [98]. For example, a computational pipeline was designed to search for proteins with splicing factor-like properties, which led to the discovery of an SR-related protein with important function in neuronal tissues [118]. Given a pool of RBPs, a previous study screened for splicing-related ones by examining the

correlation of their expression with changes in levels of alternative splicing [119]. Combined with motif analysis, the authors successfully identified known and novel splicing factors.

### 2.3.3.2 Methods for Identifying Binding Motifs of Splicing Factors

Given a splicing factor or RBP, a number of experimental methods were developed to identify their binding motifs globally. These methods can be largely categorized into two classes depending on their *in vitro* or *in vivo* nature. The Systematic Evolution of Ligands by EXponential enrichment (SELEX) approach is one of the *in vitro* methods [120]. SELEX was applied to identify ESEs and other SREs in several studies [121]. Recently, this method was combined with microarray assays to increase the throughput [122]. Another *in vitro* method called RNAcompete uses *in vitro* transcribed RNA (structured or unstructured) for pull-down with an RBP of interest, followed by microarray analysis of the bound RNA [123]. Binding motifs of over 200 RBPs were determined by this method [119]. More recently, a new *in vitro* method called RNA Bind-n-Seq (RBNS) was developed to improve quantification of the sequence and structural specificity of RBPs [124]. Besides canonical motifs, RBNS identified additional near-optimal binding motifs, which were shown to be functional *in vivo* [124].

To identify global *in vivo* binding sites of RBPs, the most widely used method is UV cross-linking and immunoprecipitation (CLIP) followed by sequencing (CLIP-Seq) [125]. Variations of this method are also used for different applications, including high-throughput sequencing of RNAs isolated by CLIP (HITS-CLIP) [126], photoactivatable-ribonucleoside-enhanced cross-linking and precipitation (PAR-CLIP) [127], and individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) [128]. Detailed discussions of these methods are provided by previous reviews [129, 130]. Briefly, CLIP-based methods have relatively high sensitivity and specificity compared to RNA immunoprecipitation alone. However, the cross-linking efficiency is generally limited in regular CLIP, which is improved in PAR-CLIP via the usage of 4-thiouridine, a photo-activated nucleotide. Deletions, substitutions, or insertions usually occur near the cross-linking sites in CLIP-Seq/HITS-CLIP [131], whereas T-to-C substitutions are observed near the cross-linking sites in PAR-CLIP. These mutations can serve as diagnostic features to pinpoint binding sites. Nonetheless, accurate read mapping tolerating such mutations is challenging. Currently, bioinformatic tools are designed to handle read mapping, cluster calling, and motif enrichment. In the future, development of tools that integrate these basic analyses with RNA secondary structure, evolutionary conservation, and *in vitro* binding data will tremendously facilitate a systematic understanding of protein–RNA interaction.



Notably, the ENCODE Consortium has devoted great efforts to generate CLIP-Seq data of about 200 RBPs. In addition, shRNA knockdown experiments of each RBP are carried out followed by RNA-Seq in cultured cells (K562 and HepG2). These data sets will facilitate identification of splicing regulatory motifs, analysis of splicing factor functions, and generation of global regulatory maps of these RBPs.

### 2.3.4 *Splicing Code*

While most existing methods focus mainly on one or a few aspects of splicing regulation, Barash et al. took a step further to assemble a “splicing code” by integrating hundreds of RNA features and the alternative splicing patterns of a wide panel of tissues [109]. This model takes as input exon sequences of interest and their flanking introns, and recursively selects for features and parameters that maximize the “code quality.” The code was later improved using Bayesian neural networks on an expanded list of RNA features [132, 133] and applied to predict splicing-altering disease mutations [108]. The above work mainly focused on analysis of alternatively skipped exons. A more recent splicing code was designed to identify RNA sequence features that categorize several major classes of alternative splicing, including exon skipping, alternative 5’ss, and alternative 3’ss exons [134]. This work demonstrated that RNA sequence features (splice sites, conservation levels, and exon/intron architecture) confer strong discriminatory contributions to classify different types of splicing.

Current versions of the splicing code are not able to predict absolute levels of exon inclusion, but rather focus on predictions of relative changes in splicing across tissues or in the presence of genetic mutations. Future development of the splicing code could be empowered by consideration of regulatory networks of multiple splicing factors, epigenetic influence, and kinetic aspects of splicing, some of which are discussed below.

### 2.3.5 *Useful Databases*

Over the years, the splicing community has built many databases and Web resources to include data on global profiling of alternative splicing and systematic analysis of splicing regulatory mechanisms. Table 2.2 summarizes some of these resources ranging from catalogs of alternative splicing events to disease-related mutations that affect splicing.

**Table 2.2** Databases (DB) of alternative splicing events and software programs (PR) for studies of splicing regulation

Type	Category	Database	URL	Function	Input	Notes
DB	AS events	DBASS	<a href="http://www.dbass.org.uk">http://www.dbass.org.uk</a>	Catalogs splice site mutations and diseases	Chosen from user drop-down menu	Includes DBASS3 and DBASS5
DB	Data	HGMD	<a href="http://www.hgmd.cf.ac.uk/ac/index.php">http://www.hgmd.cf.ac.uk/ac/index.php</a>	Provides disease-associated variants	–	Requires license to access the full database
DB	Data	ENCODE	<a href="https://www.encodeproject.org">https://www.encodeproject.org</a>	Provides various NGS data sets	Chosen from user drop-down menu	–
DB	Data	RBPDB	<a href="http://rbpdb.ccb.utoronto.ca">http://rbpdb.ccb.utoronto.ca</a>	Provides RBP PWMs	–	–
DB	Data	CisBP-RNA	<a href="http://cisbp-ma.ccb.utoronto.ca">http://cisbp-ma.ccb.utoronto.ca</a>	Provides RBP PWMs	–	–
DB	Data	CLIPdb	<a href="http://lulab.life.tsinghua.edu.cn/clipdb">http://lulab.life.tsinghua.edu.cn/clipdb</a>	Catalogs references for RBP studies and CLIP data sets	RBP names, cell types, species, or technology types	Also supports database browsing without requiring input
DB	Data	CLIPZ	<a href="http://www.clipz.unibas.ch">http://www.clipz.unibas.ch</a>	Provides CLIP data sets with target predictions of RBPs and miRNAs	Chosen from user drop-down menu	Require account sign-up; review [168]
DB	Data	doRNA 2.0	<a href="http://dorina.mdc-berlin.de">http://dorina.mdc-berlin.de</a>	Provides CLIP data sets with RBP sites annotations	Chosen from user drop-down menu	Review [168]
DB	Data	StarBase V2.0	<a href="http://starbase.sysu.edu.cn/index.php">http://starbase.sysu.edu.cn/index.php</a>	Provides CLIP data sets with annotations	Chosen from user drop-down menu	Review [168]
DB	Tools	OMICtools	<a href="http://omictools.com">http://omictools.com</a>	Databases of genomic, transcriptomic, proteomic, and metabolomic tools	Tool names, analysis type, or Web-browsing	–
DB	Regulation	RegulomeDB	<a href="http://regulomedb.org">http://regulomedb.org</a>	Displays various regulatory tracks for the input sequences	dbSNP IDs, genomic coordinates in BED files, VCF files, or GFF3 files	Links regulation to GWAS: <a href="http://regulomedb.org/GWAS/index.html">http://regulomedb.org/GWAS/index.html</a>

(continued)

Table 2.2 (continued)

Type	Category	Database	URL	Function	Input	Notes
DB	Regulation	RegRNA2.0	<a href="http://regrna2.mbc.nctu.edu.tw">http://regrna2.mbc.nctu.edu.tw</a>	Displays various regulatory tracks for the input sequences	RNA sequences in FASTA format	–
DB	Regulation	Brain RNA-Seq	<a href="http://web.stanford.edu/group/barres_lab/brain_rnaseq.html">http://web.stanford.edu/group/barres_lab/brain_rnaseq.html</a>	Provides FPKM data in various brain cells; compares gene enrichment across available cell types	Gene name, or cell types	Done in mouse; also provides data browsing
DB	Regulation	rSNPBase	<a href="http://rsnp.psych.ac.cn">http://rsnp.psych.ac.cn</a>	Provides regulatory annotation for SNPs	SNP ID or gene names	–
PR	Regulation	CRYP-SKIP	<a href="http://cryp-skip.img.cas.cz">http://cryp-skip.img.cas.cz</a>	Predicts splice site mutations	Nucleotide sequence in FASTA format	–
PR	Regulation	EX-SKIP	<a href="http://ex-skip.img.cas.cz">http://ex-skip.img.cas.cz</a>	Compares a SNV's impact on ESE/ESS to induce exon skipping	Two exonic sequences in FASTA format	Maximum 4000nt per submission
PR	Regulation	Human Splicing Finder	<a href="http://www.umd.be/HSF">http://www.umd.be/HSF</a>	Combines 12 different algorithms to predict mutations' impact on <i>cis</i> -regulatory elements	Chosen from user drop-down menu	–
PR	Regulation	MaxEntScan	<a href="http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html">http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html</a> ; <a href="http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html">http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html</a>	Predicts splice site strength	Splice site sequences in FASTA format	Web-based or command line-based analyses available

(continued)

Table 2.2 (continued)

Type	Category	Database	URL	Function	Input	Notes
PR	Regulation	WASP	<a href="http://genes.toronto.edu/wasp">http://genes.toronto.edu/wasp</a>	Predicts AS exons and the potential regulation codes	RNA sequence in FASTA format or genomic coordinates in BED format	Maximum 10 input exons per query
PR	Regulation	AVISPA	<a href="http://avispa.biociphers.org">http://avispa.biociphers.org</a>	Predicts AS exons and the potential regulation codes	FASTA or BED files containing a single putative AS exon or cassette exon triplet	–
PR	Regulation	SPANR	<a href="http://tools.genes.toronto.edu">http://tools.genes.toronto.edu</a>	Predicts SNVs effects on cassette exons	Maximum 40 SNVs per file in tab-delimited VCF format	This tool was designed for detecting exon-skipping events, so it may or may not work for other AS types

## 2.4 Ongoing Questions

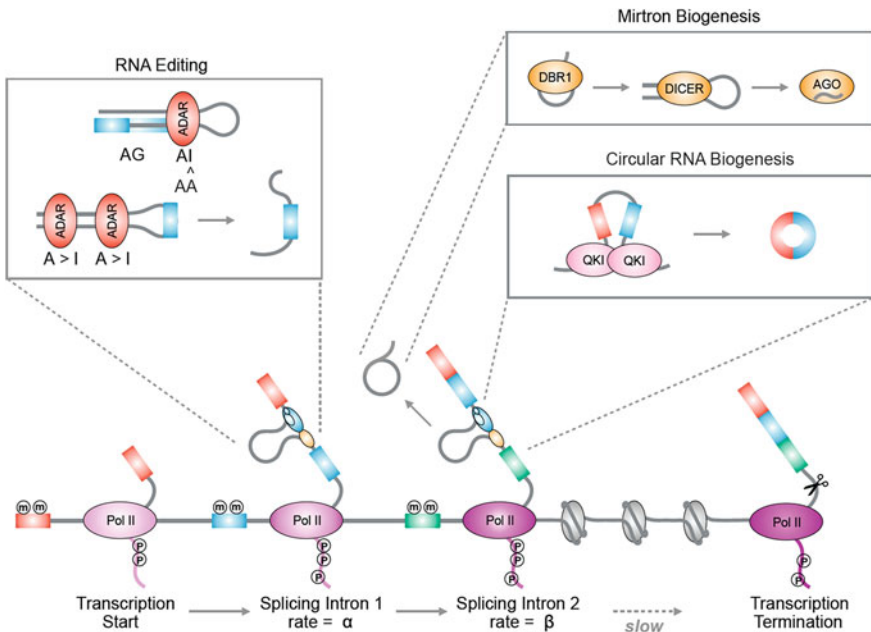
### 2.4.1 *Gene Expression Kinetics and Co-Transcriptional Splicing*

With the advent of RNA-Seq and related methodologies described previously in this chapter, it is now possible to study kinetics of gene expression and splicing on the global scale. It was recently shown that several steps in RNA processing often, but not always, occur co-transcriptionally, including capping, splicing, and polyadenylation, allowing for efficient and accurate pre-mRNA maturation (reviewed in [121, 135, 136]). In particular, co-transcriptional splicing depends on the rate of RNA Pol II elongation with the idea that slower elongation allows more time for splicing to complete. Pol II elongation can be affected by nucleosome positioning, DNA methylation, histone modifications, and chromatin remodeling [137, 138] (Fig. 2.3). Additionally, the C-terminal domain of RNA Pol II can be post-translationally and reversibly modified to guide interactions with different proteins involved in RNA processing. Thus, chromatin modifications, transcription, and splicing are all interconnected processes [136, 137].

To study dynamic regulation of gene expression and/or co-transcriptional splicing, nascent RNA must be captured. Modified RNA-Seq methods such as genomic run-on sequencing (GRO-Seq) or sequencing of 4-thiouridine-labeled RNA may be analyzed in conjunction with RNA-Seq [139–141]. Additionally, cell fractionation and selection of non-polyadenylated RNA in the chromatin fraction may be used. Recently, a native elongating transcript sequencing (NET-Seq) approach was used by two groups to identify spliceosome-mediated cleavage, Pol II dynamics related to splicing, and antisense transcription [142, 143]. Figure 2.3 illustrates co-transcriptional splicing and other events described below including RNA editing, mirtron biogenesis, and circRNA biogenesis.

### 2.4.2 *RNA Modifications*

RNA modifications such as methylation (primarily N<sup>6</sup>-methyladenosine, or m6A) and RNA editing were not extensively studied until recently. m6A, originally identified in tRNAs, rRNAs, and snoRNAs, was recently shown to be widespread in mRNAs with potential impact on splicing, mRNA degradation, and RNA secondary structures [144, 145]. The most prevalent form of RNA editing is the conversion of adenosine to inosine (A-to-I) via deamination, typically in double-stranded RNA (dsRNA) regions by adenosine deaminases acting on RNA (ADARs) (Fig. 2.3). In order for editing to affect splicing, it is expected to occur before splicing is completed. Indeed, several lines of evidence suggest editing

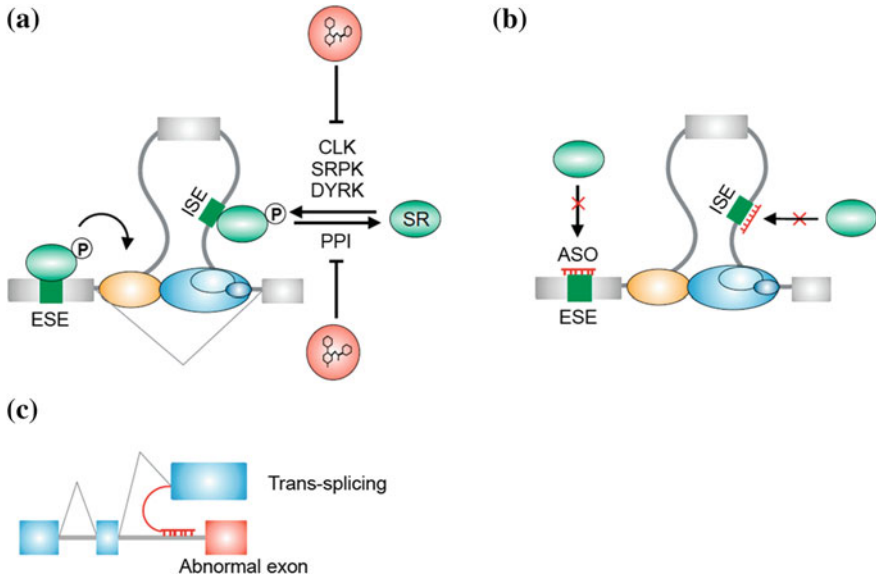


**Fig. 2.3** Co-transcriptional splicing and related RNA products. Co-transcriptional splicing of two introns with splicing rates  $\alpha$  and  $\beta$  is shown. The following epigenetic factors are illustrated: DNA methylation (m) enrichment in exons [138], dynamic phosphorylation (P) of the C-terminal domain of RNA Pol II [137], and nucleosomes slowing down Pol II transcription. Splicing coupled with RNA editing and the biogenesis of mirtrons and circRNAs are shown in the insets. RNA editing can generate new splice sites (e.g., changing A to I may create a new AG 3'ss, RNA editing inset, top [146]) and prevent circRNA biogenesis (RNA editing inset, bottom [151]). Mirtrons are derived from lariats that are debranched by DBR1 and processed by DICER (mirtron biogenesis inset [147]). QKI regulates the production of a subset of circRNAs (circRNA Biogenesis inset [155])

precedes splicing (reviewed in [146]), although exceptions do exist. These findings are only the beginning of a new era of functional and mechanistic studies of RNA modifications.

### 2.4.3 Splicing Generates Other RNA Species

Although introns are typically degraded after removal, certain introns can also be further processed to generate other RNA species. For example, biogenesis pathways of snoRNAs, mirtrons, and simtrons rely on intron splicing (reviewed in [147]). Whereas canonical miRNA biogenesis depends on the microprocessor (DGCR8 and DROSHA), mirtrons depend on lariat debranching (Fig. 2.3) and simtrons depend on U1 snRNP. Another RNA species underappreciated until recently are



**Fig. 2.4** Therapeutic approaches to modulate splicing. **a** Small molecule therapy. Phosphorylation or dephosphorylation of SR proteins are regulated by CDC2-like kinase (CLK), dual-specificity tyrosine-(Y)-phosphorylation-regulated kinase (DYRK), SR protein kinase (SRPK), and protein phosphatase-1 (PP1). Inhibitors of these kinases and phosphatase affect the associated splicing events. **b** Antisense oligonucleotides therapy. SR protein or other splicing factor binding sites can be blocked by ASO to achieve specific alternation of splicing. **c** Trans-splicing therapy. ASO linked to a restoring normal exon can rescue an abnormal splicing event that may result due to multiple mutations

circular RNAs (circRNAs) (reviewed in [148, 149]). It was shown that biogenesis of certain circRNAs depends on intronic sequence content [150–152], which may compete with pre-mRNA splicing [153]. Additionally, circRNAs can contain both exons and introns, and two of these were shown to regulate gene expression [154]. The splicing factor QKI was shown to regulate production of many circRNAs (Fig. 2.3) [155]. The biogenesis and functions of circRNAs are currently under active investigation.

#### 2.4.4 Global Misregulation of Splicing in Disease

Since splicing is required for RNA maturation, misregulation of splicing may lead to disease states [11]. In addition to well-known splicing diseases, such as myotonic dystrophy [156], there are several examples of point mutations in specific genes that cause splicing misregulation (reviewed in [121, 157]). Furthermore, global splicing

misregulation also characterizes some diseases such as cancer. The Cancer Genome Atlas (TCGA, [www.cancergenome.nih.gov](http://www.cancergenome.nih.gov)) provides a wealth of genomic data from cancer patients and controls, allowing for the study of global splicing alterations within and across cancer types [158, 159]. Splicing abnormalities were also shown in autistic brains [160]. Although splicing alterations in cancer are well established, it is difficult to identify the mechanistic cause and functional significance of these events, especially considering that up to hundreds of RBPs may be involved in the regulation of thousands of alternative splicing events in both normal and disease states [121, 157]. In the future, an understanding of the causes and functional consequences may lead to splicing-targeted therapeutics.

## 2.5 Splicing as a Therapeutic Target

Given the critical roles of splicing misregulation in disease, a number of strategies are under development to therapeutically correct aberrant splicing events. First, small molecules can be used to directly modulate the activity of splicing factors [161]. The advantage of this method is the ease of delivery and the potential for individual-specific dosage control. As examples, small molecule inhibitors were examined that target SR protein kinases (SRPKs), CDC2-like kinases (CLKs), or protein phosphatase-1 (PP1), which can then modulate phosphorylation of SR proteins (Fig. 2.4). However, such inhibitors often have off-target effects and affect splicing of many genes.

A more targeted approach involves usage of antisense oligonucleotides (ASO), reverse complementary sequences that bind to target mRNA sequences. Because ASOs are sequence-specific, they can block binding of splicing factors at specific loci and modulate alternative splicing. For example, aberrant splicing events caused by an intronic mutation in the human  $\beta$ -globin gene were corrected by ASO treatment in a  $\beta$ -thalassemia mouse model [162]. In addition, clinical trials are underway for ASO-based therapy of Duchenne muscular dystrophy and spinal muscular atrophy [163]. Although ASO therapy overcomes the nonspecificity issue of small molecules, their delivery is relatively difficult. Another method, trans-splicing, is an effective strategy for repairing multiple mutations in exons or transcripts. Also referred to as Spliceosomal-mediated RNA trans-splicing (SMaRT) [164], this method can replace the entire mRNA sequence 5' or 3' of a target splice site by trans-splicing between an ASO and the endogenous RNA [165]. This approach was proposed as a therapy for  $\beta$ -thalassemia to replace the first exon of the  $\beta$ -globin gene resulting from aberrant splicing [166]. However, the delivery of trans-splicing therapy is also challenging, as it necessitates incorporation of DNA vectors to cells (10).



## 2.6 Conclusions

In recent years, technological advances brought a fundamental shift in our approaches to splicing-related questions. Global analyses that combine high-throughput experimental assays and bioinformatic methods are becoming indispensable. As a result, numerous novel insights have been revealed regarding the landscape of alternative splicing and the regulatory mechanisms of splicing in various cell types. These global discoveries constitute a foundation for further mechanistic and functional studies in model systems and translational research. However, there still exist many challenges in handling high-throughput experiments and data analysis. We expect that these challenges will be addressed via methodology development and standardization, which will further catalyze exciting discoveries in splicing research.

## References

1. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*. 1977;74(8):3171–5.
2. Chow LT, Gelinias RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*. 1977;12(1):1–8.
3. Wahl MC, Will CL, Luhrmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell*. 2009;136(4):701–18.
4. Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470–6.
5. 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.
6. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010;463:457–63.
7. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*. 2000;101(6):671–84.
8. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*. 2005;6(5):386–98.
9. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet*. 2011;12(10):715–29.
10. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet*. 2007;8(10):749–61.
11. Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell*. 2009;136(4):777–93.
12. Poulos MG, Batra R, Charizanis K, Swanson MS. Developments in RNA splicing and disease. *Cold Spring Harb Perspect Biol*. 2011;3(1):a000778.
13. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*. 2003;72:291–336.
14. Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol*. 2004;24(24):10505–14.

15. Lee C, Roy M. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.* 2004;5(7):231.
16. Cuperlovic-Culf M, Belacel N, Culf AS, Ouellette RJ. Microarray analysis of alternative splicing. *OMICS.* 2006;10(3):344–57.
17. Blencowe BJ. Alternative splicing: new insights from global analyses. *Cell.* 2006;126(1):37–47.
18. Hu GK, Madore SJ, Moldover B, Jatkoa T, Balaban D, Thomas J, Wang Y. Predicting splice variant from DNA chip expression data. *Genome Res.* 2001;11(7):1237–45.
19. Johnson JM, Castle J, Garrett-Engel P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science.* 2003;302(5653):2141–4.
20. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell.* 2004;16(6):929–41.
21. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10(1):57–63.
22. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science.* 2008;321(5891):956–60.
23. Lee JH, Gao C, Peng G, Greer C, Ren S, Wang Y, Xiao X. Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ Res.* 2011;109(12):1332–41.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008;5(7):621–8.
25. Li G, Bahn JH, Lee JH, Peng G, Chen Z, Nelson SF, Xiao X. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* 2012;40(13):e104.
26. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* [TIG]. 2011;27(2):72–9.
27. Wulff BE, Sakurai M, Nishikura K. Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. *Nat Rev Genet.* 2011;12(2):81–5.
28. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* 2012;22(1):142–50.
29. Lee JH, Ang JK, Xiao X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA.* 2013;19(6):725–32.
30. Zhang Q, Xiao X. Genome sequence-independent identification of RNA editing sites. *Nat Methods.* 2015;12(4):347–50.
31. Kratz A, Carninci P. The devil in the details of RNA-seq. *Nat Biotechnol.* 2014;32(9):882–4.
32. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res.* 2014;322(1):12–20.
33. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P. Library construction for next-generation sequencing: overviews and challenges. *BioTech* 2014;56(suppl 2):61–4, 66, 68, passim.
34. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet.* 2010;11(10):733–9.
35. Liu Y, Ferguson JF, Xue C, Silverman IM, Gregory B, Reilly MP, Li M. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLoS ONE.* 2013; 8(6):e66883.
36. Katz Y, Wang ET, Airolidi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods.* 2010;7(12):1009–15.

37. Li H, Qiu J, Fu XD. RASL-seq for massively parallel and quantitative analysis of gene expression. In: Frederick M Ausubel et al. (Ed.) *Current protocols in molecular biology*, 2012; Chap. 4:Unit 4 13, pp 11–9.
38. Li H, Zhou H, Wang D, Qiu J, Zhou Y, Li X, Rosenfeld MG, Ding S, Fu XD. Versatile pathway-centric approach based on high-throughput sequencing to anticancer drug discovery. *Proc Natl Acad Sci USA*. 2012;109(12):4609–14.
39. Larman HB, Scott ER, Wogan M, Oliveira G, Torkamani A, Schultz PG. Sensitive, multiplex and direct quantification of RNA sequences using a modified RASL assay. *Nucleic Acids Res*. 2014;42(14):9146–57.
40. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*. 2011;8(6):469–77.
41. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Hubbard TJ, Guigo R, Harrow J, Bertone P, Consortium R. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods*. 2013;10(12):1177–84.
42. Roy CK, Olson S, Graveley BR, Zamore PD, Moore MJ. Assessing long-distance RNA sequence connectivity via RNA-templated DNA–DNA ligation. *eLife* 2015;4.
43. Zhang F, Wang M, Michael T, Drabier R. Novel alternative splicing isoform biomarkers identification from high-throughput plasma proteomics profiling of breast cancer. *BMC Syst Biol*. 2013;7(Suppl 5):S8.
44. Chen L. Statistical and computational methods for high-throughput sequencing data analysis of alternative splicing. *Stat Biosci*. 2013;5(1):138–55.
45. Hooper JE. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum Genomics*. 2014;8:3.
46. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*. 2014;32(5):462–4.
47. Zhang Z, Wang W. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics*. 2014;30(12):i283–92.
48. Aschoff M, Hotz-Wagenblatt A, Glatting KH, Fischer M, Eils R, Konig R. SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics*. 2013;29(9):1141–8.
49. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci USA*. 2014;111(51):E5593–601.
50. Kahles A, Ong CS, Ratsch G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Biorxiv*. 2015:017095.
51. Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan PF, Hammond SM, Makowski L, et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res*. 2013;41(2):e39.
52. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*. 2010;28(5):503–10.
53. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–5.
54. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52.
55. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform*. 2011;12:323.
56. Suo C, Calza S, Salim A, Pawitan Y. Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data. *Bioinformatics*. 2014;30(4):506–13.

57. Hu Y, Liu Y, Mao X, Jia C, Ferguson JF, Xue C, Reilly MP, Li H, Li M. PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. *Nucleic Acids Res.* 2014;42(3):e20.
58. Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, Lee MP. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.* 2003;63(3):655–7.
59. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. *Cell.* 2004;119(6):831–45.
60. Xiao X, Wang Z, Jang M, Burge CB. Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci USA.* 2007;104(47):18583–8.
61. Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat Struct Mol Biol.* 2009;16(10):1094–100.
62. Mark D, Haeblerle S, Roth G, von Stetten F, Zengerle R. Microfluidic lab-on-a-chip platforms: requirements, characteristics and applications. *Chem Soc Rev.* 2010;39(3):1153–82.
63. Arias MA, Lubkin A, Chasin LA. Splicing of designer exons informs a biophysical model for exon definition. *RNA.* 2015;21(2):213–29.
64. Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med: Off J Am Coll Med Genet.* 2014;16(7):497–503.
65. Desmet FO, Beroud C. Bioinformatics and mutations leading to exon skipping. *Methods Mol Biol.* 2012;867:17–35.
66. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 2009;37(9):e67.
67. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in genie. *J Comput Biol: J Comput Mol Cell Biol.* 1997;4(3):311–23.
68. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol: J Comput Mol Cell Biol.* 2004;11(2–3):377–94.
69. Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, Smith CW. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol.* 2006;7(1):R1.
70. Gao K, Masuda A, Matsuura T, Ohno K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res.* 2008;36(7):2257–67.
71. Plass M, Agirre E, Reyes D, Camara F, Eyras E. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet: TIG.* 2008;24(12):590–4.
72. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.* 2008;18(1):88–103.
73. Corvelo A, Hallegger M, Smith CW, Eyras E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol.* 2010;6(11):e1001016.
74. Pastuszak AW, Joachimiak MP, Blanchette M, Rio DC, Brenner SE, Frankel AD. An SF1 affinity model to identify branch point sequences in human introns. *Nucleic Acids Res.* 2011;39(6):2344–56.
75. Bitton DA, Rallis C, Jeffares DC, Smith GC, Chen YY, Codlin S, Marguerat S, Bahler J. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res.* 2014;24(7):1169–79.
76. Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol.* 2012;19(7):719–21.
77. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddelloh JA, Mattick JS, Rinn JL. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol.* 2012;30(1):99–104.

78. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. Genome-wide discovery of human splicing branchpoints. *Genome Res.* 2015;25(2):290–303.
79. Culler SJ, Hoff KG, Voelker RB, Berglund JA, Smolke CD. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Res.* 2010;38(15):5152–65.
80. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 2011;21(8):1360–74.
81. Wang Y, Ma M, Xiao X, Wang Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol.* 2012;19(10):1044–52.
82. Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol.* 2013;20(1):36–45.
83. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science.* 2002;297(5583):1007–13.
84. Zhang XH, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* 2004;18(11):1241–50.
85. Yeo G, Hoon S, Venkatesh B, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci USA.* 2004;101(44):15700–5.
86. Venkatesh B, Yap WH. Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *Bioessays: News Rev Mol, Cell Dev Biol.* 2005;27(1):100–7.
87. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* 2003;31(13):3568–71.
88. Zhang XH, Heller KA, Heftner I, Leslie CS, Chasin LA. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.* 2003;13(12):2637–50.
89. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet.* 2006;2(11):e191.
90. Zhang J, Kuo CC, Chen L. VERSE: a varying effect regression for splicing elements discovery. *J Comput Biol: J Comput Mol Cell Biol.* 2012;19(6):855–65.
91. Badr E, Heath LS. Identifying splicing regulatory elements with de Bruijn graphs. *J Comput Biol: J Comput Mol Cell Biol.* 2014;21(12):880–97.
92. Friedman BA, Stadler MB, Shomron N, Ding Y, Burge CB. Ab initio identification of functionally interacting pairs of cis-regulatory elements. *Genome Res.* 2008;18(10):1643–51.
93. Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Luhrmann R, Jankowsky E, Chasin LA, Nilsen TW. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell.* 2008;135(7):1224–36.
94. Ke S, Chasin LA. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biol.* 2010;11(8):R84.
95. Weyn-Vanhenenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, Xue C, Herre M, Silver PA, Zhang MQ, et al. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* 2014;6(6):1139–52.
96. Zhang C, Frias MA, Mele A, Ruggiu M, Eom T, Marney CB, Wang H, Licatalosi DD, Fak JJ, Darnell RB. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science.* 2010;329(5990):439–43.
97. Han A, Stoilov P, Linares AJ, Zhou Y, Fu XD, Black DL. De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Comput Biol.* 2014;10(1):e1003442.
98. Fu XD, Ares M Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet.* 2014;15(10):689–701.

99. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet.* 1992;90(1–2):41–54.
100. Ars E, Kruyer H, Gaona A, Serra E, Lazaro C, Estivill X. Prenatal diagnosis of sporadic neurofibromatosis type 1 (NF1) by RNA and DNA analysis of a splicing mutation. *Prenat Diagn.* 1999;19(8):739–42.
101. Teraoka SN, Telatar M, Becker-Catania S, Liang T, Onengut S, Tolun A, Chessa L, Sanal O, Bernatowska E, Gatti RA, et al. Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am J Hum Genet.* 1999;64(6):1617–31.
102. Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* 2005;579(9):1900–3.
103. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet.* 2008;40(2):225–31.
104. Zhao K, Lu ZX, Park JW, Zhou Q, Xing Y. GLIMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.* 2013;14(7):R74.
105. Monlong J, Calvo M, Ferreira PG, Guigo R. Identification of genetic variants associated with alternative splicing using sQTLseeker. *Nature Commun.* 2014;5:4698.
106. Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, Sanford JR, Mooney SD. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* 2014;15(1):R19.
107. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* 2011;21(10):1563–71.
108. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science.* 2015;347(6218):1254806.
109. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. Deciphering the splicing code. *Nature.* 2010;465(7294):53–9.
110. Barash Y, Vaquero-Garcia J, Gonzalez-Vallinas J, Xiong HY, Gao W, Lee LJ, Frey BJ. AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biol.* 2013;14(10):R114.
111. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet.* 2010;11(8):533–8.
112. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics.* 2009;25(24):3207–12.
113. Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet.* 2010;19(1):122–34.
114. Wang Y, Wang Z. Systematical identification of splicing regulatory cis-elements and cognate trans-factors. *Methods.* 2014;65(3):350–8.
115. Izquierdo JM, Majos N, Bonnal S, Martinez C, Castelo R, Guigo R, Bilbao D, Valcarcel J. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell.* 2005;19(4):475–84.
116. Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL. Homologues of the *Caenorhabditis elegans* Fox-1 protein are neuronal splicing regulators in mammals. *Mol Cell Biol.* 2005;25(22):10005–16.
117. Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, et al. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.* 2012;1(2):167–78.

118. Calarco JA, Superina S, O'Hanlon D, Gabut M, Raj B, Pan Q, Skalska U, Clarke L, Gelinas D, van der Kooy D, et al. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell*. 2009;138(5):898–910.
119. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*. 2013;499(7457):172–7.
120. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*. 1990;249(4968):505–10.
121. Lee Y, Rio DC. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu Rev Biochem*. 2015;84:291–323.
122. Reid DC, Chang BL, Gunderson SI, Alpert L, Thompson WA, Fairbrother WG. Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA*. 2009;15(12):2385–97.
123. Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* 2009;27(7):667–70.
124. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA bind-n-seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell*. 2014;54(5):887–900.
125. Ule J, Jensen K, Mele A, Darnell RB. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods*. 2005;37(4):376–86.
126. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*. 2008;456(7221):464–9.
127. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*. 2010;141(1):129–41.
128. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*. 2010;17(7):909–15.
129. McHugh CA, Russell P, Guttman M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol*. 2014;15(1):203.
130. Re A, Joshi T, Kulberkyte E, Morris Q, Workman CT. RNA-protein interactions: an overview. *Methods Mol Biol (Clifton, NJ)* 2014;1097:491–521.
131. Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol*. 2011;29(7):607–14.
132. Xiong HY, Barash Y, Frey BJ. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*. 2011;27(18):2554–62.
133. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodenic V, Kutter C, Watt S, Colak R, et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science*. 2012;338(6114):1587–93.
134. Busch A, Hertel KJ. Splicing predictions reliably classify different types of alternative splicing. *RNA (New York, NY)*. 2015;21(5):813–23.
135. de Klerk E, t Hoen PA. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet: TIG*. 2015;31(3):128–39.
136. Bentley DL. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet*. 2014;15(3):163–75.
137. de Almeida SF, Carmo-Fonseca M. Reciprocal regulatory links between cotranscriptional splicing and chromatin. *Semin Cell Dev Biol*. 2014;32:2–10.

138. Zhou HL, Luo G, Wise JA, Lou H. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.* 2014; 42(2):701–13.
139. Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, Hacohen N, Schier AF, Blackshear PJ, Friedman N, et al. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell.* 2014;159(7):1698–710.
140. Davis-Turak JC, Allison K, Shokhirev MN, Ponomarenko P, Tsimring LS, Glass CK, Johnson TL, Hoffmann A. Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. *Nucleic Acids Res.* 2015;43(2):699–707.
141. de Pretis S, Kress T, Morelli MJ, Melloni GE, Riva L, Amati B, Pelizzola M. INSPECt: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA-and 4sU-seq time course experiments. *Bioinformatics* 2015.
142. Nojima T, Gomes T, Grosso AR, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell.* 2015;161(3):526–40.
143. Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell.* 2015;161(3):541–54.
144. Chandola U, Das R, Panda B. Role of the N6-methyladenosine RNA mark in gene regulation and its implications on development and disease. *Briefings Funct Genomics.* 2015; 14(3):169–79.
145. Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature.* 2015;518(7540):560–4.
146. Rieder LE, Reenan RA. The intricate relationship between RNA structure, editing, and splicing. *Semin Cell Dev Biol.* 2012;23(3):281–8.
147. Hube F, Francastel C. Mammalian introns: when the junk generates molecular diversity. *Int J Mol Sci.* 2015;16(3):4429–52.
148. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. *Nat Biotechnol.* 2014;32(5):453–61.
149. Lasda E, Parker R. Circular RNAs: diversity of form and function. *RNA.* 2014;20(12): 1829–42.
150. Liang D, Wilusz JE. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* 2014;28(20):2233–47.
151. Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, et al. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.* 2015;10(2):170–7.
152. Wang Y, Wang Z. Efficient backsplicing produces translatable circular mRNAs. *RNA.* 2015;21(2):172–9.
153. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell.* 2014;56(1):55–66.
154. Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol.* 2015;22(3):256–64.
155. Conn SJ, Pillman KA, Toubia J, Conn VM, Salamanidis M, Phillips CA, Roslan S, Schreiber AW, Gregory PA, Goodall GJ. The RNA binding protein quaking regulates formation of circRNAs. *Cell.* 2015;160(6):1125–34.



156. Philips AV, Timchenko LT, Cooper TA. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science*. 1998;280(5364):737–41.
157. Zhang J, Manley JL. Misregulation of pre-mRNA alternative splicing in cancer. *Cancer Discovery*. 2013;3(11):1228–37.
158. Brooks AN, Choi PS, de Waal L, Sharifnia T, Imielinski M, Saksena G, Pedamallu CS, Sivachenko A, Rosenberg M, Chmielecki J, et al. A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS ONE*. 2014;9(1):e87361.
159. Dorman SN, Viner C, Rogan PK. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci Rep*. 2014;4:7063.
160. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallieres M, Tapial J, Raj B, O’Hanlon D, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*. 2014;159(7):1511–23.
161. Ohe K, Hagiwara M. Modulation of alternative splicing with chemical compounds in new therapeutics for human diseases. *ACS Chem Biol*. 2015;10(4):914–24.
162. Svasti S, Suwanmanee T, Fucharoen S, Moulton HM, Nelson MH, Maeda N, Smithies O, Kole R. RNA repair restores hemoglobin expression in IVS2-654 thalassemic mice. *Proc Natl Acad Sci USA*. 2009;106(4):1205–10.
163. Arechavala-Gomez V, Khoo B, Aartsma-Rus A. Splicing modulation therapy in the treatment of genetic diseases. *Appl Clin Genet*. 2014;7:245–52.
164. Wally V, Murauer EM, Bauer JW. Spliceosome-mediated trans-splicing: the therapeutic cut and paste. *J Invest Dermatol*. 2012;132(8):1959–66.
165. Havens MA, Duelli DM, Hastings ML. Targeting RNA splicing for disease therapy. *Wiley Interdisc Rev RNA*. 2013;4(3):247–66.
166. Kierlin-Duncan MN, Sullenger BA. Using 5'-PTMs to repair mutant beta-globin transcripts. *RNA*. 2007;13(8):1317–27.
167. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*. 2009;25(8):1026–32.
168. Reyes-Herrera PH, Ficarra E. Computational Methods for CLIP-seq Data Processing. *Bioinform Biol Insights*. 2014;8:199–207.

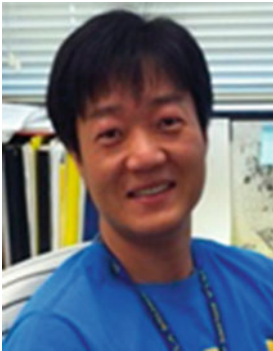
## Author Biographies



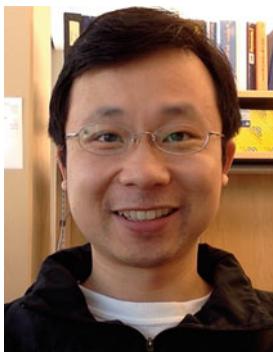
**Yun-Hua Esther Hsiao** graduated from the University of California, San Diego, with a B.S. degree in Bioengineering: Bioinformatics in 2012. She is currently a PhD student in Dr. Xinshu (Grace) Xiao’s laboratory at UCLA. Her research focuses on the development and application of bioinformatic methods for high-throughput sequencing data analysis. Her main projects aim to better understand the regulation of alternative splicing and other aspects of post-transcriptional regulation.



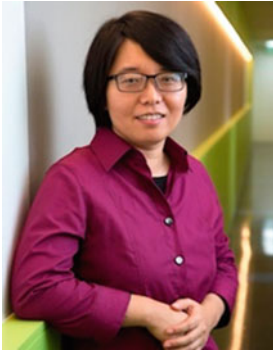
**Ashley Cass** received her B.S. in Computational and Systems Biology at UCLA in 2011. She is currently pursuing a PhD in Bioinformatics at UCLA and is a member of Dr. Xinshu (Grace) Xiao's laboratory. Her research focus is in using and developing bioinformatic methods to analyze RNA regulation and degradation, particularly mechanisms and consequences of small RNA-mediated regulation.



**Dr. Jae Hoon Bahn** received his PhD degree in Comparative and Experimental Medicine at University of Tennessee, Knoxville, in 2010. He is currently a postdoctoral researcher working in Dr. Xinshu (Grace) Xiao's laboratory at UCLA. As a bench scientist, he is working on the molecular mechanisms of post-transcriptional gene regulation, encompassing a number of topics in splicing regulation and regulatory mechanisms of RNA editing. Dr. Bahn has published 43 scientific papers with an h-index of 18 and more than 1000 citations.



**Dr. Xianzhi Lin** received his diploma in Bioengineering from the Kunming University of Science and Technology in 2004 and a PhD in Microbiology from the Institute Pasteur of Shanghai, Chinese Academy of Sciences in 2012. He is currently a postdoctoral researcher working in Dr. Xinshu (Grace) Xiao's laboratory at UCLA. His research focuses on the molecular mechanisms of RNA regulation, including alternative splicing, RNA editing, and RNA degradation.



**Dr. Xinshu (Grace) Xiao** is an associate professor in Integrative Biology and Physiology and is a member of the Bioinformatics Inter-Departmental Program (IDP), the Jonsson Comprehensive Cancer Center, and the Molecular Biology Institute of UCLA. Dr. Xiao's research focuses on the bioinformatics and genomics of RNA biology. Work in the Xiao laboratory is highly interdisciplinary, bridging bioinformatics, genomics, systems biology, and basic molecular biology. Her laboratory has focused on the computational and experimental studies of alternative splicing and its regulation, RNA editing, and small RNA regulation of gene expression. Dr. Xiao's group developed a number of new bioinformatic methods for studies of RNA regulation, including multiple methods to accurately identify A-to-I RNA editing sites using RNA-Seq data with or without genome data, new methods to predict genetically

regulated alternative splicing and polyadenylation events, and new short read aligners. The Xiao laboratory also applies existing and new methodologies to large-scale data analysis related to various diseases and biological processes, with the ultimate goal being an integrative and systematic understanding of RNA biology.

Transcriptomics and Gene Regulation

Wu, J. (Ed.)

2016, X, 185 p. 64 illus., 2 illus. in color., Hardcover

ISBN: 978-94-017-7448-2