

# Chapter 2

## Statistical and Computational Studies on Alternative Splicing

Liang Chen

**Abstract** The accumulating genome sequences and other high-throughput data have shed light on the extent and importance of alternative splicing in functional regulation. Alternative splicing dramatically increases the transcriptome and proteome diversity of higher organisms by producing multiple splice variants from different combinations of exons. It has an important role in many biological processes including nervous system development and programmed cell death. Many human diseases including cancer arise from defects in alternative splicing and its regulation. This chapter reviews statistical and computational methods on genome-wide alternative splicing studies.

### 2.1 Introduction

Alternative pre-mRNA splicing is a prevalent post-transcriptional gene regulation mechanism which has been estimated to occur in more than 90% of human genes [1, 2]. During alternative splicing, multiple transcript isoforms produced from a single gene can lead to protein isoforms with distinct functions, which greatly expands proteomic diversity in higher eukaryotes. The alternative splicing of multiple pre-mRNAs is tightly regulated and coordinated, and is an essential component for many biological processes including nervous system development and programmed cell death. The phenomenon of alternative splicing was first discovered in concept in 1978 [3], and was then verified experimentally in 1987 [4]. Alternative splicing was previously thought as a relatively uncommon form of gene regulation. With the accumulation of Expressed Sequence Tags (EST) and mRNA data sets, genome-wide studies on alternative splicing demonstrated that as many as 60% of the human genes were alternatively spliced [5–8]. The percentage was further

---

L. Chen

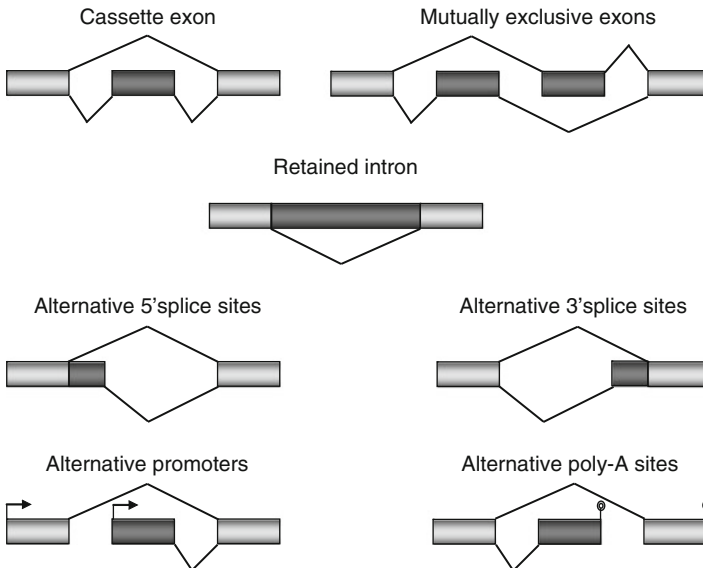
Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, California 90089, USA

e-mail: [liang.chen@usc.edu](mailto:liang.chen@usc.edu)

increased to 90% which was estimated by the most recent high-throughput sequencing technology [1, 2]. In addition, there is striking variation in alternative splicing across different tissues or different developmental stages [5]. These results indicate that alternative splicing plays an important role in increasing functional complexity in higher organisms rather than the exception in gene expression. With the availability of multiple genome sequences and high-throughput techniques, it is feasible to study alternative splicing on a genomic scale. Here we present an overview of the statistical and computational studies on alternative splicing, and important findings and challenges are highlighted and discussed.

## 2.2 Types of Alternative Splicing

Alternative splicing events can be classified into cassette exon, mutually exclusive exons, retained intron, alternative 5' splice sites, alternative 3' splice sites, alternative promoters, and alternative poly-A sites (Fig. 2.1). The most common type of alternative splicing is including or skipping a cassette exon in the mature mRNA. A pair of exons can be mutually exclusively spliced with only one exon included in the mature mRNA but not both. The excision of an intron can be suppressed, which results in the retention of the entire intron. And exons can be extended or shortened through the use of alternative 5' or 3' splice sites. Strictly speaking, alternative promoters and alternative poly-A sites are alternative selection of transcription start sites or poly-A sites and are not due to alternative splicing per se. Among these



**Fig. 2.1** Types of alternative splicing events

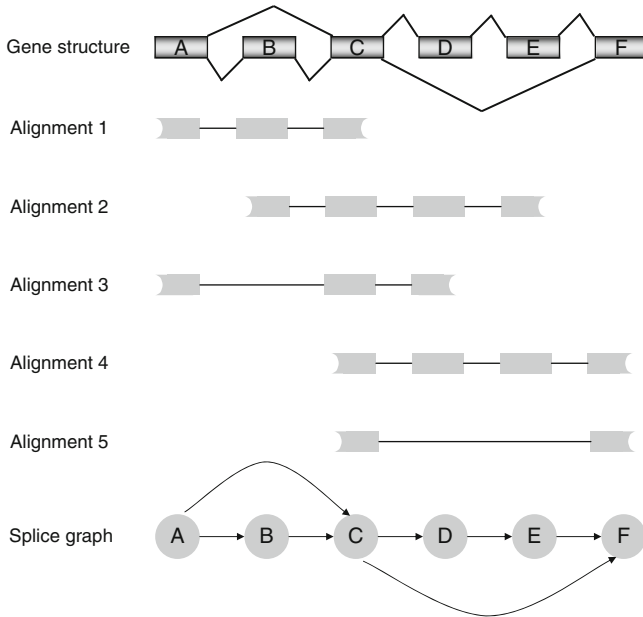
alternative splicing events, intron retention is generally the most difficult type to detect because it is hard to distinguish from experimental artifacts. For instance, incompletely spliced transcripts contain intron fragments, which could be mistakenly considered as intron retention. Many genes have multiple alternative splicing events with complex combinations of exons, producing a family of diverse transcript isoforms. For example, in *Drosophila melanogaster*, gene *Dscam* can potentially produce 38,016 different mature mRNAs by different combinations of 95 cassette exons [9–11].

## 2.3 Global Identification of Alternative Splicing Events

### 2.3.1 *Identifying Alternative Splicing by Sequence Alignment*

One way to identify alternative splicing events is based on the alignment of ESTs with genomic and mRNA sequences. EST sequences are short fragments of transcribed cDNA sequences, usually 300–400 base pair (bp). They are produced by shotgun sequencing of one or both strands of a cloned mRNA. About 61 million ESTs have been deposited in the public dbEST database (dated as April, 2009, all species). A number of programs have been developed to align ESTs against the complete genome sequences efficiently. For example, BLAT is a “BLAST-Like Alignment Tool” which uses a hashing and indexing algorithm [12]. It is about 500 times faster than BLAST for mRNA/DNA alignments. Given the alignments of ESTs and genomic sequences, we can mark the locations of exons and introns. The comparisons of exon-intron structures further distinguish the alternative splicing events. Sometimes, an EST can be mapped to multiple genomic positions with high alignment scores. These genome alignments can be further corrected by considering consensus splice sites. For example, alignment tools SIM4 [13], GMAP [14], and SPA [15] consider GT...AG consensus splice sites to generate valid alignments. Although the sequence alignment approaches have made much progress in alternative splicing detection, challenges remain in dealing with non-canonical splice junctions, detection of small exons, high EST sequencing errors, bias inherent to EST preparation, and so on. Other limitations include the insufficient sequence coverage for some transcripts and the biased sampling to a limited number of cell and tissue types.

After the identification of individual alternative splicing events, a more complicated task is the construction of full-length alternatively spliced transcripts. “Splice graph” has been introduced to facilitate the construction of full-length transcript isoforms [16–19]. The splice graph represents a gene as a directed acyclic graph in which exons are represented as vertices and each splice junction is represented as a directed edge between two exons (see example in Fig. 2.2). Splice variants can be inferred by graph algorithms to traverse the graph from a start vertex with no incoming arcs to an end vertex with no outgoing arcs. A large number of potential splice variants can be enumerated from a splice graph, but many of them may be



**Fig. 2.2** Splice graph constructed from EST alignments to reference genome. The underlying true gene structure and the observed evidence alignments are also shown

artificial constructs without biological relevance because exons are not randomly joined to produce all possible transcript isoforms. Several methods have been proposed to select or prioritize candidate transcripts which are most likely to exist given the sequence observations. For example, AIR is an integrated software system for gene and alternative splicing annotation [16]. It assigns different scores to different splicing variants based on its support by evidence such as mapping quality, the length of alignment, accuracy of splice signals, and the level of fragmentation of evidence alignments. High-scoring splice variants were further selected for the annotation. ECgene algorithm assesses each possible splice variant based on the sequence quality and the number of cDNA alignments [18]. Xing et al. applied the Expectation-Maximization algorithm to identify the most likely traversals based on the observed number of alignments along the gene [19]. The performance of these methods is limited by the contamination of ESTs with genomic fragments, alignment errors, and so on.

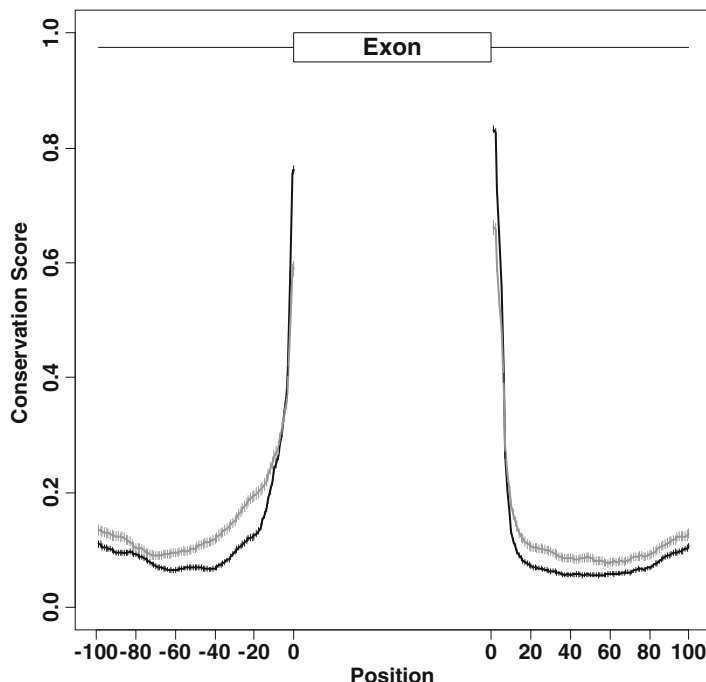
### **2.3.2 Identifying Alternative Splicing by Sequence Content and Conservation**

Because mRNA alternative splicing is a highly regulated process, comparative genomics can provide us clues about whether there is an alternative exon in sites

with high selection pressure. Alternative methods have been proposed to predict alternatively spliced exons based on machine learning algorithms incorporating features such as sequence content and sequence conservation. Lepercq et al. used splice-site sequence Markov models and a Bayesian classifier to identify cassette exons from intron sequences [20]. With additional information from sequence conservation and phosphorylation or protein-binding motifs, they successfully predicted and experimentally confirmed 26 novel human cassette exons which are involved in intracellular signaling. Sorek et al. assembled 243 alternative and 1,753 constitutive exons that are conserved between human and mouse [21, 22]. They identified several features differentiating between alternatively spliced and constitutively spliced exons. Specifically, alternative exons tend to be smaller, have length that is a multiple of 3 (to preserve the protein reading frame), have higher sequence identity between human and mouse sequences, and have higher conservation in the flanking intronic regions. The most important features are the ones based on the sequence similarity between human and mouse. Yeo et al. used sequence features to distinguish alternative splicing events conserved in human and mouse [23]. Chen et al. used the Random Forests algorithm to predict skipped exons using features like position-specific conservation scores [24]. The training data was based on the high-quality annotation of the Encyclopedia of DNA Elements (ENCODE) regions. The pilot project of the ENCODE has rigorously identified functional elements in the 1% region of the human genome. The GENCODE consortium of the ENCODE project has manually prepared a high-quality annotation for transcripts in the ENCODE regions. Chen et al. assembled the lists of skipped exons, constitutive exons and introns as training sets. Using the Random Forest algorithm [25], they were able to identify skipped exons based on the sequence content and conservation features [24]. The Random Forests consist of many decision trees and each tree is constructed by a bootstrap sample from the original data. A decision tree can be treated as a set of Boolean functions of features and these conjunctions of features partition training samples into groups with homogenous class labels. The output of the Random Forests for each test sample is the class with majority votes from these trees. The Random Forests generates an internal unbiased estimate of classification error based on the out-of-bag data during the Forests building process. There is no need for cross-validation or a separate test data.

As shown in Fig. 2.3, there are dramatic differences in the conservation scores of the flanking regions of alternative exons and constitutive exons. Alternative exons have higher conservation level in the flanking intronic regions compared to constitutive exons. These more conserved regions provide good candidates for functional regulatory motifs. The enriched sequence motifs in these regions may participate in the alternative splicing modulation which could be different from the regular splicing process.

Besides the flanking intronic regions, the exonic regions are also involved in the splicing regulation. However, the comparative genomics studies on exonic regions are more complicated, because additional selective pressure is imposed on the coding sequence in order to preserve the protein sequence. It has been shown that the evolution rate is lower for exon regions near the intron-exon boundaries than the



**Fig. 2.3** Position-specific conservation for the flanking intronic regions of constitutive exons (black) and alternative exons (grey) (Adapted from [24]). Y axis is the average conservation score at each position. The error bar indicates the standard error of the mean. Constitutive exons and alternative exons were assembled from the high-quality annotation of the ENCODE project. The conservation score is the PhastCon score from the UCSC Genome Browser (<http://genome.ucsc.edu/>)

middle part of exons, by estimating the non-synonymous substitution rate and the synonymous substitution rate from the alignment of human-mouse sequences [26]. The SNP density is the lowest near the splice sites, which also indicates that exon regions near the splice sites are under higher selection pressure [27]. These findings suggest that the exon regions near the junctions are involved in splicing regulation. Further studies are needed to distinguish the selection pressure on alternative exons, constitutive exons, and amino acid constrains.

### 2.3.3 Identify Alternative Splicing by Microarray

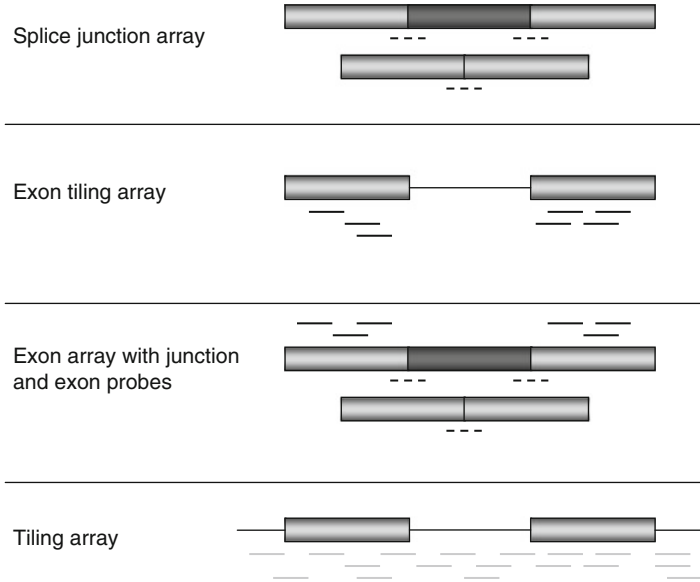
Although the sequence alignment and the comparative genomics approaches have made much progress in the prediction of alternative splicing events, they give us only a qualitative rather than a quantitative view of alternative splicing. They only provide evidence about the existence of an alternative splicing event, but cannot give

information about its temporal and spatial regulation nor the degree of alternative splicing.

The highly parallel nature of microarray platforms makes it possible to identify and quantify all of alternative splicing for a specific tissue, developmental stage, or disease versus normal conditions of the cell. Traditional microarrays are spotted with EST-derived cDNAs or 3'-clustered oligonucleotide sequences representing the total transcript abundance. These microarrays are not suitable for alternative splicing studies and special probes need to be designed instead. For example, splice junction arrays bear probes spanning annotated exon-exon junctions for individual splice variant. Johnson et al. designed a set of five Agilent microarrays containing  $\sim 125,000$  different 36-nucleotide (nt) junction probes to monitor the exon-exon junctions of 10,000 multi-exon Human RefSeq genes across 52 tissues and cell lines [5]. Boutz et al. used splice junction arrays to monitor the reprogrammed alternative splicing during neuronal development [28]. Besides splice junction arrays, alternative arrays use "exon-centric" probes. For instance, in the design of Affymetrix exon arrays, gene annotations from databases were assembled to infer transcript clusters and exon clusters. A transcript cluster roughly corresponds to a gene. In many cases, an exon cluster represents a true biological exon and it acts as one probe selection region. In other cases, an exon cluster represents the union of multiple overlapping exons possibly due to alternative splice sites. Such exon clusters were further fragmented into multiple probe selection regions according to the hard edges (e.g., splice sites). Multiple probes were designed for each probe selection region as a probe set. The Affymetrix human exon array (1.0 ST) contains approximately 1.4 million probe sets interrogating over one million exon clusters. Analysis of alternative splicing in 16 human tissues with these arrays identified a large number of tissue-specific exons [29]. Yeo et al. used Affymetrix exon arrays to identify the differential alternative splicing between human embryonic stem cells and neural progenitor cells [30]. More recent microarrays include both junction probes and exon body probes. Castle et al. designed probes targeting on exons or junctions to monitor 203,672 exons and 178,351 exon-exon junctions in 17,939 human genes across 48 diverse human tissues and cell lines [31]. In addition, tiled oligonucleotide arrays spanning whole chromosomes or genomes provide comprehensive coverage and avoid the need of prior information about exons. However, this approach is expensive and needs extremely large number of probes. These microarray designs are summarized in Fig. 2.4. In principle, all data analysis tools developed for standard gene microarrays can be used in the analysis of alternative splicing microarrays. The special challenge is how to distinguish splicing signal from transcription signal. The methods outlined below present some tools that have been used on the alternative splicing microarray data analysis.

### 2.3.3.1 Splicing Index

For the alternative splicing microarray analysis, the most straightforward approach is the splicing index calculation [32]. In the splicing index approach, exon inclusion



**Fig. 2.4** Alternative splicing microarrays. *Black dot lines* represent junction probes. *Black solid lines* represent exon probes. For tiling arrays, probes are designed along the genome disregarding gene structure (*grey lines*)

rates under two conditions are compared to identify differential alternative splicing events. Gene-level normalized exon intensity is defined as the ratio of the exon intensity to the gene intensity. For example, the normalized intensity (NI) for exon  $i$  in experiment  $j$  is:

$$NI_{ij} = E_{ij} / G_j \quad (2.1)$$

where  $E_{ij}$  is the estimated intensity level for exon  $i$  in experiment  $j$  and  $G_j$  is the estimated gene intensity. “Gene intensity” here represents the overall transcript abundance of a gene which may include a family of transcript isoforms. “Gene intensity” can be estimated by dynamic weighting of the most informative probes. It is robust to outliers due to alternative splicing. Thus, the contributions from alternative exons to “gene intensity” are trivial.

A significant difference in the normalized exon intensity indicates that this exon has different inclusion or exclusion rates (relative to the gene level) between two conditions. The splicing index for experiment 1 and experiment 2 is defined as:

$$\text{Splicing index} = \log_2(NI_{i1} / NI_{i2}). \quad (2.2)$$

Therefore, an extreme value of splicing index indicates a differential alternative splicing event.



### 2.3.3.2 ANOSVA

Analysis of splice variation (ANOSVA) uses a statistical testing principle to detect putative splicing variation from expression data [33]. It is based on a two-way analysis of variance (ANOVA) model:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \text{error}, \quad (2.3)$$

where  $y_{ijkl}$  is the observed log intensity of probe  $k$  of probe set  $i$  (or exon  $i$ ), measured in experiment  $j$  of experiment set  $l$ ;  $\mu$  is the baseline intensity level for all probes in all experiments;  $\alpha_i$  is the average probe affinity of probe set  $i$ ;  $\beta_j$  is the experiment effect; and  $\gamma_{ij}$  is the interaction term for probe set  $i$  and experiment  $j$ . A large change in splicing will result in a large interaction term  $\gamma_{ij}$ . However, due to the limited number of replicates for exon-array experiments and the resultant limited statistical power, it is difficult to identify interactions. Meanwhile, a significant interaction term does not necessarily mean a large change in splicing, because the unfitness of the single-concentration model without the interaction term may be simply due to the high noise level. Preliminary evaluation of ANOSVA on exon array data did not yield good performance (Alternative Transcript Analysis Methods for Exon Arrays Whitepaper, Affymetrix). Therefore, ANOSVA should be used with caution.

### 2.3.3.3 FIRMA

Instead of estimating the interaction term  $\gamma_{ij}$  explicitly, FIRMA (Finding isoforms using robust multichip analysis) [34] frames the problem of detecting alternative splicing as a problem of outlier detection. In FIRMA,  $y_{ijk}$  represents log intensity of probe  $k$  of exon  $i$  measured in experiment  $j$  (signal has been background-corrected and normalized). It is modeled as:

$$y_{ijk} = c_j + p_k + \text{error}, \quad (2.4)$$

where  $c_j$  is the experiment effect and  $p_k$  is the probe effect. The residual from the fitted model is:

$$r_{ijk} = y_{ijk} - \hat{c}_j + \hat{p}_k. \quad (2.5)$$

The residual describes the discrepancy of probe intensity in a given experiment from the expected expression and gives a measure of the hidden interaction term  $\gamma_{ij}$ . The final score statistic is:

$$F_{ij} = \text{median}_{k \in \text{exon } j} (r_{ijk}/s). \quad (2.6)$$

The standard error,  $s$ , is calculated by the median absolute deviation (MAD) of the residuals. Compared with ANOSVA, FIRMA can detect alternative splicing without replicates. And the interaction term is not directly inferred and reflected by a

robust measure of the residuals instead. FIRMA assumes that the interaction term has limited effect so that  $c$  and  $p$  are still well estimated in the model without the interaction term.

### 2.3.3.4 DECONV

The above methods target on each individual exon to determine whether it is differentially spliced or not. They do not require the whole exon-intron structure of a gene. A gene may have multiple positions of alternative splicing and the resulted multiple ( $>2$ ) splice variants can coexist in the same condition. Another challenging task is to estimating the relative abundance of each variant in one condition. Wang et al. developed a gene structure-based splice variant deconvolution method (DECONV) to estimate the splice variant's concentration [35]. DECONV assumes that there is linear relationship between the probe intensity and the target transcript concentration as proposed by Li and Wong [36]. In the reduced model of Li and Wong,

$$y_{ij} = PM_{ij} - MM_{ij} = a_i x_j + \varepsilon_{ij}, \quad (2.7)$$

where  $y_{ij}$  is the intensity level for probe  $i$  in experiment  $j$ ,  $a_i$  is the probe affinity, and  $x_j$  is the target transcript concentration. DECONV extends the model for multiple splice variants case:

$$\mathbf{Y} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{T} + \mathbf{E}, \quad (2.8)$$

where  $\mathbf{Y}$  is an  $I$  by  $J$  matrix with  $y_{ij}$  representing the intensity for probe  $i$  in experiment  $j$ ,  $\mathbf{A} = \text{diag}(a_{11}, \dots, a_{II})$  is the diagonal matrix of unknown affinities for all of the probes included in the gene; matrix  $\mathbf{T} = \{T_{kj}\}$  represents the unknown concentration of the  $k$ -th splice variant in the  $j$ -th experiment; the property matrix  $\mathbf{G} = \{g_{ik}\}$  relates probes with different splice variants according to whether the probe belongs to the transcript or not.

$$\begin{aligned} g_{ik} &= 1 \quad \text{if probe } i \text{ belongs to splice variant } k, \\ &= 0 \quad \text{if probe } i \text{ does not belong to splice variant } k. \end{aligned} \quad (2.9)$$

And  $\mathbf{E}$  is the error term. To estimate the unknown  $\mathbf{A}$  and  $\mathbf{T}$ , they minimize the function:

$$f(\mathbf{A}, \mathbf{T}) = (\|\mathbf{Y} - \mathbf{AGT}\|_2)^2, \quad (2.10)$$

under the constraints:

$$\sum_{i=1}^I a_{ii}^2 = \text{constant}, \quad (2.11)$$

$$a_{ii} \geq 0, \quad (2.12)$$

$$t_{kj} \geq 0. \quad (2.13)$$

The maximum likelihood estimation framework is finally used by iteratively fixing  $\mathbf{A}$  and solving for  $\mathbf{T}$ , then fixing  $\mathbf{T}$  and solving for  $\mathbf{A}$  until convergence. DECONV works well for genes with two transcript isoforms, but is less than perfect for genes with three or more isoforms. DECONV requires the complete information about the number and the structure of all possible splice variants for a gene. It is not intended for the discovery of new splice variants.

### 2.3.3.5 SPACE

A similar algorithm, SPACE (splicing prediction and concentration estimation), was proposed to predict the structures and the abundances of transcript isoforms from microarray data [37]. Besides matrices  $\mathbf{A}$  and  $\mathbf{T}$ , they also treated the gene structure matrix  $\mathbf{G}$  as unknown. A “non-negative matrix factorization” method was applied to handle the non-negative constraints and factorize  $\mathbf{Y}$  into  $\mathbf{W}$  and  $\mathbf{H}$ :

$$\mathbf{Y}_{IJ} \approx \mathbf{W}_{IK} \cdot \mathbf{H}_{KJ}. \quad (2.14)$$

Remember that  $\mathbf{Y} \approx \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{T}$ , so  $\mathbf{H}$  gives the relative concentration of each splice variant and  $\mathbf{W}$  contains information of both probe affinity and gene structure. Specifically, they used the maximum value of each row of the  $\mathbf{W}$  matrix as the affinity of the corresponding probe.

$$\begin{aligned} a_{ii} &= \max_k (W_{ik}) \\ \mathbf{G} &= \mathbf{A}^{-1} \mathbf{W}. \end{aligned} \quad (2.15)$$

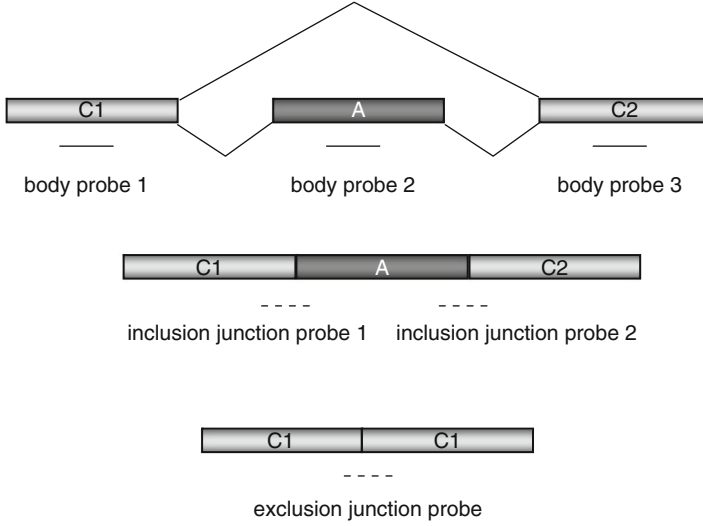
Here  $\mathbf{G}$  will be a matrix whose entries are between 0 and 1. There is a slight change in the definition of  $\mathbf{G}$ :

$$\begin{aligned} g_{ik} &= 1 \quad \text{if probe } i \text{ belongs to splice variant } k, \\ &= 0 \quad \text{if probe } i \text{ does not belong to splice variant } k, \\ &= \alpha \quad \text{if probe } i \text{ partially hybridizes with splice variant } k. \end{aligned} \quad (2.16)$$

The authors reported that the estimation of isoform structure and abundance depends on the number of experiments. When there are only a few experiments (e.g., 5), the estimation error tends to be high. They also mentioned that the model works better if the array includes more probes that are able to distinguish different isoforms or if several different experimental conditions with high variability are considered.

### 2.3.3.6 GenASAP

Shai et al. developed the GenASAP (Generative model for the alternative splicing array platform) algorithm to infer the expression levels of transcript isoforms



**Fig. 2.5** Custom microarray design for cassette exons. *Dot lines* represent junction probes. *Solid lines* represent exon body probes

including or excluding a cassette exon [38]. This was designed specifically for a custom microarray in which an exon-skipping event is represented by three exon body probes and three junction probes (see Fig. 2.5). The probe intensity  $x_i$  can be written as:

$$x_i = \lambda_{i1}s_1 + \lambda_{i2}s_2 + \varepsilon_i, \quad (2.17)$$

where  $x_i$  is one of the six intensity values for the six specially designed probes,  $s_1$  and  $s_2$  are the two unknown concentrations of the transcript isoforms,  $\lambda_{i1}$  and  $\lambda_{i2}$  are the affinity between probe  $i$  and the two transcript isoforms, and  $\varepsilon_i$  is the error term. To account for the scale-dependent noise and the outliers, the above model is changed to:

$$x_i = (r(\lambda_{i1}s_1 + \lambda_{i2}s_2 + \varepsilon_i))^{1-o_i} (\zeta_i)^{o_i}, \quad (2.18)$$

where  $r$  is the scale factor accounting for noise levels at the measured intensity,  $\zeta_i$  is a pure noise component for the outlier, and  $o_i$  is the binary indicator whether the probe measurement is an outlier or not. The conditional probability can be written as:

$$P(\mathbf{X}|\mathbf{S}, r, \mathbf{O}) = \prod_i \mathcal{N}(x_i; r(\lambda_{i1}s_1 + \lambda_{i2}s_2), r^2\psi_i)^{1-o_i} \mathcal{N}(x_i; \varepsilon_i, v_i)^{o_i}, \quad (2.19)$$

where  $\mathcal{N}(x; \mu, \sigma^2)$  indicates the density of point  $x$  under normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The variance of probe intensity is  $r^2\psi_i$ . The mean and variance for outliers are  $\varepsilon_i$  and  $v_i$ . And it assumes independence among probes. The authors used a truncated normal distribution ( $\beta \geq 0$ ) to satisfy the non-negative

constraint on isoform abundance and maximized the lower bound of the log likelihood instead of the log likelihood itself during their variational EM learning because the exact posterior cannot be computed.

GenASAP performs well on the abundance estimation and outperforms many supervised methods. It has been successfully applied to the analysis of alternative splicing in mammalian cells and tissues [39, 40]. But it is specific to the focused probe design. In addition, if a gene has more than one alternative exon and more than two transcript isoforms consequently, GenASAP cannot distinguish isoforms which all include the tested cassette exon, neither can it further distinguish isoforms which all exclude the tested cassette exon.

### 2.3.4 *Identify Alternative Splicing by High Throughput Sequencing*

Recently, high-throughput sequencing based approach (RNA-Seq) has also been developed to map and quantify transcriptomes. Poly(A)+ mRNAs are purified from cells and fragmented to small size (e.g.,  $\sim 200$  bp). Then they are converted into cDNA and sequenced by the high-throughput sequencing techniques. Sequence tags or reads (usually about 25  $\sim$  50 bp for Solexa and SOLid or 250  $\sim$  400 bp for 454, and the length expected to increase slightly) from the sequencing machines are mapped to genes and used as a quantitative measure of the expression level. RNA-seq has been successfully applied to yeast [41, 42], *Arabidopsis thaliana* [43], mouse [44, 45], and human [1, 2, 46, 47]. For RNA-seq data, inclusion or exclusion rate of an exon was calculated based on the exon body reads, the flanking inclusion junction reads, and the exclusion junction reads. For example, Wang et al. used the “percent spliced in” (PSI or  $\Psi$ ) values to determine the fraction of mRNA containing an exon [1]. The PSI value was estimated as the ratio of the density of inclusion reads (i.e. reads per position in regions supporting the inclusion isoform) to the sum of the densities of inclusion and exclusion reads. Pan et al. used the inclusion and exclusion junction reads to quantify the transcript percentage [2]. In their study, the results from RNA-seq data are consistent with results which are from custom microarrays mentioned in GeneASAP. The correlation is 0.8 when applying a threshold of 20 or more reads in one experiment that match at least one of the three splice junctions representing inclusion or skipping of a cassette exon. The correlation increases to 0.85 when a threshold of 50 or more junction reads is applied.

Besides the analysis at the individual exon level, Jiang and Wong developed a method to estimate the transcript isoform abundance from RNA-seq data [48]. This is achieved by solving a Poisson model. Suppose a gene has  $m$  exons with lengths  $\mathbf{L} = (l_1, \dots, l_m)$  and  $n$  transcript isoforms with expressions  $\Theta = (\theta_1, \dots, \theta_n)$ . If two isoforms share part of an exon, the exon was split into several parts and each part was treated as an exon respectively. The count of reads falling a specific region  $s$  (e.g., an exon or an exon-exon junction) is the observed data  $X_s$ . Let  $w$  be the total number of mapped reads. Then  $X$  follows a Poisson distribution with mean  $\lambda$ .

When  $s$  is exon  $j$ ,  $\lambda = l_j w \sum_{i=1}^n c_{ij} \theta_i$  where  $c_{ij}$  is 1 if isoform  $i$  contains exon  $j$  and 0 otherwise. When  $s$  is an exon-exon junction,  $\lambda = l w \sum_{i=1}^n c_{ij} c_{ik} \theta_i$  where  $l$  is the length of the junction region, and  $j$  and  $k$  are indices of the two exons involved in the junction. Assuming the independence among different regions, the joint log-likelihood function can be written as:

$$\log(\mathcal{L}(\Theta|x_s, s \in S)) = \sum_{s \in S} \log(\mathcal{L}(\Theta|x_s)). \quad (2.20)$$

The isoform abundance  $\theta$ 's can be obtained by the maximum likelihood estimate (MLE). When the true isoform abundance  $\theta$  is not on the boundary of the parameter space, the distribution of  $\hat{\Theta}$  can be approximated asymptotically by a normal distribution with mean  $\Theta$  and covariance matrix equal to the inverse Fisher information matrix  $I(\Theta)^{-1}$ . However, in one experimental condition, many isoforms are lowly expressed and the likelihood function is truncated at  $\theta_i = 0$ . The constraints  $\theta_i \geq 0$  for all  $i$  make the covariance matrix estimated by  $I(\Theta)^{-1}$  unreliable. Instead, they developed a Bayesian inference method based on importance sampling from the posterior distribution of  $\theta$ 's. They utilized the RefSeq mouse annotations and applied their model to a RNA-seq data set. Their results have good consistency with RT-PCR experiments (Pearson's correlation coefficient  $> 0.6$ ).

Instead of estimating the isoform abundance in each experiment, Zheng and Chen proposed a hierarchical Bayesian model, BASIS (Bayesian analysis of splicing isoforms), to identify differentially expressed transcript isoforms between two experiments. BASIS can be applied to both tiling array data and RNA-seq data [49]. For each probe  $i$  that appears in at least one transcript isoform of gene  $g$ , consider the linear model:

$$\Delta y_{gi} = \sum \Delta \beta_{gj} x_{gij} + \Delta \varepsilon_{gi}, \quad (2.21)$$

where  $\Delta y_{gi}$  is the intensity difference between two conditions for probe  $i$  of gene  $g$  ( $\Delta y_{gi} = y_{gi}^1 - y_{gi}^2$ , the intensity is background corrected and normalized),  $\Delta \beta_{gj}$  is the expression difference between two conditions for the  $j$ -th transcript isoform of gene  $g$ ,  $x_{gij}$  is the binary indicator of whether probe  $i$  belongs to isoform  $j$ 's exon region, and  $\Delta \varepsilon_{gi}$  is the error term. Within one data set,  $g$  ranges from 1 to  $G$ , where  $G$  is the total number of genes;  $i$  ranges from 1 to  $n_g$  where  $n_g$  is the total number of probes for gene  $g$ ; and  $j$  ranges from 1 to  $s_g$  where  $s_g$  is the total number of transcript isoforms for gene  $g$ . The total  $\Delta \varepsilon_{gi}$ 's ( $g = 1, \dots, G$  and  $i = 1, \dots, n_g$ ) are divided into 100 bins. Each bin contains thousands of probes with similar values. Because probe intensity variance is dependent on probe intensity mean, probes in the same bin exhibit similar variances. The same model can be specified for RNA-seq data with  $y$  representing the read coverage over each position.

A hierarchical Bayesian model is constructed as:

$$\begin{aligned} \Delta \mathbf{Y}_g | \Delta \boldsymbol{\beta}_g, \boldsymbol{\Sigma}_g &\sim \mathcal{N}_{n_g}(\mathbf{X}_g \Delta \boldsymbol{\beta}_g, \boldsymbol{\Sigma}_g), \quad g = 1, \dots, G; \\ \boldsymbol{\Sigma}_g &\equiv \text{diag}(\pi_{g1}, \dots, \pi_{gn_g}), \pi_{gi} = \delta_m \text{ if probe (or position) } i \text{ of gene } g \in \text{bin } m; \end{aligned}$$

$$\begin{aligned}
\delta_m &\sim IG(v/2, v\lambda/2), \quad m = 1, \dots, 100; \\
\Delta\beta_g | \gamma_g &\sim \mathcal{N}_{s_g}(0, \mathbf{R}_g); \\
\mathbf{R}_g &\equiv \text{diag}(\kappa_{g1}, \dots, \kappa_{gs_g}), \kappa_{gj} = \tau_{gj} \text{ if } \gamma_{gj} = 0 \text{ and } \kappa_{gj} = \psi_{gj} \text{ if } \gamma_{gj} = 1; \\
f(\gamma_g) &= \prod_{j=1}^{s_g} p^{\gamma_{gj}} (1-p)^{1-\gamma_{gj}};
\end{aligned}$$

where  $\Delta\mathbf{Y}_g$ ,  $\Delta\beta_g$ , and  $\mathbf{X}_g$  are matrices with elements described before,  $\gamma_g$  is a latent variable,  $\mathcal{N}_{n_g}$  and  $\mathcal{N}_{s_g}$  stand for multivariate normal distributions, and IG stands for the inverse gamma distribution. Given the isoform amount differences ( $\Delta\mathbf{Y}_g$ ) and the probe arrangements ( $\mathbf{X}_g$ ), the probe intensity (or read coverage) differences ( $\Delta\beta_g$ ) follow a multivariate normal distribution with mean  $\mathbf{X}_g \Delta\beta_g$  and variance  $\Sigma_g$ . For the variance  $\Sigma_g$ , specifically, if a probe (or position) is assigned to bin  $m$ , the variance of the intensity (or coverage) difference is  $\delta_m$ .  $\delta_m$  itself is a random variable following an inverse gamma distribution.  $\gamma_{gj}$  is an indicator whether the  $j$ -th isoform is differentially expressed. When  $\gamma_{gj} = 0$ , the isoform difference  $\Delta\beta_{gj} \sim \mathcal{N}(0, \tau_{gj})$  and when  $\gamma_{gj} = 1$ ,  $\Delta\beta_{gj} \sim \mathcal{N}(0, \psi_{gj})$ . Here  $\mathcal{N}$  stands for normal distribution.  $\tau_{gj}$  was set as a small value so that when  $\gamma_{gj} = 0$ ,  $\Delta\beta_{gj}$  is small enough to be estimated as 0.  $\psi_{gj}$  was set as a large value so that when  $\gamma_{gj} = 1$ ,  $\Delta\beta_{gj}$  is large enough to be included in the final model. Therefore, the latent variable  $\gamma$  can perform variable selection for the linear model. The errors for probes belonging to the same gene can be heteroscedastic and assigned to different bins. In the prior distributions for parameters  $\Delta\beta, \delta, \gamma$ , there are hyperparameters  $(\tau, \psi, v, p)$ . Model parameters were inferred based on an ergodic Markov chain generated by the Gibbs sampler.

In summary, a latent variable was introduced to perform direct statistical selection of differentially expressed isoforms. BASIS has the ability to borrow information across different probes (or positions) from the same genes and different genes. It can handle the heteroscedasticity of probe intensity or sequence read coverage, and has been successfully applied to a whole-genome human tiling array data and a mouse RNA-seq data. The authors also found that the power of BASIS is related to gene structure [49]. Specifically, if a gene has more probes (or positions), the power of BASIS is larger. If the difference among isoforms is larger, the power of BASIS is larger. BASIS does not rely on the percentage of isoform-specific positions, and it considers the joint behavior of positions. The model also depends on the completeness of the known splicing patterns of each gene. The authors utilized the Alternative Splicing and Transcript Diversity database [50]. As information accumulates and novel transcript isoforms are discovered, a more accurate and complete alternative splicing annotation database will further improve results derived from BASIS.

## 2.4 Alternative Splicing Regulation in Eukaryotes

The splicing of pre-mRNA transcripts is carried out by spliceosomes which are large ribonucleoprotein complexes with more than 100 core proteins and five small nuclear RNAs [51, 52]. Besides the core splicing factors, there are additional trans-acting splicing regulators. Consequently, in addition to the core splicing signals including the 5' splice site (5' ss), the 3' splice site (3' ss), and the branch point sequence (BPS), there is a large amount of splicing regulatory elements for both constitutive exons and alternative exons. These splicing regulatory elements (SREs) can be further classified as exonic splicing enhancers (ESEs), exonic splicing silencer (ESSs), intronic splicing enhancers (ISEs), or intronic splicing silencers (ISSs) based on their locations and functions. Due to the selective constraints, enhancers are expected to play predominant roles in the efficient constitutive splicing, and silencers are expected to play predominant roles in the control of alternative splicing [53]. Large-scale screens of exonic SREs have been conducted experimentally and computationally. Fewer screens for intronic SREs were performed although intronic SREs may have a more prominent role in the alternative splicing regulation because the intronic regions flanking alternative exons are more conserved than those flanking constitutive exons. Motif discovery methods commonly used in transcription factor binding motif identification, in principle, can also be used for the splicing regulatory motif finding. Compared with transcription factor binding sites, the SREs are usually shorter, more degenerate, and have less information content. This poses additional challenges to predict SREs. Similar as the DNA motifs for transcription factor binding, multiple copies of SREs for a single exon will increase their effect on splicing regulation [54–58]. Experimental approaches like cross-linking/immunoprecipitation (CLIP), RNP immunoprecipitation (RIP), and genomic SELEX were applied to identify the binding sites of RNA-binding proteins. Those approaches can be further extended to genome-wide studies of SREs. However, similar as transcription factors, the binding of splicing regulators may not necessarily lead to the regulation.

In the process of alternative splicing, splicing regulators bind to various pre-mRNAs and affect a large number of exons. Meanwhile the splicing pattern of a specific exon is determined by multiple pre-mRNA-binding proteins [59, 60]. Therefore, it is particularly interesting and challenging to study how the splicing of a group of exons is co-regulated; how the splicing of an exon is combinatorially controlled by multiple regulators; and what are the general rules of “splicing code” (a set of rules that can predict the splicing patterns of pre-mRNAs [60, 61]). In a recent study of alternative splicing across tissues, association links between genes and exons were identified through partial correlation studies [62]. This method was named pCastNet (partial Correlation analysis of splicing transcriptome Network). These association links can provide information about the regulation relationship between genes and the splicing of exons. It will help us to understand the gene regulation at an exon-level resolution.

We first introduce some notations. If the Pearson correlation coefficient is denoted as  $r_{ab}$  between variable  $a$  and variable  $b$ , the first-order partial correlation



coefficient between  $a$  and  $b$  conditioning on  $c$  is:

$$r_{ab \cdot c} = \frac{r_{ab} - r_{ac}r_{bc}}{\sqrt{(1 - r_{ac}^2)(1 - r_{bc}^2)}} \quad (2.22)$$

The second-order partial correlation coefficient between  $a$  and  $b$  conditioning on  $c$  and  $d$  is:

$$r_{ab \cdot cd} = \frac{r_{ab \cdot c} - r_{ad \cdot c}r_{bd \cdot c}}{\sqrt{(1 - r_{ad \cdot c}^2)(1 - r_{bd \cdot c}^2)}} \quad (2.23)$$

In pCastNet, three types of associations will be considered for a pair of genes: gene-gene (GG) association, exon-gene (EG) association, and exon-exon (EE) association. For GG association, the Pearson correlation coefficient is calculated between gene 1 ( $g_1$ ) and gene 2 ( $g_2$ ) and denoted as  $r_{g_1g_2}$ . For EG association, considering an exon ( $e_1$ ) of gene 1 and gene 2 ( $g_2$ ), besides the Pearson correlation coefficient  $r_{e_1g_2}$ , the first-order partial correlation coefficient between  $e_1$  and  $g_2$  conditioning on gene 1 ( $g_1$ ) is also calculated as  $r_{e_1g_2 \cdot g_1}$ . The partial correlation can be interpreted as the association between  $e_1$  and  $g_2$  after removing the effect of  $g_1$ . If the partial correlation is high, the association between  $e_1$  and  $g_2$  is not due to the correlation between  $g_1$  and  $g_2$ . For EE association, the correlation between an exon ( $e_1$ ) of gene 1 and an exon ( $e_2$ ) of gene 2 is calculated as  $r_{e_1e_2}$ . The partial correlations  $r_{e_1e_2 \cdot g_1}$ ,  $r_{e_1e_2 \cdot g_2}$ , and the second-order partial correlation coefficient  $r_{e_1e_2 \cdot g_1g_2}$  can also be calculated to exclude the possibility that the EE correlation is due to the EG or the GG correlation. In summary, if the p-value for  $r_{g_1g_2}$  is significant, a GG link between gene 1 and gene 2 can be declared. If the p-values for both  $r_{e_1g_2}$  and  $r_{e_1g_2 \cdot g_1}$  are significant, an EG link between  $e_1$  and  $g_2$  can be declared and the association is not due to GG association. If the p-values for  $r_{e_1e_2}$ ,  $r_{e_1e_2 \cdot g_1}$ ,  $r_{e_1e_2 \cdot g_2}$ , and  $r_{e_1e_2 \cdot g_1g_2}$  are significant, an EE link between the two exons  $e_1$  and  $e_2$  can be declared, and the association is not due to GG or EG associations.

The authors used the approach proposed by Efron [63] to control the expected FDR conditioning on a dependence effect parameter  $A$ . The sparseness of a network was estimated according to the conditional FDR and a threshold on the sparseness was then chosen. The sparseness of a network is defined as the percentage of true links among all possible node pairs. The threshold selection has several advantages: first, the corresponding correlation thresholds are data dependent; second, we can derive an accurate estimate of the number of falsely declared links taking into consideration the dependence among hypotheses; and third, we can integrate prior information about the sparseness of networks if this information is available.

By applying pCastNet to exon arrays in 11 human tissues, the authors found that gene pairs with exon-gene or exon-exon links tend to have similar functions or are present in the same pathways. More interestingly, gene pairs with exon-gene or exon-exon links tend to share cis-elements in promoter regions and microRNA binding elements in 3' untranslated regions, which suggests the coupling of co-alternative-splicing, co-transcription-factor-binding, and co-microRNA-binding.

## 2.5 Alternative Splicing, Genetic Variation, and Disease

Because of its important role in gene regulation, malfunction of alternative splicing has contributed to many human diseases [64–66]. Among point mutations associated with human genetic diseases in the Human Gene Mutation Database, about 9.5% of them are within splice sites and may cause RNA splicing defects [67]. In addition, many disease mutations that target synonymous and nonsynonymous amino acid codon positions often affect the exon splicing and cause function defects. It was estimated that as many as 50% of disease mutations in exons affect splicing [68]. Differential alternative splicing studies have been performed in many diseases such as cancers. For instances, altered transcript isoform levels have been detected for many genes in prostate and breast cancer without significant changes in total transcript abundance [69, 70]. In addition, a study of Hodgkin lymphoma tumors using custom alternative splicing microarrays found that the relative abundance of alternatively spliced isoforms correlates with transformation and tumor grade [71]. These studies suggest that alternative splicing profiling may provide additional tools for tumor diagnosis.

Kwan et al. also studied the heritability of alternative splicing in healthy people [72]. They investigated the alternative splicing variation among humans using exon array profiling in lymphoblastoid cell lines derived from the CEU HapMap population. Through family-based linkage studies and allelic association studies, they identified marker loci linked to particular alternative splicing events. They detected both annotated and novel alternatively spliced variants, and that such variation among individuals is heritable and genetically controlled.

## 2.6 Online Resources

At the end of this chapter, we provide a list of online databases for alternative splicing in Table 2.1. These databases collect alternative splicing events in different organisms or study the effect of alternative splicing on protein structures, RT-PCR, and so on.

## 2.7 Summary

Alternative splicing has been realized as one of the most important gene regulatory mechanisms. The related research has been reinvigorated by the availability of large amount of sequence data and high-throughput technologies. Nevertheless, many important questions regarding the function, the mechanism, and the regulation of alternative splicing remain unanswered. The statistical and computational analysis of alternative splicing has also emerged as an important and relatively new field.

**Table 2.1** Online databases for alternative splicing

Database	Description	Link
ASTD [50]	human, mouse, and rat	<a href="http://www.ebi.ac.uk/astd/main.html">http://www.ebi.ac.uk/astd/main.html</a>
PALSdb [73]	human, mouse, and worm	<a href="http://ymbc.ym.edu.tw/palsdb/">http://ymbc.ym.edu.tw/palsdb/</a>
SpliceInfo [74]	human	<a href="http://spliceinfo.mbc.nctu.edu.tw/">http://spliceinfo.mbc.nctu.edu.tw/</a>
ASmodeler [75]	human, mouse, and rat	<a href="http://genome.ewha.ac.kr/ECgene/ASmodeler/">http://genome.ewha.ac.kr/ECgene/ASmodeler/</a>
ECgene [76]	human, mouse, rat, dog, zebrafish, fruit fly, chick, rhesus, and C. elegans	<a href="http://genome.ewha.ac.kr/ECgene/">http://genome.ewha.ac.kr/ECgene/</a>
ASG [77]	human	<a href="http://statgen.ncsu.edu/asg/">http://statgen.ncsu.edu/asg/</a>
DEDB [78]	fruit fly	<a href="http://proline.bic.nus.edu.sg/dedb/">http://proline.bic.nus.edu.sg/dedb/</a>
EuSplice [79]	23 eukaryotes	<a href="http://66.170.16.154/EuSplice">http://66.170.16.154/EuSplice</a>
ASPicDB [80]	human	<a href="http://t.caspur.it/ASPicDB/">http://t.caspur.it/ASPicDB/</a>
HOLLYWOOD [81]	human and mouse	<a href="http://hollywood.mit.edu">http://hollywood.mit.edu</a>
AS-ALPS [82]	the effects of alternative splicing on protein structure, interaction and network in human and mouse	<a href="http://as-alps.nagahama-i-bio.ac.jp">http://as-alps.nagahama-i-bio.ac.jp</a>
SpliceCenter [83]	the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies	<a href="http://discover.nci.nih.gov/splicecenter">http://discover.nci.nih.gov/splicecenter</a>
SpliVaP [84]	changes in signatures among protein isoforms due to alternative splicing	<a href="http://www.bioinformatica.crs4.org/tools/dbs/splivap/">http://www.bioinformatica.crs4.org/tools/dbs/splivap/</a>

They will provide valuable information about the precisely regulated alternative splicing process and help us to advance our knowledge about the post-transcriptional regulation.

References

1. Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456, 470–476.

2. Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40, 1413–1415.

3. Gilbert, W. (1978). Why genes in pieces? *Nature*, 271, 501.

4. Breitbart, R. E., Andreadis, A., & Nadal-Ginard, B. (1987). Alternative splicing: A ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annual Review of Biochemistry*, 56, 467–495.

5. Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., & Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302, 2141–2144.

6. Kan, Z., Rouchka, E. C., Gish, W. R., & States, D. J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research*, 11, 889–900.
7. Mironov, A. A., Fickett, J. W., & Gelfand, M. S. (1999). Frequent alternative splicing of human genes. *Genome Research*, 9, 1288–1293.
8. Modrek, B., Resch, A., Grasso, C., & Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29, 2850–2859.
9. Graveley, B. R., Kaur, A., Gunning, D., Zipursky, S. L., Rowen, L., & Clemens, J. C. (2004). The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA*, 10, 1499–1506.
10. Missler, M., & Sudhof, T. C. (1998). Neurexins: Three genes and 1001 products. *Trends in Genetics*, 14, 20–26.
11. Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., Mueller, H. M., Dimopoulos, G., Law, J. H., Wells, M. A., Birney, E., Charlab, R., Halpern, A. L., Kokoza, E., Kraft, C. L., Lai, Z., Lewis, S., Louis, C., Barillas-Mury, C., Nusskern, D., Rubin, G. M., Salzberg, S. L., Sutton, G. G., Topalis, P., Wides, R., Wincker, P., Yandell, M., Collins, F. H., Ribeiro, J., Gelbart, W. M., Kafatos, F. C., & Bork, P. (2002). Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, 298, 149–159.
12. Kent, W. J. (2002). BLAT – the BLAST-like alignment tool. *Genome Research*, 12, 656–664.
13. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, 8, 967–974.
14. Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875.
15. van Nimwegen, E., Paul, N., Sheridan, R., & Zavolan, M. (2006). SPA: A probabilistic algorithm for spliced alignment. *PLoS Genetics*, 2, e24.
16. Florea, L., Di Francesco, V., Miller, J., Turner, R., Yao, A., Harris, M., Walenz, B., Mobarry, C., Merkulov, G. V., Charlab, R., Dew, I., Deng, Z., Istrail, S., Li, P., & Sutton, G. (2005). Gene and alternative splicing annotation with AIR. *Genome Research*, 15, 54–66.
17. Heber, S., Alekseyev, M., Sze, S. H., Tang, H., & Pevzner, P. A. (2002). Splicing graphs and EST assembly problem. *Bioinformatics*, 18(Suppl 1), S181–S188.
18. Kim, N., Shin, S., & Lee, S. (2005). ECgene: Genome-based EST clustering and gene modeling for alternative splicing. *Genome Research*, 15, 566–576.
19. Xing, Y., Yu, T., Wu, Y. N., Roy, M., Kim, J., & Lee, C. (2006). An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Research*, 34, 3150–3160.
20. Leparc, G. G., & Mitra, R. D. (2007). Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in *Caenorhabditis elegans* and human. *Nucleic Acids Research*, 35, 3192–3202.
21. Sorek, R., & Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research*, 13, 1631–1637.
22. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., & Shamir, R. (2004). A non-EST-based method for exon-skipping prediction. *Genome Research*, 14, 1617–1623.
23. Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T., & Burge, C. B. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 2850–2855.
24. Chen, L., & Zheng, S. (2008). Identify alternative splicing events based on position-specific evolutionary conservation. *PLoS One*, 3, e2806.
25. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
26. Parmley, J. L., Urrutia, A. O., Potrzebowski, L., Kaessmann, H., & Hurst, L. D. (2007). Splicing and the evolution of proteins in mammals. *PLoS Biology*, 5, e14.
27. Fairbrother, W. G., Holste, D., Burge, C. B., & Sharp, P. A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biology*, 2, e268.
28. Boutz, P. L., Stoilov, P., Li, Q., Lin, C. H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., Jr., & Black, D. L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & Development*, 21, 1636–1652.

29. Clark, T. A., Schweitzer, A. C., Chen, T. X., Staples, M. K., Lu, G., Wang, H., Williams, A., & Blume, J. E. (2007). Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biology*, 8, R64.
30. Yeo, G. W., Xu, X., Liang, T. Y., Muotri, A. R., Carson, C. T., Coufal, N. G., & Gage, F. H. (2007). Alternative splicing events identified in human embryonic stem cells and neural progenitors. *PLoS Computational Biology*, 3, 1951–1967.
31. Castle, J. C., Zhang, C., Shah, J. K., Kulkarni, A. V., Kalsotra, A., Cooper, T. A., & Johnson, J. M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics*, 40, 1416–1425.
32. Clark, T. A., Sugnet, C. W., & Ares, M., Jr. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, 296, 907–910.
33. Cline, M. S., Blume, J., Cawley, S., Clark, T. A., Hu, J. S., Lu, G., Salomonis, N., Wang, H., & Williams, A. (2005). ANOSVA: A statistical method for detecting splice variation from expression data. *Bioinformatics*, 21(Suppl. 1), i107–i115.
34. Purdom, E., Simpson, K. M., Robinson, M. D., Conboy, J. G., Lapuk, A. V., & Speed, T. P. (2008). FIRMA: A method for detection of alternative splicing from exon array data. *Bioinformatics*, 24, 1707–1714.
35. Wang, H., Hubbell, E., Hu, J. S., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M. A., Ares, M., Kulp, D. C., & Haussler, D. (2003). Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, 19(Suppl. 1), i315–i322.
36. Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 31–36.
37. Anton, M. A., Gorostiaga, D., Guruceaga, E., Segura, V., Carmona-Saez, P., Pascual-Montano, A., Pio, R., Montuenga, L. M., & Rubio, A. (2008). SPACE: An algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biology*, 9, R46.
38. Shai, O., Morris, Q. D., Blencowe, B. J., & Frey, B. J. (2006). Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics*, 22, 606–613.
39. Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., Morris, Q. D., Frey, B. J., & Blencowe, B. J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Molecular Cell*, 16, 929–941.
40. Fagnani, M., Barash, Y., Ip, J. Y., Misquitta, C., Pan, Q., Saltzman, A. L., Shai, O., Lee, L., Rozenhek, A., Mohammad, N., Willaime-Morawek, S., Babak, T., Zhang, W., Hughes, T. R., van der Kooy, D., Frey, B. J., & Blencowe, B. J. (2007). Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biology*, 8, R108.
41. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320, 1344–1349.
42. Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., & Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453, 1239–1243.
43. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133, 523–536.
44. Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., & Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5, 613–619.
45. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5, 621–628.

46. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18, 1509–1517.
47. Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., & Yaspo, M. L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321, 956–960.
48. Jiang, H., & Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25, 1026–1032.
49. Zheng, S., & Chen, L. (2009). A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Research*, 37, e75.
50. Stamm, S., Riethoven, J. J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N. L., & Thanaraj, T. A. (2006). ASD: A bioinformatics resource on alternative splicing. *Nucleic Acids Research*, 34, D46–D55.
51. Zhou, Z., Licklider, L. J., Gygi, S. P., & Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419, 182–185.
52. Jurica, M. S., & Moore, M. J. (2003). Pre-mRNA splicing: Awash in a sea of proteins. *Molecular Cell*, 12, 5–14.
53. Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14, 802–813.
54. Huh, G. S., & Hynes, R. O. (1994). Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element. *Genes & Development*, 8, 1561–1574.
55. McCullough, A. J., & Berget, S. M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Molecular Cell Biology*, 17, 4562–4571.
56. Chou, M. Y., Underwood, J. G., Nikolic, J., Luu, M. H., & Black, D. L. (2000). Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. *Molecular Cell*, 5, 949–957.
57. Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell*, 119, 831–845.
58. Zhang, X. H., & Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Development*, 18, 1241–1250.
59. Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, 72, 291–336.
60. Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: Towards a cellular code. *Nature Reviews. Molecular Cell Biology*, 6, 386–398.
61. Fu, X. D. (2004). Towards a splicing code. *Cell*, 119, 736–738.
62. Chen, L., & Zheng, S. (2009). Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biology*, 10, R3.
63. Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102, 93–103.
64. Faustino, N. A., & Cooper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes & Development*, 17, 419–437.
65. Garcia-Blanco, M. A., Baraniak, A. P., & Lasda, E. L. (2004). Alternative splicing in disease and therapy. *Nature Biotechnology*, 22, 535–546.
66. Blencowe, B. J. (2000). Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences*, 25, 106–110.
67. Krawczak, M., Thomas, N. S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., & Cooper, D. N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: Nature, distribution, and consequences for mRNA splicing. *Human Mutation*, 28, 150–158.
68. Blencowe, B. J. (2006). Alternative splicing: New insights from global analyses. *Cell*, 126, 37–47.
69. Li, H. R., Wang-Rodriguez, J., Nair, T. M., Yeakley, J. M., Kwon, Y. S., Bibikova, M., Zheng, C., Zhou, L., Zhang, K., Downs, T., Fu, X. D., & Fan, J. B. (2006). Two-dimensional transcriptome profiling: Identification of messenger RNA isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. *Cancer Research*, 66, 4079–4088.

70. Li, C., Kato, M., Shiue, L., Shively, J. E., Ares, M., Jr., & Lin, R. J. Cell type and culture condition-dependent alternative splicing in human breast cancer cells revealed by splicing-sensitive microarrays. *Cancer Research*, 66, 1990–1999 (2006).
71. Religio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R. B., & Valcarcel, J. (2005). Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *The Journal of Biological Chemistry*, 280, 4779–4784.
72. Kwan, T., Benovoy, D., Dias, C., Gurd, S., Serre, D., Zuzan, H., Clark, T. A., Schweitzer, A., Staples, M. K., Wang, H., Blume, J. E., Hudson, T. J., Sladek, R., & Majewski, J. (2007). Heritability of alternative splicing in the human genome. *Genome Research*, 17, 1210–1218.
73. Huang, Y. H., Chen, Y. T., Lai, J. J., Yang, S. T., & Yang, U. C. (2002). PALS db: Putative Alternative Splicing database. *Nucleic Acids Research*, 30, 186–190.
74. Huang, H. D., Horng, J. T., Lin, F. M., Chang, Y. C., & Huang, C. C. (2005). SpliceInfo: An information repository for mRNA alternative splicing in human genome. *Nucleic Acids Research*, 33, D80–D85.
75. Kim, N., Shin, S., & Lee, S. (2004). ASmodeler: Gene modeling of alternative splicing from genomic alignment of mRNA, EST and protein sequences. *Nucleic Acids Research*, 32, W181–W186.
76. Kim, P., Kim, N., Lee, Y., Kim, B., Shin, Y., & Lee, S. (2005). ECgene: Genome annotation for alternative splicing. *Nucleic Acids Research*, 33, D75–D79.
77. Leipzig, J., Pevzner, P., & Heber, S. (2004). The Alternative Splicing Gallery (ASG): Bridging the gap between genome and transcriptome. *Nucleic Acids Research*, 32, 3977–3983.
78. Lee, B. T., Tan, T. W., & Ranganathan, S. (2004). DEDB: A database of *Drosophila melanogaster* exons in splicing graph form. *BMC Bioinformatics*, 5, 189.
79. Bhasi, A., Pandey, R. V., Utharasy, S. P., & Senapathy, P. (2007). EuSplice: A unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. *Bioinformatics*, 23, 1815–1823.
80. Castrignano, T., D’Antonio, M., Anselmo, A., Carrabino, D., D’Onorio De Meo, A., D’Erchia, A. M., Licciulli, F., Mangiulli, M., Mignone, F., Pavesi, G., Picardi, E., Riva, A., Rizzi, R., Bonizzoni, P., & Pesole, G. (2008). ASPicDB: A database resource for alternative splicing analysis. *Bioinformatics*, 24, 1300–1304.
81. Holste, D., Huo, G., Tung, V., & Burge, C. B. (2006). HOLLYWOOD: A comparative relational database of alternative splicing. *Nucleic Acids Research*, 34, D56–D62.
82. Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K., & Go, M. (2009). AS-ALPS: A database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Research*, 37, D305–D309.
83. Ryan, M. C., Zeeberg, B. R., Caplen, N. J., Cleland, J. A., Kahn, A. B., Liu, H., & Weinstein, J. N. (2008). SpliceCenter: A suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. *BMC Bioinformatics*, 9, 313.
84. Floris, M., Orsini, M., & Thanaraj, T. A. (2008). Splice-mediated Variants of Proteins (SpliVaP) – data and characterization of changes in signatures among protein isoforms due to alternative splicing. *BMC Genomics*, 9, 453.



<http://www.springer.com/978-3-642-16344-9>

Handbook of Statistical Bioinformatics

Lu, H.H.-S.; Schölkopf, B.; Zhao, H. (Eds.)

2011, X, 630 p., Hardcover

ISBN: 978-3-642-16344-9