SplicingTypesAnno: annotating and quantifying alternative splicing events for RNA-Seq data

Xiaoyong Sun[a,*], Fenghua Zuo[b], Yuanbin Ru[c], Jiqiang Guo[d], Xiaoyan Yan[e], Gaurav Sablok[f]

[a]*Agricultural Big-Data Research Center, College of Information Science and Engineering, Shandong Agricultural University, Taian, Shandong, 271018, China*
[b]*College of Information Engineering, Taishan Medical University, Taian, Shandong 271000, China*
[c]*Department of Biomedical Informatics, Windber Research Institute, Windber, PA 15963, U.S.A.*
[d]*Applied statistics center, Columbia University, New York, NY 10027, U.S.A.*
[e]*Affiliated Hospital of Shandong University of Traditional Chinese Medicine, No. 42 Wenhua West Road, Jinan, Shandong 250011, China*
[f]*Plant Functional Biology and Climate Change Cluster (C3), University of Technology Sydney, PO Box 123 Broadway, NSW 2007 Australia*

Abstract

Alternative splicing plays a key role in the regulation of the central dogma. Four major types of alternative splicing have been classified as intron retention, exon skipping, alternative 5 splice sites or alternative donor sites, and alternative 3 splice sites or alternative acceptor sites. A few algorithms have been developed to detect splice junctions from RNA-Seq reads. However, there are few tools targeting at the major alternative splicing types at the exon/intron level. This type of analysis may reveal subtle, yet important events of alternative splicing, and thus help gain deeper understanding of the mechanism of alternative splicing. This paper describes a user-friendly R package, extracting, annotating and analyzing alternative splicing types for sequence alignment files from RNA-Seq. SplicingTypesAnno can: 1) provide annotation for major alternative splicing at exon/intron level. By comparing the annotation from GTF/GFF file, it identifies the novel alternative splicing sites; 2) offer a convenient two-level analysis: genome-scale annotation for users with high performance computing environment, and gene-scale annotation for users with personal computers; 3) generate a user-friendly web report and additional BED files for IGV visualization. SplicingTypesAnno is a user-friendly R package for extracting, annotating and analyzing alternative splicing types at exon/intron level for sequence alignment files from RNA-Seq. It is publically available at `https://sourceforge.net/projects/splicingtypes/files/` or `http://genome.sdau.edu.cn/research/software/SplicingTypesAnno.html`.

Keywords: R package, genome-scale annotation, gene-scale annotation, splicing junction

1. Background

Alternative splicing plays a key role in the central dogma. By skipping introns and linking the selective exons together, it acts as the essential component of the transcription and enables production of various transcripts. This process may generate different isoforms which may lead to different protein products, thus impacting the final phenotype. During this process, many epigenetic features have been proved to function as supplementary mechanisms, including DNA methylation, nucleosome occupancy, histone modifications [37]. Also, alternative splicing has been proven to be highly related to tissue and developmental stages [2, 3, 4]. The final transcript structure directly affects the protein production. It has been found that 15% mutations result in irregular function of alternative splicing, causing hereditary disease [5]. In addition, nonsense or missense mutations may also modify the alternative splicing, and in turn lead to diverse diseases [6].

Alternative splicing has four main types: intron retention, exon skipping, alternative 5 splice sites or alternative donor sites, and alternative 3 splice sites or alternative acceptor sites [7, 8]. Intron retention is the major type found in plants [9, 10, 11, 12], yeast [13] as well as fungus [14]. It is confirmed that intron retention is not spurious and has significant biological functions [7, 15]. Another type of alternative splicing, exon skipping, is proved to be most dominant type in the human and mouse [16, 17]. Kim et. al. compared eight eukaryotes species with expressed sequence tag (EST) data, and they found the percentage of exon skipping in alternative splicing grows slightly from invertebrates to vertebrates [18]. Through protein shifting, exon skipping changes the protein structure completely.

RNA-Seq, as a cutting-edge technology, can help researchers to investigate RNA with single-base resolution, including detection of novel isoforms [19, 20, 21, 22, 23]. Many tools have been developed to address diverse genomic features [24, 25, 26]. Specifically, a lot of software have been developed to detect splice junctions from RNA-Seq reads, including tophat [27], HMMsplicer [28] and spliceGrapher [29]. These tools are based on different algorithms to predict splicing junction sites. SwitchSeq [35] is a perl-based tool specifically designed for identifying the extreme cases of the alternative splicing. By comparing the expression levels from two isoforms, it provides a platform to screen, identify and visualize those special events. AltAnalyze [36] is a python tool originally de-

*Corresponding author
Email address: `johnsunx1@gmail.com` (Xiaoyong Sun)

signed for detecting alternative splicing in microarray. Recently it is extended to analyze RNA-Seq data and detect junction sites based on some statistical methods such as ASPIRE. Also, MISO provides isoform-level analysis to compare expression levels across samples [2] and utilizes Bayesian method to infer isoforms for the major splicing types from current known annotation. Alt Event Finder [37] and spliceR [38] are two similar tools to MISO, which take a different approach to identify the novel splicing events. Using the results from some transcript assembly tools (Cufflinks and Scripture), they help users to identify the novel transcripts, and generate *de novo* annotation for alternative splicing. However, they heavily depend on the results from those assembly tools, which were developed to quantify different isoforms based on some probability framework. This may overlook some fine, yet important novel splicing in the low-expression transcripts such as long non-coding RNAs [39].

SplicingTypesAnno is an R package that takes an event-based approach to detect the novel splicing directly from the aligned raw reads. By pinpointing the difference of the splicing junction at exon-level, it targets the individual splicing events with much higher resolution than any other software. Using this direct support from alignment file, it can easily discover any novel features instead of overlooking those splicing structures because of low coverage. Specially, it takes the alignment file, i.e., the bam file as input, and analyzes the raw reads through the pipeline with searching algorithms, and defines the related alternative splicing types, finally generates a user-friendly web report for users. In addition, it provides high flexibility for users to handle large set of data by genome-scale and gene-scale functions. In the genome-scale annotation, users can make use of computer clusters with parallel computing feature to speed up the multiple sample analysis. In the gene-scale annotation, users can conveniently extract the related alternative splicing events with a single laptop. The output also includes BED format in the web report, which can be easily imported to Integrative Genomics Viewer (IGV) [30] for visualization. SplicingTypesAnno can be applied to diverse species with the GTF files.

2. Methods

2.1. Alternative splicing types and subtypes

Alternative splicing includes exon skipping, intron retention, alternative donor sites, alternative acceptor sites, and both 5' and 3' alternative sites. There are also some other forms with small percentage, which is not discussed in this software. For EST data, researchers define these alternative splicing events based on two alternative isoforms [31, 32]. For RNA-Seq data with large quantities of short reads, the definition for these events is adapted to capture these biological features. [33] has discussed the alternative splicing events within the framework of RNA-Seq. Based on these concepts, we design our algorithms completely based on the structural properties of junction reads to define alternative splicing types. In addition, we use the known splicing sites as reference set. If the splicing sites derived from junction reads match the reference set, we consider them as "known sites"; otherwise as "novel sites". If the splicing sites are known, but the link between two sites is not reported from the annotation file, it is marked as "novel splicing link". The detailed algorithms for extracting alternative splicing events are as follows,

2.1.1. Junction reads and non-junction reads

- junction reads: the reads that have gaps when they are aligned to the reference genome. Generally they span across the splicing junctions. One part of the read matches one exon, and the other part matches the following exon.
- non-junction reads: the reads that do not have any gaps when they are aligned to the reference genome.

2.1.2. Splicing types

The description of the major splicing types is shown in Figure 1.

- Intron retention: 1) junction reads: there must be enough junction reads (> minimum number of required reads); 2) non-junction reads: there must be enough non-junction reads within the junction reads discussed above (> minimum number of required reads). In addition, the sequence contig collapsed from these non-junction reads must be more than specified percentage.
- Exon skipping: 1) junction reads: there must be three types of junction reads: one junction-read type with long gap and the other two junction-read types with short gap. These two short junction-read types are mutually exclusive and both fully contained in the long junction read type. The number of junction reads must be more than minimum number of required reads; 2) non-junction reads: there must be enough non-junction reads overlapping with the interval between two short junction-read types.
- Alternative donor sites: there are two types of junction reads, which have the same 5' junction sites, but are different in the 3' junction sites.
- Alternative acceptor sites: there are two types of junction reads, which have the same 5' junction sites, but are different in the 3' junction sites.

2.1.3. Subtypes

We further divide these alternative splicing types to two different subtypes: type I and type II (Figure 2) by comparing read information with the annotation files. Generally type I only consists of one intron or exon structure; type II consists of more than one intron or exon structures. More specifically, type I of exon skipping describes the alternative splicing events that only one exon is skipped; type II of exon skipping describes the events that multiple exons are skipped. Type I of intron retention defines the events that only one intron is retained; type II defines those that more than one intron is retained. Type I of the alternative donor or acceptor sites covers the events across only one intron; while type II defines other cases.

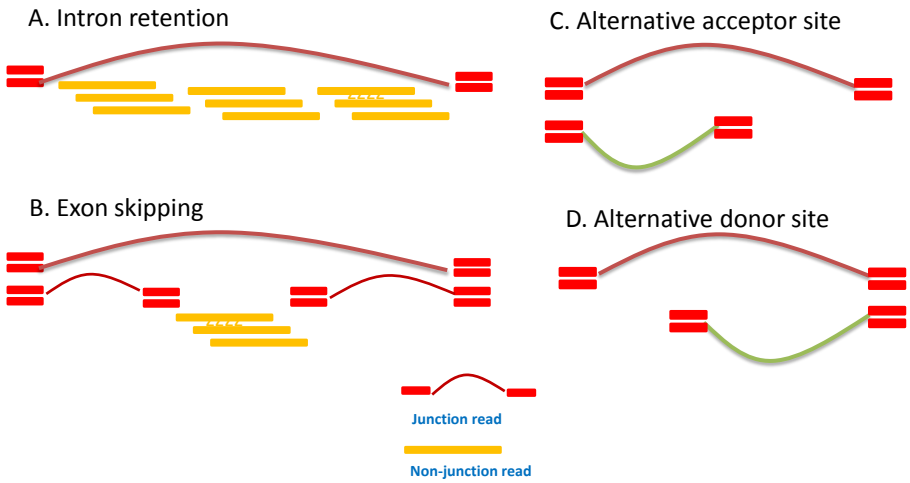

Figure 1: Major splicing types inferred from junction and non-junction reads, including intron retention (A), exon skipping (B), alternative acceptor site (C), alternative donor site (D). Junction reads have gaps, which are represented by curve. Non-junction reads do not have gaps. Red represents the junction reads; yellow represents the non-junction reads.

2.2. Genome-scale annotation and gene-scale annotation

SplicingTypesAnno is designed specifically for RNA-Seq data. To help users achieve more flexibility, SplicingTypesAnno provides two level analysis: genome-scale annotation and gene-scale annotation. The genome-scale annotation processes the RNA-Seq data for all samples, and generates a user-friendly summary report. On the other hand, the gene-scale analysis only explores the RNA-Seq data for one gene or a few genes. The advantage of this analysis is that it helps users to explore the events with deeper resolution with limited computing resources. To visualize the results, the web report from both analyses provides BED file for importing into Integrative Genomic Browser (IGV).

2.3. Novel and known splicing sites

SplicingTypesAnno identifies both novel and known splicing sites (Figure 3). This process includes two steps: 1) the first step identifies all related alternative splicing events and extracts proper splicing sites; 2) the second step marks those novel splicing sites, including novel donor sites, novel acceptor sites and novel splicing links of two sites (exon skipping does not have novel splicing links). This step is achieved by comparing the splicing sites from the first step to the boundaries of exons extracted from the GTF/GFF file.

2.4. Metrics for measuring alternative splicing

2.4.1. Normalized counts

We modify RPKM (Reads per kilobase per million mapped reads) as follows,

$$rpkm_{non-junctionReads} = \frac{readCounts}{eventLength} * \frac{10^6}{totalReads} \quad (1)$$

$$rpkm_{junctionReads} = \frac{readCounts}{readLength * 2} * \frac{10^6}{totalReads} \quad (2)$$

Specifically, for intron retention, the *eventLength* is the width of the intron inferred from the junction reads; for exon skipping, the *eventLength* is the width of the exon inferred from the junction reads; for alternative donor/acceptor site, only junction reads are utilized to identify these events, so *eventLength* is not calculated.

2.4.2. isoform percentage

We estimate the isoform percentage using the read counts at exon/intron level:

$$isoformpercentage = \frac{isoform1}{isoform1 + isoform2} \quad (3)$$

Specifically,

- for intron retention,

$$isoform1 = rpkm_{non-junctionReads} \quad (4)$$

$$isoform2 = rpkm_{junctionReads} \quad (5)$$

- for exon skipping,

$$isoform1 = rpkm_{junctionReads} \quad (6)$$

$$isoform2 = rpkm_{non-junctionReads} \quad (7)$$

- for alternative donor/acceptor sites:

$$isoform1 = rpkm_{alterSite1\_junctionReads} \quad (8)$$

$$isoform2 = rpkm_{alterSite2\_junctionReads} \quad (9)$$

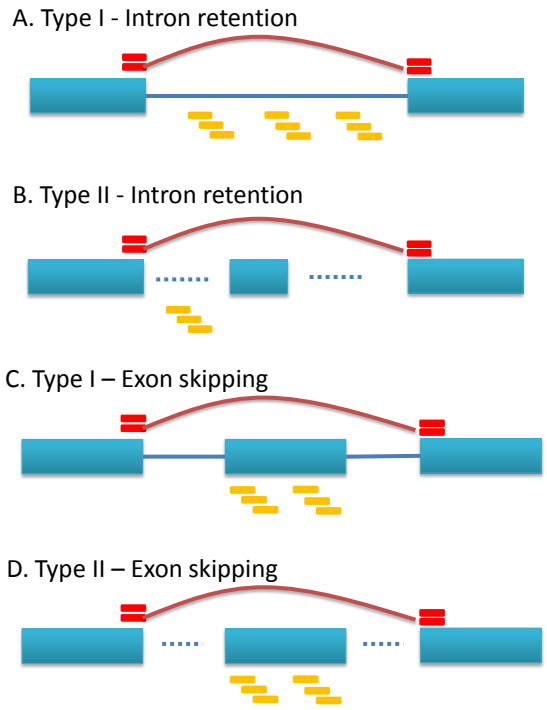

Figure 2: Subtypes for intron retention (A and B) and exon skipping (C and D). Two subtypes: type I and type II are defined based on the number of related introns or exons. Exon, junction-reads and non-junction reads are represented by blue, red and yellow colors, respectively. Solid line: there are only introns between the two exons. Dash line: some exons may exist between the two exons.

Since the equation uses the percentage approach, the ratio is not biased towards most abundant transcripts. As a matter of fact, this approach can identify the finest difference of the two isoforms. For example, the ratio of isoform 1 with 10 reads to the isoform 2 with 10 reads is equal to that of isoform 1 with 10000 reads to the isoform 2 with 10000 reads.

3. Implementation

SplicingTypesAnno is an easy to use and customized pipeline, which is capable of processing RNA-Seq read alignment files from single sample or multiple samples or multiple time points for the identification, accurate calling and annotation of the novel alternative splicing sites (Figure 4). The implemented workflow in the R-package consists of three main steps: 1) The first step deals with the filtering of the data using the following criteria: a) there should be at least two junction reads with the same beginning or ending gap boundary; b) the junction reads with one gap, i.e., one "N" in cigar string, are considered as quality reads; 2) searching the alternative splicing events; 3) generating user-friendly web report, summarizing the following descriptive information: sample description, read summary, splicing summary, alternative donor sites (type I), alternative donor sites (type II), alternative acceptor sites (type I), alternative acceptor sites (type II), BED files for visualization. To make the algorithm usable on multiple platform and High Performance Computing (HPC) SplicingTypesAnno supports parallel computing based on R package "snowfall". This configuration helps to facilitate process multiple samples when users have access to HPC environment.

To speed up the analysis, SplicingTypesAnno supports parallel computing based on R package snowfall. This configuration helps to facilitate process multiple samples when users have access to high-computing environment.

The main features of this package (Table 1) include: 1) translating GTF/GFF file to exon/intron structure; 2) counting the reads for exon/intron structure; 3) calculating and annotating the major splicing types. The first feature is implemented by `translateGTF`; the second feature requires two functions: `splicingCount` and `combineCount` and the last feature calls two functions: `splicingGene` and `splicingReport` to generate final conclusive reports.

4. Results

4.1. Data sets

To illustrate how to use this package, one data set from NCBI (SRA, GSE26561) was downloaded and aligned to mouse genome (mm9) with tophat [27]. This research studied the alternative splicing in nonsense-mediated mRNA decay-deficient mouse tissues. We selected two samples: SRR094623 (wild type sample) and SRR094624 (knockout sample) for the following demonstration.

4.2. Case study 1: managing GTF/GFF file

The main function for managing GTF/GFF file is `translateGTF`. This function takes GTF/GFF file as input, ex-

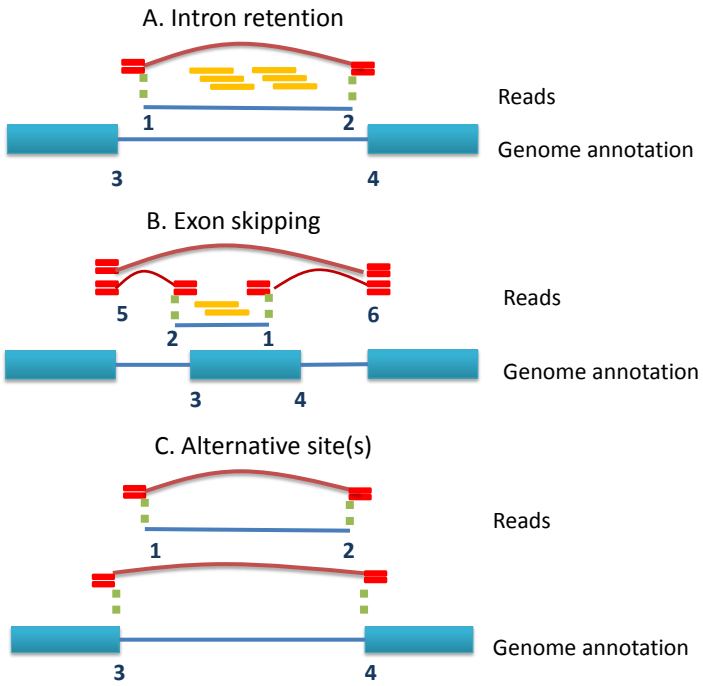

Figure 3: Major splicing types inferred from reads versus genome annotation from GTF/GFF. Nomenclature of splicing sites used in the study - 1: novel donor exon boundary; 2: novel acceptor exon boundary; 3, 4, 5, 6: known exon boundary. Exon, junction-reads and non-junction reads are represented by blue, red and yellow colors, respectively.

tracts all exon information, and produces gene, exon, and intron information as the GRanges objects.

Currently SplicingTypesAnno supports both GTF and GFF file formats, which include exon, gene or mRNA information. SplicingTypesAnno only utilizes exon information to generate exon, intron and gene annotation. To reduce ambiguity, all overlapping exons are reduced to contigs, and these contigs are used to generate genome-scale intron annotation. These intron annotations are only imported as a reference set to identify subtypes of major splicing types.

Some genes may overlap with each other in the chromosome, i.e. they share some genomic regions in the same strand. It is reported in human, mouse, rat, flies and fish [34]. Since these genes are only small percentage of all genes, we setup a parameter called "geneOverlap to control this status. In the genome-scale annotation, to reduce ambiguity, gene overlapping status is configured as "no, i.e., the overlapping genes are overlooked. In gene-scale annotation, users can determine the gene overlapping status based on specific research goals, and the following analysis is processed for each gene sequentially.

To convert the GTF/GFF file to proper gene/exon/intron objects, the following scripts are utilized,

```
library(SplicingTypesAnno)
mm9.pnpla7.gtfFile <- system.file("extdata",
                  "mm9_Pnpla7.gtf",
                  package="SplicingTypesAnno")
result.GRange <- translateGTF(mm9.pnpla7.gtfFile,
                  gtfTranscriptLabel=
                  "transcript_id")
```

4.3. Case study 2: gene-scale annotation

Gene-scale annotation is a handy tool for users who would like to check some genes quickly or do not have access to high-computing environments. It only processes the selected genes one by one, and thus does not have high requirement for hardware. Also, since it works with one gene at one time, more comprehensive analysis is provided, including determining the novel splicing links, quantifying the overlapping genes, etc.

This type of analysis includes two steps: 1) converting GTF/GFF file to gene/exon/intron structure; 2) quantifying and annotating reads at the gene-scale. The first step has been described in case study 1. The second step consists of three separate features: 1) calculating raw reads for gene, exon and intron; 2) annotating alternative splicing for single gene; 3) generating web-report for a few selected genes. The first two features are for exploratory analysis while the third feature is to generate a user-friendly and comprehensive profile hosting summary and visualization information. This web report is convenient for storage, transfer and archives.

The first feature targets quantification for gene, exon and intron. There are two functions (`splicingCount`, `combineCount`) specifically designed for this analysis. Based on the gene/exon/intron structure from `translateGTF`, `splicingCount` calculates the raw reads within gene, exon and intron respectively for single sample.

```
bam.1 <- system.file("extdata",
         "liver_ctr.sort.bam",
         package="SplicingTypesAnno")
```

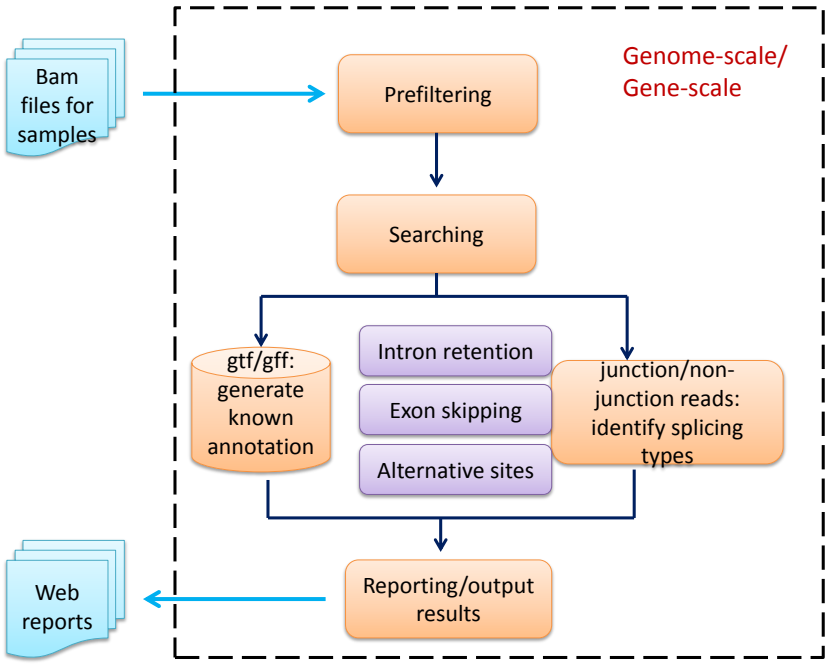

Figure 4: Schematic workflow. Working pipeline in SplicingTypesAnno. It processes alignment files (bam file) and outputs a user-friendly web report. Blue color represents the input files and output results.

```
bam.2 <- system.file("extdata",
         "liver_ko.sort.bam",
         package="SplicingTypesAnno")
selectGene <- "Pnpla7"
sample1.c <- splicingCount(selectGene,
             bam.1, result.GRange,
             sampleName="liver_ctr",
             sampleID=1)
sample2.c <- splicingCount(selectGene,
             bam.2, result.GRange,
             sampleName="liver_ko",
             sampleID=2)
```

`combineCount` combines the output from `splicingCount` and generates the count data for all samples.

```
sList.c <- list(sample1.c, sample2.c)
ccount <- combineCount(sList.c )
```

The second feature is for annotating alternative splicing for single gene. This feature is achieved by `splicingGene` function. By specifying certain alternative splicing types, `splicingGene` searches, quantifies and extracts all related information for single sample and single gene.

```
selectGene <- c("Pnpla7")
sample1 <- splicingGene(selectGene,
           bam.1, result.GRange,
           sampleName="liver_ctr", sampleID=1)
sample2 <- splicingGene(selectGene,
           bam.2, result.GRange,
           sampleName="liver_KO", sampleID=2)
```

`combineGene` combines the output from `splicingGene` and generates alternative splicing information for all samples.

```
sList <- list(sample1$type, sample2$type)
ri.1 <- combineGene(sList,
        splicingEvents="ri.1", selectGenes)
ri.2 <- combineGene(sList,
        splicingEvents="ri.2", selectGenes)
```

The last feature generates web-report for a few selected genes. `splicingReport` utilizes `splicingGene` to process selected genes one by one and then generates the summary report.

```
sampleList <- list(SampleName=c("liver_ctr",
                   "liver_ko"),
                    BamFiles=c(bam.1, bam.2),
                    SampleID=c(1,2))
splicingReport(inputData=sampleList,
               gtfFile=mm9.pnpla7.gtfFile,
               selectGenes=c("Pnpla7"))
```

4.4. Case study 3: genome-scale annotation

Based on parallel computing provided by R package snowfall, genome-scale annotation is designed for multiple samples. Users can configure the number of cpu available for this type of analysis. SplicingTypesAnno takes full advantage of the vector operation, and the memory requirement increases when users deal with the alignment files with large physical sizes.

```
sampleList <- list(SampleName=c("liver_ctr",
```

Table 1: Main functions of SplicingTypesAnno

Function names	Analysis type	Description
translateGTF	genome-scale annotation, gene-scale annotation	translate GTF/GFF file to exon/intron structure.
splicingCount	gene-scale annotation	count the reads within gene, exon and intron for selected gene.
combineCount	gene-scale annotation	combine the results from `splicingCount` for all samples.
splicingGene	gene-scale annotation	annotate and quantify the splicing types for single gene.
combineGene	gene-scale annotation	combine the results from `splicingGene` for all samples.
splicingReport	genome-scale annotation, gene-scale annotation	generate web report for splicing types.

```
                    "liver_ko"),
                    BamFiles=c(bam.1, bam.2),
                    SampleID=c(1,2))
splicingReport(inputData=sampleList,
                gtfFile=mm9.pnpla7.gtfFile,
                parallel=TRUE, cpus=2)
splicingShowReport()
splicingCleanReport()
```

The output includes a folder containing one index.html and other subfolders. Index.html is the main webpage for the final report (Figure 5), which consists of sample information (Figure 6), read summary, splicing summary, intron retention - type I and II, exon skipping - type I and II, alternative donor sites - type I and II, alternative acceptor sites type I and II (Figure 7), IGV visualization. `splicingShowReport` helps user to open the web report, and `splicingCleanReport` deletes the related report in the current working directory. In addition, the parameters ("mergeID", "nonjun", "jun", "ratio", etc) in the output have been explained in the vignettes attached in the package. Some results ("nonjun", "jun", etc) are available only in the report as csv format because of the size constraints of web format.

The visualization summary consists of a list of BED files (Figure 8), which can be imported into IGV. Each record in the BED file is one alternative splicing event, represented by a block in the IGV. Specifically, for intron retention, the block is the real retained intron(s) inferred from the gapped reads; for exon skipping, the block is the real skipped exon(s); for alternative donor/acceptor sites, there are a few blocks matching each alternative splicing event while one block represents a real intron with one alternative site. Additional information (isoform percentage, sample name, etc) for the alternative splicing event is also available for users by enabling popup text in data panels of IGV.

5. Conclusions

In this paper, we present a new R package for annotating and quantifying alternative splicing events at exon/intron level in RNA-Seq data. It can be applied to diverse species with the GTF files. The main purpose of this software is to function as the post-processing tool after RNA-Seq alignment. The main advantages of this package are as follows: 1) It provides annotation for major alternative splicing at exon/intron level. By comparing with the annotation from GTF/GFF file, it identifies the novel alternative splicing sites. In particular, since this tool uses the ratio of two isoforms, the results are not biased towards most abundant transcripts. 2) It offers a convenient two-level analysis: genome-scale annotation and gene-scale annotation. Users can analyze alternative splicing events with rather low computing requirements. A simple laptop enables users to explore alternative splicing for multiple samples. On the other hand, this package is also designed to support parallel computing, which helps speed up analysis for multiple samples at global scale. 3) It generates a user-friendly web report and additional BED files for IGV visualization.

There are some limitations of this work. The exon annotation is directly from the user input (gtf file). As a result, it does not recognize the exon structure in the novel genomic features. In addition, the software supports only exon-level annotation for alternative splicing. It does not address the isoform identification. Finally, instead of functioning as a statistical tool, it is an annotating tool for alternative splicing in RNA-Seq data. This package serves as a handy and convenient middleware between alignment software and end users.

6. Availability and requirements

- Project name: SplicingTypesAnno
- Project home page: http://sourceforge.net/projects/splicingtypes/
- Operating systems: platform independent
- Programming language: R
- other requirements: R 2.15.0
- License: GNU GPL

Figure 5: An example of the web report (results from the web page produced by **splicingReport** function). It includes the following sections: sample information, read summary, splicing summary, intron retention - type I and II, exon skipping - type I and II, alternative donor sites - type I and II, alternative acceptor sites type I and II, IGV visualization.

- Any restrictions to use by non-academics: license needed

7. Competing interests

The authors declare that they have no competing interests

8. Acknowledgements

This work was supported by Shandong Agricultural University (start-up grant to X.S.). We thank three anonymous reviewers for their constructive advices. We also thank National Supercomputing Center in Jinan for computing assistance and Nanjing Genmart Biotech Corporation for technical support.

References

[1] Zhou Y, Lu Y, Tian W: **Epigenetic features are significantly associated with alternative splicing**. *BMC Genomics* 2012, **13**:123.

[2] Wang E, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore S, Schroth G, Burge C: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**:470–476.

[3] Graveley B, et al: **The developmental transcriptome of Drosophila melanogaster**. *Nature* 2010, **471**:473–479.

[4] Bradley R, Merkin J, Lambert N, Burge C: **Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution**. *PLoS Biol* 2012, **10(1)**:e1001229.

[5] Krawczak M, Reiss J, Cooper D: **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.

[6] Caceres J, Kornblihtt A: **Alternative splicing: multiple control mechanisms and involvement in human disease**. *Trends Genet* 2002, **18(4)**:186–193.

[7] Galante P, Sakabe N, Kirschbaum-Slager N, de Souza S: **Detection and evaluation of intron retention events in the human transcriptome**. *RNA* 2004, **10(5)**:757–65.

[8] McGuire A, Pearson M, Neafsey D, Galagan J: **Cross-kingdom patterns of alternative splicing and splice recognition**. *Genome Biol* 2008, **9(3)**:R50.

[9] Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R: **Intron retention is a major phenomenon in alternative splicing in Arabidopsis**. *Plant J* 2004, **39**:877–885.

[10] Wang B, Brendel V: **Genomewide comparative analysis of alternative splicing in plants**. *Proc Natl Acad Sci USA* 2006, **103**:7175–7180.

[11] Walters B, Lum G, Sablok G, Min XJ: **Genome-wide landscape of alternative splicing events in Brachypodium distachyon**. *DNA research* 2013, **20**:163–171.

[12] Sablok G, Gupta PK, Baek JM, Vazquez F, Min XJ: **Genome-wide survey of alternative splicing in the grass Brachypodium distachyon: a emerging model biosystem for plant functional genomics**. *Biotechnology letters* 2011, **33**:629–636.

[13] Loftus B, et al: **The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans**. *Science* 2005, **307**:1321–1324.

[14] Kupfer D, Drabenstot S, Buchanan K, Lai H, Zhu H, Dyer D, Roe B, Murphy J: **Introns and splicing elements of five diverse fungi**. *Eukaryot Cell* 2004, **3**:1088–1100.

[15] Bell T, Miyashiro K, Sul J, Buckley P, Lee M, McCullough R, Jochems J, Kim J, Cantor C, Parsons T, Eberwine J: **Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations**. *Proc Natl Acad Sci USA* 2010, **107(49)**:21152–7.

[16] Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes**. *Nucleic Acids Res* 2001, **29**:2850–2859.

[17] Sugnet C, Kent W, Ares MJ, Haussler D: **Transcriptome and genome conservation of alternative splicing events in humans and mice**. *Pac Symp Biocomput* 2004, :66–77.

[18] Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes**. *Nucleic Acids Res* 2007, **35(1)**:125–131.

[19] Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using**

Annotation of the samples

SampleName	BamFiles	SampleID
liver_ctr	/home/xsun1/R/x86_64-unknown-linux-gnu-library/2.14/SplicingTypesAnno/extdata/liver_ctr.sort.bam	1
liver_ko	/home/xsun1/R/x86_64-unknown-linux-gnu-library/2.14/SplicingTypesAnno/extdata/liver_ko.sort.bam	2

Summary for splicing events

SampleName	SampleID	intronRetention.Type1	intronRetention.Type2	exonSkipping.Type1	alternativeDonor.Type1	alternativeAcceptor.Type1
liver_ctr	1	2	1	0	0	0
liver_ko	2	2	1	1	2	4

reads summary

SampleName	SampleID	Pnpla7.totalRead	Pnpla7.qualityJunRead	Pnpla7.qualityNonJunRead
liver_ctr	1	1562	867	690
liver_ko	2	2046	960	1074

Figure 6: An example for sample information, reads summary and splicing summary (results from the web page produced by **splicingReport** function). Two samples (SRR094623 and SRR094624) were included in the case study.

RNA-seq. *Curr Opin Microbiol* 2010, **13(5)**:619–624.

[20] Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10(1)**:57–63.

[21] Qian X, Ba Y, Zhuang Q, Zhong G: **RNA-Seq technology and its application in fish transcriptomics**. *OMICS* 2014, **18(2)**::98–110.

[22] Marguerat S, Bhler J: **RNA-seq: from technology to biology**. *Cell Mol Life Sci* 2010, **67(4)**:569–579.

[23] Sun X, Yang Q, Deng Z, Ye X: **Digital inventory of Arabidopsis transcripts revealed by 61 RNA sequencing samples**. *Plant Physiol* 2014, **166(2)**:869–78.

[24] Bao R, Huang L, Andrade J, Tan W, Kibbe WA, Jiang H, Feng G.: **Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing**. *Cancer Inform* 2014, **13(Suppl 2)**:67–82.

[25] Guzzi PH, Cannataro M: **Micro-Analyzer: automatic preprocessing of Affymetrix microarray data**. *Comput Methods Programs Biomed* 2013, **111(2)**:402–409.

[26] Kontopoulos DG, Glykos NM: **Pinda: a web service for detection and analysis of intraspecies gene duplication events**. *Comput Methods Programs Biomed* 2013, **111(3)**:711–714.

[27] Trapnell C, Pachter L, Salzberg S: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25(9)**:1105–1111.

[28] Dimon M, Sorber K, DeRisi J: **HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data**. *PLoS One* 2010, **5(11)**:e13875.

[29] Rogers M, Thomas J, Reddy A, Ben-Hur A: **SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data**. *Genome Biol* 2012, **13(1)**:R4.

[30] Robinson J, Thorvaldsdottir H, Winckler W, Guttman M, Lander E, Getz G, Mesirov J: **Integrative Genomics Viewer**. *Nature Biotechnology* 2011, **29**:24–26.

[31] Campbell M, Haas B, Hamilton J, Mount S, Buell C: **Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis**. *BMC Genomics* 2006, **7**:327.

[32] English A, Patel K, Loraine A: **Prevalence of alternative splicing choices in Arabidopsis thaliana**. *BMC Plant Biol* 2010, **10**:102.

[33] Katz Y, Wang E, Airoldi E, Burge C: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nature Methods* 2010, **7**:1009–1015.

[34] Sanna CR, Li WH, Zhang L: **Overlapping genes in the human and mouse genomes**. *BMC Genomics* 2008, **9**:169.

[35] Mar G, Alvis B: **Identification, annotation and visualisation of extreme changes in splicing from RNA-seq experiments with SwitchSeq** *bioRxiv* 2014.

[36] Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin BR, Albrecht M: **AltAnalyze and DomainGraph: analyzing and visualizing exon expression data** *Nucleic Acids Res.* 2010, **38(Web Server issue)**:W755-62.

[37] Zhou A, Breese MR, Hao Y, Edenberg HJ, Li L, Skaar TC, Liu Y: **Alt Event Finder: a tool for extracting alternative splicing events from RNA-seq data**. *BMC Genomics* 2012, **13(Suppl 8)**:S10.

[38] Vitting-Seerup K, Porse BT, Sandelin A1, Waage J: **spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data**. *BMC Bioinformatics* 2014, **15**:81.

[39] Sun X, Lin SM, Yan X: **Computational evidence of NAGNAG alternative splicing in human large intergenic noncoding RNA**. *BioMed Research International* 2014, 736798.

	mergeID	geneName	liver_ko.2.countOnly	liver_ko.2.junLeftEnd	liver_ko.2.junRightStart
2	chr2:24907704-24907828	gene_id Pnpla7; gene_name Pnpla7; p_id P1	111	24907704	24907828
3	chr2:24907704-24908073	gene_id Pnpla7; gene_name Pnpla7; p_id P1	16	24907704	24908073
4	chr2:24908157-24908915	gene_id Pnpla7; gene_name Pnpla7; p_id P1	74	24908157	24908915
5	chr2:24908157-24908852	gene_id Pnpla7; gene_name Pnpla7; p_id P1	10	24908157	24908852

	liver_knockOut.2.countSum	liver_knockOut.2.countMax	liver_knockOut.2.junLeftEndCollection	liver_knockOut.2.junRightStartCollection
2	127	111	24907704	24907828_24908073,111_16
3	127	111	24907704	24907828_24908073,111_16
4	84	74	24908157	24908915_24908852,74_10
5	84	74	24908157	24908915_24908852,74_10

	liver_knockOut.2.ratio	liver_knockOut.2.realJunLocus	liver_knockOut.2.note2	novelStart	novelEnd
2	0.874	chr2:24907704-24907828	0.874=24907828_24908073,111_16_liver_knockOut_adright.1	0	0
3	0.126	chr2:24907704-24908073	0.126=24907828_24908073,111_16_liver_knockOut_adright.1	0	0
4	0.881	chr2:24908157-24908915	0.881=24908915_24908852,74_10_liver_knockOut_adright.1	0	0
5	0.119	chr2:24908157-24908852	0.119=24908915_24908852,74_10_liver_knockOut_adright.1	0	24908852

Figure 7: The events for alternative acceptor sites (results from the web page produced by **splicingReport** function). The results include the following columns: mergeID, geneName, countOnly, junLeftEnd, junRigthStart, countSum, countMax, junLeftEndCollection, junRightStartCollection, ratio, realJunLocus, note2, novelStart, novelEnd. In the figure, the results for the knockout sample are shown for demonstration. The detailed descriptions for all these columns are available at vignettes with the R package.

Figure 8: Alternative splicing events visualized in IGV with BED files. 1, 3: genome annotation track for intron retention; 2: genome annotation track for exon skipping; 4: genome annotation track for alternative donor sites.