spliceR: An R package for classification of alternative splicing and prediction of coding potential from deconvoluted RNA-seq data

Kristoffer Vitting-Seerup[1,2,3,4], Bo Torben Porse[1,2,3], Albin Sandelin[2,4,*] and Johannes Waage[1,2,3,4,*]

[1]The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, DK2200 Copenhagen, Denmark

[2]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, DK-2200 Copenhagen, Denmark

[3] The Danish Stem Cell Centre (DanStem) Faculty of Health Sciences, University of Copenhagen, DK2200 Copenhagen Denmark

[4]The Bioinformatics Centre, University of Copenhagen, DK2200, Copenhagen, Denmark

ABSTRACT

Summary: With the advent of increasing depth and decreasing costs in digital gene expression technologies exemplified by RNA-sequencing, researchers are now able to profile the transcriptome with unprecedented detail. These advances not only allow for precise approximation of gene expression levels, but also for characterization of alternative isoform usage/switching between samples. Recent software improvements in full transcript deconvolution prompted us to develop `spliceR`, an R package for classification of alternative splicing. `spliceR` labels isoforms based on full-length transcripts from output by RNA-seq assemblers, detecting single- and multiple exon skipping, alternative donor or acceptor sites, intron retention, alternative first or last exon usage, and mutually exclusive exon events. Alongside, isoform fraction switch values are calculated for effective post-filtering, and genomic coordinates of differentially spliced elements are annotated for use in downstream sequence analysis. Furthermore, `spliceR` has the option to predict the coding potential and thereby the nonsense mediated decay (NMD) sensitivity of transcripts based on stop codon position.

Availability and Implementation: `spliceR` is implemented as an R package, is freely available from the Bioconductor repository (http://bioconductor.org/packages/devel/bioc/html/spliceR.html).

Contact: johannes.waage@gmail.com

Alternative splicing is an important part of the multi-layered process of RNA processing, elevating the potential number of unique products with orders of magnitude. More than 95% of all human genes undergo alternative splicing, and this is thought be a key element in driving the phenotypical complexity of mammals (Pan *et al*., 2008). Recent advances in sequencing technology of RNA (RNA-seq), combined with modern RNA-seq transcript assemblers, such as Cufflinks (Trapnell *et al*., 2010), now allows for describing the diverse RNA landscape with high resolution.

Here we present an easy-to-use tool, `spliceR`, which builds upon common RNA-seq assembly workflows by annotating multiple transcripts from the same gene with alternative splicing classes. After importing transcript data from Cufflinks or another full-length transcript assembler, alternative splicing events are classified by comparing each transcript against either the parent gene's hypothetical pre-mRNA (based on all isoforms) or against the gene's most expressed transcript. Additionally, `spliceR` allows for the characterization of the protein coding potential of each transcript by translating the full exonic sequence of each transcript with supported annotated open reading frames (ORFs). Compared to similar tools, `spliceR` offers full classification of multiple events, outputs genomic locations of differentially spliced elements for downstream sequence analysis and motif finding applications, and is fully based on object types found in the Bioconductor (Gentleman *et al*., 2004) project, such as GRanges, allowing for full flexibility and modularity, including support for all species supported in the Bioconductor annotation packages.

Classification of alternative splicing: `spliceR` takes full-length transcript information either from Cufflinks, or from data generated by any RNA-seq assembler that outputs fully deconvoluted transcripts. `spliceR` supports a number of filters, letting the user choose to classify alternative splicing only from those transcripts that have passed set of qualifiers, either given by cufflinks (transcript and/or gene model confidence), or based on NMD-sensitivity or expression thresholds.

In some scenarios, especially when looking at changes between samples where splicing patters are expected to change, users may opt to configure `spliceR` to use the major isoform as the reference transcript instead of the theoretical pre-RNA. Based on this comparison `spliceR` outputs a complete classification of splicing events as seen in figure 1a. Additionally, `spliceR` calculates isoform fraction (IF) values, representing the relative contribution of an isoform to the total expression of the parent gene, as well as the delta-IF (dIF), the absolute change in IF values between samples, indicating the change in relative importance of this isoform, and facilitating the identification of isoform switching. Examples of how these values are calculated can be found in supplementary table 1. For statistical assessment of differential splicing, users can access the isoform fidelity status and p-values of Cufflinks, or can easily apply other R-packages tailored for this purpose to the data.

*To whom correspondence should be addressed.

A

	Event type	No. of events
	Exon skipping/inclusion (ESI)	11,356
	Mult. exon skipping/inclusion (MESI)	3,670
	Intron skipping/inclusion ISI	4,336
	Alternative 5' splice site A5	9,754
	Alternative 3' splice site A3	9,129
	Alternative transcription start site ATSS	14,434
	Alternative transcription termination site ATTS	12,880
	Mutually exclusive exons MEE	206
	All events	65,765

B

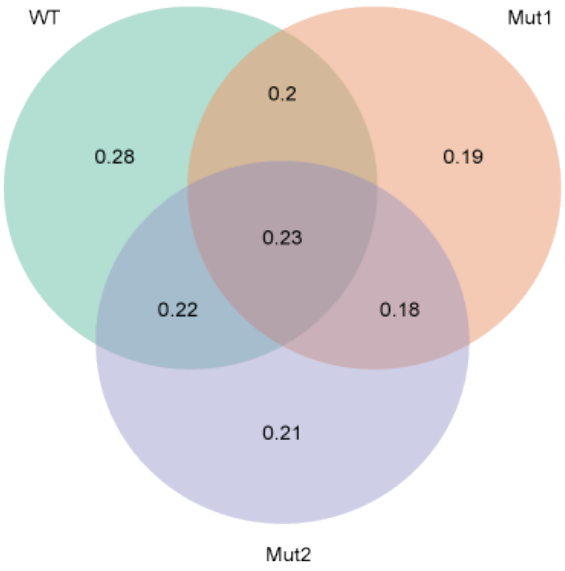

Figure 1: To demonstrate our tool and the visualization capabilities, we ran spliceR on published RNA-seq data from patients with mutations in splice factor SF3B1 (Visconte et al., 2012). A) Table displaying the total number of found alternative splicing events for all samples, based on a total of 10,042 genes found confidently expressed by cufflinks. B) An example of output from spliceR's visualization tools; a Venn diagram showing the mean number of multi exon skipping events (MESI) per transcript, as classified by spliceR from full transcript models. Overlaps between samples represent events in isoforms expressed in more than one sample.

Finally, the output of `spliceR` facilitates various downstream analyses, including filtering on isoforms that have major changes between samples, filtering for specific splicing classes, or sequence analysis on elements that are spliced in or out between samples. The latter could include detection of enriched motifs in or surrounding such elements, or identification of protein domains that are spliced in/out. `spliceR` facilitates this type of analysis by outputting the genomic coordinates of each alternatively spliced element, for each event type.

Visualization: To analyze global trends in splicing, `spliceR` can produce a range of Venn diagrams, showing the overlap of splicing events between samples for either a specific type of alternative splicing or for all events. An example is shown in figure 1b.

Analysis of coding potential: For assessment of coding potential, `spliceR` initially retrieves the genomic exon sequence using a BSGenome object. Next, ORF annotation is retrieved from the UCSC Genome Browser repository from either Refseq, Known Gene or Ensembl, or a custom ORF-table. Finally, the RNA sequences of input transcripts are assembled, and if a compatible ORF exists, translated, and positional data about the stop codon, including distance to final exon-exon junction, is recorded and returned to the user. Based on the generally accepted 50 nt rule in the NMD literature (Weischenfeldt *et al*., 2012), transcripts are marked NMD-sensitive if the stop codon falls more than 50 nt upstream of the final exon-exon junction, although this setting is user-configurable.

Conclusion: We present a Bioconductor package `spliceR`, which harnesses the power of current RNA-seq and assembly technologies, and provides a full overview of alternative splicing events and protein coding potential of transcripts. `spliceR` is flexible and easy integrated in existing workflows, supporting input and output of standard Bioconductor data types, and facilitates downstream analyses of differentially spliced elements by supplying genomic locations.

REFERENCES

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5, R80.

Pan,Q. et al. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics, 40, 1413–5

Trapnell,C. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28, 511–5.

Weischenfeldt,J. et al. (2012) Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome biology, 13, R35.

PeerJ PrePrints | https://peerj.com/preprints/80v1/ | v1 received: 11 Oct 2013, published: 11 Oct 2013, doi: 10.7287/peerj.preprints.80v1