Bioinformatics, 37(18), 2021, 3008–3010
doi: 10.1093/bioinformatics/btab142
Advance Access Publication Date: 27 February 2021
Applications Note

Genome analysis

ASimulatoR: splice-aware RNA-Seq data simulation

Quirin Manz [1], Olga Tsoy [1], Amit Fenn[1], Jan Baumbach[1,2,3], Uwe Völker[4], Markus List [1,*,†] and Tim Kacprowski [1,5,6,*,†]

[1]Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany, [2]Department of Mathematics and Computer Science, University of Southern Denmark, 5230 Odense, Denmark, [3]Chair of Computational Systems Biology, University of Hamburg, 22607 Hamburg, Germany, [4]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, 17475 Greifswald, Germany, [5]Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, 38106 Brunswick, Germany and [6]Braunschweig Integrated Centre of Systems Biology (BRICS), 38106 Braunschweig, Germany

*To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the two first authors should be regarded as Joint First Authors.

Associate Editor: Jan Gorodkin

Received on September 29, 2020; revised on February 16, 2021; editorial decision on February 23, 2021; accepted on February 25, 2021

Abstract

Summary: A plethora of tools exist for RNA-Seq data analysis with a focus on alternative splicing (AS). However, appropriate data for their comparative evaluation is missing. The R package ASimulatoR simulates gold standard RNA-Seq datasets with fine-grained control over the distribution of AS events, which allow for evaluating alternative splicing tools, e.g. to study the effect of sequencing depth on the performance of AS event detection.

Availability and implementation: ASimulatoR is freely available at https://github.com/biomedbigdata/ASimulatoR as an R package under GPL-3 license.

Contact: markus.list@wzw.tum.de or t.kacprowski@tu-braunschweig.de

1 Introduction

RNA-Seq is commonly used for alternative splicing (AS) identification and numerous tools are available for this kind of analysis (Alamancos *et al.*, 2014). To objectively assess their sensitivity and specificity together with scalability, gold standard datasets are required. Since the true number and types of AS events (such as, exon skipping, intron retention etc) in an experimental dataset is not known, suitable datasets need to be simulated. The only published tool for this purpose is BEERS, a perl script for evaluation of splice-aware alignment tools (Grant *et al.*, 2011). To introduce an AS event, BEERS randomly removes exons from a gene. This approach introduces AS event types by chance, not offering any control over the number, distribution or types of AS events. This prevents the analysis of biases that AS analysis tools might have toward the detection of particular event types. Here we present the R package ASimulatoR, which overcomes these limitations by allowing users to select the number of transcript variants with a certain AS event type, and to set RNA-Seq experiment parameters such as read length, sequencing depth, error rate, the number of replicates and technical biases. Unlike BEERS, ASimulatoR can thus produce simulated data suitable for the evaluation of any AS analysis tool.

2 Results

ASimulatoR is implemented in R 3.6.3 and tested in R 4.0.2. As input files, ASimulatoR uses a genome annotation in GTF or GFF3 format and chromosome fasta files (e.g. available at ftp://ftp.ensembl.org/pub/). Below is the detailed description of the individual steps in ASimulatoR (Fig. 1).

2.1 Exon supersets creation

For each gene, ASimulatoR collects all annotated exons from all transcript variants in the genome annotation and merges them in 'exon supersets' serving as unspliced templates (Fig. 1). Compatible AS event types are then assigned to exon supersets large enough to support them. The exon supersets are stored in the RDA format and can be reused for other simulations.

2.2 Transcript variants set creation

The next step is to create a set of transcript variants with AS events based on exon supersets. The user can define the number of overall genes in the simulated set and select AS event types to be included together with their distribution. ASimulatoR chooses the compatible exon supersets and combines exons to produce a transcript variant with the requested AS event types. Produced transcript variants serve as an input for RNA-Seq reads simulation.

ASimulatoR supports eight types of events: exon skipping, multiple exon skipping, intron retention, alternative 3'-splice site, alternative 5'-splice site, mutually exclusive exons, alternative first exon and last exon.

To define the distribution of AS events, the relative frequency or the probability can be chosen, where, e.g. the relative frequency of exon skipping to 0.1 means that 10% of genes in the resulting set will have an exon skipping event. Alternatively, setting the probability of exon skipping to 0.1 gives a 10% chance to each appropriate exon superset to create a variant with an exon skipping event, adding some randomness to the resulting dataset. To raise the complexity of simulated sets, it is further possible to introduce any combination of event types to a transcript and to allow multiple events to occur within the same exon.

This step outputs the description of all transcript variants as GTF/GFF3 genome annotation files and as a TSV file. ASimulatoR also allows to test how AS detection tools recover novel events. For this purpose the information about a user-selected fraction of transcript variants can be skipped and the truncated annotation will be written as an additional genome annotation file.

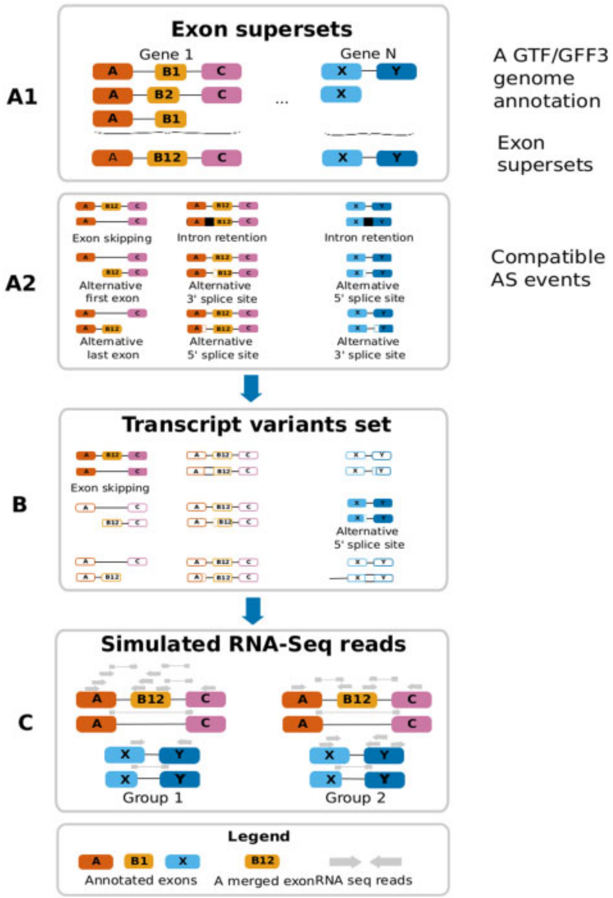

Fig. 1. ASimulatoR workflow. (**A1**) All annotated exons from a genome annotation are merged into 'exon supersets'. (**A2**) Compatible AS events are assigned to 'exon supersets'. (**B**) 'Exon supersets' are chosen based on user parameters (distribution and type of AS events) and form transcript variants. (**C**) Transcript variants serve as an input for RNA reads simulation. RNA-Seq datasets are simulated based on the custom version of *polyester* R package

2.3 RNA-Seq reads simulation

We utilized a customized version of the polyester R package (Alyssa *et al.*, 2015) from https://github.com/warrenmcg/polyester to simulate RNA-seq reads in parallel according to the transcript variants set from the previous step. General parameters from polyester (e.g. error rate or sequencing depth) are supported by ASimulatoR. We further modified *polyester* to introduce and control RNA-Seq technical biases: PCR duplicates and adapter contamination. For PCR duplication bias the user defines the fraction of reads that will be duplicated and marked as duplicates. By default the number of duplications are drawn from a Poisson distribution with $\lambda = 1$. If the user would like to have an RNA-Seq dataset with adapter contamination, the adapter sequence will be added to the fragments shorter than the defined read length.

The modified version of *polyester* can be found here: https://github.com/biomedbigdata/polyester.

For the evaluation of differential splicing tools, ASimulatoR creates groups of biological replicates and introduces fold-changes to random transcript variants.

As a result, ASimulatoR creates fastq files and a count matrix with read counts for all transcript variants.

3 Use case

As Liu *et al* have observed, the detection of AS events and in particular the sensitive detection of differential AS requires a considerably larger number of reads than the 50 million reads typically sequenced in an RNA-seq experiment (Liu *et al.*, 2013). Here, we demonstrate the usability of ASimulatoR by confirming this finding on simulated RNA-seq data.

To produce simulated data with characteristics close to a real dataset, we first analyzed 117 RNA-Seq datasets from SHIP (The Study of Health in Pomerania). This cohort comprises healthy individuals from Northeast Germany, equal number of males and females, from 40 to 69 years old (Völzke *et al.*, 2011). We mapped reads using STAR 2.7 (Dobin *et al.*, 2013). For AS event detection, we chose splAdder (Kahles *et al.*, 2016) as a relatively fast and easy to use tool which reports each event individually in a simple format. For each type of AS events we counted the proportion of genes with such an event (Table 1).

We used the mean proportions as probabilities for a gene to have this type of AS event. The other characteristics include: 200 million reads, read length 76 bp, error rate 0.1%.

We subsequently downsampled this set to 150, 100, 50 million reads to study the sensitivity of AS event detection and investigated two scenarios (Fig. 2): (i) the scenario of a well annotated genome — all transcripts were included into the genome annotation; (ii) the scenario of a poorly annotated genome — the information about one transcript per gene was retained.

We normalized transcript read counts from ASimulatoR using transcripts per kilobase million (TPM). We counted the proportion of recovered true events for all transcripts (TPM $\geq$ 0) and for transcripts with high expression (TPM $\geq$ 1) of both variants—with an event and without (Fig. 2).

Comparing results between the poorly and the well annotated genome, we observed a drop in the recovery rate, confirming that AS analysis depends on the quality of genome annotation. We further observed that even with 200 million reads not all AS events could be recovered by splAdder. This is likely due to an uneven read

Table 1. The proportion of genes with an event type in the SHIP datasets

Type	Exon skipping	Intron retention	Alternative 3'-splice site
Mean % ± standard deviation	9,8 ± 0,7	4,9 ± 0,5	7,8 ± 0,5
Type	Alternative 5'-splice site	Multiple exons skipping	Mutually exclusive exons
Mean % ± standard deviation	6,0 ± 0,5	0,4 ± 0,04	2,2 ± 0,2

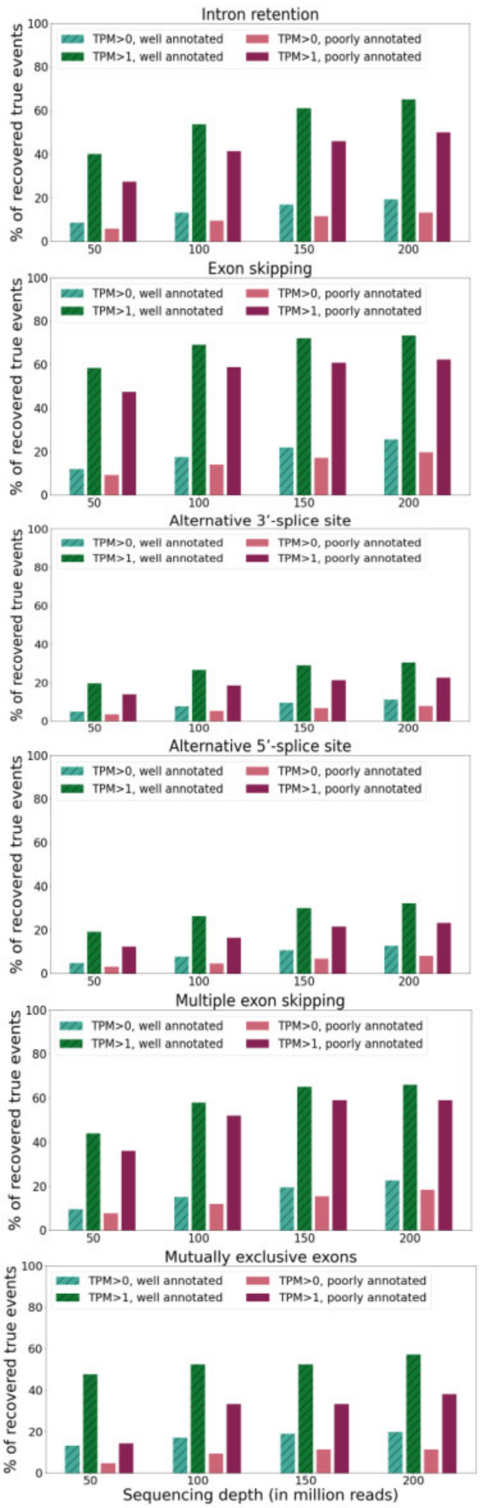

Fig. 2. The proportion of true events recovered by splAdder for six types of alternative splicing events. We considered varying sequencing depth and well and poorly annotated genome annotations

distribution affecting lowly expressed transcripts, where we observe a dramatic loss in the number of detected AS events. This loss is more pronounced for some types of AS events such as alternative splice sites (Fig. 2).

For detecting exon skipping events, splAdder requires both the junction and the exon itself to be covered by RNA-Seq reads. In contrast, tools such as ASGAL (Denti *et al.*, 2018) require only reads from the junction or the exon but not both and may thus be more sensitive in detecting rare splicing events. To test this hypothesis, we compared ASGAL and splAdder. We extracted a set of genes with AS events that were not recovered by splAdder and where at least one transcript variant in the genome annotation is supported by RNA-Seq reads and used that as input for ASGAL. ASGAL supports fewer types of events than splAdder: exon skipping and multiple exons skipping events belong to the same category and we counted them together. With 200 million reads ASGAL recovered around one-third of the events that were not found by splAdder (24.1% of exon skipping events, 34% of intron retention events, 30% of alternative 3'- and 28.2% of alternative 5'-splice sites), a trend that was shown previously for a PCR-validated dataset where this difference is even more pronounced (Denti *et al.*, 2018).

Using the simulated dataset generated by ASimulatoR, we could thus demonstrate the effect of genome annotation quality, sequencing depth and the choice of the tool on the AS analysis.

4 Conclusion

ASimulator allows for fine-grained control of AS event distributions in RNA-Seq data simulation. The resulting datasets are easily tunable and scalable, and will serve both as a standard for assessing existing and developing new AS analysis tools and pipelines.

Data Availability

Access to the SHIP data for research purposes may be requested at https://www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php.

Funding

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept [01ZX1908A (Sys_CARE) and 01ZX1910D (SyMBoD)]. SHIP is part of the Research Network Community Medicine of the University Medicine Greifswald, which is supported by the German Federal State of Mecklenburg-West Pomerania. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 395357507 – SFB 1371.

Conflict of Interest: none declared.

References

Alamancos,G. *et al.* (2014) Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol. Biol.*, **1126**, 357–397.

Denti,L. *et al.* (2018) ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events. *Bioinformatics*, **19**, 444.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Alyssa,C. *et al.* (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.

Grant,G.R. *et al.* (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.

Kahles,A. *et al.* (2016) SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics*, **32**, 1840–1847.

Liu,Y. *et al.* (2013) Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose. *PLoS One*, **8**, e66883.

Völzke,H. *et al.* (2011) Cohort Profile: the study of health in pomerania. *Int. J. Epidemiol.*, **40**, 294–307.