Gene Expression

ScanExitronLR: characterization and quantification of exitron splicing events in long-read RNA-seq data

Joshua Fry[1,2], Yangyang Li[1] and Rendong Yang[1*]

[1]Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA., [2]Bioinformatics and Computational Biology Program, University of Minnesota, Minneapolis, MN 55455, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Exitron splicing is a type of alternative splicing where coding sequences are spliced out. Recently, exitron splicing has been shown to increase proteome plasticity and play a role in cancer. Long-read RNA-seq is well suited for quantification and discovery of alternative splicing events; however, there are currently no tools available for detection and annotation of exitrons in long-read RNA-seq data. Here we present ScanExitronLR, an application for the characterization and quantification of exitron splicing events in long-reads. From a BAM alignment file, reference genome and reference gene annotation, ScanExitronLR outputs exitron events at the individual transcript level. Outputs of ScanExitronLR can be used in downstream analyses of differential exitron splicing. In addition, ScanExitronLR optionally reports exitron annotations such as truncation or frameshift type, nonsense-mediated decay status, and Pfam domain interruptions. We demonstrate that ScanExitronLR performs better on noisy long-reads than currently published exitron detection algorithms designed for short-read data.

Availability: ScanExitronLR is freely available at https://github.com/ylab-hi/ScanExitronLR and distributed as a pip package on the Python Package Index.

Contact: rendong.yang@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

An exitron is a region within an annotated coding exon that is spliced out like an intron. Exitrons are unique in that they possess both protein-coding and intronic potential while also possessing canonical splice-site signals (e.g. GT-AG). Originally described in Arabidopsis thaliana, exitrons have been shown to increase plant proteome diversity and plasticity (Marquez et al, 2015), mediate responses to heat stress (Cecchini et al, 2022) and create novel gene isoforms (Cheng et al, 2020; Aliperti et al, 2019). In humans, exitrons can alter cancer driver genes, promote tumor progression and be a potential source of neoantigens (Wang et al, 2021). Because exitrons were discovered relatively recently, nomenclature has not yet settled and they have also been called 'cryptic introns' (e.g. Dean et al, 2020) or 'intron retention loss' (e.g. Ringeling et al, 2022).

Long-read sequencing, though more error prone than short-read sequencing, is in a better position to identify novel splicing isoforms (Amarasinghe et al, 2020), such as those containing exitrons. However, there are currently no tools available for exitron detection in long-read sequencing. Moreover, novel splice site detection within noisy long-reads presents its own challenges. The higher error rate often leads to imprecision in the exon-intron boundary. To combat this, long-read aligners such as Minimap2 (Li, 2018) can utilize a BED file of annotated exons to preferentially align junctions within reads to annotated exon borders. However, while this increases the mapping accuracy of annotated transcripts, it occasionally causes exitrons to be misaligned to alternate 3' or 5' splice-sites (Supplementary Figure S1, S2).

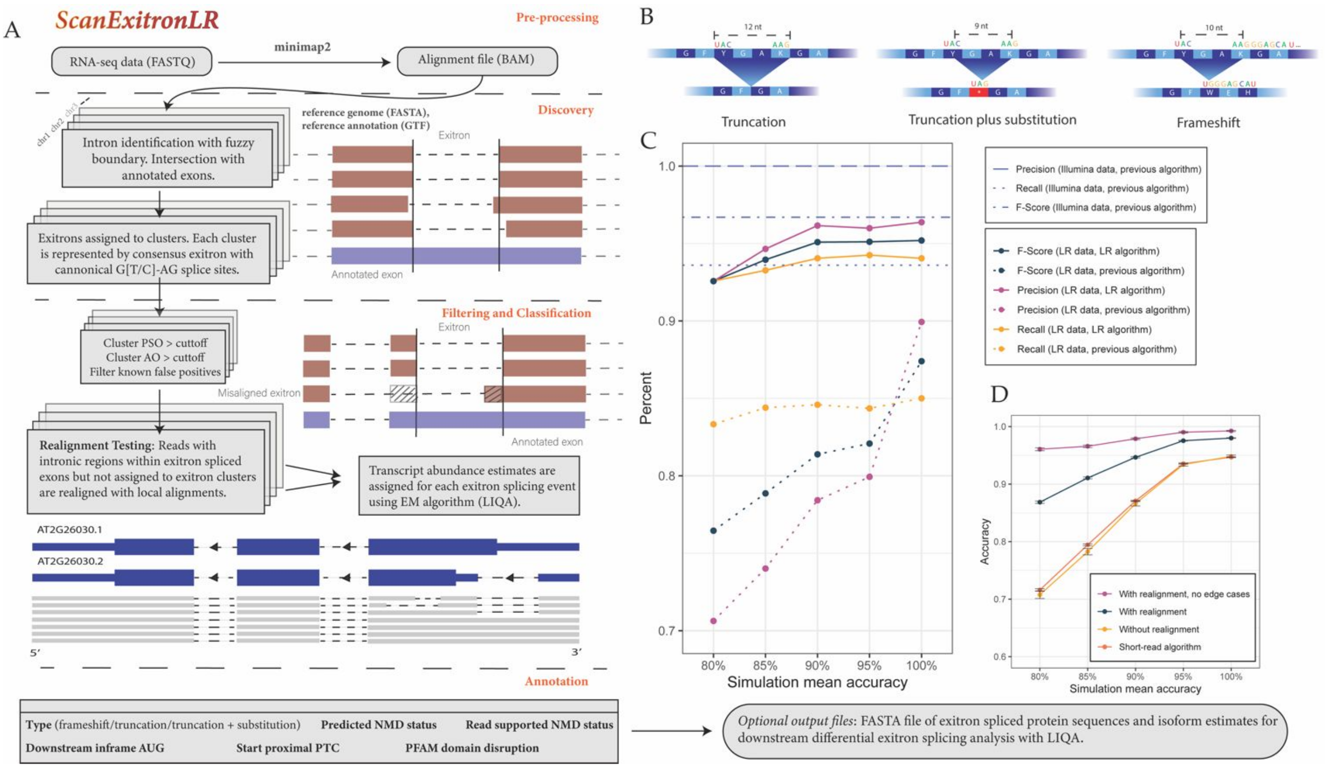
A
ScanExitronLR
Pre-processing
RNA-seq data (FASTQ)
minimap2
Alignment file (BAM)
Discovery
chr1 chr2 chr3
reference genome (FASTA),
reference annotation (GTF)
Intron identification with fuzzy boundary. Intersection with annotated exons.
Exitron
Exitrons assigned to clusters. Each cluster is represented by consensus exitron with cannonical G[T/C]-AG splice sites.
Annotated exon
Filtering and Classification
Cluster PSO > cuttoff
Cluster AO > cuttoff
Filter known false positives
Exitron
Misaligned exitron
Annotated exon
Realignment Testing: Reads with intronic regions within exitron spliced exons but not assigned to exitron clusters are realigned with local alignments.
Transcript abundance estimates are assigned for each exitron splicing event using EM algorithm (LIQA).
AT2G26030.1
AT2G26030.2
5′
3′
Annotation
Type (frameshift/truncation/truncation + substitution)
Predicted NMD status
Read supported NMD status
Downstream inframe AUG
Start proximal PTC
PFAM domain disruption
Optional output files: FASTA file of exitron spliced protein sequences and isoform estimates for downstream differential exitron splicing analysis with LIQA.
B
12 nt
UAC AAG
G F Y G A K G A
G F G A
Truncation
9 nt
UAC AAG
G F Y G A K G A
UAG
G F * G A
Truncation plus substitution
10 nt
UAC AAGGGAGCAU...
G F Y G A K G A
UGGGAGCAU
G F W E H
Frameshift
C
1.0
0.9
0.8
0.7
Percent
80%
85%
90%
95%
100%
Simulation mean accuracy
Precision (Illumina data, previous algorithm)
Recall (Illumina data, previous algorithm)
F-Score (Illumina data, previous algorithm)
F-Score (LR data, LR algorithm)
F-Score (LR data, previous algorithm)
Precision (LR data, LR algorithm)
Precision (LR data, previous algorithm)
Recall (LR data, LR algorithm)
Recall (LR data, previous algorithm)
D
1.0
0.9
0.8
0.7
0.6
Accuracy
80%
85%
90%
95%
100%
Simulation mean accuracy
With realignment, no edge cases
With realignment
Without realignment
Short-read algorithm

Figure 1. (A) Description of ScanExitronLR and companion tool AnnotateExitron. (B) Schematic representation of the difference between truncation, truncation plus substitution and frameshift inducing exitron splicing events. (C) Simulation results from ONT long-reads at 80%, 85%, 90%, 95% and 100% mean read accuracy levels. Dotted lines indicate simulation results from ScanExitron on 75bp read length Illumina short-reads. (D) Mean accuracy for each detected exitron, defined as (algorithm reported AO)/(ground-truth AO). Error bars represent

To solve these issues, we present ScanExitronLR, an application for the discovery and annotation of exitron splicing events in long-reads. In addition to overcoming the higher sequencing error of long-reads, ScanExitronLR utilizes read length to match exitron splicing events with specific annotated transcripts using an expectation maximization (EM) algorithm provided by LIQA (Hu et al, 2021).

2 Algorithm Description

The input to ScanExitronLR is a BAM alignment file, along with a FASTA reference genome and a GTF reference gene annotation file (Fig 1a). For alignment, we suggest using Minimap2 (Li, 2018) ScanExitronLR first identifies introns within the BAM alignments and intersects them with annotated exons to find introns completely contained within an annotated exon. Because unannotated splice-sites may be noisy, we add a small amount of nucleotide jitter (10 as default) and treat splice-sites in the alignment file as fuzzy. Thus, a splice-site will be considered novel only if it occurs +/- jitter away from annotated splice-sites. We then assign exitrons into clusters such that every exitron within the cluster is no more than 2 * jitter away from each other. For each cluster, we nominate a consensus exitron that (1) has canonical G[T/C]-AG splice sites and (2) has the highest number of supporting reads. We then assign all exitrons within the cluster to the consensus exitron, thus treating the cluster as one splicing event.

After clustering, we filter exitrons based on AO (number of reads supporting the exitron), PSO (percent spliced out, a measure of the percentage of transcripts for which the exitron is spliced), and, optionally, cluster purity (measured as the proportion of reads with the consensus exitron splice-sites). A PSO cutoff allows the user to identify exitrons at a desired splicing frequency. A low cluster purity indicates low confidence in the reported exitron splice-sites (Supplementary Figure S3).

Nearby insertions and deletions may cause exitron spliced reads to be misaligned with alternative 3' and 5' splice-sites and thus not detected in the previous steps (Fig 1a; see Supplementary Figure S2a,b for examples). In order to correctly identify these reads as exitron spliced, ScanExitronLR undergoes realignment testing. For each exon in which an exitron was detected in the previous step, ScanExitronLR examines reads with intronic regions within the exon but not assigned to exitron clusters. Using two rounds of local alignments, a read is realigned as supporting an exitron splicing event if, first, the exitron sequence does not appear in the region of the read aligned to the annotated exon and, second, a top scoring local alignment contains a gap with roughly the correct length and flanking splice sequences (Supplementary Methods). Thus the realignment step increases the AO estimate accuracy but does not identify new exitron splicing events.

In order to identify which gene transcripts contain called exitrons, we separate exitron spliced reads from unspliced reads and run LIQA (Hu et al, 2021) for exitron specific transcript quantification. Based on this quantification, ScanExitronLR reports annotated transcript abundance estimates for each exitron splicing event. Additionally, ScanExitronLR optionally determines the type of each exitron: frameshift, truncation, or truncation plus substitution which have the potential to substitute a novel amino acid or even stop codon into the truncation site (Fig 1b). Predicted and read supported nonsense-mediated decay (NMD) features (Supplementary Methods) and Pfam protein domain disruption are also reported. The user can optionally save a FASTA file of exitron spliced protein sequences and isoform estimates for further downstream splicing analysis.

3 Experimental Results

In order to assess the performance of ScanExitronLR, we performed simulation experiments with simulated long-reads at mean read accuracies of 80%, 85%, 90%, and 100% using PBSIM2 (Ono et al, 2021; Supplementary Methods). For each read accuracy level, we chose 10,000 protein coding transcripts from the GENCODE v37 annotation at random with replacement. For each transcript we chose a random CDS exon region and found random G[T/C]-AG splice sites at least 30 nt away from each other (with a 9:1 GT to GC splice-site ratio). We aligned the resulting reads with minimap2 (Li, 2018). Because there are currently no published tools to detect exitrons in long-reads, we compared the performance of ScanExitronLR with a previously published tool, ScanExitron, which has been tested to detect exitrons in short-read RNA-seq (Wang et al, 2021a; Wang et al, 2021b). In addition, we similarly simulated 10,000 exitron splicing events in 75bp read length short-reads with Illumina error profiles using Rsubread and mapped the resulting reads using the STAR (Dobin et al, 2013) aligner.

Our simulation results show that the performance of ScanExitronLR is stable across all accuracy levels (Fig 1c). At 80% read accuracy the precision and recall is 92.57% and 92.58% respectively, while at 100% accuracy it is 94.05% and 96.38%. In contrast, the short-read algorithm precision is significantly lower for noisy long-reads, 70.62% at 80% accuracy and 89.93% at 100% accuracy. The 100% to 80% percent difference of the F-score for the short-read algorithm is 10.95% compared to only 2.62% for ScanExitronLR. This shows that our algorithm is indeed able to correct for the errors unique to long-read sequencing. ScanExitron achieves an F-score of 96.7% on Illumina short-reads compared to an F-score average of 94.3% for ScanExitronLR on long-reads. This shows that exitron detection in long-reads is just as reliable as exitron detection in short-reads.

We also computed the accuracy of each true exitron detection event as (algorithm reported AO)/(ground-truth AO) (Fig 1d). Without our realignment step, ScanExitronLR is just as accurate as the short-read algorithm. However, with realignment, the accuracy is significantly increased, especially at low read accuracy levels. We observed that many of the exitron spliced reads not detected by ScanExitronLR were due to faulty alignments when the exitron splicing occurred close to the exon border (Supplementary Figure S2b). Excluding edge cases where an exitron occurs within 50 nt of an exon border, ScanExitronLR is more than 95% accurate across all read accuracies.

4 Example

We ran ScanExitronLR on a recently published direct RNA sequencing dataset of Arabidopsis samples (Zhang et al, 2020). With an AO cutoff of 2 and PSO cutoff of 0.05, we found 172 exitrons across four biological replicates, two buds and two flowers (Supplementary Methods). As an example, we identified an exitron with a length of 90 nt in gene AT2G26030, an F-box containing protein. This exitron was not detected in the original Arabidopsis exitron dataset through short-read RNA-seq analysis (Marquez, et al 2015), though longer exitrons were found in this same gene. Because ScanExitronLR can identify exitrons at the transcript level, using the output of ScanExitronLR in downstream analysis we were able to discover that, in bud samples, exitrons within this gene are differentially expressed in a shortened transcript, AT2G26030.2, with alternative start codon (p = 1.88x10-4, chi-squared test; Fig 1a). Interestingly, this shortened transcript splices out the F-box binding domain. Thus, this exitron is most likely associated with alternative functions of this gene--an insight one could not obtain without transcript level quantification of exitron splicing events.

Funding

This work was supported by the National Institute of General Medical Sciences [R35GM142441].

Conflict of Interest: none declared.

References

Aliperti, Vincenza et al. (2019) Identification, Characterization, and Regulatory Mechanisms of a Novel EGR1 Splicing Isoform. International journal of molecular sciences, 20, 1548.

Amarasinghe, Shanika L et al. (2020) Opportunities and challenges in long-read sequencing data analysis. Genome biology, 21, 30.

Cecchini, Nicolás Miguel et al. (2022) Alternative splicing of an exitron determines the subnuclear localization of the Arabidopsis DNA-glycosylase MBD4L under heat stress. The Plant journal: for cell and molecular biology, 110, 377-388.

Chen, Ying, et al. (2021) A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. bioRxiv.

Cheng, Qiang et al. (2020) Conserved exitrons of FLAGELLIN-SENSING 2 (FLS2) across dicot plants and their functions. Plant science: an international journal of experimental plant biology, 296, 110507.

Dean, Dexter N, and Jennifer C Lee. (2020) Modulating functional amyloid formation via alternative splicing of the premelanosomal protein PMEL17. The Journal of biological chemistry, 295, 7544-7553.

Dobin A, Davis CA, Schlesinger F, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics.29, 15-21.

Hu, Yu et al. (2021) LIQA: long-read isoform quantification and analysis. Genome biology, 22, 182.

Li, Heng. (2018) Minimap2: pairwise alignment for nucleotide sequences." Bioinformatics, 34, 3094-3100.

Marquez, Yamile et al. (2015) Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. Genome research 25, 995-1007.

Ono, Yukiteru et al. (2021) PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. Bioinformatics 37, 589-595.

Ringeling, Francisca Rojas et al. (2022) Partitioning RNAs by length improves transcriptome reconstruction from short-read RNA-seq data. Nature biotechnology, 40, 741-750.

Wang, Ting-You et al. (2021) A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. Molecular cell, 81, 2246-2260.

Wang, Ting-You, and Rendong Yang. (2021) Integrated protocol for exitron and exitron-derived neoantigen identification using human RNA-seq data with ScanExitron and ScanNeo. STAR protocols, 2, 100788.

Zhang, Shoudong et al. (2020) New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. Nucleic acids research, 48, 7700-7711.