S-IRFindeR: an R package for detecting and measuring Intron Retention using RNA-seq data

Lucile Broséus

Introduction

Abstract

Accurate quantification of intron retention (IR) levels is currently the crux for detecting and interpreting the function of retained introns. S-IRFindeR implements our new approach to measuring intron retention levels using second generation RNA-seq data, the Stable Intron Retention ratio (SIRratio). In addition to this, the package also enables to detect IR events and compute observed IR rates using third generation RNA-seq data.

Summary

Package Installation

The R package *SIRFindeR* can be installed from GitHub by copy-pasting the following code line:

```
devtools::install_github("lbroseus/SIRFindeR")
```

Then, load *SIRFindeR*:

```
suppressPackageStartupMessages( library(SIRFindeR) )
```

Estimating IR-levels using second generation RNA-seq data

Prerequisite: short read alignment using STAR

In order to compute SIRratio values, we will need two files from *STAR* alignments: - the bam file with read alignments; - the *SJ.out.tab* file.

Here is a typical command line to perform the required genomic alignment step using *STAR*:

```
STAR  --genomeDir $STARindex  \                    # Path to the STAR index
      --readFilesIn  Reads_1.fq,Reads_2.fq\        # Read files
      --outFileNamePrefix $mySample  \             # A prefix for output files
      --runThreadN $nthreads  \                    # Number of threads
      --outStd BAM_Unsorted --outSAMtype BAM   Unsorted \
      --outSAMstrandField intronMotif --outSAMunmapped None --outFilterMultimapNmax 1
```

Note: SIRFindeR was tested using reference data from *ENSEMBL*: https://www.ensembl.org/index.html.

Computing SIRratios

Once the alignment step is completed, SIRratios can be obtained from a wrapper function as follows:

```
#Input data
bamFile <- "Unsorted.bam"
junctionFile <- "SJ.out.tab"

#Read length:
readLength <- 100
#Indicate whether your read are single-end ("SE") or paired-end ("PE")
libraryType <- "SE"

computeSIRratio(gtf,
                bamFile = bamFile,
                readLength = readLength,
                libraryType = libraryType,
                junctionFile = junctionFile,
                saveDir = saveDir)
```

This will create several sample-specific files in the directory *saveDir*.
Among whose:
- *SIRratio.txt*: which contains final results with SIRratio values for each sample-curated independent intron;
- *ResultsByIntron.txt*: with the SIRratio values per independent intron. Independent introns are reference genomic intervals common to all samples from the same organism. This is the file you will need if you want to aggregrate and compare several samples.

Estimating IR-levels using third generation RNA-seq data

Several functions for estimating (observed) IR rates are available in SIRFindeR.
Our proposed method is reference-based (evaluated intron are extracted from a reference transcriptome annotation, ie: a gtf file), and makes use of (spliced) genome alignments.
We intend to make it more flexible in the near future.

Prerequisite: long read alignment

Thus, first, you will need to align long reads onto a reference genome. There exist several long read splice-aware aligners. We mainly tested two of them (*Minimap2* and *GMAP*). For our applications, *Minimap2* achieved better results on raw data; while being definitely faster. They both showed comparable performances on corrected data.
If, for some reason, you want to use *GMAP*, we strongly advise to perform long read correction before aligning the reads; this may improve significantly alignment rates and accuracy (cf: reference *4*).
In case you also have matched short read data, you can perform hybrid correction using *TALC*: https://github.com/lbroseus/TALC.

Here are typical command lines for aligning *Oxford Nanopore* long read using *Minimap2* (https://github.com/lh3/minimap2, reference *3*):

```
minimap2 -ax splice -uf -k14 \
                    $yourReferenceGenome.fa \ # A fasta file with the reference genome (eg: from ENSEMI
                    $yourSample.fa \          # A fasta or fastq file containing long read sequences
                    -t $nthreads \            # Number of threads
                    >$yourSample.mmap2.sam    # A name for the output file
```

Computing observed IR rates

```
# Input data and paths

gtf <- "MyBelovedTranscriptAnnot.gtf"
bamFile <- "yourSample.bam"
saveDir <- "Where/To/Output/Results"

# Calculate IR rates

computeIRrates(gtf = gtf, bamFile = bamFile, saveDir = saveDir, keepSecondaryAlignments = FALSE)
```

This will:
1. extract annotated intron intervals from the reference transcriptome;
2. detect intron retention events (output in *IRevents.txt*);
3. compute observed IR rates (output in *IRrates.txt*).
Both *.txt* files will be written in the directory *saveDir*.

References

1. *Broseus and Ritchie* ***SIRFindeR: stable and accurate quantification of intron retention levels*** *biorXiv. (2020)*
2. *Dobin et al.* ***STAR: ultrafast universal RNA-seq aligner.*** *Bioinformatics. (2013)*
3. *Li.* ***Minimap2: pairwise alignment for nucleotide sequences.*** *Bioinformatics. (2013)*
4. *Broseus et al.* ***TALC: Transcript-level Aware Long Read Correction.*** *Bioinformatics. (2020)*