SpliceLauncher

SpliceLauncher is a pipeline tool to study the alternative splicing. It works in three steps: * Get a read count matrix from fastq files, by a dedicated RNAseq pipeline (A step in diagram below). * Generate data files used hereafter (B step in diagram below) * Run SpliceLauncher from a read count matrix (C step and furthermore in diagram below).

SpliceLauncher

Table of contents

Repository contents

- dataTest: example of input files
- scripts: complementary scripts to run SpliceLauncher

Prerequisites to install SpliceLauncher

The SpliceLauncher pipeline needs to install the following tools and R librairies:

- STAR (v2.6 or later)
- samtools (v1.3 or later)
- BEDtools (v2.17 or later)

- R with *WriteXLS* and *Cairo* packages
- Perl

STAR

Following instruction were from the STAR manual

Get the g++ compiler for linux

```
sudo apt-get update
sudo apt-get install g++
sudo apt-get install make
```

Download the latest release and uncompress it

```
# Get latest STAR source
wget https://github.com/alexdobin/STAR/archive/2.7.0c.tar.gz
tar -xzf 2.7.0c.tar.gz
cd STAR-2.7.0c

# Alternatively, get STAR source using git
git clone https://github.com/alexdobin/STAR.git
```

Compile under Linux

```
# Compile
cd STAR/source
make STAR
```

Samtools

Download the samtools package at: https://github.com/samtools/samtools/releases/latest

Configure samtools for linux:

```
cd samtools-1.x
./configure --prefix=/where/to/install
make
make install
```

For more information, please see the samtools manual

BEDtools

Installation of BEDtools for linux:

```
wget https://github.com/arq5x/bedtools2/releases/download/v2.25.0/bedtools-2.25.0.tar.g
tar -zxvf bedtools-2.25.0.tar.gz
cd bedtools2
make
```

For more information, please see theBEDtools tutorial

Install R libraries

Open the R console:

```
install.packages("WriteXLS")
install.packages("Cairo")
```

Installing SpliceLauncher

Download the latest release from of SpliceLauncher source using git

```
git clone https://github.com/raphaelleman/SpliceLauncher
cd ./SpliceLauncher
```

Download the reference files

The reference files are the genome (Fasta) and the corresponding annotation file (GFF3):

1. Reference genome in fasta format
2. The annotation file in GFF v3 format

Steps: 1. Download Fasta genome: from RefSeq FTP server or from Gencode.

For example, human hg19 genome file from RefSeq:

```
#the ftp URL depends on your assembly genome choice
wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh37_latest/refseq_identif
gunzip ./GRCh37_latest_genomic.fna.gz
```

2. Download the GFF annotation file, either from RefSeq FTP server or from Gencode.

For example, human hg19 annotation file from RefSeq:

```
wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh37_latest/refseq_identi
gunzip ./GRCh37_latest_genomic.gff.gz
head ./GRCh37_latest_genomic.gff
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build GRCh37.p13
#!genome-build-accession NCBI_Assembly:GCF_000001405.25
#!annotation-date
#!annotation-source
##sequence-region NC_000001.10 1 249250621
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606
NC_000001.10    RefSeq  region  1   249250621   .   +   .   ID=id0;Dbxref=taxon:9606;Nam
NC_000001.10    BestRefSeq  gene    11874   14409   .   +   .   ID=gene0;Dbxref=GeneID:1
NC_000001.10    BestRefSeq  transcript  11874   14409   .   +   .   ID=rna0;Parent=gene0
NC_000001.10    BestRefSeq  exon    11874   12227   .   +   .   ID=id1;Parent=rna0;Dbxre
```

3. [Optionnal] To reduce needed memory, we can also restrict the analysis to the primary assembly, without unplaced contigs:

```
grep ">" GRCh37_latest_genomic.fna | grep -v "unplaced genomic contig"| grep -v "unloca
seqtk subseq GRCh37_latest_genomic.fna chr_names > GRCh37_latest_genomic.sub.fna
cut -f 1 -d ' ' chr_names > chr_names_id
head -n 9 GRCh37_latest_genomic.gff > GRCh37_latest_genomic.sub.gff
grep -f chr_names_id GRCh37_latest_genomic.gff >> GRCh37_latest_genomic.sub.gff
```

Configure SpliceLauncher with INSTALL mode

SpliceLauncher comes with a ready to use config.cfg file. It contains the paths of software and files used by SpliceLauncher. The INSTALL mode of SpliceLauncher updates this config.cfg file. If you define the path to GFF (v3) file and path to the FASTA genome, the INSTALL mode will extract all necessary information from this GFF and indexing the STAR genome. This informations are stored in a BED file that contains the exon coordinates, in a sjdb file that contains the intron coordinates and a text file that contains the details of transcript structures. You need to define where these files will saving by the `-O, --output` argument

Use INSTALL mode of SpliceLauncher:

```Bash
cd /path/to/SpliceLauncher/
mkdir ./refSpliceLauncher # Here this folder will contain the reference files used by Sp
bash ./SpliceLauncher.sh --runMode INSTALL \
    -O ./refSpliceLauncher \
    --STAR /path/to/STAR \
    --samtools /path/to/samtools \
    --bedtools /path/to/bedtools \
    --gff /path/to/gff \
    --fasta /path/to/fasta
```

Docker image

SpliceLauncher and all its dependencies are also integrated in a Docker image:

1. To build it:

```
docker build . -t evolbioinfo/splicelauncher
```

2. To download an already existing Docker image from Docker hub:

```
docker pull evolbioinfo/splicelauncher
```

or with singularity

```
singularity pull docker://evolbioinfo/splicelauncher
```

3. To open a shell in the container:

```
docker run -ti -w $PWD -v $PWD:$PWD --entrypoint bash evolbioinfo/splicelauncher
```

or

```
singularity shell splicelauncher-latest.simg
```

Running the SpliceLauncher tests

The example files are provided in dataTest, with the example data provided in single end RNAseq (1x75pb) on *BRCA1* and *BRCA2* transcripts:

`Bash cd /path/to/SpliceLauncher bash ./SpliceLauncher.sh --runMode Align,Count,SpliceLauncher -F ./dataTest/fastq/ -O ./testSpliceLauncher/`

After running, the BAM files from alignment are in a *Bam* folder, the count files are in *getClosestExons* and the

results of SpliceLauncher analysis are in *testSpliceLauncher_result*.

The final results are displayed in the file*testSpliceLauncher_outputR.xlsx*, this last is in *testSpliceLauncher_result* folder. The scheme of this file is:

Column names	Example	Description
Conca	chr13_32915333_32920963	The junction id (chr_start_end)
chr	chr13	Chromosome number
start	32915333	Genomic coordinate of start junction End if on reverse strand
end	32920963	Genomic coordinate of end junction Start if on reverse strand
strand	+	Strand of the junction ('+': forward; '-':reverse)
Strand_transcript	forward	Strand of transcript
NM	NM_000059	The transcript id according RefSeq nomenclature
Gene	BRCA2	Gene symbol
Sample	2250	Read count
P_Sample	15.25659623	% of relative expression
event_type	SkipEx	The nature of junction: Physio: Natural junction SkipEx: Exon skipping 5AS: Donor splice site shift 3AS: Acceptor splice site shift NoData: Unannotated juntion
AnnotJuncs	Δ12	The junction names
cStart	c.6841	Transcriptomic start coordinate of the junction
cEnd	c.6938	Transcriptomic end coordinate of the junction
mean_percent	12.60242	Average in % of relative expression across samples
read_mean	2683.769231	Average of read count across samples
nbSamp	11	Number of time that the junction has been seen in samples
DistribAjust	-	The Distribution of junction expression (Gamma/N.binom)
		If a sample shown an abnormal expression of the

Column names	Example	Description
Significative	NO	junction

SpliceLauncher options

--runMode INSTALL,Align,Count,SpliceLauncher * The runMode defines the steps of analysis with: * INSTALL: Updates the config.cfg file for SpliceLauncher pipeline * Align: Generates BAM files from the FASTQ files * Count: Generates the matrix read count from the BAM files * SpliceLauncher: Generates final output from the matrix read count

Option for INSTALL mode

-C, --config /path/to/configuration file/ * Path to the config.cfg file,**only** if you want to use your own config file

-O, --output /path/to/output/ * Directory to save the reference files (BED, sjdb, txt) and the indexed genome

--STAR /path/to/STAR * Path to the STAR executable

--samtools /path/to/samtools * Path to the samtools executable

--bedtools /path/to/bedtools * Path to the bedtools executable

--gff /path/to/gff file * Path to the GFF file (v3)

--fasta /path/to/fasta * Path to the genome fasta file

-t, --threads N * Nb threads used to index the STAR genome

Option for Align mode

-F, --fastq /path/to/fastq/ * Repository of the FASTQ files

-O, --output /path/to/output/ * Repository of the output files

-p * Processes to paired-end analysis

-t, --threads N * Nb threads used for the alignment

-g, --genome /path/to/genome * Path to the genome directory,**only** if you to use a genome directory different of the genome defined in config.cfg file

--STAR /path/to/STAR * Path to the STAR executable,**only** if you to use a STAR software different of the STAR defined in config.cfg file

--samtools /path/to/samtools * Path to the samtools executable,**only** if you to use a samtools software different of the samtools defined in config.cfg file

Option for Count mode

-B, --bam /path/to/BAM files * Repository of the BAM folder

-O, --output /path/to/output/ * Repository of the output files

--samtools /path/to/samtools * Path to the samtools executable,**only** if you to use a samtools software different of the samtools defined in config.cfg file

--bedtools\t/path/to/bedtools * Path to the bedtools executable,**only** if you to use a bedtools software different of the bedtools defined in config.cfg file

-b, --BEDannot /path/to/your_annotation_file.bed * Path to exon coordinates file (in BED format),**only** if you to use exon coordinates different of the coordinates defined in config.cfg file

Option for SpliceLauncher mode

-I, --input/path/to/inputFile * Read count matrix (.txt)

-O, --output /path/to/output/ * Directory to save the results

--TranscriptList /path/to/transcriptList.txt * Set the list of transcripts to use as reference

--txtOut * Print main output in text instead of xls

--bedOut * Get the output in BED format

--Graphics * Display graphics of alternative junctions (Warnings: increase the runtime)

-n, --NbIntervals 10 * Nb interval of Neg Binom (Integer)

--SampleNames name1|name2|name3 * Sample names, '|'-separated, by default use the sample file names

If list of transcripts (--TranscriptList): **--removeOther** * Remove the genes with unselected transcripts to improve runtime

If graphics (-g, --Graphics): **--threshold** 1 * Threshold to shown junctions (%)

-R, --RefSeqAnnot /path/to/RefSpliceLauncher.txt * Transcript information file,**only** if you to use a transcript information file different of file defined in config.cfg file

Authors

- Raphael Leman - raphaelleman
    - You can contact me at: r.leman@baclesse.unicancer.fr or raphael.leman@orange.fr

> **Cite as:** SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data. Raphaël Leman, Valentin Harter, Alexandre Atkinson, Grégoire Davy, Antoine Rousselin, Etienne Muller, Laurent Castéra, Fréderic Lemoine, Pierre de la Grange, Marine Guillaud-Bataille, Dominique Vaur, Sophie Krieger

License