1. OVERVIEW

Trans-splicing detection tool (TSD) was developed to reliably identify intra-chromosomal intergenic *trans*-splicing events (iTSEs) from RNA-seq data. TSD only takes NGS sequencing reads as input and could able to exactly filter false positives due to PCR template switching during library preparation and read-through events.

2. ENVIRONMENTS

1) TSD requests python 2.7.11 or later version, Tophat v2.0.12 or later version but before V2.1.0), bowtie1.1.2, blat V.36x2 or later version, bedtools v2.26.0 or later version and be sure that the python module of subprocess, os, re, Biopython, sys, getopt have been installed.

2) The Tophat, bowtie and blat packages are enclosed in the TSD package.

3) The bedtools can be downloaded from https://github.com/arq5x/bedtools2/releases

3. USAGE

3.1 Use Tophat to find chimeric transcripts from RNA-seq data

3.1.1 Required data

TSD takes RNA-seq raw reads (".fastq" format) which can be either single-end or pair-end as input.

3.1.2 Build Index

bowtie-build hg19.fa hg19

3.1.3 For example, when processing a illumina sequencing data in fastq format with read length of 100bp, we can set the parameters as follows:

sh PATH_to_software/TSD.tophat.sh -i $index -x test_1.fastq -y test_2.fastq

3.1.4 There are two parameters that have a relatively important impact.

(1) max_mismatch: The maximum mismatch for the supporting junction reads. For RNA-seq data at high sequencing depth (over 20-fold), 0 is suggested. Default:1.

(2) RSL: The span length of junction read over the junction site (RSL). For RNA-seq data with read-length (less than 75bp), 15 is suggested. Default: 25.

3.2 Use TSD to find trans-splicing events

3.2.1 Simple running code

python ${PATHtoTSD}/TSD.py -i config_test.txt -o test-filtered

3.2.2 Configuration options

Options	Request
genefile	Annotations of genes. The first seven columns are same as first seven columns of GENCODE GTF. The eighth column is the "gene_name" of genes. The ninth column is name of gene family for genes and "none" means there is no gene family for the corresponding gene. The GENCODE GTF can be downloaded from GENCODE (https://www.gencodegenes.org/). The gene family information can be downloaded from HGNC (https://www.genenames.org/).
gapfile	Annotations of gaps of assembly sequence in bed format. Which can be downloaded from UCSC (https://genome.ucsc.edu/).
fastafile	Corresponding genome in fasta format. Which can be downloaded from UCSC (wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz) .
fasta2bitfile	Corresponding genome in .2bit format. Which can be downloaded from UCSC (wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit).
junctions_inputfile	Tophat output file names: junctions.bed for detecting iTSEs.

junctions_repswitch	Determined whether using the replicates option. When it is “Y”, the corresponding files of replicates need to be filled in correctly; while when it is “N”, it can be filled in freely, such as “na” or you can fill in nothing.
junctions_repfile1	Tophat output junctions.bed file of replicates.
junctions_repfile2	Tophat output junctions.bed file of replicates.
fusions_inputfile	Tophat output file names: fusions.out for detecting iTSEs.
fusions_repswitch	Determined whether using the replicates option. When it is “Y”, the corresponding files of replicates need to be filled in correctly; while when it is “N”, it can be filled in freely, such as “na” or you can fill in nothing.
fusions_repfile1	Tophat output fusions.out file of replicates.
fusions_repfile2	Tophat output fusions.out file of replicates.
gro_switch	Controls the method used in filtering read-through events. When it is “Y”, TSD will filter read-through using GRO-seq data. The corresponding folder containing processed GRO-seq data (grofiledir option) needs to be filled in correctly. The downloaded GRO-seq data in bedgraph format could be processed by:
	python2.7 pipeline/prepare_GROseq_data.py -i GRO-seq.bedGraph -b 200.
	You can only process the GRO-seq data with the desired chromosome, while the data on different strands of same chromosome needed to be available.
	When it is “N”, the TSD will filter read-through independent of GRO-seq data (a rough filtering method), the grofiledir option can be filled in freely, such as “na” or you can fill in nothing.
grofiledir	Folder containing processed GRO-seq data.

delSV_switch	Determines whether to remove possible false positive iTSEs caused by chromosomal variations. When it set to “Y”, you need provide an annotation file in vcf format in SVVCF option. when it is “N”, the SVVCF option can be filled in freely, such as “na” or you can fill in nothing.
SVVCF	Annotations of chromosomal variations.
min_score_filter	The minimum score for alternative matches over 80bp. default: 97; suggested: 95, 96, 97.
min_sup	The minimum support junction reads for fusions; default:4; suggested: 1-6.
Softwaredir	The path to TSD.

Detailed introduction of Data format please refer to the UCSC. (https://genome.ucsc.edu/FAQ/FAQformat.html)

3.3 Output files

Files named as final_iTSEs_removedSVs.txt containing final results after structure variation (SV) removal. The output is txt file containing detected iTSEs with twelve columns.

1) The first three columns are coordinates of iTSEs.
2) The fourth column is the splicing types when it was "– "or "+", while it represents the orientation of chromosomes when it was" ff", "fr", "rf" or "rr". "f" means in forwarding direction and "r" means in reversing direction. “ff” means the orientations of the two chromosomes - both chromosomes are in forwarding direction.
3) The fifth column is the number of reads spanning the junction sites.
4) The sixth and seventh columns are the strand information of two junction sites.
5) The eighth and ninth columns are the relative coordinate between the donor or accepter transcript with their junction site. “minus” means the coordinate of transcript is less than its corresponding junction site. “plus”

means the coordinate of transcript is more than its corresponding junction site.

6) The tenth and eleventh columns are the gene annotation of junction sites.
7) The twelfth column is the possibility of iTSEs involved in the SVs.

4. NOTEWORTHY

The warning message below can be ignored:

Warning: malformed line 1, missing columns

track name=junctions description="Tophat junctions"

5. GENERATE SIMULATED DATA

5.1 All of the fasta file used to generate simulated RNA-seq data in TSD are loaded in the TSD package. Users can use the art_illumina to simulated RNA-seq data with different sequencing depthand read-length.

Also, the simulated RNA-seq datasets are available from the authors upon request.

5.2 Using art_illumina:

cd PATHtoTSD

cd tools-package

tar -xvzf artbinmountrainier2016.06.05linux64.tgz

cp art_bin_MountRainier/art_illumina ../tools

5.3 Examples:

1. simulated RNA-seq data with read-length 50 and at sequencing depth20

${Softwaredir}/tools/art_illumina -ss HS20 -i ${Softwaredir}/simulated-fasta/positive_simu1.fa -o ${prefix}_ -l 50 -f 20 -p -m 200 -s 20

2. simulated RNA-seq data with read-length 75 and at sequencing depth40

${Softwaredir}/tools/art_illumina -ss HS20 -i ${Softwaredir}/simulated-fasta/positive_simu1.fa -o ${prefix}_ -l 75 -f 40 -p -m 250 -s 25

3. simulated RNA-seq data with read-length 100 and at sequencing depth60

${Softwaredir}/tools/art_illumina -ss HS20 -i ${Softwaredir}/simulated-fasta/positive_simu1.fa -o ${prefix}_ -l 100 -f 60 -p -m 300 -s 30

4. simulated RNA-seq data with read-length 150 and at sequencing depth80

${Softwaredir}/tools/art_illumina -ss HS20 -i ${Softwaredir}/simulated-fasta/positive_simu1.fa -o ${prefix}_ -l 150 -f 80 -p -m 400 -s 40

Detailed introduction of simulation data please refer to ART : https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm

6. Please cite:

Kim, D. and Salzberg, S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biology, 12.

Huang, W.C., Li, L.P., Myers, J.R. and Marth, G.T. (2012) ART: a next-generation sequencing read simulator. Bioinformatics, 28, 593-594.