TEToolkit Manual

A package for including transposable elements in differential enrichment analysis of sequencing datasets.

Created by Ying Jin, Eric Paniagua, Oliver Tam & Molly Hammell, February 2014

Copyright (C) 2014-2018 Ying Jin, Eric Paniagua, Oliver Tam & Molly Hammell

Contact: Ying Jin ([yjin@cshl.edu](mailto:yjin@cshl.edu))

# Brief Description

TEToolkit contains two standalone tools, TEtranscripts and TEcount. Both programs take RNA-seq (and similar data) and annotates reads to both genes & transposable elements, handling both uniquely mappable and ambiguously mapped reads. TEtranscripts then performs differential analysis using DESeq2.

**NOTE:** TEtranscripts and TEcount rely on specially curated GTF files, which are not packaged with this software due to their size. Please go to [our website](http://hammelllab.labsites.cshl.edu/software#TEToolkit) for instructions to download the curated GTF files.

# Requirements

**Python**: 2.6.x or 2.7.x (not tested in Python 3.x)

**pysam**: 0.9.x or greater

**R**: 2.15.x or greater

**DESeq2**: 1.10.x or greater

# Installation

1. Download compressed tarball from Github or Pypi
2. Unpack tarball.
3. Navigate into unpacked directory.
4. Run the following:

$ python setup.py install

If you want to install locally (e.g. /local/home/usr), run this command instead:

$ python setup.py install --prefix /local/home/usr

**NOTE** In the above example, you must add /local/home/usr/bin to the PATH variable, and /local/home/usr/lib/python2.X/site-packages to the PYTHONPATH variable, where python2.X refers to the python version (e.g. python2.7 if using python version 2.7.x).

# TEtranscripts

## Usage

TEtranscripts -t treatment sample [treatment sample ...]

-c control sample [control sample ...]

--GTF genic-GTF-file

--TE TE-GTF-file

[optional arguments]

### Required arguments:

-t | --treatment [treatment sample 1 treatment sample 2...]

Sample files in group 1 (e.g. treatment/mutant), separated by space

-c | --control [control sample 1 control sample 2 ...]

Sample files in group 2 (e.g. control/wildtype), separated by space

--GTF genic-GTF-file GTF file for gene annotations

--TE TE-GTF-file GTF file for transposable element annotations

### Optional arguments:

#### Input/Output options

--format [input file format]

Input file format: BAM or SAM. DEFAULT: BAM

--stranded [option] Is this a stranded library? (yes, no, or reverse).

DEFAULT: yes.

--sortByPos Input file is sorted by chromosome position.

--project [name] Prefix used for output files (e.g. project name)

DEFAULT: TEtranscript\_out

#### Analysis options

--mode [TE counting mode]

How to count TE:

uniq (unique mappers only)

multi (distribute among all alignments).

DEFAULT: multi

--minread [min\_read] read count cutoff. DEFAULT: 1

-L | --fragmentLength [fragLength]

Average length of fragment used for single-end sequencing

DEFAULT: For paired-end, estimated from the input alignment file.

For single-end, ignored by default.

-i | --iteration

max number of iterations used to optimize multi-reads assignment.

DEFAULT: 100

-p | --padj [pvalue]

FDR cutoff for significance. DEFAULT: 0.05

-f | --foldchange [foldchange]

Fold-change ratio (absolute) cutoff for differential expression.

DEFAULT: 1

#### DESeq1 compatibility options

--DESeq

Use DESeq (instead of DESeq2) for differential analysis.

-n | --norm [normalization]

Normalization method :

DESeq\_default (default normalization method of DESeq)

TC (total annotated read counts)

quant (quantile normalization)

Only applicable if DESeq is used instead of DESeq2.

DEFAULT: DESeq\_default

#### Other options

-h | --help

Show help message

--verbose [number]

Set verbose level.

0: only show critical messages

1: show additional warning messages

2: show process information

3: show debug messages

DEFAULT: 2

--version Show program's version and exit

**NOTE:** BAM files must be either unsorted or sorted by queryname. If the BAM files are sorted by position, please use the '--sortByPos' option

## Example Command Lines

If BAM files are unsorted, or sorted by queryname:

$ TEtranscripts --format BAM --mode multi -t RNAseq1.bam RNAseq2.bam -c CtlRNAseq1.bam CtlRNAseq.bam --project sample\_nosort\_test

If BAM files are sorted by coordinates/position:

$ TEtranscripts --sortByPos --format BAM --mode multi -t RNAseq1.bam RNAseq2.bam -c CtlRNAseq1.bam CtlRNAseq.bam --project sample\_sorted\_test

# TEcount

## Usage

TEcount -b RNAseq BAM

--GTF genic-GTF-file

--TE TE-GTF-file

[optional arguments]

### Required arguments:

-b | --BAM alignment-file RNAseq alignment file (BAM preferred)

--GTF genic-GTF-file GTF file for gene annotations

--TE TE-GTF-file GTF file for transposable element annotations

### Optional arguments:

#### Input/Output options

--format [input file format]

Input file format: BAM or SAM. DEFAULT: BAM

--stranded [option]

Is this a stranded library? (yes, no, or reverse).

DEFAULT: yes.

--sortByPos

Input file is sorted by chromosome position.

--project [name]

Prefix used for output files (e.g. project name)

DEFAULT: TEcount\_out

#### Analysis options

--mode [TE counting mode]

How to count TE:

uniq (unique mappers only)

multi (distribute among all alignments).

DEFAULT: multi

-L | --fragmentLength [fragLength]

Average length of fragment used for single-end sequencing

DEFAULT: For paired-end, estimated from the input alignment file.

For single-end, ignored by default.

-i | --iteration

max number of iterations used to optimize multi-reads assignment.

DEFAULT: 100

#### Other options

-h | --help

Show help message

--verbose [number]

Set verbose level.

0: only show critical messages

1: show additional warning messages

2: show process information

3: show debug messages

DEFAULT: 2

--version Show program's version and exit

**NOTE** BAM files must be either unsorted or sorted by queryname. If the BAM files are sorted by position, please use the '--sortByPos' option

## Example Command Lines

If BAM files are unsorted, or sorted by queryname:

$ TEcount --format BAM --mode multi -b RNAseq.bam --project sample\_nosort\_test

If BAM files are sorted by coordinates/position:

$ TEtranscripts --sortByPos --format BAM --mode multi -b RNAseq.bam --project sample\_sorted\_test