TAPIS Documentation

Release 1.2.1

Mike Hamilton

Jul 17, 2017

CONTENTS

Contents:

CHAPTER

ONE

GET TAPIS

1.1 Required packages

The following packages are required for TAPIS core functions. Version numbers correspond to those tested during development.

- SpliceGrapher (v0.2.4)
- Pysam (v0.8.1)
- matplotlib (v1.3.1)
- bx-python (v0.5.0)
- NumPy (v1.8.2)
- GMAP (2016-04-01) note: version 2015-07-23 is incompatible

1.2 Download

current release: v1.2.1

TAPIS is hosted on bitbucket https://bitbucket.org/comp_bio/tapis

1.3 Install

```
$ tar zxvf tapis_<version>.tgz
$ cd tapis_<version>.tgz
$ python setup.py install
```

Note: To install in a user directory, use the option:

```
--home=/Path/To/Local/Library
```

CHAPTER

TWO

TUTORIAL

This tutorial is meant as a complete walk-through for identifying transcripts and poly(A) sites from PacBio reads.

1. Align and clean reads
2. Cluster reads and analyze transcripts and poly(A) sites

2.1 Align reads

TAPIS accepts any sorted, indexed BAM file for long reads but it provides a method that cleans and aligns reads with high accuracy and efficiency. To align and clean reads use the following provided script **`alignPacBio.py`**. Before running the script, you will need to run **`gmap_build`** to make a genome reference index. Since version 1.2.1, the alignment script creates two files, an aligned bamfile and unaligned reads in a FASTA file (previous versions required merging of individual iterations of alignment/cleaning runs).

```
usage: alignPacBio.py [-h] [-v] [-i ITERATIONS] [-e EDR] [-o OUTDIR]
                      [-p PROCS] [-K MAXINTRON]
                      indexesDir indexName reference fasta

Iteratively fix aligned reads using reference genome

positional arguments:
  indexesDir            directory to gmap indexes
  indexName             name of gmap index
  reference             Reference sequence
  fasta                 Reads to align

optional arguments:
  -h, --help            show this help message and exit
  -v, --verbose         Verbose mode
  -i ITERATIONS, --iterations ITERATIONS
                        Number of aligment iterations, default=3
  -e EDR, --edr EDR     Edit distance ratio, default=10
  -o OUTDIR, --outdir OUTDIR
                        Output directory, default=./filtered
  -p PROCS, --procs PROCS
                        Number of processors, default=1
  -K MAXINTRON, --maxIntron MAXINTRON
                        maximum intron length for gmap, default=8000
```

2.2 Creating indexed, sorted BAM files

If your cleaned/aligned reads are in the form of a SAM file, you can convert it to a indexed, sorted BAM file using **convertSam.py**.

```
usage: convertSam.py [-h] [-o BAMFILE] [-p PROCS] [-m MEMORY] [-v] samfile

Generate sorted BAM and index files for given SAM file

positional arguments:
  samfile               Samfile to convert

optional arguments:
  -h, --help            show this help message and exit
  -o BAMFILE, --outfile BAMFILE
                        Name of converted BAM file [default=<sambase>.bam]
  -p PROCS, --procs PROCS
                        Number of processors to use for BAM sorting (default
                        1)
  -m MEMORY, --memory MEMORY
                        Max memory (in GBs) for each processor used for BAM
                        sorting (default 2)
  -v, --verbose         Print verbose output
```

2.3 Running TAPIS

```
$ run_tapis.py --help
usage: run_tapis.py [-h] [-v] [-p] [-o OUTDIR] [-t TRIMMAX] [-w W]
                    [-m MINDIST] [-s MINSUPPORT]
                    geneModel bamfile

Assemble transcripts from PacBio alignments

positional arguments:
  geneModel             Gene models annotation file (GFF/GTF)
  bamfile               Aligned reads file (sorted and indexed)

optional arguments:
  -h, --help            show this help message and exit
  -v, --verbose         Verbose mode
  -p, --plot            Plot novel gene graphs and poly(A) figures, default is
                        no plotting
  -o OUTDIR, --outdir OUTDIR
                        Output directory for TAPIS results, default=tapis_out
  -t TRIMMAX, --trimMax TRIMMAX
                        Maximum length of read trimming to tolerate on 3' end
                        of reads, default=5
  -w W, --w W           Width of peaks when searching for poly(A) sites,
                        default=5
  -m MINDIST, --minDist MINDIST
                        Minimum distance between any two poly(A) sites,
                        default=20
  -s MINSUPPORT, --minSupport MINSUPPORT
                        Minimum number of trusted reads supporting a poly-A
                        site, default=2
```

While **TAPIS** offers many options, default values should work for most cases.

2.4 Interpreting TAPIS output

TAPIS builds an output directory as follows:

```
$ tree my_result
  tapis_out
  |-- polyAFigures
  |   |-- gene1.png
  |   |-- gene2.pbg
  |   |-- ...
  |   |-- geneN.png
  |-- novelGraphs
  |   |-- chrom_start_end_strand.pdf
  |   |-- ...
  |-- assembled.gtf
  |-- novelGenes.csv
  |-- novelGenes.fa
  |-- polyA_summary.csv
```

- **polyAFigures** - contains poly(A) site depictions for genes with at least one poly(A) site supported by long reads.
- **novelGraphs** - contains splice graph figures for transcripts not found in within any annotated gene.
- **assembled.gtf** - gene models for transcripts detected in long reads
- **novelGenes.csv** - tab-delimited file containing summary of novel genes detected

CHAPTER

THREE

CONTACT

TAPIS is developed by Mike Hamilton at Colorado State University.

Bug reports and feature requests can be submitted through bitbucket.

- search

BIBLIOGRAPHY

[SG] Rogers, MF, Thomas, J, Reddy, AS, Ben-Hur, A (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.*, 13, 1:R4.