TECDetec Manual

TECDetec is written in perl, thus can be run in any system that have installed Perl 5.12 or higher. TECDetec contains a main script tecdetec.pl and a package PerlLib which needs to be kept in the same directory as tecdetec.pl.

TECDetec implements Bowtie 2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) (Hatem, Bozdağ, Toland, & Çatalyürek, 2013), Tophat 2 (http://ccb.jhu.edu/software/tophat/index.shtml) (Kim et al., 2013) and Samtools (http://samtools.sourceforge.net) (Li et al., 2009). Please install these 3 softwares before starting TECDetec.

TECDetec accepts paired-end RNA-Seq dataset in FASTQ format and reports TE-related transcript active genomic regions in GTF format. It requires an index of TE sequence assembly, an index of masked reference genome and an index of unmasked reference genome. Custom repetitive sequence assembly and reference genome can be indexed using `bowtie2-build` in Bowtie2.

```
$ bowtie2-build <sequence_in_fasta_format> <basename>
```

Usage of TECDetec:

```
$ perl tecdetec.pl [options] -idx_te <TE index> -idx_mskge <masked reference genome index> -idx_ge <reference genome index> -r1 <reads1 file> -r2 <reads2 file>
```

Option	Description
-p/-threads	Use parallel threads (default is 1)
-idx_te	The basename of the index for the transposon sequence assembly
idx_mskge	The basename of the index for the masked genome

-idx_ge	The basename of the index for the transposon sequence assembly
-lib_type	Library type, unstranded if not specified (default is unstranded). fr - first read is sequenced in the sense orientation (e.g. Ligation) rf - first read is sequenced in the antisense orientation (e.g. dUTP)
-r1	The first read file in a paired end dataset
-r2	The second read file in a paired end dataset
-min_intron	minimum intron length (default is 100 bps)
-min_cov	minimum depth of coverage (default is 1)
-I/-min	minimum insert size (default is 0 bp)
-X/-max	maximum insert size (default is 500 bps)
-o/-out_dir	output dir (default is ./tecdetec_out)
-h	help
-v	version

Reference

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*, 184. doi: 10.1186/1471-2105-14-184

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. doi:10.1186/gb-2013-14-4-r36

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352