Fondecyt Grant #11140869

Center for Bioinformatics and
Molecular Simulation

TEcandidates

TEcandidates is a pipeline to include transposable elements in RNA-seq differential expression analysis.

INSTALLATION INSTRUCTIONS

---

1. Dependencies

TEcandidates is implemented in Bash, and requires no installation. However, other softwares that are part of the pipeline are required. The following are the required softwares, along with some minimum instructions to install them in a computer with Linux. For additional help and/or troubleshooting, the source page of each software is listed for more detailed instructions.

- **BEDtools v2.25**

-Source page: https://github.com/arq5x/bedtools2/releases/tag/v2.25.0

```
$ wget https://github.com/arq5x/bedtools2/releases/download/v2.25.0/bedtools-
2.25.0.tar.gz
$ tar -xvzf bedtools-2.25.0.tar.gz
$ cd bedtools2/
$ make
$ sudo make install
```

Check if installation was correct using

```
$ bedtools --version
bedtools v2.25.0
```

- **BioPerl**

-Source page: http://bioperl.org/INSTALL.html

```
$ sudo perl -e shell -MCPAN
Password:
Terminal does not support AddHistory.

cpan shell -- CPAN exploration and modules installation (v1.9800)
Enter 'h' for help.

cpan[1]> install C/CJ/CJFIELDS/BioPerl-1.007001.tar.gz
```

Verify installation with:

```
$ perl -MBio::Root::Version -e 'print $Bio::Root::Version::VERSION,"\n"'
```

- **Bowtie v2.3**

-Source page: https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.2

Download the appropiate file for your distribution and follow these instructions:

```
$ unzip bowtie2-2.3.2-linux-x86_64.zip
$ cd bowtie2-2.3.2/
$ pwd
```

The above will show the full path to the bowtie2 binaries. Copy it and add it to the $PATH environment variable:

```
$ export PATH=$PATH:path_to_bowtie2-2.3.2
```

Check correct installation with

```
$ bowtie2 --version
bowtie2-align-s version 2.3.2
64-bit
Built on dde45b53bd81
Sat May  6 02:39:52 UTC 2017
Compiler: gcc version 4.1.2 20080704 (Red Hat 4.1.2-55)
Options: -O3 -m64 -msse2 -funroll-loops -g3 -DPOPCNT_CAPABILITY -DWITH_TBB -
DNO_SPINLOCK -DWITH_QUEUELOCK=1
Sizeof {int, long, long long, void*, size_t, off_t}: {4, 8, 8, 8, 8, 8}
```

- **Samtools v1.4.1**

-Source page: https://github.com/samtools/samtools

```
$ wget https://github.com/samtools/samtools/releases/download/1.4.1/samtools-
1.4.1.tar.bz2
$ tar -xvjf samtools-1.4.1.tar.bz2
$ cd samtools-1.4.1
$ ./configure
$ make
$ sudo make install
```

Check correct installation:

```
$ samtools --version
samtools 1.4.1
Using htslib 1.4.1
Copyright (C) 2017 Genome Research Ltd.
```

- **Trinity v2.4**

-Source page: https://github.com/trinityrnaseq/trinityrnaseq/

```
$ wget https://github.com/trinityrnaseq/trinityrnaseq/archive/master.zip
$ unzip master.zip
$ cd trinityrnaseq-master/
$ make
$ pwd
```

Copy the path to the Trinity directory and add it to the $PATH environment variable:

```
$ export PATH=$PATH:path_to_Trinity
```

Check correct installation with

```
$ Trinity --version
Trinity version: Trinity-v2.4.0
```

2. Installing the TEcandidates pipeline

Download the TEcandidates tarball, and uncompress it:

```
$ wget
https://github.com/TEcandidates/TEcandidates/blob/master/TEcandidates_LATEST_STABLE.tar.

$ tar -xvzf TEcandidates_LATEST_STABLE.tar.gz
```

Grant execution permissions to the pipeline script:

```
$ chmod u+x TEcandidates_LATEST_STABLE/TEcandidates.sh
```

For simplicity of use, add the TEcandidates full path to your PATH environment variable. First get the full path:

```
$ readlink -f TEcandidates_LATEST_STABLE
```

Then copy the output of the previous command, and add it to the PATH variable:

```
$ export PATH=$PATH:/path/to/TEcandidates_LATEST_STABLE
```

SAMPLE USAGE

Once TEcandidates is in your PATH variable, you can execute it as

```
$ TEcandidates.sh -t=Number_of_threads -r=RAM_to_use -g=Genome_Fasta_File -
fq=Path_to_FASTQ_files -m=Mode -te=TE_Annotation -c=Coverage -l=TE_length -
N=Number_of_candidateTEs

-t Number of threads to use in the softwares executed during the pipeline
-r Maximum amount of RAM assigned to Trinity (Trinity's --max_memory option)
-g Genome to use (FASTA format, .fasta extension)
-fq Path to FASTQ files to use (all files must have .fastq extension)
-m Mode of FASTQ files, SE for Single-end reads and PE for Paired-end reads
-te Transposable Element annotation file
-c Minimum coverage in which a Transposable Element must be covered by a de-novo
transcript in order to be selected as candidate
```

```
-l Minimum length of Transposable Element to be considered in the selection step
-N Number of candidate Transposable Elements to output
```

Important considerations

- Reads files must have ".fastq" extension.
- TEcandidates can be used with either single-end reads or paired-end reads. Paired-end reads **must have** "_1.fastq" and "_2.fastq" extensions.

SAMPLE DATA

In order to check that TEcandidates is working correctly, please test it with a dataset from *Drosophila melanogaster* (Ohtani et al., 2013). The dataset is available at Gene Expression Omnibus, accession no. GSE47006, and must be downloaded with The *fastq-dump* tool from the SRA toolkit . To install the SRA toolkit, please copy the link of the appropriate version for your Operating System from https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/, and download it:

```
$ wget https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz
```

Check if fastq-dump is working:

```
$ tar -xvzf sratoolkit.current-ubuntu64.tar.gz
$ ./sratoolkit.2.8.2-1-ubuntu64/bin/fastq-dump -V

  ./sratoolkit.2.8.2-1-ubuntu64/bin/fastq-dump : 2.8.2
```

For ease of use, add the SRA toolkit to your $PATH environment variable. Get the full path to the SRA toolkit binaries with

```
$ readlink -f sratoolkit.2.8.2-1-ubuntu64/bin/
```

and append it to $PATH like this:

```
$ export PATH=$PATH:path_to_SRAtoolkit
```

Create a new work directory

```
$ mkdir TEcandidates_test
$ cd TEcandidates_test
```

Download the *Drosophila melanogaster* control dataset with:

```
$ nohup fastq-dump --defline-seq '@$sn[_$rn]/$ri' --split-files --accession SRR851837
> SRR851837.fastq-dump.log &
```

Download the *Drosophila melanogaster* treatment dataset with:

```
$ nohup fastq-dump --defline-seq '@$sn[_$rn]/$ri' --split-files --accession SRR851838
> SRR851838.fastq-dump.log &
```

Once these processes are done, check that they were downloaded correctly with

```
$ tail *log
==> SRR851837.fastq-dump.log <==
Read 42134407 spots for SRR851837
Written 42134407 spots for SRR851837

==> SRR851838.fastq-dump.log <==
Read 48277060 spots for SRR851838
Written 48277060 spots for SRR851838
```

Download the genome and the annotation file:

```
$ wget http://mobilomics.cl/tecandidates/files/dm3.fasta
$ wget http://mobilomics.cl/tecandidates/files/dm3_rmsk_TE.gff3
```

Execute the pipeline script afterwards:

```
nohup TEcandidates.sh -t=10 -r=128 -c=0.3 -l=900 -te=dm3_rmsk_TE.gff3 -g=dm3.fasta -
fq=. -m=SE -N=1 > TEcandidates.log &
```

Once it's done, you should have a folder named **candidateTE_analysis_coverage-0.3_length-900_N-1**, that contains the following files:

```
$ ls -lht candidateTE_analysis_coverage-0.3_length-900_N-1
total 2.1G
drwxr-xr-x 2 bvaldebenito bvaldebenito 4.0K Apr 28 08:22 trinity_assemblies
-rw-r--r-- 1 bvaldebenito bvaldebenito  44M Apr 28 08:23
dm3.fasta.masked_BT2.rev.1.bt2
-rw-r--r-- 1 bvaldebenito bvaldebenito  30M Apr 28 08:23
dm3.fasta.masked_BT2.rev.2.bt2
-rw-r--r-- 1 bvaldebenito bvaldebenito  44M Apr 28 08:22 dm3.fasta.masked_BT2.1.bt2
-rw-r--r-- 1 bvaldebenito bvaldebenito  30M Apr 28 08:22 dm3.fasta.masked_BT2.2.bt2
-rw-r--r-- 1 bvaldebenito bvaldebenito 505K Apr 28 08:22 dm3.fasta.masked_BT2.3.bt2
-rw-r--r-- 1 bvaldebenito bvaldebenito  30M Apr 28 08:22 dm3.fasta.masked_BT2.4.bt2
-rw-r--r-- 1 bvaldebenito bvaldebenito 165M Apr 28 08:22 dm3.fasta.masked
-rw-r--r-- 1 bvaldebenito bvaldebenito 5.8M Apr 28 08:22 repeatsToMask_coverage-
0.3_length-900.gff3
-rw-r--r-- 1 bvaldebenito bvaldebenito 7.4K Apr 28 08:22 allcandidates_coverage-
0.3_length-900_N-1.gff3
-rw-r--r-- 1 bvaldebenito bvaldebenito 1.1G Apr 28 08:01 SRR851838_1_filtered.fastq
-rw-r--r-- 1 bvaldebenito bvaldebenito  223 Apr 28 07:57 SRR851838_1.bt2_summary
-rw-r--r-- 1 bvaldebenito bvaldebenito 718M Apr 28 07:39 SRR851837_1_filtered.fastq
-rw-r--r-- 1 bvaldebenito bvaldebenito  222 Apr 28 07:36 SRR851837_1.bt2_summary
```

The candidate Transposable Elements can be found in the **allcandidates_coverage-0.3_length-900_N-1.gff3**, those that were removed from the genome in the **repeatsToMask_coverage-0.3_length-900.gff3** file. The genome, with the repeats in **repeatsToMask_coverage-0.3_length-900.gff3** removed, is in the **dm3.fasta.masked**. Additional files for Bowtie 2 are also generated (***_BT2*.bt2**).

CONTACT

Please send any inquiries about usage and/or bugs to TEcandidates@gmail.com

REFERENCES

Ohtani H, Iwasaki YW, Shibuya A, Siomi H, Siomi MC, Saito K. (2013). DmGTSF1 is necessary for Piwi-piRISC-mediated transcriptional transposon silencing in the Drosophila ovary. Genes Dev. 2013 Aug 1;27(15):1656-61. doi: 10.1101/gad.221515.113.