T. brucei co-expression network analysis

Kennedy Mwangi

March 26, 2019

Introduction

This document contains the workflow used in the analysis of T. brucei gene co-expression network analysis. It contains code used in each step of the analysis.

Data acquisition and quality assessment

The first step involves downloading the RNASeq data and checking its quality through the FASTQC tool.

Here, RNASeq data is downloaded from ENA database using FTP.

```
cat ../scripts/fastq_download.sh
```

```
## #!/bin/bash
## #
## #Script to download fastq files from European Nucleotide Archive
## #Created March 28,2019 by Kennedy Mwangi
## #
## FILE=$1 #File containing fastq url links to ENA FTP site
##
## OUT_DIR=../data/raw_data/
##
## cat ${FILE} | xargs -n1 wget$2 -P ${OUT_DIR}
```

Some of the downstream tools require that FASTQ files downloaded in zipped form are unzipped.

```
cat ../scripts/unzip.sh
```

```
## #/bin/bash
## #
## #Script to decompress fastq.gz files
## #Created March 28,2019 by Kennedy Mwangi
## #
## FASTQ_FILES=../data/raw_data/*.fastq.gz
##
## for file in ${FASTQ_FILES}; do
##     gunzip ${file}
## done
```

Next, the quality of the FASTQ files is checked through the FASTQC tool whose output is a report in HTML format.

```
cat ../scripts/fastqc_reports.sh
```

```
## #/bin/bash
## #
## #Script to run FastQC reports using the FastQC tool
## #Written on March 28, 2019 by Kennedy Mwangi
## #
## #load fastqc module
```

```
## module load fastqc/0.11.4
##
## FASTQ_DIR=../data/raw_data/*.fastq
##
## #create output directory if it doesn't exist.
## mkdir -p ../results/fastqc_reports
##
## REPORTS_DIR=../results/fastqc_reports/
##
## for file in ${FASTQ_DIR}; do
##    fastqc -f fastq -o ${REPORTS_DIR} ${file}
## done
```

Downloading *T. brucei* genome and GFF files

After determining that the data is of good quality and that no trimming is required, the reads are aligned on the *T. brucei* genome obtained from TriTryDB database.

The genome and the Gene Feature Format (GFF) files are downloaded from the TriTryDB database as follows:

```
#Downloading the genome

#wget https://tritrypdb.org/common/downloads/release-42/TbruceiTREU927/fasta/data/\
#TriTrypDB-42_TbruceiTREU927_Genome.fasta -P ../data/tbrucei_genome/


#Downloading the GFF file
#wget https://tritrypdb.org/common/downloads/release-42/TbruceiTREU927/gff/data/\
#TriTrypDB-42_TbruceiTREU927.gff -P ../data/genome_annotations_GFF/
```

Alignment of reads on the genome

Here, HISAT2 is used to align reads on the *T. brucei* genome. The first step is indexing the genome using HISAT2 followed by alignment of the reads. The output is SAM files.

Indexing the genome

```
cat ../scripts/hisat2_index.sh
```

```
## #!/bin/bash
## #
## #Script to index T. brucei genome using HISAT2
## #Created April 8, 2019 by Kennedy Mwangi
## #
## module load hisat/2-2.1.0
##
## #Create directory for HISAT2 indexed genome
## mkdir -p ../data/HISAT2_indexed_genome
##
## GENOME_FILE=$1
##
```

```
## cd ../data/HISAT2_indexed_genome/
##
## hisat2-build ${GENOME_FILE} tbrucei_genome_index_hisat2
```

Aligning the reads to the genome

```
cat ../scripts/hisat2_align.sh
```

```
## #!/bin/bash
## #
## #Script to align T. brucei reads to the indexed genome using HISAT2
## #created April 9, 2019
## #
## module load hisat/2-2.1.0
##
## #change directory to that of the indexed genome
## cd ../data/HISAT2_indexed_genome/
##
## for fastq in ../raw_data/*.fastq; do
##     fqname=$(echo $fastq | cut -f1 -d '.')
##
##      hisat2 \
##       -x tbrucei_genome_index_hisat2 \
##       -U ${fastq} \
##       -S ${fqname}.sam \
##       -p 8 \
##      --summary-file ${fqname}.txt \
##      --new-summary
## done
##
## #make directories and move created files into them
##
## mkdir -p ../processed_data
## mkdir -p ../../results/hisat2_alignment_summary
##
## mv ../raw_data/*.sam ../processed_data/
## mv ../raw_data/SRR*.txt ../../results/hisat2_alignment_summary/
```

Reads quantification

HTSeq tool is used to count reads that aligned to the *T. brucei* genome. The output is a text file for each sample that contains the number of reads that were counted for each gene.

```
cat ../scripts/htseq_counts.sh
```

```
## #!/bin/bash
## #
## #Script to counts the number of reads aligned to T. brucei genome using HTSeq.
## #Resource: HTSeq documentation https://htseq.readthedocs.io/en/latest/count.html
## #Created on April 2, 2019 by Kennedy Mwangi
## #
## module load htseq/0.11.2
##
```

```
## #create output directory if it doesn't exist
## mkdir -p ../results/HTSeq_count_results
##
## GFF_FILE=$1
##
## for sam_file in ../data/processed_data/*.sam; do
##     sam_file_name=$(echo $sam_file | cut -f1 -d '.')
##
##         python /opt/apps/htseq/0.11.2/bin/htseq-count \
##             -f sam \
##             -s no \
##             -t exon \
##             -i Parent \
##             $sam_file \
##             $GFF_FILE \
##             > ../results/HTSeq_count_results/${sam_file_name}.counts.txt
## done
```