Analysis of PBMC Transcriptome in COVID-19 Symptomatic vs. Uninfected Patients

Yuchen Sun

2023-04-20

Contents

Introduction

Results

Methods

Data download

The data used in this project is linked to this paper[1]. The data description metadata is retrieved from the SRA Run Selector, whereas the actual sequencing data with `.fastq.gz` format is retrieved from the ENA. The original data contains five groups: uninfected (22), recovering (15), symptomatic (13), re-detectable positive (RP) (12), and asymptomatic (8). However, in this project, only data from uninfected and symptomatic patients is used. A bash script is used to download data by first retrieving the the corresponding sample SRR ID, which is then used to create an `ftp` connection with the specific ENA web address hosting the paired-end read files of that sample, and using `wget` to download the files. Each read file is automatically put into the corresponding directory (uninfected or symptomatic) after the download is complete.

Read preprocessing

After downloading all read files, a FastQC run is used to determine the quality of the raw reads. It was discovered that most of the paired-end reads has high adapter content that failed the quality check. Hence, Trim-Galore was used to trim the raw reads by the command

```
trim_galore --illumina --paired \
            --output_dir $trim_out_dir \
            --stringency 13 $file $file2
```

Notice that the parameter `--illumina` and `--paired` was added to the command given that the original raw reads are produced by illumina paired-end sequencing, and that the parameter `--stringency 13` is set as it is a length threshold more likely to prevent potentially wrong adaptor trimming due to small overlaps (this threshold was chosen according to previous runs of Trim-Galore on the same data). `$trim_out_dir` represents the directory that holds the trimmed reads output of Trim-Galore, whereas `$file` and `file2` represents the two paired-end reads. After running Trim-Galore, another FastQC run is used to determine the quality of the trimmed reads, which are all proven to have low adatper content and are available for further processing. All scripts for this part of the preprocessing is put into a bash script to enable automatic FastQC run before and after Trim-Galore trimming.

Sequence alignment

In this project, STAR alignment tool is used to align the trimmed reads to the genome. Before running actual alignments, an index for STAR alignment is created by the command

```
STAR --runMode genomeGenerate \
     --runThreadN 1 \
     --genomeDir $ref_dir \
     --genomeFastaFiles $genome_seq \
     --sjdbGTFfile $genome_annot \
     --sjdbOverhang 149
```

Notice that the parameter `--sjdbOverhang 149` was added to the command instead of the default value (99) in order to reflect the fact that the original raw reads are produced by 2 $\times$ 150bp paired-end sequencing protocal, given that the best value for this parameter is usually the sequence length - 1. `--genomeDir $ref_dir`

represents the directory that holds the STAR index output, where as `--genomeFastaFiles $genome_seq` and `--sjdbGTFfile $genome_annot` represents the actual genome sequence file (in `FASTA` format) and the genome annotation file (in `GTF` format), respectively. Human genome `hg38` is used in this project, and a bash script is used to download both genome files (`hg38.fa.gz` and `hg38.ncbiRefSeq.gtf.gz`) from the UCSC Genome Data website. The script also contains the commands to execute the STAR index creation after verifying that the required genome files are downloaded and gzipped in the correct directory.

Quality control

Feature counts

Differential gene analysis

Discussion

Code Availability

Reference

1. Zhang, J., Lin, D., Li, K., Ding, X., Li, L., Liu, Y., Liu, D., Lin, J., Teng, X., Li, Y., Liu, M., Shen, J., Wang, X., He, D., Shi, Y., Wang, D., & Xu, J. (2021). Transcriptome analysis of peripheral blood mononuclear cells reveals distinct immune response in asymptomatic and re-detectable positive COVID-19 patients. *Frontiers in Immunology, 12*. https://doi.org/10.3389/fimmu.2021.716075