University of TEHRAN

Algorithms in Bioinformatics

Project 2: RNA-Seq analysis

Due date:Khordad 20, 1402

Teaching assistants: Sajedeh Bahonar and Mohammad Sadegh Vafaei

Instructors: Alireza Fotuhi Siahpirani and Hesam Montazeri

Department of Bioinformatics, IBB, University of Tehran

Objective

In this project, you will learn how to extract a gene expression matrix from fastq files and perform advanced analysis on the resulting data. Specifically, you will analyze the next-generation sequencing expression profiles of unpaired normal and COVID-19 FFPE bronchoalveolar lavage or lung samples from a study with accession number GSE190496. The table below provides information about the selected samples from this study that will be used in your analysis.

Tissue	SRR Accession number
FFPE Normal human lung	SRR17172463, SRR17172465, SRR17172466, SRR17172467
FFPE COVID19 patient lung	SRR17172468, SRR17172469, SRR17172470, SRR17172471
FFPE Normal human brochoalveolar lavage cells	SRR17172480, SRR17172481, SRR17172482
FFPE COVI19 Patient brochoalveolar lavage cells	SRR17172485, SRR17172486, SRR17172487

University of TEHRAN

Data preparation

Each student is supposed to preprocess paired data of normal and covid19 lung tissue samples. To do this project, you need to download fastq files from SRA using sratoolkit using the *fastq-dump* (Hint: use *fastq-dump [options] file.fastq.gz*)

Part a- Quality control and trimming

In this step, first, assess the read qualities using the *FastQC* software. Then use the *Trimmomatic* software to improve the read qualities through the read trimming. Recheck the read qualities to make sure the problems are solved. (Hint: you may use *fastqc* and *TrimmomaticPE*)

Tips on trimming:

Note that the trimming command first deletes the Adaptor sequences and then removes low-quality bases. Eliminate bases with a quality of less than 30 and then eliminate reads shorter than 80 bp.

Please answer the following questions.

1. What is the average number of reads across samples before and after the read trimming?
2. Compare the read length averages in different samples before and after the read trimming?
3. Compare the read quality distributions over all sequences before and after the read trimming.
4. What does the Adaptor Content warning indicate?
5. Why do we first remove the Adapter sequences for the reads and then the low-quality bases?
6. What does the quality of bases mean, and how is it obtained?

Part b- Read mapping

In the second step, map the reads to the reference genome using the *HISAT2* software. To map reads to the reference, the *HISAT2* software uses a graph-based alignment and a variety of metrics. The output of this step will be a SAM file for each sample. (Hint: you may use *hisat2 [options]* -x <ht2-idx> {-1 <m1> -2 <m2> } [-S <sam>]*)

Use the *Samtools* software to convert the *HISAT2* SAM files to BAM files (Hint: you may use *samtools view [options] <in.sam>*). Make sure you delete the SAM files afterwards.

This step uses the *hisat2-build* software to index the Homo sapiens (GRCh38) reference genome (available at this link). The fasta files of the genome have already been downloaded and placed in the path " */home/studmin/Proj_02/Homo_Genome/Homo_sapiens.GRCh38.dna.toplevel.fa*" . Do not copy the genome fasta file into your directory. Refer to the HISAT2 manual for a more detailed explanation. Please answer the following questions. (Hint: you may use *hisat2-build [options]* <reference_in> <ht2_index_name>*)

1. What is the difference between SAM and BAM files?
2. What is the purpose of indexing the genome?
3. Report mapping percentages of all samples in a table . Please explain why a low percentage of reads cannot be mapped.

Part c- Building gene expression matrix

In the third step, run htseq-count on count aligned reads for differential expression analysis. You can use the HTSeq documentation for further explanation. In this step, you need a gene/transcript annotation file that you can download from this link. Merge results files into a single matrix for use in the *edgeR* package.

Please answer the following question after completing this step.

1. How many genes are not expressed in control and covid samples? Explain the results.
2. Compare the matrix obtained at this stage with the corresponding gene expression submatrix of the main study. Discuss the differences.
3. What are other software available to do this step? Name two other software and discuss their advantages and disadvantages.

Part d- Differential gene expression analysis

Use edgeR to perform differential gene expression analysis. The edgeR is an open-source Bioconductor package designed for differential expression analysis based on count-based RNA-seq data. Use the expression matrix of the main study (all samples) to answer the following questions.

1. How many genes are given to edgeR? How many of them are differentially expressed in covid versus normal samples? How do you define statistical significance in this context?
2. Determine the percentage of differentially expressed genes with |log2FoldChange| > 1.5.
3. Explain the difference between P-value and FDR?

Part e- Gene Ontology enrichment analysis

In the final step, to obtain the distinct biological functions present in cancer samples, use the GOseq package in R to perform Gene Ontology (GO) enrichment analysis. The GO enrichment analysis statistically assesses the overrepresentation of differentially expressed genes in common GO functional branches. To accomplish this step, select the genes with FDR < 0.1 and an absolute value of Log2FoldChange > 1.5. Please answer the following question after completing this step.

1. Display results related to Biological Process, Molecular Function, Cellular Component, and KEGG as separate plots using an R package of your choice.
2. Do a brief study of each of the significant terms and discuss which terms you think may play an important role.
3. Write a general biological conclusion about the final results of the project.

Important note: report all the results in a directory named ***Project2*** containing the subdirectories for QC results, the output of the trimmed data, the output of the mapping to the reference genome, the output of the expression, the output of the differential expression gene (DEG), and the results of the Gene Ontology analysis.