BENG 183 Applied Genomic Technologies
Homework #3

Due: Monday November 6, 2023 by 5:00pm (PST)

Notes:

Please **ensure your terminal's username line is visible in all screenshots**; otherwise, the particular question will **NOT** be credited. (Not required for Problem 4)

Tutorial:

Tutorial for BENG 183 Homework 3 RNA-seq Analysis - 2023Fall
All problems will follow closely from steps in the tutorial.

Problem 1

Please download the sequencing files from the following link:
https://drive.google.com/file/d/12qNoEYrJk6xbInxLHX3jOZXimvWLVSB6/view?usp=sharing

Or from github repository for Homework 3:
https://github.com/Gaoyuan-Li/BENG183_HW3_2023Fall/tree/main/data_chrX

To check whether your downloaded files are intact, please provide a screenshot of the first 8 rows of ERR188044_chrX_1.fastq. (5 pts)

Please calculate how many reads are there in ERR188044_chrX_1.fastq. The screenshot of your command is required for full credit. (5 pts)

Problem 2

Use FastQC to check the sequence quality of sample ERR188044_chrX_1.fastq. Please attach a screenshot of the FastQC output from your terminal (5 pts). In the html report, please find and attach Per base sequence quality plot and Sequence duplication levels. (5 pts)

Problem 3

You should see there are a lot of replicated sequences being detected in the ERR188044_chrX_1.fastq and ERR188044_chrX_2.fastq. Most of them could be sequence adapters and we would like to remove them before the next step.

We can use fastp to remove the adapters. The output (cleanned fastq) should be named ERR188044_chrX_1_clean.fastq and ERR188044_chrX_2_clean.fastq. Please attach a screenshot of the statistics of Duplication rate. (10pts)

Problem 4

Run fastqc on the trimmed sequence. Discuss the changes you noticed between the trimmed and raw data with figures attached. (10pts)

Problem 5

Build reference genome index: You will next align these clean reads to a chromosome X genome.

Attach a screenshot of the STAR generated genome index folder (use --runMode genomeGenerate). The genome index folder should include SAindex, chrNameLength.txt, sjdbInfo.txt and other files. (10pts)

Problem 6

Next, you will align your cleaned reads to the reference genome you just built.

After you finish mapping each sample clean reads to the reference, you are supposed to get a few reports from STAR.

For example, you are supposed to get the following results:

ERR188044_STAR_Aligned.sortedByCoord.out.bam
ERR188044_STAR_Log.final.out
ERR188044_STAR_Log.out
ERR188044_STAR_Log.progress.out
ERR188044_STAR_SJ.out.tab

Report “Number of input reads” (5 pts) and the percentage of the “Uniquely mapped reads” (5 pts) with a screenshot of ERR188044_STAR_Log.final.out file. Please attach a screenshot of the STAR output from your terminal (5 pts).

Problem 7

Use samtools view command to open your bam file and attach a screenshot of the first 3 lines of this file. (10 pts)

Problem 8

In the bam file you opened in the previous question, please find reads ‘ERR188044.27155124’. You’re supposed to see two reads that share the same name because our sample library is a pair-end library. Please attach the screenshot of these reads from samtools view (10 pts).

Report the number of bases that matched the reference genome for each read. (5 pts)

Problem 9

After you get the bam file from STAR alignment, you will use featureCounts to quantify how many reads are mapped to each gene.

Please report the number of Unassigned_NoFeatures reads from ERR188044_count.txt.summary file and attach a screenshot of this summary file (5pts). Please attach a screenshot of the featureCounts output from your terminal (5 pts).

Problem 10 (BONUS, 15 pts)

So far, we only finished the RNA-seq pipeline from raw reads processing to gene abundance quantification. Next step for this practice is to find out the differentially expressed genes between different samples.

To achieve this goal, you will need to finish the above analysis for all of the samples. After you get the raw read counts for each gene, you could use DESeq2 packages in R to identify the differentially expressed genes.

Here, we will use the Tutorials in Discussion Session 3 (Please see Media Gallery on Canvas) and Homework 3 Github Repository BENG 183 Homework 3 RNA-seq Analysis - 2023Fall.

Report the top10 DE genes (sort based on the adjusted p-value from low to high) between conditions(10 pts).

Use clusterProfiler to understand the biological related functions of the top100 DE genes. Please attach the figure generated for full credit (5 pts).

Note: The count table created by featureCounts is available in the Homework 3 Github Repository BENG 183 Homework 3 RNA-seq Analysis - 2023Fall. So you can finish Problem 10 without any foundation from Problem 1-9.