BENG 183 Applied Genomic Technologies
Homework #3

Due: Monday October 31, 2022 by 5:00pm (PST)

Notes:
For every screenshot please include your username line in the terminal (If not, we will NOT give any credit for that question)

Problem 1

Please download **female_midgut1_R1_raw.fastq** and **female_midgut1_R2_raw.fastq** from Homework 3 Github Repository .

1. To check whether your downloaded files are intact, please provide a screenshot of the first 8 rows of **female_migut1_R2_raw.fastq** (0.5 pt)
2. How long is the first read? (0.5 pt)

   Hint: You could use linux command like head ,wc

Problem 2

Please calculate how many reads are there in **female_migut1_R2_raw.fastq** (1 pt). The screenshot of your command is required for full credit.

Hint: Each read starts with ">"symbol. You could use grep command to grab sequence with assigned pattern. You could use wc -l to count the number of lines

Problem 3

Use fastqc to check the raw sequence quality of sample **female_midgut1_R1_raw.fastq**. Please attach the screenshot of **Per base sequence quality plot** and **Sequence duplication levels** from fastqc report (2*0.5=1 pt).

Problem 4

You should see there are a lot of replicated sequence been detected in the **female_midgut1_R1_raw.fastq**. Most of them could be sequence adapters and we would like to remove them before next step.

We can use fastp to remove the adapters. The output (cleanned fastq) should be named
female_midgut1_R1_clean.fastq and **female_midgut1_R2_clean.fastq**

For the usage of fastq, please check the Homework 3 Github Repository **Tutorial2_RawData.md** or the documentation of fastp.

Please attach a screenshot of the statistics of Duplication rate. (1pt)

Problem 5

Pre mapping —build reference genome index: You will next align these clean reads to a drosophila genome reference. Follow the instructions related to download reference genome; building index of reference genome; using STAR align to reference genome on Homework 3 Github Repository **Tutorial3_Mapping_and_qualification.md**.

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. (1pt)

Problem 6

Mapping. Next, you will be asked to align your cleaned reads to the reference genome you just built. Follow the instruction from Homework 3 Github Repository **Tutorial3_Mapping_and_qualification.md**.

After you finished mapping each sample clean reads to the reference, you are supposed to get a few reports from STAR. For example, for Female_midgut1 sample you are supposed to get the following results:

1. Female_midgut1_STAR_genomeAligned.sortedByCoord.out.bam
2. Female_midgut1_STAR_genomeLog.final.out
3. Female_midgut1_STAR_genomeLog.out
4. Female_midgut1_STAR_genomeLog.progress.out
5. Female_midgut1_STAR_genomeSJ.out.tab

Report the unique mapping rate and Number of input reads with a screenshot of **F_midgut1_STAR_genomeLog.final.out** file (1 pt).

Problem 7

Use **samtools view** command to open you bam file and attach a screenshot of the first 3 lines of this file (hint: you might need head command) (1pt)

Problem 8

For the bam file you open in previous question, please find reads D00261:358:C9JAVANXX:1:2303:17663:2997 . 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**(1 pt).

Report the number of bases that matched to the reference genome for each read (0.5 pt) and which chromosome(s) they mapped to (0.5 pt). hint: you might need grep command.

Problem 9

Gene abundance quantification: After you got the bam file from STAR alignment, you will use featureCounts to quantify how many reads are mapped to each gene. Follow the same tutorial Homework 3 Github Repository **Tutorial3_Mapping_and_qualification.md**. The final output should include: (1pt)

1. F_midgut1_count.txt
2. F_midgut1_count.txt.summary

Please report the number of **Unassigned_NoFeatures** reads from **F_midgut1_count.txt.summary** file. Attach a screenshot of this summary file.

Hint: use cat to open your file so that your username line will be included.

Problem 10

So far, we only finished RNA-seq pipeline from raw reads processing to gene abundance quantification. However, the goal for this practice is to find out the differentially expressed genes between drosophila head tissue and midgut tissue.

To achieve this goal, you will need to finish the above analysis for EACH of the four samples (head1, head2, midgut1, midgut2). After you get the raw read counts for each gene, you could use DESeq2 packages in R to identify the differentially expressed genes.

Tutorials are in Discussion Session 3 (Please see Canvas Files and Media Gallery) and Homework 3 Github Repository **Tutorial4_DE.md**).

Report the top10 DE genes (sort based on the adjust pvalue from low to high, here genes with smallest pval) between head and midgut (2 pts).