Bios 555 High-throughput data analysis using R and Bioconductor

Homework 4

Due on **Oct 13th, 2020 (Wednesday) before 11:59pm**.

I. Read Wikipedia pages for “DNA sequencing”, “RNA-seq”, “ChIP-seq” and “negative binomial distribution”.

II. Short answer questions, 5 points each. Be creative in answering the questions.

1. What is DNA sequencing? Why do we want to do DNA sequencing?
2. Compared to traditional sequencing method (Sanger sequencing), what are the pros and cons of second-generation sequencing?
3. Briefly describe the workflow of second-generation sequencing data analysis.
4. Suppose after one run, the sequencing machine generated 1 million sequence reads, each of 50 base pairs long. What will be the dimension of the raw intensity data?
5. What’s the difference between sequence alignment and assembly?
6. Compared to the gene expression microarrays, what additional information can RNA sequencing provide?
7. What are the major differences for RNA-seq and expression microarray data? How are they modeled in DE test procedures?

III. Based on the results from lab, answer the following questions:

1. (10 pts) Based on the bowtie alignment results for bacteriophage, how many reads can be aligned to the reference genome? What about the results from Rsubread?
2. (30 pts) Write a short report for the integrative analysis of RNA-seq and C-Myc ChIP-seq data for K562 cell lines. (Hint: briefly describe the procedures of getting read counts, and illustrate that C-Myc binding and gene expressions are correlated.)
3. (15 pts) Compare the results of DE test from DEseq, edgeR and DSS for the simulated data.