Bios 560R High-throughput data analysis using R and Bioconductor

Homework 6

Due on **Dec 6th, Tuesday before class at 1pm**.

I. Read Wikipedia pages for “RNA-seq”, “alternative splicing”, “negative binomial distribution”.

II. Short answer questions, 10 points each. Be creative in answering the questions.

1. Compare to gene expression microarrays, what additional information can RNA sequencing provide?
2. In read counts summarization, why are people not satisfied by counting the number of reads within the gene bodies? (hint: give example of biases and potential problems.) Can you come up with a summarization algorithm yourself?
3. Why data normalization is necessary for RNA-seq counts?
4. What are the major differences for RNA-seq and expression microarray data? How are they modeled in DE test procedures?
5. What is alternative splicing? What is isoform? What is the goal in isoform expression estimation?

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

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