BIMS 8701 Assignment 5: Due March 18, 2024

Stefan Bekiranov

March 2, 2024

For this Differential Expression Assignment/Lab, you will demonstrate the equivalence of linear modeling and performing a t-test, generate log normal and negative binomial distributed synthetic RNA-seq read count data and compare use of linear modeling and DESeq2 for determining the set of differenially expressed genes using R.

1. Problem 1 (2 points): Demonstrate the equivalence of assessing significance using a linear model and a t-test.

    - Generate 5 replicate treatment samples using `rnorm` with `mean = 4` and `sd = 1` and 5 replicate control samples using `rnorm` with `mean = 2` and `sd = 1`.
    - Concatenate the treatment and control samples to create the output of the linear model and use a vector of `1`s and `0`s as input to `lm`. Use `summary` to display the results.
    - Perform a t-test with `var.equal = T` and compare the resuling p-value and difference in means derived from using `lm`.

2. Problem 2 (2 points): Generate count matrices that DESeq2 will accecpt as input with 10,000 genes, 5 replicate treatment samples and 5 replicate control samples which are log normal and negative binomial distributed and 100 genes are upregulated in treatment compared to control.

    - log normal case: For 100 genes, generate 5 treatment replicates using `rlnorm` with `meanlog = 8` and `sdlog = 0.2` and 5 control replicates using `rlnorm` with `meanlog = 6.9` and `sdlog = 0.2`. For the remaining 9900 genes, use `rlnorm` with `meanlog = 6.9` and `sdlog = 0.2`. You may find functions `floor`, `matrix`, `paste` and `as.integer` useful.
    - negative binomial case: For 100 genes, generate 5 treatment replicates using `rnbinom` with `size = 24.7` and `prob = 0.008` and 5 control replicates using `rnbinom` with `size = 25.1` and `prob = 0.024`. For the remaining 9900 genes, use `rnbinom` with `size = 25.1` and `prob = 0.024`. Again, you may find functions `floor`, `matrix`, `paste`, and `as.integer` useful.
    - For the log normal and negative binomial distributed cases, plot the log transformed distributions of the 100 upregulated treatment and control genes on the same plot with separate treatment and control labels. Do the same with the 9900 genes that are not differentially expressed.

3. Problem 3 (3 points): Generate a dataframe containing statistics associated with differential expression and determine the number of differentially expressed genes for both the log normal and negative binomial distributed cases by performing the equivalent of a t.test using linear modeling.

- The goal is to generate a dataframe containing columns with gene names, log2 fold changes, p-values and FDR adjusted p-values associated with differential expression between treatment and control samples for the log normal and negative binomial distributed cases.
- Use log2 transformed count data as output and a vector of `1`s and `0`s as input to `lm` to derive the log2 fold change and p-value for each gene and `summary` to extract these values. The easiest way to do all of this is in a `for` loop with an empty dataframe initialized before the loop.
- Once all 10,000 log2 fold changes and p-values are derived, use `p.adjust` with option `"BH"` to derived the FDR adjusted p-values. Add this as a column to your dataframe and apply a 0.05 cutoff to your adjusted p-values to determine the number of differentially expressed genes.
- Plot -log10 p-value (y-axis) versus the gene index (x-axis) as well as -log10 adj p-value (y-axis) versus gene index (x-axis).

4. Problem 4 (3 points): Apply DESeq2 to your log normal and negative binomial distributed count matrices, derive the results table and apply a 0.05 cutoff to the adjusted p-values to determine the number of differentially expressed genes. Compare these numbers to those derived from application of `lm`. Do your results match expectations? Briefly, why or why not?