BENG 183 Applied Genomic Technologies (Fall 2019)
Due: November 19, 2019

Homework 4

Total: 10 points

Online Tutorial 4: Analysis of differentially expressed genes
Please go through Tutorial 4 to learn how to check the quality of samples and conduct differential analysis using DESeq2, and answer the following questions:

1. **Replicates/experiments quality check.** Among the four samples of Drosophila, there are two replicate samples for head and two replicates for midgut. Based on the raw-count matrices generated by *featureCounts*, you are now able to check whether the gene expression correlates between replicates.

   a. Plot the log2 transformed TPM values of one replicate on X axis and the log2 transformed TPM values of the other replicate on Y axis. Thus, you have created a scatter plot. Attach your scatter plots below, one for the head samples and one for the midgut samples. Make sure you have clear axis labels to specify which samples you plot. (1 pt)

   (Scatter plot for head samples)

   (Scatter plot for midgut samples)

   b. To have a quantitative measurement of how the replicates correlate, you can calculate the Pearson correlation coefficients ρ of two replicates using the formula below:

$$\rho = \frac{\mathrm{E}[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

   where X and Y represent log2 TPM values of two replicates. μ_x and σ_x are the mean and standard deviation of log2 TPM values in X, respectively. Same as σ_y and μ_y for Y. Expectation is calculated like mean. List the Pearson correlation coefficients for the replicates of head and midgut samples. (1 pt; hint: function “mean()” and “sd()” in R can be useful for calculation)

   c. After doing hierarchical clustering for the four samples on log2 TPM values, attach the cluster dendrogram below. (1 pt)

   d. What is your conclusion regarding the quality of the four samples after seeing all the results from a-c? Briefly describe your reasoning. (1 pt)

BENG 183 Applied Genomic Technologies (Fall 2019)
Due: November 19, 2019

2. **Differential gene analysis.** DESeq2 is currently the most widely used tool to identify differentially expressed genes across samples. It generates more reliable results than other existing tools.

   a. The better performance of DESeq2 is largely due to how its algorithm models gene expression. What factors of the gene expression data does DESeq2 correct/normalize for? (1 pt; hint: there are two factors)

   b. On the log fold change vs. mean of normalized counts plot, we do not see many differentially expressed genes (red dots) with less than 10 normalized counts even if they have large fold changes. Which of the corrected factors in *question 3a* accounts for this observation? Briefly describe why it is important to correct for this factor. (1 pt)

   c. To identify differential genes, the best practice is to set cutoffs for both fold change and significance value. How many genes are differentially expressed by at least 2-fold changes (absolute log2 fold change greater than one) with adjusted p-values less than 0.05 between head samples and midgut samples? (1 pt)

   d. Attach a heatmap showing log2 TPM values of the differentially expressed genes in *question 3c* for the four samples. (1 pt)

   e. How many of the differential genes in *question 3c* have higher gene expression in head samples and how many have higher expression in midgut samples? (1 pt)

   f. List the top three pathways that the two groups of differentially expressed genes in *question 3e* are enriched for. (1 pt)

Top pathways	Genes with higher expression in head	Genes with higher expression in midgut
1		
2		
3		