Day 8: Advanced DESeq2 Experimental Designs Homework
By: Samuel Hunter sahu0957@colorado.edu

The homework data and metadata are located in /scratch/Shares/public/sread2022/day8/homework_data_files/
The homework will be completed on your local computer using R. Download all files before you start your homework.

Question 1. Load in the files labeled h8_data1.txt and homework_day8_q1_metadata1.csv as data frames in R.

A. How many different batches are there?
B. How many different values are in the treatment column?
C. Create a DESeq2-compatible matrix from the count data. Run DESeq2. Use a design that includes a batch effect correction and the treatment.
D. What are the names of the design matrix columns for your DESeq2 object?
E. Which gene is the top hit in arsenic treatment? What is the adjusted p-value?

Question 2: Load in the file labeled h8_data2.txt

Uh-oh! You accidentally deleted your metadata file! You'll have to re-create it before running DESeq2. Luckily, you named all of your columns in your counts file using the same pattern: sex_treatment.
You used three different treatments: control, LPS, and PolyI:C (PIC)

A. Rebuild the metadata file. You can use any program of your choice (R, Excel, vim, etc.)
B. You decide not to look at differences between the sexes in treatment. Write the design formula that investigates both treatment effects but not differences between the sexes.
C. Run DESeq2. What is the top hit (lowest padj) for LPS treatment? How about PIC? Hint: Use a contrast
D. You decide to look at the generalized immune response by averaging the results of the LPS and PIC treatments. What numeric contrast vector would you use to accomplish this? Hint: you can print the design matrix used in your DESeq2 object using attr(dds, "modelMatrix")
E. What gene is the top hit (lowest padj) from the comparison in part D?
F. Reviewer 2 thinks you weren't thorough enough in your analysis above- after all, you didn't correct for variations (not interactions) due to sex differences. What design formula would you use to satisfy their request?
G. Did the top genes change from part C using the new design formula?

Question 3: Load in the file h8_data3.txt and h8_day8_q3_metadata.csv

A. This is time series data, with data spanning over 3 days in both a wild-type and a mutant background. What design formula would you use to query the interaction between the genotype and the treatment, correcting for batch effects?
B. You want to run a likelihood ratio test to test whether the interaction effects are significant for any gene. What reduced model would you use?
C. Run DESeq2 using your design above. What is the top gene (lowest padj)?
D. You want to generate a heatmap, using the top 100 gene hits from part C.
    a. First, fetch the top 100 hits by adjusted p-value
    b. Next, normalize the counts of your DESeq2 object
    c. Filter your normalized counts to the top 100 genes
    d. Rescale your counts to Z-scores
    e. Generate the heatmap using pheatmap