MED263: Homework
Winter 2017
RNA-Seq
Instructor: Kathleen Fisch, Ph.D.
Email: KFisch@ucsd.edu

Instructions:
For this assignment, you will be running an RNA-Seq differential expression analysis on a human disease dataset (TCGA LUAD), using the Jupyter Notebook provided (Filename: RNASeq123-MED263HW.ipynb). The code in the notebook has already been configured to run this analysis on a different dataset than the one provided during the in-class practical.

Step 1: Download the notebook here http://ccbb-analysis.s3.amazonaws.com/Katie/MED263/RNASeq123-MED263HW.ipynb .
Step 2: Open the new analysis notebook and go to File → Make a copy → Rename the file with _yourname.
Step 3: Start at the top of the notebook and execute each cell in order. Some cells will have Homework Question cells as the next cell. Pause at each of these cells and answer the questions below.
Step 4: For each Homework Question, analyze the output from the desired step in the analysis and answer the questions in at least 2-3 full sentences, including figures when necessary (you can right click on the images in the notebook to Save As).
Step 5: Save this document with your homework question answers and save the fully executed Jupyter notebook as an html file (File → Download As → Html) for submission.

Questions:

1. We are using a subset of the TCGA LUAD RNAseqV2 dataset, that comprises 32 samples from 16 patients (Tumor/Normal). Open the counts file and meta file downloaded in the step above to view them. Navigate in your web browser to the GDC Data Portal (https://gdc-portal.nci.nih.gov/projects/TCGA-LUAD) and Firebrowse website http://firebrowse.org/ and type in the TCGA cohort = LUAD (Lung Adenocarcinoma) to read about the dataset. Read the information provided on the websites, and navigate to the provided publication https://tcga-data.nci.nih.gov/docs/publications/luad_2014/ article: http://www.nature.com/nature/journal/v511/n7511/full/nature13385.html. A. What is the title of the study? B. What is the design of the RNAseq part of the study? (i.e. number of samples, disease type, conditions, PE vs SE) C. What was the goal of the study? D. What RNA-seq pipeline did they use (i.e. what tools did they use for alignment and quantification)? E. What are two biological questions you can answer with the RNA-Seq experimental design for the 32 paired samples?
2. A. What are two reasons for filtering lowly expressed genes? B. What proportion of genes were filtered out that had 0 counts across all 32 samples? C. How many genes were retained after filtering for genes with a CPM >1 across at least 10 samples? D. Insert the raw data and filtered data Log-cpm vs Density.

3. A. Describe TMM normalization in 2-3 sentences. B. Insert the Bar graph of TMM Normalized samples. C. Interpret the graph in words.
4. Explore the MDS plots, including the interactive MDS-Plot. A. What variable seems to account for the proportion of variation explained by dimension 1 (by looking at it)? B. Does library size seem to have an effect on the variance in the samples? C. Identify patient paired samples. Do you expect from looking at this that patient will have a larger effect on differentially expressed genes compared to condition (tumor normal)? Why or why not?
5. Explain why we want to remove heteroscedasticity from the data and how we do this with voom. Refer to the voom publication if necessary https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29.
6. A. How many differentially expressed genes between tumor and normal samples are there at an adjusted p value < 0.05? B. How many differentially expressed genes between tumor and normal samples are there at and adjusted p value < 0.05 and a log2 fold change cutoff of 1? C. Open the results spreadsheet called DE_tumorvsnormal.csv and view the differentially expressed genes. Sort them by adjusted p-value and examine the logFC values. What is the highest logFC? What is the lowest logFC? What genes do these correspond to? D. Cut and copy the entrezIDs from this file that have an adjusted pvalue of < 0.01. Input this list into the ToppFun tool on the ToppGene website https://toppgene.cchmc.org/enrichment.jsp and perform a gene set enrichment analysis with default settings. What are the top three significant pathways and their B&H pvalues? Are these hits relevant to the disease under study? How do they compare to the Camera results?