**BIOM262 Midterm Exam 2017**

**Part 1: Short answer questions (Only answer 3 out of the 5 questions, 3-5 sentences per response)**

1. What are different features of alignment algorithms that we can consider in choosing one versus another?
2. How would one benchmark alignment algorithms?
3. What are PCR duplicates? How would you recognize that and when do you remove PCR duplicates? And when do you not?
4. What is the difference between FPKM and RPKM? FPKM and TPM? What is the advantage of using TPM?
5. How is stranded information maintained in an RNA-Seq library prep? What is the difference between a stranded and unstranded library?

**Part 2: Quantify Differential expression in an RNA-Seq Dataset**

1. Download a dataset from ENCODE, GEO, or use a dataset that you have available from your research project. You must have at least two replicates per condition and two conditions to compare. Write 3-5 sentences summarizing the dataset that you have chosen (read depth, read length/type, how the libraries were generated, where the RNA came from, sequencing instrument, etc.), and what a differential expression analysis will tell you about your biological system.
2. Run fastqc on your dataset and summarize what you have learned in a few sentences. Include your directory location on tscc where I can find the results of your analysis (make sure I have read permissions on those files – this is especially important if you are using your own account, I have access to all of the ucsd-train accounts regardless of permissions…. Big brother is watching).
3. Perform the alignment and quantify differential expression. Make a table with the most important metrics that you get from the Log file in STAR summarizing the mapping results (copy your table directly into the word document with the rest of your responses). Provide the directory where I can find your STAR results (again… read permissions). Write a few sentences commenting on the parameters you chose for your alignment and why.
4. Quantify gene expression and use those quantifications to perform a differential expression analysis with DESeq2. Generate a MA-plot to summarize your differential expression results (copy the MA plot directly into the word document with your responses). Comment on the significance cutoffs you chose for your analysis and why. Summarize what you have learned about differential expression in your dataset from the MA-plot.
5. Make a clustermap of TPM specifically for the genes that you identified as differentially expressed. How many genes are included in your analysis? What can you conclude from this figure? Perform this analysis in one notebook, download the .ipynb file (or.ipynb.json) and name it lastname\_firstname\_tpm\_clustering.ipynb

**Part 3 Statistics – See R notebook**

**Submission Instructions:**

The midterm is due on Sunday February 19 at 11:59PM. Submit your answers as a word document with the file name: lastname\_firstname\_midterm\_exam along with any additional files specified in the instructions (TPM clustermap notebook and R notebook). Email all files to Emily and Jamison.