CSCI 5461 Homework 1 (100 points, Due date: Monday, March 13th, 11:59 PM):

For our first homework assignment, we will apply the statistical methods we learned in class to identify differentially expressed genes with patients that have been diagnosed with Glioblastoma Multiforme (GBM). We have separated GBM patients into two cohorts of short- (<= 1 year) and long-term (>1 year) survivors. This RNAseq data set was generated as a part of the Cancer Genome Atlas (TCGA). The detailed description of the dataset can be found under the following link (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers/glioblastoma).

The raw data is publicly available from the Gene Expression Omnibus (GEO) database, which will provide you with links to the raw data, but we have also preprocessed the data for you so you can instead just focus on the downstream analysis. We have used an R script to preprocess the data, but please feel free to implement your solution in any language you choose.

Note about working in teams: As described in class, you have the option of completing this homework individually or as part of a team of **at most** 2 students. Teams must consist of students with complementary background/expertise (e.g. one student with a primarily computational background, one with a primarily biology background). If you are unsure whether you are a complementary pairing, please confirm with Prof. Myers in advance of completing the homework. Teams will need to answer a few additional questions in the homework (marked with **) and will need to submit a **single** copy of the report/code implementing the solution. Both students will be assigned the **same grade** for the assignment.

1. **Acquiring the gene expression data:** Go to the GSE6944 repository page and read the descriptions of the dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62944). We have already downloaded the count and clinical data in the file `getRNASeq_data.R` and included it for you in `HW1-GSE62944-count.csv` and `HW1-GSE62944-clinical.csv` respectively. For the count data, the rows are the genes and the columns are the patient identifiers. For the clinical data, there are two columns, the “sampleName” and “Group” that identifies patients as “short” or “long” survivors. The final piece of data is the analysis from the *DESeq2* method. We have completed that analysis and saved it to the file `HW1-DESEQ2.csv`. If you’re interested in the details of the pipeline that was used to process the raw RNAseq data, you can read about those details in this paper (https://www.ncbi.nlm.nih.gov/pubmed/26209429).

** **Additional question for teams:** Read the Rahman et al. paper (https://www.ncbi.nlm.nih.gov/pubmed/26209429) and describe the computational pipeline they used to produce the gene-level read-count data linked above.

2. **Exploring the dataset (10 points):** Process the data file using whatever programming language you have chosen and answer the following questions:
   (a) How many genes are included in the dataset?
   (b) How many patient samples are in the dataset? We have divided patients into two "Groups" based on their survival time. How many patients survived short- (<= 1 year) and long-term (>1 year)?

   **** Additional question for teams:** Read McLendon et al. (https://www.nature.com/articles/nature07385), the paper that originally published this data. What was the goal of their study? Briefly summarize what they did and what they concluded from this gene expression dataset.

3. Data processing and normalization (30 points):

(a) First, compute the total number of reads per sample and plot a bar graph that summarizes these values (one value per sample). Describe what you observe.

(b) For this assignment, we are going to perform a "total count" normalization for read-depth in each sample. To do this, normalize each gene's read count in each sample as follows:

$$g'_{ij} = g_{ij} * (\frac{N_{med}}{N_i})$$

where g'_{ij} is the normalized count for gene j in sample i, N_i is the total number of reads mapping to all genes in sample i, and N_{med} is the median number of total reads taken across all samples. For our question, we are only interested in comparing each gene's expression to its own expression levels in other samples, so we are not normalizing by gene length. For other applications, you would also want to normalize by gene length (e.g., by computing either TPM or FPKM). See this review for more details (https://academic.oup.com/bib/article/14/6/671/189645). Once you've completed this normalization, remake the plot in (a) to verify that it produces the expected result.

(c) Log-transform the normalized count data of the complete dataset (all genes, all samples). To allow for genes with 0 counts, add 1 "pseudocount" to each gene before log transforming, $g'_{ij} = log\ (g_{ij} + 1)$. Plot a histogram of this log-transformed data (x-axis: log-transformed expression levels, y-axis: frequency). Use the log-transformed data for all analysis that follows.

(d) Plot individual histograms of the log-transformed data for the first five samples. How do the distributions compare from sample to sample?

(e) Implement and perform quantile normalization **on your log-transformed data** (from **part c**) across all samples such that each has the same empirical distribution. Use the mean quantile corresponding to each data point across the whole set of samples as the reference distribution for this normalization. The mean quantile for a given data point in a dataset can be derived by sorting each sample's expression values and then computing the mean across all samples of these sorted values at the corresponding point in the sorted lists. This paper contains a complete description of the quantile normalization algorithm: Bolstad et al. Bioinformatics (2003)
https://academic.oup.com/bioinformatics/article/19/2/185/372664
Make sure to implement the quantile normalization yourself and do not use a built-in function.

Plot a histogram of the quantile-normalized data for each of the first five samples. Use the quantile-normalized data for the remaining problems.

4. Analysis of differential expression (30 points): Use the Wilcoxon rank-sum statistic to identify differentially expressed genes with a per-gene significance level of p < 0.05. You should divide the gene expression data into two groups (short- vs. long-term survival), using the "Group" variable, and test each gene independently. *You do not need to implement the Wilcoxon rank-sum test yourself– you may use a built-in function or package for this*.

(a) List the top 10 genes, the corresponding gene names, and the p-values associated with them. Pick 1-2 of these genes and discuss what is known about their function and how it might relate to GBM. There are many databases such as GeneCards (http://www.genecards.org), NCBI (http://www.ncbi.nlm.nih.gov/gene), and WikiGenes (http://www.wikigenes.org) that you can use to learn more about a gene.

**** Additional question for teams:** Comment on how well-established the genes that you discovered are in terms of their relevance to glioblastoma. Are there other genes that are well-characterized as playing a major role in the progression of glioblastoma, but that do not appear in your list?

(b) Report the number of significant genes for the test.

(c) What is the overlap between the set of genes deemed significant at an uncorrected p < 0.05 cutoff by the two different approaches (*DESeq2* and Wilcoxon rank-sum statistics)? You do not need to report the entire list associated with each approach. You only need to report the total number of genes that overlap and a list of those overlapping genes.

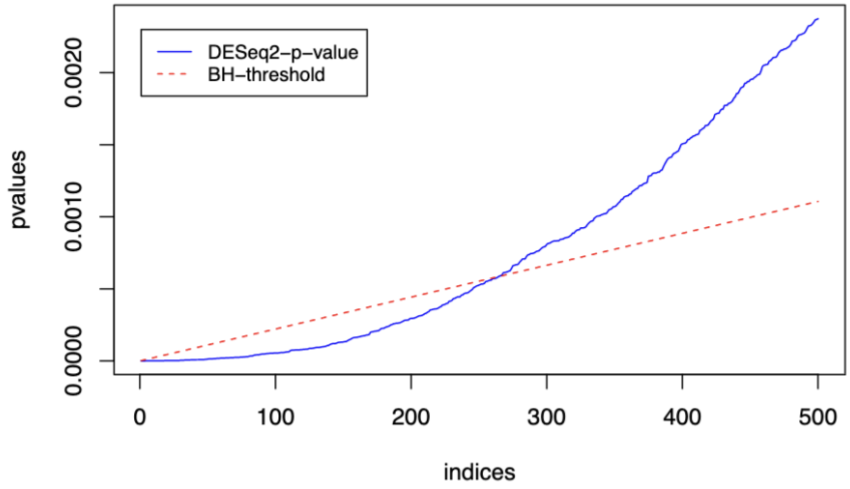
5. Multiple hypothesis correction (30 points):

(a) Use Bonferroni correction with the *DESeq2* results to identify differentially expressed genes at a global significance level p < 0.05 and report the number of significant genes.

(b) Use the Benjamini-Hochberg step-up procedure to control the False Discovery Rate (FDR) with the *DESeq2* statistics to identify differentially expressed genes at an FDR < 0.05. Report the number of significant genes after applying the BH procedure (specify which independence assumption you are using for the BH procedure).

(c) Rank the genes by their *p*-values in ascending order and plot the *p*-values and the threshold used for adjusted *p*-values as a function of the index of the ranked genes. More specifically, with the x-axis as the index of the ranked genes from 1 to the number of genes, plot the first 500 genes (*i* on the x-axis, the *p*-value of the i^{th} gene on the y-axis). As a separate color, plot the adjusted p-value threshold at each point (*i* on the x-axis, threshold for adjusted p-value on the y-axis), where *i* is the index of the genes and alpha=0.05.

HINT: Your plot should look like this:



Submission Instructions:

Zip all files into a single lastname.zip (individual submission) or lastname1_lastname2.zip (group submission) file and submit your .zip file on the Canvas site. Please avoid including the raw data files we provided in your .zip file. Your homework submission should only include:

1. Any source code you used to complete the assignment.
2. Report.pdf: A file with all of your plots and answers to questions.