CSC 315, Fall 2018
Bioinformatics Project

In this project, you will perform a bioinformatics analysis on a Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) dataset of interest to you. You may analyze any GEO series that has data obtained by *Expression profiling by array*. Possible datasets include those looking at *cancer*, *diabetes, heart disease, autism, asthma*, *cocaine addiction*, and many others. Many datasets are very specific and experimental in nature. Don't hesitate to contact me if you have any questions about a dataset you are interested in.

The datasets you select must have two groups that will be compared, with at least 3 samples in each group, and you cannot compare males and females. You should perform the following analyses using *R*, and turn in an *R* notebook that answers the following questions and/or completes the steps below. In addition, you will be expected to post a link to your dataset and a brief description of your intended analysis to Piazza (I will make a Piazza post shortly with more information). Each person must analyze their own unique dataset.

1. At the top of your R script, include your name and a title describing the analysis (e.g., Identifying differentially expressed genes between males and females)

2. Download the processed data from GEO, and pull out the expression data and phenotype data.

3. If necessary, take the log2 of the expression data. Then generate a boxplot of your data to show that the data is normalized. If there are more than 50 samples, you may choose to display the first 50 only. Give the boxplot a meaningful title and y-axis label.

4. How many samples were profiled, and how many probes are there in the dataset?

5. Pull out the column containing the data that you would like to compare (e.g., the gender column), and use *R* to output the number of samples in each group. Note: in some cases, you may have to process the data first, for example if the values were Male1, Male2, Female1, Male3, Female2, etc, the number would need to be removed. If you need help with this step, let me know.

6. Using *limma*, find the probes that are differentially expressed across the groups you are comparing, using a false discovery rate (FDR) of 20%. Output the number of probes identified, using the *nrow* function. If there are less than 50 probes with an FDR of 20%, find the top 50 probes. In either case, use *R* to output the overall FDR of your results, by outputting the *last* value of

the adj.P.val column of your results table. (Note: although we want that value to be low (such as 20%) for some datasets with small samples sizes, this value can be very high, even 1).

7. Create and output a data frame containing the top 3 probes, along with each logFC, and FDR (adjusted p-value) only.

8. For the probe with the lowest adjusted p-value, construct a boxplot showing the expression of that probe across the groups. The title of the boxplot should consist of the fold change (FC) and the FDR for the probe, and the boxplot must be constructed using ggplot. Note that the title includes the FC and not the log FC.

9. Use R to output the annotation (GPL platform) of the data that you are analyzing.

10. Using the *getGEO* function, download the platform (GPL) for this data.

11. Find the gene names corresponding to all probes, and create a table containing the corresponding gene names, the probe names, the logFC, and the adjusted p-values only. Hint: you should start with a table containing the probe name, logFC, and adjusted p-values, then add the corresponding gene names. Then, output the rows for the first 5 probes.

12. Using DAVID, identify Gene Ontology (GO) terms and KEGG pathways that are associated with the differentially expressed probes identified in (6). A screenshot of these results should be submitted with your R notebook.

13. Summarize your results based on your analysis. Your summary should include the GSE number analyzed, a description of the samples or individuals, including the number of samples and number of probes that were profiled, the number of samples in each group, the number of differentially expressed probes (and the corresponding FDR), the names of the top 3 genes, and the top GO terms or pathways associated with the phenotype that you analyzed.