**CSC 315, Fall 2021**
**Bioinformatics Project**

In this project, you will perform a bioinformatics analysis on an RNA-Seq dataset of your choice. All of the datasets available from Xena are posted on https://xenabrowser.net/datapages/. You must analyze *count* data (which will generally be labeled "HTSeq-Counts"). If you are interested, you may also find data from other sources, such as the Gene Expression Omnibus (GEO). RNA-Seq datasets can be found using the following link: https://www.ncbi.nlm.nih.gov/gds/. To find valid datasets, search for (with the quotes) `"expression profiling by high throughput sequencing"[DataSet Type]`. In order for us to analyze a dataset from GEO, there must be a csv file of gene counts that we can download (I can show you how to get the phenotype data). Xena datasets are generally limited to cancer data, while GEO will have other conditions (such as heart disease), but will be a little more challenging to analyze. Many GEO datasets are specific and experimental in nature, and don't hesitate to contact me if you have any questions about a dataset you are interested in.

The dataset you select must have two groups that will be compared, with at least 5 samples in each group. You cannot compare males and females. For tumor samples, common variables to be compared are stage and grade (https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html). You are also welcome to compare tumor samples (01-09) with normal samples (10-19). However, in this case you must account for the fact that individuals are sampled multiple times (and therefore their expression values will be correlated). I will guide you through this if you choose to do this comparison; an example can also be found on page 51 of the Limma User Guide.

***You may work in groups of up to 3 on this project.***

## <u>Assignment</u>

**By 5:00 PM on Monday, 11/29/21, e-mail me a link to your dataset (from Xena or GEO) and a description of your intended analysis.** Your dataset and analysis must be approved.

**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.**

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

2. Retrieve the expression and phenotype data

3. Pre-process the data as appropriate (remove unwanted samples, make sure the order of samples in the clinical data matches the order of samples in the expression data).

4. Process the expression data by removing probes with low expression and using TMM normalization. If your data is on the scale x = log2 (count + 1), then you will need to get the count data back by calculating $2^x$ – 1.

5. Generate a boxplot of the first 10 samples of your data, to show that the data using `boxplot(logCPM[,1:10])`, to show that your data have been normalized. Your boxplot should have a meaningful title (using the `main` argument) and y-axis label (using the `ylab` argument).

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

7. Extract the column that contains the categories 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.

8. Using *limma*, find the probes that are differentially expressed across the groups you are comparing, using a false discovery rate (FDR) of 10%. Output the number of probes identified, using the *nrow* function. Note: it is possible for this number to be 0. For some datasets, particularly when the sample size is small, you may end up with an FDR of 100% for all probes!)

9. ***Only complete this step if you have less than 30 probes in your result from (8)***. If you have found less than 30 probes with an FDR of 10% in the previous step, then find the top 30 probes.

10. For the top probe (with the lowest adjusted p-value), construct a boxplot (using *ggplot*) to compare the expression of that probe across 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.

11. Find the gene names corresponding to all probes from question #8 (or 9 if this was completed) and create a data frame with the following columns: gene names, probe names, logFC, adjusted p-values. Your data frame should

not contain any other columns. Output the first 5 rows to display the top 5 probes.

12. Using DAVID, identify Gene Ontology (GO) terms and KEGG pathways that are associated with the differentially expressed genes identified in (11). A file containing the gene names and 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 name of the dataset you've analyzed, with a link, 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 with an FDR of 10%, the names of the top 3 genes, and the top GO terms or pathways associated with the phenotype that you analyzed.