# CIS4930 Bioinformatics - Assignment 2

# Directions: Following the steps below, analyze the GEO RNA-seq data you selected in Assignment 1. Save each image/table you create and write a short summary (3-5 sentences) of how they were made and what you discovered. Create an assignment document with your team name and GitHub repository at the top followed by these results and summaries, to be submitted on Canvas.

*You will submit a document (doc/docx/pdf) that lists the names of everyone on your team, your team’s GitHub repository with all code used in the assignment, and each image/table plus their short summaries. These writeups will be used to create your final project report, so it is to your advantage to do a thorough job now.*

1. *Download the expression data and matching metadata from GEO that you selected in Assignment 1.*
   1. *You should have either a matrix of samples by genes, or a set of counts files*
   2. *If your matrix has Ensembl IDs (e.g.* [*ENSG00000141510*](http://useast.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000141510;r=17:7661779-7687538)*) instead of Hugo gene names (e.g. TP53), convert the names following these directions:* 
      1. [*alexslemonade.github.io/refinebio-examples/03-rnaseq/gene-id-annotation\_rnaseq\_01\_ensembl.html*](https://alexslemonade.github.io/refinebio-examples/03-rnaseq/gene-id-annotation_rnaseq_01_ensembl.html)
      2. *Also helpful:* [*bioconductor.org/help/course-materials/2019/BSS2019/05\_Annotations.html - org.hs.eg.db*](https://www.bioconductor.org/help/course-materials/2019/BSS2019/05_Annotations.html#org.hs.eg.db)
   3. *Load the data into R. What size is your expression matrix? How many genes does it include? How much variation do you see in the data? Calculate per-gene expression ranges and generate a density plot showing those results. Summarize your findings.*

The size of the expression data is 12 experiments compared with 47313 data sets. Within the data-sets, there is a lot of variation in terms of values. Per-Gene expression ranges were made with turning it into data frames. The density plot is shown in densityPlot.pdf and the results that have been found are that there is a lot of variation in the density between 1 and 1.5 \* e^-7. Metadata is available in Metadata.txt.

1. *Now that you have loaded the expression data into R, generate a PCA plot:* 
   1. *If you have counts file(s), follow these* [*DESeq2 directions*](http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#count-matrix-input) *to generate an expression matrix.*
   2. *Use the DESeq2 function plotPCA() to generate your plot (see* [*here*](https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html)*)*
   3. *Color your plot by the 2 groups you identified in assignment 1 (e.g., cancer vs normal)*
   4. *Make sure you include a legend and label the axes!*
   5. ***If you have 3 or 4 students in your group****, also generate either t-SNE or UMAP plot, and summarize the differences and similarities between your two plots.*
      1. *t-SNE (*[*example here*](https://www.r-bloggers.com/2019/05/quick-and-easy-t-sne-analysis-in-r/)*)*
      2. *UMAP (*[*example here*](https://cran.r-project.org/web/packages/umap/vignettes/umap.html)*)*
   6. *Save your plot(s) and summarize your findings.*
2. *Perform differential analysis on the samples from your two groups, following the directions below*
   1. [*alexslemonade.github.io/refinebio-examples/03-rnaseq/differential-expression\_rnaseq\_01.html*](https://alexslemonade.github.io/refinebio-examples/03-rnaseq/differential-expression_rnaseq_01.html)
   2. *Create a volcano plot of your data, following the directions above*
   3. *Create a table of differentially expressed genes.*
   4. *Save and summarize your findings.*
3. *Extract the list of significantly differentially expressed genes, and generate a heatmap using* [*ComplexHeatmap*](file:///C:\Users\arunj\Downloads\ComplexHeatmap)
   1. *Package reference (*[*https://jokergoo.github.io/ComplexHeatmap-reference/book/*](https://jokergoo.github.io/ComplexHeatmap-reference/book/)*)*
   2. *Add a side bar colored by sample groupings (cancer vs not, etc)*
4. *Extract the list of differentially expressed genes and run enrichment analysis.* ***Each student in your team should run a different method OR ontology*** *(e.g., if there are 4 students on the team, there should be results for 4 applications in your assignment writeup).*
   1. *Choose a method:*
      1. [*clustProfiler*](http://bioconductor.org/packages/release/bioc/vignettes/clusterProfiler/inst/doc/clusterProfiler.html)
      2. [*gProfiler2*](https://cran.r-project.org/web/packages/gprofiler2/vignettes/gprofiler2.html)
      3. [*GenomicSuperSignature*](http://bioconductor.org/packages/release/bioc/html/GenomicSuperSignature.html)
   2. *Choose an ontology (e.g. Disease Ontology, Gene Ontology)*
   3. *Run enrichment analysis on your data using your selected method and ontology*
5. *Create a table of the enriched processes found for each method in step 4 (one table per method). Create a table showing statistically significantly enriched terms and any characteristics shared by the method you used (e.g., q-value, p-value, log fold change)*
6. *Write a short summary to go with each plot/table you create. Describe what you did, what parameters you used (if any) and an interesting result from it.*
7. *Combine all results into a single file, submit on Canvas. Make sure that all of your code is added to your github repository.*