# CGS4144 Bioinformatics - Assignment 3

# Directions: Following the directions below, analyze your data. Save the images and write a short summary of how they were made and what you discovered. Create a work document with your team’s name at the top followed by your results, to be submitted on Canvas.

1. *Load into R the expression data and matching metadata from GEO that you processed in Assignment 2.*
2. *Unsupervised Analysis*
   1. *Subset your data to the 5,000 most variable genes*
   2. *Using that subset of the data, select and run a clustering algorithm from this list:*
      1. *K-means*
      2. *Hierarchical clustering (hclust)*
      3. [*ConsensusClusterPlus*](https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html)
      4. [*PAM clustering*](https://cran.r-project.org/web/packages/cluster/cluster.pdf)
      5. [*Gaussian mixture models*](https://cran.r-project.org/web/packages/mclust/index.html)
   3. ***Each student in your team should run a different clustering method*** *(e.g., if there are 4 students on the team, there should be results for 4 clustering methods in your assignment writeup).*
   4. *How many clusters did each method find?* 
      1. *If you ran a method that requires you to select the number of clusters (k), how did changing this value change the results? Compare cluster membership at each k to investigate this.*
      2. *If you ran a method that selects k for you, describe why it chose that k.*
   5. *Rerun each clustering method using different numbers of genes. Try 10, 100, 1000, and 10000 genes.* 
      1. *How did the number of genes affect clustering?*
      2. *Create an alluvial diagram (Sankey plot) to visualize this. (*[*https://cran.r-project.org/web/packages/ggalluvial/vignettes/ggalluvial.html*](https://cran.r-project.org/web/packages/ggalluvial/vignettes/ggalluvial.html)*)*
3. *Heatmaps and Dendrograms*
   1. *Create a heatmap of the 5,000 genes used in clustering, with an annotation sidebar showing the clusters you identified from each clustering method and the sample groups from Assignment 1.*
      1. *Either create a different heatmap for each method or include all in the same heatmap.*
      2. *Make sure to include a legend and axis labels*
      3. *Include row and column dendrograms*
4. *Statistics*
   1. *Does cluster membership correlate with the groups you chose in Assignment 1? Perform a chi-squared test of independence to statistically compare the two. Do this for each clustering method you used.*
   2. *Compare your different clustering results. Run a chi-squared test comparing each pair.*
   3. *Adjust all statistical test results for multiple hypothesis testing (p.adjust).*
   4. *Create a table with statistical test results that includes adjusted and un-adjusted p-values.*
   5. *Create an enrichment plot showing your statistical test results between all metadata and clustering results (*[*https://www.r-bloggers.com/2011/03/five-ways-to-visualize-your-pairwise-comparisons/*](https://www.r-bloggers.com/2011/03/five-ways-to-visualize-your-pairwise-comparisons/) *and* [*https://ggobi.github.io/ggally/articles/ggpairs.html*](https://ggobi.github.io/ggally/articles/ggpairs.html)*). Use any R package to do this.*
5. *Write a short summary for each plot/table you created in this assignment. In 3-5 sentences for each, describe what you did, what parameters you used (if any) and an interesting result from it.*
6. *Combine all results into a single file, submit on Canvas. Make sure that all your code is added to your GitHub repository.*