**KMG060 – Systems Biology – Fall 2022**

**Exercise 1 Report**

Work in your assigned group for this exercise. Do not “team up” with another pair of students. Each question will be graded with one point if the correct answer is provided. This report should be uploaded to Canvas by **Sunday 18September at 23:59**.

**Date:**

**Names (Group Number):**

**Condition:**

1. Download the required data and scripts from: <https://github.com/SysBioChalmers/KMG060-Systems-Biology-course>

Take a look at your RNA-seq counts dataset:

1. How many genes are represented in your dataset?
2. How many replicates per experimental condition were obtained in the study?
3. How many genes were actually measured by RNA-seq (non-zero reads)?
4. Visualize your read counts distributions using boxplots for the transformed log2 values (before and after normalization). What can you say about your dataset in terms of spanning (orders of magnitude), median expression values? Are there any evident effects on the dataset with the proposed normalization method? If so, try to explain them.
5. Show your PCA results (PC1 vs PC2, displaying the percentage of variance for each principal component in their respective axis labels).
   1. Explain what the PCA results show.
   2. Are there any outlier samples in the dataset for any of the analyzed conditions?
6. Show the number of differentially expressed genes between the reference condition and the assigned stress condition. Use absolute log2 fold-change = 2 and 0.01 corrected p-value as your threshold parameters.
   1. Plot your results in a volcano plot, highlighting with a different color the differentially expressed genes. Explain in some lines how a volcano plot can be interpreted.
   2. How many down and up regulated genes were obtained?
   3. How are the results from these plots similar to or different from the results from the PCA plot?
7. Explore your DE analysis results and combine them with gene descriptions information available in the exercise data subfolder (see the MATLAB script).
8. Provide the gene identifiers and functions for your top 10 differentially expressed genes.
9. Can you infer any interesting/meaningful biological pattern from these top DE genes?
10. GO terms analysis.
11. Provide the associated GO Terms and descriptions for the top DE gene for your stress condition.
12. What can you learn from looking up the associated GO Terms for the top 10 DE genes? Compare the knowledge you obtained with the summary paragraph on the *S. cerevisiae* Genome Database.
13. Find enriched GO terms in differentially expressed genes subset.
14. Take the first associated GO term for the top DE gene in the dataset, provide the number of DE genes that this GO term is also associated with.
15. Provide the number of total genes (DE and non-DE) that this GO term is associated with. Run a hypergeometric test in order to assess if the enrichment of this GO term in DE genes is statistically significant.
16. Provide a list of GO term IDs and descriptions for all the significantly enriched GO terms (adjusted p-Value <=0.01) in your DE genes. (Hint: if you do not see any results, perhaps consider adjusting the p-Value threshold)