**KMG060 – Systems Biology – Fall 2021**

**Exercise 1 Report**

Work in pairs 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 19September at 23:59**.

**Date**:

**Names**:

**Selected stress 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:
2. How many genes are represented in your dataset?
3. How many biological replicates per experimental conditions were obtained in this study?
4. How many genes were actually measured by RNA-seq (non-zero reads)?
5. Normalize your read counts dataset, by first dividing the values by the mean expression level of each gene across samples, and then dividing each library (column) by the median of its previously normalized values.
6. 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.
7. Show your PCA results (PC1 vs PC2, displaying the percentage of variance for each 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?
8. Show the number of differentially expressed genes reference and a stress condition of your choice (High temperature, low pH, osmotic pressure). 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?
9. Explore your DE analysis results and combine them with gene descriptions information available in the exercise data subfolder (see the MATLAB script).
10. Provide the gene identifiers, names and functions for your top 10 differentially expressed genes.
11. Can you infer any interesting/meaningful biological pattern from these top DE genes?
12. GO terms analysis.
13. Provide the associated GO term IDs and descriptions for the top DE gene for your stress condition.
14. What can you learn from it? Compare the knowledge you got with the summary paragraph on the *S. cerevisiae* Genome Database.
15. Find enriched GO terms in differentially expressed genes subset.
16. 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.
17. provide the number of total genes (DE and non-DE) that this GO term is associated with. Run a hypergeometric test in order to asses if the enrichment of this GO term in DE genes is statiscally significant.
18. 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.
19. **Bonus track (optional, value 1.5 points):** As you may have realized, DE analysis together with GO terms enrichment analysis is a powerful tool for finding biological phenotype differences across different environmental conditions, however it is usually argued that results depend largely on the selected differential expression thresholds.

Try to repeat point #4 in this exercise for different p-value and log2 fold-change threshold values (see the MATLAB script suggestion). What can you say about this, in terms of number of obtained DE genes? Which of the two parameters is affecting your results the most?