**Differential Gene Expression in Acute Myocardial Infraction**

# Introduction

# Gene expression describes the process in which genes that are coded in the DNA of living organisms are transcribed into mRNA. This is part of the bigger process in which genes are being copied (transcribed), processed, translated and modified into the final product, usually a protein. Gene expression profiling measures the levels at which mRNA molecules pertaining to the genes profiled are observed in a sample.

# In this exercise, we will perform analysis over a data set that compares expression profiles of circulating endothelial cells (CECs) in patients with acute myocardial infraction to CECs in healthy controls.

# The Data Set

# The data set was taken from:

1. Dataset record in NCBI:

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66360>

1. Published paper: Muse et al, Sci Rep 2017 <https://www.nature.com/articles/s41598-017-12166-0>

# We extracted the data matrix and provide it as a separate csv attachment. The csv file needs to be pre-processed before moving to the main analysis steps. Some information should be removed but make sure that you keep all information important for the analysis.

# The paper describes a study that seeks to develop an expression based signature that can detect AMI in patients in a non-invasive manner, by profiling CECs.

# Analysis

# Describing the data

1. How many genes profiled?
2. How many samples (patients) in total?
3. How many samples in each class?
4. If there are missing values, then remove the entire row (gene) from the data matrix.  
   How many rows left now?

# WRS for differential expression (DE)

1. Consider some gene, g. Under the null model (which assumes that for g there is no M vs H DE), what is the expected sum of ranks of g’s expression levels measured for samples labeled M?
2. Denote this sum of ranks by RS(g). What is the minimal value, m, that RS(g) can take?
3. Under the null model, what is the probability of RS(g) = m? (provide a formula for this and explain it)
4. Under the null model, what is the probability of RS(g) = m+1? (provide a formula for this and explain it)
5. Under the null model, what is the probability of RS(g) = m+2? (provide a formula for this and explain it)
6. Draw a histogram of the values of RS(g) when g ranges over all genes in the data (after the clean-up)

# Differential Expression

# Determine the statistical significance of differential expression (DE) observed for each gene in H vs M. Evaluate the DE in both one-sided directions. That is, report genes overexpressed in M vs H and separately genes underexpressed in M vs H.

# Use a Student t-test and a WRS test for this analysis.

# 

# Correlations Select 60 genes from each one of the DE lists you computed in 3.c, using WRS. Generate a set of 120 genes, D, which is the union of the above two sets. Compute Spearman rho correlations in all pairs within D.

1. What can you report about co-expression of genes in D (co-expression is the correlation of the expression levels of genes)? Do we observe any significant co-expression? If so how many pairs, etc.
2. What would have been advantages and disadvantages of computing co-expression for all genes in the study rather than only for genes in D?
3. Perform the above steps but restrict attention only to samples labeled H. What do you see now? Can you explain this?

# Plots and Conclusions

1. Construct the DE overabundance plots (blue and green lines as shown in class) for the DE comparisons you performed (Section 3.c).

State for each comparison, the number of genes, k, at which we observe:

1. FDR = 0.1
2. FDR = 0.05
3. FDR = 0.001

If these events are not observed at any k, then make that statement.

1. What can you say about the difference in results obtained in WRS vs those obtained by Student t-test?
2. Select 3 differentially expressed genes, from D (from 3.d), and produce a graphical representation of their expression patterns that demonstrates the observed DE.
3. Heatmap  
   Draw a heatmap representation of the expression values of the genes in D (from 3.d), across the entire cohort.