Ben Galili,

Dr Leon Anavy,

Prof Zohar Yakhini

HW4 – statistics and data analysis.

Differential Gene Expression in Acute Myocardial Infraction

1. 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 guided analysis, comparing expression profiles of circulating endothelial cells (CECs) in patients with acute myocardial infraction to CECs in healthy controls. A comparison of two sample classes.

2. 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
2) 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 (link to download). 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 that is important for the analysis. Specifically, all expression values should be kept and the label of each sample (H – Healthy, M - Myocardial Infraction).

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.

3. Analysis

a. High level description of the data and some pre-processing

1) How many genes profiled?
2) How many samples (subjects/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?
5) Pick 20 genes at random. Draw 20 pair boxplots in one figure comparing expression levels of each of these genes in the two classes M and H.

b. 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 maximal value, c, that RS(g) can take?
3) Under the null model, what is the probability of RS(g) = c? (Provide a formula for this and explain it)
4) Under the null model, what is the probability of RS(g) = c-1? what is the probability of RS(g) = c-2?
   (Provide formulas and explain them)
5) Draw a histogram of the values of RS(g) in the dataset. Here g ranges over all genes in the data (after the clean-up). Compute the IQR for this distribution and present it on the plot with the histogram.

c. **Differential Expression**

The purpose is to determine the statistical significance of differential expression (DE) observed for each gene in H vs M. Evaluate the DE in both one-sided directions for every gene, using both Student t-test and WRS test.

Report the number of genes overexpressed in M vs H (M > H) at a p-value better ($\leq$) than 0.07 and separately genes underexpressed in M vs H (M < H) at a p-value better than 0.07. For both directions use both a Student t-test and a WRS test.

d. **Correlations**

Select the 80 most significant genes from each one of the one-sided WRS DE lists you computed in 3c. Generate a set of 160 genes, D, which is the union of the above two sets.

1) Compute Kendall τ correlations in all pairs within D (160 choose 2 numbers). Represent the correlation matrix as a 160x160 heatmap.

2) Under a NULL model that assumes that genes are pairwise independent, what is the expected value for τ?

3) Now compute the Kendall τ correlations in all 80 choose 2 pairs from the overexpressed genes in D. Present your results on a histogram. What is the average value you observed? Compare it to the value in the above section. Explain.

4) What can you report about co-expression of genes in D (co-expression is inferred from the correlation of the expression levels of genes, across a set of samples)?

5) What can you say about how many co-expressed pairs we would observe (in the entire dataset) at FDR=0.05? Explain your answer.

6) What would have been advantages and disadvantages of computing co-expression for all genes in the study rather than only for genes in D?

e. Plots and Conclusions of the DE and correlation analysis

1) Construct the DE overabundance plots (blue and green lines as shown in class) for M vs H overexpression (higher expression levels in M) using WRS and t-test using the results you had computed in Section 3c.

   State, for each comparison, the number of genes, k, at which we observe with an FDR threshold of:

   a) $\tau = 0.05$

   b) $\tau = 0.01$

   c) $\tau = 0.005$

   If these events are not observed at any k>0, then make that statement.

2) For any given gene, g, consider the following set of p-values:

$$S(g) = \left\{ p(\lambda) \middle| \begin{array}{c} p(\lambda)\ is\ a\ WRS\ p-value \\ for\ overexpression\ in\ M \\ after\ swapping\ one\ label \end{array} \right\}$$

   (In the above definition λ represents a label swap – there are 99 such swaps).
   Let

$$p_U(g) = \max_{\lambda} S(g)$$

   In words: $p_U(g)$ represents the maximum p-value that could have been obtained for g, assuming a single labeling error.

   a) For all genes, compute $p_U(g)$.

   b) Run the FDR procedure with $\tau = 0.05$ using $p_U(g)$ (instead of the original p-values as computed in section 3e1a above). How many genes can you report now?

   c) What is the intersection of the genes you can report above with the results of 3e1a. Explain.

   Comment: The process you ran in this section yields a set of genes called Robust Differentially Expressed Genes (RDEG).

3) Select any 3 differentially expressed genes, from D (which was defined in 3d), and produce a graphical representation of their expression patterns that demonstrates the observed DE.

4) Heatmap
Draw a heatmap representation of the expression values of the genes in D (from 3d), across the entire cohort (all samples). Order the genes and the samples to produce the maximal visual effect.