Statistics and Data Analysis - HW4

Gal Vizner - ID:201247681
Ronen Eytan - ID:036183879

a. Describing the data

1. There are 54675 genes overall
2. There are 99 patients in total
3. The samples in each class are:
49 samples for M (acute myocardial infarction)
50 samples for H (healthy)
4. After removing 47 genes we are left with 54628

b. WRS for differential expression (DE)

1. We have 99 patients. Since the ranks go from 1 to 99 and, under the null model, ranks have no preference between M and H, the expected rank value for M is the average rank:

$$E(X) = \frac{1}{2}(99 + 1) = 50$$

Considering the fact that the number of M samples are 49 we get:
50 * 49 = 2450

2. The minimal value for RS(g) is as if we "arrange" the M genes to be the lowest on the list - meaning they'll have the smallest values,
1 + 2 + 3 +...+ 49 = 1225.

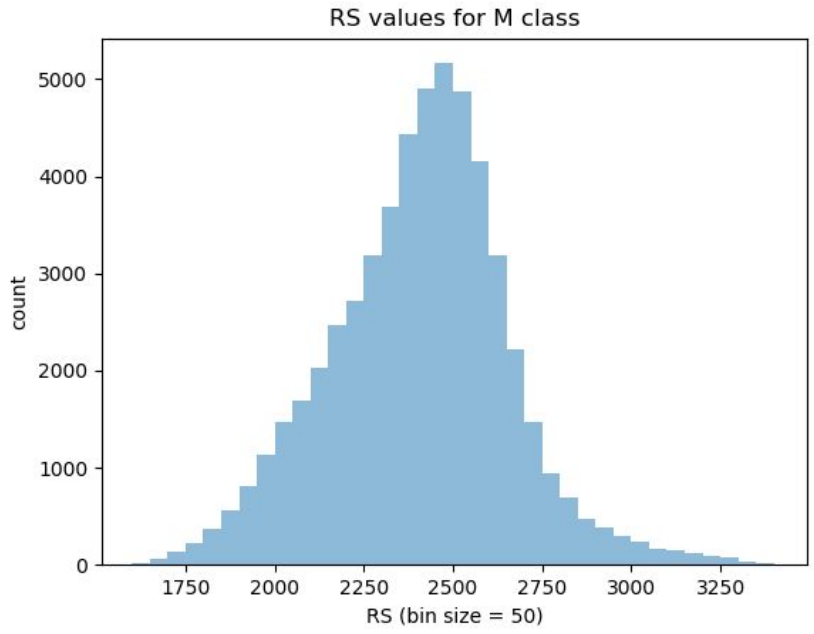
3. Since under the null model all M vs H rank orders are equiprobable and only one ordering gives us m the probability for RS(g) = m is $1/\binom{99}{49}$ = 1.98233060428366781347534993937 67e-29

4. Similarly to the above, and again we have only a single M vs H ordering that gives m+1. We get this ordering by switching rank 49 to H and rank 50 to M. Same probability as above.
Note: Starting with the RS(g) = m ordering, our only option for changing RS(g) is to switch ranks between H and M. Any other switch will increase

RS(g) by more than 1 and there's no option to reduce. So indeed there is only one ordering.

5. Again similarly to the above, we need to find the number of M vs H rank orderings that will give us RS(g) = m + 2. This time we have 2 orderings where the M ranks sum to m + 2.
#1: switch between M49 and H50 (+1). Then, switch between M48 and H49 (+1).
#2: switch between M48 and H50 (+2).
Since we have 2 orderings vs one above, the probability is twice the value above $2/\binom{99}{49}$

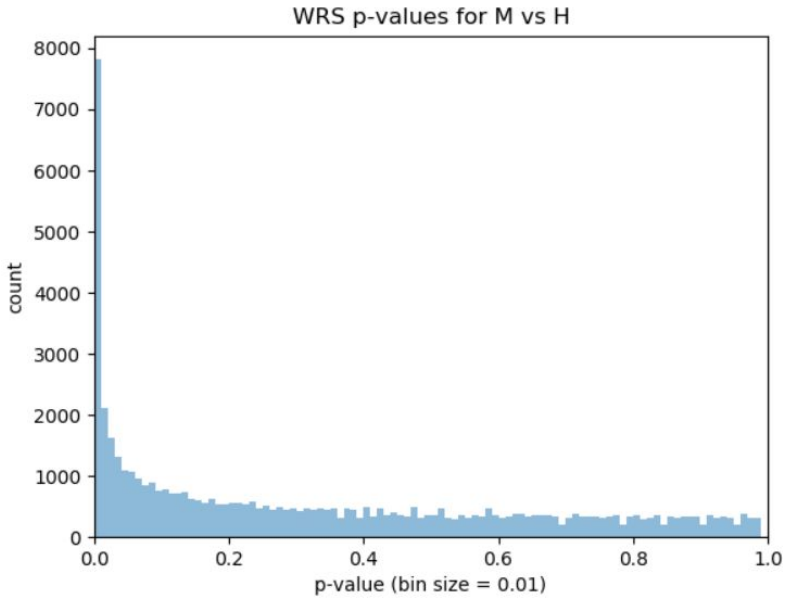
6. Histogram of the values of RS(g):

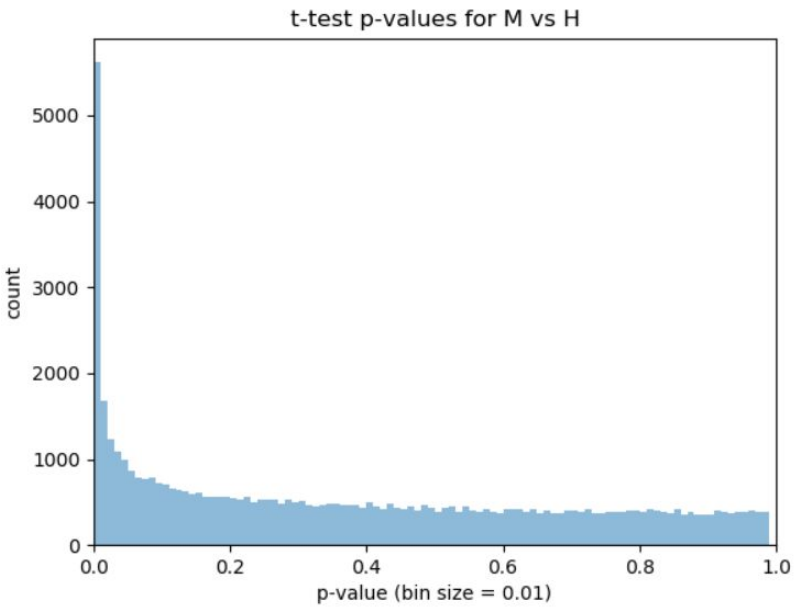


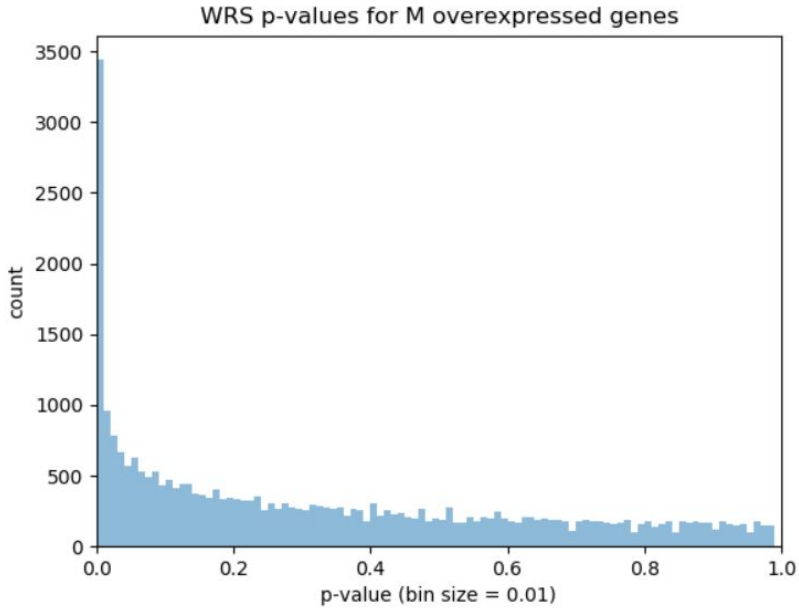
c. Differential expression

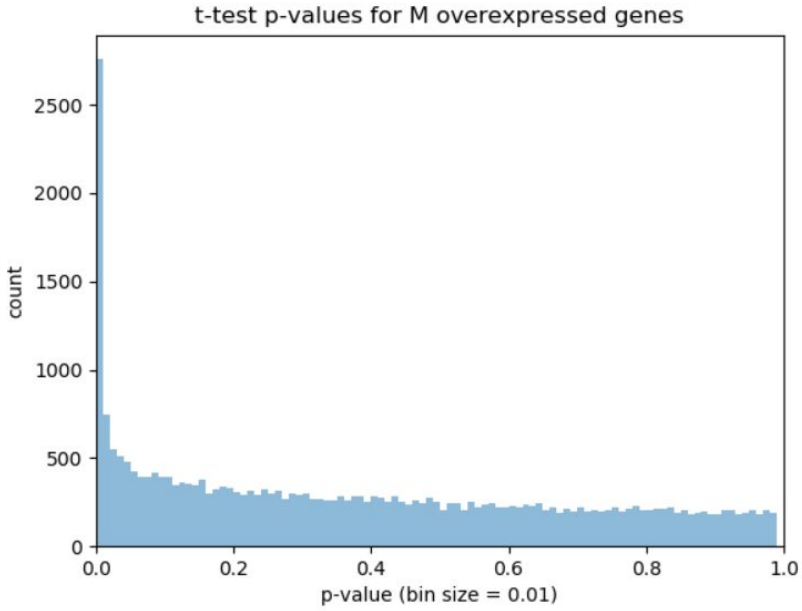
Note: Due to the large number of genes our report will not include names of genes and overexpression p-value levels. Instead we give a quantitative report. And since DE wasn't defined by a minimal required p-value (threshold), we include results data of all p-value, the lower the p-value the more significant is the DE (the null model where the gene is arbitrarily expressed in the 2 classes is less likely).

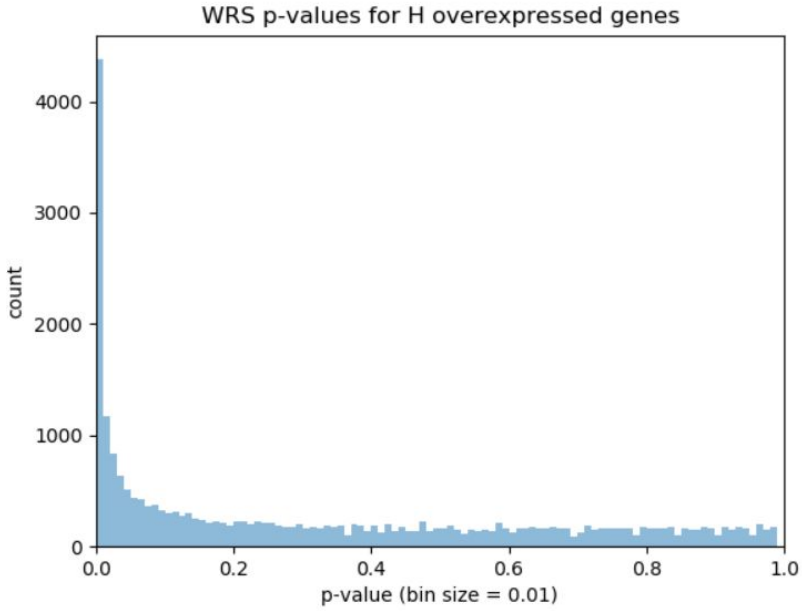
The following histograms show the p-value distribution. First, the general p-value distribution (WRS, t-test) and then the p-value distribution for genes overexpressed in class M followed by the same for genes overexpressed in class H (underexpressed in class M).

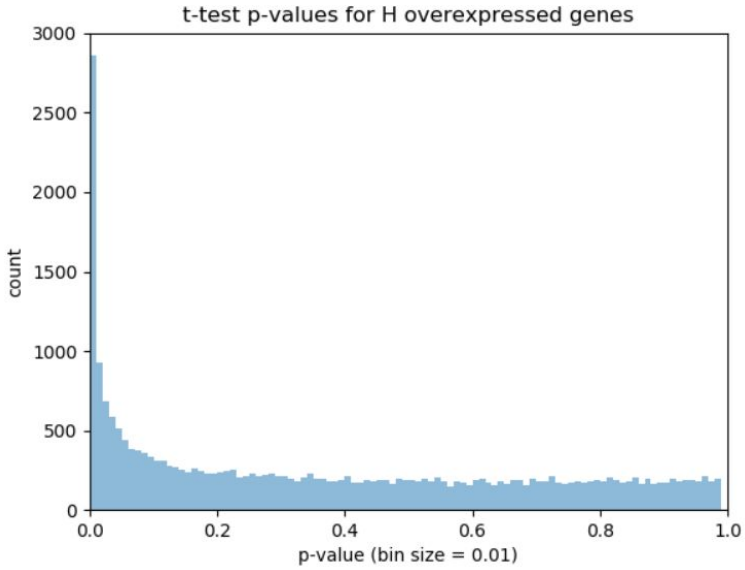



t-test p-values for M vs H
count
p-value (bin size = 0.01)
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0


WRS p-values for M overexpressed genes
count
p-value (bin size = 0.01)
0
500
1000
1500
2000
2500
3000
3500
0.0
0.2
0.4
0.6
0.8
1.0


t-test p-values for M overexpressed genes
count
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
p-value (bin size = 0.01)

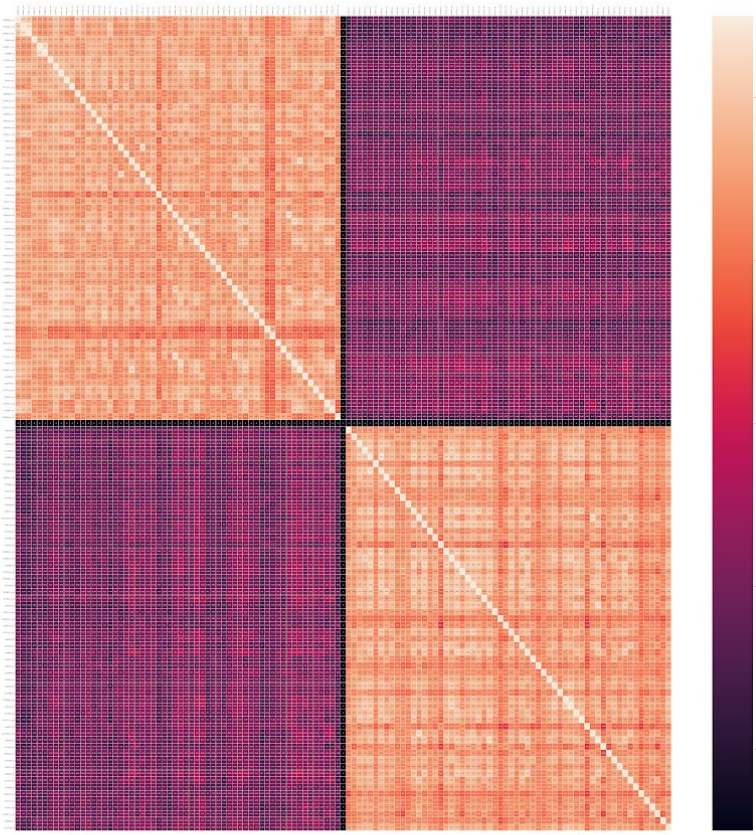

WRS p-values for H overexpressed genes
count
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
p-value (bin size = 0.01)



d. Correlations

The following heatmap (for a closer look, also added in file: heatmapD120.png) shows the spearman correlation values (on each square) in a 120x120 matrix. The matrix rows and columns start (from top to bottom and from left to right) with the 60 genes that are overexpressed in class M and got the lowest p-values. Following those 60 genes are the 60 genes that are overexpressed in class H and got the lowest p-values.
The heatmap shows clearly a relatively positive co-expression between genes that are overexpressed in the same class. And a relatively negative co-expression between genes that are overexpressed in different classes.

*It is apparent that there are more close-to-white areas in the HxH area of the matrix (bottom right) compared to the MxM area of the matrix (top left).



d.1.

Since genes that are overexpressed in class M have higher expression values in M class and lower expression values in H class and the opposite in genes that are overexpressed in class H, it is expected to see low co-expression between pairs of genes from both groups. On the other hand and due to the same reason, in pairs of genes where both are overexpressed in the same class it is expected to see relatively high co-expression values.

What is “significant” is unclear and may change in different scenarios. Once a spearman rho value threshold is given, the number of significantly co-expressed pairs can be found by simply counting the number of values equal or above the threshold under/above the diagonal line in the heatmap.

For example, let’s say that the threshold is 0.9. There are 12 values equal or greater to 0.9 under the diagonal line in the heatmap. So we can conclude that in this case there are 12 significantly co-expressed genes.

Alternatively, we could run the following code to find the number of pairs: If we assume that significant co expression is when we get spearman rho >= 0.9. After calculating spearman rho on each one of the 120 pairs the code returns 12 as the number of pairs with significant co expression.

```
all_120 = m_60 + h_60
spearman_correlations = []
for i in range(0, 120):
  for j in range(i + 1, 120):
    rho, pvalue = spearmanr(all_120[i], all_120[j])
    spearman_correlations.append({'i': i, 'j': j, 'rho': rho, 'pvalue': pvalue})

co_expressions = []
for i in range(0, len(spearman_correlations)):
  if (spearman_correlations[i]['rho']  >= 0.9):
    co_expressions.append(spearman_correlations[i])

print(len(co_expressions))
```

d.2.

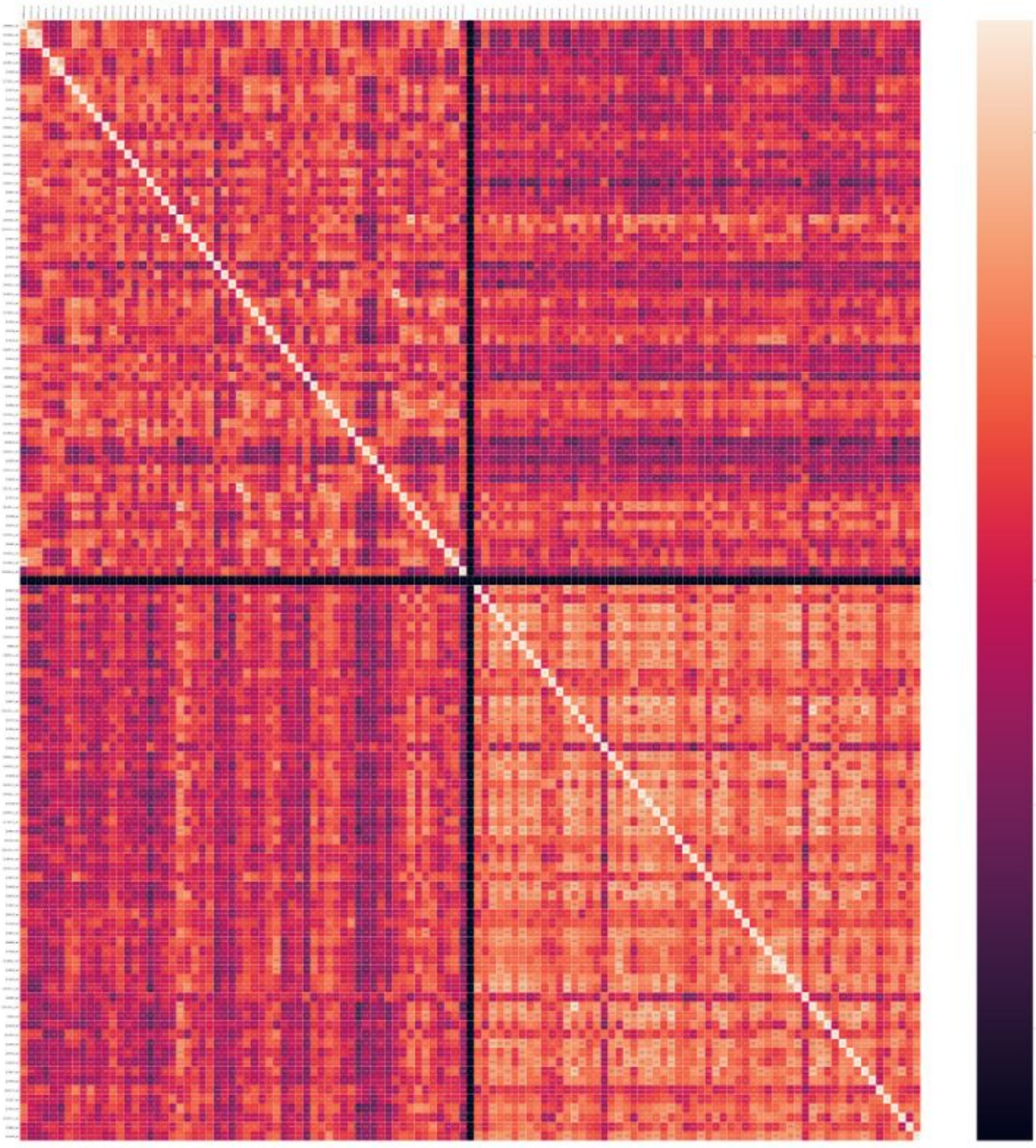
If you are looking for just a few random co-expressed pairs of genes, it may save you time to start with pairs of genes overexpressed in the same class as explained above. But if you are looking for all the co-expressed pairs or the the most co-expressed pair, you will have to go through all the gene pairs in the study. The reason is that co-expressed pairs of genes may not be differentially expressed in M vs H. For example, 2 different genes with the same constant expression value or a pair of genes with expression

values [a, 2a,..,49a] in class M samples and [b, 2b, 3b,...,50b] in class H samples (each gene may have a different a,b values).

The main advantage of the full computation is that we will not miss any important data because we will compute the co expression on all the data we have, the main disadvantage is that it will take a lot of time computing it, and we will need more cpu power (generally depending on the amount of data of course) while in some cases we could have just looked in D.

d.3

After computing the co expression again this time only for the samples labeled H. As can be seen in the heatmap below (for a closer look, also added in file: h_only_heatmapD120.png), the spearman values dropped in pairs of genes overexpressed in the same class and increased in pairs of genes overexpressed in different classes. This change took away the expected co-expression advantage discussed above (d.1) in pairs of genes overexpressed in the same class and the expected co-expression disadvantage of pairs of genes overexpressed in different classes.
Again assuming 0.9 threshold, we found that there are now only 3 pairs with significant co-expression.

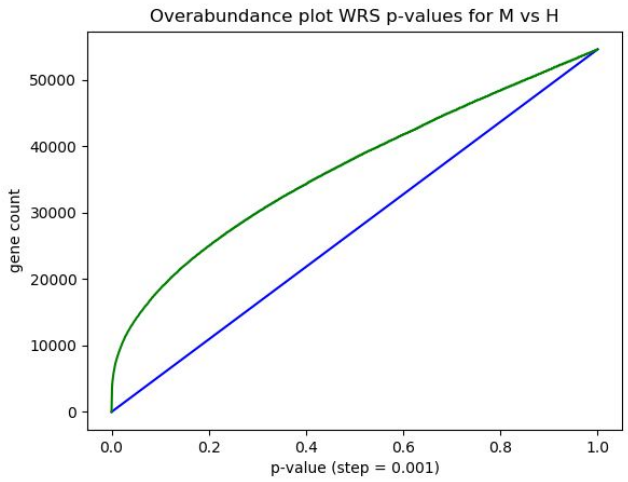


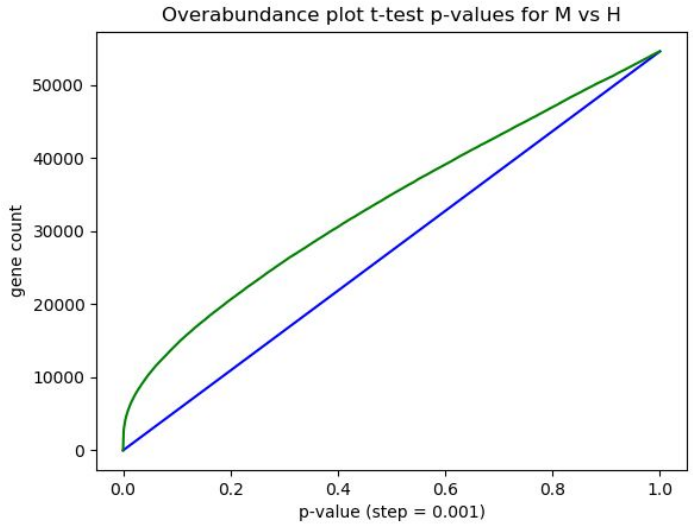
e. Plots and conclusion

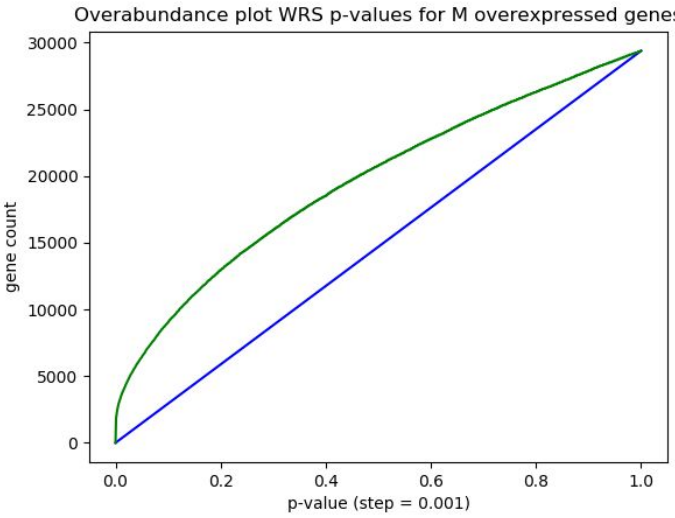
e.1.

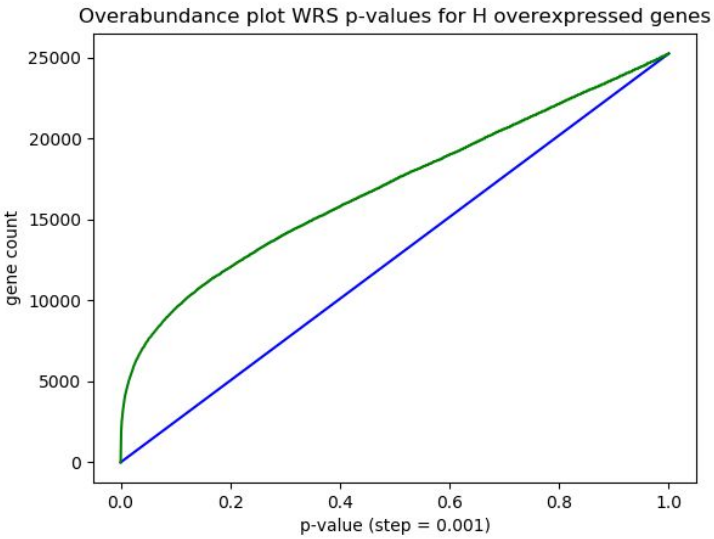
We want to plot 2 lines. Blue line: per p-value, the number of spurious genes we expect (expected distribution). Green line: per p-value, the actual number of genes we found having not greater p-value (actual distribution).

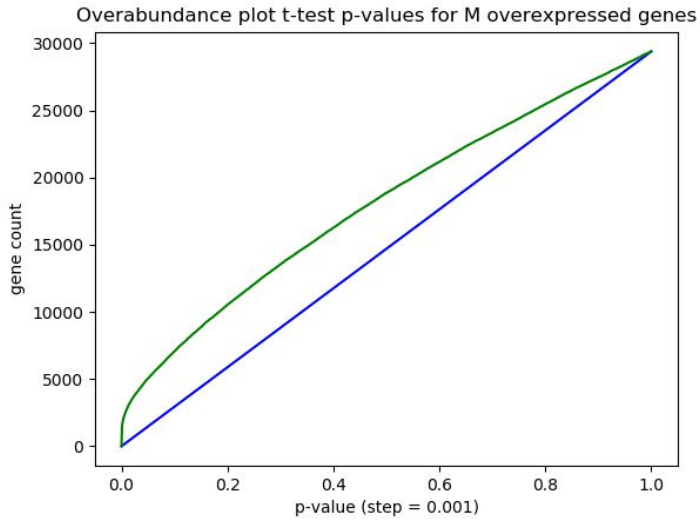
We will calculate a blue and green dot for every 0.001 progress of the p-value from 0 to 100 and plot the lines that pass through those dots.

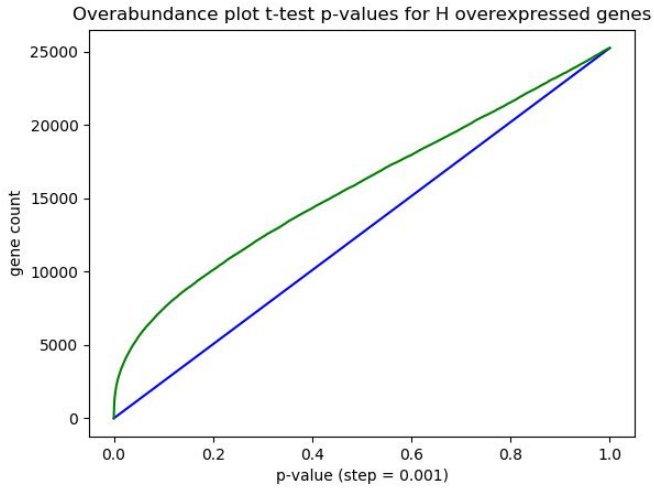





Overabundance plot WRS p-values for M overexpressed genes
gene count
p-value (step = 0.001)
0
5000
10000
15000
20000
25000
30000
0.0
0.2
0.4
0.6
0.8
1.0


Overabundance plot WRS p-values for H overexpressed genes
gene count
p-value (step = 0.001)
0
5000
10000
15000
20000
25000
0.0
0.2
0.4
0.6
0.8
1.0





*As shown in class, we will look for the max k with the required (robust) FDR.

<u>General WRS p-values for M vs H (all differentially expressed included):</u>
For FDR = 0.1, maximal k = 9477
For FDR = 0.05, maximal k = 6678
For FDR = 0.001, maximal k = 1311

<u>General t-test p-values for M vs H (all differentially expressed included):</u>
For FDR = 0.1, maximal k = 5700

For FDR = 0.05, maximal k = 3970
For FDR = 0.001, maximal k = 963

WRS p-values for M overexpressed genes:
For FDR = 0.1, maximal k = 3723
For FDR = 0.05, maximal k = 2706
For FDR = 0.001, maximal k = 883

t-test p-values for M overexpressed genes:
For FDR = 0.1, maximal k = 2691
For FDR = 0.05, maximal k = 2041
For FDR = 0.001, maximal k = 756

WRS p-values for H overexpressed genes (M underexpressed):
For FDR = 0.1, maximal k = 5906
For FDR = 0.05, maximal k = 4129
For FDR = 0.001, maximal k = 378

t-test p-values for H overexpressed genes:
For FDR = 0.1, maximal k = 3095
For FDR = 0.05, maximal k = 1959
For FDR = 0.001, maximal k = 155

e.2

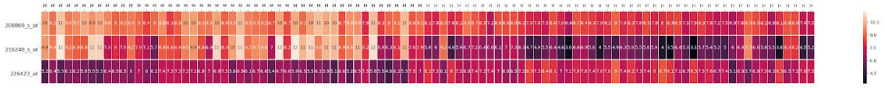
The results indicate lower (more significant) p-values in WRS compared to t-test. As we know, t-test is a a-parametric test while WRS is parametric. The lower p-values in t-test compared to WRS indicate relatively close mean value of the empirical values (expression levels) and similar variance between the genes in the study.

e.3.

Selected genes:

Overexpressed in M (most differentially expressed):
208869_s_at WRS p-value = 1.4553617006745165e-14
216248_s_at WRS p-value = 1.8106915682495881e-14

Underexpressed in M (most differentially expressed):
226423_at WRS p-value = 7.6647822285559772e-10
Following is a heatmap of the results (for a closer look, also added in file: de3.png):



In the first 2 rows the heatmap clearly shows brighter(higher expression) colors on the left (M class) and darker(lower expression) colors on the right (H class) - demonstrating the over expression of M. And the other way around in the 3rd row.
*Exact values also included in the heatmap

e.4.
The heatmap (for a closer look, also added in file: de_exp_120.png):