Report assignment 10
Breast cancer gene expression analysis

Matteo Bianco, Filippo Grobbo
300781, 305723

January 10, 2023

Abstract

In this report we analyze a gene expression dataset with two phenotypes: breast cancer and normal. Two analysis are carried out: over-representation enrichment analysis and gene set enrichment analysis (GSEA). Results show differentially expressed genes between tumor and healthy patients are maily associated to cellular processes. Less correlation is found also with some diseases like COVID 19.

1 Dataset and Preprocessing

The purpose of this assignment is to carry a breast cancer gene expression analysis. Looking at Gene Expression Omnibus (GEO) database we found a paper about gene expression of different types of cancer with control group [Stathopoulos GP(2008)]. In order to consider only the problem about breast cancer we discard every patient but the ones having breast cancer or being healthy. Therefore, we get a dataset with 20 samples, of which 9 healthy and 11 with tumor. For each sample we have expression values for 32878 different genes. Gene expression profiles are derived from RNA samples from the blood and are studied using the whole genome microarray platform AB1700 from Applied Biosystems. In order to shrink the dataset and to perform subsequent analysis we decide to keep only the genes with a gene symbol associated on [Biosystems(2006)]. In the end we have a dataset with 20 samples and 17916 features.

The final dataset has no missing values. In order to look for outliers we calculate the z-score of the samples and find that every patient has got a mean absolute z-score less or equal to 1. This implies that there are not any outliers between our samples. Moreover, we discover that there are very few features with more than one observation with z-score greater than 3 or lower than -3. We can visualize the distribution of some features with outliers data in figure 1. Summarizing all these considerations, we decide to keep the whole dataset for subsequent analysis.

2 Statistical and over-representation enrichment analysis

We decide to do an analysis of genes differentially expressed between the breast cancer patients and the control group through a multiple t-test. Without any p-value adjustment we find that there are 2828 genes statistically significant: 1171 of which are overexpressed in people with breast cancer and 1657 underexpressed. We afterwards perform a Benjamini-Hochberg adjustment procedure for multiple t-testing, obtaining 433 genes statistically significant, 92 of which are overexpressend while 341 are underexpressed.

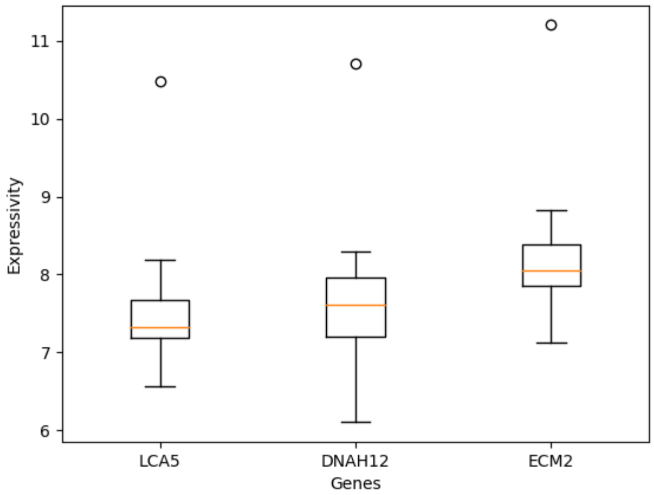

Figure 1: Boxplot distribution of gene expression for 3 genes with outliers

We perform over-representation enrichment analysis using the last obtained list of genes. In order to do this we study the guide of python package gseapy [Fang(2022)] and use the function ”enrich”. With this tool we are comparing the list of differentially expressed genes with some biological pathways to find if there is statistically significant overlapping between them. ”enrich” requires gene symbols in a specific format. This is why we initially kept only genes with a gene symbol available on [Biosystems(2006)]. We conduct three separate analysis on over-expressed genes, under-expressed and both together, as done in [Hong G(2013)]. Actually, as this paper shows:

> ”gene pairs with functional links in pathways tend to have positively correlated expression levels” [...] ”analysing up- and downregulated genes separately is more powerful than analysing all of the [...] genes together”.

We therefore perform the analysis on all genes together just to look for differences with prevoius two analysis.
What the function ”enrich” does is conducting many hyper-geometric tests (one for every pathway) to understand if the presence of our differentially expressed genes in that pathway is just random or statistically significant. In particular, p.values are calculated assuming independence for probability of any gene belonging to any pathway. The formula for the p-value is:

$$\text{p-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where:

- N is the total number of genes in the background distribution (17916 in our case): this is the total number of genes for which we performed a t-test to find differentially expressed genes
- M is the number of genes in the background distribution that are present in a given pathway
- n is the total number of differentially expressed genes (in our case 92 for over-expression, 341 for under-expression and 433 for the whole list)
- k is the number of differentially expressed genes that are present in a given pathway

Than, every p-value is corrected for multiple testing using Benjamini-Hochberg adjustment procedure. Concerning libraries of pathways to use in our analysis, we refer to the guides [Chen EY()], [Kuleshov MV(2016)] and use: BioCarta_2016, BioPlanet_2019, WikiPathway_2021_Human, Elsevier_Pathway_Collection, KEGG_2021_Human. Moreover, since in the paper where we took data from ([Stathopoulos GP(2008)]) the preprocessing was conducted looking at [Thomas lab(2022)], we decide to add 3 more libraries: Panther_2016, GO_Biological_Process_2021 and Reactome_2022.

3 Results

Performing analysis on under and over differentially expressed genes we summarize the results in Figures 2, 3. Looking at the plots we can see on the y axis the pathways and on the x axis the libraries of pathways. Only the top 3 pathways are represented for the most important pathways libraries. Importance is based on the adjusted p-value, which depends on the percentage of overlapping genes. Bigger and darker dots correspond to pathways with high overlapping and low p-values. It is important to observe that colours and dots size scales of the two plots are different. At first sight it could seem more relevant the over-expression of Fluoroacetic acid toxicity WP4966 rather than the under-expression of Translation R-HSA-72766. Actually they are both relevant and this depends if we care more about overlapping genes or lower p-values.
Taking some of the pathways in the y axis and trying to understand their meaning we analyze the most significants:

- **Translation R-HSA-72766**, under-expressed. This pathway deals with protein synthesis through the process of translation of an mRNA sequence into a polypeptide chain. The fact that in patients with tumor we find many genes of this pathway under-expressed follows the intuition about cancer cells anomalies.

- **Coronavirus disease**,under-expressed. SARS-CoV-2 activates the innate immune system which results in systemic inflammatory response syndrome and multiple organ failure. Further study can be conducted about possible correlation between this pathway and breast cancer.

- **Fluoroacetic acid toxicity WP4966**, over-expressed. This pathway controls fluoroacetic acid toxicity describing how it ultimately leads to disturbance of the Krebs cycle. As found in many studies, Krebs cycle is often altered in cancer cells [Karishma Sajnani(2017)]

Concerning the test with all genes togheter, we obtain very similar result to the case with just under-expressed genes. This is beacuse the large majority of differentially expressed genes were under-expressed. Moreover, we saw in [Hong G(2013)] that analyzing all genes together can be less meaningful. This is why we do not report results of this test in details.

4 A different approach: Gene Set Enrichment Analysis (GSEA)

Another way to perform a functional enrichment analysis is through the function ”gsea” of the python package gseapy[Fang(2022)]. Differently from ”enrich”, this function does not just perform a statistical over-representation analysis of differentially expressed genes. Actually, it focuses on cumulative changes in the expression of multiple genes as a group. In particular, it first ranks the genes basing on a t-test measuring each gene’s differential expression with respect to the two phenotypes (tumor versus normal). Then the entire ranked list is used to assess how the genes of each pathway are distributed across the ranked list. To do this, ”gsea” walks down the ranked list of genes, increasing a

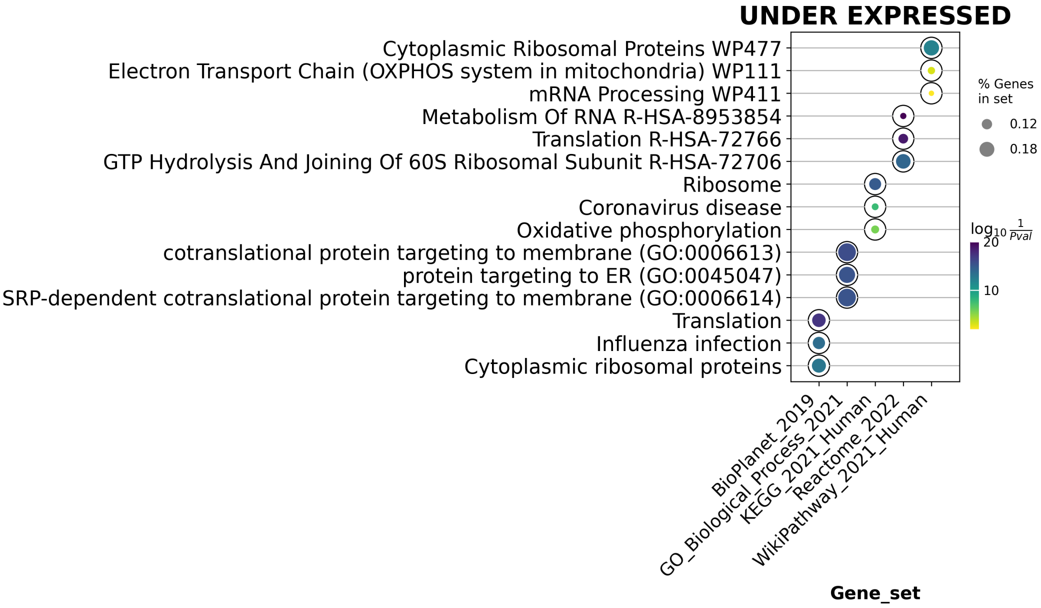

Figure 2: Most enriched pathways with under-expressed genes

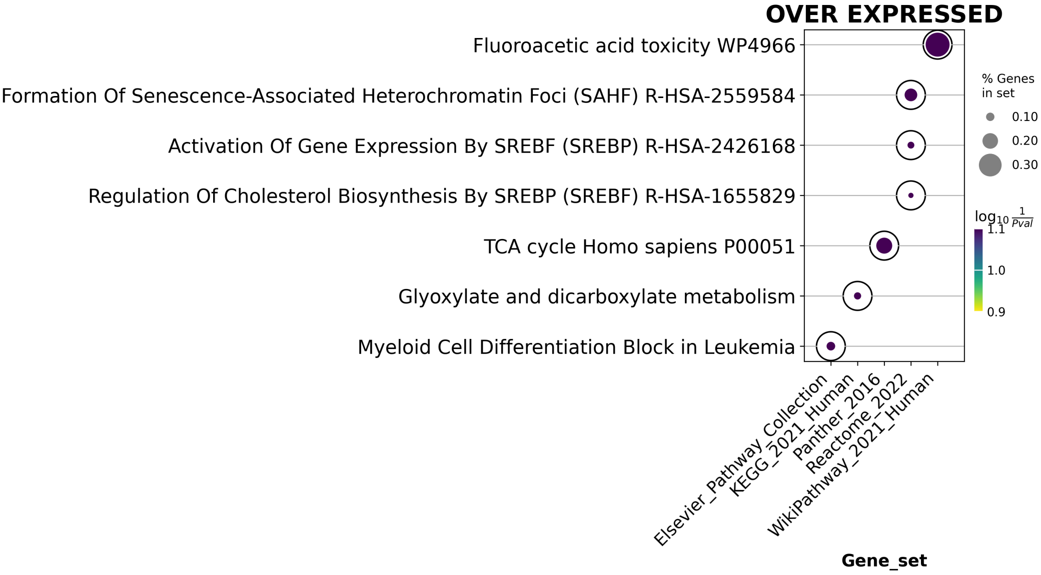

Figure 3: Most enriched pathways with over-expressed genes

running-sum statistic when a gene belongs to the pathway and decreasing it when the gene does not.

Since we have already an idea about correlation between genes in our dataset and pathways, we decide to make further analysis only with one of those pathways.
"gsea" function gives different scores to characterize its results:

- Enrichment Score (ES): is the maximum deviation from zero encountered by the running-sum statistic during the walk. It reflects the degree to which the genes in a pathway are overrepresented at the top or bottom of the entire ranked list of genes.
- Normalized Enrichment Score (NES): is the ES adjusted to account for differences in the pathway sizes and in correlations between pathways and the expression data set. Values over 1 or under -1 imply a significant enrichment.
- Nominal p-value: is the probability under the null distribution (H0: every gene has the same probability of being in any point of the ranked list), of obtaining an ES value that is as strong or stronger than that observed in our experiment.
- False Discovery Rate (FDR) q-value: is a different way (from Benjamini-Hochberg procedure) to adjust p-values for multiple testing.

In figure 4 and 5 we analyze the result obtained for the pathway "Translation R-HSA-72766" that we have already discussed with the over-representation analysis:

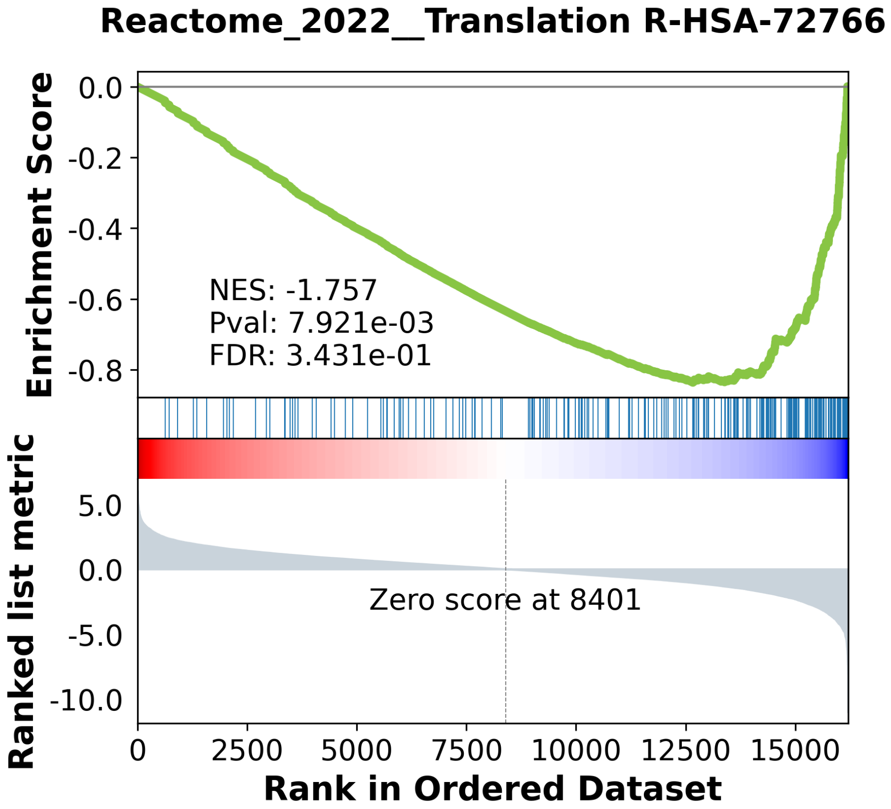

Figure 4: GSEA for the pathway "Translation R-HSA-72766"

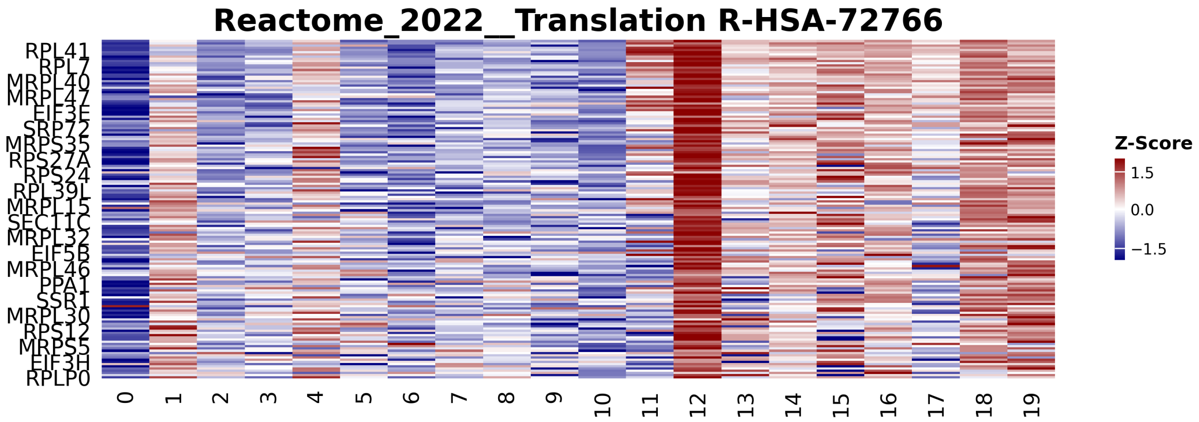

Figure 5: Heatmap for genes in pathway ”Translation R-HSA-72766” (Notice that on the y axis there are all the dataser genes found in the pathway even if just few names are shown)

Figure 4 shows how genes belonging to this pathway are mainly at the bottom of the ranked list. This is coherent with what we found in the previous analysis: most of the pathway genes are underexpressed. Moreover, in Figure 5, we plot the heatmap showing correlation between genes expression and patients. In particular, we evaluate only genes belonging to the pathway we are considering. The first 11 samples are breast cancer patients, while last 9 are healthy ones. It is quite clear that genes are differentially correlated among this two phenotypes.
To conclude, the results found in this analysis confirm what already obtained before.

5 Conclusion

The analysis we carried out shows that the majority of enriched pathways are related with cellular processes. Intuitively this is correct, since cancer is usually associated with unusual attractors in epigenetic landscape. Results also confirm what already said in [Hong G(2013)], since analyzing all genes together does not bring us new meaningful results. Looking at the different techniques used we notice that it could be adavantageous to concentrate more on GSEA in future analysis. In fact, this technique is able to understand deeper the intrisic mechanisms underlying gene expression. Finally, we suggest also to better analyze the relationship between breast tumor and diseases such as COVID 19, since they seem to be correlated.

References

[Biosystems(2006)] Applied Biosystems. Abi human genome survey microarray version 2, 2006. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2986.

[Chen EY()] Kou Y et al. Chen EY, Tan CM. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool.

[Fang(2022)] Zhuoqing Fang. gseapy documentation, 2022. URL https://gseapy.readthedocs.io/en/master/index.html.

[Hong G(2013)] Li H Shen X Guo Z. Hong G, Zhang W. Separate enrichment analysis of pathways for up- and downregulated genes. 2013. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3899863/.

[Karishma Sajnani(2017)] Robert Anthony Smith Vinod Gopalan Alfred King-Yin Lam Karishma Sajnani, Farhadul Islam. Genetic alterations in krebs cycle and its impact on cancer pathogenesis. *Biochimie*, 135:164–172, 2017. ISSN 0300-9084. doi: https://doi.org/10.1016/j.biochi.2017.02.008. URL https://www.sciencedirect.com/science/article/pii/S0300908416303480.

[Kuleshov MV(2016)] Rouillard AD et al. Kuleshov MV, Jones MR. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. 2016. URL https://pubmed.ncbi.nlm.nih.gov/27141961/.

[Stathopoulos GP(2008)] Armakolas A. Stathopoulos GP. Gene expression analysis of whole blood samples from patients with single and double primary tumors and healthy controls. *National Library of Medicine*, 2008. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11545.

[Thomas lab(2022)] University of Southern California Thomas lab. Panther classification system, 2022. URL http://www.pantherdb.org/.