**CGS4144 Bioinformatics - Assignment 2**

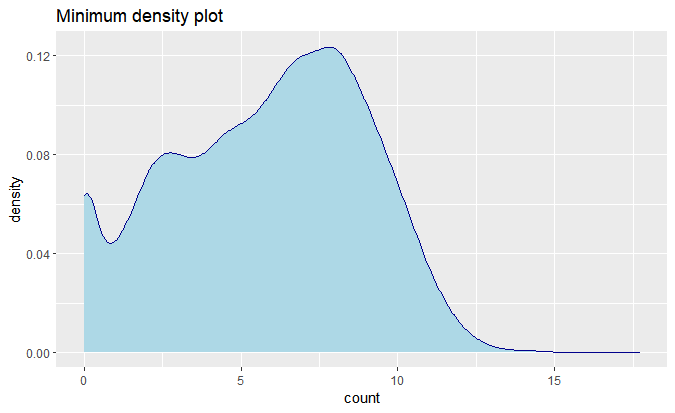
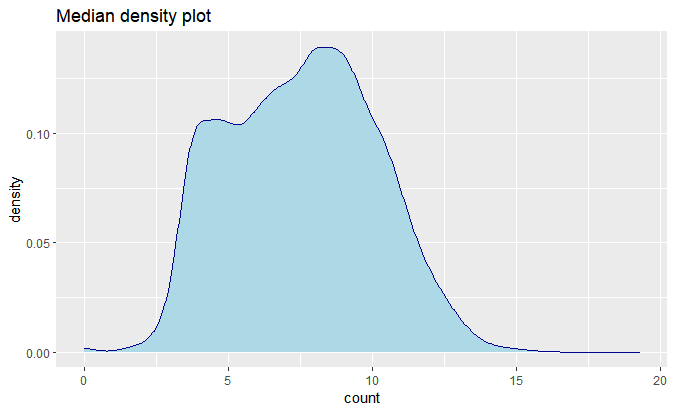
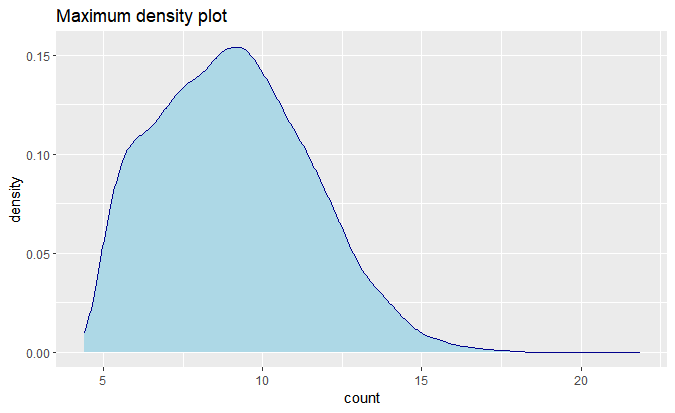
By: Liron Naccache, Seggev Haimovich, Jayden Johnson, and William Manuel

1. **Expression matrix size**: 15929 rows and 69 columns,

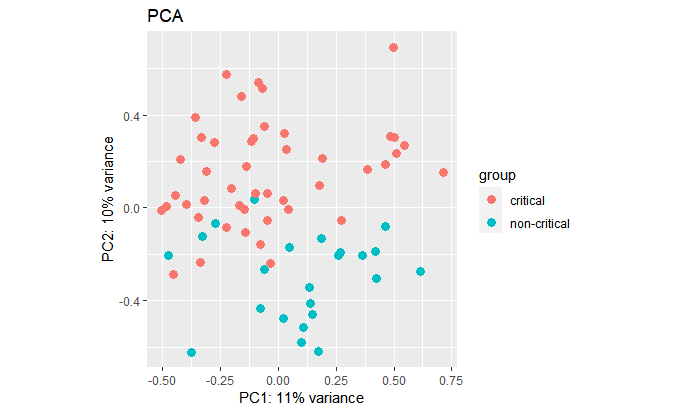
**The number of genes**: Each row represents a gene, meaning there are 15929 genes.

**Variations in the data:** We observed that there are some genes with more counts than others, however, there aren’t many genes with significantly high counts.

We calculated for each gene the median, minimum, and maximum values of counts and produced the following density plots.

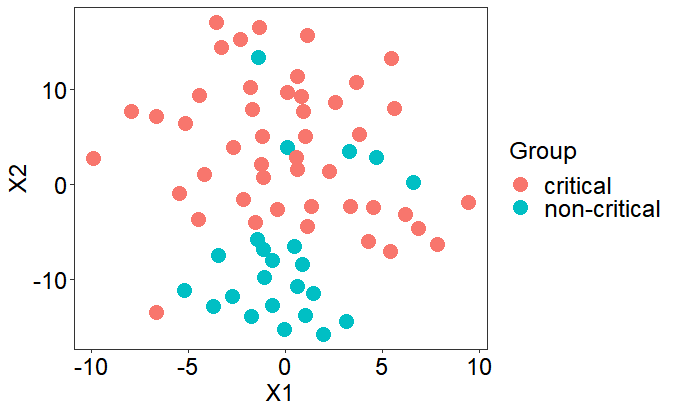


1. In this section, we used dimensional reduction algorithms to check if there’s a signal in our data. As mentioned above, each sample has more than 15,000 features (aka genes) meaning it’s very hard to learn something from the raw data. That’s why we reduced the data dimensions to 2. Using PCA we got the following result:



From the graph, we can infer that the main components that describe our two groups are different.

Using T-SNE we got the graph below:

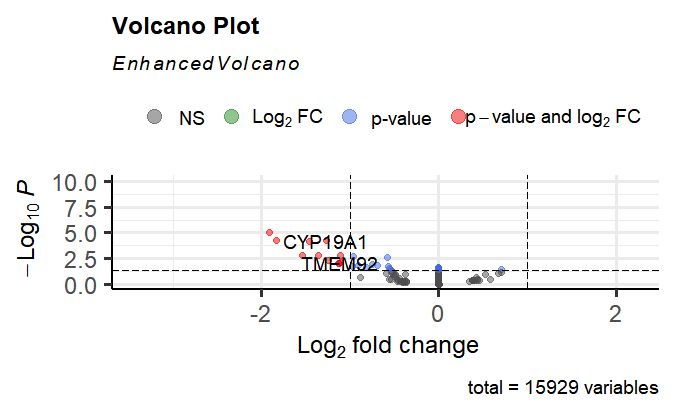


Here, the separation is less clear but still exists.

We see that in both graphs the groups are almost linear separatable and in both graphs, the critical group’s values are higher than the non-critical group’s values.

Using both methods, we infer that there might be a signal in our data. Nevertheless, it’s known that when reducing dimension one can lose data, so it might be a false signal due to a reduction that was too radical and destroyed it.

1. In this section, we created a table of differentially expressed genes and plotted the volcano graph:



The y-axis represents the -log10 of the P-value: if a gene is statistically significant, its p-value will be close to zero, which means a higher position in the y-axis on the Volcano plot.

The x-axis represents the log2 of the fold change, meaning the ratio between the counts of a gene in the group of critical patients and the group of non-critical patients.

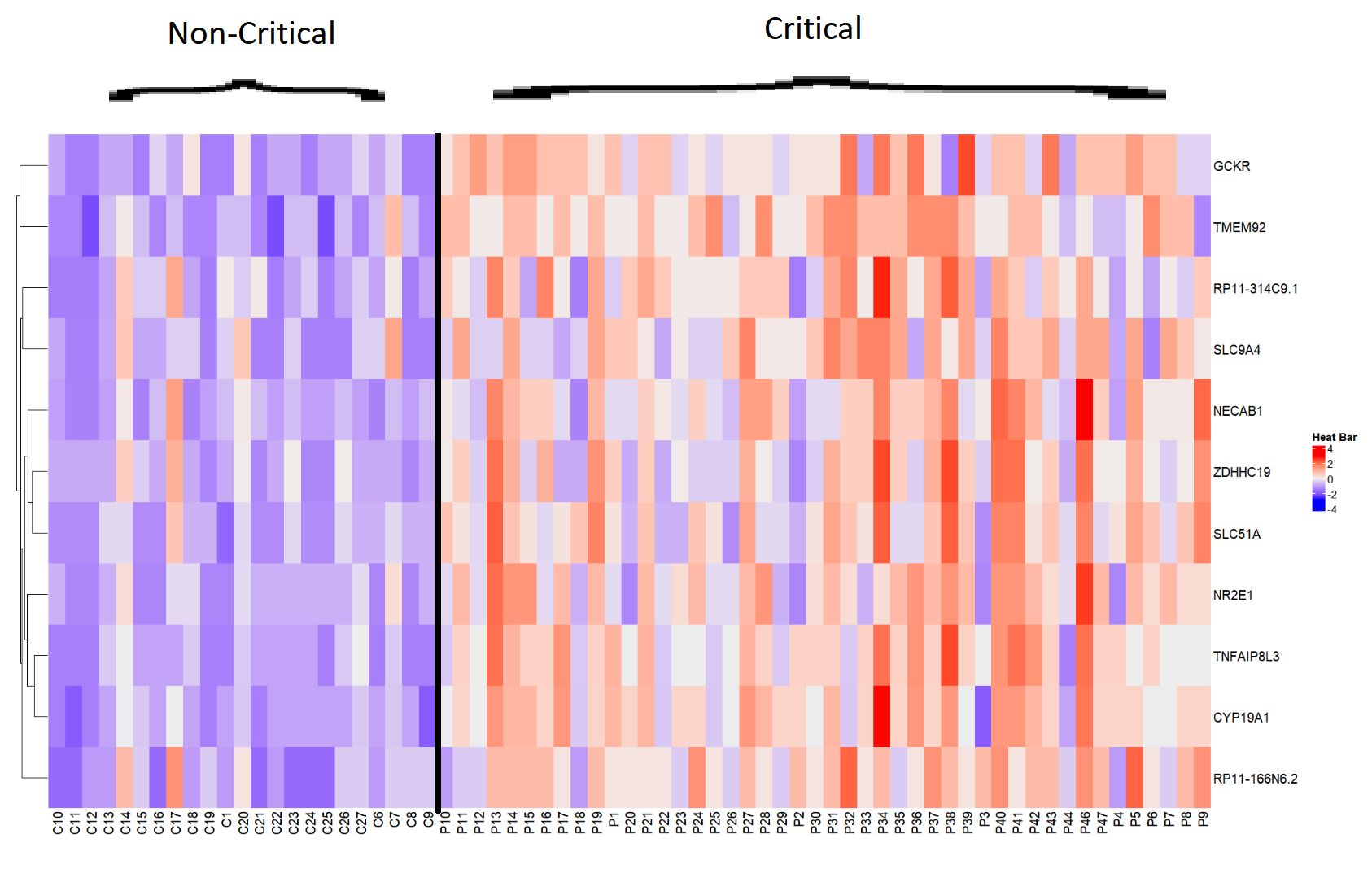
We can see that only 11 genes passed the thresholds.

Our threshold for log\_2 fold change is the default threshold (1), we didn’t change it because it seems like a small change will add a small number of genes passing the threshold and a big change will cause too many genes to pass the threshold and make the results irrelevant.

Our p-value threshold is “p<0.05”. We followed the tutorial in the exercise PDF, which indicated that due to the use of adjusted p-values we can loosen the default p-value threshold.

These are the first 20 genes in our table (the full table is on GitHub)

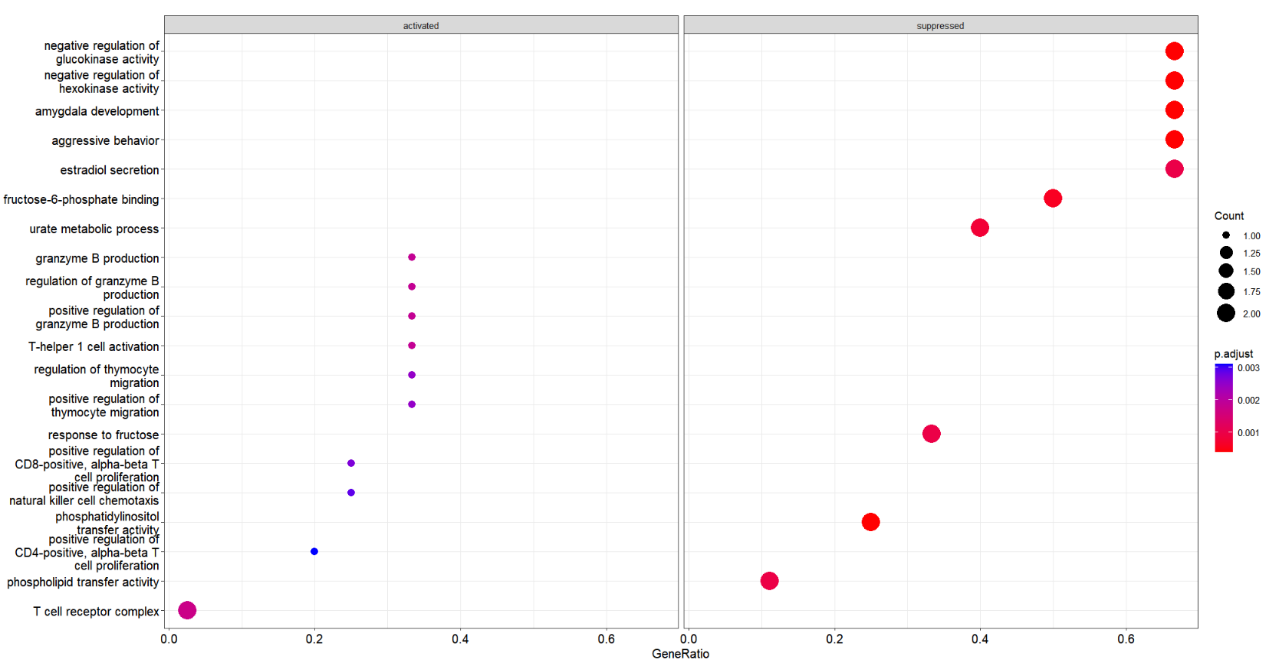


1. In this section, we extracted the 11 differentially expressed genes and created a heatmap:

There aren’t so many differentially expressed genes but in the little ones that we do have, the difference in the expression is very clear. The black line represents the separation between the two sample groups (critical and non-critical patients).

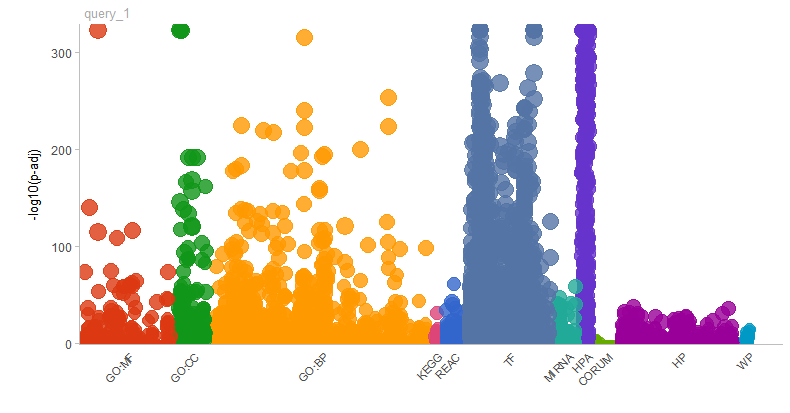
5+6) In this section, we ran an enrichment analysis with 4 different methods:

* 1. topGo
  2. clusterProfiler

By looking at the graph below, we see that some of our genes relate to families that have a connection with the Immune System and are activated and that some genes that have a connection with the Metabolism Process that is suppressed. Since our database relates to Covid-19, it's not surprising to see that it has something to do with the Immune System.

The table below represents the different families, sorted by p-value (The full table is on GitHub). We can see for each family the Ontology it belongs to, the official ID, and more features like set size and qvalue.

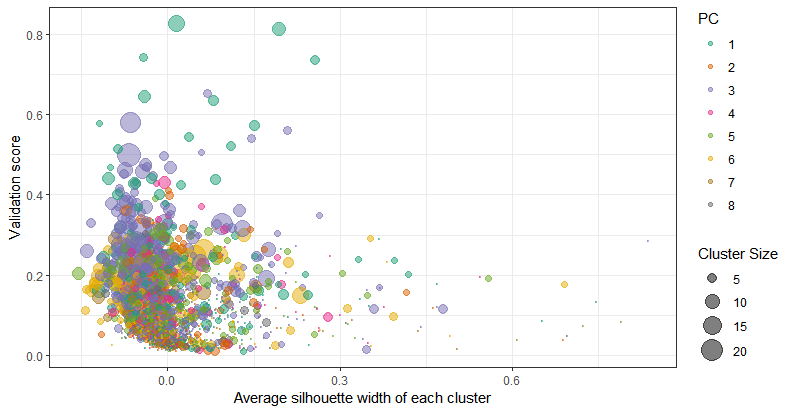
* 1. gProfiler2

Using this graph, we wanted to see if we can connect our genes with specific families of genes, for further exploration. The y-axis is like in the volcano plot, meaning the higher the point is, the gene it represents is more statistically significant. The x-axis represents the different ontologies and pathways. We can infer that many of our statistically significant genes relate to the TF and HPA pathways.

In the table below we can see all the terms and their relevant values such as p-value and source.

* 1. GenomicSuperSignature

In this method, we used 8 principal components from our data and tried to cluster them. The x-axis represents the silhouette width of each cluster, which is a way of understanding how close the points of each cluster are – as being close to 0 is the best. The y-axis represents the Validation score, which is a way of understanding if our prediction is good – being close to 1 is the best. With that being said, we only have a few clusters with small silhouette widths and high validation scores. We think we can not infer much from this graph.



Below, we can see the first 20 rows of the data table used to create the graph above (The full chart is on GitHub). Each line represents a cluster and has information like cluster size and the PC it relates to.

