# Assignment 2

Michael Mezzina

Zhiqing Qu

### GitHub Repository

Get the source: https://github.com/michaelmez39/CIS4930CovidInLungs

### Data Loading

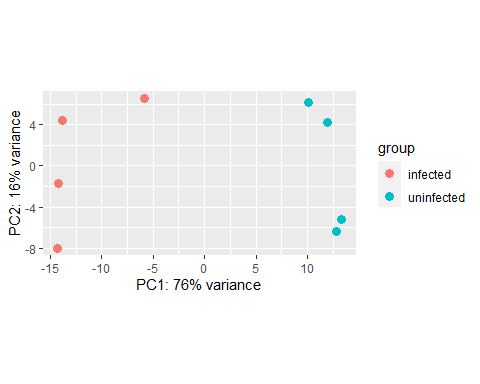
The steps we took to load our data:

1. Removing some extra columns that the authors had computed in our counts matrix

2. Merging repeated rows of genes using the median count value

3. Merging the infected cell counts with uninfected cell counts

### PCA Plot



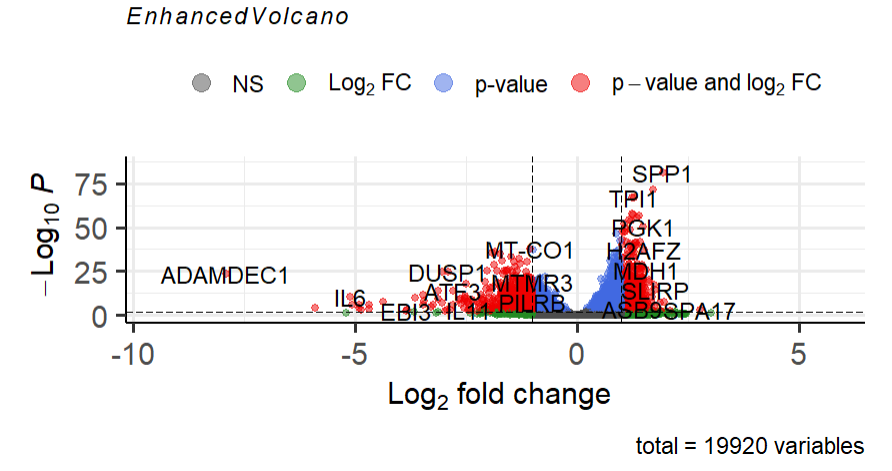
In this plot, we plotted rlog transformed data to check results. To compare the infected and uninfected genes, we plotted one with plotPCA() function. An interesting result is that the plot produced huge and distinct differences between infected and uninfected genes

### Volcano Plot

#### DSEA Analysis Results

## Gene baseMean log2FoldChange lfcSE stat pvalue padj threshold  
## 1 LCK 3.485863 4.231088 1.331835 3.176887 0.001488651 NA NA  
## 2 CD7 3.003281 4.019607 2.257319 1.780700 0.074961537 NA NA  
## 3 GRM6 2.699829 3.859555 1.805952 2.137131 0.032587355 NA NA  
## 4 ABHD16B 1.979928 3.412723 2.028978 1.681991 0.092570644 NA NA  
## 5 C1orf95 1.979928 3.412723 2.028978 1.681991 0.092570644 NA NA  
## 6 VSIG8 1.963892 3.396306 1.804804 1.881814 0.059861329 NA NA  
## 7 SLC2A7 1.898840 3.355788 2.431937 1.379883 0.167622608 NA NA  
## 8 ANP32C 3.112576 3.302454 1.643594 2.009289 0.044506546 NA NA  
## 9 ACTG2 3.086690 3.297262 1.856547 1.776019 0.075729839 NA NA  
## 10 CD84 1.815817 3.291125 2.449052 1.343836 0.179001272 NA NA

In this results table, we used the wrapper function DESeq() to do the differential expression analysis. We also used tidyverse to do sort or filter before saving our results to a file since it is not sorted or filtered. Finally, we printed out the top results. An interesting result is that LCK and CD7 had similar log2FoldChange results.



In the volcano plot, we plotted the Log2FoldChange on the x-axis and padj on the y-axis. The padj variable is the p values corrected with Benjamini-Hochberg. An interesting result is that pCutoff argument can be made 0.01, which the default value is 1e-05, since we were using adjusted p values.

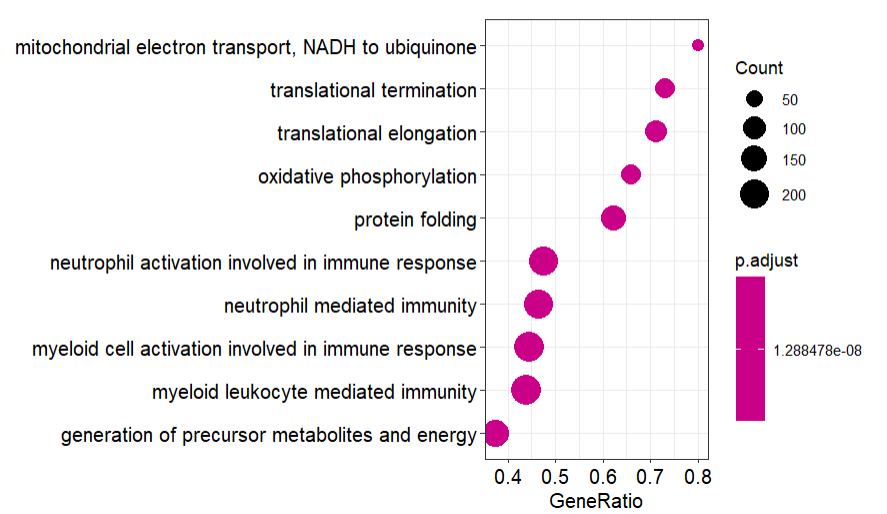
### ClustProfiler Gene Ontology

Uses the Human genome annotation in the analysis. Using the biological processes ontology option, to see effects outside cell.

### Analysis Results

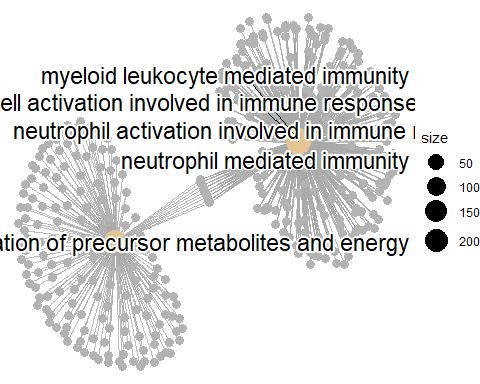
## ID Description  
## GO:0002275 GO:0002275 myeloid cell activation involved in immune response  
## GO:0002283 GO:0002283 neutrophil activation involved in immune response  
## GO:0002444 GO:0002444 myeloid leukocyte mediated immunity  
## GO:0002446 GO:0002446 neutrophil mediated immunity  
## GO:0006091 GO:0006091 generation of precursor metabolites and energy  
## GO:0006119 GO:0006119 oxidative phosphorylation  
## GO:0006120 GO:0006120 mitochondrial electron transport, NADH to ubiquinone  
## GO:0006414 GO:0006414 translational elongation  
## GO:0006415 GO:0006415 translational termination  
## GO:0006457 GO:0006457 protein folding  
## setSize enrichmentScore NES pvalue p.adjust qvalues  
## GO:0002275 468 0.3583013 1.964671 1e-10 1.288478e-08 1.029062e-08  
## GO:0002283 421 0.4010586 2.208797 1e-10 1.288478e-08 1.029062e-08  
## GO:0002444 475 0.3496213 1.918329 1e-10 1.288478e-08 1.029062e-08  
## GO:0002446 433 0.3821367 2.098519 1e-10 1.288478e-08 1.029062e-08  
## GO:0006091 442 0.4142780 2.288052 1e-10 1.288478e-08 1.029062e-08  
## GO:0006119 103 0.6541312 3.003821 1e-10 1.288478e-08 1.029062e-08  
## GO:0006120 45 0.7818669 3.088561 1e-10 1.288478e-08 1.029062e-08  
## GO:0006414 125 0.5788764 2.743235 1e-10 1.288478e-08 1.029062e-08  
## GO:0006415 100 0.6610471 3.003169 1e-10 1.288478e-08 1.029062e-08  
## GO:0006457 204 0.5092289 2.587248 1e-10 1.288478e-08 1.029062e-08  
## rank leading\_edge  
## GO:0002275 4686 tags=44%, list=24%, signal=35%  
## GO:0002283 4686 tags=48%, list=24%, signal=37%  
## GO:0002444 4686 tags=44%, list=24%, signal=34%  
## GO:0002446 4686 tags=46%, list=24%, signal=36%  
## GO:0006091 3391 tags=37%, list=17%, signal=32%  
## GO:0006119 3091 tags=66%, list=16%, signal=56%  
## GO:0006120 2718 tags=80%, list=14%, signal=69%  
## GO:0006414 5485 tags=71%, list=28%, signal=52%  
## GO:0006415 4522 tags=73%, list=23%, signal=57%  
## GO:0006457 5876 tags=62%, list=30%, signal=44%  
##

#### Dot Plot



In the dot plot, we see many of the relevant genes found in the analysis. The p-values are quite small, which is good, but seems to have caused a problem with the coloring scale. Many of the genes related to the immune response have a high number of counts, indicating that the infected cells are responding to the covid-19 infection.

#### CNet Plot



For the gene ontology using clusterProfiler we looked at the biological processes domain. Many of the genes that are being expressed are related to the immune response. There seems to be two large clusters, possibly one cluster related to immune response and the other to metabolites and energy.