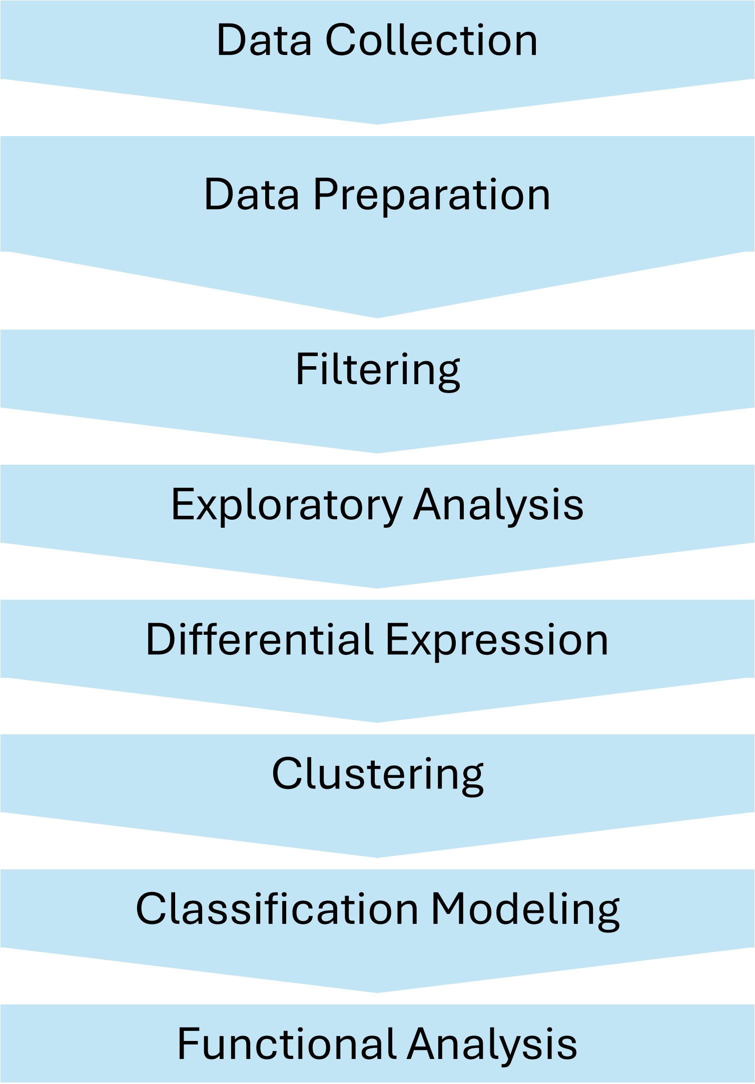
Gene Expression and Functional Analysis of SARS Dataset (GDS1028)

AS.410.671 Gene Expression Data Analysis and Visualization

Hendyco Pratama

Workflow Overview



Data Collection

- Dataset: GEO Dataset (GDS1028).
- Data type: Expression data using Affymetrix Human HG-Focus Target Array.

Data Preparation

- Wrangling: Removal of unused columns, renaming, handling missing values.
- Normalization: Tested Quantile and Cyclic Loess methods. Chose Quantile normalization.
- Transformation: Log2 transformation for data homogeneity.

Filtering

- Noise filtering: Expression levels <5.0 and expressed in <25% of samples removed.
- Outlier removal: Identified SARS_3 as an outlier using graphical assessment.

Exploratory Analysis

- Histograms: Distribution assessment
- Boxplots: Variance analysis between SARS and control groups
- F-test: Confirmed no significant variance differences, enabling parametric testing.

Differential Expression

- Student's T-Test: Identified 1923 significant genes ($p < 0.05$).
- Benjamini-Hochberg adjustment: Reduced to 256 significant genes.

Clustering

- PCA and Hierarchical clustering: Visualized group separation and identified anomalies (e.g., SARS_9 resembled controls).

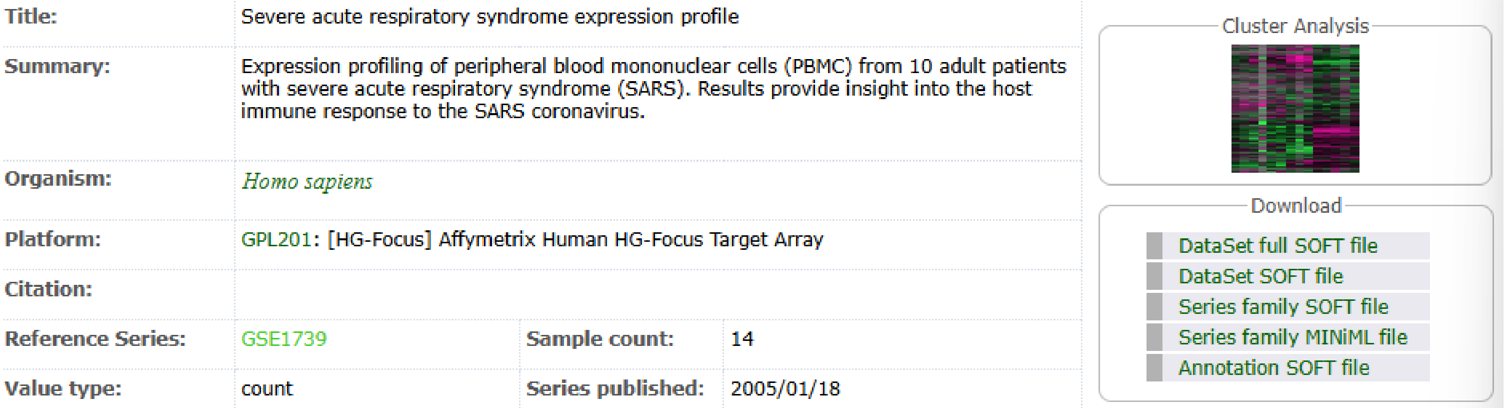
Classification Modeling

- Linear Discriminant Analysis (LDA): Achieved perfect classification on the test set.

Functional Analysis

- Analyzed gene function and pathways using NCBI DAVID.

Data set description



Title:	Severe acute respiratory syndrome expression profile		
Summary:	Expression profiling of peripheral blood mononuclear cells (PBMC) from 10 adult patients with severe acute respiratory syndrome (SARS). Results provide insight into the host immune response to the SARS coronavirus.		
Organism:	*Homo sapiens*		
Platform:	GPL201: [HG-Focus] Affymetrix Human HG-Focus Target Array		
Citation:			
Reference Series:	GSE1739	Sample count:	14
Value type:	count	Series published:	2005/01/18

- The data set used in this project is a GEO Dataset (GDS1028)[1]
- Severe Acute Respiratory Syndrome Expression Profile Dataset
- Expression data was gathered using GPL201: Affymetrix Human HG-Focus Target Array
- Overall, the data contains 14 samples (4 control and 10 SARS) with over 8000 genes.
- Data could be accessed from: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1028

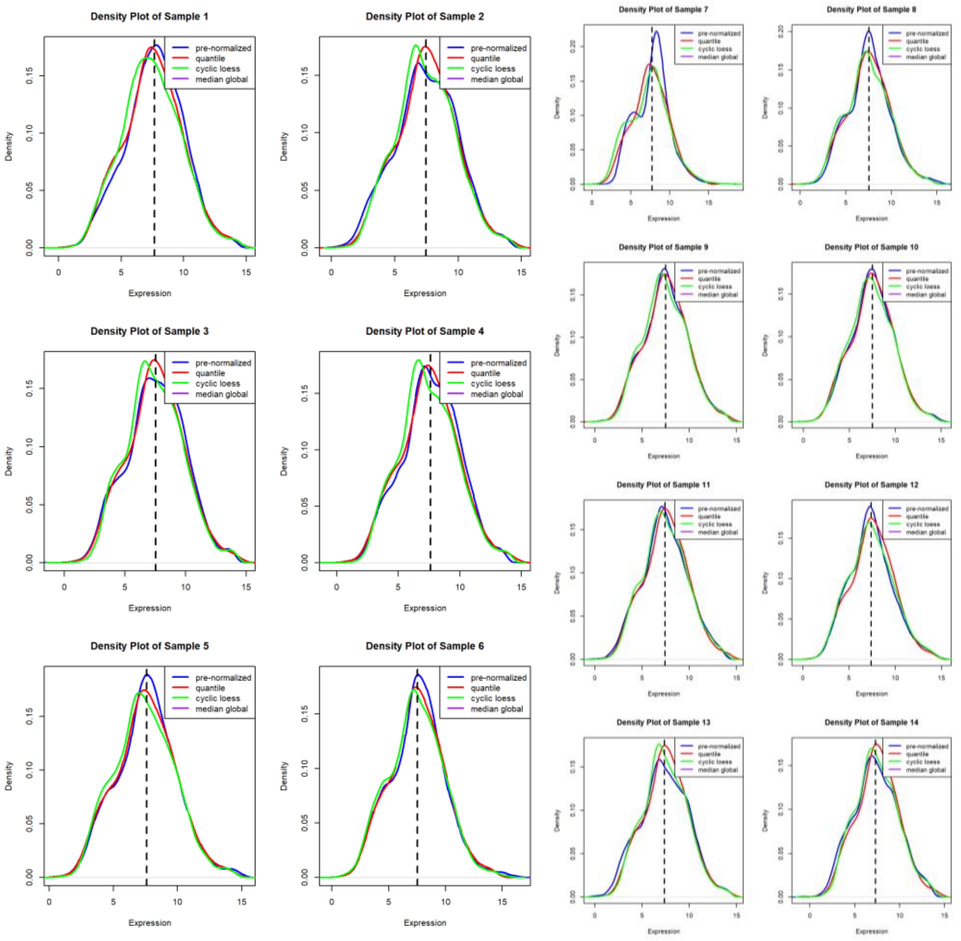
Data Wrangling, Transformation and Normalization

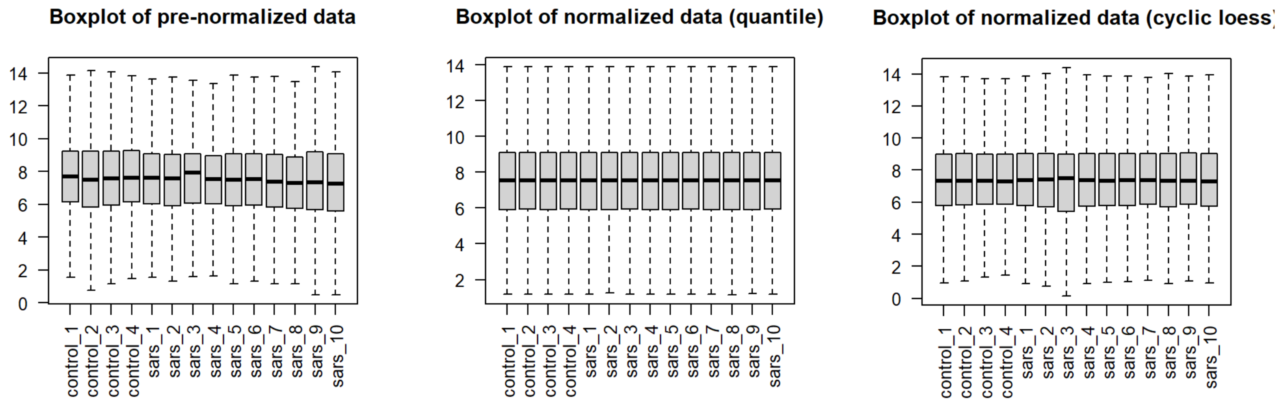
- Includes checking for NAs or 0s in the data set
- Removing IDNETIFIER column as it contains gene name and will not be used in the analysis
- Renaming column into group name rather than sample ID to make analysis easier

- The raw data has not been transformed yet
- Log2 transformation was done to increase homogeneity

- 2 normalization method were tested and compared
  - Quantile normalization
  - Cyclic Loess Normalization

Normalization Result

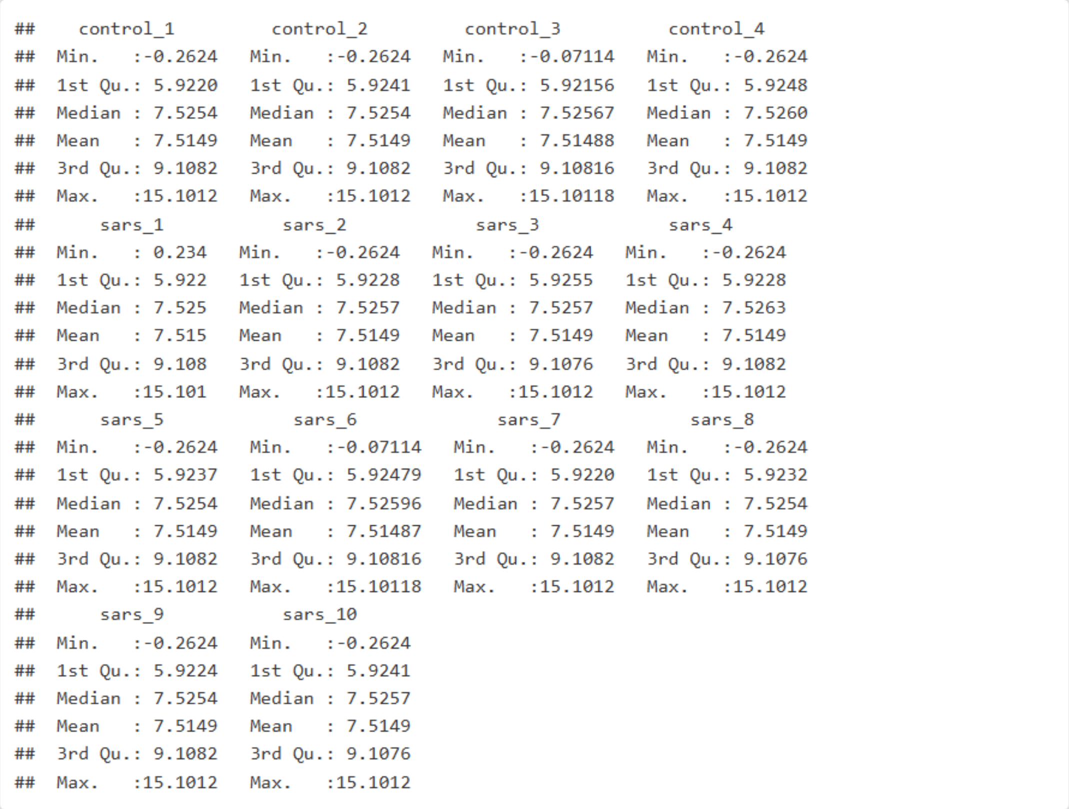




- Normalization results were visualized with both density and box plot.
- Density plot – especially on Sample 2, 3, and 7 – shows that quantile normalization works better
- Boxplot results also suggest the same thing with quantile normalization mean lines and box size being more uniform across samples
- Hence, **quantile normalized data will be used**

Noise Filtering

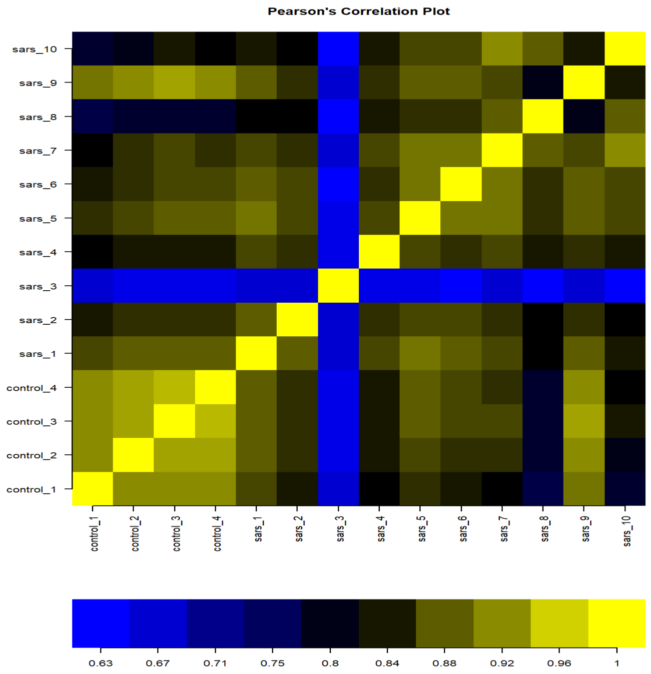
- Noise filtering criteria:
  - Expression level of less than 5.0 since the 1st quantile across samples are around 5.9
  - Expressed in at least 25% of the samples
- Result after filtration:
  - Reduced gene number from 8793 to 8286



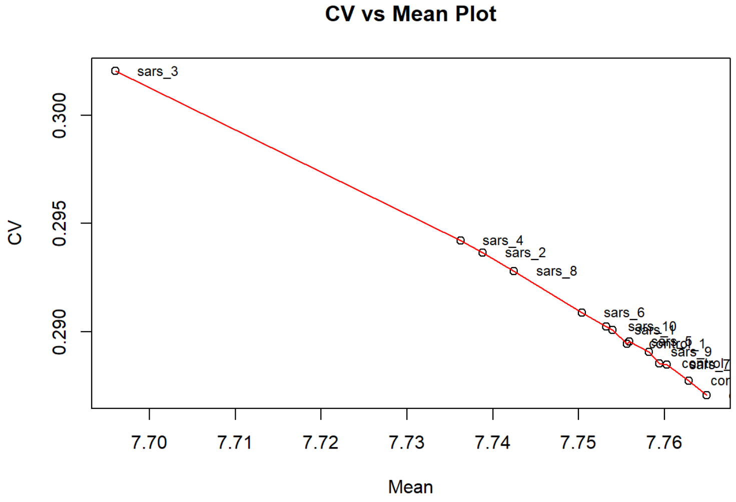
```
##    control_1         control_2         control_3           control_4      
##  Min.   :-0.2624   Min.   :-0.2624   Min.   :-0.07114   Min.   :-0.2624  
##  1st Qu.: 5.9220   1st Qu.: 5.9241   1st Qu.: 5.92156   1st Qu.: 5.9248  
##  Median : 7.5254   Median : 7.5254   Median : 7.52567   Median : 7.5260  
##  Mean   : 7.5149   Mean   : 7.5149   Mean   : 7.51488   Mean   : 7.5149  
##  3rd Qu.: 9.1082   3rd Qu.: 9.1082   3rd Qu.: 9.10816   3rd Qu.: 9.1082  
##  Max.   :15.1012   Max.   :15.1012   Max.   :15.10118   Max.   :15.1012  
##      sars_1           sars_2            sars_3            sars_4       
##  Min.   : 0.234   Min.   :-0.2624   Min.   :-0.2624   Min.   :-0.2624  
##  1st Qu.: 5.922   1st Qu.: 5.9228   1st Qu.: 5.9255   1st Qu.: 5.9228  
##  Median : 7.525   Median : 7.5257   Median : 7.5257   Median : 7.5263  
##  Mean   : 7.515   Mean   : 7.5149   Mean   : 7.5149   Mean   : 7.5149  
##  3rd Qu.: 9.108   3rd Qu.: 9.1082   3rd Qu.: 9.1076   3rd Qu.: 9.1082  
##  Max.   :15.101   Max.   :15.1012   Max.   :15.1012   Max.   :15.1012  
##      sars_5             sars_6              sars_7            sars_8       
##  Min.   :-0.2624   Min.   :-0.07114   Min.   :-0.2624   Min.   :-0.2624  
##  1st Qu.: 5.9237   1st Qu.: 5.92479   1st Qu.: 5.9220   1st Qu.: 5.9232  
##  Median : 7.5254   Median : 7.52596   Median : 7.5257   Median : 7.5254  
##  Mean   : 7.5149   Mean   : 7.51487   Mean   : 7.5149   Mean   : 7.5149  
##  3rd Qu.: 9.1082   3rd Qu.: 9.10816   3rd Qu.: 9.1082   3rd Qu.: 9.1076  
##  Max.   :15.1012   Max.   :15.10118   Max.   :15.1012   Max.   :15.1012  
##      sars_9           sars_10       
##  Min.   :-0.2624   Min.   :-0.2624  
##  1st Qu.: 5.9224   1st Qu.: 5.9241  
##  Median : 7.5254   Median : 7.5257  
##  Mean   : 7.5149   Mean   : 7.5149  
##  3rd Qu.: 9.1082   3rd Qu.: 9.1076  
##  Max.   :15.1012   Max.   :15.1012  
```

Outlier Assessment

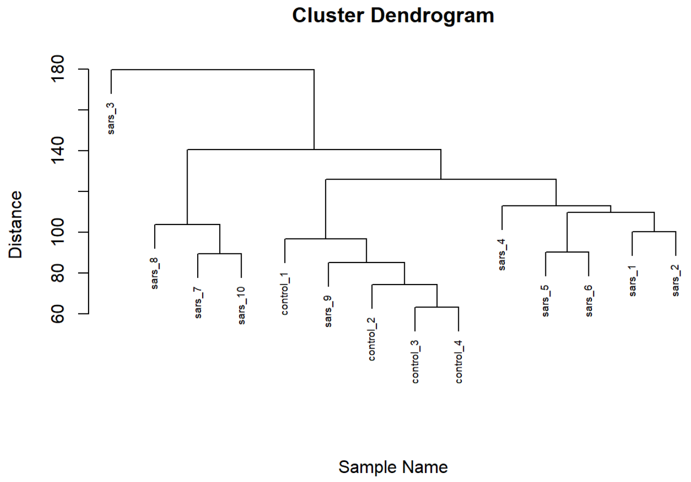
Done with multiple method to assess presence of outliers.



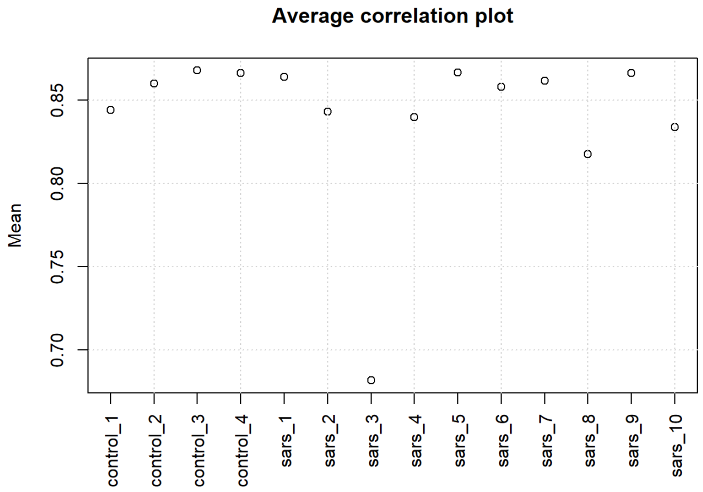
Pearson's Correlation Plot



CV vs Mean Plot



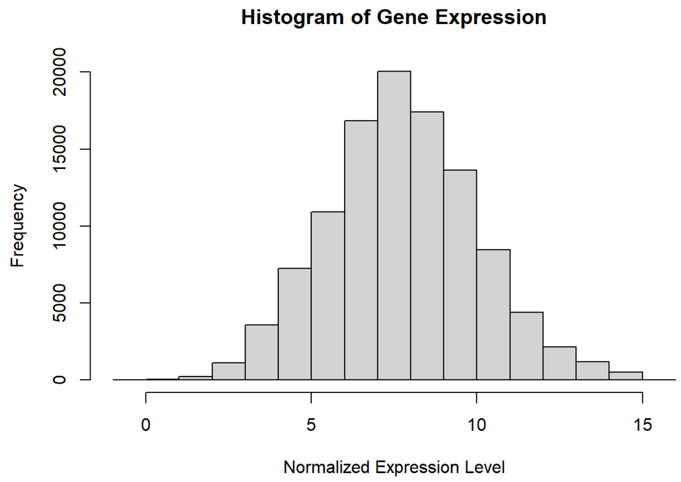
Cluster Dendrogram

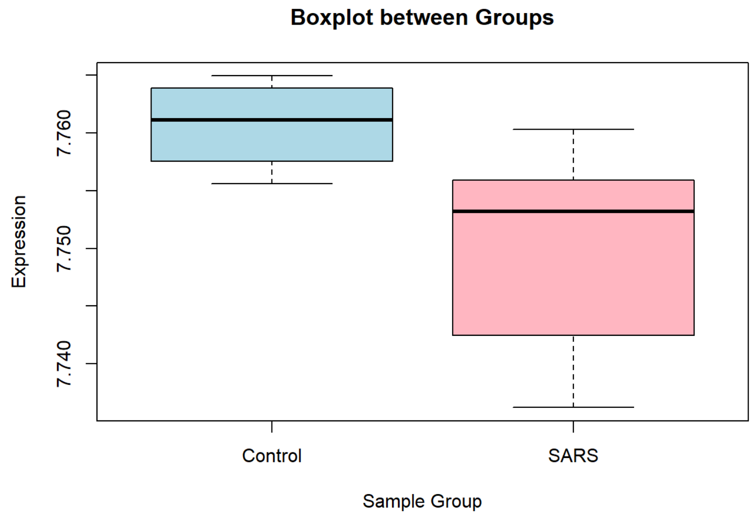


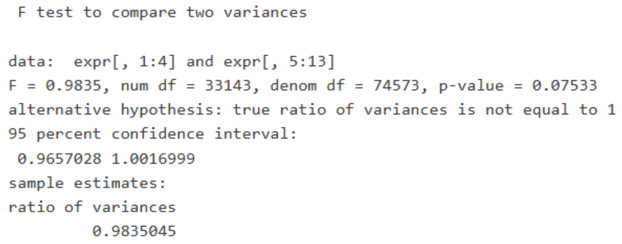
Average correlation plot

Based on the graphs, it seems like SARS_3 sample is an outlier, hence it will be removed

Exploratory Analysis







```
 F test to compare two variances

data:  expr[, 1:4] and expr[, 5:13]
F = 0.9835, num df = 33143, denom df = 74573, p-value = 0.07533
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9657028 1.0016999
sample estimates:
ratio of variances
         0.9835045
```

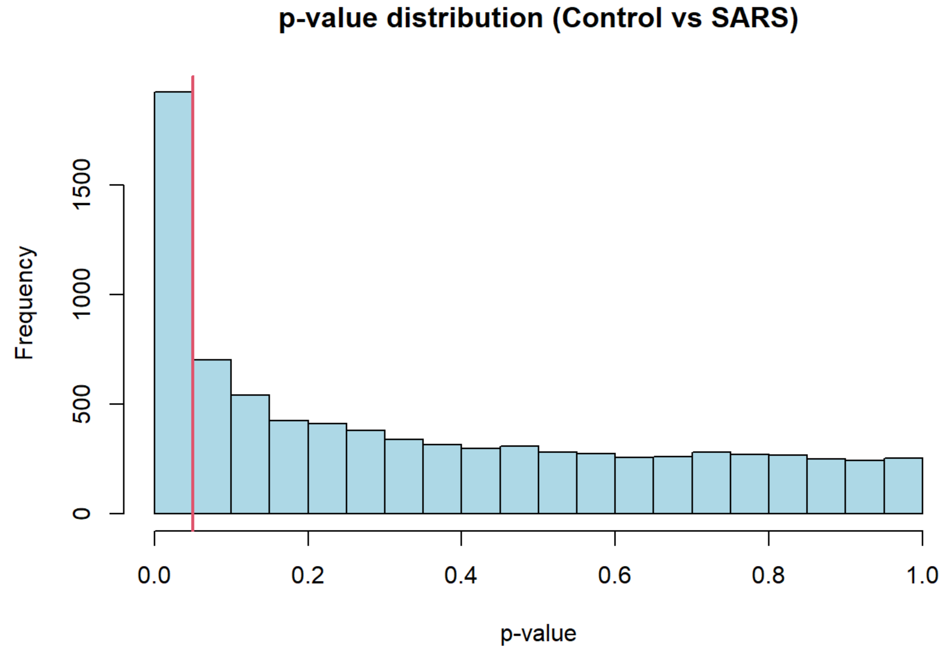
- Histogram was generated to see the distribution of the data

- Boxplot was generated to see how each sample group behaves
- F-test were also done to see the variance behavior

The histogram of the data shows that it follow normal distribution pattern hence it can be considered parametric. Meanwhile, the boxplot shows that the SARS group has more variance compared to the control group - shown by the size of the boxes. Despite that, the F test result (F = 0.983, p-value = 0.0753) shows that there is no significant difference between the variance, hence student t-test could be used.

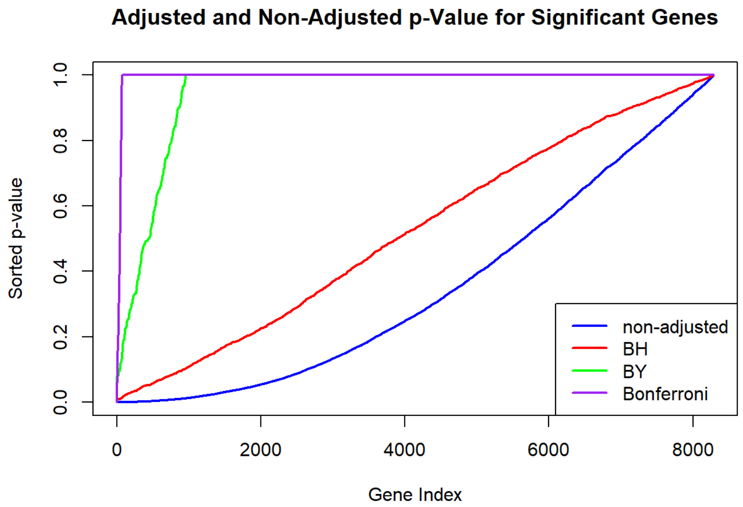
Differential Testing (Student’s T-Test)

- Student’s T-Test shows that there are 1923 significant differentially expressed genes (p < 0.05)



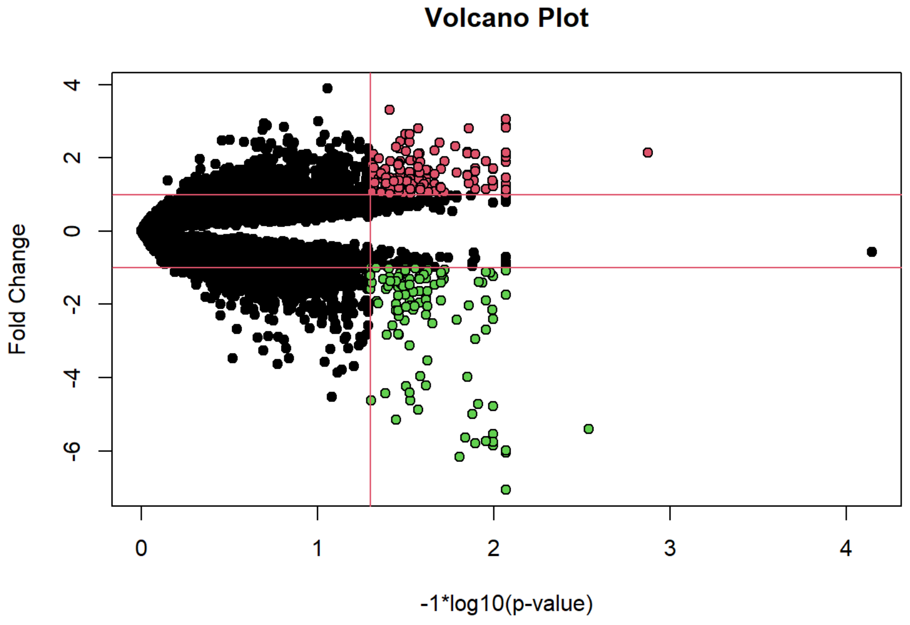
Multiple Testing (Benjamini-Hochberg)

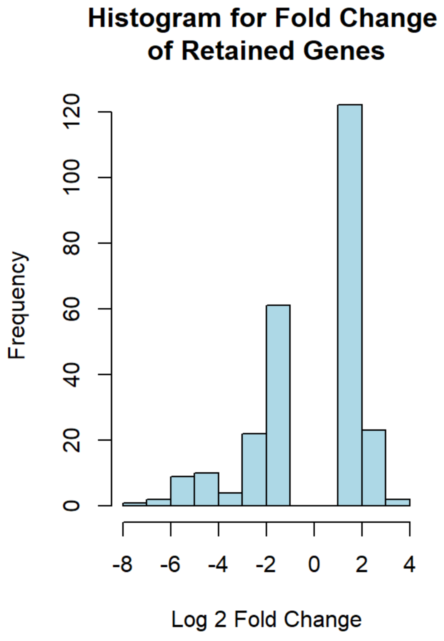
- Multiple testing was done to accommodate increased likelihood of false positive form just using Student’s t-test.
- Benajmini-Hochberg was used because based on the comparison, it is the least conservative compared to BY, and bonferroni

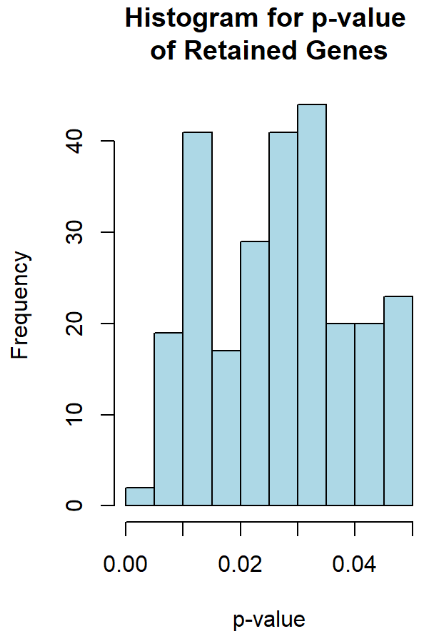


Multiple Testing (con’t.)

- After performing differential testing and using BH adjustment and utilizing fold change, the number of differentially expressed genes drop from 1923 to 256.

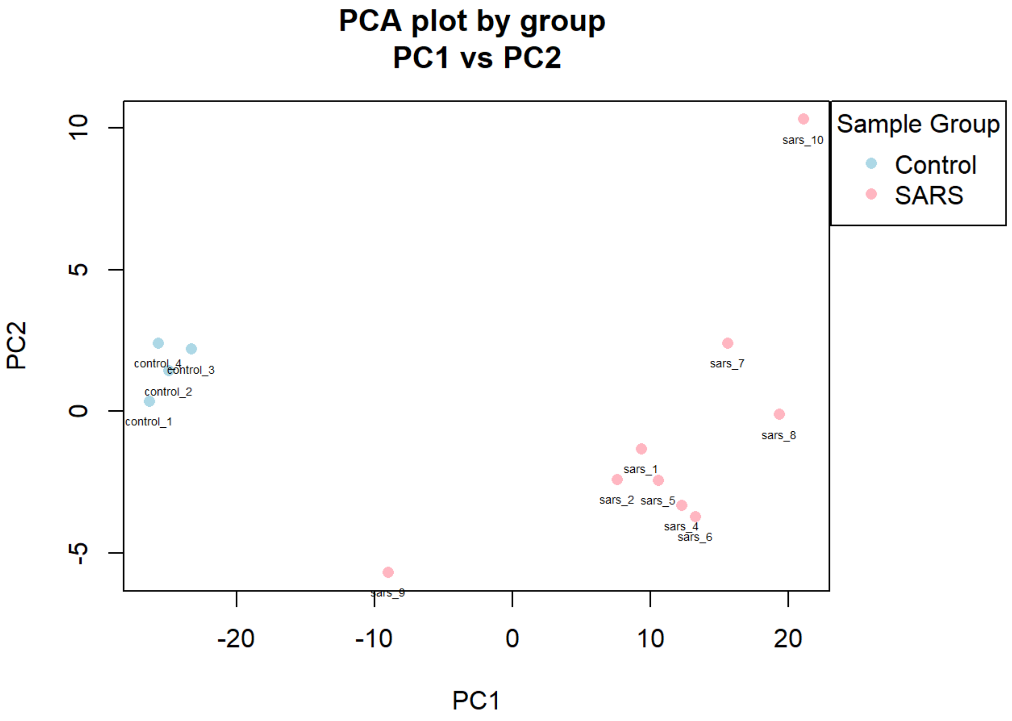

Volcano Plot


Histogram for Fold Change of Retained Genes

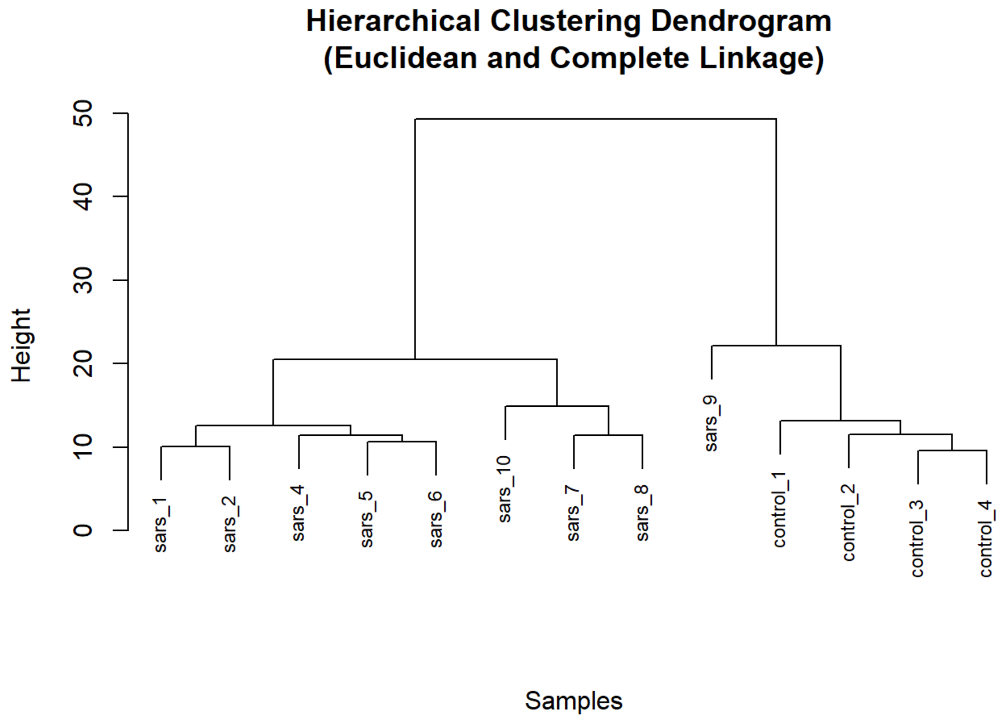

Histogram for p-value of Retained Genes

Clustering

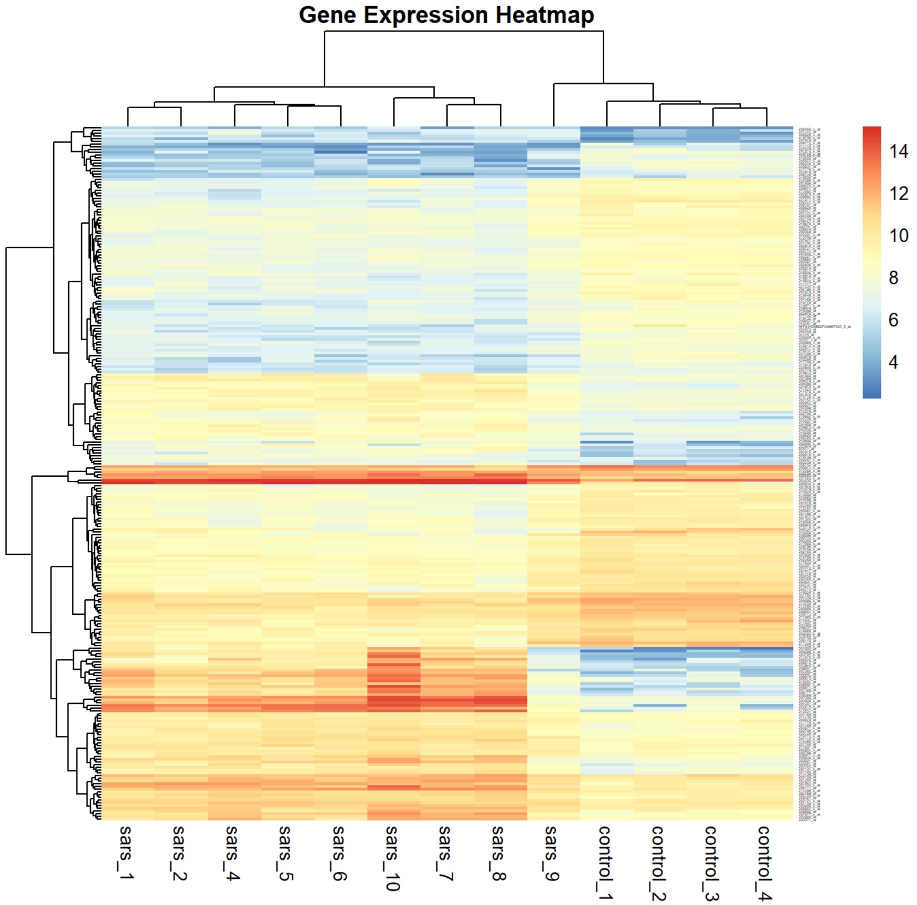
By Dimensional Reduction (PCA)



By hierarchical clustering (HCA)



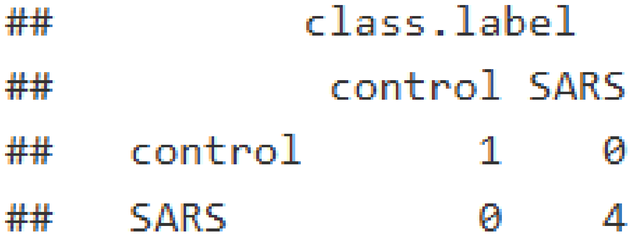
Clustering (Con't.)



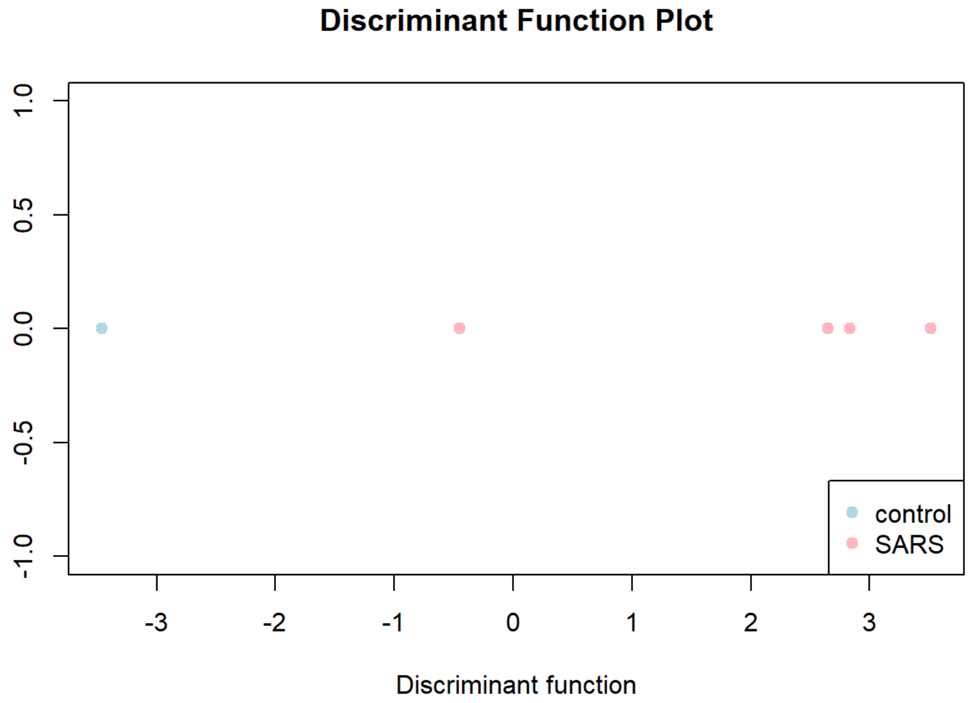
- The PCA scatter plot (PC1 vs. PC2) clearly distinguished SARS and control groups.
- Hierarchical clustering dendrogram revealed SARS_9 grouped with control samples.
- Heatmap of gene expression showed SARS_9 having expression patterns resembling control samples.
- This indicates possible biological variation in SARS_9 compared to other SARS samples.
- Differential expression analysis used the BH (Benjamini-Hochberg) method for FDR control.
- Potential false positives from the analysis could explain the unexpected clustering of SARS_9.

Classification Modeling

- Classification modeling was done by:
  - Dividing data set into training (3 control and 5 SARS samples) and test (1 control and 4 SARS sample) set.
  - Performed using lda
- Confusion Matrix when performed on test set:



```
##          class.label
##           control SARS
##   control       1    0
##   SARS          0    4
```



LDA successfully classified the test set without any misclassification

Functional Analysis (NCBI DAVID)[2, 3]

Gene Symbol	Gene Name	Chromosome Location	GO Term			KEGG Pathway	OMIM Disease
			Biological process	Cellular Component	Molecular Function		
AKAP11	A-kinase anchoring protein	13	Renal Water Homeostasis, Protein Localization, Cortical Actin Cytoskeleton Organization	Nucleus, Cytoplasm, Centrosome, Cytosol, Plasma Membrane	Protein Binding, Protein Phosphastase 1 Binding, Protein Kinase A Regulatory Subunit Binding		
FBXO3	F-box protein 3	11	Proteolysis, Protein Ubiquitination, SCF-dependent protasomal ubiquitin-dependent protein catabolic process	Nucleoplasm, Centrosome, Cytosol, SCF ubiquitin ligase complex	ubiquitin-protein transferase activity, protein binding, ubiquitin=likaligase-substrate adaptor activity		
RBL2	RB Transcriptional corepressor like 2	16	Chromatin Organization, cell cycle, regulation of lipid kinase activity	Chromatin, nucleus, nucleoplasm, transcription regulator complex, chromosome, nucleolus, cytosol, extracellular exosome	RNA polymerase II transcription regulatory sequence-specific DNA Binding, protein binding, promoter-specific chromatin binding	FoxO signaling pathway, Cell cycle, PI3K-Akt signaling pathway, Cellular senescence, Human papillomavirus infection, Viral carcinogenesis,	Brunet-Wagner neurodevelopmental syndrome,
S100A9	S100 calcium bindung protein A9	1	leukocyte migration involved in inflammatory response, chronic inflammatory response, autophagy, apoptotic process, activation of cysteine-type endopeptidase activity involved in apoptotic process, inflammatory response, cell-cell signaling,	extracellular region, extracellular space, nucleus, cytoplasm, cytosol, cytoskeleton, plasma membrane, secretory granule lumen, collagen-containing extracellular matrix, extracellular exosome, calprotectin complex, S100A9 complex,	calcium ion binding, protein binding, microtubule binding, zinc ion binding, antioxidant activity, Toll-like receptor 4 binding, calcium-dependent protein binding, arachidonic acid binding, RAGE receptor binding	IL-17 signaling pathway,	
CASP2	Caspase 2	7	luteolysis, neural retina development, proteolysis, apoptotic process, activation of cysteine-type endopeptidase activity involved in apoptotic process, DNA damage response, DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest,	nucleus, nucleolus, cytoplasm, cytosol, endopeptidase complex,	protease binding, cysteine-type endopeptidase activity, protein binding, enzyme binding, protein domain specific binding, identical protein binding, death domain binding, cysteine-type endopeptidase activity involved in apoptotic signaling pathway, cysteine-type endopeptidase activity involved in execution phase of apoptosis,	Apoptosis,	Intellectual developmental disorder, autosomal recessive 80, with variant lissencephaly,
CHMP2A	Charged multivesicular body protein 2A	19	plasma membrane repair, autophagy, nucleus organization, mitotic metaphase chromosome alignment, membrane invagination, exit from mitosis, regulation of centrosome duplication, protein transport,	autophagosome membrane, kinetochore, chromatin, ESCRT III complex, nuclear envelope, nuclear pore, lysosomal membrane, multivesicular body, kinetochore microtubule, cytosol, plasma membrane, membrane, membrane coat, midbody, multivesicular body membrane, extracellular exosome, amphisome membrane,	protein binding, protein domain specific binding, phosphatidylcholine binding,	Endocytosis, Necroptosis,	
MIR1248	MicroRNA 1248	3	RNA processing	Nucleus			
MPO	Myeloperoxidase	17	response to yeast, hypochlorous acid biosynthetic process, respiratory burst involved in defense response, defense response, response to oxidative stress,	extracellular region, extracellular space, nucleus, nucleoplasm, lysosome, secretory granule, azurophil granule lumen, azurophil granule, intracellular membrane-bounded organelle, extracellular exosome, phagocytic vesicle lumen,	chromatin binding, peroxidase activity, protein binding, heparin binding, heme binding, metal ion binding, lactoperoxidase activity,	Drug metabolism - other enzymes, Phagosome, Neutrophil extracellular trap formation, Transcriptional misregulation in cancer, Acute myeloid leukemia,	Alzheimer disease, susceptibility to, Myeloperoxidase deficiency, Lung cancer, protection against, in smokers,
SRSF5	Serine and arginen rich splicing factor	14	mRNA splicing, via spliceosome, mRNA splice site recognition, mRNA processing,	nucleoplasm, nucleolus, cytosol, nuclear speck,	RNA binding, mRNA binding, protein binding,	Spliceosome, Herpes simplex virus 1 infection,	
TRMT11	tRNA methyltransferase 11 homolog	6	RNA methylation, tRNA processing, methylation,	cytoplasm	tRNA binding, protein binding, methyltransferase activity, tRNA (guanine(10)-N2)-methyltransferase activity,		

Conclusion

- Normalization and filtering were critical in ensuring data quality and reliability.
- Exploratory analysis and clustering revealed distinct differences between SARS and control groups, with minor anomalies (e.g., SARS_9 clustering with controls) attributed to biological variation.
- Differential expression testing identified 256 genes with significant expression changes, which were further explored for functional relevance.
- Classification modeling (LDA) proved highly effective, demonstrating the ability to distinguish SARS from controls with no misclassification.
- Functional analysis linked significant genes to key biological processes and pathways, offering insights into SARS pathogenesis and potential therapeutic targets.

Reference

[1] Jayapal, M., Regunathan, R., Melendez, A. J., Tai, D., Leung, B. P., Reghunathan, R., Hsu, L. Y., & Chng, H. H. (2004). *Expression profile of immune response genes in patients with Severe Acute Respiratory Syndrome* [Data set]. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1739 (Accession No. GSE1739)

[2] Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, 50(W1), W216–W221. https://doi.org/10.1093/nar/gkac194

[3] Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44–57. https://doi.org/10.1038/nprot.2008.211