DATA3001: Project Report

Discovering Characteristic Gene Expressions of COVID-19 Patients

By *Beyond Analysis:*

Ayra Islam (z5255744)
Lukasz Lewandowski (z5257195)
Nadya Darmawan (z5231267)
Tiarne Lo (z5254839)
Yvanne Shalauddin (z5259532)

1. Team name

Beyond Analysis

2. Mission Statement

Our mission is to verify whether or not there are genes which make certain individuals more susceptible to COVID-19 than others, and develop predictive methods based on our findings.

3. Executive Summary

Over the past two years, COVID-19 has become a global concern and has led to worldwide disaster. Spanning from global economic recession to direct and indirect health effects, the pandemic has disrupted the way humans live, work and interact, leading scientists around the world to rapidly study the virus in an attempt to find the most effective treatment.

An interesting approach to the study of the COVID-19 virus is analysis of gene expression, as it is believed gene occurrences could play an important role in helping researchers understand the virus. The purpose of our project is to identify and investigate genes that make us more vulnerable to COVID-19, perform various analytical methods on data provided to us by *Immunogenomics Lab UNSW,* and relay our findings, conclusions and recommendations to the client in the format of a report.

Our team took advantage of highly robust methods such as Uniform Manifold Approximation Projection (UMAP) and Louvain Clustering to reduce the multicollinearity effect existent between hundreds of genetic variables, and enhance visualisation of both the B cells and CD8+ T cells datasets. The UMAP and Louvain algorithm improved the overall interpretation of the genetic variables by reducing the two datasets into five clusters, enabling quicker derivation of statistical results, which were displayed by volcano plots. Volcano plots were used to observe statistically significant gene expression differences between COVID-19 and healthy patients.

Finally, a conceptual model based on LASSO regression was built to subset our results further by determining which gene expressions were sufficient in predicting a COVID-19 status for a patient.

Out of the 26,361 genes that were analysed, we found that 12 genes from the B cells dataset and 7 genes from the CD8+ T cells dataset were statistically significant and fit for predictability. These gene expressions included IFI44, IGLC3, S100A8, HLA-DQB1, IFI27, IGLV3-10 and IFI44L.

Following our analysis and determined conclusions, we recommend that the client firstly considers the genes *Beyond Analysis* have found to be significant in future studies involving differences in gene expression of healthy and COVID-19 subjects. Following this, we suggest that the client reruns our analysis with a larger sample of patients that perhaps includes greater levels of variation to ensure results remain similar and consistent. Finally, we recommend the undertaking of further extended research into genes that play a vital role in the immune system, particularly IFI44, IFI44L, IGLC3, IGLV3-10 and HLA-DQB1.

4. Introduction

Immunogenomics Lab UNSW has the main goal of understanding how immune cells respond to pathogens and cancer through the combination of immunology, bioinformatics, molecular biology, and mathematics and statistics. The lab has been focusing on T cell and B cell responses to understand the mechanisms of the human immune system. One of *Immunogenomics Lab UNSW*'s past projects involves the investigation of single-cell multi-omics analysis for breast cancer T and B cells which analysed scRNA-seq using clustering methods [2].

Beyond Analysis has been tasked with conducting extensive research surrounding the relatively new COVID-19 virus. Our main objectives (2nd objective) involve identifying any differences in the gene expression of COVID-19 donors and healthy donors, and building a conceptual predictive model that can be implemented and tested by *Immunogenomics Lab UNSW* to predict whether a patient has COVID-19. We will be focusing on the B cells and CD8+ T cells of our patients. B cells play an important role in the immune system, responsible for producing antibodies against invasive pathogens [1A]. Similarly, CD8+ T cells are necessary for the body's immune defence. It has three mechanisms in place to kill virus/bacteria infected or malignant cells [1B]. At the end of our study, our findings will be presented to *Immunogenomics Lab UNSW*.

4.1. Background on COVID-19

On March 11 2020, the World Health Organisation (WHO) recognised the COVID-19 virus as a global pandemic; a virus which resulted in horrendous ramifications across the globe [3]. Not only has the disease led to worldwide economic strain, it has also been the cause of countless deaths in over 221 countries, as well as a major source of struggle within health care systems attempting to deal with the increasing number of infected patients seeking urgent medical attention and help. As a result of months of various lockdowns and social distancing, border and trade restrictions, millions around the world have lost jobs, and many companies, enterprises and businesses have been forced to either shut down or struggle to survive [4].

As of the present time, over 250 million COVID-19 cases and 5 million deaths have been confirmed across the globe. While the majority of regions worldwide are relievingly facing a decrease in COVID-19 cases and deaths, many of the issues mentioned above are still prevalent in one form or another. This highlights the necessity for the diagnosis and recovery of all those affected by the virus, in order to alleviate many of the health (both physical and mental), economic and social issues still circulating worldwide.

4.2. Link between COVID-19 and gene expression

According to an article that will be studied in more depth later on in our discussion of the link between gene expression and COVID-19, gene expression can be thought of and defined as "information stored in DNA that regulates how cells respond to changing environments … [including] controlling when and how much response is made against an invading virus." [4] Why are gene expression studies important? A number of risk factors have been found to contribute to the adverse effects of the COVID-19 virus, including but not limited to: ethnicity, smoking-status, physical health (eg. obesity), age, and host genetics. It has been proven that the expression level of genes that interact with the virus can in fact influence the level of one's susceptibility to COVID-19, and is causal for interindividual variability amongst patients. For example, genes that have been frequently mentioned in similar studies include ACE2, IL6 and TNF[6].

From thorough background research and analysis of an article entitled, *Expectations, Validity, and Reality in Gene Expression Profiling,* which attempts to provide an overview, analysis and in-depth discussion of the methodology of gene expression profiling and its development, our team has come to understand the extensive use of gene expression profiling specifically in biological research. As stated by the authors, it "has resulted in significant advances in the understanding of the molecular mechanisms of complex disorders, including cancer, heart disease, and metabolic disorders." [7]

Interestingly, gene expressions have also been scrutinised to understand the origins of long-term COVID-19 symptoms such as memory loss and coughing. A study revealed that certain individuals who had been infected with COVID-19 faced notable modifications to their immune systems [5]. Immune cells and gene expressions affected during the post-infection period were examined to determine causality of long-term COVID-19 symptoms present in some patients. The study assessed a group of patients from South Australia by inspecting and examining immune systems of 69 patients aged between 20-80 years. Of this group of participants, 47 had been currently facing a mild infection, 6 were facing a moderate infection, and 13 were in critical condition, attempting to recover from a severe form of COVID-19. After analysis of antibody responses, different types of immune cells, and an in-depth analysis of thousands of gene expressions within the blood (all via blood samples taken at 12, 16, and 24 weeks post-infection), results concluded that the immune systems of those who had been infected with SARS-CoV-2 faced considerable changes until a minimum of six months after being originally infected. These changes included significant dysregulation of gene expressions and a growth in the number of immune cells and antibodies [5].

Gene expression analysis continues to play a vital role in our ever-increasing understanding of the impact COVID-19 can have on individuals during the infection and post-infection phase. We hope that our study may, in some shape or form, help professionals to better understand the mechanisms underpinning COVID-19 complications, and establish molecular foundations for the development of therapies against the virus.

4.3. Link to Similar Studies on this Subject

There have been multiple studies on this subject surrounding COVID-19. For example, our datasets were collected from a study published in a *Nature* article on June 8, 2020 [6]. The objective of the study was to find differentially expressed genes in different cell types such as B cells and Neutrophils which could help distinguish between healthy and sick patients. Single cell sequencing was performed on 8 peripheral blood samples from 7 hospitalised patients and 6 healthy controls. The patients were all male, aged 20-80 years, whereas the control group were asymptomatic, consisting of 4 males and 2 females aged 30-50 years.

In our study, we implemented UMAP to visualise clusters of gene expressions on a 'Status' level. The new variable 'Status' was created to identify cell observations belonging to either a COVID-19 or healthy patient. In the referenced study, cell observations were observed at a patient level in order to account for patient variability. The study also had a larger sample of observations and thus was able to visualise cell observations at a cellular level, for example, most clusters could be categorised by cell type groups such as B cells, Neutrophils and Monocytes.

The referenced study performed statistical analysis such as observing average log fold change of differentially expressed genes by using a software known as QIAGEN Ingenuity Pathway Analysis (IPA) [6]. Genes were filtered through a Wilcoxon rank-sum test against the two-sided p-value < 0.05 threshold, and heat maps were created by plotting average log fold change for the remaining differentially expressed genes. Instead of creating heat maps and using the software, our team decided to create volcano plots (as recommended by our client), which clearly showcased average log changes, upregulation and downregulation of gene expressions, and most importantly, identified differentially expressed genes based on a log p-value threshold of 0.05.

5. Plan/Methodology

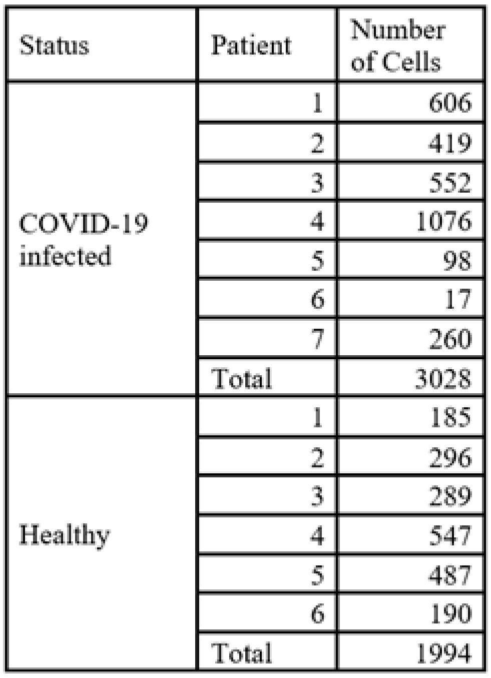
5.1. Assumptions

Before undertaking this study, there were certain assumptions which had to be made in order to make a fair interpretation of the results obtained from our analysis. A necessary assumption that we needed to consider when carrying out this project is that the only systematic difference across the two groups of patients is their COVID-19 status - in all other regards concerning their gene expression profile, the two groups are comparable (for instance, they are assumed to be sampled from the same population). This assumption is vital as it is required in order to be able to confidently state that any difference between COVID-19 and healthy patients is due to COVID-19 itself. Additionally, we have also made the assumption that the number of provided cells is not at all reflective of the true number of cells in each patient (ie. it is purely a sample). Another assumption made is that, while fairly intuitive, all COVID-19 patients actually have the COVID-19 virus (and thus, an accurate COVID-19 positive status), and likewise, all patients classed as 'healthy' are in fact healthy.

5.2. Exploratory Data Analysis

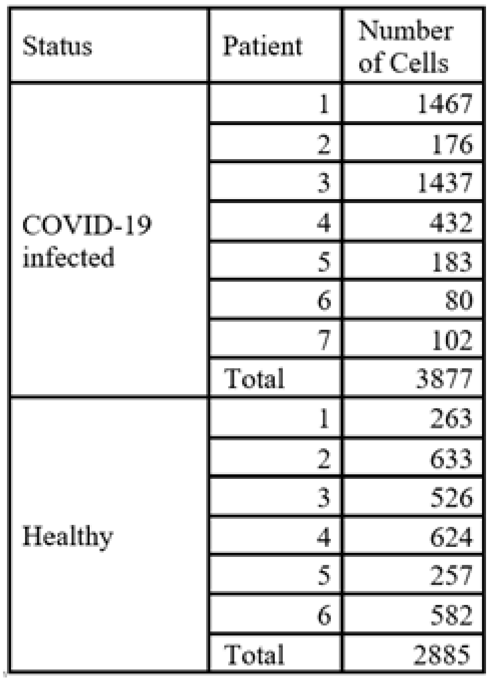
The beginning stages of our plan consisted of exploratory data analysis. *Beyond Analysis* were provided two datasets, both in RData format, by *Immunogenomics Lab UNSW* to assist with gene analysis and modelling. One dataset contains gene expression data for the B cells of both healthy and COVID-19 infected donors, while the other contains gene expression data for the CD8+ T cells of both healthy and COVID-19 infected donors.

The data for both datasets are categorised into two lists, one for healthy donors (labelled "list_healthy"), containing 6 elements to represent each of the 6 healthy donors, and one for COVID-19 infected donors (labelled "list_covid"), containing 7 elements to represent each of the 7 COVID-19 infected donors. Each element of the list consists of a separate matrix encompassing the gene expression of the donors' cells, with rows representing individual cells and columns representing different genes. Every cell across all matrices and datasets provided had a record of its gene expression from the same range of 26,361 genes. However, the number of cells provided in total for each patient's cell types differed, as summarised in the two tables below:



Status	Patient	Number of Cells
COVID-19 infected	1	606
	2	419
	3	552
	4	1076
	5	98
	6	17
	7	260
	Total	3028
Healthy	1	185
	2	296
	3	289
	4	547
	5	487
	6	190
	Total	1994

Figure 1: Cell count of B cells



Status	Patient	Number of Cells
COVID-19 infected	1	1467
	2	176
	3	1437
	4	432
	5	183
	6	80
	7	102
	Total	3877
Healthy	1	263
	2	633
	3	526
	4	624
	5	257
	6	582
	Total	2885

Figure 2: Cell count of CD8+ T cells

Overall, it can be noted that there were more COVID-19 cells provided for analysis in comparison to healthy cells for both types of cells. In total, there were 5022 B cells and 6762 CD8+ T cells across healthy and COVID-19 positive donors.

Additionally, on account of the significantly smaller values obtained in the original count data, and to allow modelling and analysis of proportional changes as opposed to additive changes, the original gene expression values were already normalised by *Immunogenomics Lab UNSW*. A log transformation of $log_{10}(1 + x)$ was applied, where x is the original count data.

Due to our team's expertise in the Python programming language as opposed to the R programming language for statistical and exploratory data analysis, the datasets were converted into four CSV files. One file was created for each cell type (B cells and CD8+ T cells) per healthy and COVID-19 infected donor for ease of supplementary analysis utilising Jupyter Notebooks, which operates primarily in Python, and Google Colab.

To prepare our data for our own analysis and modelling via methods such as logistic regression, a dataset combining the healthy and COVID-19 positive donors was created for both cell types, and a dummy variable was introduced to the datasets to indicate whether the patient was COVID-19 positive. We chose a result of 1 to represent COVID-19 positive patients and 0 to represent healthy patients.

Next, we manually filtered our data into a smaller subset for ease of further filtering utilising other methods. This involved the removal of genes that are present in 0 cells, and the identification of a threshold to filter the remaining genes, removing the genes which fall under the chosen threshold. Our cell-filtering choice was justified by the previously mentioned *Nature* article, which states "genes with very low counts across all samples provide little evidence for differential expression and interfere with some of the statistical approximations that are used later in the pipeline. They also add to the multiple testing burden when estimating false discovery rates, reducing power to detect differentially expressed genes" [6]. In this particular study, genes that were expressed in fewer than 10 cells were removed from the final count matrix. Additionally, cells were analysed for their gene reads as cells with low gene reads may have been unsuccessful in capturing their true gene composition, thus being faulty, and should be removed [8].

5.3. Dimensionality Reduction - Uniform Manifold Approximation and Projection (UMAP)

UMAP is a non-linear, dimensionality reduction method which takes a high dimension dataset and reduces it into simpler dimensions (typically 2-3 dimensions) without losing most of the information obtained from the original dataset. The B cells and CD8+ T cells dataset contained thousands of gene variables. Therefore, it was necessary to reduce our datasets into smaller datasets in order to maximise computational efficiency, visualisation and clustering capability.

UMAP is based on the kth nearest neighbourhood algorithm. A general topology of all data points from the original dataset is taken and the densities of different regions are assessed according to how close (or far) a set of data points are from each other. For example, lower density regions have larger gaps between data points and are more likely to be spread out on a 2 (or more) dimensional plane, while higher density regions have closer gaps between data points and are more likely to stay in place.

UMAP has numerous advantages over other notable methods including Principal Component Analysis (PCA) and T-distributed Stochastic Neighbour Embedding (tSNE). These advantages include computational efficiency and its ability to preserve the global and local structure of the original dataset [16]. One study compared computational methods on evolutionary data of a large set of 203,344 SARS CoV-2 genomic sequences. The performance and computational efficiency of three dimension reduction methods were compared: PCA, tSNE and UMAP. After conducting comparative analysis on a number of datasets it was found that "UMAP is the best-suited technique due to its stable, reliable, and efficient performance, its ability to improve clustering accuracy, especially for large Jaccard distance-based datasets, and its superior clustering visualization." [17]

5.4. Clustering (Louvain Method for Community Detection)

The Louvain clustering method was used to compartmentalise the UMAP results into several clusters. The Louvain algorithm is a hierarchical clustering algorithm which partitions a pre-computed neighbourhood graph into clusters with the highest modularity scores. Modularity measures the extent of interconnectedness (or the number of edges) within a community of data points in comparison to the overall connectivity of the graph [20].

In the first phase of the Louvain algorithm,

- Each node i is initially categorised as a distinct community.
- Modularity of a community is calculated when node i moves into the community of its neighbouring node j, such that there is an edge between node i and node j.

- The node is moved into the community that yields the greatest change in modularity when node i is included in the community.
- All steps are repeated until no node is able to move into a new community.

In the second phase, each community is summarised into a single node, whereby its modularity is the sum of all modularities within the community. First phase is repeated until communities are merged into super clusters with maximum modularity.

In our study, the algorithm reduced the network graph based on UMAP results into five clusters. Each cluster consisted of a subset of cell observations with similar gene expression results. Our team decided to use the Louvain method as numerous studies have noted its ease of implementation, computational speed and superior performance over other clustering methods [20].

5.5. Volcano Plots

After the conclusion of exploratory data analysis, we used various statistical and visualisation tools to identify key gene expression differences between healthy and COVID-19 infected donors. We attempted to pinpoint statistically significant genes by generating visualisations of volcano plots for the two cell types. A volcano plot is a form of scatter plot, 2-dimensional in nature, aptly named for its resemblance to the shape of a volcano. They are prominently utilised in gene analysis for their ease in identifying gene expressions with statistically significant changes amongst thousands of genes when comparing two differing experimental circumstances (such as COVID-19 patients and healthy patients, in our case) [9].

The x-axis of a volcano plot represents the log fold changes (ie. the magnitude of change) whilst the y-axis measures the statistical significance of the genes through the negative log of their p-values. Due to the application of a negative log transformation on the p-values, more significant changes will generate a smaller p-value and thus, a larger negative log of the p-value. In addition to selecting genes which produce more statistically significant change, a threshold is set on the log fold change value (the x-axis) to ensure only genes which reach a certain magnitude of change are taken into consideration. As a result, gene expressions which are deemed the most important when interpreting volcano plots are situated higher up in the plots and to the left or right side [10].

Our decision to incorporate volcano plots in our procedure was largely due to their ability to pinpoint and showcase important genes in a clear and understandable manner. It was further influenced and reinforced by a study carried out in China where the immune cell profiling of recovering COVID-19 patients was analysed. Volcano plots were utilised multiple times throughout the study to compare and highlight key differences between the cells of COVID-19 patients and healthy controls, when analysing them relative to each other [11]. This aligned well with our objective of identifying gene expression changes between healthy and COVID-19 infected donors.

We performed significance testing and hypothesis testing (eg. calculating p-values and the log fold change for two independent samples of scores, COVID-19 and healthy samples), in order to identify the most significant subsets of genes which distinguish healthy patients from sick patients. We implemented differential expression analysis by choosing a threshold of p-value < 0.05 as statistically significant, as it indicated strong evidence against the null hypothesis (ie. there was less than 5% probability the null is correct, and 95% probability of not rejecting the null hypothesis when it is correct). From a biological point of view, fewer genes change drastically and the significance difference observed at $p < 0.05$ is related to a possible whole human response to treatment. Additionally, we needed to consider patients as a batch.

5.6. Conceptual Model/ LASSO Regression

For the final stage of our project, we addressed our second objective of creating a conceptual predictive model which could be used to predict whether a patient has COVID-19 or not. In order to gather procedures on a predictive model for determining which patient has COVID-19 based on their gene expression, we required a logistic regression model. LASSO regression was trialled as a means of regularising the parameters in the dataset, allowing it to successfully run and produce meaningful results in the presence of a large number of predictors.

We aimed to construct a logistic regression model with predictors corresponding to the top 20 significant gene expressions identified in our analysis. We performed LASSO regression to compare the significance of these genes with each other, and we compared cross validation accuracy scores of the LASSO regression models with various penalty coefficients to determine the final models.

LASSO (Least Absolute Shrinkage and Selection Operator) regression can be defined as a specific variant of linear regression that uses the method of shrinkage - a process by which data values are "shrunk" towards a particular point of centre. The aim of LASSO regression is to essentially acquire the optimal subset of predictors from a model, 'optimal' meaning those that minimise the prediction error for the predetermined response variable. The LASSO method achieves this by imposing a penalty on the model based on the sum of the values of the coefficients' magnitudes, otherwise known as *L1 regularisation.* In some cases, coefficients can be minimised all the way to zero, which would allow its corresponding predictor to be eliminated from the model entirely. In the case of large penalties imposed by the LASSO regression method, values for coefficients will get shrunk closer to zero, which is what leads to the desired simpler model (which in turn accounts for better interpretability than a complex model). [12]

The LASSO regression method is admired by many due to its encouragement of a simple and 'neat' model (ie. a less-complex model with fewer parameters). LASSO regression is preferred over alternate regression methods particularly for models that display high amounts of multicollinearity, as well as for models where you may want to automate otherwise lengthy and complex processes during model selection, such as variable selection and/or the elimination of unnecessary parameters. [12]

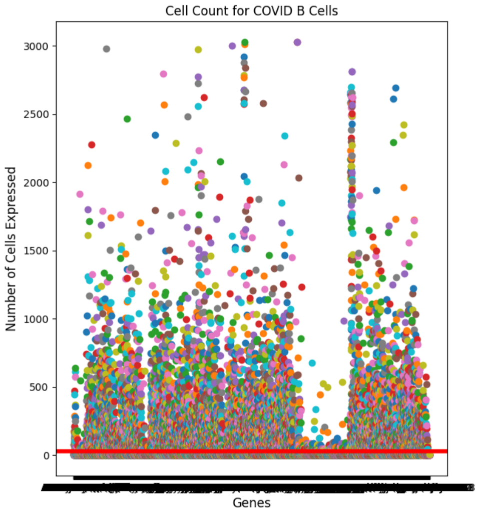
An example of where this method is used is in a study conducted approximately a decade ago that used "LASSO regression to detect predictive aggregate effects in genetic studies" [11]. In this study, LASSO regression was used to select "genetic markers and phenotypic features that are most informative with respect to a trait of interest" [11]. It is interesting to note that in the discussion of this particular study, the authors mention that while there are a large range of machine learning methods available at our disposal for genome-wide association studies (GWAS), regression methods that involve penalisation techniques (such as LASSO) are the most desirable due to their flexibility. In this case, LASSO was used to find common and uncommon variants using "sets of markers grouped by pathway and gene" [11], and there have been variations of the method used in an attempt to improve accuracy (for example, setting a p-value cutoff which will lead to a more sparse and interpretable LASSO model with fewer genetic variants included). In this particular study, LASSO regression was used to firstly select subsets of genetic variants for each pathway and gene, and then used to obtain an optimal model based on those chosen marker sets [12].

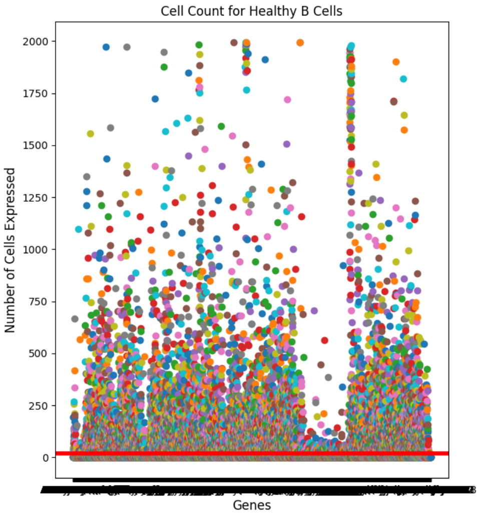
Particularly, in our case, we implemented L1 regularisation methods (i.e., LASSO) for logistic regression of the binary patient status on several sets of shortlisted genes, in order to quantify the effects of each gene in determining which patients have COVID-19. The success of the model produced was then measured using leave-one-out cross validation methods, rather than creating training and test data due to the small amount of data available on a patient level. The summarisation of the cell level data was done by calculating the mean of the gene expressions for each patient. We then compared the accuracies of different logistic regression models with different sets of shortlisted genes as the regressors and confirmed the most significant genes and their quantitative measures in determining which patients have COVID-19.

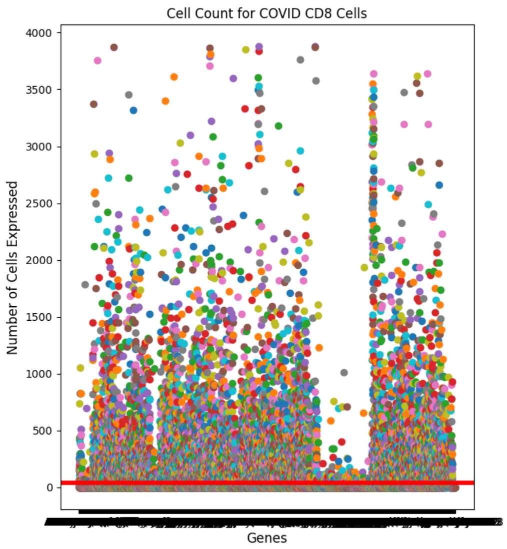
6. Findings/Results

6.1. Exploratory Data Analysis

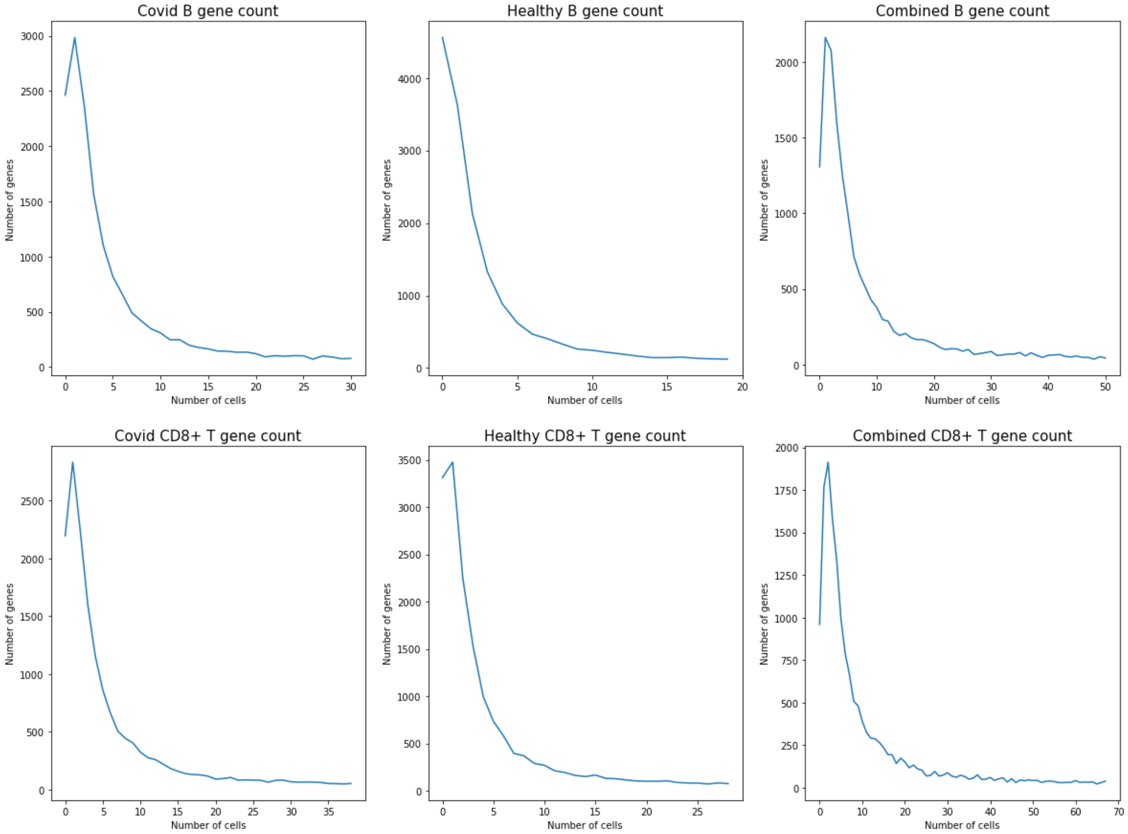
Upon initial analysis of value counts for each gene among the cells available for analysis, it became evident that several thousands of genes were not expressed in any cells at all, or in an insignificant number of cells. The plots below express the count of cells per gene, with a benchmark line at 1% of the total number of cells for each cell and donor type.







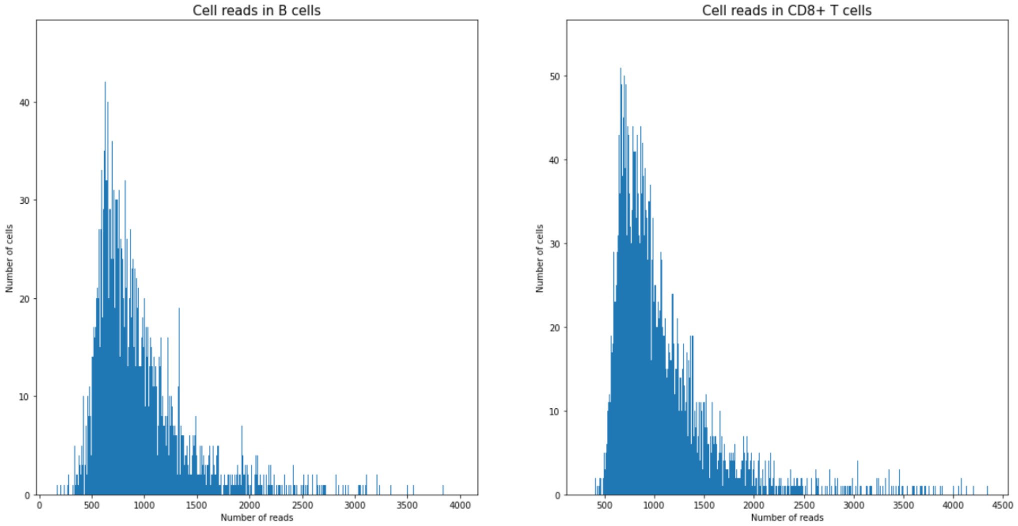
There are a large number of genes with a cell count below the 1% threshold, and it could be argued that these genes would generate noise as their values are relatively small compared to the rest of the observations. Therefore, a further analysis was conducted to better understand the distribution of these genes. The count distribution of genes expressed in less than 1% of the total number of cells being analysed can be observed in the plots below for each cell type:



The range of the x-axis differs for each gene count due to the varying sizes of the datasets provided for each cell type. It can be noted that across the combined datasets of healthy and COVID-19 infected patients for both B cells and CD8+ T cells, there were 1307 and 960 genes respectively that were not expressed in any cells at all, and 16,094 and 16,248 genes respectively that were expressed in less than only 1% of the total number of cells provided.

For our study, we chose the threshold of 1% to filter our dataset, after trialling several different thresholds, as a large percentage of genes fell under this threshold. This allowed a much smaller dataset to be created, generating less noise, which helped facilitate clearer analysis whilst still ensuring the threshold was not too large to prevent important genes from being eliminated in this study. The full count of genes under every threshold trialled is provided in Appendix A. It should be noted that the filtering of the dataset occurred after the COVID-19 positive and healthy donor datasets were combined. As a result, a relatively small percentage of genes were eliminated which would not have been eliminated had the filtering been carried out prior to merging the two datasets for each cell type.

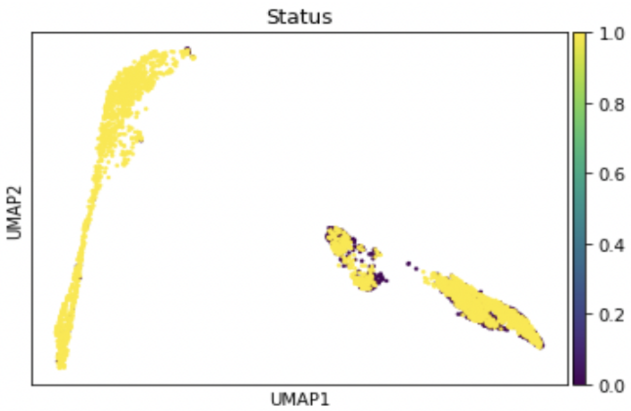
Alongside analysing the count of cells that genes were expressed in, the total count of genes that cells had expressed in them were also examined. The histogram below summarizes the cell reads for both cell types:



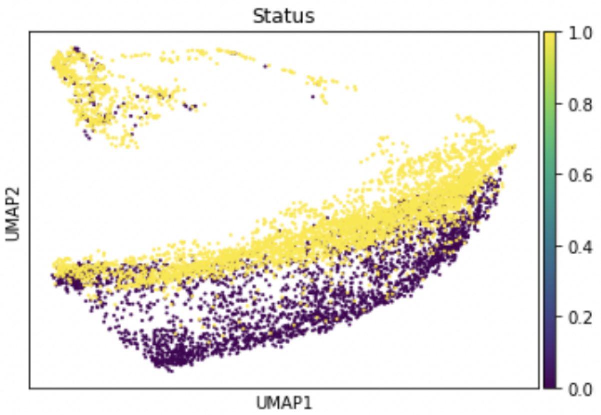
The histogram was created with 1000 intervals within which all the cells of the two datasets were partitioned into. It was found that the minimum number of gene reads from a cell in B cells was 160 and the minimum number of gene reads from a cell in CD8+ T cells was 409. Due to the relatively high minimum, no cells were deemed insignificant or faulty in either dataset and were all retained for further analysis.

6.2. Dimensionality Reduction - Uniform Manifold Approximation and Projection (UMAP) and Louvain Method for Community Detection

UMAP was implemented on the normalised datasets for B and CD8+ T cells:



UMAP Plot (B cells)



UMAP Plot (CD8+ T cells)

UMAP clusters were observed on a 'Status' level (i.e. '1' for a patient with COVID-19 and '0' otherwise). Yellow dots correspond to cell observations from COVID-19 positive patients and purple dots correspond to cell observations from healthy patients. Our team also produced UMAP results on a 'Patient' level, however we eventually decided to not consider patient variability within our analysis to avoid the risk of losing information that was captured in the original dataset. Thus, those results are not shown.

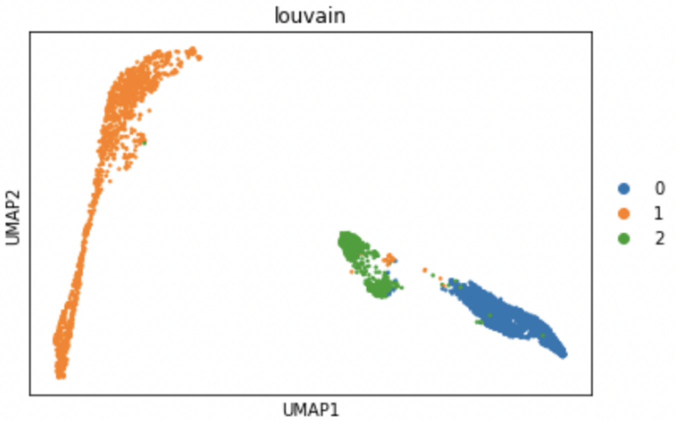
Our team experimented with different values for two different parameters: 'min_dist' and 'n_neighbours'. These parameters were used to control the local and global structure in the final projections. 'Min_dist' refers to the minimum distance between data points in low-dimensional space, with higher values summarising the broader global structure of the data. 'N_neighbours' refers to the number of nearest neighbours used to determine the final graph, with higher values summarising the broader structure of the dataset whilst losing finer detail [21].

After experimenting with various thresholds, it was decided that for both datasets, 'min_dist' would be set as 0.1 for both the B cells and CD8+ T cells dataset. It also turns out that setting minimum distance to values near 0.1 is common practice, particularly in population genomic studies. One article mentioned how "studies varied in their parameter selection, but … the minimum distance was usually 0.1 < MD < 0.5; values of MD close to 0 create very tight clusters, which can be appropriate for downstream processes such as cluster analysis." [22]

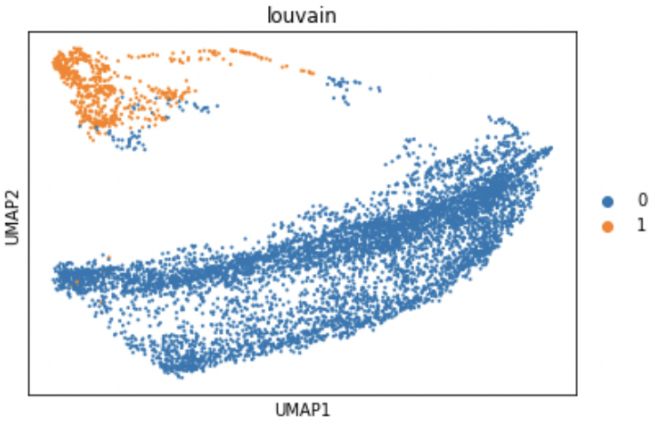
As for 'n_neighbours', picking an appropriate value was less obvious and required a bit of trial and error. Eventually, we decided that the parameter would be 20 for the B cells dataset and 30 for the CD8+ T cells dataset.

As shown above, the UMAP algorithm identified three groups of cell observations from the B cells dataset. The left cluster has mostly cell observations from COVID-19 patients whereas the smaller clusters consist of a mixture of cell observations from both healthy and sick patients. The second picture consists of two clusters of cell observations. Unsurprisingly, the bottom hand cluster provided the most interesting results in the analysis section of our study because it had an abundant mixture of cell observations from both types of patients.

Finally, the Louvain clustering method was used to definitively group UMAP results into individual clusters. Overall, two datasets were reduced into five smaller datasets/clusters.



Clusters 0, 1 and 2 (B cells)

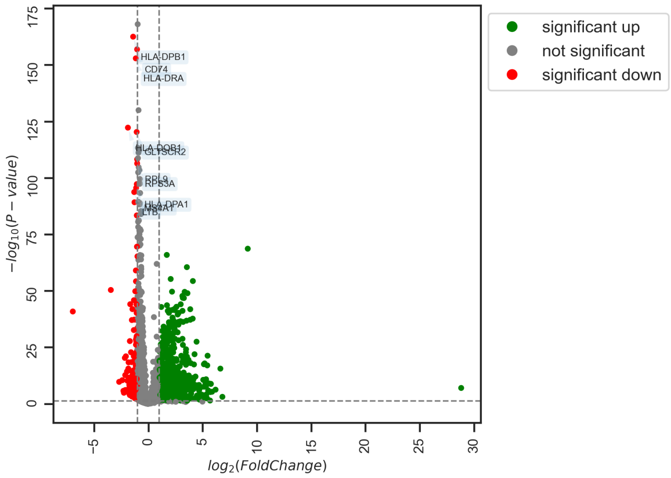


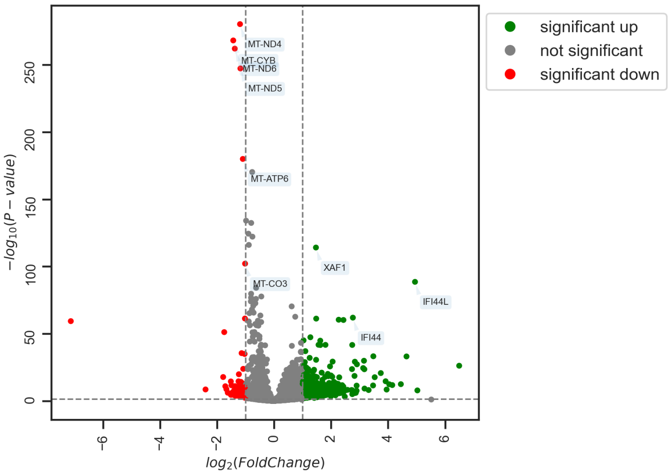
Clusters 0 and 1 (CD8+ T cells)

6.3. Volcano Plots

6.3.1. Initial Volcano Plots for Data without UMAP Clustering

The volcano plots generated by our team, for both B and CD8+ T cells, are as seen below:





The plots were generated via the bioinfokit package in Python, using the combined dataset of the COVID-19 and healthy B cells to create the plot (on the left) and the combined dataset of the COVID-19 and healthy CD8+ T cells to create the plot (on the right). In these volcano plots, the most upregulated genes (“significant up”) are towards the right, the most downregulated genes (“significant down”) are towards the left, and the most statistically significant genes are towards the top. The thresholds chosen to determine the most significant genes were the default log fold-change thresholds of -1 and 1 (dotted vertical lines) on the x-axis and an adjusted log P-value threshold of 0.05 (dotted horizontal line) on the y-axis. The two sided test for 2 independent samples is assumed to have 2 unequal variances. We also chose the threshold of the absolute value of the log fold change value of > 1 as significant, because it appears to eliminate background noise and it is a common choice for a sensible threshold (refer to Extended Data Fig.5b from article [13] and from blogs posts [14A], [14B] and [15A], [15B]). The grey points are in the region of low log fold change significance between -1 < logFC < 1, however some points have low p-values, indicating some significance. We decided to consider genes as significant if both |logFC| > 1 and p values < 0.05. Thus, the red and green points meet the criteria of |logFC| > 1 and p values < 0.05, rendering them as statistically significant.

The combined dataset of COVID-19 and healthy B cells has 5316 genes that are statistically significant under the 5% level, whereas the combined dataset of COVID-19 and healthy CD8+ T cells has 4329 genes. In the combined dataset of the COVID-19 and healthy B cells, we selected the top 10 genes that have the lowest p-values, plotted it, and compared it with the top 10 most significant genes from the combined dataset of the COVID-19 and healthy CD8+ T cells.

Below are the 10 most significant genes of the combined datasets of COVID-19 and healthy patients, filtered by just the p-value (refer to Appendix B for labelled volcano plot):

B cells: ['MT-CO1', 'HLA-DPB1','CD74', 'HLA-DRA', 'RPL13A', 'HLA-DQB1', 'GLTSCR2', 'RPS4X', 'RPL3', 'RPS3']

CD8+ T cells: ['MT-CO1', 'MT-ND4', 'MT-CYB', 'MT-ND6', 'MT-ND5', 'MT-ATP6' ,'MALAT1', 'MT-CO2', 'RPS3', 'RPS4X']

Based on the volcano plots and the given results, we have genes ‘MT-CO1’ and ‘RPS3’ that are very statistically significant with the lowest p-values in both B cells and CD8+ T cells.

Moreover, below are the top 20 most significant genes of the combined datasets of COVID-19 and healthy patients, filtered by p-value and log2 fold change:

B cells: ['CD74', 'HLA-DPB1', 'HLA-DRA', 'GLTSCR2', 'IFI27', 'MS4A1', 'AC009501.4', 'HLA-DQB1', 'RPS3A', 'RPL9', 'IFI6', 'HLA-DPA1', 'IGHG4', 'MX1', 'LTB', 'CD79B', 'CD79A', 'XBP1', 'ELL2', 'CALR']

CD8+ T cells: ['MT-CO1', 'MT-ND4', 'MT-CYB', 'MT-ND6', 'MT-ND5', 'MT-ATP6', 'XAF1', 'IFI44L', 'MT-CO3', 'IFI44', 'MX1', 'OAS3', 'PARP9', 'SUN2', 'EIF2AK2', 'IFIT3', 'EPSTI1', 'DTX3L', 'IFI6', 'ISG15']

B cells and CD8+ T cells: ['MX1']

Based on the volcano plots and the given results, we have gene ‘MX1’ that is very statistically significant, with the lowest p-value in both B cells and CD8+ T cells. MX1 is a protein encoding gene which plays a role in the cellular antiviral responses of the body. It has antiviral activity against several RNA viruses and some DNA viruses. It is associated with the flu and viral encephalitis [18], the latter of which has been noted as a severe neurological complication that can be caused by COVID-19 [19].

Note: The top 20 significant genes are heavily based on the p-values and not so much the log fold change [23A]. This is mainly due to the fact that the p-value already takes into account the fold change. In order for a gene to be differentially expressed at a statistically significant level, it needs to be quantitatively different between the 2 biological groups, COVID-19 and healthy, and it has to be relatively consistent for each gene [23B]. Though we focus on limiting the number of genes for variable selection by placing the threshold log fold change of > 1 or < -1.

Correct Multiple Testing

After filtering out genes that are expressed in less than only 1% of the total number of cells provided, we obtain 10267 genes. By performing 1 hypothesis test, we have a 5% chance of declaring false positives. We realised that just doing hypothesis testing for 1 gene is insufficient and would not necessarily help in determining whether a patient is susceptible to COVID-19. Instead, we needed to explore all the genes and their hypothesis tests by undergoing a method called **Correct Multiple Testing**. There are 2 types of corrections (given below), Bonferroni and Benjamini/Hochberg correction, though we mainly focus on Benjamini/Hochberg correction. Thus, when testing for potential differential expressions, each gene is considered independently from one another.

Since we have 10267 genes, we perform 10267 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we would expect 514 genes to be deemed "significant" by chance.

The Bonferroni correction adjusts the p-values by taking into account the number of genes and the position of the p-values before correction, i.e.:

$$Bonferroni\ Correction\ =\ pvalue_{before} * (N - Rank + 1)$$

Similarly, Benjamini and Hochberg correction also utilises the number of genes and the position of the p-values before correction but with a different formula, which is:

$$Benjamini - Hochberg\ Correction\ =\ pvalue_{before} * \frac{N}{Rank},$$

where for both formulas:

- $pvalue_{before}$is the p-value calculated before correction
- $N$is the Number of genes
- $Rank$is the designated position of the p-values before correction. E.g. The smallest p-value before correction is of *Rank = 1*. The second smallest p-value before correction is of *Rank = 2*, and so on, while the largest p-value before correction is of *Rank = N*.

The reason why we implemented correct multiple testing is because it is important to correct the p-value of each gene when performing a statistical test on a group of genes. Multiple testing correction adjusts the individual p-value for each gene to keep the overall error rate (or false positive rate) to less than or equal to the user-specified p-value cutoff or error rate [24].

	B cells	CD8+ T cells
Number of statistically significant genes under the 5% level	5291	4324
Bonferroni correction to control FWER	2444	1699
Benjamini and Hochberg to control FDR	4684	3485

Based on the table above, Type I error rate is calculated as 5291/10269 = 0.52. After Bonferroni, we have Type I error rate of 2444/10269 = 0.24 and after performing Benjamini/Hochberg correction, we obtain Type I error rate = 4684/10269 = 0.27.

The Type II error rate (false negative) is 1-0.36 = 0.64. After Bonferroni correction, the rate of false negatives is 1-0.11= 0.89, whereas with Benjamini/Hochberg correction, the rate of false negatives is 1-0.27 = 0.73.

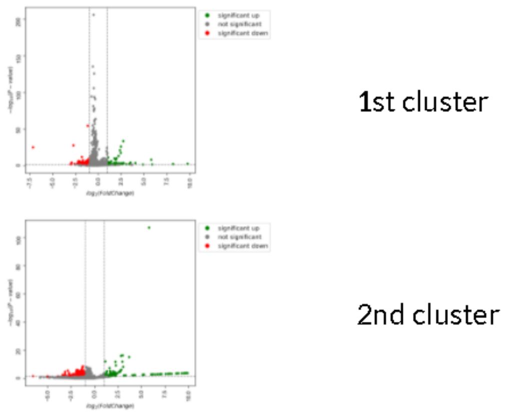
False Positives vs. False Negatives

The number of False positives (Type 1 error) is equivalent to the number of genes that are declared significant and when the null hypothesis is true, i.e. when we reject the null hypothesis [25]. This occurs when we reject the true null hypothesis incorrectly or when we accidentally reject the null hypothesis 5% of the time. In other words, our findings are significant when in fact they have occurred by chance. We will only get a 5% chance overall of falsely rejecting the null hypothesis. Therefore using the threshold p-value of 0.05 indicates that we are willing to accept a 5% chance that we are wrong when we reject the null hypothesis [26].

The number of False negatives (Type II error) is equivalent to the number of genes that are declared non-significant and the alternative hypothesis is true, i.e. where we fail to reject the null hypothesis when it is not true. A Type II error occurs when a researcher fails to reject a null hypothesis which is really false. In other words, the researcher concludes there is not a significant effect, when in fact there is.

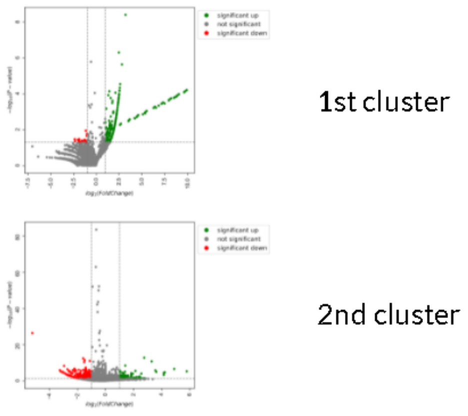
Even though using a lower p-value of 0.01 would reduce the type I error, this however, would give us a lower chance of detecting a true difference if one really exists as there is a trade off between Type I error and Type II error. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists (thus risking a Type II error).

6.3.2. Volcano Plots for Data Filtered by UMAP


1st cluster

2nd cluster

Volcano plots for B cells


1st cluster

2nd cluster

Volcano plots for CD8+ T cells

Benjamini and Hochberg is a more suitable method for correcting p-values than Bonferroni correction in gene expression analysis because it controls the False discovery rate, that is, the expected proportion of false positives among the variables for which we claim the existence of a difference [27]. For example, if FDR controlled to 5% and 20 tests are positive, we would expect only one of these tests to be a false positive [28].

Choosing Benjamini and Hochberg correction over Bonferroni correction was due to the fact that Bonferroni controls the FWER (the probability of making a Type I error) while Benjamini and Hochberg controls FDR (controls how many of the Type I errors we make in proportion to the true positives). Even though Benjamini and Hochberg is more lenient than Bonferroni, the FDR has a higher power because it has a higher Type I error rate. Therefore it is still a better "contender" than Bonferroni correction. It has also been widely used in cases where a large number of hypotheses are simultaneously tested.

	B cells			**CD8+ T cells**	
	1st cluster	2nd cluster	3rd cluster	1st cluster	2nd cluster
Number of statistically significant genes under the 5% level	2404	1958	1262	2955	1420
Bonferroni correction to control FWER	False positives = 289	False positives = 442	False positives = 61	False positives = 2373	False positives = 115

Benjamini and Hochberg to control FDR	False positives = 0	False positives = 0	False positives = 0	False positives = 0	False positives = 0

The null hypothesis is assumed to be correct, that is, identifying statistically significant genes to determine which patient is susceptible to COVID-19. Then the Type errors (I and II) goes as follows:

Type 1 error (False Positive): The patient who is immune to COVID-19 after identifying statistically significant genes.

Type II error (False Negative): The patient who is susceptible to COVID-19 after identifying non-significant genes.

Interpretation of False Positives = 0

Thus, by looking at the Benjamini and Hochberg and control FDR row, we interpret False positives = 0 as "there is a 0% chance that the patient who is immune to COVID-19 after identifying statistically significant genes". This was the optimal outcome for us as, even after identifying statistically significant genes and which patients are immune to COVID-19, we cannot be sure the patients were immune as a result of being a COVID-19 patient given a cure or a healthy patient being given a cure. Thus, we are better off having a higher percentage chance for a Type II error by having Type 1 error as low as possible.

Number of genes (N)	**Probability of getting one or more false positives (Type 1 error) by chance** $100(1 - 0.95^N)$
1	5%
2	9.75%
10	40.13%
20	64.14%
100	99.41%

By conducting the hypothesis test: null hypothesis $H_0: \beta_i = 0$ against the alternative hypothesis, $H_1: \beta_i \neq 0$ for the model, we reject the null hypothesis under a 5% level for 1 gene. Since this occurs for multiple genes, we conduct multiple hypothesis tests. Across 1 test, there is a 5% chance of declaring false positives, and for 2 tests, there is a 10% chance. So, the chance of declaring false positives increases with more tests (referring to the table above).

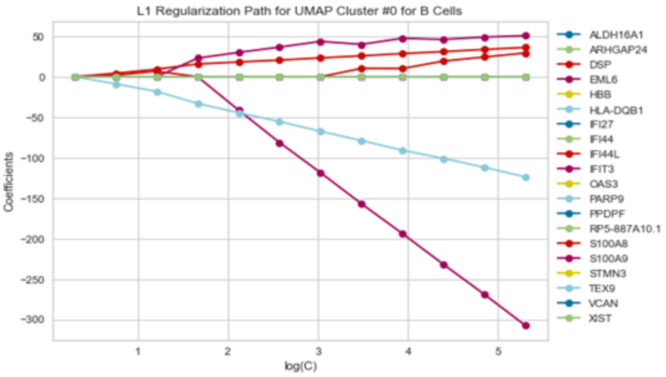
We conclude that there is a small number of genes in the COVID-19 patient cells that are statistically significant compared to the genes in Healthy patients cells. Thus, these genes show more variability in the COVID-19 patients cells rather than the Healthy patient cells. The genes are also more dominant in the COVID-19 patient cells, helping us to determine that COVID-19 patients are more useful for the creation of the predictive model, and making it easier to predict which patient has COVID-19 based on the statistically significant gene expression.

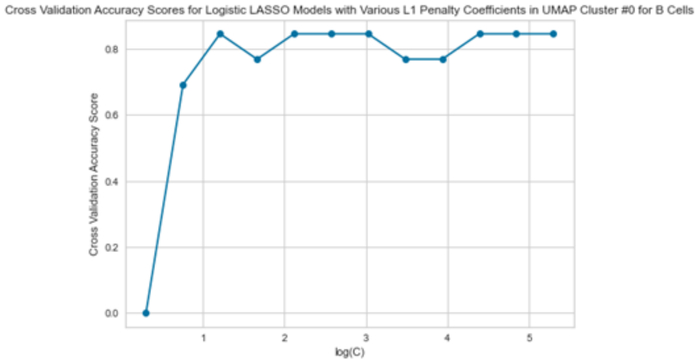
	Cluster	**Top 20 statistically significant genes (lowest p-values and |logFC| > 1)**
B cells	Cluster 0	['HLA-DQB1', 'IFI44L','DSP', 'S100A8', 'XIST', 'S100A9', 'IFI44', 'IFIT3', 'EML6', 'PARP9', 'VCAN', 'ARHGAP24', 'OAS3', 'PPDPF', 'HBB', 'IFI27', 'RP5-887A10.1', 'STMN3', 'ALDH16A1', 'TEX9']
	Cluster 1	['IFI27','CLU', 'CD3D', 'F13A1', 'HIST1H2AC', 'ANAPC1', 'BZW2', 'IGHV3-9', 'FASTKD1', 'SIGLEC1', 'CHSY1','TCTN3', 'IGLV1-40', 'COLGALT1', 'MKKS', 'VWA8', 'ITGB3', 'PTP4A3','ZCCHC8', 'IGLV3-10']
	Cluster 2	['IL32', 'IFI44L', 'MX1', 'DSP', 'S100A8', 'IGLC2', 'SPON2', 'PDIA4', 'CD300A', 'SOCS3', 'MATK', 'IFI44', 'LINC00623', 'CHST2', 'TAPSAR1', 'PILRB', 'PPP1R12C', 'P4HA1', 'METRNL', 'MYOM2']
CD8+ T cells	Cluster 0	['MAPKAPK2', 'RP11-284N8.3', 'ZFAND5', 'SRSF9', 'CASP8AP2', 'IFI44', 'BRD7', 'HECTD4', 'ATXN2L', 'IPO5', 'IFIH1', 'GIMAP8', 'MON2', 'SLC15A4', 'ENDOD1', 'PAIP2', 'ANO6', 'ITPR1','NXF1', 'IGHG4']
	Cluster 1	['XIST', 'IGLC3', 'PDZD4', 'ZNF683', 'RASSF1','S100A8', 'BCR', 'ZNRD1-AS1', 'HLA-DQB1', 'CMKLR1', 'PCNXL3', 'IGHG1', 'AGAP3', 'HCP5', 'NPC1', 'PILRB', 'AGAP1', 'MCM7', 'MED12','PLXND1']

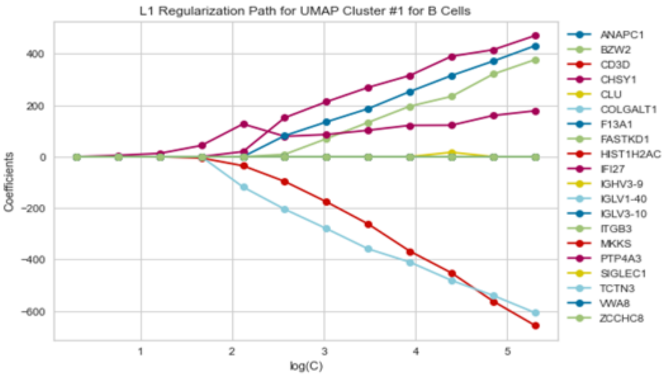
6.4. Conceptual model/ LASSO Regression

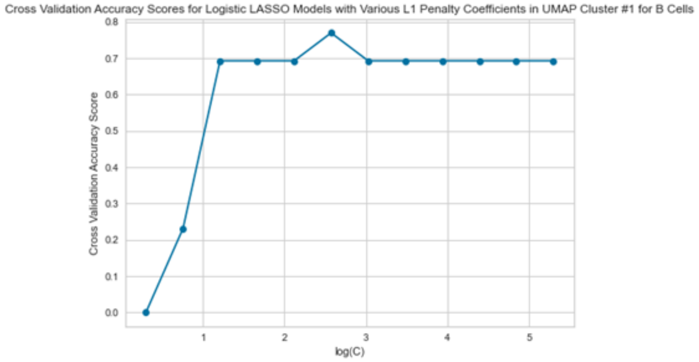
After summarising the cell-level data by taking the mean of the gene expressions for each patient, we implemented logistic LASSO regression and leave-one-out cross validation methods on the most significant genes generated from the previous step to further determine the quantitative effects of these genes. The dependent variable for the models is the binary status of the patient, i.e. ‘1’ for a COVID-19 patient and ‘0’ for a healthy patient.

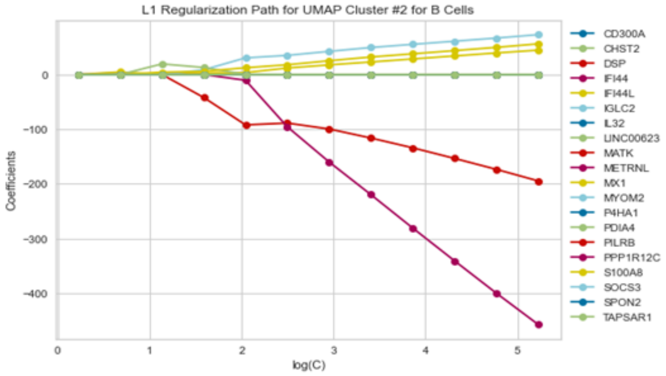
The plots below represent the results of the logistic LASSO regressions with different penalty coefficients (i.e. ‘C’) for the B cells. The independent variables for this model are the twenty significant genes filtered by the volcano plots. The first plot illustrates how smaller 'C' values (x-axis) are associated with a stronger penalty, resulting in regression coefficients (y-axis) being closer to zero. The second plot summarises the performance of each model of different penalty coefficients using the leave-one-out cross validation accuracy scores. This step was utilised for the purpose of choosing the optimal penalty coefficients. Each pair of plots represent the results for each UMAP cluster.

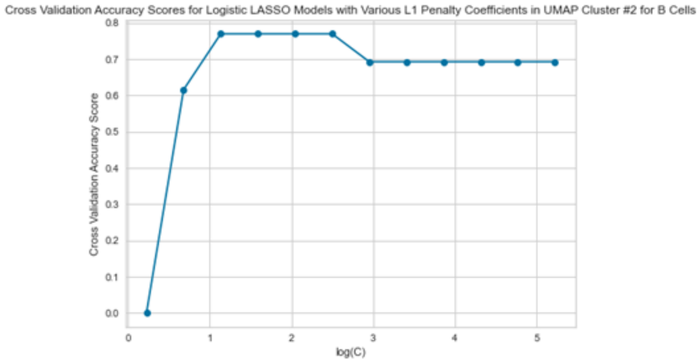




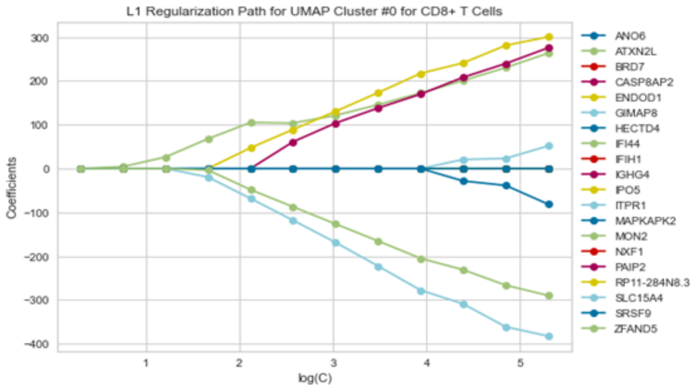


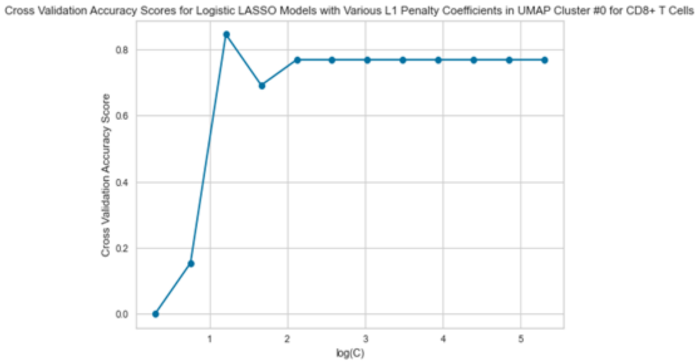


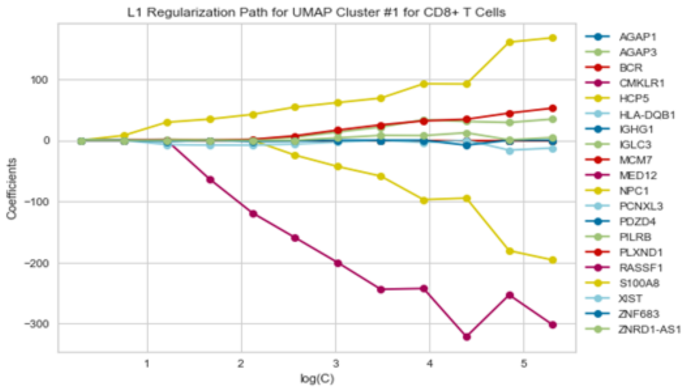


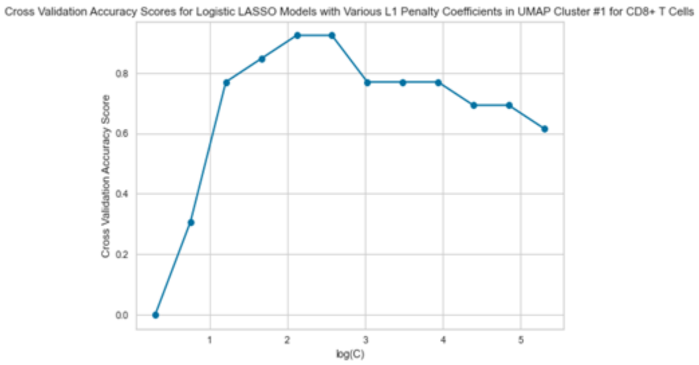


Similarly, the results for the CD8+ T cell clusters are presented below.









The penalty coefficients being compared were in the range of $[0,\ 9 * 10^{5}]$. For each cluster, we chose the simplest model with the highest accuracy. This is due to our objective in distinguishing the significant genes in a finer way. The specific values varied across multiple runs due to the stochastic nature of the learning algorithm. However, it could be argued that the essence of the system remained the same across trials.

7. Discussion

7.1. Results obtained

The tables below summarise the logistic LASSO regression results for each UMAP cluster, for both B cells and CD8+ T cells.

	Gene	Logistic LASSO Regression Coefficient
Cluster 0 - B Cells	HLA-DQB1	-18.255
	IFI44L	9.319
	S100A8	6.862

Cluster 1 - B Cells	BZW2	9.030
	CD3D	-96.323
	COLGALT-1	-204.355
	IFI27	79.167
	IGLV3-10	81.551
	PTP4A3	151.682
Cluster 2 - B Cells	IFI44L	1.644
	PDIA4	19.398
	S100A8	4.282

	Gene	**Logistic LASSO Regression Coefficient**
Cluster 0 - CD8+ T Cells	IFI44	26.210
Cluster 1 - CD8+ T Cells	IGLC3	1.305
	MCM7	1.465
	RASSF1	-120.245
	S100A8	42.054
	XIST	-7.727
	ZNF683	-1.742

The regression coefficients could help us gain understanding regarding how the expression of each gene affects the probability of a patient having COVID-19 or not. Based on the tables above, it could be seen that a number of genes have positive correlations with the binary patient status, meaning that the presence of these genes would increase the patient's probability of having COVID-19, ceteris paribus. On the other hand, several other genes were predicted to have negative correlations with the patient status. This could be interpreted as patients with these genes present would be less likely to have COVID-19, holding other factors constant.

Between clusters, the magnitudes of the regression coefficients could vary to an arguably great extent, due to different choices of the penalty coefficients. Nevertheless, the magnitudes of the regression coefficients could help us to determine and compare the significance of each gene relative to the other genes in its respective cluster.

In terms of the functions of these genes, we found that the majority of the statistically significant genes are associated with the immune system and immune responses. Interestingly, some of them (such as IFI44, MCM7, RASSF1) are phenotypes for increased vaccinia virus infection, which has been noted to generate antibodies with good activity towards COVID-19 variants [31]. The full list of individual gene functions and associated phenotypes can be found in Appendix C.

7.2. Problems encountered and how they were approached

There were several problems we encountered throughout our analysis. Prior to beginning any analysis, there was one problem we had to consider:

- The age/gender of patients (the metadata):
  - We had considered old patients as outliers, but we believed removing them would be bad practice for the sake of producing a better model or statistically significant results.
  - Another reason is that although old patients have extreme values/variability, we did not remove them because they are a natural part of the population. We believed that tampering with the number of patients in the dataset would lead to unrealistic results.

Once we began analysis, problems that were encountered during our exploratory data analysis stage included:

- The redundancy of using PCA because UMAP was also used. Both methods are dimensionality reduction techniques which would produce similar results. However, we decided UMAP was a better choice for our particular analysis as there is overwhelming evidence to suggest its advantages over PCA such as computational speed.
- We noticed that some genes, such as MX1, were seen as significant under the 5% level when UMAP was not implemented. However, when UMAP was implemented, it was less significant. There is a good chance that we may have lost some information along the way by not selecting the appropriate number of nearest neighbours and/or minimum distance between data points. This meant that some genes were clumped with more/less genes than they should have, influencing the final results shown in the LASSO modelling and volcano plots section.

Following exploratory data analysis, issues we ran into during the creation of our volcano plots were:

- The mean of some genes in healthy patients was zero in the combined dataset. This was an issue due to it being required in the denominator when calculating fold change for our volcano plots. The issue was approached by setting the value of the mean to an arbitrary small number close to zero (our choice was 0.0001).
- The mean of genes in the COVID-19 patients is smaller than the mean of genes in the healthy patients. Therefore the Fold Change calculation: mean(COVID)/mean(Healthy) - 1 gave a negative result. E.g. 7SK; mean(COVID)/mean(healthy) - 1 = 0.063738/0.109328 - 1 = -0.41700 and logFC(-0.417) = NaN, so we used mean(COVID)/mean(Healthy) instead.

During the last stage of our analysis, there were three problems we addressed whilst building the LASSO regression model:

- The first problem regarding the use of the LASSO method was how the number of selected variables could not exceed the number of observations [29]. Since we implemented LASSO to the patient-level data with only 13 observations, the models would apply penalties such that the maximum number of variables with non-zero regression coefficient was 13. While this could be advantageous in terms of the feature selection function, it could be argued that this limiting feature could result in loss of important information. We approached this problem by implementing LASSO to smaller and filtered groups of genes. Hence, the limitation of 13 variables overall was extended to 13 variables per cluster, thus resulting in an arguably more reliable result.

- The second problem related to LASSO regression was finding the optimal regularization/penalty coefficient. Choosing a penalty coefficient that is too strict would result in insufficient amount of resulting variables. On the other hand, a loose penalty coefficient would result in the logistic LASSO regression being statistically equivalent to a regular logistic regression. Therefore, we approached this by comparing the leave-one-out cross validation accuracy score of each model with different penalty coefficients, and chose the model with the highest accuracy score.

- LASSO regression also suffers from a grouped variables problem [29]. Several genes that have similar biological pathways would have high multicollinearity with each other and could be thought of as forming groups [30]. Ideally, a gene selection method should include every gene in the group if one of the genes in the group is selected [29]. However, LASSO algorithms would only select one variable from the group for such a scenario [29]. Therefore, we suggest that our client, who has a stronger background in bioinformatics, to also analyse the genes that have similar biological pathways to the significant genes we nominated.

7.3. Recommendations for client

Based on our findings, there are a few recommendations we can provide to our client. Firstly, we recommend that the *Immunogenomics Lab UNSW* consider the significant genes as mentioned in the LASSO output for future studies around the gene expression differences between healthy and COVID-19 patients. We also recommend conducting further research on genes that share the same biological pathways with the genes we considered significant. Moreover, it is also recommended that *Immunogenomics Lab UNSW* conduct supplemental research on genes which play a role in the immune system, such as IFI44, IFI44L, IGLC3, IGLV3-10 and HLA-DQB1. Lastly, in the interest of analysing consistency, we also recommend that our client repeat the pipeline of methods above with a larger sample of patients, as well as a higher variability between patients, and fine-tune parameters accordingly. For example, it would prove useful to include female COVID-19 patients in the sample since the data we obtained from COVID-19 patients were all male.

8. Conclusion

Overall, we have implemented several statistical methods such as UMAP, Louvain clustering method and volcano plots to verify that there are genes that make certain individuals more susceptible to COVID-19 than others. We had also developed predictive methods based on the findings using logistic LASSO regression and leave-one-out cross validation methods.

We have concluded that the genes that are associated with the immune system are the most impactful and the most beneficial genes in verifying that an individual is more susceptible to COVID-19 than others. We hope our findings prove to be valuable to the *Immunogenics Lab UNSW* in their analysis surrounding the COVID-19 virus and its impacts.

9. Technical appendix

9.1. Peer Review Feedback

1. *"The draft report is well presented and expertly steps the reader through the team's process, especially in some areas which are more nuanced to the field of bioinformatics. An introduction to what B cells and CD8+ T cells are and why they are considered over other cell types as a good indicator of gene expression would help to highlight the predictive power of your model in the given context. Without seeing your results from modelling it is hard to gauge your model's predictive power, however, it may be interesting to compare a variety of models against the baseline presented. In saying this there is always the question of predictive power vs interpretability, which would need to be balanced in achieving your desired goals and considering the limitations on number of patients in the dataset. Also, it appears the mission statement may be somewhat different to the scope as the mission states determining those susceptible to COVID-19 as the objective, while the scope says its aim is to predict those who have COVID-19."*

- Resolution: We added 1-2 sentences introducing what B cells and CD8+ T cells are to address the first half of this feedback. However, it is difficult to resolve the second half of this statement. We were simply provided with the B cells and CD8+ T cells - we did not really consider them over other cell types.

- Resolution: Even though Ridge regression, elastic net, KNN (k nearest neighbours) models are not presented in the report, we believe that these models will converge to the same results. LASSO was not only recommended by our supervisor, who has a broad knowledge of specific models, but is in general more effective when shrinking insignificant coefficients automatically to 0, in comparison to Ridge or other models.

- Resolution: We would refer to the "no free lunch theorem": neither ridge regression or the lasso universally outperform the other. The choice of method should be data driven. The lasso has better interpretability since it can lead to a sparse solution. Ridge is more likely to choose all the predictors, hence losing the interpretability but often improving predictability. There is always a trade-off between interpretability and predictive power. Some would say that knowing the predictive performance on a dataset is sufficient, but knowing "why" can help us learn more about these problems, the data and the reason why a model might fail. Interpretability is needed because the model should also be able to explain how it came to the prediction as a correct prediction only partially solves the original problem [32]. Limitations on the number of patients in the dataset already answered under 7.2 Problems encountered and how they were approached.

- Resolution: We mentioned one objective in the mission statement and another objective in the scope, so we have two objectives. The objectives were mentioned in both parts but we decided to include "Our main objective (2nd objective)..." for more clarity.

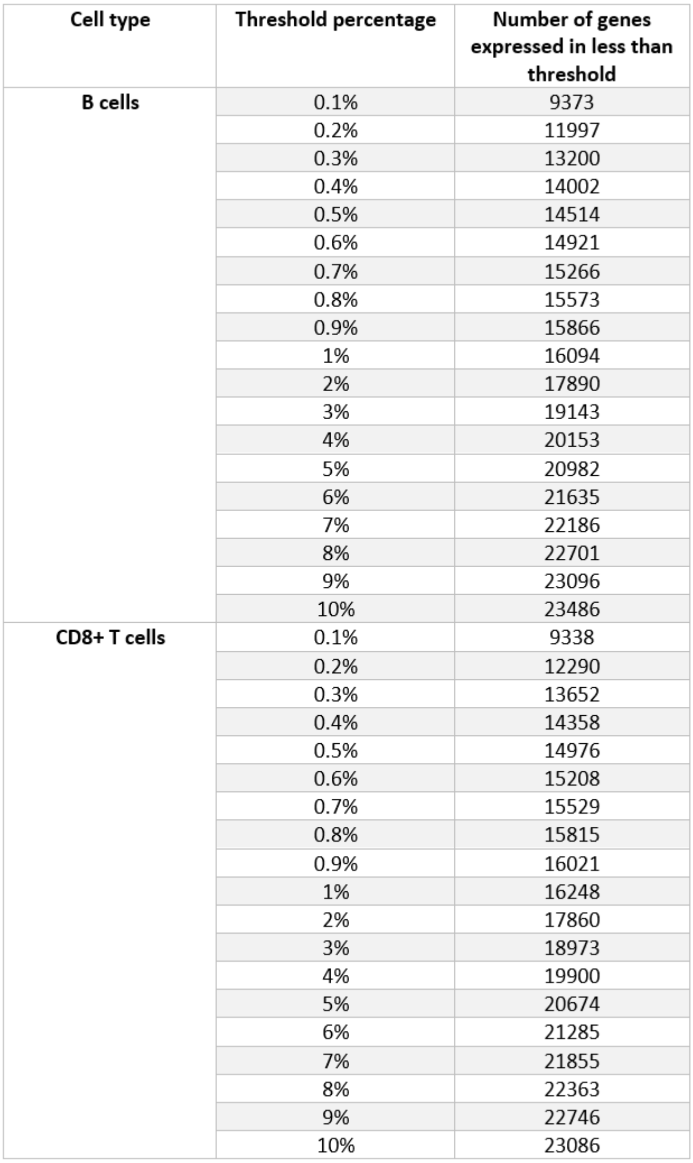
2. *"The analysis of gene expression and its effects on the susceptibility on COVID-19 was very well done. Overall, the report has a clear structure with detailed information that is easy to understand. The data cleaning of the cell filtering was supported through evidence from an article. Choosing the threshold of 1% and removing cells with low gene reads allowed for a reasonable sample size with sufficiently high gene reads within cells to be analysed. The*

decision to choose UMAP over PCA for clustering was also supported through performance results. The UMAP plots for status clearly portray findings but patient plots are difficult to understand. Perhaps some further analysis can be done with the patient category. In the findings and results section, you could perhaps explain whether or not the findings on the UMAP plot for patients is relevant or not with analytical evidence. Volcano plots clearly explain significant data. The todo notes already outline a suitable course of action."

- Resolution: Admittingly, the patients plots were difficult to understand as they were done incorrectly. Furthermore, we decided to omit the patient-level UMAP plots from the final report because we eventually decided not to consider patient variability within our analysis.

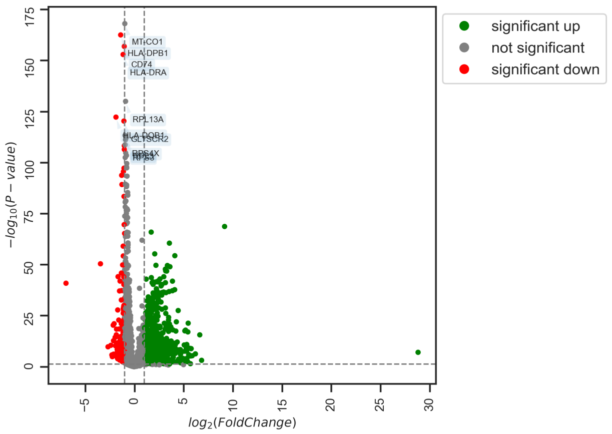
9.2. Appendices referenced throughout report

Appendix A: Trial of different thresholds for gene filtering.

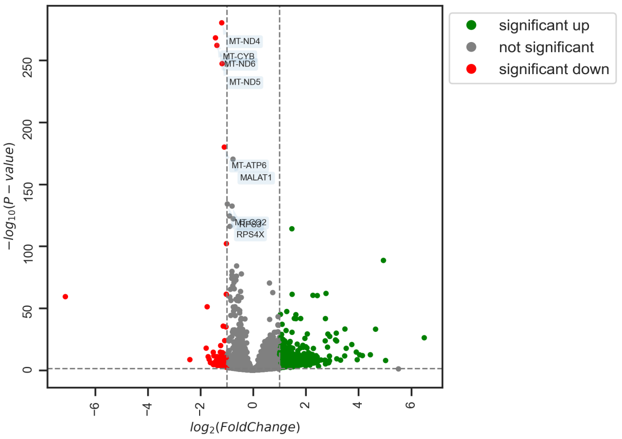


Cell type	Threshold percentage	Number of genes expressed in less than threshold
B cells	0.1%	9373
	0.2%	11997
	0.3%	13200
	0.4%	14002
	0.5%	14514
	0.6%	14921
	0.7%	15266
	0.8%	15573
	0.9%	15866
	1%	16094
	2%	17890
	3%	19143
	4%	20153
	5%	20982
	6%	21635
	7%	22186
	8%	22701
	9%	23096
	10%	23486
CD8+ T cells	0.1%	9338
	0.2%	12290
	0.3%	13652
	0.4%	14358
	0.5%	14976
	0.6%	15208
	0.7%	15529
	0.8%	15815
	0.9%	16021
	1%	16248
	2%	17860
	3%	18973
	4%	19900
	5%	20674
	6%	21285
	7%	21855
	8%	22363
	9%	22746
	10%	23086

Appendix B: Labelled volcano plot with most significant genes based on p-value alone.

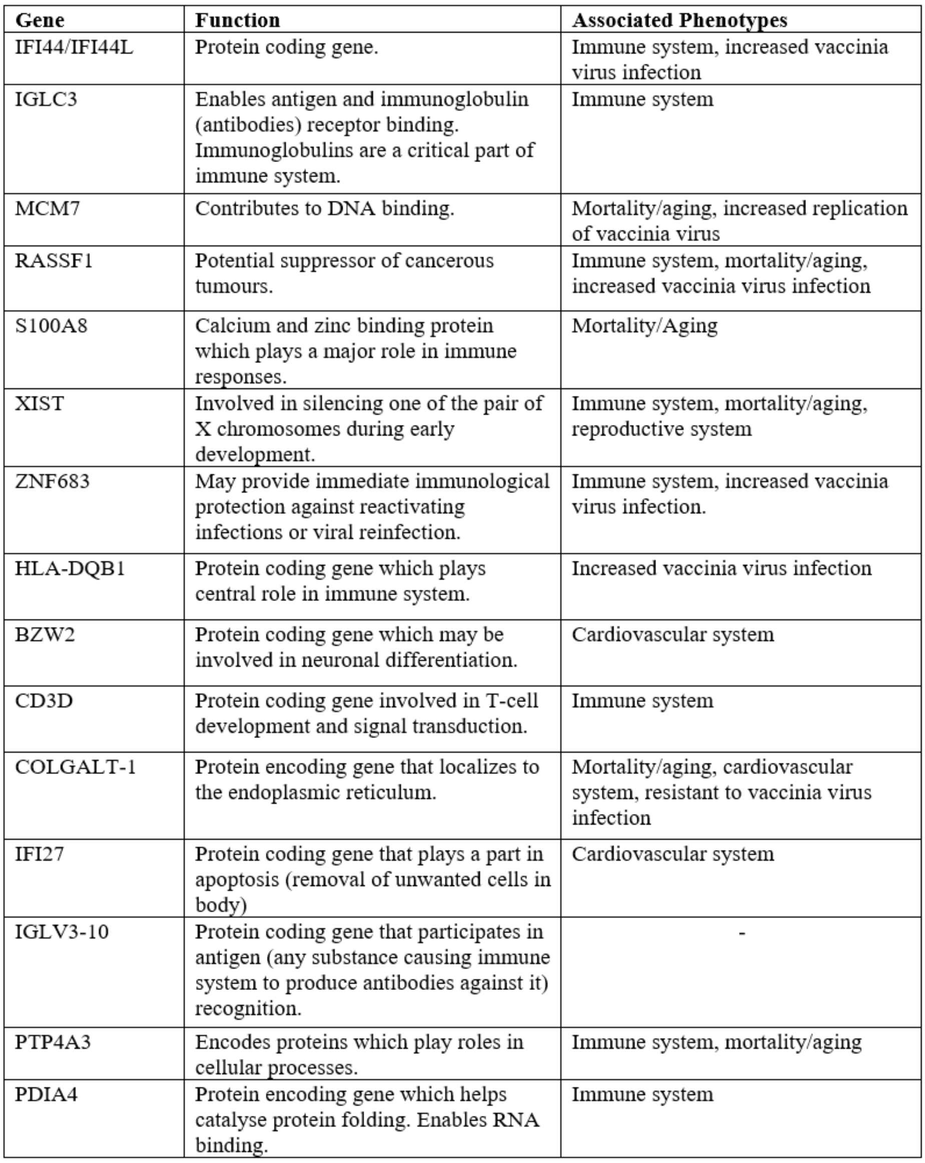


Volcano Plot using combined B cells dataset



Volcano Plot using combined CD8+ T cells dataset

Appendix C: Gene functions* of most statistically significant genes.



Gene	Function	Associated Phenotypes
IFI44/IFI44L	Protein coding gene.	Immune system, increased vaccinia virus infection
IGLC3	Enables antigen and immunoglobulin (antibodies) receptor binding. Immunoglobulins are a critical part of immune system.	Immune system
MCM7	Contributes to DNA binding.	Mortality/aging, increased replication of vaccinia virus
RASSF1	Potential suppressor of cancerous tumours.	Immune system, mortality/aging, increased vaccinia virus infection
S100A8	Calcium and zinc binding protein which plays a major role in immune responses.	Mortality/Aging
XIST	Involved in silencing one of the pair of X chromosomes during early development.	Immune system, mortality/aging, reproductive system
ZNF683	May provide immediate immunological protection against reactivating infections or viral reinfection.	Immune system, increased vaccinia virus infection.
HLA-DQB1	Protein coding gene which plays central role in immune system.	Increased vaccinia virus infection
BZW2	Protein coding gene which may be involved in neuronal differentiation.	Cardiovascular system
CD3D	Protein coding gene involved in T-cell development and signal transduction.	Immune system
COLGALT-1	Protein encoding gene that localizes to the endoplasmic reticulum.	Mortality/aging, cardiovascular system, resistant to vaccinia virus infection
IFI27	Protein coding gene that plays a part in apoptosis (removal of unwanted cells in body)	Cardiovascular system
IGLV3-10	Protein coding gene that participates in antigen (any substance causing immune system to produce antibodies against it) recognition.	-
PTP4A3	Encodes proteins which play roles in cellular processes.	Immune system, mortality/aging
PDIA4	Protein encoding gene which helps catalyse protein folding. Enables RNA binding.	Immune system

*Gene functions and phenotypes courtesy of Gene Cards.

9.3. All code used for analysis

Github Link to Jupyter Notebook containing code.

10. References

[1A] Wissinger, E. (2016, March 17). CD8+ T Cells. *British Society for Immunology.* https://www.immunology.org/public-information/bitesized-immunology/cells/b-cells

[1B] Roghanian, A., Newman, R. (2021, March 18). B cells. *British Society for Immunology.* https://www.immunology.org/public-information/bitesized-immunology/cells/cd8-t-cells

[2] Samir, J., Rizzetto, S., Gupta, M., & Luciani, F. (2020). Exploring and analysing single cell multi-omics data with VDJView. *BMC medical genomics*, *13*(1), 1-9. https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-020-0696-z

[3] A. Fouad, N. (2020). Editor in Chief's Introduction to Essays on the Impact of COVID-19 on Work and Workers. *Journal of Vocational behaviour*, *119*. https://doi.org/10.1016/j.jvb.2020.103441

[4] Chriscaden, K. (2020, October 13). Impact of COVID-19 on people's livelihoods, their health and our food systems. *World Health Organisation*. Retrieved from https://www.who.int/

[5] Black, E. (2021, August 9). New Research shows lasting impact of COVID-19 infection on immune system. *The University of Adelaide.* Retrieved from https://www.adelaide.edu.au/

[6] Wilk, A.J., Rustagi, A., Zhao, N.Q. *et al*. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine.* https://doi.org/10.1038/s41591-020-0944-y

[7] Kim, K., Zakharkin, S.O., & Allison, D.B. (2010). Expectations, Validity and Reality in Gene Expression Profiling. *Journal of Clinical Epidemiology.* https://doi.org/10.1016%2Fj.jclinepi.2010.02.018

[8] Data 100 at UC Berkeley (2019). *Quality control.* https://chanzuckerberg.github.io/scRNA-python-workshop/preprocessing/01-basic-qc.html

[9] Renesh Bedre. (2021, March 7). Volcano Plot on Python. [Web log post]. Retrieved from https://www.reneshbedre.com/blog/volcano.html

[10] *CLC bio*. (2019). Volcano plots - inspecting the result of the statistical analysis https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/754/index.php?manual=Volcano_plots_inspecting_result_statistical_analysis.html

[11] Fontanarosa, J.B., & Dai, Y. (2011). Using LASSO regression to detect predictive aggregate effects in genetic studies. *BMC Proceedings.* https://doi.org/10.1186/1753-6561-5-S9-S69

[12] Stephanie Glen (2015). LASSO Regression: Simple Definition. Retrieved from https://www.statisticshowto.com/lasso-regression/

[13] Stephenson, E., Reynolds, G., Botting, R.A. *et al.* (2021). Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine.* https://doi.org/10.1038/s41591-021-01329-2

[14A] cms72. (2016). *LogFC: how do you determine the cutoff for differentially expressed genes?* [web log post] https://support.bioconductor.org/p/98367/

[14B] Aaron Lun. (2016). *RE: LogFC: how do you determine the cutoff for differentially expressed genes?* [web log comment] https://support.bioconductor.org/p/98367/

[15A] Ankush Dehlia. (2021, June 1). *What is an ideal threshold for log2(Fold Change)?.* [web log post] https://www.researchgate.net/post/What_is_an_ideal_threshold_for_log2Fold_Change

[15B] Xuanyi Chen. (2021, June 26). Re: *What is an ideal threshold for log2(Fold Change)?.* [web log comment] https://www.researchgate.net/post/What_is_an_ideal_threshold_for_log2Fold_Change

[16] Hozumi, Y., Wang, R., Yin, C., Wei, GW. (2021, February 22). UMAP - assisted K-means clustering of large scale SARS-CoV-2 mutation datasets. *US National Library of Medicine National Institutes of Health.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7897976/

[17] Nikolay Oskolkov (2019). *How Exactly UMAP Works.* https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

[18] GeneCards Human Gene Database. (2013). *MX1 Gene - MX Dynamin Like GTPase 1.* https://www.genecards.org/cgi-bin/carddisp.pl?gene=MX1

[19] Siow, I., Lee, K.S., Zhang, J.J.Y *et al.* (2021). Encephalitis as a neurological complication of COVID-19: A systematic review and meta-analysis of incidence, outcomes, and predictors. *Wiley Online Library.* https://doi.org/10.1111/ene.14913

[20] Traag, V.A., Waltman, L., & van Eck, N.J. (2019, March 26). From Louvain to Leiden: guaranteeing well-connected communities. *Nature scientific reports.* https://doi.org/10.1038/s41598-019-41695-z

[21] Andy Coenen, Adam Pearce (2021). Understanding UMAP [online] https://pair-code.github.io/understanding-umap/

[22] Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021, September 10). A review of UMAP in population genetics. *Journal of Human Genetics* https://doi.org/10.1038/s10038-020-00851-4

[23A] Parviz Heidari. (2015, December 14). *FDR or log fold change: which one is the priority for selecting the DEGs?* [web log post] https://www.researchgate.net/post/FDR_or_log_fold_change_which_one_is_the_priority_for_selecting_the_DEGs

[23B] Jaroslav Zak**.** (2015, December 14). Re: *FDR or log fold change: which one is the priority for selecting the DEGs?* [web log comment] https://www.researchgate.net/post/FDR_or_log_fold_change_which_one_is_the_priority_for_selecting_the_DEGs

[24] Silicon Genetics. (2003). *Multiple Testing Corrections* https://physiology.med.cornell.edu/people/banfelder/qbio/resources_2008/1.5_GenespringMTC.pdf

[25] Douglas Bowman, Darren Delaye. (2006 July 14).Multiple Testing P-Value Corrections in Statsmodels. Minima Lefty Stretch. *JOEPY.* http://jpktd.blogspot.com/2013/04/multiple-testing-p-value-corrections-in.html

[26] Dr. Saul McLeod. (2019 July 4). What are Type I and Type II Errors? *Simply Psychology.* https://www.simplypsychology.org/type_I_and_type_II_errors.html

[27] Renesh Bedre. (2021, January 25). Multiple hypothesis testing problem in Bioinformatics. *Data science blog.* https://www.reneshbedre.com/blog/multiple-hypothesis-testing-corrections.html

[28] user4673. (2012). Is Benjamini-Hochberg correction more conservative as the number of comparisons increases? *StackExchange* [web blog comment] https://stats.stackexchange.com/questions/21193/is-benjamini-hochberg-correction-more-conservative-as-the-number-of-comparisons

[29] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[30] Segal, M. R., Dahlquist, K. D., & Conklin, B. R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology,* 10(6), 961-980. https://doi.org/10.1089/106652703322756177

[31] Greenwood, M. (2021, August 5). Vaccinia-based COVID-19 vaccines confer immunity in hamster model. *News Medical.* Retrieved from
https://www.news-medical.net/

[32] Christoph Molnar. (2021 November 11). 3.1 Importance of Interpretability. *Interpretable Machine Learning.*
https://christophm.github.io/interpretable-ml-book/interpretability-importance.html