Regulatory response to infection by SARS-CoV-2

Pedro Rodrigues
IST, IDMEC, INESC-ID
Lisboa, Portugal
pedro.p.rodrigues@tecnico.ulisboa.pt

***Abstract*—Covid-19, the disease caused by the novel coronavirus, SARS-CoV-2, has already affected over 241 million individuals and caused the deaths of over 4.9 million. However, the knowledge of the impacts of this virus on infected cells is still incomplete. Thus, the present work aims to identify and analyse the main cell regulatory processes affected and induced by SARS-CoV-2, using transcriptomic data from several infectable cell lines available in public databases. We propose a new class of statistical models to handle three major challenges, namely the scarcity of observations, the high dimensionality of the data, and the complexity of the interactions between genes. Additionally, we analyse the function of these genes and their interactions within cells to compare them to ones affected by IAV (H1N1), RSV and HPIV3 in the target cell lines. Gathered results show that the usage of clustering, biclustering and predictive algorithms significantly improve the number and quality of the detected biological processes. Additionally, a comparative analysis of these processes is performed in order to identify potential pathophysiological characteristics of Covid-19. These are further compared to those identified by other authors for the same virus as well as related ones such as SARS-CoV-1. This approach is particularly relevant due to a lack of other works utilizing more complex machine learning tools within this context.**

***Index Terms*—COVID-19; SARS-CoV-2; Discriminative Regulatory Patterns; Cell Transcriptomics; Biclustering; Gene Expression Data Modeling.**

I. Introduction

The infection of humans by Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) represents a major global health concern, with deaths having surpassed 4.9 million according to the World Health Organization (WHO) [1]. Due to the situation, there has been a focus on making data relating to this virus publicly available. This has provided an opportunity for researchers to utilize public data to draw novel insights into the infectious disease, which have enabled continuous breakthroughs in the understanding of how the virus can enter and utilize the cellular machinery to replicate itself and infect other cells. The knowledge relating to these mechanisms has been pushed forward mainly by a generic understanding of the process of viral replication, the transcriptomic properties of the virus, and by the study of differentially expressed genes after infection and subsequent comparison to ones affected by other viral strains. These genes have generally been identified by the usage of recent sequencing technologies, such as RNA-seq, which have been applied to certain types of cells, chosen according to their level of permissivity to infection, as well as cells collected from organisms susceptible to infection, such as humans and ferrets [1].

Despite the ongoing breakthroughs, the cellular responses to SARS-CoV-2 are still considerably unknown. For instance, the role played by genes with moderate differential expression, and how interactions between multiple genes support or prevent viral replication are still being actively updated. In addition to this, most works in this field do not make use of more complex techniques such as clustering, predictive models and biclustering to aid in the identification of differentially expressed genes and related biological processes.

II. Related Work

The primary focus of this work is to detect differentially expressed genes when cells are infected by SARS-CoV-2, as well as identifying defining traits when compared with other viruses. Though the present section is focused on this particular area, it composes only a fraction of the existing body of work, with the main focus being on the identification of host genomic factors which may affect clinical outcomes of COVID-19 [2] and the usage of the transcriptome of SARS-CoV-2 [3] to identify particular characteristics of the virus.

Blanco-Melo D. et al. [1] utilized high-throughput sequencing (RNA-Seq) to characterize the transcriptional response of cells to infection by SARS-CoV-2 and against other respiratory viruses, including RSV, IAV and HPIV3 from data collected by the authors and MERS-CoV and SARS-CoV-1 from data collected by Frieman et al. [4] and available on the GEO website (GSE56192). The cells analysed consisted in three main groups: cell lines consisting of NHBE cells, A549 cells and Calu-3 cells; human respiratory tract cells extracted from infected and non-infected individuals; and cells extracted from infected and non-infected ferrets. The second and third groups were used to ascertain if the gene signatures matched the ones found in vitro. Aditionally, the authors treated cells with universal IFNβ to determine whether or not SARS-CoV-2 is sensitive to IFN-I. The treatment resulted in highly decreased viral replication, which indicates that it is.

Then, to investigate how infection affects the cell transcriptome, the authors performed a differential expression analysis on NHBE cells, which revealed significant differences between the response to infection by SARS-CoV-2 and other viral strains, with PCA also revealing significant differences. Functional enrichment was also performed on the resulting genes, to better understand the cellular functions affected by SARS-CoV-2 infection. The main factors consistent throughout the

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019, accessed on the 16th of October 2021

various models tested was the production of cytokines and the corresponding transcriptional response, as well as the induction of a subset of interferon stimulated genes (ISGs).

Ochsner et al. [5] analyzed multiple publicly archived transcriptomic datasets to better identify the transcriptional response of human cells to SARS-CoV-2 infection as well as comparing it with MERS-CoV, SARS-CoV-1 and IAV in order to identify possible common impacts between viral strains. The authors generated consensomes by analysing how frequently the corresponding genes were differentially expressed throughout the various datasets. Similarly to Blanco-Melo D. et al. the authors found ISGs had significant induction levels.

Wei et al. [6] performed a genome-wide CRISPR screen on an African green monkey cell line (Vero-E6), a method used for identifying genes or genetic sequences that have a certain physiological effect, in this case, aiding (pro-viral) or preventing (anti-viral) infection. To this end, surviving cells from populations either healthy or infected with SARS-CoV-2 were harvested 7 days post-infection. Then a genome-wide screen was performed and a z-score was calculated to identify which genes could be associated with increased or decreased resistance to SARS-CoV-2-induced cell death. The gene with the strongest pro-viral effect was ACE2, associated with the protein which allows viral entry into the cell. TMPRSS2, another gene posited to play a role in the entry of SARS-CoV-2 into the cell, was not identified significantly as pro or anti-viral, whereas the CTSL gene, which encodes the Cathepsin L protease and can also play a role in viral entry, was identified as pro-viral.

Similarly to Blanco-Melo et al., Wyler et al. [7] performed a comprehensive analysis of the transcriptional response of three cell lines, Caco-2 (a gut cell line), Calu-3 and H1299 (both lung cell lines). The authors began by identifying the susceptibility of each cell line to SARS-CoV-2 infection, which revealed H1299 cells had the lowest percentage of viral reads. Caco-2 and Calu-3 cells had comparable levels, despite the latter revealing visible signs of impaired growth and cellular death, as opposed to the former. Additionally, Calu-3 cells showed a strong induction of interferon-stimulated genes, with cytokines among these, in agreement with the findings of others.

Due to thrombotic complications being common among COVID-19 patients, Manne et al. [8] investigated the functional and transcriptional changes elicited by SARS-CoV-2 infection in platelets. The data showed that SARS-CoV-2 infection does indeed alter the platelet transcriptome. To detect these changes, when comparing two groups with normal distributions, a paired t-test was used and when comparing two groups with non-normal distributions a Mann-Whitney test was used, considering a two-tailed p-value < 0.05 as statistically significant. Additionally, COVID-19 induces functional and pathological changes to platelets, including thrombocytopenia (abnormally low numbers of platelets), despite the platelets not presenting detectable levels of ACE2. This may be a contributing factor to the pathophysiology of COVID-19.

Golden et al. [9] tested the pathogenesis of the SARS-CoV-2 virus on transgenic mice presenting the human ACE2 gene. The infection of these mice by SARS-CoV-2 resulted in high mortality rates, especially in male mice. The transcriptional analysis of the lungs of infected animals revealed increases in transcripts involved in lung injury and inflammatory cytokines, in agreement with findings for humans.

Though there are multiple authors applying machine learning and more complex statistical models to COVID-19 patient biometric data, in order to analyse the characteristics and the outcome of the disease, these approaches have been more scarcely applied to transcriptomic data. The objective of this work is to fill this gap, addressing the question of whether the application of those models to this data can yield novel insights into the disease.

III. Exploratory Analysis

A. Data Description

The target dataset, identified as GSE147507 [2], was collected by Blanco-Melo D. et al. [1] using RNA-Seq, which means the resulting dataset is numeric, with the values representing the number of RNA transcripts of each gene detected in the sample.

We began by checking the available samples. These are subdivided into different *Series* (a subset of samples), each of which aim to compare the behavior of a single cell line among different sets of experimental conditions. A schematic of the structure of the dataset is presented in . These also correspond to particular experiments being run, with each experiment containing multiple replicas of each experimental condition being tested. As such, the assumed independence between replicas is an important factor to test, since being able to use samples from multiple experiments simultaneously could significantly increase the amount of data available, and thus improve the reliability of the analysis.

For NHBE (normal human bronchial epithelial) cells, there are a total of 9 samples of healthy cells (3 belonging to *Series* 1 and 4 to *Series* 9), 3 samples of SARS-CoV-2 infection (all part of *Series* 1), 4 samples of IAV infection (all in *Series* 9), 4 samples of infection by an IAV strain which lacks the NS1 protein and, finally, 2 samples of cells treated with IFNβ 4, 6 and 12 hours post treatment.

For A549 (adenocarcinomic human alveolar basal epithelial) cells, there are 13 samples of healthy cells (3 each of *Series* 2, 5 and 8, 2 each of *Series* 3 and 4), 6 samples of SARS-CoV-2 infection (3 each of *Series* 2 and 5), 2 samples of IAV infection (*Series* 4), 2 samples of RSV infection (*Series* 3) and 3 samples of HPIV3 infection (*Series* 8). Blanco-Melo et al. [1] noted A549 cells had low viral counts, which was posited, in agreement with others, to be due to the low expression of ACE2 in these cells. Thus, data of A549 cells with added ACE2 (A549-ACE2) was also made available. In particular, 6 samples of healthy cells (3 each of *Series* 6 and 16), 6 samples of cells infected by SARS-CoV-2 (3 each of *Series* 6 and 16)

[2]Available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147507

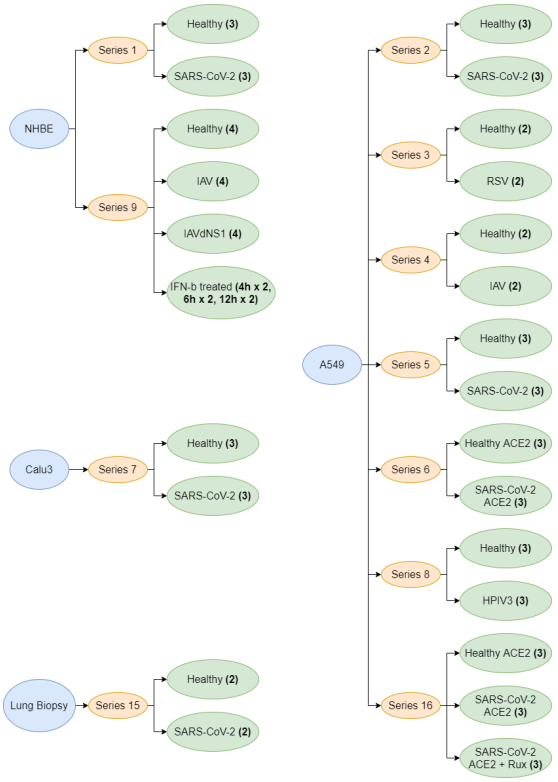

Fig. 1. Overview of the structure of the dataset

and, finally, 3 samples of cells after treatment with Ruxolitinib (*Series* 16).

For Calu3 cells (generated from a bronchial adenocarcinoma), there are 3 samples of healthy cells and 3 samples of cells infected by SARS-CoV-2 (all belonging to *Series* 7).

There are an additional 2 samples from a lung biopsy of two healthy human donors (one male, one female), as well as 2 samples from a single deceased male patient of COVID-19.

B. Preliminary analysis

Since the original data is highly skewed, which is the norm for transcriptomic data, a log-transform was applied for all subsequent analysis, which resulted in less skewed distributions.

From the initial distributions, we observed the various *in-vitro* cell lines to be fairly similar, whereas lung biopsy cells appear to show lower overall transcription levels (Figure 2).

Subsequently, the standard deviation of gene expression among healthy cells and among infected cells was computed to verify if there are significant differences between healthy and infected cells (Figure 3).

Despite there being clear differences in the distributions, there seems to be no clear pattern between the different types of cells. For NHBE and lung biopsy cells, infected cells seem to have more variation, whereas for Calu3 and A549 cells the opposite seems to be the case. From this we can derive the hypothesis that there is a hierarchy in the cells when it comes to variability of gene expression, though we cannot

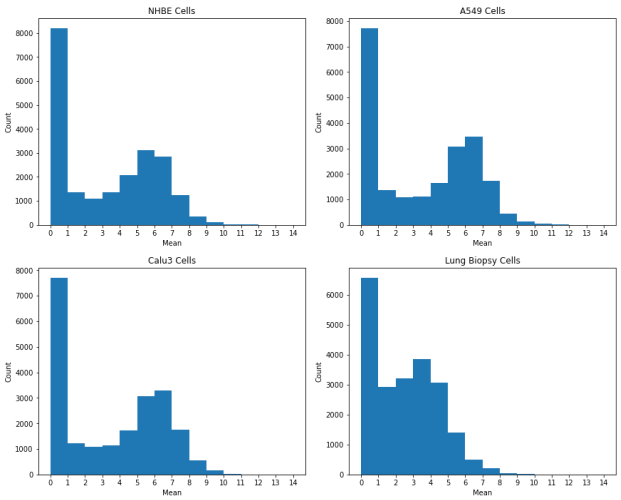

Fig. 2. Distribution of gene expression (mean among samples) after applying a log transform (N = 21797 genes)

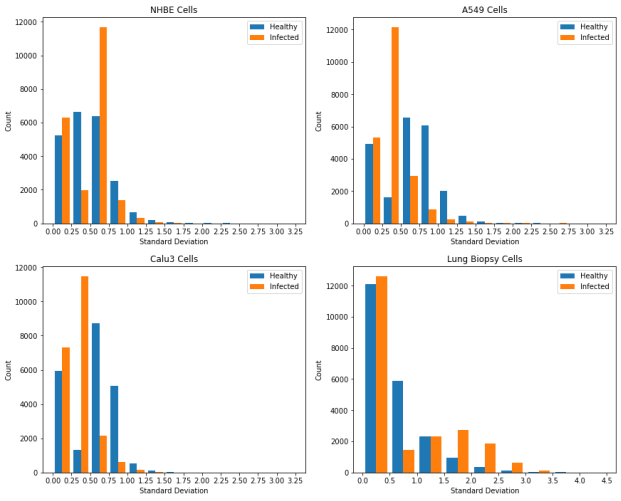

Fig. 3. Standard deviation of gene expression within healthy and within infected cells

posit whether this is due to the level of susceptibility of each cell type to infection and/or due to certain types responding better to infection.

In order to select an appropriate statistical test for the initial feature selection, a number of assumptions need to be checked. Firstly, we perform a median based Levene's test [10], which is used, in the context of this work, to assess the equality of variances each pair of conditions (in particular for the pairs presented in Table I). For these pairs, out of 19967 genes with non-null expression levels, 18990 had unequal variance for at least one pair of conditions, with $p < 0.01$.

Additionally, a Shapiro-Wilk test [11] is used to assess whether these genes follow a normal distribution, applied in this case only to healthy and SARS-CoV-2 infected cells for each cell type (since these will be the main focus of our analysis and this test is only defined for at least 3 samples). A $p < 0.05$ was used. It is important to note that overall 32.8%,

TABLE I
TESTED PAIRS OF CONDITIONS

First Condition	Second Condition
NHBE Healthy	NHBE SARS-CoV-2
NHBE Healthy	NHBE IAV
NHBE Healthy	NHBE IAVdNS1
A549 Healthy	A549 SARS-CoV-2
A549 Healthy	A549 IAV
A549 Healthy	A549 RSV
A549 Healthy	A549 HPIV3
Calu3 Healthy	Calu3 SARS-CoV-2
Biopsy Healthy	Biopsy SARS-CoV-2

46.1% and 27.4% of genes for NHBE, A549 and Calu3 cells respectively are non-normal.

The results of Levene's test suggest that an assumption of equal variance cannot be made. As such, either an unequal variance (Welch) t-test or it's non-parametric alternative, the Mann-Whitney U test, are more suitable for variable selection. With the results for non-normality still including a significant percentage of the genes the Mann-Whitney U test seems more appropriate.

IV. Solution

As previously stated, our work aims to find relevant biological processes involved in the infection of cells by SARS-CoV-2. To this end, we propose a methodology for the selection and discovery of correlated groups of DEG composed of 5 major steps. First, we begin with preprocessing techniques and preliminary gene selection. Then we proceed to pattern detection techniques, namely clustering, predictive modeling and biclustering. For each of these techniques, we apply functional enrichment to the obtained groups of genes, in order to identify related biological functions. Finally, we analyse and interpret the identified functions, relating them to known characteristics of the disease as well as work by other authors. These steps are summarized in Figure 4. In the present chapter, we motivate their need and explore each of the steps in more detail.

V. Preprocessing and Gene Selection

Given the highly skewed distribution of the data (with a vast majority of genes having very low transcription), we first apply a log transform. Then, since the data is high-dimensional, with transcription values for over 20.000 genes, we need to select a set of DEG to be analysed. To this end, due to the non-normal nature of the data and the unequal variance between the control and test groups (as seen in section III), we use a Mann-Whitney U test, with a $p < 0.05$ and $p < 0.01$. By default a $p < 0.01$ is used, however for certain cell types this does not provide a sufficient amount of genes for analysis, so in those cases (as well as for biclustering, in order to provide a comparison between the two values) a $p < 0.05$ is used. The Mann Whitney U test tests for the null hypothesis that the two populations tested are equal. Therefore, this test can only be applied for pairs of conditions. We can define the following settings, in which this method is applied:

1) **Pair Setting** - Single pairs of conditions, such as, for instance, healthy and SARS-CoV-2 infected NHBE cells or healthy and IAV infected A549 cells;
2) **Multi-condition Setting** - A set of pairs of conditions, presented in Table I. For each of these pairs, a p-value is calculated using a Mann-Whitney U test for each gene. Then, all genes with $p < 0.01$ or $p < 0.05$ are chosen.

Additionally, for the biclustering algorithms, we also used an ANOVA test. This was mainly included to provide a contrast in the biclustering analysis to the default preprocessing method, as well as due to this method still being robust with non-normal the data [12].

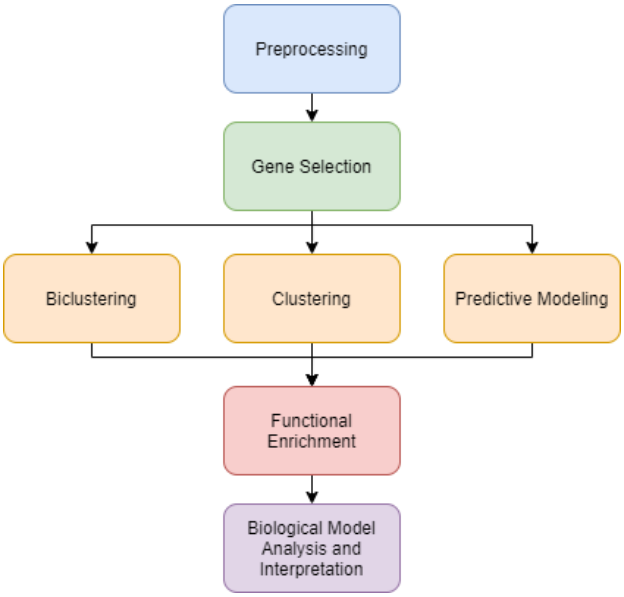

Fig. 4. Schematic of the steps composing our proposed solution

VI. Pattern Detection

The usage of complete data with a simple statistical preselection of genes yields results which, depending on the chosen level of statistical significance, can surpass 1.000 genes. Applying functional enrichment to these results delivers none or very few enriched processes, which, when they exist, tend to be very generic cell functions. This is due to problems with the predictive models used to obtain relevant biological processes. Thus, by first finding smaller sets of DEG, we can obtain more specific biological processes, as well as better statistical significance for each one found.

To achieve this goal, we present three main methods, Clustering, Predictive Modeling and Biclustering.

A. Clustering

The notion of cluster in our data can assume two distinct forms. First, a subset of correlated genes along a given set of samples, and second a subset of correlated samples along a given set of genes.

The latter is mainly interesting to understand which samples may be more closely related, though since it does not subdivide genes it cannot identify which genes may be better at distinguishing between different conditions.

The former is the option most useful to identify gene sets with correlated expression, though it has considerable limitations. Namely, that each grouping found will use all selected samples, which means, if multiple conditions are used simultaneously, this information will not be taken into account and will bias the detected patterns. However, by selecting different sets of conditions for each run of the algorithms, we can obtain relevant patterns for each specific condition and, though this doesn't allow for a direct comparison between different conditions, it can provide sets of correlated genes which may have biological relevance.

The main clustering method we propose is Agglomerative Clustering, with Euclidean affinity and Ward linkage. This is due to two main reasons, the easy visualization of the proximity between genes (using a dendrogram, which can also help in the selection of the number of clusters) and the flexibility of the algorithm, which allows for multiple parameters to be adjusted according to the provided data.

B. Predictive Modeling

Classifiers generally use training data to produce predictive models, which are then used on test data to classify samples. In our work, since we seek to better understand potential signaling pathways and gene ontologies involved in the infection by SARS-CoV-2, we mainly focus on which genes are chosen to classify each of the samples, by inpecting the learned model. Thus, we mainly propose associative classifiers which can be easily interpreted, namely decision trees, random forests and XGBoost. While not directly interpretable, both random forests and XGBoost provide a metric of the relevance of each gene, which can be used to obtain the set of genes with the highest difference in expression level. In both cases, this metric corresponds to the impurity-based feature importances, which are calculated using the Gini criterion and then averaged across all trees within the model.

C. Biclustering

By using biclustering algorithms, we can detect patterns spanning particular sets of conditions, as well as patterns spanning multiple conditions, allowing for a more comprehensive view of the genes associated with not only SARS-CoV-2 infection but also the main differences when compared to other infections. In particular, when compared to the other proposed methods, biclustering allows for the detection of more specific patterns, such as a set of genes with higher or lower expression levels for a particular set of conditions, which are in turn easier to interpret and provide better results with functional enrichment.

We tested several algorithms, as well as different gene selection options, to assess differences between the detected biclusters, namely the Cheng and Church [13], plaid [14], xMotifs [15] and BicPAMS [16] algorithms.

VII. Functional Enrichment and Biological Analysis

To obtain potential biological processes associated with the gene groups found using the aforementioned methods, we used the EnrichR [3] [17], [18]. This tool provides a set of metrics to evaluate each of the enrichment results. These are the p-value, which can be calculated using Fisher's exact test; the q-value, which adjusts the p-value to control the False Discovery Rate; the z-score, which takes into account that Fisher's exact method to calculate the p-value produces lower values for longer lists even if they are random. Furthermore, the tool also provides a combined score, which combines the z-score and the p-value as follows: $c = \ln(p) \times z$.

Given the available metrics and the results by the authors of the tool [18], we propose the usage of both the adjusted p-value and the combined score to compare the results of the enrichment analysis.

Additionally, this tool provides access to multiple knowledge bases (a list is available here). For our analysis, we mainly use the Gene Ontology (GO) Biological Process knowledge base [19], [20], in particular the 2021 revision. This is due to the fact that it covers a large amount of genes (14937) and also includes a high number of terms (6036), as well as that it provides biological processes in which a given set of genes is involved, which aligns with the goals of this work, namely the understanding of the biological processes elicited in response to and by the infection by SARS-CoV-2. Additionally, we use the Kyoto Encyclopedia of Genes and Genomes (KEGG) [21] to analyse enriched pathways and diseases. The identified biological processes are then analysed and compared to known characteristics of the disease and work by other authors, in order to identify potential new insights into the effects of the virus and verify existing ones.

VIII. Key Findings

To solve the problem of identifying smaller and more internally correlated sets of DEG, as overviewed in section IV, we use three methods: Clustering, Predictive Modeling and Biclustering. These methods allow us, when performing functional enrichment, to identify more statistically significant biological processes. In the present section, we present the key findings resulting from the application of each of these methods to the dataset, as well as an analysis of the identified biological processes within the context of viral infection. In particular, we will begin with clustering, then classification and finally biclustering.

To assess the effectiveness of the methods explored later in the chapter, we begin by presenting, in Table II, the result of performing functional enrichment on the set of genes obtained directly through preprocessing (in the Multi-Condition Setting, $p < 0.01$), in the previously (section IV) defined **Multi-Condition Setting**.

As we can see in Table II, there is a considerable number of processes with low p-value. However, the c-score is signifi-

[3]Freely available at https://maayanlab.cloud/Enrichr/

TABLE II
TOP 8 GO BIOLOGICAL PROCESSES ORDERED BY COMBINED SCORE, USING JUST PREPROCESSING (MULTI-CONDITION SETTING, $p < 0.01$)

GO Biological Process	p-value	c-score
cellular response to type I interferon (GO:0071357)	2.35E-10	324.03
type I interferon signaling pathway (GO:0060337)	2.35E-10	324.03
cytokine-mediated signaling pathway (GO:0019221)	3.80E-26	319.38
protein mono-ADP-ribosylation (GO:0140289)	3.22E-04	319.17
receptor signaling pathway via STAT (GO:0097696)	2.73E-06	299.82
receptor signaling pathway via JAK-STAT (GO:0007259)	2.50E-06	250.15
exogenous peptide antigen, TAP-independent (GO:0002480)	7.04E-03	219.44
negative regulation of bone remodeling (GO:0046851)	2.62E-03	212.68

TABLE III
TOP 8 KEGG PATHWAYS ORDERED BY COMBINED SCORE, FOR A549 CELLS

KEGG Pathway	p-value	c-score	Cluster
Measles (map05162)	1.02E-08	295.14	2
Influenza A (map05164)	1.02E-08	252.18	2
Herpes simplex virus 1 infection (map05168)	7.16E-10	177.20	2
Epstein-Barr virus infection (map05169)	4.82E-07	156.14	2
TNF signaling pathway (map04668)	1.11E-07	153.86	0
Coronavirus disease (map05171)	3.78E-07	150.32	2
RIG-I-like receptor signaling pathway (map04622)	3.68E-04	120.34	2
NOD-like receptor signaling pathway (map04621)	9.37E-06	119.63	2

cantly lower when compared to the same genes after clustering (see Table II). This is likely due to the higher number of genes being analysed together when compared to the proposed methods, since clustering and biclustering identify smaller subgroups of genes with correlated expression and predictive models select a smaller number of genes. Additionally, terms such as *negative regulation of bone remodeling* (GO:0046851) and *negative regulation of bone resorption* (GO:0045779), which seem to be more generic and less related to the viral infection appear in this analysis, but do not seem to reoccur within the terms found for clustering, classification or biclustering.

A. Clustering

Starting with a **Multi-Condition Setting** ($p < 0.01$) in Table II, a high percentage of the top identified processes are related to response to viral infection, as well as to immune responses. The annotation *cytoplasmic pattern recognition receptor (PRR) signaling pathway in response to virus*, GO:0039528 (directly related to the annotations GO:0140546 and GO:0051607, also within the top 25 enriched processes) corresponds to a set of molecular signals associated with the detection (by binding of viral RNA molecules to certain cytoplasmic receptors) of a virus. In particular, the detection seems to be performed by the RIG-I PRR, responsible for the detection of RNA synthesized during the process of viral replication, since there are 3 child processes (GO:0039529 with $p = 2.91 * 10^{-3}$ and $c = 905.29$; GO:0039535 with $p = 7.67 * 10^{-4}$ and $c = 526.08$; GO:0039526 with $p = 5.26 * 10^{-3}$ and $c = 513.51$) associated with this receptor which are still statistically relevant. This receptor, along with others, has been identified as part of the inflammatory response to SARS-CoV-2 as well as other coronaviruses [22]. Additionally, the signaling cascade resulting from the detection of viral proteins is associated with the production of Type I interferons and pro-inflammatory cytokines [23], which can also be observed within the top enriched processes (for instance, terms GO:0060337, GO:0071357 and GO:0060333).

In particular, the term *type I interferon signaling pathway* (GO:0060337), which has several related terms also present within the top 25 processes (for instance, *type I interferon signaling pathway*, GO:0060337 and *cytokine-mediated signaling pathway*, GO:0019221, both direct ancestors) are related to type I interferons. The association between these and the process of viral infection is further bolstered by the presence of terms *response to interferon-beta* (GO:0035456) and *response to interferon-alpha* (GO:0035455), which are both type I interferons.

It is also interesting to note the presence of the term *negative regulation of type I interferon-mediated signaling pathway* (GO:0060339) as well as *negative regulation of chemokine production* (GO:0032682). Chemokines are involved in inflammation and the control of viral infections, and they and their receptors are sometimes mimicked by viruses in order to evade host antiviral immune responses [24]. The presence of these is noteworthy mainly due to directly opposing the other processes related to the activation of an immune response.

Additionally, there are multiple processes directly related to cellular response to viruses, namely *defense response to symbiont* (GO:0140546), *defense response to virus* (GO:0051607), *negative regulation of viral genome replication* (GO:0045071, also associated with GO:0045069), *antiviral innate immune response* (GO:0140374), *negative regulation of viral process* (GO:0048525) and *cellular response to virus* (GO:0098586). These indicate that NHBE cells were able to identify that they had been infected by a virus and induce an immune response.

For A549 cells, the identified terms can be seen in **??**. The genes composing all detected processes have higher expression levels for infected cells than for control. Similarly to NHBE cells, there seems to be a prevalence of type I interferon and cytokine related terms. Multiple processes, such as *cellular response to type I interferon* (GO:0071357), *type I interferon signaling pathway* (GO:0060337), *response to interferon-beta* (GO:0035456) are repeated, with most of the common processes having to do with interferon and general cytokine response as well as responses to viral infection.

The terms *STAT cascade* (GO:0097696), *positive regulation of JAK-STAT cascade* (GO:0046427) and *JAK-STAT cascade* (GO:0007259), are not present for NHBE cells. These are all related to the JAK-STAT signaling pathway, which is associated with a wide variety of cytokines. Not triggering signaling or not regulating it properly, can lead to inflammatory disease [25], among other issues.

Interestingly, similarly to the NHBE cells the process *negative regulation of type I interferon production* (GO:0032480) seems to suggest a potential attempt to reduce immune re-

TABLE IV
TOP 25 GO BIOLOGICAL PROCESSES ORDERED BY COMBINED SCORE (MULTI-CONDITION SETTING, $p < 0.01$).

GO Biological Process	p-value	c-score
type I interferon signaling pathway (GO:0060337)	4.40E-27	9111.11
cellular response to type I interferon (GO:0071357)	4.40E-27	9111.11
negative regulation of viral genome replication (GO:0045071)	1.10E-16	3820.31
defense response to symbiont (GO:0140546)	7.74E-22	3260.22
cytoplasmic PRR signaling pathway [4](GO:0039528)	1.74E-06	3253.53
negative regulation of viral process (GO:0048525)	5.16E-17	3163.10
defense response to virus (GO:0051607)	2.34E-21	2930.10
endogenous peptide antigen, TAP-independent (GO:0002486)	4.42E-05	2797.75

sponse. However, the opposite term, *positive regulation of type I interferon production* (GO:0032481) is also within the top 25 (though with higher p-value and lower c-score). This may be due to both pathways being active simultaneously, although it may also reveal overlap in the genes that produce each process (2 out of 5 genes in common between the two processes).

In Table III we present the pathways identified when using the **KEGG Pathway database** (2021 version) instead of the GO database. The results, similarly to the GO database, include multiple virus related pathways. These are all composed by genes with higher expression values for infected cells than control. Within the top identified terms, there is a prevalence of virus related pathways. *Coronavirus disease* (map05171), the sixth term, is directly associated with SARS-CoV-2, which provides more confidence that the terms identified thus far are indeed related to the viral infection.

Additionally, the term *RIG-I-like receptor signaling pathway* (map04622), which is related to the previously mentioned RIG-I receptor, helps solidify the idea of it being involved in the anti-viral immune response.

The KEGG pathways *Antigen processing and presentation* (map04612), *JAK-STAT signaling pathway* (map04630), the principal signaling mechanism for a variety of cytokines, *IL-17 signaling pathway* (map04657), a subset of cytokines with various roles related to inflammatory responses and defence against external pathogens, and *NF-kappa B signaling pathway* (map04064), a signaling pathway which is activated by the aforementioned cytokines and is related to immune responses, all support the processes identified previously in the role played by inflammatory cytokines and related signaling pathways in the infection by SARS-CoV-2.

B. Predictive Models

For predictive models, we begin once again with a Multi-Condition Setting ($p < 0.01$), for the Random Forest and xGBoost algorithms. With XGBoost, 94 genes are identified. With the Random Forest, 356 genes are selected. These algorithms have 69 genes in common. There are multiple terms present in both models, mostly related to immune system activity. However, there are several processes uniquely identified by each of the algorithms. Processes identified only by XGBoost are particularly interesting, since most genes selected by XGBoost are also selected by the Random Forest and the extra genes selected by the Random Forest may mask relevant information.

ISG15-protein conjugation (GO:0032020), a term identified only within XGBoost selected genes, is related to the cellular protein modification process of ISG15. This protein has an important role in host antiviral response, with several different actions depending on the infecting virus. Most significantly among these actions is the inhibition of viral replication in addition to the modulation of the damage and repair as well as the immune responses [26].

Also within the terms identified only by XGBoost, there are multiple related to chemotaxis, the movement of a cell or organism towards a higher or lower concentration of a given substance, and migration of various types of immune cells. In particular, macrophages [27], [28] (GO:0048246 and GO:1905517), natural killer cells [29] (GO:2000501)), eosinophils [30] (GO:0072677 and GO:0048245), neutrophils [31] (GO:0030593 and GO:1990266), which are all types of white blood cells involved with the innate immune response to viral infection.

Additionally, there are multiple terms in both cases associated with cytokine production and related signaling pathways, as well as response to different types of interferons. In addition to these, terms such as *regulation of fever generation* (GO:0031620), *negative regulation of viral process* (GO:0048525), *inflammatory response* (GO:0006954) and *negative regulation of viral genome replication* (GO:0045071) are also associated with immune response. Together with the previously mentioned signaling of white blood cells, these results show the significant, both innate and adaptive, immune responses by cells infected by this virus.

Among the top processes in both tables is *chronic inflammatory response* (GO:0002544). Similarly to what was mentioned for the combined data, there are multiple terms related to the recruitment of certain types of white blood cells. In particular, *positive regulation of monocyte chemotactic protein-1 production* (GO:0071639), the top term for the Random Forest, is associated to a protein which plays a key role in the migration of monocytes [32].

It is also important to note that multiple terms associated with the apoptotic process are present, namely *positive regulation of intrinsic apoptotic signaling pathway* (GO:2001244), *regulation of intrinsic apoptotic signaling pathway* (GO:2001242) and *positive regulation of apoptotic signaling pathway* (GO:2001235). This process, responsible for causing the death of a cell when a certain internal or external stimulus is received, may indicate that the cell was able to detect that it was infected by SARS-CoV-2. This hypothesis is further supported by the presence of the term *pattern recognition receptor signaling pathway* (GO:0002221). These receptors, as previously explained for the related term present in Table IV, have been associated with the inflammatory response to SARS-CoV-2 [22].

[4]Some names have been shortened in favor of succinctness, with full definitions available in the accompanying hyperlink

There are several terms related to the response to virus by the host. In particular, *positive regulation of defense response to virus by host* (GO:0002230), *regulation of defense response to virus by host* (GO:0050691), *defense response to symbiont* (GO:0140546) and *defense response to virus* (GO:0051607), although these are only present within the Random Forest selected genes. It is also worth noting once again the abundance of interferon related processes, as well as some cytokine related terms. Among these, *negative regulation of cytokine production* (GO:0001818) and *positive regulation of cytokine production* (GO:0001819), which are contradicting, may indicate an attempt to modulate the immune response by the cell or potentially a mechanism of the virus to defend itself from the immune response.

The term *RIG-I signaling pathway* (GO:0039529) which is associated with the Pattern Recognition Receptor RIG-I, and the term *cytoplasmic pattern recognition receptor signaling pathway in response to virus* (GO:0039528) were also identified in Table IV as well as for NHBE cells using the Random Forest algorithm. These receptors play crucial roles in the detection of viruses by cells and the resulting signaling cascade, which in turn leads to the production of Type I interferons and pro-inflammatory cytokines [23].

C. Biclustering

In order to allow for the detection of more complex patterns, we now present the results of applying several biclustering algorithms to our data. In particular, these algorithms, unlike clustering, can identify patterns which span only certain conditions. This means that by analyzing the resulting biclusters and functionally enriching them, we can obtain processes associated with any particular subset of conditions.

We begin in Table V by presenting several metrics for each algorithm and preprocessing option used. It is important to note that $|\mathcal{B}|$ corresponds to the number of biclusters; $\overline{|I|}$ corresponds to the average number of genes per bicluster; $\sigma_{|I|}$ corresponds to the standard deviation of genes per bicluster; $\overline{|J|}$ corresponds to the average number of conditions per bicluster; $\sigma_{|J|}$ corresponds to the standard deviation of the number of conditions per bicluster; and finally $\overline{\text{Terms}}$ corresponds to the average number of enriched terms per bicluster. BicPAMS and Cheng and Church present the highest average number of biclusters, with the Plaid and xMotifs algorithms significantly less for most preprocessing conditions. It is also important to note that BicPAMS selects a larger amount of genes for a much smaller amount of conditions. This is particularly relevant to better understand the comparatively much larger amount of average enriched terms per bicluster with BicPAMS, since having too many conditions can lead to the identification of more generic genes and having too few genes can lead to the identification of less significant processes.

Using these methods, we obtain a set of biclusters, each consisting of a subset of genes and a subset of conditions. By performing functional enrichment on these genes, a set of biological processes associated with those genes is then produced. In order to analyze these results and obtain a more generic view of how often certain processes occur for each condition, a count is performed for each process identified. This allows for the identification of the most commonly occurring processes, and thus provides a better view of which processes are most closely related with a certain condition, while also potentially reducing the amount of more generic biological processes. In addition to this, it provides a direct element of comparison between different cell types for the same condition, or between the same cell type and different viruses. In addition to the number of occurrences of each process, the best c-score and p-value are also provided, in order to compare the statistical relevance of different processes.

TABLE V
METRICS FOR COMPARING THE PERFORMANCE OF THE TESTED BICLUSTERING ALGORITHMS WITH DIFFERENT PREPROCESSING TECHNIQUES

Algorithm	Preprocessing	$\lvert\mathcal{B}\rvert$	$\overline{\lvert I\rvert}$	$\sigma_{\lvert I\rvert}$	$\overline{\lvert J\rvert}$	$\sigma_{\lvert J\rvert}$	$\overline{\text{Terms}}$
BicPAMS	$p < 0.01$	80	208.03	18.54	3.16	0.53	28.91
	$p < 0.05$	79	3526.66	301.50	3.24	0.64	341.70
	ANOVA (top 200)	7	188.29	5.95	10.00	9.70	10.57
	ANOVA (top 1000)	20	676.05	29.13	5.00	4.22	55.75
	ANOVA (top 5000)	57	2106.18	128.36	3.61	1.25	131.32
Cheng and Church	$p < 0.01$	50	15.60	12.59	12.92	5.90	3.46
	$p < 0.05$	100	55.90	16.23	34.79	9.96	1.68
	ANOVA (top 200)	8	25.00	23.49	21.38	12.56	6.50
	ANOVA (top 1000)	56	17.86	15.27	17.89	10.67	4.41
	ANOVA (top 5000)	100	34.54	24.10	22.76	11.25	2.47
Plaid	$p < 0.01$	10	64.70	55.53	14.20	5.60	29.90
	$p < 0.05$	10	776.40	922.43	11.60	6.89	24.10
	ANOVA (top 200)	8	44.00	30.76	12.88	8.43	9.88
	ANOVA (top 1000)	10	159.50	100.72	12.20	7.29	18.70
	ANOVA (top 5000)	10	739.20	530.04	13.10	7.48	43.40
xMotifs	$p < 0.01$	10	31.90	17.17	8.20	2.86	1.70
	$p < 0.05$	10	654.50	365.82	6.00	0.00	6.30
	ANOVA (top 200)	6	30.33	34.30	24.50	9.73	10.67
	ANOVA (top 1000)	10	71.90	103.54	11.10	4.28	7.60
	ANOVA (top 5000)	10	326.00	538.95	6.20	0.60	5.30

We now proceed to a comparative analysis of the biological processes associated with SARS-CoV-2 for all cell types, using biclustering. In order to provide an ordering for the processes taking into account all cell types, each enriched term is first ranked by the number of occurrences it has related to a given condition. Then a fused rank is computed by multiplying the resulting ranks. The multiplication allows for a higher penalization of terms which contain a single very low rank but high ranks for other cell types.

There are several identified processes which have been previously described with clustering and predictive models. In particular, there are multiple terms related to cytokine activity, for instance *cytokine-mediated signaling pathway* (GO:0019221), which possesses a high number of occurrences for A549 (1.00), NHBE (0.75) and Calu3 (1.00) cells and a lower count for Biopsy cells (0.60). It is interesting to note a seeming tendency for the normalized number of occurrences for Biopsy cells to be lower for most processes, with more generic DNA related processes, such as *DNA metabolic process* (GO:0006259), *DNA repair* (GO:0006281) and *cellular response to DNA damage stimulus* (GO:0006974), possessing higher values. This may be due to biopsy results possibly containing multiple cell types as well as due to the very

low number of samples of this type of cell (2 healthy and 2 infected).

Other cytokine associated processes include *cellular response to cytokine stimulus* (GO:0071345), *chemokine-mediated signaling pathway* (GO:0070098) followed also by *cellular response to chemokine* (GO:1990869). Chemokines in particular play an important role in multiple processes related with host immune response against viral infection, namely the attraction of leukocytes to the infected tissue. The presence of the terms *neutrophil mediated immunity* (GO:0002446), *neutrophil activation involved in immune response* (GO:0002283) and *neutrophil degranulation* (GO:0043312), further supports this hypothesis. Neutrophils are leukocytes which are the first responders to sites of infection, and have also been identified as the main infiltrating cell population in IAV infection [31]. Despite containing somewhat lower counts than other processes, this set of enriched terms still possess p-values and c-scores well within the range of statistical significance.

Another previously identified set of processes which is also present is interferon related terms. Interferons are a potent type of cytokines which are associated with antiviral response, with most viruses having developed adaptations to at least partially avoid this mechanism [33]. In particular, *cellular response to interferon-gamma* (GO:0071346) and *interferon-gamma-mediated signaling pathway* (GO:0060333).

We now proceed to a comparative analysis of the processes associated with different viruses. There are many processes in common with the SARS-CoV-2 analysis, which is to be expected, since most identified processes are related to immune response.

cellular response to interferon-gamma (GO:0071346) has somewhat fewer occurrences when compared to the other viruses (0.66 vs 0.85 for RSV, 0.92 for HPIV3 and 0.86 for IAV). *cytokine-mediated signaling pathway* (GO:0019221) has a somewhat higher number of occurrences for SARS-CoV-2 and HPIV than others (1.00 and 1.00 vs 0.74 for RSV and 0.92 for IAV). *inflammatory response* (GO:0006954) is somewhat muted for SARS-CoV-2 when compared to the other viruses, for both A549 (0.45 vs 0.97 for RSV, 0.92 for HPIV3, 1.00 for IAV) and NHBE cells (0.75 vs 0.91 for IAV, 1.00 for IAVdNS1). These differences are consistent with those found by Blanco-Melo D. et al. [1], who found SARS-CoV-2 to induce a limited interferon response when compared with the other viruses but a strong production of cytokines and resulting processes. Overall, there seems to be a tendency for the other viruses to have comparatively higher counts, especially IAV.

In Table VI, we can see a compilation of the number of GO Biological Processes detected for each of the applied methods. As we can see, biclustering provided, by a considerable margin, a highest amount of biological processes, followed by clustering. The predictive models provided the worst results, with Random Forests providing somewhat better results for the Multi-Condition Setting as well as for NHBE cells. Overall, these results seem to suggest pattern-based algorithms are better suited for this application.

TABLE VI
NUMBER OF PROCESSES FOUND, FOR DIFFERENT p VALUES, FOR EACH OF THE METHODS APPLIED. MCS - MULTI-CONDITION SETTING.

Method	Setting	Number of GO Biological Processes		
		$p < 0.05$	$p < 0.01$	$p < 0.001$
Clustering	MCS (p <0.01)	463	215	76
	NHBE	234	75	20
	A549	182	38	19
Random Forests	MCS (p <0.01)	215	109	44
	NHBE	110	22	3
	A549	21	0	0
xGBoost	MCS (p <0.01)	60	41	15
	NHBE	34	0	0
	A549	36	0	0
BicPAMS	MCS (p <0.01)	4440	2086	1184
	NHBE	2912	685	305
	A549	3926	779	273

IX. CONCLUSION

This dissertation proposed a set of novel principles to identify putative regulatory modules associated with the response to SARS-CoV-2, while also presenting an analysis of the biological processes associated with them, as well as a comparison to other viruses. A particular focus was placed on the relevance of pattern-centric views for gene set enrichment analysis. The source of data used is an RNASeq dataset which provides gene expression levels for a set of genes and samples, healthy and infected by SARS-CoV-2 and other viruses.

A novel methodology was proposed combining different approaches, which when consolidated provide a more robust view of the putative processes associated with the infection by SARS-CoV-2. In particular, the complete gene set is initially filtered using a Mann-Whitney U Test, which allows for the selection of genes with statistically relevant differences in expression between healthy and infected cells.

Other authors perform feature enrichment directly on the set of genes obtained using simplistic statistical tests. However, this stance results in a smaller amount of biological processes detected, as well as a decrease in their quality (measured using Fisher's Exact Test and the combined c-score). So a three-fold, pattern-centric approach was proposed, using hierarchical clustering, decision tree based predictive algorithms and bi-clustering algorithms on the resulting genes to identify groups of genes with correlated expression. With both clustering and predictive algorithms, a mostly individual approach was taken, separating cell type and analyzing only two conditions at a time.

Under this methodology, we were able to validate and identify potentially novel biological processes associated with SARS-CoV-2 infection. Among the various enriched terms, the high cytokine induction, Type I interferon related terms, as well as signaling pathways related to these were reoccurring in all analysis performed. Additionally, comparing these results to existing literature on SARS-CoV-2, other viruses and also on the biological function of certain terms

allowed for the identification of characteristics of the disease. In particular, SARS-CoV-2 was found to induce a limited interferon response when compared with the other viruses but a strong production of cytokines and associated processes (namely interferon induction and response to these stimuli). These findings were consistent with Blanco-Melo D. et al. [1]. Additionally, we found in multiple analysis the involvement of Pattern Recognition Receptors (with particular emphasis on RIG-I) in the process of infection. This was not identified by Blanco-Melo D. et al., however it is consistent with other literature on coronaviruses, and further supports the hypothesis that a pattern-centric view of the gene enrichment process can result in novel information.

X. Future Work

As potential directions for future work, we suggest the:

- application of this methodology to different SARS-CoV-2 datasets to cross-validate, expand and improve the robustness of the provided findings;
- application of this methodology to datasets pertaining to other viruses, to better assess it's capability to offer new insights into the unique biological processes associated with each virus;
- addressing of the issue of sample interdependence by:
  - obtaining more samples by using other RNASeq datasets, which allows for the underlying relationships between the different conditions to be diluted;
  - designing a novel biclustering approach more tailored to this type of data, which takes into account the underlying relationships between each set of experimental conditions.

XI. Code Availability

The code utilized to obtain the presented results can be obtained in the following GitHub repository: https://github.com/PRodrigues98/Analysis-of-regulatory-response-to-SARS-CoV-2-infection. It utilizes python 3.8 mainly with the NumPy, pandas, scikit-learn and matplotlib libraries.

References

[1] D. Blanco-Melo, B. E. Nilsson-Payant, W.-C. Liu, *et al.*, "Imbalanced host response to sars-cov-2 drives development of covid-19," *Cell*, 2020.

[2] M. F. Murray, E. E. Kenny, M. D. Ritchie, *et al.*, "Covid-19 outcomes and the human genome," *Genetics in Medicine*, pp. 1–3, 2020.

[3] R. L. Tillett, J. R. Sevinsky, P. D. Hartley, *et al.*, "Genomic evidence for reinfection with sars-cov-2: A case study," *The Lancet Infectious Diseases*, 2020.

[4] M. Frieman and R. Baric, "Mechanisms of severe acute respiratory syndrome pathogenesis and innate immunomodulation," *Microbiology and Molecular Biology Reviews*, vol. 72, no. 4, pp. 672–685, 2008.

[5] S. A. Ochsner, R. T. Pillich, and N. J. McKenna, "Consensus transcriptional regulatory networks of coronavirus-infected human cells," *Scientific Data*, vol. 7, no. 1, pp. 1–20, 2020.

[6] J. Wei, M. Alfajaro, R. Hanna, *et al.*, "Genome-wide crispr screen reveals host genes that regulate sars-cov-2 infection," *Biorxiv*, 2020.

[7] E. Wyler, K. Mösbauer, V. Franke, *et al.*, "Bulk and single-cell gene expression profiling of sars-cov-2 infected human cell lines identifies molecular targets for therapeutic intervention," *bioRxiv*, 2020.

[8] B. K. Manne, F. Denorme, E. A. Middleton, *et al.*, "Platelet gene expression and function in patients with covid-19," *Blood, The Journal of the American Society of Hematology*, vol. 136, no. 11, pp. 1317–1329, 2020.

[9] J. Golden, C. Cline, X. Zeng, *et al.*, "Human angiotensin-converting enzyme 2 transgenic mice infected with sars-cov-2 develop severe and fatal respiratory disease," *bioRxiv*, 2020.

[10] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

[11] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[12] M. J. Blanca Mena, R. Alarcón Postigo, J. Arnau Gras, R. Bono Cabré, and R. Bendayan, "Non-normal data: Is anova still a valid option?" *Psicothema, 2017, vol. 29, num. 4, p. 552-557*, 2017.

[13] Y. Cheng and G. M. Church, "Biclustering of expression data.," in *Ismb*, vol. 8, 2000, pp. 93–103.

[14] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," *Statistica sinica*, pp. 61–86, 2002.

[15] T. Murali and S. Kasif, "Extracting conserved gene expression motifs from gene expression data," in *Biocomputing 2003*, World Scientific, 2002, pp. 77–88.

[16] R. Henriques, F. L. Ferreira, and S. C. Madeira, "Bicpams: Software for biological data analysis with pattern-based biclustering," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–16, 2017.

[17] E. Y. Chen, C. M. Tan, Y. Kou, *et al.*, "Enrichr: Interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–14, 2013.

[18] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, *et al.*, "Enrichr: A comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, W90–W97, 2016.

[19] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[20] "The gene ontology resource: Enriching a gold mine," *Nucleic Acids Research*, vol. 49, no. D1, pp. D325–D334, 2021.

[21] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[22] Y. Liang, M.-L. Wang, C.-S. Chien, *et al.*, "Highlight of immune pathogenic response and hematopathologic effect in sars-cov, mers-cov, and sars-cov-2 infection," *Frontiers in immunology*, vol. 11, p. 1022, 2020.

[23] E. De Wit, N. Van Doremalen, D. Falzarano, and V. J. Munster, "Sars and mers: Recent insights into emerging coronaviruses," *Nature Reviews Microbiology*, vol. 14, no. 8, pp. 523–534, 2016.

[24] J. Melchjorsen, L. N. Sørensen, and S. R. Paludan, "Expression and function of chemokines during viral infections: From molecular mechanisms to in vivo function," *Journal of leukocyte biology*, vol. 74, no. 3, pp. 331–343, 2003.

[25] J. S. Rawlings, K. M. Rosler, and D. A. Harrison, "The jak/stat signaling pathway," *Journal of cell science*, vol. 117, no. 8, pp. 1281–1283, 2004.

[26] Y.-C. Perng and D. J. Lenschow, "Isg15 in antiviral immunity and beyond," *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 423–439, 2018.

[27] P. K. Pribul, J. Harker, B. Wang, *et al.*, "Alveolar macrophages are a major determinant of early responses to viral lung infection but do not influence subsequent disease development," *Journal of virology*, vol. 82, no. 9, pp. 4441–4448, 2008.

[28] C. Schneider, S. P. Nobs, A. K. Heer, *et al.*, "Alveolar macrophages are essential for protection from respiratory failure and associated morbidity following influenza virus infection," *PLoS pathogens*, vol. 10, no. 4, e1004053, 2014.

[29] A. R. French and W. M. Yokoyama, "Natural killer cells and viral infections," *Current opinion in immunology*, vol. 15, no. 1, pp. 45–51, 2003.

[30] H. F. Rosenberg, K. D. Dyer, and J. B. Domachowske, "Eosinophils and their interactions with respiratory virus pathogens," *Immunologic research*, vol. 43, no. 1-3, pp. 128–137, 2009.

[31] I. E. Galani and E. Andreakos, "Neutrophils in viral infections: Current concepts and caveats," *Journal of leukocyte biology*, vol. 98, no. 4, pp. 557–564, 2015.

[32] S. L. Deshmane, S. Kremlev, S. Amini, and B. E. Sawaya, "Monocyte chemoattractant protein-1 (mcp-1): An overview," *Journal of interferon & cytokine research*, vol. 29, no. 6, pp. 313–326, 2009.

[33] G. C. Sen, "Viruses and interferons," *Annual Reviews in Microbiology*, vol. 55, no. 1, pp. 255–281, 2001.