TÉCNICO LISBOA

Regulatory response to infection by SARS-CoV-2

Pedro Miguel Santos Pereira Rodrigues

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Rui Miguel Carrasqueiro Henriques
Dr. Rafael Sousa Costa

Examination Committee

Chairperson: Prof. José Carlos Martins Delgado
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Andreia Sofia Monteiro Teixeira

October 2021

Acknowledgments

Em primeiro lugar, quero agradecer aos meus orientadores Rui Henriques e Rafael Costa. As nossas reuniões semanais e trocas de emails foram fundamentais para que conseguisse desenvolver este trabalho. A sua disponibilidade para esclarecer dúvidas, paciência com o trabalho que fui desenvolvendo e abertura para conversar foram extremamente importantes para que pudesse não só desenvolver esta dissertação como também decidir o que quero fazer no futuro.

Quero também agradecer aos meus pais, cujo apoio foi total e imprescindível para que pudesse desenvolver este trabalho e ter sucesso em todo o meu percurso académico.

Finalmente gostaria de agradecer aos meus amigos e colegas, em especial à Joana, que com troca de ideias, apoio e partilha de experiências tornaram este processo muito mais fácil e agradável, e contribuíram também muito para melhorar a qualidade desta dissertação.

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.

Keywords

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

Resumo

COVID-19, a doença causada pelo novo coronavírus, SARS-CoV-2, afetou já mais de 241 milhões de pessoas e causou a morte de mais de 4.9 milhões. No entanto, o conhecimento relativo ao impacto deste vírus sobre as células infetadas é ainda incompleto. Portanto, este trabalho procura identificar e analisar os principais processos regulatórios afetados e induzidos pelo SARS-CoV-2, utilizando dados transcriptómicos publicamente disponíveis, respetivos a várias linhas celulares suscetíveis à infeção. Propomos um novo conjunto de modelos estatísticos para lidar com três principais desafios, nomeadamente o reduzido número de amostras, a elevada dimensionalidade dos dados e ainda a complexidade das interações entre genes. Adicionalmente, analisamos a função destes genes e as suas interações nas células, comparando-os aos afetados por IAV (H1N1), RSV e HPIV nas linhas celulares analisadas. Os resultados obtidos mostram que o uso de clustering, biclustering e modelos preditivos aumentam significativamente o número e qualidade dos processos biológicos detetados. Adicionalmente, é feita uma análise comparativa destes processos, de forma a identificar potenciais características patofisiológicas da Covid-19. Estes são ainda comparados aos identificados por outros autores para o mesmo vírus e ainda para outros relacionados, como o SARS-CoV-1. Esta abordagem é particularmente relevante pela falta de outros trabalhos que utilizem ferramentas mais complexas de Machine Learning neste contexto.

Palavras Chave

COVID-19; SARS-CoV-2; Padrões Regulatórios Discriminativos; Biclustering; Transcriptómica; Modelação de Dados de Expressão de Genes.

Contents

List of Figures

List of Tables

1

Introduction

Contents

The infection of humans by the 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 present 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 [2].

Despite the ongoing breakthroughs, the cellular responses to SARS-CoV-2 are still being explored. 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 explore the utility of more complex techniques such as clustering, predictive models and biclustering to aid in the identification of differentially expressed genes and related biological processes.

1.1 Major Contributions

The main problems to be addressed by this dissertation are:

1. Identification of the main biological processes involved in the infection process, which allows for a better understanding of:
    - The viral life-cycle and interactions with the cell;
    - The defence mechanisms employed by the cell against the virus;
2. Verify whether clustering, predictive modeling and biclustering can aid in the identification of biological functions from transcriptomic data.

Thus, the main contributions of this dissertation are the application of machine learning algorithms, with particular emphasis on biclustering, within the context of the identification of biological functions in gene expression data, as well as the analysis of the main biological functions involved in the infection

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

by SARS-CoV-2. Additionally, the comparison between different biclustering algorithms according to several metrics when applied in this setting, which gives a better idea of which algorithms provide the best results for this use case.

The analysis and identification of biological processes is particularly relevant since it allows for an easier and faster identification of the characteristics of a particular disease. Since sequencing technologies are becoming increasingly cheap and available, these methods could allow for more prompt response to the threat of new viruses, by revealing characteristics of the disease without directly requiring the study of infected individuals. In particular, the usage of clustering and biclustering significantly aids the identification of biological processes by identifying underlying patterns comprising multiple genes and thus supporting the association of those to corresponding biological functions.

1.2 Organization of the Document

The organization of the remaining document is as follows: chapter 2 introduces the concepts from biology and machine learning which support the rest of the dissertation; chapter 3 provides a selection of related work; chapter 4 explores some preliminary information on the analyzed dataset; chapter 5 presents the methodology proposed to achieve the previously mentioned contributions; chapter 6 presents the obtained results, including an analysis of the identified biological processes. Finally, chapter 7 finishes this dissertation with the concluding remarks and future work.

2

Background

Contents

The present section introduces concepts from two broad areas of knowledge: biology and machine learning. Accordingly, the background is divided into two subsections: SARS-CoV-2 Biology (section 2.1) and Data Modeling and Statistical Analysis (section 2.2).

2.1 SARS-CoV-2 Biology

2.1.1 Cell Biology and Behavior

Genes are the underlying units responsible for encoding information necessary to cell function and are constituted by DNA. An important consideration regarding most cellular functions is their dependence on fine interactions to produce a given result. This means that, when trying to identify which genes are responsible for a certain behavior, a modular view is generally necessary.

For the cell to respond to internal needs and external stimuli, it produces RNA molecules through the process of DNA transcription which, depending on their type, can then be translated into proteins. In particular, RNA can be divided into coding RNA, which directly encodes proteins, and non-coding RNA (ncRNA), whose function will be expanded later in this section. There are several sub-types of RNA with different functions, including messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). mRNA is responsible for the transfer of genetic information from DNA in the nucleus to the ribossomes, to be translated into proteins. tRNA participates in protein synthesis by allowing for specific aminoacids, the constituents of proteins, to be brought to ribosomes to form proteins. rRNA is one of the main constituents of ribosomes. The transcription of DNA into RNA can vary throughout the life cycle of a cell, in an attempt to respond to it's environment. This process is mediated by three main factors: epigenetic markers (changes to the DNA structure which do not change the genetic sequence, but can affect gene activity), proteins (which, among many other functions, act as receivers of chemical signals, thus allowing the cell to respond to changes in both the environment and it's internal state) and other regulatory molecules, including ncRNA.

2.1.2 Viral Infection and SARS-CoV-2

A virion, an individual viral particle, consists of genetic material, which can be either DNA or RNA, enclosed by a protein shell. The process of viral infection of a cell begins with the viral particle entering the cell, which can happen, depending on the type of virus, in essentially two ways: receptor-mediated fusion or the endocytic pathway [3].

For viruses in the coronavirus family [4], both begin with a spike (S) protein in the surface of the virion binding to a receptor on the host cell's surface. The particular type of receptor varies as a function of the particular viral chain, the angiotensin-converting enzyme 2 (ACE2) for SARS-CoV-2 [5]. In receptor-

mediated fusion, TM protease serine 2 (TMPRSS2) activates of the S protein, promoting viral entry into the cell [6]. Viral RNA is thus introduced into the cell, resulting in the transcription of coding RNA into viral proteins as well as the inhibition of certain genes by ncRNAs. These processes ultimately result in the assembly of new virions which proceed to exit the cell [7]. This process is illustrated with more detail in Figure 2.1.

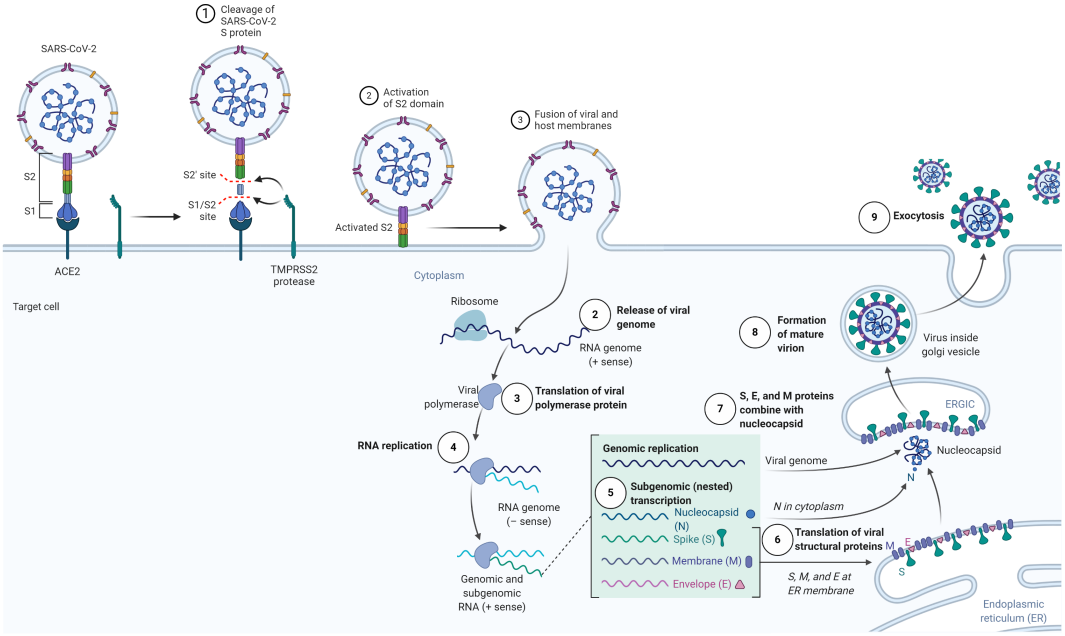

Figure 2.1: Diagram of the SARS-CoV-2 life cycle [1].

The knowledge of the receptor responsible for allowing the entry of SARS-CoV-2 into cells has allowed for the assessment of the risk of infection for various human organs, as well as the identification of potential targets for *in vitro* experiments. As we can see in Figure 2.2, the respiratory system has a particularly high risk of infection (with all organs from the nasal mucosa to the lungs having at least some risk of infection), which is consistent with the symptoms associated with COVID-19.

Non-coding RNA (ncRNA) is RNA which is not translated into proteins. It has various functions within cells, such as post-transcriptional gene regulation. These molecules, in particular microRNAs (miRNAs), a subtype of small ncRNAs, can play crucial roles in viral infection, such aiding viral replication by promoting changes in the cellular life cycle and interfering with immune responses [8]. Additionally, since RNA viruses co-opt the cell's ribosomes to produce the viral proteins necessary for viral replication, miRNAs in the host cell can prevent the translation of these proteins as well as potentially degrade viral RNA. These types of mechanisms have been proposed as a potential explanation for the lack of clinical symptoms of infection in bats [9], by limiting the extent of viral infection. These animals have been

[1] Adapted from "Coronavirus Replication Cycle" and "Mechanism of SARS-CoV-2 Viral Entry", by BioRender.com (2020). Retrieved from https://app.biorender.com/biorender-templates.

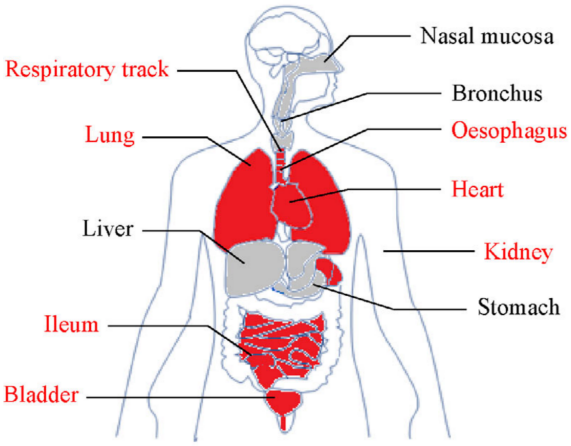

Figure 2.2: Vulnerable organs to SARS-CoV-2 infection, with higher-risk organs in red and lower-risk in gray (Zou et al. [1]).

proposed as a potential original host of SARS-CoV-2 [10] and established as a major source of novel infectious diseases [9].

2.1.3 Transcriptomics

Transcriptomics consists in the study of the transcriptome, the complete set of RNA transcripts that are produced by a given cell under a set of circumstances. This information also includes the amounts of transcripts present for each gene, thus allowing not only for the study of which genes are expressed, but also the levels of expression of each gene.

This is generally achieved by the usage of high throughput methods, such as microarray analysis [11] and deep-sequencing, with RNA-Seq being a technology of particular interest [12]. RNA-Seq is a quantitative and high-throughput sequencing technique, which means, when comparing to alternative methods such as microarrays, it can determine RNA expression levels more accurately, especially for lowly and highly expressed genes. Additionally, it removes the necessity to know the target genome *a priori*.

The analysis of the transcriptome allows the tracking of changes to gene transcription in cells caused by certain conditions, in particular changes caused by viral infection. This technique can be used, for instance, to identify genes that positively or negatively correlate to clinical outcomes, which can then be conducive to the development of novel therapies and thus the potential improvement of patient care [13]. Additionally, by applying functional enrichment analysis, which consists in the combination of previously available information about the function of certain groups of genes, it is possible to identify the biologic mechanisms involved in the response to a certain disease.

Since the transcriptome associated with SARS-CoV-2 is widely known, our work in particular is fo-

cused on the regulatory response of the cell to infection, not on the level of replication of the viral transcriptome or the translation of proteins which compose the virus.

2.2 Data Modeling and Statistical Analysis

Transcriptomic data is generally available in two distinct formats: absolute, with RNA counts for thousands of genes, or relative, against reference gene expression values. The data can refer to cell cultures or to cells obtained, through biopsy or autopsy, directly from target organisms. Cell cultures have the advantage of easier control of experimental conditions, as well as allowing the selection of specific cell lines with known characteristics. However, it can be hard to generalize certain models from cell cultures to the complete organism. For instance, hydroxychloroquine was shown to be effective at reducing viral load *in vitro* [14], which lead to an increase in intake, despite it's efficacy in patients not proving to be the same [15].

When testing for different conditions or phenotypes, samples can be annotated, where if we take X as the set of all samples, X_k is the subset of samples with annotation k. The dataset is then costructed as a set of samples of genes $Y = y_1, ..., y_m$, where each entry a_{ij} corresponds to the expression level (relative or absolute) of gene $y_j \in Y$ for sample $\mathbf{x}_i \in X$.

This type of data poses a particular set of challenges, since it has high dimensionality (high m) and, generally, a low number of samples [16]. Additionally, it can be highly skewed, with many genes having null or close to null transcription, and very few having very high transcription, thus requiring the application of preprocessing techniques.

To identify differentially expressed genes, potentially related to the biological mechanisms underlying the disease, several approaches can be taken, namely the usage of statistical models to directly identify genes whose variance among samples is statistically abnormal and the application of classification models to the data and subsequent analysis of the genes used by these.

Additionally, clustering techniques [17] and biclustering [18] techniques can be applied to identify patterns within samples and genes.

Finally, dimensionality reduction techniques can be applied to the data as a preprocessing method, as well as to visualize the differences between the various viral strains and the non-infected cells.

2.2.1 Gene Relevance

To assess gene relevance and identify differentially expressed genes, statistical tests which compare two populations can be used. There are two main types depending on the nature of the data. If certain parameters of the distribution of both populations can be assumed, a parametric test can be used, otherwise a non-parametric test is better suited [19].

A t-Student test [20] is a parametric test which compares two populations and identify if they differ from each other. As such, it can be used for several purposes, such as to evaluate the discriminatory power of a model, to compare the behavior of two models and to perform feature selection, with the latter being particularly relevant in the domain of transcriptomic data for the identification of differentially expressed genes. This type of test assumes that the scale of measurement follows a continuous or ordinal scale, that the sample is representative of the population and finally that the data approximately follows a normal distribution. It requires three data values, the standard deviation, the difference between means and the number of values of each sample. The result of the test is a t-value (T) and the degrees of freedom (df), which can then be compared to the t-distribution to obtain a p-value (level of significance). The type of distribution depends on the purpose of the test being applied, if it is to assess whether the mean of one of the populations is greater than the other, a one-tail distribution is appropriate, if it is to test if the two populations are different, a two-tail distribution should be used. There are three types of t-tests, depending on the characteristics of the data:

- ***Dependent t-test*** - Assumes the samples are paired and have the same size and variance,

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\frac{s_{diff}}{\sqrt{N}}}, \tag{2.1}$$

$$df = N - 1,\ N = N_1 = N_2, \tag{2.2}$$

  where N_k is the number of elements of the sample set X_k, $\overline{X}_k$ is the corresponding mean and s_{diff} is the standard deviation of the difference between both sets.

- ***Pooled t-test*** - Used when samples are independent, assuming both have the same size or variance,

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(N_1-1)*s_1^2+(N_2-1)*s_2^2}{N_1+N_2-2}} \times \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}, \tag{2.3}$$

$$df = N_1 + N_2 - 2, \tag{2.4}$$

  where s_k is the standard deviation of each sample.

- ***Unequal Variance t-test*** - Used when samples are independent and have unequal variance and/or sample size,

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{2.5}$$

$$df = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}}. \tag{2.6}$$

The aforementioned tests have non-parametric alternatives, which should be used when the distribution of the data cannot be assumed to be normal. It is important to note that, in cases where the assumptions can be made, parametric tests generally offer better results.

- ***Wilcoxon signed-rank test*** - assumes the two samples are paired (i.e. if we consider the annotations $k = 1, 2$, each pair would be $(a_{i_1 j}, a_{i_2 j})$, with $\mathbf{x}_{i_1} \in X_1, \mathbf{x}_{i_2} \in X_2$, y_j a specific gene and $\#X_1 = \#X_2$). To perform this test, we must begin by removing pairs whose difference is zero, i.e. $|a_{i_2 j} - a_{i_1 j}| = 0$, with N_r being the new number of pairs. Then the pairs should be assigned a rank, $R_{i_1 i_2}$, starting by the one with the smallest absolute difference being assigned 1. If multiple pairs have equal absolute differences, all pairs should be assigned a rank equal to the median of the ranks these pairs span. Finally the test statistic can be calculated as follows:

$$W = \sum_{i=1}^{N_r} \left[sgn(a_{i_2 j} - a_{i_1 j}) \times R_{i_1 i_2} \right], \tag{2.7}$$

  where W is the test statistic, which can either be compared to the exact distribution, or, if $N > 20$ to the Normal Distribution to obtain a p-value.

$$sgn(x) = \begin{cases} -1, & \text{if } x < 0 \\ 0, & \text{if } x = 0 \\ 1, & \text{if } x > 0 \end{cases}. \tag{2.8}$$

- ***Mann-Whitney U test*** - can be applied to independent samples. To apply this test to a pair of samples, we start by joining both samples in a single set and assigning ranks to them, starting with 1 for the smallest value and, in case of ties, the median of the ranks the tied values span. Then, we obtain the sum of the ranks ($R_k,\ k = 1, 2$) for each sample (note that, by obtaining one of these values and since there are N ranks, the remaining value can be directly calculated). Finally, the test statistic can be computed as the lowest, for $k = 1, 2$, of:

$$U_k = R_k - \frac{N_k \left(N_k + 1\right)}{2}. \tag{2.9}$$

Dimensionality reduction techniques allow a dataset to be mapped onto a lower dimension space

[21]. These can be divided in two main types: unsupervised and supervised. Unsupervised methods do not use the labels associated to the samples to compute the transformation, as opposed to supervised ones which do.

These techniques include:

- ***Principal Component Analysis (PCA)*** - PCA is an unsupervised technique that computes the set of components which explain the highest possible variance of the data, utilizing the eigenvectors and eigenvalues of the covariance matrix. Since it is sensitive to differences in the variance of the input variables, it is important to normalize the data prior to using this technique.
- ***Linear Discriminant Analysis (LDA)*** - LDA is a supervised technique which calculates the linear combination of input variables that result in the highest possible separation between classes.
- ***Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)*** - UMAP is a recently developed manifold learning technique for dimensionality reduction [22], which can be utilized in both a supervised and unsupervised manor. It has provided good results when applied to single-cell RNA data [21].

2.2.2 Classification

Classification models map observations to a set of possible categorical values, called labels. Multiple use cases exist for these models with transcriptomic data, including, for instance, classification of clinical outcomes from transcriptomic data [23], identification of cell-type specific genes and expression rules [24], among others. In our work, the main focus is in the identification of gene groups from transcriptomic data.

Focusing on the class of associative (pattern centric) approaches to classification, decision tree classifiers are composed by decision nodes and leaves. Decision nodes are conditions which divide the data into multiple subsets, whereas leaves identify the label attributed to a particular pattern, described by the set of decision nodes in the path between the root of the tree and that leaf. The tree is usually built top-down by, at each node, selecting the variable which best divides the input data, according to a certain metric. If the depth of the tree is too large, this type of classification method is susceptible to over-fitting (i.e. fitting the data too tightly and thus becoming unable to correctly classify new inputs), and if it is too small to under-fitting (i.e. fitting the data too loosely, resulting in low accuracy). These issues can be addressed by using ensemble methods, which use a collection of simpler models, such as decision trees, combining them and allowing for improved performance and robustness, especially when dealing with overfitting.

xGBoost is an ensemble classification model based on decision trees that uses gradient boosting [25] to iteratively improve upon previously built trees.

Random Forests are also an ensemble learning technique based on decision trees, which uses bagging as opposed to gradient boosting in xGBoost. This consists in building trees based on the random sampling with replacement of the data, which results in different trees. This allows for errors in each individual tree to be masked by the remaining ones.

2.2.3 Clustering and Biclustering

Clustering refers to the task of grouping a set of samples in a dataset, where the elements of any given group are more similar amongst each other, according to a given metric, than they are compared to elements of other groups. In other words, the similarity within each of these groups, called clusters, is maximized and the similarity between different groups is minimized. When analysing transcriptomic data, clustering techniques can either group samples according to the similarity between all genes or group genes according to the similarity of their inter-sample variation. The first view can be useful to verify if the clusters correspond to the different types of samples, namely if the clustering algorithm can successfully separate different viral strains and non-infected cells [26]. The second view, can be useful to detect if two different genes show similar levels of variation across all samples, which might indicate they perform related functions.

This method, however, has significant drawbacks, namely the fact that it considers all samples when grouping genes, which may hide possible relationships between genes across only some samples, and that it considers all genes when grouping samples, which can similarly hide relationships between between samples that span only a few genes. Additionally, each gene can only belong to one group, whereas genes can belong to several groups depending on their corresponding biological function. These limitations are addressed by biclustering, thus making it a good candidate for detecting patterns in this type of data [18]. More specifically, biclustering can be utilized to detect patterns of co-expression in genes, allowing a more direct comparison to available biological models, which associate sets of genes to given functions.

In particular, considering a data matrix, $D = (X, Y)$, with a set of rows $X = \{x_1, ..., x_n\}$, a set of columns $Y = \{y_1, ..., x_m\}$ and composed of elements $d_{ij} \in \mathbb{R}$ such that $i \in \{1, ..., n\}$ corresponds to a row and $j \in \{1, ..., m\}$ corresponds to a column. For gene expression data, a_{ij} represents the expression level of gene i in condition j. A bicluster $\mathcal{B} = (I, J)$ is an r by s submatrix of D, with $I = (i_1, ..., i_r) \subseteq X$ and $J = (j_1, ..., j_s) \subseteq Y$. Thus, the biclustering task is the identification of a set of biclusters, $\boldsymbol{\mathcal{B}}$, which satisfy a certain criteria of homogeneity.

Cheng and Church [27, 28] defined the criteria as a high similarity score. The score introduced was called mean squared residue, and it measures the coherence of the columns and rows in the bicluster. This algorithm assumes that each element a_{ij} can be defined by a background value a_{IJ}, a row constant a_{iJ} and a column constant a_{Ij}, as follows:

$$a_{ij} = a_{iJ} + a_{Ij} - a_{IJ} \tag{2.10}$$

Since the elements cannot be perfectly defined as above, a residue is defined such that:

$$a_{ij} = r(a_{ij}) + a_{iJ} + a_{Ij} - a_{IJ} \tag{2.11}$$

Then, the approach aims at minimizing the mean squared residue loss:

$$H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} r(a_{ij})^2 \tag{2.12}$$

The plaid algorithm [29] considers that the data matrix is composed as a sum of terms called layers, and proposes a homogeneity criteria to find biclusters in those conditions.

The xMotifs algorithm [30] aims to find conserved gene expression motifs, which correspond to biclusters. It begins by dividing the genes into a set number of states (if two states are considered, for instance, then these would be just upregulated and downregulated). Then it finds genes with a conserved state across multiple conditions.

The BicPAM algorithm [31] is capable of finding multiple types of biclusters. In this context, we choose constant patters, similar to xMotifs, since these types of patterns allow us to find, for instance, genes for which expression levels are similar for infected cells. This algorithm starts by normalizing and discretizing gene expression levels and then applying pattern mining techniques to obtain a set of biclusters. It further offers the option to extend, merge and filter the obtained biclusters.

3

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 [13] and the usage of the transcriptome of SARS-CoV-2 [32] to identify particular characteristics of the virus.

Blanco-Melo D. et al. [2] 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. [33] 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 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. [34] 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. [35] 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. [36] 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. [37] 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. [38] 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.

4

Exploratory Analysis

Contents

In the present chapter, we explore some preliminary information on the main dataset analysed in our work. Namely a description of the dataset, the method used to collect the data, how the method used impacts the data, and finally how particular characteristics of this dataset may affect the analysis.

4.1 Data Description

The target dataset, identified as GSE147507 [1], was collected by Blanco-Melo D. et al. [2] using RNA-Seq (a technique explained in subsection 2.1.3), 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. [2] 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) 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.

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

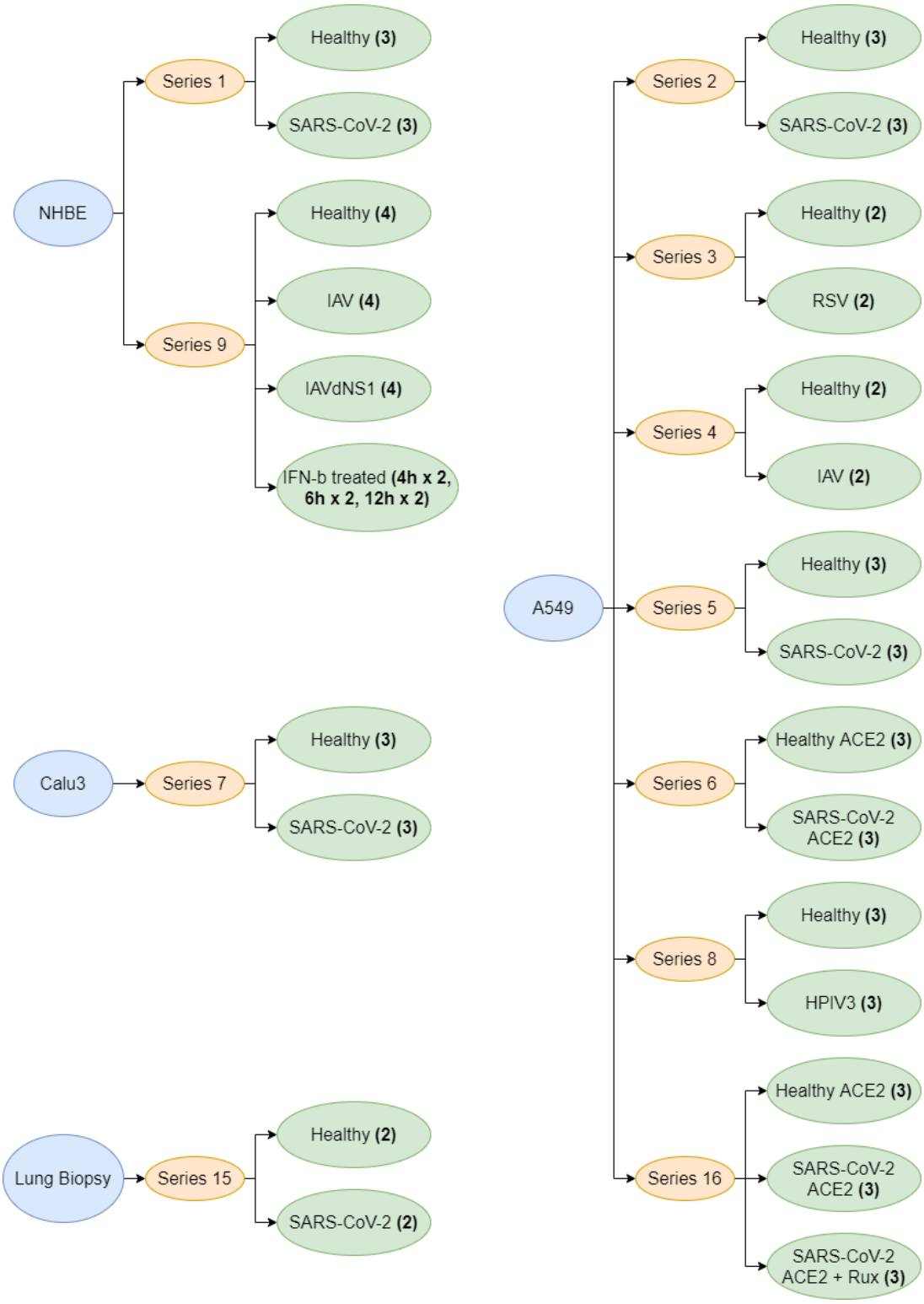

Figure 4.1: Overview of the structure of the dataset.

4.2 Preliminary analysis

4.2.1 Human data

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.

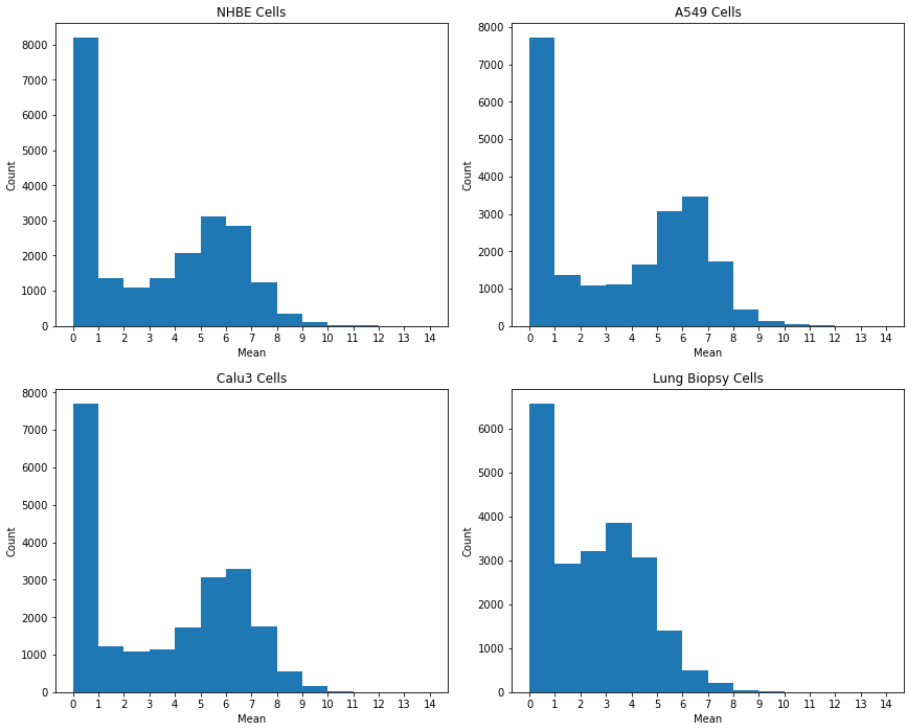

Figure 4.2: Distribution of gene expression (mean among samples) after applying a log transform (N = 21797 genes).

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 4.2).

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 4.3).

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 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.

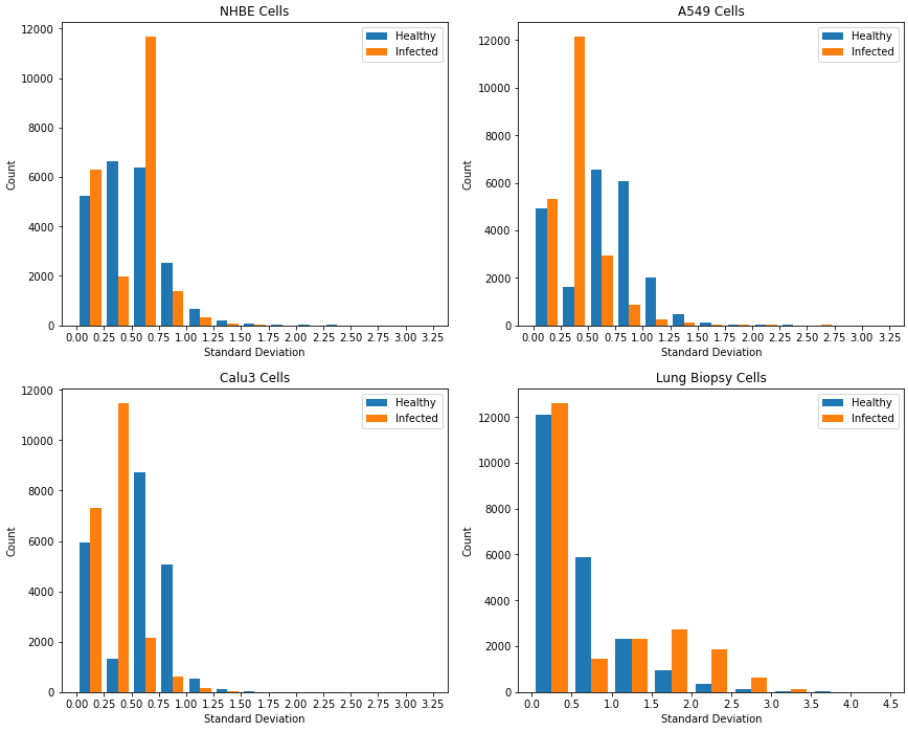

Figure 4.3: Standard deviation of gene expression within healthy and within infected cells.

We then proceeded to use LDA and PCA to visualize the data for NHBE cells and to verify if significant differences between healthy and infected cells could be observed.

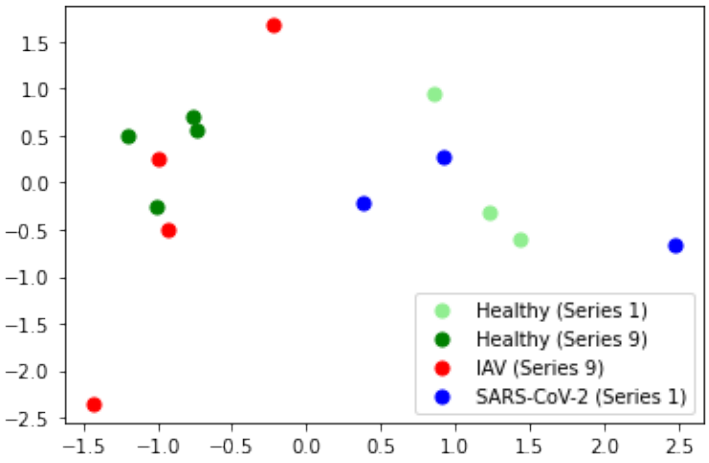

Figure 4.4: 2-Dimensional visualisation using LDA of the diferences between samples of NHBE cells.

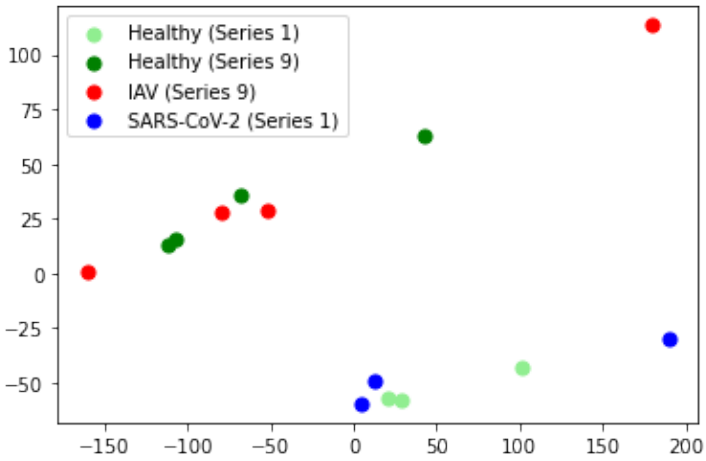

Figure 4.5: 2-Dimensional visualisation using PCA of the diferences between samples of NHBE cells.

From Figure 4.4 and Figure 4.5, we can observe that samples from the same experiment (which corresponds to a particular *Series*) tend to cluster together, with this tendency somewhat clearer in Figure 4.5 (for PCA). This indicates that samples within a given experiment are not truly independent, which was identified as one of the challenges in this work.

Table 4.1: 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

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 [39], 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 4.1). 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 [40] 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. The overlap between each set of non-normal genes is shown in Figure 4.6, with the the selected genes also provided for reference. These, further elaborated upon in chapter 5, consist of genes with

$p < 0.01$ for any pair of conditions in Table 4.1, using a Mann-Whitney U test. As we can see, there are 493 (63.2%) non-normal selected genes (spread across the different cell types) and 287 normal genes. Additionally, it is important to note that overall 32.8%, 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, (both explored in chapter 2) 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.

Figure 4.6: Venn diagram of overlapp between genes considered non-normal for each cell type.

5

Solution

Contents

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 5.1. In the present chapter, we motivate their need and explore each of the steps in more detail.

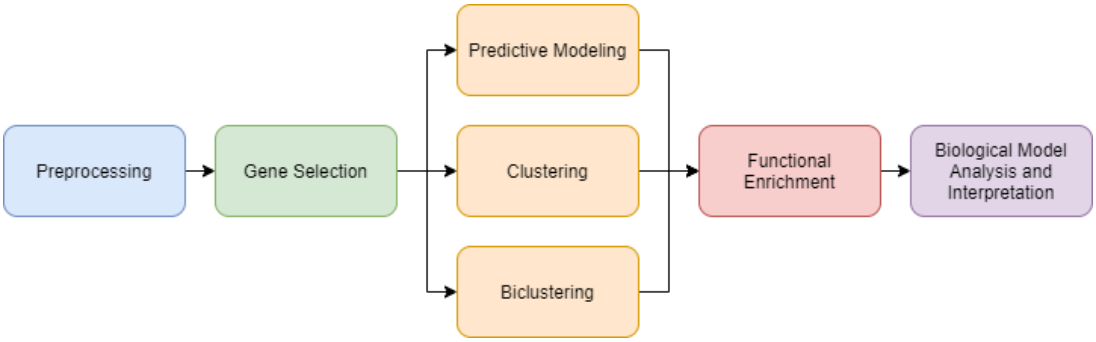

Figure 5.1: Schematic of the steps composing our proposed solution.

5.1 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 chapter 4), 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, as mentioned in chapter 2, 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 4.1. 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 [41].

5.2 Pattern Detection

The usage of complete data with a simple statistical pre-selection 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.

5.2.1 Clustering

The notion of cluster (subsection 2.2.3) 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.

5.2.2 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.

5.2.3 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 [27], plaid [29], xMotifs [30] and BicPAMS [42] algorithms.

5.3 Functional Enrichment and Biological Analysis

To obtain potential biological processes associated with the gene groups found using the aforementioned methods, we used the EnrichR [1] [43, 44]. 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

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

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 [44], 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 Biological Process knowledge base [45,46], 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) [47] 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.

6

Results

Contents

To solve the problem of identifying smaller and more internally correlated sets of DEG, as overviewed in the previous chapter, we use three methods, Clustering (section 6.1), Predictive Modeling (section 6.2) and Biclustering (section 6.3). These allow us, when performing functional enrichment, to identify more statistically significant biological processes.

In the present chapter, we present the results of applying 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 6.1, the result of performing functional enrichment on the set of genes obtained directly through preprocessing, in the previously (chapter 5) defined **Multi-Condition Setting**.

Table 6.1: Top 25 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
interferon-gamma-mediated signaling pathway (GO:0060333)	3.48E-08	211.38
cellular response to interferon-gamma (GO:0071346)	5.98E-10	192.56
cellular response to cytokine stimulus (GO:0071345)	6.64E-16	177.02
negative regulation of bone resorption (GO:0045779)	9.92E-03	164.52
positive regulation of tyrosine phosphorylation of STAT protein (GO:0042531)	1.39E-06	163.05
negative regulation of viral genome replication (GO:0045071)	2.50E-06	159.30
positive regulation of defense response (GO:0031349)	6.15E-08	155.70
regulation of tyrosine phosphorylation of STAT protein (GO:0042509)	1.39E-06	146.93
defense response to symbiont (GO:0140546)	3.48E-08	141.92
negative regulation of viral process (GO:0048525)	2.02E-06	138.13
response to interferon-gamma (GO:0034341)	2.11E-06	125.13
defense response to virus (GO:0051607)	1.19E-07	121.67
response to interferon-beta (GO:0035456)	8.82E-04	116.74
interleukin-7-mediated signaling pathway (GO:0038111)	3.67E-03	111.23
cellular response to interleukin-7 (GO:0098761)	3.67E-03	111.23
positive regulation of response to external stimulus (GO:0032103)	2.73E-07	110.34
protein kinase B signaling (GO:0043491)	2.10E-03	108.05

As we can see, there is a considerable number of processes with low p-value. However, the c-score is significantly lower when compared to the same genes after clustering (see Table 6.2). 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.

6.1 Clustering

In chapter 5, we mention the possibility that testing multiple conditions simultaneously, when using clustering, poses significant challenges. This is mainly due to this method being unsupervised, which means we cannot input the data labels to aid the algorithm in discovering more relevant patterns. Thus, we first seek to understand if, even with this limitation, the clustering algorithm can identify clear distinctions between healthy and infected cells from the various cell types simultaneously. To this end, we performed agglomerative clustering (using ward linkage) on the various samples for each condition, the results of which can be seen in the dendrogram presented in Figure 6.1. For the joined analysis (Multi-Condition Setting), similarly to the biclustering algorithms, the approach for gene selection is to select all genes deemed relevant for at least one condition. The identification of relevant genes, as mentioned in chapter 5, is done using a Mann-Whitney U test, selecting only genes for which $p < 0.01$.

As we can see in Figure 6.1, some of the conditions are accurately grouped together, such as the healthy Biopsy samples. However, there are issues, such as the healthy samples of NHBE cells belonging to *Series* 9 being closer to ones infected by IAV than they are to healthy cells belonging to other *Series*.

Despite these issues, by running the clustering algorithm on genes, with all conditions present in the dendrogram and then performing functional enrichment on the clusters, significant enriched terms are still present, presented in Table 6.2. These terms, when compared to the analysis performed on a cell by cell basis, have a very high c-score, which may be due to the presence of a higher number of genes, as well as a higher number of samples, and thus more information per gene, to support each cluster.

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, despite not within the top 25, are still statistically relevant. This receptor, along with others, has been identified as part of the inflammatory response to

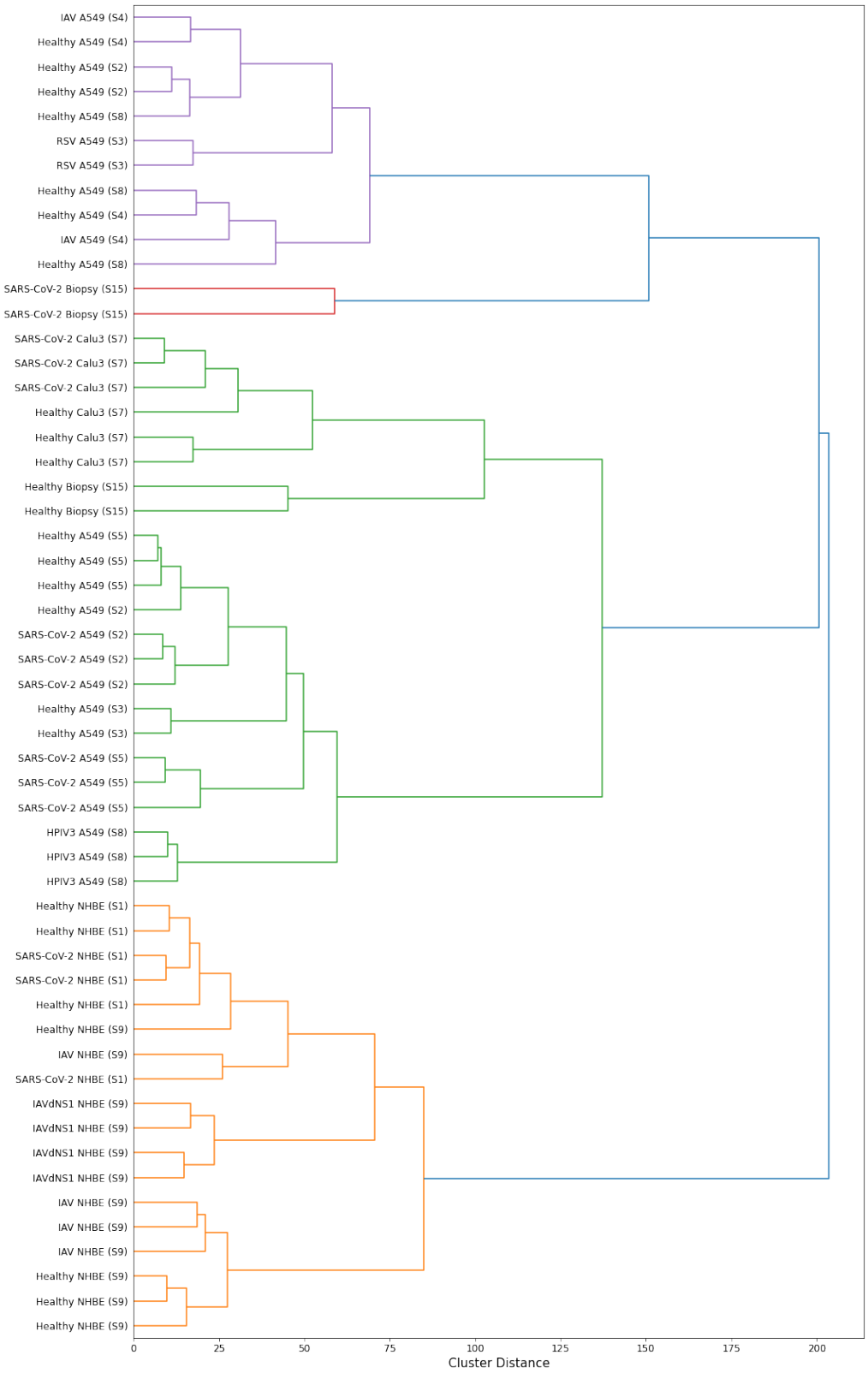

Figure 6.1: Dendrogram of various samples for each condition available in the dataset (Sn means the sample belongs to *Series* n).

Table 6.2: Top 25 GO biological processes ordered by combined score (Multi-Condition Setting, $p < 0.01$).

GO Biological Process	p-value	c-score	Cluster
type I interferon signaling pathway (GO:0060337)	4.40E-27	9111.11	2
cellular response to type I interferon (GO:0071357)	4.40E-27	9111.11	2
negative regulation of viral genome replication (GO:0045071)	1.10E-16	3820.31	2
defense response to symbiont (GO:0140546)	7.74E-22	3260.22	2
cytoplasmic receptor signaling pathway in response to virus [1] (GO:0039528)	1.74E-06	3253.53	2
negative regulation of viral process (GO:0048525)	5.16E-17	3163.10	2
defense response to virus (GO:0051607)	2.34E-21	2930.10	2
endogenous peptide antigen, TAP-independent (GO:0002486)	4.42E-05	2797.75	2
endogenous peptide antigen (GO:0002484)	4.42E-05	2797.75	2
regulation of viral genome replication (GO:0045069)	1.64E-15	2717.94	2
interferon-gamma-mediated signaling pathway (GO:0060333)	1.79E-15	2656.13	2
protein mono-ADP-ribosylation (GO:0140289)	3.35E-06	2318.71	2
exogenous peptide antigen, TAP-independent (GO:0002480)	6.69E-05	2159.41	2
cellular response to interferon-gamma (GO:0071346)	5.16E-17	2028.44	2
negative regulation of lipid localization (GO:1905953)	8.90E-05	1742.91	2
response to interferon-beta (GO:0035456)	6.88E-08	1678.71	2
regulation of ribonuclease activity (GO:0060700)	1.81E-03	1644.95	2
positive regulation of glial cell proliferation (GO:0060252)	1.81E-03	1644.95	2
interleukin-27-mediated signaling pathway (GO:0070106)	7.76E-06	1582.01	2
cytokine-mediated signaling pathway (GO:0019221)	6.11E-25	1457.27	2
protein poly-ADP-ribosylation (GO:0070212)	1.23E-04	1451.28	2
positive regulation of natural killer cell proliferation (GO:0032819)	5.80E-03	1445.70	7
regulation of mononuclear cell proliferation (GO:0032944)	5.80E-03	1445.70	7
regulation of NK T cell proliferation (GO:0051140)	5.80E-03	1445.70	7
positive regulation of mononuclear cell proliferation (GO:0032946)	5.80E-03	1445.70	7

SARS-CoV-2 as well as other coronaviruses [48]. Additionally, the signaling cascade resulting from the detection of viral proteins is associated with the production of Type I interferons and pro-inflammatory cytokines [49], which can also be observed within the processes identified in Table 6.2 (for instance, terms GO:0060337, GO:0071357 and GO:0060333, among others).

Since with clustering we cannot be sure of which cell types or conditions are responsible for a certain process, we now proceed, as explained in chapter 5, to analyse each cell type **individually**. For each cell type, two conditions are compared by performing Agglomerative Clustering with Euclidean affinity and Ward linkage, using a distance metric as well as pearson and spearman correlation. Each of these metrics proved to be very similar after functional enrichment, so the clusters correspond only to the distance metric. Additionally, the number of clusters is 3, which also does not seem to affect the results of functional enrichment when changed to reasonable values (between 2 and 10), which is mainly due to the presence, in all cases, of a dominant cluster which contains most of the enriched processes, with usually only one other cluster with some, significantly less, processes. Each cluster of genes, was then functionally enriched using the EnrichR API [43, 44]. These tables contain the name of the GO Biological Process, the adjusted p-value and the combined score. We first removed all results with $p >= 0.01$ and

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

Table 6.3: Top 25 GO biological processes ordered by combined score, for NHBE cells.

GO Biological Process	p-value	c-score	Cluster
regulation of calcidiol 1-monooxygenase activity (GO:0060558)	1.83E-03	1610.66	1
pantothenate metabolic process (GO:0015939)	1.83E-03	1610.66	1
cellular response to type I interferon (GO:0071357)	1.95E-12	1330.17	2
type I interferon signaling pathway (GO:0060337)	1.95E-12	1330.17	2
postsynaptic neurotransmitter receptor internalization (GO:0098884)	9.96E-03	865.07	2
postsynaptic endocytosis (GO:0140239)	9.96E-03	865.07	2
regulation of ribonuclease activity (GO:0060700)	9.96E-03	865.07	2
response to interferon-beta (GO:0035456)	2.28E-06	830.41	2
defense response to symbiont (GO:0140546)	9.02E-12	717.21	2
defense response to virus (GO:0051607)	1.94E-11	641.59	2
response to interferon-alpha (GO:0035455)	2.93E-04	593.97	2
negative regulation of viral genome replication (GO:0045071)	4.90E-06	434.32	2
regulation of lipid storage (GO:0010883)	5.88E-04	430.27	2
antiviral innate immune response (GO:0140374)	3.39E-03	426.70	2
cytokine-mediated signaling pathway (GO:0019221)	9.05E-15	395.62	2
interleukin-27-mediated signaling pathway (GO:0070106)	3.68E-03	382.07	2
negative regulation of type I interferon-mediated signaling pathway (GO:0060339)	4.10E-03	344.91	2
negative regulation of chemokine production (GO:0032682)	4.81E-03	313.53	2
regulation of viral genome replication (GO:0045069)	2.03E-05	309.60	2
negative regulation of viral process (GO:0048525)	2.50E-05	289.01	2
regulation of complement activation (GO:0030449)	1.88E-03	233.87	1
regulation of lipid storage (GO:0010883)	9.15E-03	233.65	1
inflammatory response (GO:0006954)	3.32E-07	215.82	2
positive regulation of NIK/NF-kappaB signaling (GO:1901224)	1.88E-03	213.72	1
cellular response to virus (GO:0098586)	2.89E-03	208.65	2

then ordered the obtained terms by combined score, taking the top 25.

In Table 6.3, we present the identified processes when comparing **healthy and infected NHBE cells**. The genes composing all detected enriched terms have higher expression levels for infected cells when compared to control. The top two processes do not seem to be related to viral infection, however, after those, there seems to be a prevalence of type I interferon and generic cytokine related processes, which are associated with immune response.

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 [50]. The pres-

Table 6.4: Top 25 GO biological processes ordered by combined score, for A549 cells.

GO Biological Process	p-value	c-score	Cluster
cellular response to type I interferon (GO:0071357)	2.23E-15	1540.94	2
type I interferon signaling pathway (GO:0060337)	2.23E-15	1540.94	2
negative regulation of viral genome replication (GO:0045071)	3.07E-09	792.37	2
response to interferon-beta (GO:0035456)	4.98E-05	637.94	2
negative regulation of viral life cycle (GO:1903901)	1.99E-08	569.74	2
regulation of viral genome replication (GO:0045069)	2.32E-08	540.29	2
regulation of interferon-alpha production (GO:0032647)	6.90E-04	472.38	2
positive regulation of interferon-alpha production (GO:0032727)	1.26E-03	354.12	2
interferon-gamma-mediated signaling pathway (GO:0060333)	1.05E-06	344.39	2
cytokine-mediated signaling pathway (GO:0019221)	2.23E-15	327.50	2
cellular response to interferon-gamma (GO:0071346)	4.59E-08	321.58	2
positive regulation of defense response to virus by host (GO:0002230)	1.81E-03	300.45	2
STAT cascade (GO:0097696)	9.26E-04	211.79	0
chemokine-mediated signaling pathway (GO:0070098)	4.73E-04	195.27	2
response to interferon-gamma (GO:0034341)	1.78E-04	187.71	2
regulation of leukocyte chemotaxis (GO:0002688)	5.77E-03	169.19	2
regulation of defense response to virus by host (GO:0050691)	5.77E-03	169.19	2
negative regulation of type I interferon production (GO:0032480)	2.18E-03	160.73	2
response to cytokine (GO:0034097)	3.49E-05	147.64	2
positive regulation of JAK-STAT cascade (GO:0046427)	1.26E-03	139.37	2
regulation of type I interferon production (GO:0032479)	6.82E-04	130.07	2
neutrophil migration (GO:1990266)	6.25E-03	102.79	2
JAK-STAT cascade (GO:0007259)	4.73E-03	95.44	0
positive regulation of leukocyte chemotaxis (GO:0002690)	7.10E-03	94.66	2
positive regulation of type I interferon production (GO:0032481)	7.40E-03	92.17	2

ence 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 Table 6.4. 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-

Table 6.5: Top 25 GO biological processes ordered by combined score, for A549-ACE2 cells.

GO Biological Process	p-value	c-score	Cluster
positive regulation of heat generation (GO:0031652)	8.32E-03	3921.08	0
regulation of fever generation (GO:0031620)	8.32E-03	3921.08	0
positive regulation of fever generation (GO:0031622)	8.32E-03	2825.38	0
regulation of vascular wound healing (GO:0061043)	8.32E-03	1765.21	0
positive regulation of steroid biosynthetic process (GO:0010893)	8.32E-03	1765.21	0
aerobic electron transport chain (GO:0019646)	3.06E-20	833.55	2
mitochondrial ATP synthesis coupled electron transport (GO:0042775)	3.06E-20	801.88	2
mitochondrial electron transport, NADH to ubiquinone (GO:0006120)	2.72E-13	692.38	2
L-phenylalanine catabolic process (GO:0006559)	4.28E-03	637.69	2
amino acid catabolic process (GO:1902222)	4.28E-03	637.69	2
ribose phosphate metabolic process (GO:0019693)	4.28E-03	637.69	2
quinone catabolic process (GO:1901662)	4.28E-03	637.69	2
cellular glucuronidation (GO:0052695)	9.25E-03	426.04	1
acyl-CoA biosynthetic process (GO:0071616)	2.56E-05	292.54	2
acetyl-CoA biosynthetic process (GO:0006085)	7.63E-04	291.06	2
NADH dehydrogenase complex assembly (GO:0010257)	3.07E-10	286.82	2
mitochondrial respiratory chain complex I assembly (GO:0032981)	3.07E-10	286.82	2
mitochondrial respiratory chain complex assembly (GO:0033108)	4.09E-10	198.40	2
L-phenylalanine metabolic process (GO:0006558)	5.27E-03	194.75	2
secondary alcohol biosynthetic process (GO:1902653)	8.13E-06	177.46	2
mitochondrial electron transport (GO:0006122)	2.48E-03	171.57	2
cholesterol biosynthetic process (GO:0006695)	1.04E-05	165.42	2
heme biosynthetic process (GO:0006783)	2.74E-04	148.92	2
fatty-acyl-CoA metabolic process (GO:0035337)	2.74E-04	148.92	2
sterol biosynthetic process (GO:0016126)	2.56E-05	135.90	2

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 [51], 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 response. 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 6.5, we present the identified processes **for A549-ACE2 cells**. The genes composing the top five detected enriched terms have higher expression levels for infected cells when compared to control, and all others have higher levels for control. As mentioned in section 4.1, this culture consists of A549 cells with added ACE2, since viral counts were low in A549 cells. This addition seems to have significantly impacted the amount of unrelated processes found. However the top terms, *positive regulation of heat generation* (GO:0031652), *regulation of fever generation* (GO:0031620) and *positive regulation of fever generation* (GO:0031622) are all associated with acute inflammatory response (the term GO:0002526, which is an ancestor), as well as one of the most common symptoms associated

Table 6.6: Top 25 GO biological processes ordered by combined score, for Calu3 cells.

GO Biological Process	p-value	c-score	Cluster
secondary alcohol biosynthetic process (GO:1902653)	7.95E-16	1990.21	0
regulation of ribonuclease activity (GO:0060700)	7.19E-05	1921.48	2
negative regulation of viral process (GO:0048525)	2.57E-26	1907.45	2
negative regulation of viral genome replication (GO:0045071)	8.48E-23	1896.45	2
cholesterol biosynthetic process (GO:0006695)	7.95E-16	1858.91	0
sterol biosynthetic process (GO:0016126)	2.75E-15	1541.71	0
defense response to symbiont (GO:0140546)	9.30E-31	1498.91	2
type I interferon signaling pathway (GO:0060337)	7.48E-22	1405.26	2
cellular response to type I interferon (GO:0071357)	7.48E-22	1405.26	2
defense response to virus (GO:0051607)	9.30E-31	1396.78	2
regulation of viral genome replication (GO:0045069)	9.17E-19	1016.38	2
regulation of nuclease activity (GO:0032069)	1.78E-04	880.78	2
positive regulation of extrinsic apoptotic signaling pathway (GO:1902043)	1.78E-04	880.78	2
cytokine-mediated signaling pathway (GO:0019221)	9.36E-44	869.33	2
isopentenyl diphosphate biosynthetic process (GO:0009240)	6.97E-03	735.60	0
negative regulation of lymphocyte differentiation (GO:0045620)	3.64E-04	546.31	2
positive regulation of gliogenesis (GO:0014015)	3.64E-04	546.31	2
positive regulation of smooth muscle cell differentiation (GO:1905065)	3.64E-04	546.31	2
negative regulation of innate immune response (GO:0045824)	9.07E-10	500.91	2
cellular response to interferon-gamma (GO:0071346)	2.10E-16	485.61	2
cellular response to cytokine stimulus (GO:0071345)	3.66E-28	483.06	2
positive regulation of heat generation (GO:0031652)	2.52E-03	479.71	2
exocyst localization (GO:0051601)	2.52E-03	479.71	2
regulation of fever generation (GO:0031620)	2.52E-03	479.71	2
positive regulation of glial cell proliferation (GO:0060252)	2.52E-03	479.71	2

with Covid-19.

The top 25 processes associated with **Calu3 cells** are shown in Table 6.6. The genes composing the terms *secondary alcohol biosynthetic process* (GO:1902653), *cholesterol biosynthetic process* (GO:0006695), *sterol biosynthetic process* (GO:0016126), *isopentenyl diphosphate biosynthetic process* (GO:0009240) are downregulated in infected cells, whereas the rest are upregulated. For these cells, similar processes to the NHBE and A549 cells were found, so predominantly interferon and cytokine related ones. It is also interesting to note the presence of the term *regulation of fever generation* (GO:0031620), which was also present for A549-ACE2 cells.

The term *regulation of ribonuclease activity* (GO:0060700), which was identified for NHBE cells (Table 6.3) as well, is also noteworthy, since ribonuclease (RNase) is an enzyme which catalyzes the breakdown of RNA into smaller components. Particularly RNase L is associated with innate immune response, and certain viruses have been shown to block this pathway in order to prevent viral RNA degradation [52].

An individual analysis for biopsy samples was not performed due to there being only 2 healthy and 2 infected samples.

In Table 6.7 we present the pathways identified when using the **KEGG Pathway database** (2021

Table 6.7: Top 25 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
Legionellosis (map05134)	1.77E-05	118.99	0
Hepatitis C (map05160)	1.79E-05	118.31	2
TNF signaling pathway (map04668)	8.61E-05	116.25	2
Primary immunodeficiency (map05340)	2.17E-03	116.00	2
Allograft rejection (map05330)	2.17E-03	116.00	2
Amoebiasis (map05146)	2.13E-06	111.38	0
African trypanosomiasis (map05143)	1.50E-04	111.14	0
Rheumatoid arthritis (map05323)	3.12E-06	110.93	0
Legionellosis (map05134)	1.12E-03	105.75	2
Antigen processing and presentation (map04612)	5.58E-04	100.66	2
JAK-STAT signaling pathway (map04630)	1.07E-06	96.92	0
NF-kappa B signaling pathway (map04064)	3.68E-04	92.76	2
Kaposi sarcoma-associated herpesvirus infection (map05167)	8.97E-05	82.21	2
Insulin resistance (map04931)	1.68E-05	81.67	0
AGE-RAGE signaling pathway in diabetic complications (map04933)	2.87E-05	77.90	0
IL-17 signaling pathway (map04657)	1.29E-03	73.62	2
Type II diabetes mellitus (map04930)	5.14E-04	72.61	0

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.

Using the same method, we are not able to find relevant terms when comparing SARS-CoV-2 with

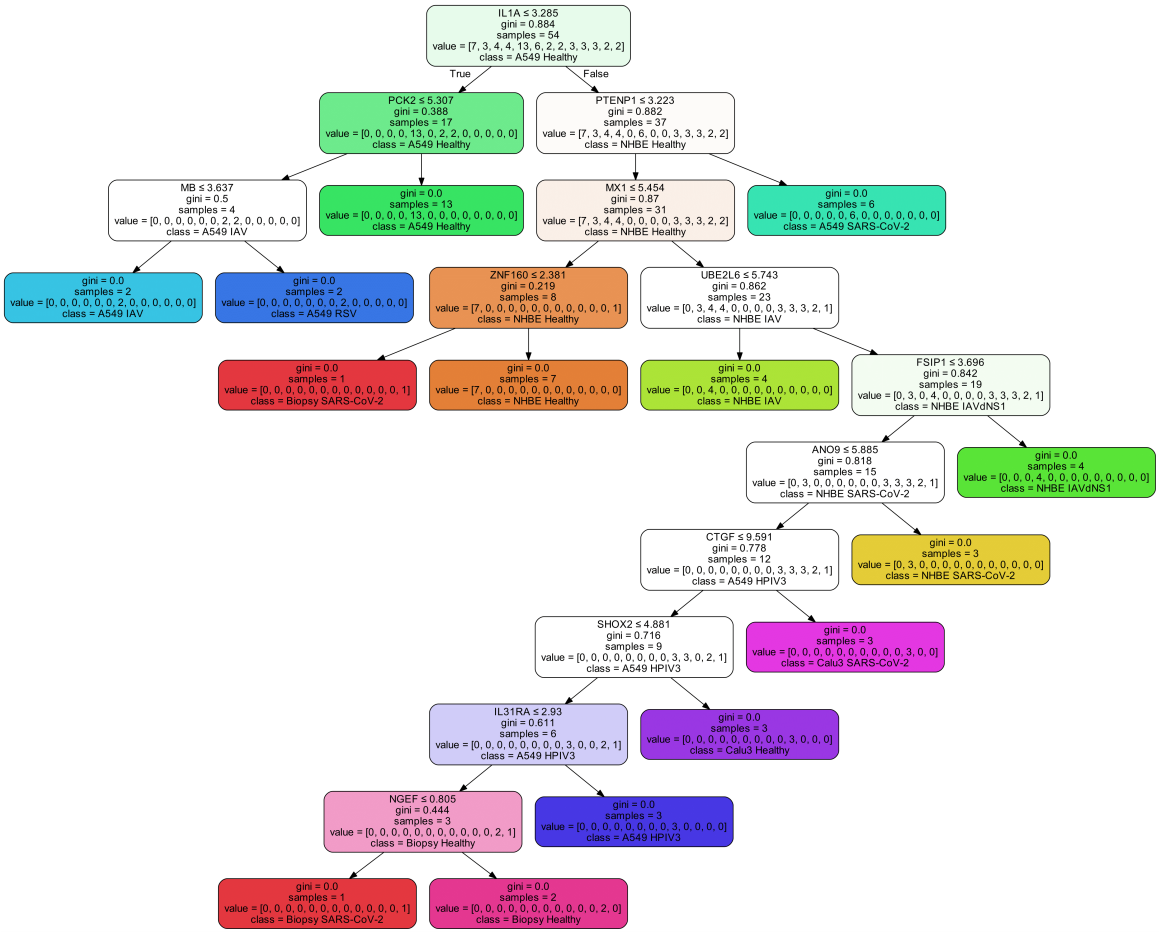

Figure 6.2: Decision Tree with gini criterion (Multi-Condition Setting, $p < 0.01$). The colors represent each class, with a node's color corresponding to the combination of the colors of all classes belonging to it.

other viruses. This may be due to these viruses having subtler differences, which are not captured by the clustering algorithm.

6.2 Predictive Modeling

We present in Figure 6.2 a decision tree produced using gini criterion, in the Multi-Condition Setting. Similarly to the method described for Table 6.2, the approach for gene selection is to select all genes deemed relevant for at least one condition, with the identification of relevant genes performed using a Mann-Whitney U test, selecting only genes for which $p < 0.01$. In this decision tree, we can see that a few genes are sufficient to discriminate between classes. While in a classification problem this can be desirable, for the problem of discovering regulatory modules in transcriptomic data it is not, since in the latter case regulatory modules are usually composed of multiple interacting genes. However, the

presence of certain genes in the tree related to immune and inflammatory responses is useful to further understanding of which genes are associated with the response to viral infection. Namely, *IL1A*, Interleukin 1 Alpha, a protein-encoding gene associated with cytokine activity and inflammatory response; *MX1*, which is a protein-encoding gene associated with antiviral activity against a variety of RNA viruses; *IL31RA*, a type I cytokine receptor. Despite these genes being potentially associated, we can see multiple other genes which don't seem to have a direct connection to this process, which is to be expected, since with the amount of genes available and the low number of samples the decision tree can focus on a single gene which may be altered merely due to chance, and not due to any of the conditions being investigated. This issue is even more clear when applying decision trees to single cell type analysis, like the one performed with clustering earlier in the present chapter, where the tree tends to be able to discriminate between healthy and infected cells using a single gene, which is not sufficient to discover significant processes.

Due to the aforementioned issues, we proceed to ensemble algorithms, which are less prone to overfitting. The genes obtained using the feature importances of the Random Forests and XGBoost models are functionally enriched using the Enrichr API, as explained for clustering. It is important to note that using these algorithms a significantly lower amount of genes is selected, which leads to higher p-values in the obtained terms. Due to this, all terms use $p_{\text{adjusted}} < 0.05$, as opposed to $p_{\text{adjusted}} < 0.01$ which is used for clustering and biclustering.

As with clustering, we first begin by presenting the processes identified when using the complete data, with multiple cell types. The process to obtain this data is explained in chapter 5.

In Table 6.8, we have the GO Biological processes identified when selecting genes using a Random Forest algorithm, and in Table 6.9 we have the enriched terms found when using XGBoost. 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 [53].

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

Table 6.8: Top 25 GO biological processes ordered by combined score (Random Forest, Multi-Condition Setting, $p < 0.01$).

GO Biological Processes	p-value	c-score
protein mono-ADP-ribosylation (GO:0140289)	3.44E-06	980.50
type I interferon signaling pathway (GO:0060337)	3.70E-14	833.99
cellular response to type I interferon (GO:0071357)	3.70E-14	833.99
regulation of fever generation (GO:0031620)	2.02E-03	819.46
positive regulation of glial cell proliferation (GO:0060252)	2.02E-03	819.46
cytokine-mediated signaling pathway (GO:0019221)	7.67E-24	420.28
interferon-gamma-mediated signaling pathway (GO:0060333)	3.65E-09	372.42
negative regulation of viral genome replication (GO:0045071)	3.35E-08	368.56
antigen processing via MHC class I via ER pathway (GO:0002484)	5.02E-03	358.52
antigen processing via MHC class I via ER pathway, TAP-independent (GO:0002486)	5.02E-03	358.52
positive regulation of gliogenesis (GO:0014015)	5.02E-03	358.52
positive regulation of podosome assembly (GO:0071803)	5.02E-03	358.52
cellular response to interferon-gamma (GO:0071346)	1.87E-11	354.96
negative regulation of viral process (GO:0048525)	4.92E-09	353.10
interleukin-27-mediated signaling pathway (GO:0070106)	3.24E-04	344.30
defense response to symbiont (GO:0140546)	2.53E-11	339.26
receptor signaling pathway via STAT (GO:0097696)	2.01E-05	337.92
positive regulation of epidermal growth factor-activated receptor activity (GO:0045741)	1.26E-03	332.52
receptor signaling pathway via JAK-STAT (GO:0007259)	6.49E-06	329.14
defense response to virus (GO:0051607)	8.56E-11	297.98
protein auto-ADP-ribosylation (GO:0070213)	1.74E-03	280.02
response to interferon-beta (GO:0035456)	4.34E-05	273.57
interleukin-21-mediated signaling pathway (GO:0038114)	6.94E-03	271.56
exogenous peptide antigen via MHC class I, TAP-independent (GO:0002480)	6.94E-03	271.56
cellular response to interleukin-21 (GO:0098757)	6.94E-03	271.56

various types of immune cells. In particular, macrophages [54, 55] (GO:0048246 and GO:1905517), natural killer cells [56] (GO:2000501)), eosinophils [57] (GO:0072677 and GO:0048245), neutrophils [58] (GO:0030593 and GO:1990266) are all types of white blood cells involved with the innate immune response to viral infection.

Additionally, there are multiple terms in both tables associated with cytokine production and related signaling pathways, as well as response to diferent 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.

Similarly to clustering, there is no way to verify for which types of cells a particular set of genes is significant. Thus, we now proceed to a cell type specific analysis.

In Table 6.10 and Table 6.11 we can see the terms identified **for NHBE cells** using a Random Forest and XGBoost, respectively. As was previously mentioned, the p-value for these processes is significantly

Table 6.9: Top 25 GO biological processes ordered by combined score (XGBoost, Multi-Condition Setting, $p < 0.01$).

GO Biological Processes	p-value	c-score
ISG15-protein conjugation (GO:0032020)	7.61E-03	869.12
macrophage chemotaxis (GO:0048246)	1.20E-03	783.48
response to interferon-gamma (GO:0034341)	1.20E-07	672.59
nicotinamide nucleotide biosynthetic process (GO:0019359)	9.89E-03	666.41
regulation of natural killer cell chemotaxis (GO:2000501)	9.89E-03	666.41
macrophage migration (GO:1905517)	2.00E-03	548.36
eosinophil migration (GO:0072677)	2.01E-03	495.85
eosinophil chemotaxis (GO:0048245)	2.01E-03	495.85
lymphocyte migration (GO:0072676)	8.21E-05	435.11
neutrophil chemotaxis (GO:0030593)	8.53E-06	432.97
chemokine-mediated signaling pathway (GO:0070098)	2.76E-05	412.00
granulocyte chemotaxis (GO:0071621)	9.17E-06	406.08
lymphocyte chemotaxis (GO:0048247)	1.24E-04	376.49
neutrophil migration (GO:1990266)	1.11E-05	374.27
cellular response to chemokine (GO:1990869)	3.60E-05	370.93
cellular response to interferon-gamma (GO:0071346)	1.66E-06	355.72
type I interferon signaling pathway (GO:0060337)	4.53E-05	328.36
cellular response to type I interferon (GO:0071357)	4.53E-05	328.36
NAD biosynthetic process (GO:0009435)	3.89E-03	327.14
monocyte chemotaxis (GO:0002548)	2.00E-03	232.75
response to interferon-beta (GO:0035456)	7.50E-03	212.45
inflammatory response (GO:0006954)	1.66E-05	168.96
cytokine-mediated signaling pathway (GO:0019221)	2.53E-07	161.46
negative regulation of viral genome replication (GO:0045071)	3.87E-03	159.21
response to tumor necrosis factor (GO:0034612)	7.93E-04	145.80

higher when compared to those obtained using clustering on the same data. It is worth noting however, that the c-score, while comparable for terms identified using a Random Forest, is significantly higher for terms identified using XGBoost. This is due to the previously mentioned fact that the XGBoost algorithm selects much fewer genes when compared to a Random Forest.

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 [59].

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*

Table 6.10: Top statistically relevant GO biological processes ordered by combined score, for NHBE cells (Random Forest).

GO Biological Processes	p-value	c-score
positive regulation of monocyte chemotactic protein-1 production (GO:0071639)	4.06E-03	718.81
chronic inflammatory response (GO:0002544)	2.24E-02	558.06
positive regulation of glial cell proliferation (GO:0060252)	2.24E-02	558.06
positive regulation of heat generation (GO:0031652)	2.24E-02	558.06
response to salt stress (GO:0009651)	2.24E-02	558.06
regulation of fever generation (GO:0031620)	2.24E-02	558.06
regulation of monocyte chemotactic protein-1 production (GO:0071637)	8.10E-03	402.78
positive regulation of fever generation (GO:0031622)	2.70E-02	395.36
ISG15-protein conjugation (GO:0032020)	2.70E-02	395.36
positive regulation of histone phosphorylation (GO:0033129)	2.70E-02	395.36
toll-like receptor 4 signaling pathway (GO:0034142)	3.02E-03	312.35
positive regulation of gliogenesis (GO:0014015)	3.17E-02	300.93
regulation of calcidiol 1-monooxygenase activity (GO:0060558)	3.17E-02	300.93
negative regulation of MyD88-independent toll-like receptor signaling pathway (GO:0034128)	3.17E-02	300.93
intermediate filament bundle assembly (GO:0045110)	3.17E-02	300.93
positive regulation of granulocyte macrophage colony-stimulating factor production (GO:0032725)	1.33E-02	268.34
interleukin-21-mediated signaling pathway (GO:0038114)	3.55E-02	239.87
cellular response to interleukin-21 (GO:0098757)	3.55E-02	239.87
vascular associated smooth muscle cell differentiation (GO:0035886)	3.55E-02	239.87
regulation of MyD88-independent toll-like receptor signaling pathway (GO:0034127)	3.55E-02	239.87
positive regulation of osteoclast differentiation (GO:0045672)	1.57E-02	239.65
positive regulation of alpha-beta T cell proliferation (GO:0046641)	1.60E-02	215.79
regulation of granulocyte macrophage colony-stimulating factor production (GO:0032645)	1.60E-02	215.79
regulation of gap junction assembly (GO:1903596)	4.07E-02	197.46
cellular response to interleukin-9 (GO:0071355)	4.07E-02	197.46

Table 6.11: Top statistically relevant GO biological processes ordered by combined score, for NHBE cells (XGBoost).

GO Biological Processes	p-value	c-score
regulation of integrin biosynthetic process (GO:0045113)	1.32E-02	8352.53
chronic inflammatory response (GO:0002544)	1.32E-02	8352.53
peptidyl-cysteine S-nitrosylation (GO:0018119)	1.32E-02	8352.53
astrocyte development (GO:0014002)	1.32E-02	6499.56
regulation of macromolecule biosynthetic process (GO:0010556)	1.32E-02	6499.56
astrocyte differentiation (GO:0048708)	1.32E-02	5287.72
leukocyte aggregation (GO:0070486)	1.32E-02	4436.86
peptidyl-cysteine modification (GO:0018198)	1.32E-02	3808.55
defense response to fungus (GO:0050832)	3.05E-02	1111.12
glial cell development (GO:0021782)	3.05E-02	1006.18
positive regulation of intrinsic apoptotic signaling pathway (GO:2001244)	3.92E-02	589.61
regulation of intrinsic apoptotic signaling pathway (GO:2001242)	3.92E-02	425.07
inorganic anion transport (GO:0015698)	3.92E-02	405.46
positive regulation of apoptotic signaling pathway (GO:2001235)	3.92E-02	362.85
pattern recognition receptor signaling pathway (GO:0002221)	3.92E-02	347.96
antimicrobial humoral immune response (GO:0061844)	3.92E-02	327.57
neutrophil chemotaxis (GO:0030593)	3.92E-02	292.57
response to molecule of bacterial origin (GO:0002237)	3.92E-02	277.46
granulocyte chemotaxis (GO:0071621)	3.92E-02	277.46
chloride transport (GO:0006821)	3.92E-02	263.66
positive regulation of growth (GO:0045927)	3.92E-02	263.66
neutrophil migration (GO:1990266)	3.92E-02	259.33
regulation of organelle organization (GO:0033043)	3.92E-02	247.05
activation of endopeptidase activity involved in apoptotic process (GO:0006919)	3.92E-02	243.19
positive regulation of neuron projection development (GO:0010976)	3.92E-02	218.84

signaling pathway (GO:0002221). These receptors, as previously explained for the related term present in Table 6.2, have been associated with the inflammatory response to SARS-CoV-2 [48].

In Table 6.12 and Table 6.13 we present the biological processes identified for A549 cells, using a Random Forest and XGBoost respectively.

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*

Table 6.12: Top statistically relevant GO biological processes ordered by combined score, for A549 cells (Random Forest).

GO Biological Processes	p-value	c-score
RIG-I signaling pathway (GO:0039529)	2.01E-02	819.47
positive regulation of dendritic cell cytokine production (GO:0002732)	2.08E-02	658.60
cytoplasmic pattern recognition receptor signaling pathway in response to virus (GO:0039528)	3.15E-02	463.95
positive regulation of epidermal growth factor-activated receptor activity (GO:0045741)	3.65E-02	401.14
positive regulation of vascular endothelial growth factor production (GO:0010575)	1.52E-02	314.50
regulation of vascular endothelial growth factor production (GO:0010574)	1.52E-02	280.69
positive regulation of nuclear division (GO:0051785)	1.52E-02	280.69
positive regulation of defense response to virus by host (GO:0002230)	1.52E-02	266.02
response to interferon-beta (GO:0035456)	1.52E-02	266.02
regulation of interleukin-2 production (GO:0032663)	1.52E-02	240.53
regulation of defense response to virus by host (GO:0050691)	2.02E-02	183.70
positive regulation of mitotic nuclear division (GO:0045840)	2.02E-02	183.70
positive regulation of interleukin-6 production (GO:0032755)	1.75E-02	120.65
regulation of protein localization to plasma membrane (GO:1903076)	1.97E-02	111.57
defense response to symbiont (GO:0140546)	1.52E-02	98.43
defense response to virus (GO:0051607)	1.52E-02	88.09
negative regulation of cytokine production (GO:0001818)	1.52E-02	84.74
regulation of interleukin-6 production (GO:0032675)	4.07E-02	68.02
cellular response to cytokine stimulus (GO:0071345)	1.52E-02	63.15
positive regulation of cytokine production (GO:0001819)	2.02E-02	45.00
cytokine-mediated signaling pathway (GO:0019221)	1.52E-02	38.92

Table 6.13: Top statistically relevant GO biological processes ordered by combined score, for A549 cells (XGBoost).

GO Biological Processes	p-value	c-score
negative regulation of substrate adhesion-dependent cell spreading (GO:1900025)	1.49E-02	3304.67
negative regulation of cell morphogenesis involved in differentiation (GO:0010771)	1.49E-02	3304.67
protein localization to vacuole (GO:0072665)	1.49E-02	3012.37
regulation of lymphocyte activation (GO:0051249)	1.49E-02	2764.28
negative regulation of T cell receptor signaling pathway (GO:0050860)	1.49E-02	2062.22
regulation of protein localization to cell periphery (GO:1904375)	1.49E-02	1935.63
negative regulation of protein localization to plasma membrane (GO:1903077)	1.49E-02	1822.54
negative regulation of protein localization to cell periphery (GO:1904376)	1.49E-02	1822.54
negative regulation of interleukin-2 production (GO:0032703)	1.49E-02	1720.94
negative regulation of antigen receptor-mediated signaling pathway (GO:0050858)	1.49E-02	1470.20
negative regulation of protein localization to membrane (GO:1905476)	1.49E-02	1400.90
regulation of calcium-mediated signaling (GO:0050848)	1.49E-02	1278.76
regulation of B cell activation (GO:0050864)	1.49E-02	1278.76
regulation of protein localization to membrane (GO:1905475)	1.50E-02	1174.65
regulation of T cell receptor signaling pathway (GO:0050856)	1.60E-02	971.56
regulation of sodium ion transport (GO:0002028)	1.60E-02	938.42
negative regulation of cell-substrate adhesion (GO:0010812)	1.68E-02	824.08
cellular response to tumor necrosis factor (GO:0071356)	1.49E-02	773.33
regulation of interleukin-2 production (GO:0032663)	1.85E-02	657.83
negative regulation of ERK1 and ERK2 cascade (GO:0070373)	1.85E-02	625.39
regulation of substrate adhesion-dependent cell spreading (GO:1900024)	1.85E-02	610.23
interferon-gamma-mediated signaling pathway (GO:0060333)	2.35E-02	426.61
regulation of ion transport (GO:0043269)	2.45E-02	353.55
response to interferon-gamma (GO:0034341)	2.45E-02	348.01
regulation of protein localization to plasma membrane (GO:1903076)	2.45E-02	348.01

to virus (GO:0039528) were also identified in Table 6.2 as well as in Table 6.10. 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 [49].

6.3 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 6.14 by presenting several metrics for each algorithm and preprocessing option used. 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.

To better understand how different conditions compare when it comes to associated biological processes, the Spearman correlation coefficient for each pair of columns is presented as a heatmap, both for the number of occurrences of each enriched term (Figure 6.3) and the best c-score among those occurrences (Figure 6.4). It is interesting to note the high correlation between NHBE IAV and NHBE SARS-CoV-2 (0.86 for the number of occurrences and 0.67 for the c-scores), particularly due to the somewhat lower correlation coefficient (0.65) between the NHBE and A549 SARS-CoV-2 conditions. This may be partly explained by NHBE and A549 being different cell types and thus possessing some differences in baseline processes which lead to the lower correlation. However, the lower value of the

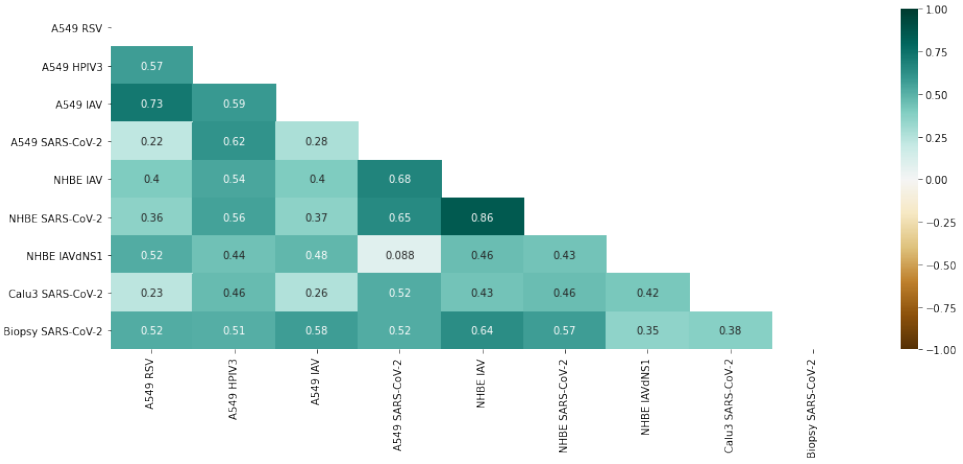

Figure 6.3: Comparison between the processes identified for each condition, using Spearman correlation between the number of occurrences of each GO biological process.

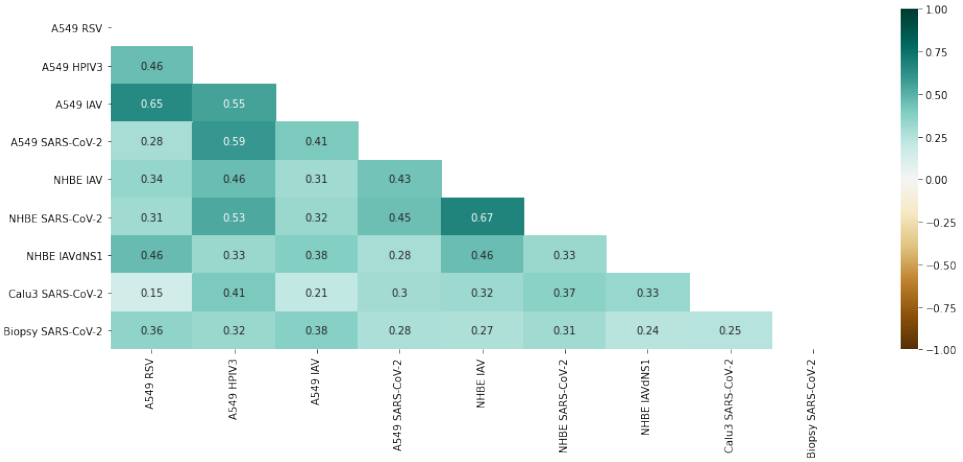

Figure 6.4: Comparison between the processes identified for each condition, using Spearman correlation between the c-score of each GO biological process.

Table 6.14: Metrics for comparing the performance of the tested biclustering algorithms with different preprocessing techniques. $|\mathcal{B}|$ - Number of biclusters; $\overline{|I|}$ - Average number of genes per bicluster; $\sigma_{|I|}$ - Standard deviation of genes per bicluster; $\overline{|J|}$ - Average number of conditions per bicluster; $\sigma_{|J|}$ - Standard deviation of the number of conditions per bicluster; $\overline{\text{Terms}}$ - Average number of enriched terms per bicluster.

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

coefficient between the SARS-CoV-2 and IAVdNS1 (both for NHBE cells) suggests there may also be some similarities between these two viruses.

We now proceed to a comparative analysis of the biological processes associated with SARS-CoV-2 for all cell types, which is presented in Table 6.15. 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

Table 6.15: GO Biological processes with highest joint ranks for SARS-CoV-2 conditions. Counts correspond to the normalized number of occurrences of each process within each condition.

GO Biological Processes	A549 SARS-CoV-2			NHBE SARS-CoV-2		
	count	p-value	c-score	count	p-value	c-score
cytokine-mediated signaling pathway (GO:0019221)	1.00	6.95E-03	6.06E+05	0.75	2.79E-04	6.06E+05
cellular response to interferon-gamma (GO:0071346)	0.66	1.15E-07	5.61E+02	0.88	4.31E-03	3.97E+02
cellular response to cytokine stimulus (GO:0071345)	0.61	2.50E-04	6.55E+05	0.92	2.50E-04	6.55E+05
inflammatory response (GO:0006954)	0.45	5.30E-05	6.46E+02	0.75	3.90E-04	1.47E+02
protein modification by small protein removal (GO:0070646)	0.49	7.88E-03	1.74E+02	0.46	7.88E-03	1.68E+02
regulation of immune response (GO:0050776)	0.34	3.64E-07	7.11E+02	0.71	5.14E-03	1.85E+02
mRNA splicing, via spliceosome (GO:0000398)	0.51	2.08E-06	7.38E+02	0.62	1.69E-04	6.89E+02
mRNA processing (GO:0006397)	0.51	2.88E-06	7.02E+02	0.62	3.98E-04	6.99E+02
DNA metabolic process (GO:0006259)	0.45	5.44E-04	2.63E+02	0.38	7.56E-04	4.22E+01
interferon-gamma-mediated signaling pathway (GO:0060333)	0.54	1.03E-03	6.97E+02	0.38	9.63E-04	6.97E+02
epidermis development (GO:0008544)	0.54	1.83E-04	1.44E+03	0.88	2.21E-22	2.15E+03
RNA splicing, with bulged adenosine as nucleophile (GO:0000377)	0.48	3.31E-06	7.02E+02	0.62	4.30E-04	6.75E+02
positive regulation of response to external stimulus (GO:0032103)	0.39	3.65E-03	5.26E+02	0.67	1.82E-03	1.36E+02
chemokine-mediated signaling pathway (GO:0070098)	0.39	6.02E-05	3.99E+03	0.67	1.55E-04	5.04E+02
DNA repair (GO:0006281)	0.45	4.48E-03	3.13E+02	0.25	4.65E-03	3.81E+01
extracellular matrix organization (GO:0030198)	0.32	1.13E-03	1.05E+02	1.00	3.31E-08	1.70E+02
neutrophil mediated immunity (GO:0002446)	0.53	3.68E-03	2.39E+02	0.38	2.14E-24	1.93E+02
cellular response to DNA damage stimulus (GO:0006974)	0.47	2.65E-03	1.29E+02	0.29	4.65E-03	1.10E+02
neutrophil activation involved in immune response (GO:0002283)	0.49	6.01E-21	2.39E+02	0.50	9.44E-03	1.93E+02
neutrophil degranulation (GO:0043312)	0.49	2.12E-21	2.46E+02	0.50	8.99E-03	2.00E+02
cellular response to chemokine (GO:1990869)	0.36	8.06E-05	3.65E+03	0.62	1.94E-04	4.57E+02
protein ubiquitination (GO:0016567)	0.45	9.45E-03	1.70E+02	0.29	9.45E-03	1.70E+02
cellular protein modification process (GO:0006464)	0.49	1.52E-19	1.39E+02	0.29	1.93E-03	1.20E+02
antigen receptor-mediated signaling pathway (GO:0050851)	0.49	6.48E-06	8.87E+01	0.67	3.95E-03	8.87E+01
defense response to symbiont (GO:0140546)	0.39	7.76E-06	9.11E+05	0.38	7.76E-06	9.11E+05

GO Biological Processes	Calu3 SARS-CoV-2			Biopsy SARS-CoV-2		
	count	p-value	c-score	count	p-value	c-score
cytokine-mediated signaling pathway (GO:0019221)	1.00	2.29E-03	1.02E+03	0.60	2.79E-04	6.06E+05
cellular response to interferon-gamma (GO:0071346)	0.83	2.36E-04	1.13E+03	0.56	1.37E-18	1.56E+03
cellular response to cytokine stimulus (GO:0071345)	0.67	4.14E-04	3.67E+02	0.44	2.50E-04	6.55E+05
inflammatory response (GO:0006954)	0.50	3.23E-05	2.30E+02	0.72	3.23E-23	3.84E+02
protein modification by small protein removal (GO:0070646)	0.67	6.62E-03	1.39E+02	0.63	2.74E-04	1.25E+02
regulation of immune response (GO:0050776)	0.69	3.32E-03	2.62E+02	0.72	2.84E-25	5.29E+02
mRNA splicing, via spliceosome (GO:0000398)	0.42	3.83E-05	3.18E+02	0.65	1.11E-05	5.68E+02
mRNA processing (GO:0006397)	0.42	3.83E-05	3.67E+02	0.65	1.27E-05	6.39E+02
DNA metabolic process (GO:0006259)	0.22	7.56E-04	1.40E+02	1.00	2.59E-06	1.97E+02
interferon-gamma-mediated signaling pathway (GO:0060333)	0.75	1.14E-06	1.49E+03	0.21	5.67E-08	1.28E+03
epidermis development (GO:0008544)	0.08	8.41E-06	5.33E+02	0.35	6.74E-04	6.99E+02
RNA splicing, with bulged adenosine as nucleophile (GO:0000377)	0.42	3.83E-05	3.36E+02	0.56	1.27E-05	5.35E+02
positive regulation of response to external stimulus (GO:0032103)	0.69	8.96E-03	3.29E+02	0.35	2.81E-10	3.26E+02
chemokine-mediated signaling pathway (GO:0070098)	0.67	1.16E-03	4.40E+02	0.37	1.19E-08	4.94E+02
DNA repair (GO:0006281)	0.22	4.65E-03	1.36E+02	0.95	9.43E-03	1.36E+02
extracellular matrix organization (GO:0030198)	0.28	1.81E-03	6.33E+01	0.44	2.96E-06	9.18E+01
neutrophil mediated immunity (GO:0002446)	0.67	2.16E-05	2.48E+02	0.28	1.57E-08	1.12E+02
cellular response to DNA damage stimulus (GO:0006974)	0.11	2.73E-17	1.42E+02	0.95	6.21E-05	1.42E+02
neutrophil activation involved in immune response (GO:0002283)	0.67	2.16E-05	2.47E+02	0.28	4.96E-09	8.95E+01
neutrophil degranulation (GO:0043312)	0.67	2.16E-05	2.55E+02	0.26	5.39E-08	9.08E+01
cellular response to chemokine (GO:1990869)	0.72	1.75E-03	3.99E+02	0.26	4.05E-08	4.45E+02
protein ubiquitination (GO:0016567)	0.19	9.45E-03	8.06E+01	0.81	9.45E-03	1.55E+02
cellular protein modification process (GO:0006464)	0.25	1.93E-03	8.82E+01	0.70	2.97E-06	1.19E+02
antigen receptor-mediated signaling pathway (GO:0050851)	0.14	3.73E-06	1.63E+02	0.26	4.15E-11	1.63E+02
defense response to symbiont (GO:0140546)	0.75	1.15E-10	1.66E+03	0.23	7.76E-06	9.11E+05

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 [58]. 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 [60]. 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. In Table 6.16 we present the results from A549 cells, and in Table 6.17. There are many processes in common with Table 6.15, 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, 0.92 and 0.86).

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 and 0.92).

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, 0.92, 1.00) and NHBE cells (0.75 vs 0.91, 1.00).

These differences are consistent with those found by Blanco-Melo D. et al. [2], 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 6.18, 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 6.16: GO Biological processes with highest joint ranks for all viruses for the A549 cell type. Counts correspond to the normalized number of occurrences of each process within each condition.

	A549 SARS-CoV-2			A549 RSV		
	count	p-value	c-score	count	p-value	c-score
epidermis development (GO:0008544)	0.54	1.83E-04	1.44E+03	1.00	4.96E-06	7.01E+02
cellular response to interferon-gamma (GO:0071346)	0.66	1.15E-07	5.61E+02	0.85	1.15E-07	2.91E+02
cytokine-mediated signaling pathway (GO:0019221)	1.00	6.95E-03	6.06E+05	0.74	6.95E-03	9.57E+01
inflammatory response (GO:0006954)	0.45	5.30E-05	6.46E+02	0.97	1.23E-07	2.33E+02
interferon-gamma-mediated signaling pathway (GO:0060333)	0.54	1.03E-03	6.97E+02	0.74	1.03E-03	2.19E+02
antigen receptor-mediated signaling pathway (GO:0050851)	0.49	6.48E-06	8.87E+01	0.68	2.38E-04	8.87E+01
complement activation, classical pathway (GO:0006958)	0.39	6.59E-03	8.66E+03	0.91	4.52E-05	8.66E+03
skin development (GO:0043588)	0.46	9.85E-04	3.80E+02	0.71	8.82E-06	3.80E+02
cellular response to cytokine stimulus (GO:0071345)	0.61	2.50E-04	6.55E+05	0.41	5.74E-05	8.36E+01
chemokine-mediated signaling pathway (GO:0070098)	0.39	6.02E-05	3.99E+03	0.76	6.02E-05	5.04E+02
humoral immune response mediated by circulating immunoglobulin (GO:0002455)	0.37	8.06E-05	5.89E+03	0.91	8.06E-05	5.89E+03
positive regulation of response to external stimulus (GO:0032103)	0.39	3.65E-03	5.26E+02	0.56	3.65E-03	1.43E+02
epidermal cell differentiation (GO:0009913)	0.36	9.67E-04	9.29E+02	0.88	1.21E-07	9.29E+02
keratinocyte differentiation (GO:0030216)	0.36	1.51E-03	1.29E+03	0.85	1.31E-06	1.29E+03
regulation of immune response (GO:0050776)	0.34	3.64E-07	7.11E+02	0.94	3.06E-09	7.11E+02
cellular response to chemokine (GO:1990869)	0.36	8.06E-05	3.65E+03	0.68	8.06E-05	4.57E+02
positive regulation of defense response (GO:0031349)	0.36	2.07E-03	8.24E+02	0.38	2.07E-03	1.36E+02
exogenous peptide antigen via MHC class II (GO:0019886)	0.48	6.36E-03	6.51E+02	0.41	6.36E-03	6.51E+02
peptide antigen via MHC class II (GO:0002495)	0.48	6.81E-03	6.34E+02	0.38	6.81E-03	6.34E+02
positive regulation of chemotaxis (GO:0050921)	0.31	6.20E-04	3.42E+02	0.88	3.37E-03	3.42E+02
extracellular matrix organization (GO:0030198)	0.32	1.13E-03	1.05E+02	0.53	1.13E-03	3.98E+01
T cell receptor signaling pathway (GO:0050852)	0.41	2.70E-03	1.16E+02	0.29	8.46E-04	5.79E+01
antigen processing and presentation of exogenous peptide antigen (GO:0002478)	0.45	7.80E-03	6.10E+02	0.26	7.80E-03	6.10E+02
positive regulation of protein phosphorylation (GO:0001934)	0.33	4.81E-03	5.19E+01	0.32	4.81E-03	5.19E+01
mRNA processing (GO:0006397)	0.51	2.88E-06	7.02E+02	0.18	2.34E-08	3.67E+02

	A549 HPIV3			A549 IAV		
	count	p-value	c-score	count	p-value	c-score
epidermis development (GO:0008544)	1.00	1.83E-04	1.44E+03	1.00	2.38E-14	9.94E+02
cellular response to interferon-gamma (GO:0071346)	0.92	1.15E-07	7.32E+02	0.86	4.74E-06	2.91E+02
cytokine-mediated signaling pathway (GO:0019221)	0.90	6.95E-03	4.40E+02	1.00	2.87E-10	4.36E+02
inflammatory response (GO:0006954)	0.92	5.30E-05	1.97E+02	1.00	3.90E-04	2.53E+02
interferon-gamma-mediated signaling pathway (GO:0060333)	0.92	1.03E-03	9.72E+02	0.61	8.73E-03	2.19E+02
antigen receptor-mediated signaling pathway (GO:0050851)	0.63	6.48E-06	8.87E+01	0.59	1.92E-03	8.87E+01
complement activation, classical pathway (GO:0006958)	0.86	6.59E-03	8.66E+03	0.78	4.54E-05	8.66E+03
skin development (GO:0043588)	0.59	4.62E-04	3.80E+02	0.57	9.85E-04	3.80E+02
cellular response to cytokine stimulus (GO:0071345)	0.55	5.74E-05	1.49E+02	0.71	5.74E-05	1.69E+02
chemokine-mediated signaling pathway (GO:0070098)	0.71	6.02E-05	8.65E+02	0.67	7.71E-03	5.04E+02
humoral immune response mediated by circulating immunoglobulin (GO:0002455)	0.82	8.06E-05	5.89E+03	0.76	5.90E-05	5.89E+03
positive regulation of response to external stimulus (GO:0032103)	0.71	3.65E-03	1.95E+02	0.76	4.51E-03	1.47E+02
epidermal cell differentiation (GO:0009913)	0.71	8.76E-05	9.29E+02	0.76	9.67E-04	9.29E+02
keratinocyte differentiation (GO:0030216)	0.71	6.21E-04	1.29E+03	0.76	1.51E-03	1.29E+03
regulation of immune response (GO:0050776)	0.88	3.64E-07	2.29E+02	0.88	1.18E-08	7.11E+02
cellular response to chemokine (GO:1990869)	0.67	8.06E-05	7.92E+02	0.59	1.94E-04	4.57E+02
positive regulation of defense response (GO:0031349)	0.45	2.07E-03	2.17E+02	0.75	2.07E-03	1.37E+02
exogenous peptide antigen via MHC class II (GO:0019886)	0.43	6.36E-03	6.51E+02	0.29	9.02E-03	6.51E+02
peptide antigen via MHC class II (GO:0002495)	0.43	6.81E-03	6.34E+02	0.27	9.02E-03	6.34E+02
positive regulation of chemotaxis (GO:0050921)	0.80	6.20E-04	4.14E+02	0.82	6.71E-03	3.42E+02
extracellular matrix organization (GO:0030198)	0.59	1.13E-03	3.90E+01	0.57	1.13E-03	4.23E+01
T cell receptor signaling pathway (GO:0050852)	0.35	2.70E-03	6.64E+01	0.24	2.46E-03	3.68E+01
antigen processing and presentation of exogenous peptide antigen (GO:0002478)	0.31	7.80E-03	6.10E+02	0.16	9.02E-03	6.10E+02
positive regulation of protein phosphorylation (GO:0001934)	0.29	4.81E-03	5.19E+01	0.31	5.56E-03	5.19E+01
mRNA processing (GO:0006397)	0.29	3.83E-05	5.11E+02	0.20	2.88E-06	3.67E+02

Table 6.17: GO Biological processes with highest joint ranks for all viruses for the NHBE cell type. Counts correspond to the normalized number of occurrences of each process within each condition.

GO Biological Processes	NHBE SARS-CoV-2			NHBE IAV		
	count	p-value	c-score	count	p-value	c-score
extracellular matrix organization (GO:0030198)	1.00	3.31E-08	1.70E+02	1.00	3.31E-08	1.70E+02
cardiac muscle tissue development (GO:0048738)	0.79	4.78E-03	3.00E+02	0.91	4.78E-03	3.00E+02
inflammatory response (GO:0006954)	0.75	3.90E-04	1.47E+02	0.91	3.90E-04	2.71E+02
dendritic cell migration (GO:0036336)	0.75	6.79E-03	5.92E+02	0.86	6.79E-03	5.92E+02
dendritic cell chemotaxis (GO:0002407)	0.75	6.71E-03	7.08E+02	0.86	6.71E-03	7.08E+02
cellular response to interferon-gamma (GO:0071346)	0.88	4.31E-03	3.97E+02	0.79	4.31E-03	5.12E+02
extracellular structure organization (GO:0043062)	0.75	1.10E-07	8.85E+01	0.86	1.10E-07	8.85E+01
external encapsulating structure organization (GO:0045229)	0.75	4.67E-08	9.64E+01	0.86	4.67E-08	9.64E+01
regulation of immune response (GO:0050776)	0.71	5.14E-03	1.85E+02	0.79	5.14E-03	1.85E+02
negative regulation of T cell activation (GO:0050868)	0.75	9.83E-03	1.13E+03	0.77	3.11E-03	4.28E+02
phospholipase receptor signaling pathway (GO:0007200)	0.67	9.33E-03	2.60E+02	0.81	9.33E-03	2.60E+02
regulation of T cell proliferation (GO:0042129)	0.67	7.93E-04	4.13E+02	0.77	7.93E-04	4.13E+02
chemokine-mediated signaling pathway (GO:0070098)	0.67	1.55E-04	5.04E+02	0.67	1.55E-04	5.04E+02
positive regulation of chemotaxis (GO:0050921)	0.62	6.71E-03	3.42E+02	0.79	6.71E-03	9.51E+02
cellular response to cytokine stimulus (GO:0071345)	0.92	2.50E-04	6.55E+05	0.70	2.90E-03	1.42E+02
positive regulation of lymphocyte proliferation (GO:0050671)	0.54	3.72E-03	1.66E+02	0.70	3.72E-03	1.66E+02
complement activation, classical pathway (GO:0006958)	0.58	4.54E-05	8.66E+03	0.70	4.54E-05	9.67E+03
nervous system development (GO:0007399)	0.54	3.84E-03	3.69E+01	0.74	3.84E-03	3.74E+01
heart development (GO:0007507)	0.54	2.69E-03	7.04E+01	0.74	2.69E-03	7.04E+01
positive regulation of MAPK cascade (GO:0043410)	0.54	9.24E-06	2.95E+02	0.70	9.24E-06	2.95E+02
cellular response to chemokine (GO:1990869)	0.62	1.94E-04	4.57E+02	0.63	1.94E-04	4.57E+02
regulation of calcium ion-dependent exocytosis (GO:0017158)	0.54	4.08E-03	1.91E+02	0.72	4.08E-03	1.91E+02
positive regulation of ERK1 and ERK2 cascade (GO:0070374)	0.54	6.12E-05	2.77E+02	0.70	6.12E-05	2.77E+02
B cell receptor signaling pathway (GO:0050853)	0.54	7.09E-04	5.07E+02	0.70	7.09E-04	5.07E+02
calcium-mediated signaling (GO:0019722)	0.54	6.85E-03	1.07E+02	0.70	6.85E-03	1.07E+02

GO Biological Processes	NHBE IAVdNS1		
	count	p-value	c-score
extracellular matrix organization (GO:0030198)	0.84	3.31E-08	1.70E+02
cardiac muscle tissue development (GO:0048738)	0.80	4.78E-03	3.00E+02
inflammatory response (GO:0006954)	1.00	3.90E-04	2.71E+02
dendritic cell migration (GO:0036336)	0.77	6.79E-03	5.92E+02
dendritic cell chemotaxis (GO:0002407)	0.77	6.71E-03	7.08E+02
cellular response to interferon-gamma (GO:0071346)	0.75	7.60E-04	1.13E+03
extracellular structure organization (GO:0043062)	0.70	1.10E-07	8.85E+01
external encapsulating structure organization (GO:0045229)	0.70	4.67E-08	9.64E+01
regulation of immune response (GO:0050776)	0.82	5.14E-03	1.85E+02
negative regulation of T cell activation (GO:0050868)	0.66	3.11E-03	4.28E+02
phospholipase receptor signaling pathway (GO:0007200)	0.66	9.33E-03	2.60E+02
regulation of T cell proliferation (GO:0042129)	0.66	7.93E-04	4.13E+02
chemokine-mediated signaling pathway (GO:0070098)	0.68	4.01E-04	5.04E+02
positive regulation of chemotaxis (GO:0050921)	0.59	6.71E-03	9.51E+02
cellular response to cytokine stimulus (GO:0071345)	0.52	2.90E-03	2.98E+02
positive regulation of lymphocyte proliferation (GO:0050671)	0.70	8.75E-03	3.05E+02
complement activation, classical pathway (GO:0006958)	0.61	4.54E-05	9.67E+03
nervous system development (GO:0007399)	0.61	3.84E-03	3.56E+01
heart development (GO:0007507)	0.61	2.69E-03	7.04E+01
positive regulation of MAPK cascade (GO:0043410)	0.66	9.24E-06	2.95E+02
cellular response to chemokine (GO:1990869)	0.75	4.14E-04	4.57E+02
regulation of calcium ion-dependent exocytosis (GO:0017158)	0.59	4.08E-03	1.91E+02
positive regulation of ERK1 and ERK2 cascade (GO:0070374)	0.61	6.12E-05	2.77E+02
B cell receptor signaling pathway (GO:0050853)	0.61	7.09E-04	5.07E+02
calcium-mediated signaling (GO:0019722)	0.61	6.85E-03	1.07E+02

Table 6.18: 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
Biclustering	MCS (p <0.01)	4440	2086	1184
	NHBE	2912	685	305
	A549	3926	779	273

7

Conclusion

Contents

7.1 Concluding Remarks

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 biclustering 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.

These gene sets were then analyzed using functional enrichment tools, and the resulting biological mechanisms were compared to those identified by other authors, in addition to those identified using the three aforementioned methods. In particular, with biclustering we were able to more directly and robustly compare the enriched terms obtained using the described process.

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. [2]. 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. and co-workers, however it is consistent with other literature on coronaviruses (cited throughout chapter 6), and further supports the hypothesis that a pattern-centric view of the gene enrichment process can result in novel information.

7.2 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.

7.3 Scientific Communication

The main contributions presented in this dissertation have been submitted to scientific journals. Additionally, the code used to produce the presented results is available in this thesis' 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.

Bibliography

[1] X. Zou, K. Chen, J. Zou, P. Han, J. Hao, and Z. Han, "Single-cell rna-seq data analysis on the receptor ace2 expression reveals the potential risk of different human organs vulnerable to 2019-ncov infection," *Frontiers of medicine*, pp. 1–8, 2020.

[2] D. Blanco-Melo, B. E. Nilsson-Payant, W.-C. Liu, S. Uhl, D. Hoagland, R. Møller, T. X. Jordan, K. Oishi, M. Panis, D. Sachs *et al.*, "Imbalanced host response to sars-cov-2 drives development of covid-19," *Cell*, 2020.

[3] F. S. Cohen, "How viruses invade cells," *Biophysical Journal*, vol. 110, no. 5, p. 1028, 2016.

[4] B. J. Bosch, R. Van der Zee, C. A. De Haan, and P. J. Rottier, "The coronavirus spike protein is a class i virus fusion protein: structural and functional characterization of the fusion core complex," *Journal of virology*, vol. 77, no. 16, pp. 8801–8811, 2003.

[5] H. Zhang, J. M. Penninger, Y. Li, N. Zhong, and A. S. Slutsky, "Angiotensin-converting enzyme 2 (ace2) as a sars-cov-2 receptor: molecular mechanisms and potential therapeutic target," *Intensive care medicine*, vol. 46, no. 4, pp. 586–590, 2020.

[6] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche *et al.*, "Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor," *Cell*, 2020.

[7] P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, and V. Thiel, "Coronavirus biology and replication: implications for sars-cov-2," *Nature Reviews Microbiology*, pp. 1–16, 2020.

[8] R. L. Skalsky and B. R. Cullen, "Viruses, micrornas, and host interactions," *Annual review of microbiology*, vol. 64, pp. 123–141, 2010.

[9] C. Cowled, C. R. Stewart, V. A. Likic, M. R. Friedländer, M. Tachedjian, K. A. Jenkins, M. L. Tizard, P. Cottee, G. A. Marsh, P. Zhou *et al.*, "Characterisation of novel micrornas in the black flying fox (pteropus alecto) by deep sequencing," *BMC genomics*, vol. 15, no. 1, p. 682, 2014.

[10] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of sars-cov-2," *Nature medicine*, vol. 26, no. 4, pp. 450–452, 2020.

[11] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Hughes, E. Snesrud, N. Lee, and J. Quackenbush, "A concise guide to cdna microarray analysis," *Biotechniques*, vol. 29, no. 3, pp. 548–562, 2000.

[12] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[13] M. F. Murray, E. E. Kenny, M. D. Ritchie, D. J. Rader, A. E. Bale, M. A. Giovanni, and N. S. Abul-Husn, "Covid-19 outcomes and the human genome," *Genetics in Medicine*, pp. 1–3, 2020.

[14] C. K. Kang, M.-W. Seong, S.-J. Choi, T. S. Kim, P. G. Choe, S. H. Song, N.-J. Kim, W. B. Park, and M.-d. Oh, "In vitro activity of lopinavir/ritonavir and hydroxychloroquine against severe acute respiratory syndrome coronavirus 2 at concentrations achievable by usual doses," *The Korean journal of internal medicine*, vol. 35, no. 4, p. 728, 2020.

[15] J. Andreani, M. Le Bideau, I. Duflot, P. Jardot, C. Rolland, M. Boxberger, N. Wurtz, J.-M. Rolain, P. Colson, B. La Scola *et al.*, "In vitro testing of combined hydroxychloroquine and azithromycin on sars-cov-2 shows synergistic effect," *Microbial pathogenesis*, vol. 145, p. 104228, 2020.

[16] P. A. McGettigan, "Transcriptomics in the rna-seq era," *Current opinion in chemical biology*, vol. 17, no. 1, pp. 4–11, 2013.

[17] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi, "Clustering algorithms: their application to gene expression data," *Bioinformatics and Biology insights*, vol. 10, pp. BBI–S38 316, 2016.

[18] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *Journal of biomedical informatics*, vol. 57, pp. 163–180, 2015.

[19] C. M. Kitchen, "Nonparametric vs parametric tests of location in biomedical research," *American journal of ophthalmology*, vol. 147, no. 4, pp. 571–572, 2009.

[20] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908. [Online]. Available: http://www.jstor.org/stable/2331554

[21] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[22] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

[23] F. Deng, L. Shen, H. Wang, and L. Zhang, “Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models,” *American journal of cancer research*, vol. 10, no. 12, p. 4624, 2020.

[24] F. Yuan, X. Pan, T. Zeng, Y.-H. Zhang, L. Chen, Z. Gan, T. Huang, and Y.-D. Cai, “Identifying cell-type specific genes and expression rules based on single-cell transcriptomic atlas data,” *Frontiers in bioengineering and biotechnology*, vol. 8, p. 350, 2020.

[25] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[26] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, “Challenges in unsupervised clustering of single-cell rna-seq data,” *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282, 2019.

[27] Y. Cheng and G. M. Church, “Biclustering of expression data.” in *Ismb*, vol. 8, no. 2000, 2000, pp. 93–103.

[28] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: a survey,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

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

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

[31] R. Henriques and S. C. Madeira, “Bicpam: Pattern-based biclustering for biomedical data analysis,” *Algorithms for Molecular Biology*, vol. 9, no. 1, pp. 1–30, 2014.

[32] R. L. Tillett, J. R. Sevinsky, P. D. Hartley, H. Kerwin, N. Crawford, A. Gorzalski, C. Laverdure, S. C. Verma, C. C. Rossetto, D. Jackson *et al.*, “Genomic evidence for reinfection with sars-cov-2: a case study,” *The Lancet Infectious Diseases*, 2020.

[33] 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.

[34] 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.

[35] J. Wei, M. Alfajaro, R. Hanna, P. DeWeirdt, M. Strine, W. Lu-Culligan, S.-M. Zhang, V. Graziano, C. Schmitz, J. Chen *et al.*, "Genome-wide crispr screen reveals host genes that regulate sars-cov-2 infection," *Biorxiv*, 2020.

[36] E. Wyler, K. Mösbauer, V. Franke, A. Diag, L. T. Gottula, R. Arsie, F. Klironomos, D. Koppstein, S. Ayoub, C. Buccitelli *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.

[37] B. K. Manne, F. Denorme, E. A. Middleton, I. Portier, J. W. Rowley, C. Stubben, A. C. Petrey, N. D. Tolley, L. Guo, M. Cody *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.

[38] J. Golden, C. Cline, X. Zeng, A. Garrison, B. Carey, E. Mucker, L. White, J. Shamblin, R. Brocato, J. Liu *et al.*, "Human angiotensin-converting enzyme 2 transgenic mice infected with sars-cov-2 develop severe and fatal respiratory disease," *bioRxiv*, 2020.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–14, 2013.

[44] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[45] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

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

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

[48] Y. Liang, M.-L. Wang, C.-S. Chien, A. A. Yarmishyn, Y.-P. Yang, W.-Y. Lai, Y.-H. Luo, Y.-T. Lin, Y.-J. Chen, P.-C. Chang *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.

[49] 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.

[50] 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.

[51] 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.

[52] L. Zhao, B. K. Jha, A. Wu, R. Elliott, J. Ziebuhr, A. E. Gorbalenya, R. H. Silverman, and S. R. Weiss, “Antagonism of the interferon-induced oas-rnase l pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology,” *Cell host & microbe*, vol. 11, no. 6, pp. 607–616, 2012.

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

[54] P. K. Pribul, J. Harker, B. Wang, H. Wang, J. S. Tregoning, J. Schwarze, and P. J. Openshaw, “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.

[55] C. Schneider, S. P. Nobs, A. K. Heer, M. Kurrer, G. Klinke, N. Van Rooijen, J. Vogel, and M. Kopf, “Alveolar macrophages are essential for protection from respiratory failure and associated morbidity following influenza virus infection,” *PLoS pathogens*, vol. 10, no. 4, p. e1004053, 2014.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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