Computers in Biology and Medicine 146 (2022) 105443

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Enrichment analysis on regulatory subspaces: A novel direction for the superior description of cellular responses to SARS-CoV-2

Pedro Rodrigues [a,b], Rafael S. Costa [a,c], Rui Henriques [b,*]

[a] IDMEC, Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal
[b] INESC-ID and Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal
[c] LAQV-REQUIMTE, DQ, NOVA School of Science and Technology, Caparica, Portugal

ARTICLE INFO

Keywords:
SARS-CoV-2
Biclustering
Discriminative regulatory patterns
Transcriptomics
Machine learning
Computational biology
COVID-19

ABSTRACT

Statement: Enrichment analysis of cell transcriptional responses to SARS-CoV-2 infection from biclustering solutions yields broader coverage and superior enrichment of GO terms and KEGG pathways against alternative state-of-the-art machine learning solutions, thus aiding knowledge extraction.
Motivation and methods: The comprehensive understanding of the impacts of SARS-CoV-2 virus on infected cells is still incomplete. This work aims at comparing the role of state-of-the-art machine learning approaches in the study of cell regulatory processes affected and induced by the SARS-CoV-2 virus using transcriptomic data from both infectable cell lines available in public databases and in vivo samples. In particular, we assess the relevance of clustering, biclustering and predictive modeling methods for functional enrichment. Statistical principles to handle scarcity of observations, high data dimensionality, and complex gene interactions are further discussed. In particular, and without loos of generalization ability, the proposed methods are applied to study the differential regulatory response of lung cell lines to SARS-CoV-2 (α-variant) against RSV, IAV (H1N1), and HPIV3 viruses.
Results: Gathered results show that, although clustering and predictive algorithms aid classic stances to functional enrichment analysis, more recent pattern-based biclustering algorithms significantly improve the number and quality of enriched GO terms and KEGG pathways with controlled false positive risks. Additionally, a comparative analysis of these results is performed 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. The findings are particularly relevant given the lack of other works utilizing more complex machine learning algorithms within this context.

1. Introduction

The infection of humans by Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) represents a major global health concern, with deaths having surpassed 5.8 million according to the World Health Organization.[1] Worldwide initiatives to publicly share data related to the virus provides an opportunity to draw novel insights on the infection by the family of coronaviruses, enabling continuous breakthroughs on the understanding of how the virus can enter and use the cellular machinery to replicate. The knowledge of these mechanisms has been propelled by a generic understanding of the process of viral replication, the transcriptomic properties of the virus, and the study of differentially expressed genes after infection [1,2]. This later line of research has been primarily assisted by the sequencing of RNA transcripts in infectable cell lines, chosen according to the level of permissivity to infection, as well as cells collected from organisms susceptible to infection, including humans and ferrets [1]. Along this line of research, comparisons of infection in different cells and tissues have been pursued, as well as between different viral strains and families of virus [1].

Despite the ongoing breakthroughs, the regulatory responses to SARS-CoV-2 infection are still not comprehensively known [3]. For instance, the role played by genes with moderate differential expression in response to infection, or how interactions between multiple genes support or prevent viral replication, are still being actively updated [4]. In addition, most works in this field do not explore the role of machine learning approaches, such as clustering, predictive modeling and

* Corresponding author.
E-mail address: rmch@tecnico.ulisboa.pt (R. Henriques).

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019, accessed on Dec. 5th, 2021.

https://doi.org/10.1016/j.compbiomed.2022.105443
Received 16 December 2021; Received in revised form 13 March 2022; Accepted 20 March 2022
Available online 25 April 2022

biclustering, to aid in the identification of differentially expressed regulatory modules [5].

This work aims to address these challenges by assessing the extent to which clustering, predictive modeling and biclustering methods aid the identification of elicited biological functions and pathways from transcriptomic data in response to infection. The aforementioned machine learning approaches are placed to model differential regulatory responses after infection from both infectable lung cell lines and in vivo tissue samples. In addition, a comprehensive analysis of the enriched processes and pathways using these methods is undertaken to better understand the viral life-cycle and interactions with the cell, as well as the defence mechanisms employed by the cell against the virus. In particular, SARS-CoV-2 infection is assessed against non-infected cells and infection by other respiratory viruses such as the RSV, IAV (H1N1), and HPIV3. SARS-CoV-2 (α-variant) is targeted in this study, yet the underlying methodological principles extensible to other variants and viruses.

Four major contributions are provided. *First*, the role of different machine learning approaches to produce relevant gene sets for enrichment analysis is experimentally compared. *Second*, state-of-the-art biclustering algorithms are assessed and compared against alternative descriptive and predictive stances. The gathered results reveal that biclustering significantly assists the knowledge acquisition process. In particular, the recent class of pattern-based biclustering approaches [6, 7] show distinctive ability to produce a comprehensive set of superiorly enriched biological annotations in well-established knowledge bases without an increase on false positive discoveries. *Third*, grounded on the previous findings, a novel methodology is provided for a robust and comprehensive analysis of putative regulatory modules associated with virus infection. *Fourth*, under this methodology, we further identify and highlight the putative role of less-studied biological processes associated with SARS-CoV-2 infection, some consistent with literature on other coronaviruses. In particular, cell- and virus-specific regulatory differences are further identified.

The manuscript is organized as follows: section 2 covers related contributions; section 3 explores the datasets; section 4 presents the proposed methodology; section 5 experimentally compares the role of state-of-the-art machine learning approaches to aid enrichment analysis, together with the description of the identified biological processes. Finally, major concluding remarks are drawn.

2. Related work

Blanco-Melo et al. [1] profiled the transcriptional response of different cell lines to infection by SARS-CoV-2 and other respiratory viruses, including RSV, IAV and HPIV3. In their work, these diverse transcriptomic data sources are consolidated and further integrated with experimental data from MERS-CoV and SARS-CoV-1 infection collected by Frieman et al. [8]. The profiled cells consisted in three main groups: i) respiratory cell lines, including NHBE, A549 and Calu-3 cells; ii) human respiratory tract cells extracted from infected and non-infected individuals; and iii) cells extracted from infected and non-infected ferrets [1]. The second and third groups were used to ascertain if the gene signatures matched the ones found *in vitro*. Additionally, the authors treated some of the cell lines with universal IFNβ to determine whether or not SARS-CoV-2 is sensitive to IFN-I. The treatment resulted in significantly decreased viral replication, confirming the hypothesized sensitivity in earlier works [9]. To investigate how infection affects the cell transcriptome, the authors performed a differential expression analysis on NHBE cells, revealing significant differences between the response from SARS-CoV-2 infection and other viral strains. Functional enrichment was further performed on the differentially expressed genes to better understand the cellular functions affected by SARS-CoV-2 infection. Consistently across experiments, the authors highlight the production of cytokines and associated transcriptional responses as pivotal pathways, as well as the induction of a subset of interferon-stimulated genes (ISGs).

The transcriptional response of human cells to SARS-CoV-2 infection, and its comparison them with MERS-CoV, SARS-CoV-1 and IAV, has been complementarily analysed to identify possible common impacts between viral strains [10]. The authors generated consensomes by analysing how frequently the corresponding genes were differentially expressed throughout the various datasets. Similarly to Blanco-Melo et al. [1], the authors found ISGs to have significant induction levels.

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) was conducted in Ref. [11]. 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 revealed 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, among other cytokines, in agreement with the findings of other authors.

In [12], genome-wide CRISPR screening is performed on an African green monkey cell line (Vero-E6) to identify genetic sequences aiding (pro-viral) or preventing (anti-viral) infection. To this end, surviving cells from populations infected with SARS-CoV-2 were harvested 7 days post-infection. A genome-wide screen was performed and a z-score applied to identify genes associated with increased or decreased resistance to SARS-CoV-2-induced cell death. The gene with the strongest pro-viral effect was ACE2, the protein facilitating viral entry into the cell. TMPRSS2, another gene with an established role in the SARS-CoV-2 entry, was not identified significantly as pro or anti-viral, whereas the CTSL gene, which encodes the Cathepsin L protease with an identified role in viral entry, was identified as pro-viral.

Due to thrombotic complications being common among COVID-19 patients, the functional and transcriptional changes elicited by SARS-CoV-2 infection in platelets have been further explored [13]. The conducted analysis shows that SARS-CoV-2 infection does indeed alter the platelet transcriptional activity [13]. To assess the significance of differential changes, paired t-Student and Mann-Whitney tests are considered. The authors observed that COVID-19 further 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.

In [14], the authors 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 in humans.

Though there are multiple authors applying machine learning and complex statistical models to COVID-19 patient biometric data, 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 these approaches to this data can assist functional enrichment analysis, yielding novel insights into the disease.

3. Dataset

In order to assess the proposed methods, we use the transcriptomic data (RNA-Seq) collected by Blanco et al. (Gene Expression Omnibus, GEO accession GSE147507[2]) [1]. A schematic of its structure is presented in Fig. 1. The samples are divided into different *series* (a subset of samples), each comprising the behavior of a single cell line among different sets of experimental conditions. These also correspond to

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

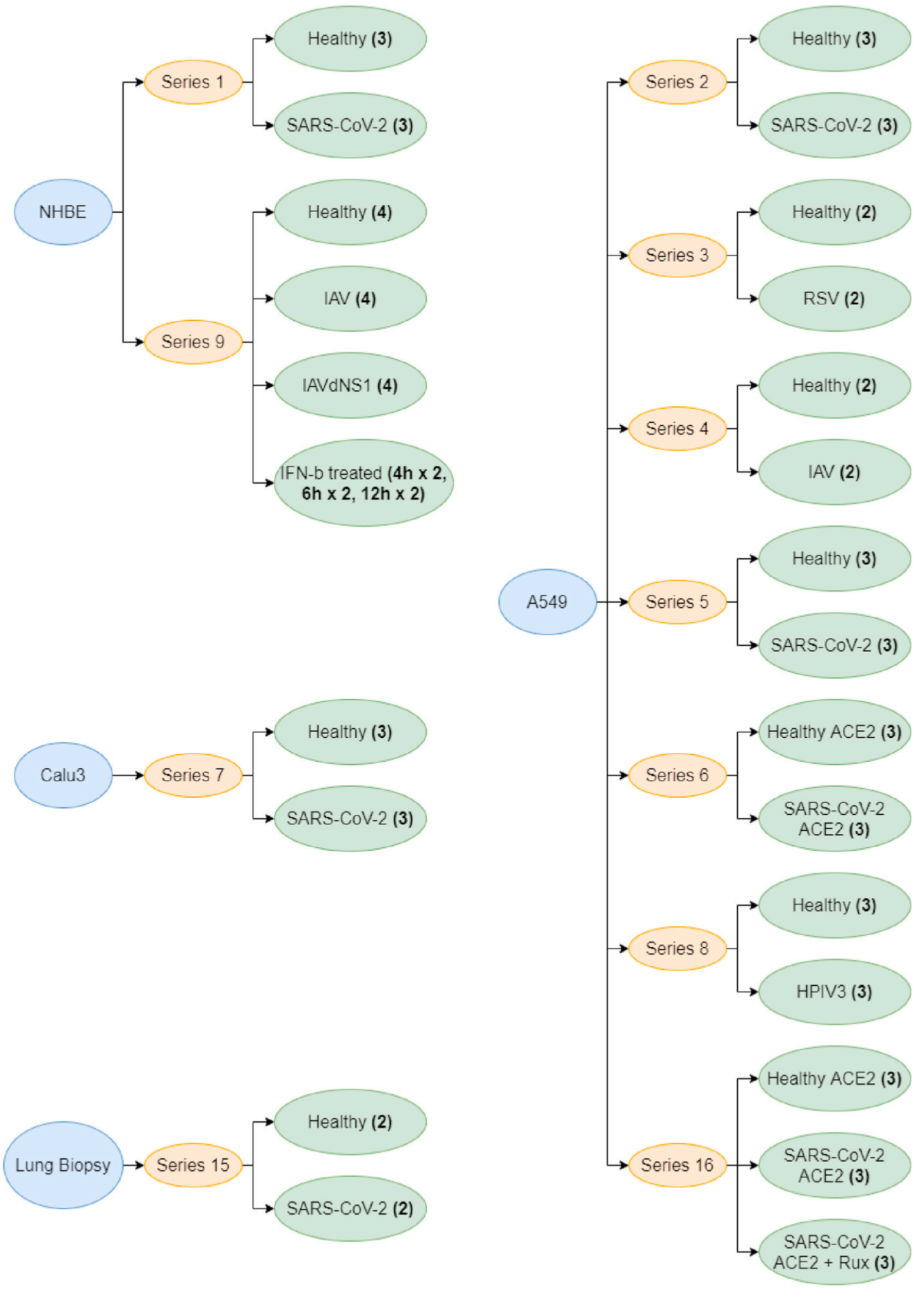

Fig. 1. Overview of the structure of the dataset used in this study. Numbers between parentheses represent the number of data points.

particular experiments being run, with each experiment containing multiple replicas of each experimental condition being tested.

Three major cell lines are considered: NHBE (normal human bronchial epithelial), A549 (adenocarcinomic human alveolar basal epithelial) and Calu3 (generated from a bronchial adenocarcinoma) cells. Considering NHBE cells, there is a total of 7 samples of healthy cells (spanned along three *series*), 3 samples of SARS-CoV-2 α-variant 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 h post treatment. Considering A549 cells, there are 13 samples of healthy cells (distributed along five *series*), 6 samples of SARS-CoV-2 α-variant infection (three each in *series* 2 and 5), 2 samples of IAV infection (*series* 4), 2 samples of RSV infection (*series* 3) and 3 samples of HPIV3 infection (*series* 8). Blanco-Melo et al. [1] notes that A549 cells show low viral counts, a fact posited, in agreement with other studies, to be due to the low expression of ACE2 in these cells. Thus, data of A549 cells with added ACE2 (A549-ACE2) is also available. In particular, 6 samples of healthy cells (*series* 6 and 16), 6 samples of cells infected by SARS-CoV-2 α-variant (*series* 6 and 16) and, finally, 3 samples of cells after treatment with Ruxolitinib (*series* 16). For Calu3 cells, there are 3 samples of healthy cells and 3 samples of cells infected by SARS-CoV-2 α-variant (all belonging to *series* 7). Complementarily to infectable lung cell lines, there is an additional set of 2 samples from a lung biopsy of two healthy human donors (one male, one female), and 2 samples from a single deceased male patient of COVID-19.

Since the distribution of transcript counts is understandably significantly skewed, gene expression was adjusted by a log2-transform for all subsequent analysis. Fig. 2 depicts the distribution of expression levels among *in-vitro* cell lines and lung biopsies. The standard deviation of gene expression within healthy and infected cells was subsquently computed to preliminarily verify if there are significant differences between healthy and infected cells (Fig. 3).

In order to select an appropriate statistical test to identify

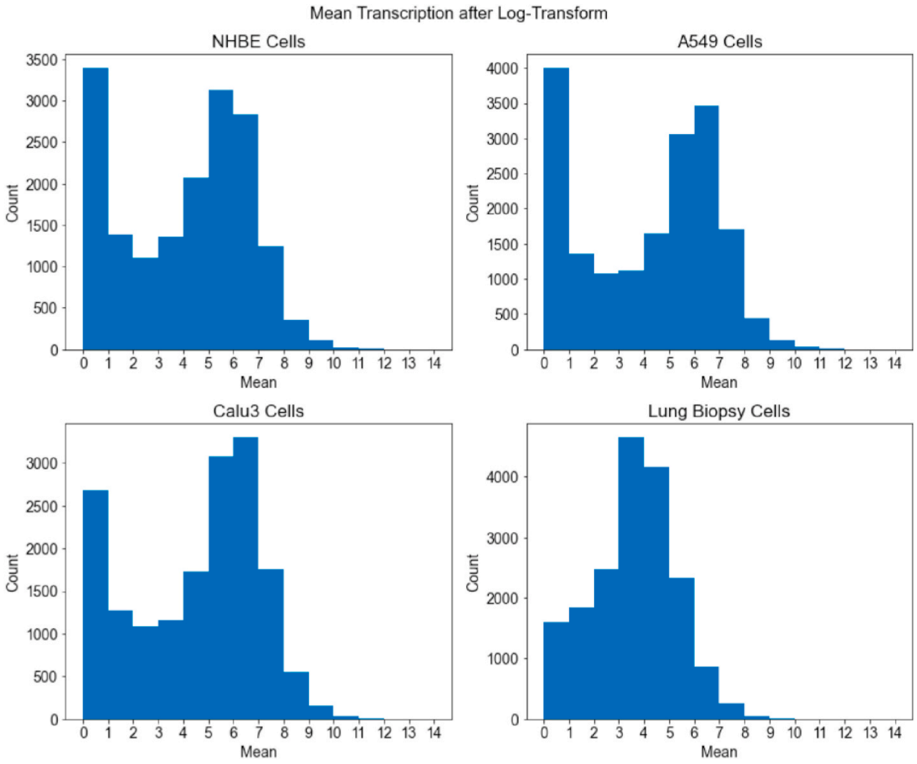

Fig. 2. Distribution of gene expression (mean among samples) after applying a log2 transform (N = 21797 genes).

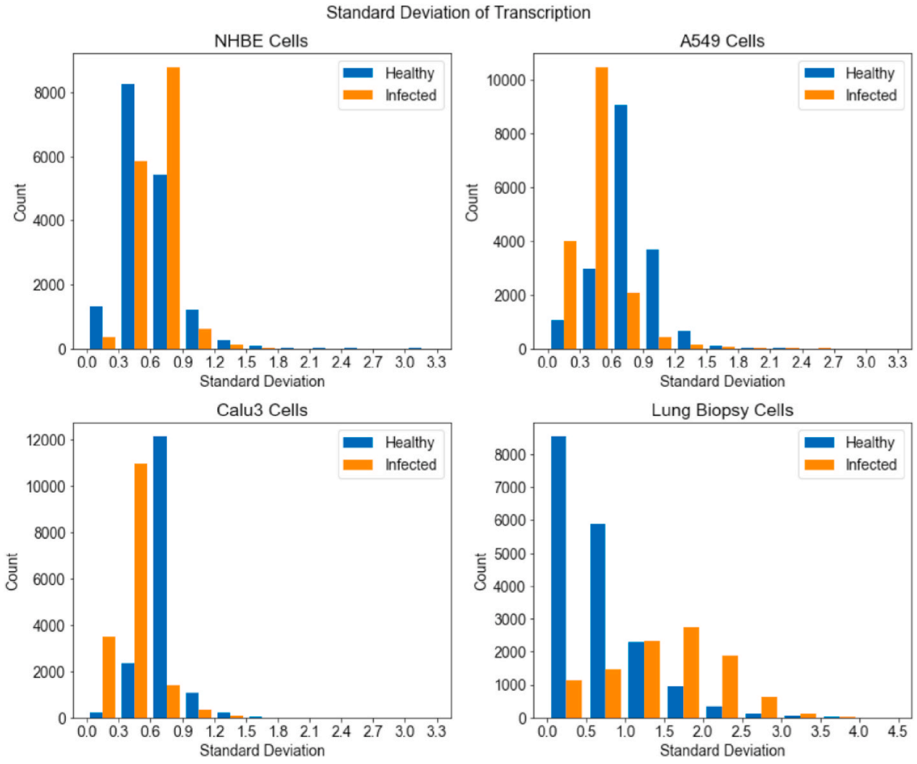

Fig. 3. Variability of gene expression within healthy and within infected cells.

differentially expressed genes, a number of assumptions needs to be assessed. Firstly, we performed a median-based Levene's test [15] to assess the equality of variances for each pair of conditions (Table 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, we applied the Shapiro-Wilk test [16] to assess whether these genes follow a normal distribution, applied to healthy and infected lines for each cell type with $p < 0.05$. Overall, 32.8%, 46.1% and 27.4% of all genes in NHBE, A549 and Calu3 cells, respectively, are non-normally distributed.

Table 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

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 its non-parametric alternative, the Mann-Whitney *U* test [17], are more suitable for variable selection. As results reveal a significant percentage of non-normally distributed genes, Mann-Whitney *U* tests are selected as a baseline to identify differentially expressed genes.

4. Methodology

To comprehensively unravel the biological processes involved in the cell response to SARS-CoV-2 infection, we explore the role of state-of-the-art machine learning approaches to find discriminative transcriptional modules. In this context, we propose a methodology for the selection and discovery of correlated groups of differentially expressed genes (DEG) composed of three major steps. First, preprocessing and preliminary gene selection are undertaken. Second, we proceed with pattern recognition 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 putative 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 Fig. 4.

4.1. Preprocessing and gene selection

Given the skewed distribution of gene expression (with a vast majority of genes having low expression levels), we first apply a log2 transform. Then, from the high-dimensional set of over 20 000 genes, we select a subset of DEGs. Due to the non-normal nature of data and unequal variance between control and test groups (as seen in section 3), Mann-Whitney *U* test is applied with $p < 0.05$ and $p < 0.01$. By default a $p < 0.01$ is used, however for certain cell types $p < 0.05$ is suggested to guarantee a better coverage of gene candidates to the subsequent learning stage. The null hypothesis of Mann Whitney *U* test is that the two assessed populations are equal. Therefore, this test can only be applied for pairs of conditions. In particular, we consider the following settings:

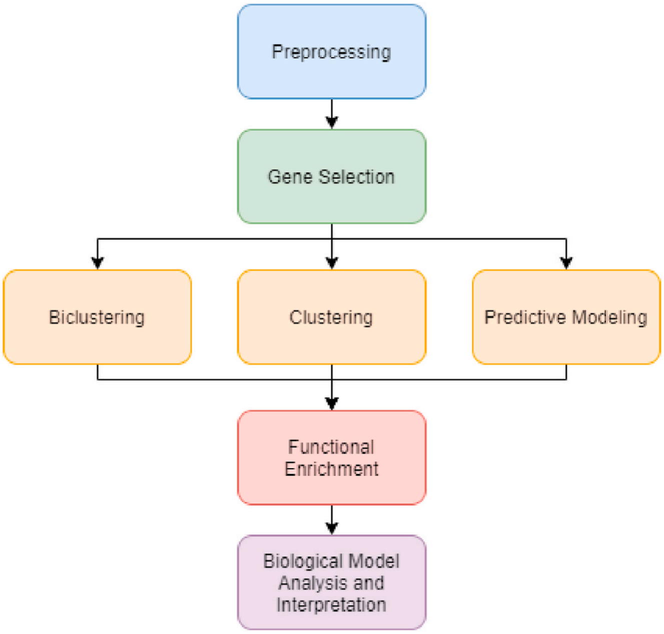

Fig. 4. Schematic diagram of the steps composing the proposed computational pipeline.

- *Paired setting*: single pairs of conditions, e.g. healthy and SARS-CoV-2 infected NHBE cells or healthy and IAV infected A549 cells;
- *Multi-condition setting*: genes are selected if they show differential expression in one of the various pairs of conditions presented in Table 1. For each set of pairs, a Mann-Whitney *U* test is applied for each gene. Genes satisfying $p < 0.01$ or $p < 0.05$ are selected.

Additionally, ANOVA test [18] can be optionally applied to further ensure the discriminative power of the resulting set of differentially expressed genes. This is suggested if the subsequent learning step benefits from a reduced dimensionality by requiring candidate genes to satisfy two distinct statistical criteria.

4.2. Pattern recognition

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 1000 genes. Applying functional enrichment to these results delivers none or very few enriched processes, which, when present, tend to be very generic cell functions. In this context, the pursue of putative transcriptional modules given by smaller sets of DEG is attempted to obtain more specific biological processes, as well as better statistical significance for each one found. To achieve this goal, three major approaches are applied: *clustering*, *predictive modeling* and *biclustering*.

4.2.1. Clustering

Given a set of genes *G* and sample *X*, the *clustering* task aims to find groups (clusters) of genes, $\{J_1, .., J_k\}$ where $J_i \subseteq G$, or conditions, $\{I_1, .., I_k\}$ where $I_i \subseteq X$, maximizing intra-cluster similarity and inter-cluster dissimilarity. As introduced, the notion of a cluster can assume two distinct forms – a subset of correlated genes along a given set of samples or a subset of correlated samples along a given set of genes. The former is suggested to identify sets of co-expressed genes, which further satisfy delineate discriminative criteria satisfied in the precedent feature selection step (section 4.1). Agglomerative clustering is considered in this work with Euclidean affinity and Ward linkage for two main reasons: the easy visualization of gene proximity using dendrogram, which can also help with the selection of the number of clusters; and inherent algorithmic flexibility, allowing for parameters to be adjusted according to the provided data.

Despite the relevance of clustering for enrichment analysis [19,20], it has considerable limitations. Namely, similarity between genes is assessed across all samples. If multiple conditions are used simultaneously, such information will not be taken into account, imposing similarity across all conditions and biasing the detected patterns. To ameliorate this effect, clusters can be found on subsets of conditions or individual conditions. As a result, multiple clustering solutions can be acquired, providing a wide diversity of sets of correlated genes with potential biological relevance. However, such disaggregation prevents a direct comparison between different conditions.

4.2.2. Machine learning models for classification

Consider the target multivariate data, described by a set of samples (observations), X, with expression measured along a high-dimensional set of genes, G, and each sample annotated with the corresponding condition (e.g. IAV infection), $c \in C$. Given a set of annotated data samples, the classification task aims at learning a mapping between samples and conditions, $X \rightarrow C$, on the basis of the underlying transcriptional activity.

As 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 inspecting the learned predictive model. For this reason, we focus on the family associative classifiers given their easier explainability, namely decision trees [21], random forests [22], and XGBoost [23] in Python. 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 importance [24], which is calculated using the Gini criterion and then averaged across all trees within the model.

4.2.3. Biclustering

Given a set of observations (samples), $X = \{x_1, .., x_n\}$, genes $G = \{g_1, .., g_m\}$, a bicluster, B=(I, J), is a subspace defined by a subset genes, $J \subseteq G$, co-expressed on a subset of conditions, $I \subseteq X$. The *biclustering* task aims at identifying a set of biclusters, $\mathcal{B}$, such that each bicluster, B_k=(I_k, J_k), satisfies specific criteria of *homogeneity*, *dissimilarity* and *statistical significance*.

Homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster [25]. Merit functions are typically applied to guide the formation of biclusters in greedy and exhaustive searches. In stochastic approaches, a set of parameters that describe the biclustering solution are learned by optimizing a merit (likelihood) function. The pursued homogeneity determines the coherence (co-expression patterning), quality (noise tolerance) and structure (number, size and positioning) of the subspaces in the biclustering solution [7]. A putative regulatory module is in this context given by a subspace of co-expressed genes, i.e. expression pattern on observations (Fig. 5). A co-expressed subspace, B=(I, J), can be described by an *order-preserving coherence* when the ordering of a subset of genes according to their expression values, π_J, is preserved for each sample in I. In alternative, co-expression can be defined by *constant coherence* where expression values in a bicluster, $a_{ij} \in B$, are described by $a_{ij} = c_j + \eta_{ij}$, where c_j is the expected value of gene g_j and η_{ij} is the noise factor, generally a bounded deviation from expectations, $\eta_{ij} \in [-\delta/2, \delta/2]$.

In addition to homogeneity criteria, *dissimilarity* criteria can be further placed to guarantee the discovery of non-redundant biclusters [6]. Finally, *statistical significance* criteria guarantee that the probability of a bicluster's occurrence (against a null data model) deviates from expectations [26].

With biclustering algorithms, we can detect gene sets co-expressed on particular subsets of conditions, allowing for a more comprehensive modular view of regulatory responses to infection by SARS-CoV-2 and other viruses. In particular, when compared to the other proposed methods, biclustering allows for the detection of more specific patterns, such as gene groups 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.

To this end, we tested several biclustering algorithms to assess differences between solutions, namely the Cheng and Church [27], plaid [28] and xMotifs [29] algorithms. In recent years, a clearer understanding of the synergies between biclustering and pattern mining paved the rise of a new class of biclustering algorithms, generally referred to as *pattern-based biclustering* [7]. Pattern-based biclustering algorithms are inherently prepared to efficiently find exhaustive solutions of biclusters and offer the unprecedented possibility to affect their structure, coherency and quality [30]. This behavior explains why this class of biclustering algorithms are receiving an increasing attention in recent years [7]. In this context, we additionally assess the role of BicPAMS (Biclustering based on PAttern Mining Software), which consistently combines state-of-the-art contributions on pattern-based biclustering [6].

4.3. Functional enrichment and biological analysis

To study the biological processes associated with the gene groups found with the aforementioned methods, we used the EnrichR tool [31,32]. To assess enrichment of terms in the target knowledge bases, we focus on three major criteria: p-value from Fisher's exact test; the q-value, which adjusts the p-value to control the False Discovery Rate; and the z-score, which takes into account that Fisher's exact method to calculate the p-value produces lower values for longer lists even if they are random. Furthermore, the z-score and p-value are combined as follows: $c = \ln(p) \times z$. We prioritize both the adjusted p-value and the combined score to compare the results of the enrichment analysis.

Additionally, EnrichR web tool provides access to multiple knowledge bases.[3] For our analysis, we prioritize Gene Ontology (GO) Biological Process knowledge base [33,34] (ver. 2021) as it comprehensively characterizes a large amount of genes (14937) against 6036 terms, including recently augmented biological processes on viral infection and immune responses. Additionally, we use the Kyoto Encyclopedia of Genes and Genomes (KEGG) [35] to analyse enriched pathways and diseases. The identified terms are then analysed and compared to known characteristics of the infectious disease and other previous studies, in order to identify potential new insights into the effects of the virus and verify existing ones.

4.4. Code availability

The code used to obtain the results can be obtained in the following GitHub repository: https://github.com/PRodrigues98/Analysis-of-regulatory-response-to-SARS-CoV-2-infection. Dependencies: *Python* version 3.8, *NumPy*, *pandas*, *scikit-learn* and *matplotlib* libraries.

5. Results and discussion

To address the challenges of classic functional enrichment analyses, we introduced three approaches for identifying DEGs associated with modular regulatory views (section 4): clustering, predictive modeling and biclustering. In the present section, we present the key findings resulting from the application of these methods to study regulatory responses to viral infection, as well as an analysis of the identified biological processes within the context of SARS-CoV-2 infection.

To assess the effectiveness of the methods, we begin by presenting, in Table 2, the functional enrichment on the baseline set of genes obtained directly through gene selection using the Multi-Condition Setting (section 4), $p < 0.01$. As we can see in Table 2, there is a considerable number of processes with low p-value and c-scores are significantly lower when compared to the gene sets formed using machine learning methods (Tables 3–8). This is likely due to the higher number of genes being analysed together, an observation that supports the need for complementary methods, such as clustering, predictive modeling and biclustering, better suited to identify smaller subgroups of co-expressed genes. Additionally, terms such as *negative regulation of bone remodeling* (GO:0046851) and *negative regulation of bone resorption* (GO:0045779), less related to viral infection appear in this analysis, yet do not seem to reoccur within the terms found when using machine learning methods. Subsequent sections 5.1 to 5.3 assess the role of clustering, classification and biclustering searches to functional enrichment analysis, establishing a comparative appraisal.

[3] https://maayanlab.cloud/Enrichr.

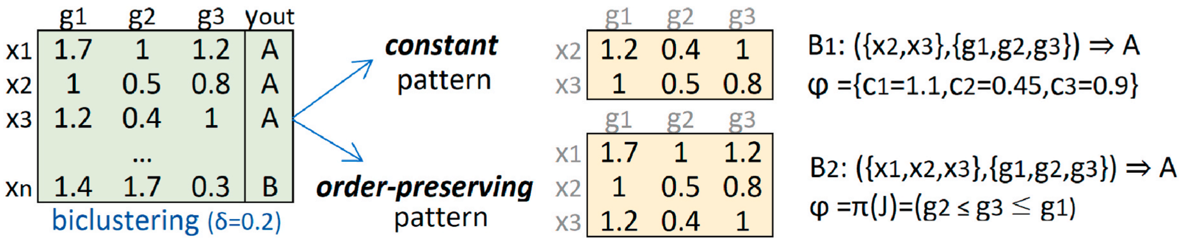

Fig. 5. Biclustering with constant and order-preserving coherence assumptions. The illustrative constant bicluster has pattern (value expectations) $\varphi_B = \{c_1 = 1.1, c_2 = 0.45, c_3 = 0.9\}$ on x_2 and x_3 observations, while the order-preserving bicluster satisfies the $\pi_J = (g_1 \geq g_3 \geq g_2)$ permutation on $\{x_1, x_2, x_3\}$ observations.

Table 2
Top 15 GO biological processes ordered by combined score after selecting differentially expressed genes using the Multi-Condition Setting ($p < 0.01$). GO IDs are linked to descriptions of the biological processes in the QuickGO web browser.

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

Table 3
Top 20 GO biological processes ordered by combined score (clustering applied over the multi-condition gene selection setting, $p < 0.01$). See GO IDs for links to biological processes in the QuickGo web browser.

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

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

5.1. Clustering analysis

Considering the application of agglomerative clustering (Ward linkage, Pearson correlation affinity) over differentially expressed genes obtained using the Multi-Condition Setting ($p < 0.01$), seven clusters of co-expressed genes were produced under the Elbow method, and all clusters subsequently subjected to functional enrichment analysis using the EnrichR API [31,32] and ordered by combined score. The gathered results from GO term enrichment, in Table 3, show that a high percentage of the top enriched processes are related to response to viral infection, as well as complementary immune-related 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 enriched processes) corresponds to a set of molecular signals associated with the detection of a virus (binding of viral RNA molecules to certain cytoplasmic receptors). 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 three child processes (GO:0039529 with $p = 2.91 \times 10^{-3}$ and $c = 905.29$; GO:0039535 with $p = 7.67 \times 10^{-4}$ and $c = 526.08$; GO:0039526 with $p = 5.26 \times 10^{-3}$ and $c = 513.51$) associated with this receptor which are statistically relevant. This receptor, along with others, has been identified as part of the inflammatory response to SARS-CoV-2 as well as other coronaviruses [36]. Additionally, the signaling cascade resulting from the detection of viral proteins is associated with the production of Type I interferons and pro-inflammatory cytokines [37], which can also be observed within the top enriched processes (for instance, terms GO:0060337, GO:0071357 and GO:0060333).

In order to assess cell-specific transcriptional responses, clustering was also performed separately per cell line after gene selection using the

Paired Setting (section 5). Appendix tables 12, 13, 14 and 15 list the top enriched GO terms from healthy *versus* infected expression for NHBE cells, A549 cells, A549 cells with added ACE2, and Calu3 cells, respectively. In particular, *type I interferon signaling pathway* (GO:0060337), which has several related terms also present within the top enriched processes (for instance, *type I interferon signaling pathway*, GO:0060337 and *cytokine-mediated signaling pathway*, GO:0019221, both direct ancestors), appear to be strongly associated to the process of viral infection, further bolstered by the presence of terms *response to interferon-beta* (GO:0035456) and *response to interferon-alpha* (GO:0035455).

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 [38]. The presence of these is noteworthy mainly due to directly opposing the other processes related to the activation of an immune response.

Considering normal bronchial epithelial (NHBE) cells (Table 12), there are multiple additional 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 adenocarcinomic alveolar basal (A549) cells (Table 13), the genes composing all detected processes show 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) reoccur, with most of the common processes linked to interferon and general cytokine responses as well as general responses to viral infection. Interestingly, and also 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 terms (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).

The terms *STAT cascade* (GO:0097696), *positive regulation of JAK-STAT cascade* (GO:0046427) and *JAK-STAT cascade* (GO:0007259) are also considerably enriched in A549 cells, although not significantly enriched in NHBE cells. These terms are related to the JAK-STAT signaling pathway, mediated by a wide variety of cytokines. Not triggering signaling or not regulating it properly, can lead to inflammatory disease [39], among other issues.

The addition of ACE2 to A549 cell cultures (Table 14) seems to increase the number of processes not directly related with viral infection. Nevertheless, top terms, including *positive regulation of heat generation* (GO:0031652), *regulation of fever generation* (GO:0031620) and *positive regulation of fever generation* (GO:0031622), all associated with acute inflammatory response (the term GO:0002526, which is an ancestor), underlie common physiological symptoms of COVID-19 (α-variant) and mediate immune responses.

Predominantly interferon and cytokine related processes are observed for Calu3 cells (Table 15), similarly to NHBE and A549 cells. The term *regulation of ribonuclease activity* (GO:0060700), also identified for NHBE cells (Table 12), is noteworthy as ribonuclease (RNase) is an enzyme that catalyzes the decomposition of RNA into smaller components. Particularly, RNase L is associated with innate immune response, and certain viruses have been shown to block this pathway for preventing viral RNA degradation [40].

Finally, in Table 4, we present the top enriched pathways from the KEGG knowledge base (ver. 2021). 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 enriched term, confirms the association to the target viral infection. 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. Additionally, the term *RIG-I-like receptor signaling pathway* (map04622), which is related to the previously mentioned RIG-I receptor, solidifies the relevance of its putative involvement with the anti-viral immune response.

5.2. Predictive modeling analysis

Fig. 6 presents a decision tree produced for the Multi-Condition Setting (Mann-Whitney *U* test with $p < 0.01$) using the Gini criterion. We notice the presence of selected genes related to immune and inflammatory responses, 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; *IL31*RA, a type I cytokine receptor. We observe that a few genes appear to be sufficient to discriminate conditions, including infection by different viruses. While in a classification problem this can be desirable, it does not support the comprehensive discovery of putative regulatory modules from transcriptomic data. To address this limitation, we now proceed with ensemble algorithms.

Random Forest and xGBoost algorithms are selected as ensemble predictive models. As with clustering, we first begin by presenting the processes identified when using the complete data, with multiple cell types and viruses. Tables 5 and 6 gather the top enriched terms for Random Forests and XGBoost, respectively. Using impurity-based

Table 4
Top 20 KEGG pathways ordered by combined score for A549 cells (paired healthy versus SARS-CoV-2 setting). See KEGG IDs for links to the pathway maps description.

KEGG Pathway	*p*-value	*c*-score
Measles (map05162)	1.02E-08	295.14
Influenza A (map05164)	1.02E-08	252.18
Herpes simplex virus 1 infection (map05168)	7.16E-10	177.20
Epstein-Barr virus infection (map05169)	4.82E-07	156.14
TNF signaling pathway (map04668)	1.11E-07	153.86
Coronavirus disease (map05171)	3.78E-07	150.32
RIG-I-like receptor signaling pathway (map04622)	3.68E-04	120.34
NOD-like receptor signaling pathway (map04621)	9.37E-06	119.63
Legionellosis (map05134)	1.77E-05	118.99
Hepatitis C (map05160)	1.79E-05	118.31
TNF signaling pathway (map04668)	8.61E-05	116.25
Primary immunodeficiency (map05340)	2.17E-03	116.00
Allograft rejection (map05330)	2.17E-03	116.00
Amoebiasis (map05146)	2.13E-06	111.38
African trypanosomiasis (map05143)	1.50E-04	111.14
Rheumatoid arthritis (map05323)	3.12E-06	110.93
Legionellosis (map05134)	1.12E-03	105.75
Antigen processing and presentation (map04612)	5.58E-04	100.66
JAK-STAT signaling pathway (map04630)	1.07E-06	96.92
NF-kappa B signaling pathway (map04064)	3.68E-04	92.76

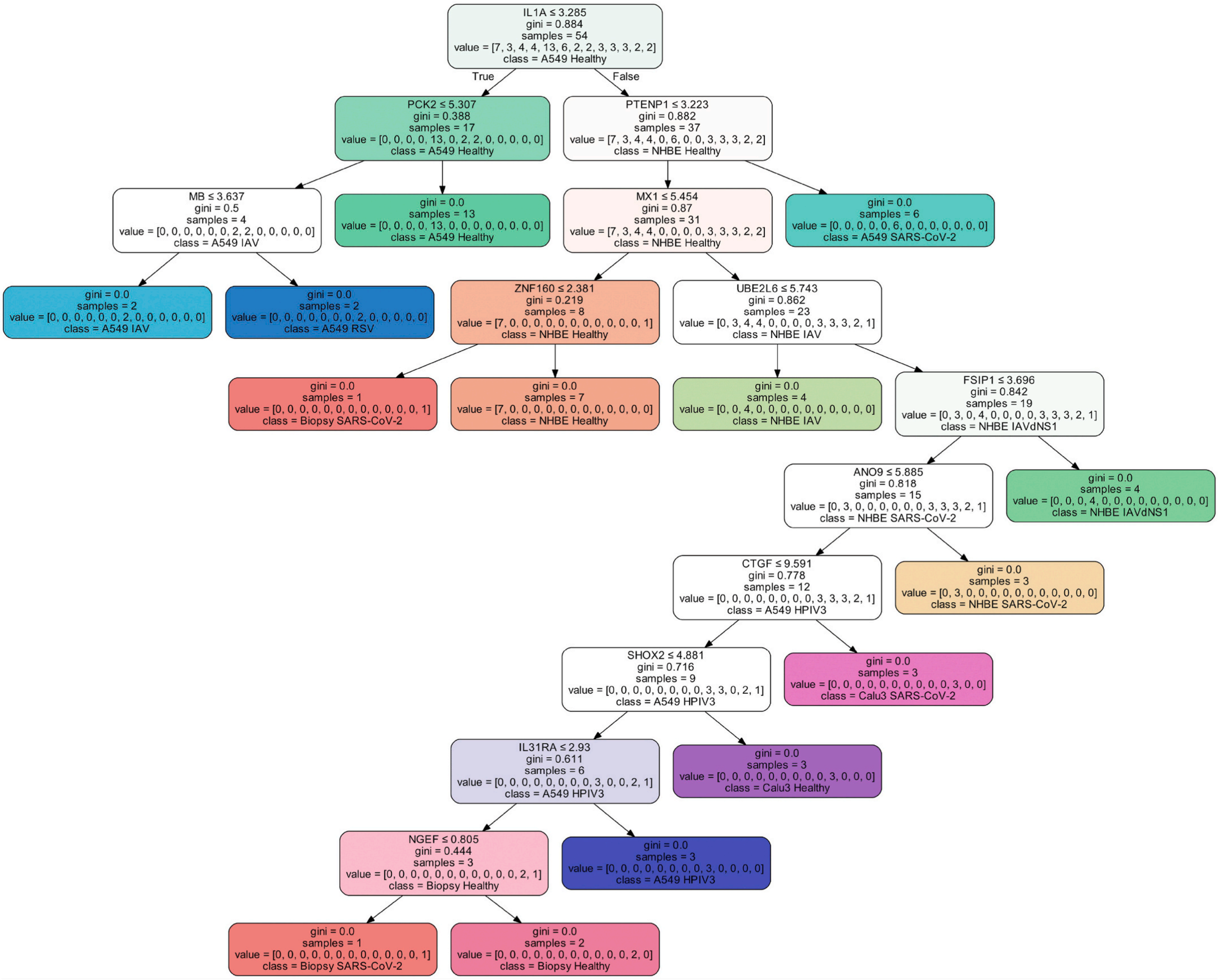
IL1A ≤ 3.285
gini = 0.884
samples = 54
value = [7, 3, 4, 4, 13, 6, 2, 2, 3, 3, 3, 2, 2]
class = A549 Healthy
True
False
PCK2 ≤ 5.307
gini = 0.388
samples = 17
value = [0, 0, 0, 0, 13, 0, 2, 2, 0, 0, 0, 0, 0]
class = A549 Healthy
PTENP1 ≤ 3.223
gini = 0.882
samples = 37
value = [7, 3, 4, 4, 0, 6, 0, 0, 3, 3, 3, 2, 2]
class = NHBE Healthy
MB ≤ 3.637
gini = 0.5
samples = 4
value = [0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0]
class = A549 IAV
gini = 0.0
samples = 13
value = [0, 0, 0, 0, 13, 0, 0, 0, 0, 0, 0, 0, 0]
class = A549 Healthy
MX1 ≤ 5.454
gini = 0.87
samples = 31
value = [7, 3, 4, 4, 0, 0, 0, 0, 3, 3, 3, 2, 2]
class = NHBE Healthy
gini = 0.0
samples = 6
value = [0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0]
class = A549 SARS-CoV-2
gini = 0.0
samples = 2
value = [0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0]
class = A549 IAV
gini = 0.0
samples = 2
value = [0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0]
class = A549 RSV
ZNF160 ≤ 2.381
gini = 0.219
samples = 8
value = [7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
class = NHBE Healthy
UBE2L6 ≤ 5.743
gini = 0.862
samples = 23
value = [0, 3, 4, 4, 0, 0, 0, 0, 3, 3, 3, 2, 1]
class = NHBE IAV
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
class = Biopsy SARS-CoV-2
gini = 0.0
samples = 7
value = [7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
class = NHBE Healthy
gini = 0.0
samples = 4
value = [0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
class = NHBE IAV
FSIP1 ≤ 3.696
gini = 0.842
samples = 19
value = [0, 3, 0, 4, 0, 0, 0, 0, 3, 3, 3, 2, 1]
class = NHBE IAVdNS1
ANO9 ≤ 5.885
gini = 0.818
samples = 15
value = [0, 3, 0, 0, 0, 0, 0, 0, 3, 3, 3, 2, 1]
class = NHBE SARS-CoV-2
gini = 0.0
samples = 4
value = [0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0]
class = NHBE IAVdNS1
CTGF ≤ 9.591
gini = 0.778
samples = 12
value = [0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 3, 2, 1]
class = A549 HPIV3
gini = 0.0
samples = 3
value = [0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
class = NHBE SARS-CoV-2
SHOX2 ≤ 4.881
gini = 0.716
samples = 9
value = [0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0, 2, 1]
class = A549 HPIV3
gini = 0.0
samples = 3
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0]
class = Calu3 SARS-CoV-2
IL31RA ≤ 2.93
gini = 0.611
samples = 6
value = [0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 2, 1]
class = A549 HPIV3
gini = 0.0
samples = 3
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0]
class = Calu3 Healthy
NGEF ≤ 0.805
gini = 0.444
samples = 3
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1]
class = Biopsy Healthy
gini = 0.0
samples = 3
value = [0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0]
class = A549 HPIV3
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
class = Biopsy SARS-CoV-2
gini = 0.0
samples = 2
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0]
class = Biopsy Healthy

Fig. 6. Decision Tree for Multi-Condition Setting ($p < 0.01$) using Gini criterion. Colors represent each class (condition), with a node's color corresponding to the combination of the colors of all associated conditions.

feature importance, 94 genes are identified with xGBoost and 356 genes with Random Forest. 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 algorithm.

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 [41].

Amongst the terms identified by XGBoost, multiple ones are related to chemotaxis, the movement of a cell or organism towards a higher or lower concentration of a given substance, and migration of various types of immune cells. In particular, macrophages [42,43] (GO:0048246 and GO:1905517), natural killer cells [44] (GO:2000501)), eosinophils [45] (GO:0072677 and GO:0048245), neutrophils [46] (GO:0030593 and GO:1990266), which are all types of white blood cells involved with the innate immune response to viral infection.

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

NHBE cell-specific terms are highlighted in Tables 16 and 17 in appendix for Random Forest and XGBoost, respectively. The *p*-values for these processes are significantly higher than those obtained using clustering on the same data. 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 with a pivotal role in the migration of monocytes [47].

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 cell death

Table 5
Top 20 GO biological processes ordered by combined score using 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, 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

Table 6
Top 20 GO biological processes ordered by combined score using 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

programming, may indicate that the cell was able to detect the infection by SARS-CoV-2. This hypothesis is further supported by the presence of the term *pattern recognition receptor signaling pathway* (GO:0002221). These receptors, as previously explained for the related term present in Table 3, have been associated with the inflammatory response to SARS-CoV-2 [36].

Considering A549 cells (Tables 18 and 19), we observe the presence of several terms related with the host response to the virus. 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) within the selected DEG by the Random Forest. 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 terms *RIG-I signaling pathway* (GO:0039529) and *cytoplasmic pattern recognition receptor signaling pathway in response to virus* (GO:0039528), also enriched for NHBE cells, are once more identified. 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 [37].

5.3. Biclustering analysis

With the aim of modeling more complex regulatory patterns to acquire novel knowledge, biclustering is now applied. In particular, biclustering, unlike clustering, can identify regulatory co-expression profiles spanning a subset of overall conditions. In addition, we can go beyond classic correlation assumptions, and accommodate less-trivial (yet relevant) forms of subspace coherence, such as additive and order-preserving expression [6].

We begin in Table 7 by presenting multiple statistics per algorithm when considering different preprocessing options. BicPAMS and Cheng and Church algorithms present the highest average number of biclusters, while Plaid and xMotifs algorithms provide significantly less for most preprocessing conditions. These differences are driven by the varying coherence, positioning constraints, and underlying searches (greedy in Cheng and Church and xMotifs, stochastic in Plaid, and exhaustive in BicPAMS). It is also important to note that BicPAMS presents delineatedly higher enrichments, and selects a larger amount of genes and lower number of conditions per putative regulatory module. Comparatively, this behavior is particularly relevant since having too many conditions can lead to the identification of more generic genes, while having too few genes can lead to the identification of less significant processes.

The observed differences are further hypothesized to be driven by four unique properties of the pattern-based biclustering searches implemented in BicPAMS. First, the exhaustive nature of the searches combined with the possibility to mask regions of the data space with greater likelihood in an attempt to find a more diversified set of non-redundant biclusters [6]. Second, the ability to consider varying levels of coherence strength and quality, allowing the discovery of regulatory modules with varying degrees of homogeneity [30]. Third, the ability to statistically test biclusters with varying coherence assumptions, ensuring deviations from expectations and therefore minimizing false positive discoveries [26]. Finally, the absence of structural constraints, enabling the discovery of an arbitrarily-high number of putative regulatory modules with flexible positioning [7], including overlapping genes and conditions.

Using these methods, we obtain a set of biclusters per algorithm, where each bicluster consists of a subset of genes and a subset of conditions. By performing functional enrichment on these genes, a set of biological processes is then produced (Tables 8–10). In order to analyse 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 the identification of the most commonly occurring processes, and thus provides a better view of which processes are

Table 7
Statistics for comparing the performance of the tested biclustering algorithms with different preprocessing techniques. Note: $|\mathcal{B}|$ corresponds to the number of biclusters; $\overline{|I|}$ is the average number of genes per bicluster; $\sigma_{|I|}$the standard deviation of genes per bicluster; $\overline{|J|}$ the average number of conditions per bicluster; $\sigma_{|J|}$ the standard deviation of the number of conditions per bicluster; and finally $\overline{\text{Terms}}$ the average number of enriched terms per bicluster.

Algorithm	Preprocessing	$\|\mathcal{B}\|$	$\overline{\|I\|}$	$\sigma_{\|I\|}$	$\overline{\|J\|}$	$\sigma_{\|J\|}$	$\overline{\text{Terms}}$
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

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.

We now proceed to a comparative analysis of the biological processes associated with SARS-CoV-2 for all cell types using biclustering (Table 8). 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.

Some of the identified processes have been analogously retrieved with clustering and predictive models. In particular, terms related to cytokine activity, for instance *cytokine-mediated signaling pathway* (GO:0019221), showing 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 given 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 [48,49], 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 [46]. 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 set of previously identified processes 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 avoiding this mechanism [50]. 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 SARS-CoV-2, RSV, HPIV3 and IAV viruses. In Table 9 we present the results from A549 cells, and in Table 10 from NHBE cells. We observe a considerable number of processes in common with the analysis provided in Table 8, which is to be expected, since most identified processes are related to immune response.

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

Table 11 offers a compilation of the number of GO biological processes detected for each of the applied machine learning approaches. As we can see, biclustering provides, by a considerable margin, a highest amount of biological processes, followed by clustering. The predictive models generally provided worse results, with Random Forests providing somewhat better results amongst predictors for the Multi-Condition Setting as well as for NHBE cells. Overall, these results provide initial empirical evidence in favor of pattern-based algorithms to promote the coverage and statistical significance of functional enrichment analysis, offering a way of unraveling less-trivial yet relevant regulatory behavior in knowledge bases.

6. Conclusion

This work assesses the impact of different modular views on

Table 8
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			Calu3 SARS-CoV-2			Biopsy SARS-CoV-2		
	count	p-value	c-score	count	p-value	c-score	count	p-value	c-score	count	p-value	c-score
cytokine-mediated signaling pathway (GO:0019221)	1.00	6.95E-3	6.06E+5	0.75	2.79E-4	6.06E+5	1.00	2.29E-3	1.02E+3	0.60	2.79E-4	6.06E+5
cellular response to interferon-gamma (GO:0071346)	0.66	1.15E-7	5.61E+2	0.88	4.31E-3	3.97E+2	0.83	2.36E-4	1.13E+3	0.56	1.37E-18	1.56E+3
cellular response to cytokine stimulus (GO:0071345)	0.61	2.50E-4	6.55E+5	0.92	2.50E-4	6.55E+5	0.67	4.14E-4	3.67E+2	0.44	2.50E-4	6.55E+5
inflammatory response (GO:0006954)	0.45	5.30E-5	6.46E+2	0.75	3.90E-4	1.47E+2	0.50	3.23E-5	2.30E+2	0.72	3.23E-23	3.84E+2
protein modification by small protein removal (GO:0070646)	0.49	7.88E-3	1.74E+2	0.46	7.88E-3	1.68E+2	0.67	6.62E-3	1.39E+2	0.63	2.74E-4	1.25E+2
regulation of immune response (GO:0050776)	0.34	3.64E-7	7.11E+2	0.71	5.14E-3	1.85E+2	0.69	3.32E-3	2.62E+2	0.72	2.84E-25	5.29E+2
mRNA splicing, via spliceosome (GO:0000398)	0.51	2.08E-6	7.38E+2	0.62	1.69E-4	6.89E+2	0.42	3.83E-5	3.18E+2	0.65	1.11E-5	5.68E+2
mRNA processing (GO:0006397)	0.51	2.88E-6	7.02E+2	0.62	3.98E-4	6.99E+2	0.42	3.83E-5	3.67E+2	0.65	1.27E-5	6.39E+2
DNA metabolic process (GO:0006259)	0.45	5.44E-4	2.63E+2	0.38	7.56E-4	4.22E+1	0.22	7.56E-4	1.40E+2	1.00	2.59E-6	1.97E+2
interferon-gamma-mediated signaling pathway (GO:0060333)	0.54	1.03E-3	6.97E+2	0.38	9.63E-4	6.97E+2	0.75	1.14E-6	1.49E+3	0.21	5.67E-8	1.28E+3
epidermis development (GO:0008544)	0.54	1.83E-4	1.44E+3	0.88	2.21E-22	2.15E+3	0.08	8.41E-6	5.33E+2	0.35	6.74E-4	6.99E+2
RNA splicing, with bulged adenosine as nucleophile (GO:0000377)	0.48	3.31E-6	7.02E+2	0.62	4.30E-4	6.75E+2	0.42	3.83E-5	3.36E+2	0.56	1.27E-5	5.35E+2
positive regulation of response to external stimulus (GO:0032103)	0.39	3.65E-3	5.26E+2	0.67	1.82E-3	1.36E+2	0.69	8.96E-3	3.29E+2	0.35	2.81E-10	3.26E+2
chemokine-mediated signaling pathway (GO:0070098)	0.39	6.02E-5	3.99E+3	0.67	1.55E-4	5.04E+2	0.67	1.16E-3	4.40E+2	0.37	1.19E-8	4.94E+2
DNA repair (GO:0006281)	0.45	4.48E-3	3.13E+2	0.25	4.65E-3	3.81E+1	0.22	4.65E-3	1.36E+2	0.95	9.43E-3	1.36E+2
extracellular matrix organization (GO:0030198)	0.32	1.13E-3	1.05E+2	1.00	3.31E-8	1.70E+2	0.28	1.81E-3	6.33E+1	0.44	2.96E-6	9.18E+1
neutrophil mediated immunity (GO:0002446)	0.53	3.68E-3	2.39E+2	0.38	2.14E-24	1.93E+2	0.67	2.16E-5	2.48E+2	0.28	1.57E-8	1.12E+2
cellular response to DNA damage stimulus (GO:0006974)	0.47	2.65E-3	1.29E+2	0.29	4.65E-3	1.10E+2	0.11	2.73E-17	1.42E+2	0.95	6.21E-5	1.42E+2
neutrophil activation involved in immune response (GO:0002283)	0.49	6.01E-21	2.39E+2	0.50	9.44E-3	1.93E+2	0.67	2.16E-5	2.47E+2	0.28	4.96E-9	8.95E+1
neutrophil degranulation (GO:0043312)	0.49	2.12E-21	2.46E+2	0.50	8.99E-3	2.00E+2	0.67	2.16E-5	2.55E+2	0.26	5.39E-8	9.08E+1
cellular response to chemokine (GO:1990869)	0.36	8.06E-5	3.65E+3	0.62	1.94E-4	4.57E+2	0.72	1.75E-3	3.99E+2	0.26	4.05E-8	4.45E+2
protein ubiquitination (GO:0016567)	0.45	9.45E-3	1.70E+2	0.29	9.45E-3	1.70E+2	0.19	9.45E-3	8.06E+1	0.81	9.45E-3	1.55E+2
cellular protein modification process (GO:0006464)	0.49	1.52E-19	1.39E+2	0.29	1.93E-3	1.20E+2	0.25	1.93E-3	8.82E+1	0.70	2.97E-6	1.19E+2
antigen receptor-mediated signaling pathway (GO:0050851)	0.49	6.48E-6	8.87E+1	0.67	3.95E-3	8.87E+1	0.14	3.73E-6	1.63E+2	0.26	4.15E-11	1.63E+2
defense response to symbiont (GO:0140546)	0.39	7.76E-6	9.11E+5	0.38	7.76E-6	9.11E+5	0.75	1.15E-10	1.66E+3	0.23	7.76E-6	9.11E+5

regulation for gene set enrichment analysis using transcriptional responses to SARS-CoV-2 infection, and further presents non-trivial biological processes associated with virus infection. Amongst state-of-the-art machine learning approaches, particular focus is placed on the pattern-centric views given the observed role of subspace clustering methods to improve the coverage and quality of enriched terms from knowledge bases.

A novel methodology is proposed, combining different computational approaches, which when consolidated provide a more robust view of the putative processes associated with virus infection. To guarantee the discriminative power of the pursued regulatory modules, 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, as well as between cells infected by different viruses. 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 – composed by clustering, associative predictive modeling and biclustering algorithms – is suggested to identify DEGs with correlated expression.

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

As directions for future work, we aim at: i) extending the conducted experimental analysis towards other transcriptomic data sources, in particular SARS-CoV-2 related sources, to cross-validate, expand and

Table 9
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.

GO Biological Processes	A549 SARS-CoV-2			A549 RSV			A549 HPIV3			A549 IAV		
	count	p-value	c-score	count	p-value	c-score	count	p-value	c-score	count	p-value	c-score
epidermis development (GO:0008544)	0.54	1.83E-4	1.44E+3	1.00	4.96E-6	7.01E+2	1.00	1.83E-4	1.44E+3	1.00	2.38E-14	9.94E+2
cellular response to interferon-gamma (GO:0071346)	0.66	1.15E-7	5.61E+2	0.85	1.15E-7	2.91E+2	0.92	1.15E-7	7.32E+2	0.86	4.74E-6	2.91E+2
cytokine-mediated signaling pathway (GO:0019221)	1.00	6.95E-3	6.06E+5	0.74	6.95E-3	9.57E+1	0.90	6.95E-3	4.40E+2	1.00	2.87E-10	4.36E+2
inflammatory response (GO:0006954)	0.45	5.30E-5	6.46E+2	0.97	1.23E-7	2.33E+2	0.92	5.30E-5	1.97E+2	1.00	3.90E-4	2.53E+2
interferon-gamma-mediated signaling pathway (GO:0060333)	0.54	1.03E-3	6.97E+2	0.74	1.03E-3	2.19E+2	0.92	1.03E-3	9.72E+2	0.61	8.73E-3	2.19E+2
antigen receptor-mediated signaling pathway (GO:0050851)	0.49	6.48E-6	8.87E+1	0.68	2.38E-4	8.87E+1	0.63	6.48E-6	8.87E+1	0.59	1.92E-3	8.87E+1
complement activation, classical pathway (GO:0006958)	0.39	6.59E-3	8.66E+3	0.91	4.52E-5	8.66E+3	0.86	6.59E-3	8.66E+3	0.78	4.54E-5	8.66E+3
skin development (GO:0043588)	0.46	9.85E-4	3.80E+2	0.71	8.82E-6	3.80E+2	0.59	4.62E-4	3.80E+2	0.57	9.85E-4	3.80E+2
cellular response to cytokine stimulus (GO:0071345)	0.61	2.50E-4	6.55E+5	0.41	5.74E-5	8.36E+1	0.55	5.74E-5	1.49E+2	0.71	5.74E-5	1.69E+2
chemokine-mediated signaling pathway (GO:0070098)	0.39	6.02E-5	3.99E+3	0.76	6.02E-5	5.04E+2	0.71	6.02E-5	8.65E+2	0.67	7.71E-3	5.04E+2
humoral immune response via immunoglobulin (GO:0002455)	0.37	8.06E-5	5.89E+3	0.91	8.06E-5	5.89E+3	0.82	8.06E-5	5.89E+3	0.76	5.90E-5	5.89E+3
positive regulation of external stimulus response (GO:0032103)	0.39	3.65E-3	5.26E+2	0.56	3.65E-3	1.43E+2	0.71	3.65E-3	1.95E+2	0.76	4.51E-3	1.47E+2
epidermal cell differentiation (GO:0009913)	0.36	9.67E-4	9.29E+2	0.88	1.21E-7	9.29E+2	0.71	8.76E-5	9.29E+2	0.76	9.67E-4	9.29E+2
keratinocyte differentiation (GO:0030216)	0.36	1.51E-3	1.29E+3	0.85	1.31E-6	1.29E+3	0.71	6.21E-4	1.29E+3	0.76	1.51E-3	1.29E+3
regulation of immune response (GO:0050776)	0.34	3.64E-7	7.11E+2	0.94	3.06E-9	7.11E+2	0.88	3.64E-7	2.29E+2	0.88	1.18E-8	7.11E+2
cellular response to chemokine (GO:1990869)	0.36	8.06E-5	3.65E+3	0.68	8.06E-5	4.57E+2	0.67	8.06E-5	7.92E+2	0.59	1.94E-4	4.57E+2
positive regulation of defense response (GO:0031349)	0.36	2.07E-3	8.24E+2	0.38	2.07E-3	1.36E+2	0.45	2.07E-3	2.17E+2	0.75	2.07E-3	1.37E+2
exogenous peptide antigen via MHC class II (GO:0019886)	0.48	6.36E-3	6.51E+2	0.41	6.36E-3	6.51E+2	0.43	6.36E-3	6.51E+2	0.29	9.02E-3	6.51E+2
peptide antigen via MHC class II (GO:0002495)	0.48	6.81E-3	6.34E+2	0.38	6.81E-3	6.34E+2	0.43	6.81E-3	6.34E+2	0.27	9.02E-3	6.34E+2
positive regulation of chemotaxis (GO:0050921)	0.31	6.20E-4	3.42E+2	0.88	3.37E-3	3.42E+2	0.80	6.20E-4	4.14E+2	0.82	6.71E-3	3.42E+2
extracellular matrix organization (GO:0030198)	0.32	1.13E-3	1.05E+2	0.53	1.13E-3	3.98E+1	0.59	1.13E-3	3.90E+1	0.57	1.13E-3	4.23E+1
T cell receptor signaling pathway (GO:0050852)	0.41	2.70E-3	1.16E+2	0.29	8.46E-4	5.79E+1	0.35	2.70E-3	6.64E+1	0.24	2.46E-3	3.68E+1
presentation of exogenous peptide antigen (GO:0002478)	0.45	7.80E-3	6.10E+2	0.26	7.80E-3	6.10E+2	0.31	7.80E-3	6.10E+2	0.16	9.02E-3	6.10E+2
positive regulation of protein phosphorylation (GO:0001934)	0.33	4.81E-3	5.19E+1	0.32	4.81E-3	5.19E+1	0.29	4.81E-3	5.19E+1	0.31	5.56E-3	5.19E+1
mRNA processing (GO:0006397)	0.51	2.88E-6	7.02E+2	0.18	2.34E-8	3.67E+2	0.29	3.83E-5	5.11E+2	0.20	2.88E-6	3.67E+2

improve the robustness of the provided findings; ii) assessing the validity of the methodological contributions over proteomic and metabolomic data; iii) addressing the issue of sample interdependence by testing the underlying relationships and designing pattern-centric approaches tailored to the presence of replicates; and iv) explicitly combining available background knowledge [51] to guide the enrichment analysis towards novel findings. In particular, and beyond virus infections, the proposed methodology can be straightforwardly extended towards the study of other pathologies as long as distinct phenotypes or morphological features of interest are present. Paradigmatic examples are cancer and cardiovascular disease analysis. As such, we believe that the proposed computational pipeline will prove to be a useful workflow for generating hypotheses across biomedical domains.

Funding

This work is supported by Fundação para a Ciência e Tecnologia (FCT) under IPOscore (DSAIPA/DS/0042/2018), LAETA (UIDB/50022/2020), and ILU (DSAIPA/DS/0111/2018). This work was further supported by LAQV (UIDB/50006/2020 and UIDP/50006/2020) and INESC-ID plurianual (UIDB/50021/2020). FCT distinction (CEECIND/01399/2017) to RSC under the CEEC Individual program is also acknowledged.

Declaration of competing interest

None declared.

Table 10
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			NHBE IAVdNS1		
	count	p-value	c-score	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	0.84	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	0.80	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	1.00	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	0.77	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	0.77	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	0.75	7.60E-04	1.13E+03
extracellular structure organization (GO:0043062)	0.75	1.10E-07	8.85E+01	0.86	1.10E-07	8.85E+01	0.70	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	0.70	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	0.82	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	0.66	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	0.66	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	0.66	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	0.68	4.01E-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	0.59	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	0.52	2.90E-03	2.98E+02
positive regulation of lymphocyte proliferation (GO:0050671)	0.54	3.72E-03	1.66E+02	0.70	3.72E-03	1.66E+02	0.70	8.75E-03	3.05E+02
complement activation, classical pathway (GO:0006958)	0.58	4.54E-05	8.66E+03	0.70	4.54E-05	9.67E+03	0.61	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	0.61	3.84E-03	3.56E+01
heart development (GO:0007507)	0.54	2.69E-03	7.04E+01	0.74	2.69E-03	7.04E+01	0.61	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	0.66	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	0.75	4.14E-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	0.59	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	0.61	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	0.61	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	0.61	6.85E-03	1.07E+02

Table 11
Overview of the number of processes found, for different *p* values and for each of the methods applied (Multi-Condition Setting).

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

Appendix A. Clustering: functional enrichment per cell line

Table 12
Top 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

(*continued on next page*)

Table 12 (*continued*)

GO Biological Process	***p***-value	***c***-score	Cluster
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

Table 13
Top 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

Table 14
Top 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

Table 15Top 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

Predictive modeling: enrichment per cell line

Table 16
Top statistically relevant GO biological processes ordered by combined score for NHBE cells using Random Forests.

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 (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 (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 (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 17
Top statistically relevant GO biological processes ordered by combined score for NHBE cells using 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

(*continued on next page*)

Table 17 (*continued*)

GO Biological Processes	***p*-value**	***c*-score**
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

Table 18
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 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 19
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

(*continued on next page*)

Table 19 (*continued*)

GO Biological Processes	***p***-value	***c***-score
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

References

[1] 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.

[2] D. Kim, J.-Y. Lee, J.-S. Yang, J.W. Kim, V.N. Kim, H. Chang, The architecture of sars-cov-2 transcriptome, Cell 181 (4) (2020) 914–921.

[3] X. Chen, E. Saccon, K.S. Appelberg, F. Mikaeloff, J.E. Rodriguez, B.S. Vinhas, T. Frisan, Á. Végvári, A. Mirazimi, U. Neogi, et al., Type-i interferon signatures in sars-cov-2 infected huh7 cells, Cell Death Discov. 7 (1) (2021) 1–15.

[4] T.A. Taz, K. Ahmed, B.K. Paul, F.A. Al-Zahrani, S.H. Mahmud, M.A. Moni, Identification of biomarkers and pathways for the sars-cov-2 infections that make complexities in pulmonary arterial hypertension patients, Briefings Bioinf. 22 (2) (2021) 1451–1465.

[5] Y.-H. Zhang, H. Li, T. Zeng, L. Chen, Z. Li, T. Huang, Y.-D. Cai, Identifying transcriptomic signatures and rules for sars-cov-2 infection, Front. Cell Dev. Biol. 8 (2021) 1763.

[6] R. Henriques, F.L. Ferreira, S.C. Madeira, Bicpams: software for biological data analysis with pattern-based biclustering, BMC Bioinf. 18 (1) (2017) 1–16.

[7] R. Henriques, C. Antunes, S.C. Madeira, A structured view on pattern mining-based biclustering, Pattern Recogn. 4 (12) (2015) 3941–3958.

[8] M. Frieman, R. Baric, Mechanisms of severe acute respiratory syndrome pathogenesis and innate immunomodulation, Microbiol. Mol. Biol. Rev. 72 (4) (2008) 672–685.

[9] K.G. Lokugamage, A. Hage, M. de Vries, A.M. Valero-Jimenez, C. Schindewolf, M. Dittmann, R. Rajsbaum, V.D. Menachery, Type i interferon susceptibility distinguishes sars-cov-2 from sars-cov, J. Virol. 94 (23) (2020) e01410–20.

[10] S.A. Ochsner, R.T. Pillich, N.J. McKenna, Consensus transcriptional regulatory networks of coronavirus-infected human cells, Sci. Data 7 (1) (2020) 1–20.

[11] 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), https://doi.org/10.1101/2020.05.05.079194.

[12] 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).

[13] 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, J. Am. Soc. Hematol. 136 (11) (2020) 1317–1329.

[14] 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).

[15] M.B. Brown, A.B. Forsythe, Robust tests for the equality of variances, J. Am. Stat. Assoc. 69 (346) (1974) 364–367.

[16] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (3/4) (1965) 591–611.

[17] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. (1947) 50–60.

[18] J.W. Tukey, Comparing individual means in the analysis of variance, Biometrics (1949) 99–114.

[19] J.M. Freudenberg, V.K. Joshi, Z. Hu, M. Medvedovic, Clean: clustering enrichment analysis, BMC Bioinf. 10 (1) (2009) 1–15.

[20] C. Mclean, X. He, I.T. Simpson, D.J. Armstrong, Improved functional enrichment analysis of biological networks using scalable modularity based clustering, J. Proteonomics Bioinf. 9 (1) (2016) 9–18.

[21] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[22] K. Fawagreh, M.M. Gaber, E. Elyan, Random forests: from early developments to recent advancements, Syst. Sci. Contr. Eng.: Open Access J. 2 (1) (2014) 602–609.

[23] T. Chen, 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.

[24] E. Scornet, Trees, Forests, and Impurity-Based Variable Importance, 2021, 04295 arXiv:2001.

[25] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE ACM Trans. Comput. Biol. Bioinf 1 (1) (2004) 24–45.

[26] R. Henriques, S.C. Madeira, Bsig: evaluating the statistical significance of biclustering solutions, Data Min. Knowl. Discov. 32 (1) (2018) 124–161.

[27] Y. Cheng, G.M. Church, Biclustering of expression data, Ismb 8 (2000) 93–103.

[28] L. Lazzeroni, A. Owen, Plaid models for gene expression data, Stat. Sin. (2002) 61–86.

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

[30] R. Henriques, S. Madeira, Bicpam: pattern-based biclustering for biomedical data analysis, Algorithm Mol. Biol. 9 (1) (2014) 27.

[31] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative html5 gene list enrichment analysis tool, BMC Bioinf. 14 (1) (2013) 1–14.

[32] 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 Res. 44 (W1) (2016) W90–W97.

[33] 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, Nat. Genet. 25 (1) (2000) 25–29.

[34] The gene ontology resource: enriching a gold mine, Nucleic Acids Res. 49 (D1) (2021) D325–D334.

[35] M. Kanehisa, S. Goto, Kegg: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.

[36] 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, Front. Immunol. 11 (2020) 1022.

[37] E. De Wit, N. Van Doremalen, D. Falzarano, V.J. Munster, Sars and mers: recent insights into emerging coronaviruses, Nat. Rev. Microbiol. 14 (8) (2016) 523–534.

[38] J. Melchjorsen, L.N. Sørensen, S.R. Paludan, Expression and function of chemokines during viral infections: from molecular mechanisms to in vivo function, J. Leukoc. Biol. 74 (3) (2003) 331–343.

[39] J.S. Rawlings, K.M. Rosler, D.A. Harrison, The jak/stat signaling pathway, J. Cell Sci. 117 (8) (2004) 1281–1283.

[40] L. Zhao, B.K. Jha, A. Wu, R. Elliott, J. Ziebuhr, A.E. Gorbalenya, R.H. Silverman, 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 11 (6) (2012) 607–616.

[41] Y.-C. Perng, D.J. Lenschow, Isg15 in antiviral immunity and beyond, Nat. Rev. Microbiol. 16 (7) (2018) 423–439.

[42] P.K. Pribul, J. Harker, B. Wang, H. Wang, J.S. Tregoning, J. Schwarze, P. J. Openshaw, Alveolar macrophages are a major determinant of early responses to viral lung infection but do not influence subsequent disease development, J. Virol. 82 (9) (2008) 4441–4448.

[43] C. Schneider, S.P. Nobs, A.K. Heer, M. Kurrer, G. Klinke, N. Van Rooijen, J. Vogel, M. Kopf, Alveolar macrophages are essential for protection from respiratory failure and associated morbidity following influenza virus infection, PLoS Pathog. 10 (4) (2014), e1004053.

[44] A.R. French, W.M. Yokoyama, Natural killer cells and viral infections, Curr. Opin. Immunol. 15 (1) (2003) 45–51.

[45] H.F. Rosenberg, K.D. Dyer, J.B. Domachowske, Eosinophils and their interactions with respiratory virus pathogens, Immunol. Res. 43 (1–3) (2009) 128–137.

[46] I.E. Galani, E. Andreakos, Neutrophils in viral infections: current concepts and caveats, J. Leukoc. Biol. 98 (4) (2015) 557–564.

[47] S.L. Deshmane, S. Kremlev, S. Amini, B.E. Sawaya, Monocyte chemoattractant protein-1 (mcp-1): an overview, J. Interferon Cytokine Res. 29 (6) (2009) 313–326.

[48] L. Glaser, P.J. Coulter, M. Shields, O. Touzelet, U.F. Power, L. Broadbent, Airway epithelial derived cytokines and chemokines and their role in the immune response to respiratory syncytial virus infection, Pathogens 8 (3) (2019) 106.

[49] B.A. Khalil, N.M. Elemam, A.A. Maghazachi, Chemokines and chemokine receptors during covid-19 infection, Comput. Struct. Biotechnol. J. 19 (2021) 976–988.

[50] G.C. Sen, Viruses and interferons, Annu. Rev. Microbiol. 55 (1) (2001) 255–281.

[51] R. Henriques, S. Madeira, Bic2pam: constraint-guided biclustering for biological data analysis with domain knowledge, Algorithm Mol. Biol. : Assoc. Méd. Bras. (São Paulo) (AMB) 11 (2016).