Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

CHAPTER TEN

Exploring COVID-19 pathogenesis on command-line: A bioinformatics pipeline for handling and integrating omics data

Janaina Macedo-da-Silva[a], João Victor Paccini Coutinho[a], Livia Rosa-Fernandes[a], Suely Kazue Nagahashi Marie[b], and Giuseppe Palmisano[a,c,*]

[a]GlycoProteomics Laboratory, Department of Parasitology, ICB, University of São Paulo, São Paulo, Brazil
[b]Cellular and Molecular Biology Laboratory (LIM 15), Neurology Department, Faculdade de Medicina FMUSP, Universidade de Sao Paulo, Sao Paulo, Brazil
[c]School of Natural Sciences, Macquarie University, Sydney, NSW, Australia
*Corresponding author: e-mail addresses: palmisano.gp@gmail.com; palmisano.gp@usp.br

Contents

Advances in Protein Chemistry and Structural Biology, Volume 131
ISSN 1876-1623
https://doi.org/10.1016/bs.apcsb.2022.04.002

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in late 2019 in Wuhan, China, and has proven to be highly pathogenic, making it a global public health threat. The immediate need to understand the mechanisms and impact of the virus made omics techniques stand out, as they can offer a holistic and comprehensive view of thousands of molecules in a single experiment. Mastering bioinformatics tools to process, analyze, integrate, and interpret omics data is a powerful knowledge to enrich results. We present a robust and open access computational pipeline for extracting information from quantitative proteomics and transcriptomics public data. We present the entire pipeline from raw data to differentially expressed genes. We explore processes and pathways related to mapped transcripts and proteins. A pipeline is presented to integrate and compare proteomics and transcriptomics data using also packages available in the Bioconductor and providing the codes used. Cholesterol metabolism, immune system activity, ECM, and proteasomal degradation pathways increased in infected patients. Leukocyte activation profile was overrepresented in both proteomics and transcriptomics data. Finally, we found a panel of proteins and transcripts regulated in the same direction in the lung transcriptome and plasma proteome that distinguish healthy and infected individuals. This panel of markers was confirmed in another cohort of patients, thus validating the robustness and functionality of the tools presented.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in late 2019 in Wuhan, China, and has proven to be highly pathogenic, making it a global public health threat (Mahalmani et al., 2020). There are a variety of manifestations, from asymptomatic to severe cases, that include symptoms such as muscle damage, prolonged tiredness, shortness of breath, loss of taste/smell, and severe pneumonia (Dixon et al., 2021; Nalbandian et al., 2021). Cumulative reports also associate post-acute effects of the infection, including pulmonary, hematological, cardiovascular, neuronal, renal, endocrine, and gastrointestinal sequelae (Hayes, Ingram, & Sculthorpe, 2021; Nalbandian et al., 2021). The quest to understand the infectious mechanisms of SARS-CoV-2, diagnostic alternatives, and treatments have been approached by several techniques (Das, Ahmed, Akhtar, Begum, & Banu, 2021). Among them, omics sciences generate large

amounts of information to enable holistic understanding of cell, tissue or organism function and reaction against a disease. These are high-performance techniques able to explore an organism at the level of genes (genomics), proteins (proteomics), metabolites (metabolomics), and lipids (lipidomics). Integration of this vast information provides a comprehensive and powerful tool for exploring the infection mechanism of SARS-CoV-2 (Overmyer et al., 2021; Wu et al., 2021).

In the last 10 years, with the advancement of technologies and processing capacity, proteomics and transcriptomics were readily implemented in several laboratories (Martens & Vizcaíno, 2017), while metabolomics and lipidomics were integrated along the years (Wenk, 2005). The more sophisticated data becomes, the greater the demand for new computational methods to deal with them. Currently, numerous tools can assess protein-protein interaction, calculate correlation between expression of different genes, identify enriched pathways and biological processes, and thus, drive data to find the most relevant points (Mangul et al., 2019; Mishra et al., 2021).

2. Applications

Omics sciences combined with bioinformatics technologies, in particular, contributed to the identification of therapeutic targets (Bojkova et al., 2020; Li et al., 2021); to the elucidation of virus pathophysiological mechanisms (Desai et al., 2020; Rosa-Fernandes et al., 2019; Wu et al., 2020); pathogen–host interactions (Terracciano et al., 2021); detection of patients predisposed to manifestations of severe symptoms (Lazari et al., 2021), selection of prognostic markers (Rosa-Fernandes et al., 2020; Terracciano et al., 2021) and to determine differential gene/protein/metabolite expression profiles (Desai et al., 2020; Leng et al., 2020; Wu et al., 2020). In addition to these fundamental applications, omics tools are also helpful in stratifying clinical manifestations in viral infections (Macedo-da-Silva et al., 2020). Furthermore, integrating data from different biological matrices may offer a comprehensive overview of disease pathogenesis, helping in prioritizing biomarkers and therapeutic targets (Wu et al., 2021; Zhu et al., 2022). Storing, analyzing, and interpreting these data can be a big challenge for researchers. However, it is a worthwhile task to overcome as omics approach enable answering multiple biological questions.

Notably the generated raw data in previous studies are cumulatively stored and accessible in public repositories, such as PRIDE (Vizcaíno et al., 2013) and SRA (Kodama et al., 2012), allowing reanalysis by different research groups. Therefore, mastering bioinformatics tools to process, analyze, integrate, and interpret these omic data is a powerful knowledge to build and validate hypotheses. The objective of this study is to present bioinformatics tools that can be used to reanalyze transcriptomic and proteomic data deposited in public databases and integrate them to increase the understanding of the pathophysiology of SARS-CoV-2.

3. Methodologies

3.1 Selected datasets

Four datasets, including proteomics and transcriptomics, were selected from public repositories. Shu et al. (PXD019106) evaluated plasma samples from patients infected with SARS-CoV-2, subdivided into fatal (FT, $n=5$), severe (SV, $n=7$), and moderate (MD, $n=10$) groups by proteomics. In addition, the study included samples from healthy patients (HE, $n=8$) who tested negative on throat swab and serological tests (Shu et al., 2020). Zhong et al. (S-BSST719) determined the plasma proteome profile of 50 individuals with mild and moderate disease. Collections were performed in the first 24 h (D0) after the positive PCR test and after 14 days (D14) with negative PCR result (Zhong et al., 2021). Kwan et al. (PRJNA692253) conducted RNA sequencing analyses of whole blood samples from 64 patients, of whom 45 tested positive for COVID-19 (INF) and 19 healthy participants not exposed to the virus (CTRL) (Kwan et al., 2021). Wu et al. (PRJNA646224) explored the lung and colon tissue transcriptome of patients who died due to COVID-19 (INF, $n=8$) and healthy counterparts were obtained from cancer patients who underwent surgical resection or biopsy (CTRL) (Wu et al., 2020).

3.2 Hardware features

The analyzes presented here were performed on an Ubuntu 20.04.3 LTS system (https://www.ubuntu.com/), with a total memory of 62GB, 40 CPUs, 2 threads per core, 10 cores per socket, and 2 sockets. The local storage of raw files required a large amount of space. However, the size of the files varied substantially as it depended on the quality of the data and equipment used.

3.3 Data download

In this protocol, we did not address the processing of raw files from proteomic approaches. The protein quantification tables provided in the selected manuscripts were downloaded and analyzed. The detailed pipelines for processing proteomic raw data are available in Kong, Leprevost, Avtonomov, Mellacheruvu, and Nesvizhskii (2017), Carvalho et al. (2016), Guangcan et al. (2021), and Tyanova, Temu, and Cox (2016).

For transcriptomic analysis, the fastq raw files were downloaded through the SRA Explorer platform (https://sra-explorer.info/) with the access numbers PRJNA692253 and PRJNA646224. Information regarding each file, such as SRA Accession, instrument, and total bases (Mb) were accessed. The SRA repository stores raw high-throughput sequencing data from different search fields. SRA files consist of raw data and require quality scores per basis for the submitted data. Tools like the sratoolkit download sequencing data from this database. Fastq files are raw data that have a definition line (defline) that contains a read identifier and nucleotide base information, all in text form.

After selection, all files from PRJNA692253 were downloaded by clicking on "Add to collection." A total of 64 SRA files were added. The platform offers the option of downloading .fastq files directly, but also .sra files and metadata. In the "FastQ Downloads" tab, the corresponding URL for all files could be downloaded. Kwan et al. used paired-end sequencing, resulting in 128 fastq files. Accession number PRJNA646224 provides the possibility to download colon and lung transcriptome data. Here, we downloaded only data referring to the lung (CTRL and INF), resulting in a total of 19 fastq files since Wu et al. used single-end sequencing. After creating a directory named "fastq_files" in the terminal console, files were downloaded using the "wget" command.

```
# Creating a directory named "fastq_files"
mkdir fastq_files

cd fastq_files


# Download the files

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR134/015/SRR13440815/SRR13440815_1.fastq.gz

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR134/015/SRR13440815/SRR13440815_2.fastq.gz
```

3.4 FastQC

FastQC is a highly recommended software for checking the quality of raw data coming from high-throughput sequencing pipelines. To download the software *via* the terminal console, "wget" command must be used.

The download is also possible by using directly the web browser from the Brabraham Institut website (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). After unzipping the file, turned the tool executable and added FastQC to the path. A directory to allocate FastQC outputs files was created and then the software was executed.

```
# FastQC software download
cd ..
wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.9.zip
unzip fastqc_v0.11.9.zip
chmod +x FastQC

# Creating a directory named "fastqc_output"
mkdir fastqc_output
# Run FastQC for all files in the "fastq_files" directory
fastqc -o fastqc_output fastq/*fastq.gz
```

At the end of the process, two output files will be created for each fastq file. By opening the generated .html file, the user can access the sample quality control report. The first module presents the results of basic statistics, including Total Sequences, Sequences flagged as poor quality, Sequence length, and %GC. Below is a BoxWhisker chart, which on the y-axis indicates the quality scores and on the x-axis the position in the read. The background indicates the quality of the sequences: green (very good), orange (fair), and red (poor). The "Per sequence quality scores" graph shows the average quality score on the x-axis and the number of sequences with that average on the y-axis. Generally, most samples' readings are expected to have a high-quality score. The parameter "Per base sequence content" reveals the proportion of each position of a base. If the difference between A and T or G and C bases is greater than 10%, the software will report a warning. If the difference is greater than 20%, this module will fail. In "Per sequence GC content," the GC content measurement is reported. A normal distribution profile is expected. In "Overrepresented Sequences" frequent sequences are reported, indicating that it has some biological relevance or that the library is not diverse or contaminated.

3.5 Trim Galore!

Trim Galore! is a platform that performs sample quality control by efficiently removing poor-quality parts of the readings and trimming the adapter.

This tool requires the previous installation of cutadapt and FastQC tools. The step of removing adapter sequences may or not be adopted in a transcriptome data analysis pipeline. In the selected datasets, Wu et al. chose to apply Trim Galore! while Kwan et al. started the sequence alignment right after running FastQC. In this pipeline, we applied the tool only to the dataset from Wu et al. study, based on the obtained FastQC report. The command used followed the structure for single-end sequencing data.

```
# Running Trim Galore
Command: trim_galore [options] <filename(s)>
trim_galore -j 2 -e 0.1 -q 20 --stringency 1 -O SRR12816734 SRR12816734.fastq.gz

# Running in a bash mode
#!/usr/bin/bash
SAMPLES="SRR12816734 SRR12816735"
for SAMPLE in $SAMPLES; do
trim_galore -j 2 -e 0.1 -q 20 --stringency 1 -O Trimmed {SAMPLE}.fastq.gz
done
```

In the above command, the "-j" option indicates the number of cores to be used for trimming; "-e" corresponds to the maximum allowed error rate; "-q" allows the trimming of low-quality ends from reads in addition to adapter removal; "--stringency" points out the overlap with adapter sequence required to trim a sequence; the "-O" option creates a directory to save the output files. The command "-a" is used to add the adapter sequence. The Trim Galore! can automatically detect whether the Illumina universal, Nextera transposase, or Illumina small RNA adapter sequence was used. However, it is possible to add the adapter sequence manually. When the command is missing, the software auto-detects the adapter.

3.6 Reference genome

Before starting the sequence alignment, it is necessary to download the reference genome. As indicated in the original manuscripts, Wu et al. used the h19 genome and Kwan et al. used the GRCh38 genome. The downloaded file will be allocated in the created directory. Then the HISAT2 sequence alignment software will be downloaded, unzipped, and added to the path. The genome will be indexed using the "hisat2-build" script, allowing HISAT2 to perform the read alignment faster in the next step. At the end of the indexing, eight .ht2 files will be created.

```
# Creating a directory and downloading the reference genome

mkdir reference_genome
cd reference_genome
#Kwan et. al
wget ftp://ftp.ensembl.org/pub/release-
86/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz
gunzip Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz
#or
#Wu et. al
wget https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz
gunzip hg19.fa.gz
cd ..

# Downloading the HISAT2 software and building the reference genome

mkdir hisat2
cd hisat2
wget https://cloud.biohpc.swmed.edu/index.php/s/oTtGWbWjaxsQ2Ho/download
unzip hisat2-2.2.1-Linux_x86_64.zip
cd ..
mkdir index
cd index
Command: hisat2-build [options]* <reference_in> <ht2_index_base>
hisat2-build -p 4 /home/reference_genome/hg19.fa hg19_index
#or
hisat2-build -p 4 /home/reference_genome/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa GRCh38_index
```

3.7 Alignment of raw readings

Sequence alignment was performed with HISAT2 software (Kim, Paggi, Park, Bennett, & Salzberg, 2019). Other programs are available for free, such as STAR (Dobin et al., 2013), Bowtie2 (Langmead & Salzberg, 2012), and TopHat2 (Kim et al., 2013). A comparison between the methods was recently reported and can be accessed at Schaarschmidt, Fischer, Zuther, and Hincha (2020) and Musich, Cadle-Davidson, and Osier (2021).

```
# For single-end alignment
Command: hisat2  -x genome_index -U input_file.fq -S output_name.sam --no-spliced-alignment
# For paired-end alignment
Command: hisat2 -x reference_genome -1 input_file_1.fq -2 input_file_2.fq -S output_name.sam

cd ..
mkdir aligned
#!/usr/bin/bash
SAMPLES="SRR12816730_trimmed SRR12816729_trimmed SRR12816728_trimmed SRR12816727_trimmed SRR12816726_trimmed
SRR12816725_trimmed SRR12816735_trimmed SRR12816732_trimmed SRR12816724_trimmed SRR12816723_trimmed
SRR12816722_trimmed SRR12816721_trimmed SRR12816736_trimmed SRR12816720_trimmed SRR12816719_trimmed
SRR12816718_trimmed SRR12816731_trimmed SRR12816733_trimmed SRR12816734_trimmed"
for SAMPLE in $SAMPLES; do
hisat2 -p 6  -x /home/index/hg19_index -U /home/Trimmed/${SAMPLE}.fq.gz -S /home/aligned/${SAMPLE}.sam
done
#or
hisat2 -p 8 -x /home/index/GRCh38_index -1 /home/fastq/{SAMPLE}_1.fastq.gz -2 /home/fastq/{SAMPLE}_2.fastq.gz -S {SAMPLE}.sam
done
```

The "-p" command indicates the number of threads to perform the alignment; "-x" is used to enter the path of previously generated reference

files. It is important not to add the file extension (e.g., .ht2). The "-U" option is used to show the path of unpaired files. To perform paired-end sequencing alignment, the synthesis used is "-1" and "-2" to assign files _1 and _2. The "-S" command designates the .sam file that will be generated. SAM files store aligned sequences and take up much space, so these files can be converted to compressed version .BAM by the samtools software.

3.8 Samtools to convert SAM to BAM files

Samtools is a set of programs for interacting with high-throughput sequencing data. It is helpful for converting SAM, BAM and CRAM files. One of the most used commands is the "samtools view," which takes .BAM/.SAM files as input and converts them to .SAM/.BAM, respectively. The "-S" and "-b" commands are used. The alignment of fastq files occurs in random order with the position in the reference genome. Therefore, in "samtools sort," the BAM files sorting is performed. The "-o" command indicates the output file.

```
# Running samtools

samtools view -@ 4 -S -b SRR12816736_trimmed.sam > SRR12816736_trimmed.bam
samtools sort -@ 4 SRR12816736_trimmed.bam -o SRR12816736_trimmed.sorted.bam
```

3.9 FeatureCounts

FeatureCounts is a platform that performs reading counting, also called reading summarization, by gene (Liao, Smyth, & Shi, 2014). The htseq-count tool (Anders, Pyl, & Huber, 2015), and featureCounts, are widely used for this purpose. Both share the same file input format (BAM or SAM) and need an annotation file that includes the chromosomal coordinates of features. Among the mandatory arguments required to run the software are "-a" and "-o," which specifies the name/path of an annotation file and the name of the output file including read counts. Among the optional arguments, we used the "-T" command to indicate the number of threads to be used in the analysis; the "-t" option specifies the feature type in GTF annotation, and "-g" specifies the attribute type in the GTF annotation.

```
# Running featureCounts

featureCounts -T 4 -t exon -g gene_id -a Homo_sapiens.GRCh38.86.gtf -o SRR12816736.txt SRR12816736_trimmed.sorted.bam
```

3.10 Exploring data with R packages

After obtaining the read counts for each sample and downloading the quantitative proteomics tables, a set of R packages were used to handle

the data and identify the relevant biological findings. The packages used are: limma, Glimma, edgeR, Homo.sapiens, metboAnalystR, Enhancedvolcano, combiroc, pROC, clusterProfiler, flashClust, uwot, NbClust, richplot, ggplot2, WCGNA, gprofiler2, ggvenn, DOSE, wordcloud, FactoMineR, factoextra, ggstatsplot, among others. The codes used to generate the data and figures will be available in Supplementary File 1 in the online version at https://doi.org/10.1016/bs.apcsb.2022.04.002.

The gene expression count files for each sample mapped by the transcriptomics pipeline were gathered in a single directory and, with the "readDGE" function available in the edgeR package, they were merged into a single object DGEList class. The raw counts were transformed into counts per million (CPM) so that the size of the libraries was taken into account and then scaled to log (log CPM). Most samples genes that did not have sufficiently large counts were removed with the "filterByExpr" function. The resulting matrix was normalized by a trimmed mean of M-values (TMM) (Law et al., 2016).

The differentially expressed genes and proteins were determined by applying the standard analysis pipeline of the limma package, with *p*-value correction by the Benjamini-Hochberg method (q-value). The Homo.sapiens package performed the transformation of ENSMBL ID into Gene symbol. Interactive MAplots were plotted with the Glimma package. The table resulting from the analysis of differentially regulated genes/proteins was used as input to visualize of volcano plots by Enhancedvolcano package. The x and y values were the columns logFC and adjusted *p*-value, respectively. Principal component analysis (PCA) was conducted by applying the FactoMineR package using the "PCA" function. The "fviz_pca_ind()" function, available in the factoextra package, was applied to visualize the result. Samples were identified as outliers using the MetaboAnalystR package, with the "PlotRF.Outlier" function. The input data used were filtered and normalized matrices. Venn diagrams were built with the ggvenn package. The gprofiler2 package was used to perform gene ontology (GO). The "gost" function was applied to access functional enrichments of gene lists by applying a threshold of q-value < 0.05 and "gostplot" to visualize the results, with the capped and interactive options set to TRUE. Only ontologies with 5 or more proteins/genes were considered. The wordcloud package was used to build the word cloud based on the name frequency of ontologies identified by gprofiler2. Bubble plots were built with the ggplot package. The x-axis shows the gene count in each ontology shown on the y-axis. The size of the bubbles indicates the -log of the q-value and the color indicates the ontology category. To access pathways related to differentially

expressed genes/proteins, the clusterProfiler and enrichplot packages were used. The "GSA" function with the options minGSSize set to 5 and maxGSSize to 100. The p-value correction method was indicated as "BH" (Benjamini-Hochberg). The regulated pathways network was built with the "cnetplot" function.

Data integration was performed by applying the "merge" function to select Gene symbols common between lung transcriptomics and plasma proteomics normalized tables. The "boxplot" function was applied to verify data distribution. Then the "removeBatchEffect" function, available in limma, was applied. The data were then normalized by the "scale" function (z-score). The "umap" function, available in the uwot package, was used to produce a low-dimensional embedding that summarizes the overall structure of high-dimensional data. The parameters adopted were min_dist-=0.01, n_neighbors=20, n_epochs=10,000, verbose=2 and spread=2. The NbClust package was applied to determine the best number of clusters. Then the data were clustered by the "eclust" function. The cluster number was set to k=8. The "silhouette" function, available in the cluster package, was applied to assess the clustering quality. Genes and proteins from the same cluster, which have logFC $> |1|$ (same direction) and q-value < 0.05 were selected and applied to construct ROC curves based on data from other datasets. The WGCNA package was applied to the expression data from Wu et al. and Kwan et al. The intersection of the ENSEMBL IDs mapped in the two datasets was performed by the "intersect" function. Correlation and connectivity between two datasets were determined using the softPower parameter set to 5. The number of modules was set to 4 and minClusterSize to 30. The "combiroc" package was used to determine the best marker combinations, and the "pROC" package was applied to plot and build the ROC curves.

4. Results

Given the increasing availability of public omics data, we present a robust and open access pipeline for extracting insights from quantitative biomolecular data (Fig. 1). Initially, raw files generated by RNA sequencing were downloaded and processed. The proteomics and transcriptomics data were then analyzed to find differently regulated features between the groups. In addition, outlier samples, biological processes, cellular components, molecular functions, and pathways of interest were identified. Transcriptomics data from different matrices were correlated, and proteomics

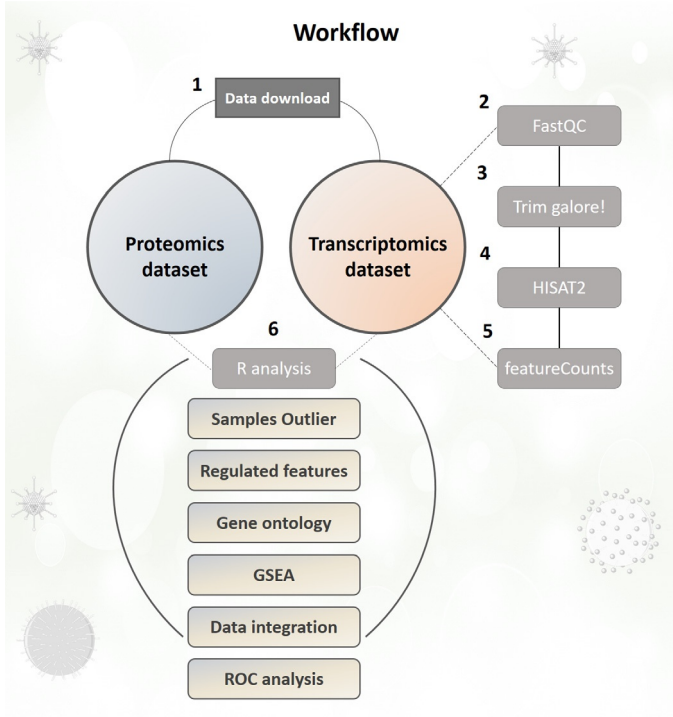

Fig. 1 Computational workflow and tools adopted to explore quantitative proteomics and transcriptomics data from plasma, whole blood, and lung of healthy individuals and patients infected with SARS-CoV-2.

and transcriptomics data from different biological samples were fitted and integrated. Finally, we evaluated through ROC curves the ability to separate samples from infected and healthy individuals based on the panel of relevant proteins and transcripts. Proteins and transcripts with better resolution between the clinical conditions were prioritized and discussed.

4.1 Plasma proteome presents different patterns according to the patient's COVID-19 grade

After filtering the original data provided by the authors, the PCA plot demonstrated that the plasma proteome of infected patients presented a different pattern than that of healthy patients (HE). The groups SV (severe) and FT (fatal) showed a very similar pattern, which differed from the MD (moderate) group (Fig. 2A). The protein profile of plasma from the same patient collected on day 0 (D0) of infection (PCR positive) and day

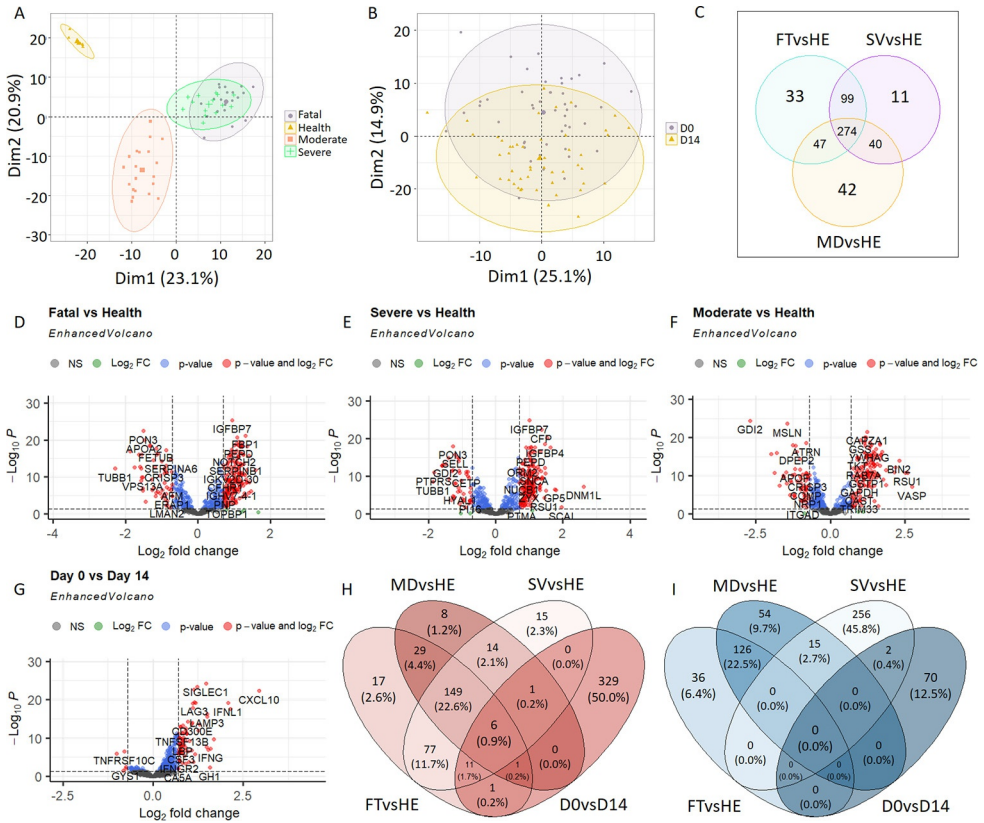

Fig. 2 Exploring the plasma proteome alterations resulting from SARS-CoV-2 infection. (A) Principal component analysis (PCA) indicating the separation between the groups that developed fatal (FT), severe (SV), moderate (MD) COVID-19 disease, and healthy donors (HE); (B) PCA indicating the separation between the same patients, with samples collected on day 0 of infection (D0) and day 14 (D14); (C) Venn diagram showing differently regulated proteins between FTvsHE, SVsvHE and MDvsHE comparisons; (D–G) Volcanos plot of differentially expressed proteins among FTvcHE, SVvsHE, MDvsHE, and D0vsD14 conditions; (H) Venn diagram showing upregulated proteins in the SASRS-CoV-2 infected group in both evaluated studies and (I) Venn diagram showing downregulated proteins in the SASRS-CoV-2 infected group in both evaluated studies.

14 (D14) (PCR negative) showed similarity in the majority of cases, however few presented distinct recovery pattern (Fig. 2B).

A total of 453 proteins were differentially regulated in the comparison between FT and HE (Fig. 2D); 424 between SV and HE (Fig. 2E), and 403 between MD and HE (Fig. 2F). Among patients in groups D0 and D14, 421 regulated proteins were identified (Fig. 2G). Comparing the proteins upregulated and downregulated in all comparisons (Fig. 2H and I), 6 proteins (C1QA, CD93, FLT4, INHBC, ENPP2, and NID1) are identified upregulated in the infected groups, indicating a possible role in the SARS-CoV-2 pathology (Fig. 3).

The boxplot plots (Fig. 3) showed that in the three degrees of infectious disease severity (FT, SV, MD) the expression level of the six proteins were higher than the healthy group (HE). The intra-patient comparisons (Fig. 3B) showed that after 14 days (D14) of disease onset (negative PCR), the levels of these proteins were lower than at the time of diagnosis (D0).

4.2 Plasma proteome analysis indicates dysregulation of immune system and cholesterol metabolism

Enriched GO terms were associated to the dysregulation of immune system, proteasome, cytoskeleton, cell adhesion, and lipoprotein metabolism (Fig. 4A). Upregulated proteins in the infected groups (FT, SV, MD, and D0) are related to the activation of the immune system, especially with the response mediated by leukocytes. Cytoskeleton dynamics and proteasomal degradation were also upregulated (Fig. 4B). Processes related to lipid metabolism and cell adhesion and migration are downregulated (Fig. 4C). Pathway analysis was performed using differentially regulated proteins and confirmed findings from GO analysis (Fig. 4D and E). In fact, downregulated proteins in the infected groups (FT, SV, MD, and D0) are related to cholesterol and lipoprotein metabolism. The ECM-proteoglycans pathway and cell adhesion molecules were also associated with downregulated features. On the other hand, the pathways of infection, immune system, apoptosis, and extracellular matrix (ECM) organization are upregulated.

4.3 Lung and whole blood transcriptome reinforce findings in the plasma proteome

PCA analysis based on blood transcriptome from CTRL and INF patients with different degrees of symptoms did not show a clear separation (Fig. 5A). On the other hand, analysis of lung tissue from patients with fatal

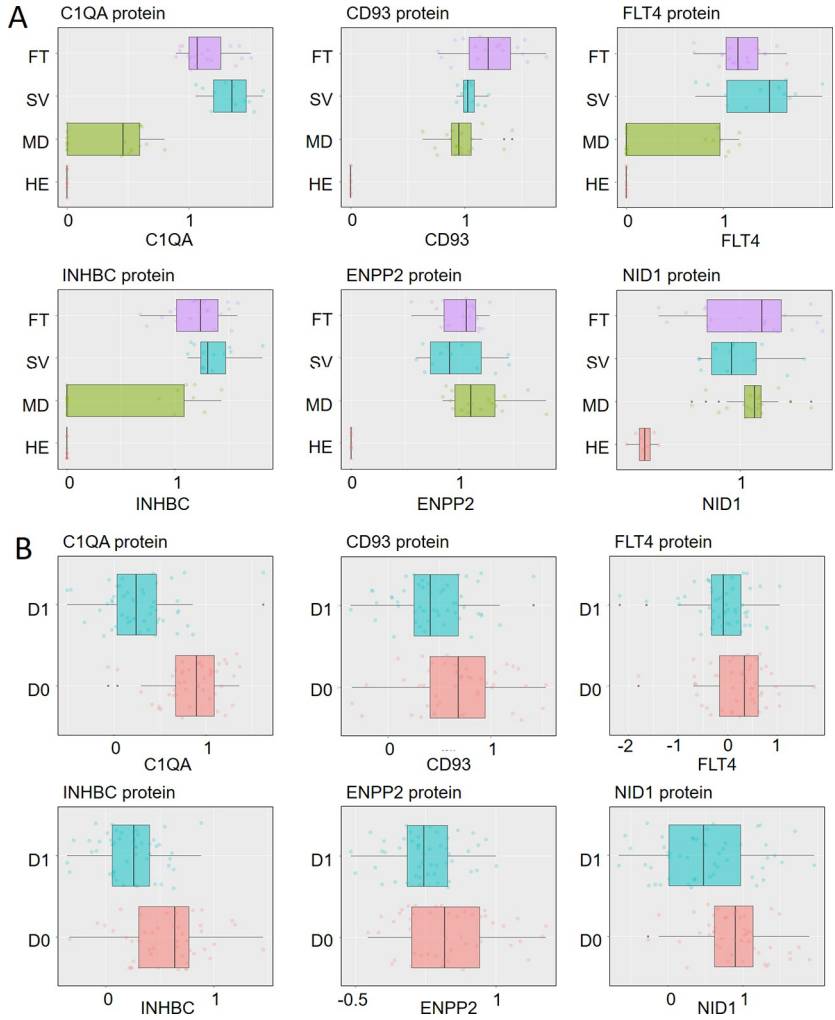

Fig. 3 Boxplot of upregulated proteins in SARS-CoV-2 infection. (A) Upregulated proteins in the groups that developed fatal (FT), severe (SV), and moderate (MD) COVID-19; (B) Upregulated proteins at the onset of COVID-19 (D0).

COVID-19 (INF) and healthy counterparts (CTRL) showed separation between groups (Fig. 5B). A total of 1039 genes were identified differentially expressed in the blood transcriptome (Fig. 5C**)** and 1034 in the lung tissue analysis (Fig. 5D) (logFc $|1|$ and q-value < 0.05). A higher number of enriched GO terms (q value < 0.05) were found in the lung tissue compared to the blood transcriptome giving more information about the COVID-19 pathophysiology. Immune response mediated by leukocytes, apoptotic

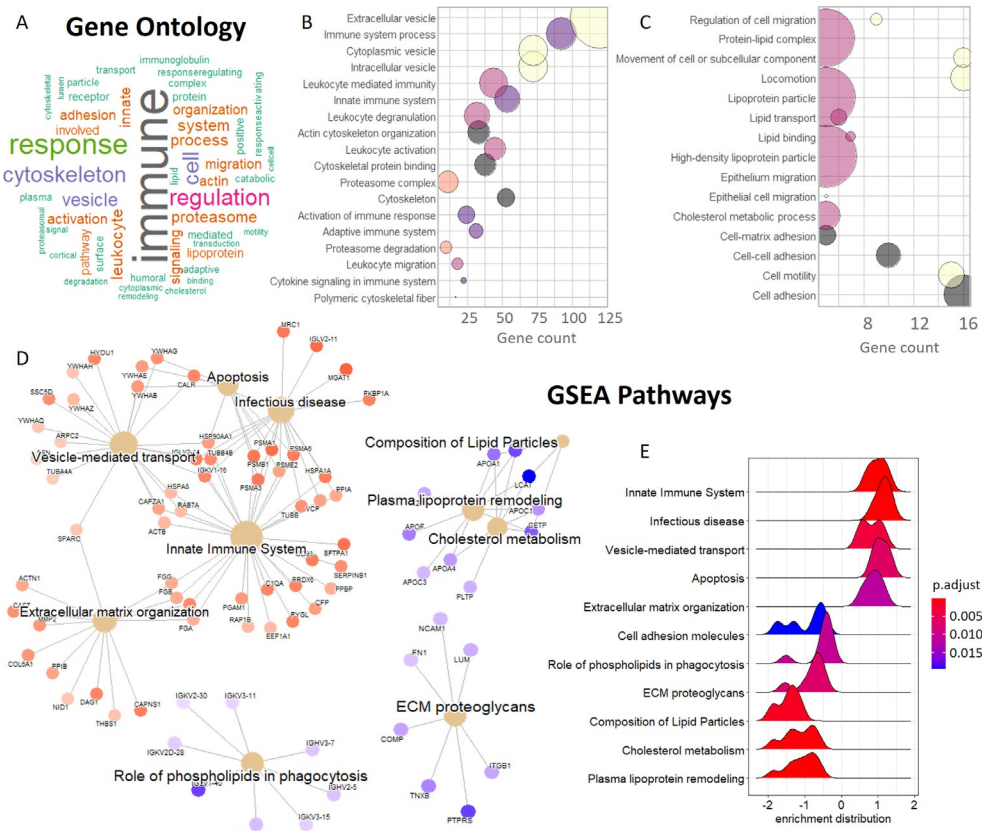

Fig. 4 Gene ontology and GSEA analysis of differentially regulated plasma proteins. (A) Word cloud generated from differentially regulated identified terms (gene count > 5 and q-value < 0.05); (B–C) Biological processes, cellular components, and molecular functions of upregulated and downregulated proteins in infected groups, respectively; (D) Network of proteins that participate in the main enriched pathways. Red and blue dots indicate upregulated and downregulated proteins, respectively; (E) Pathways associated with differentially regulated proteins.

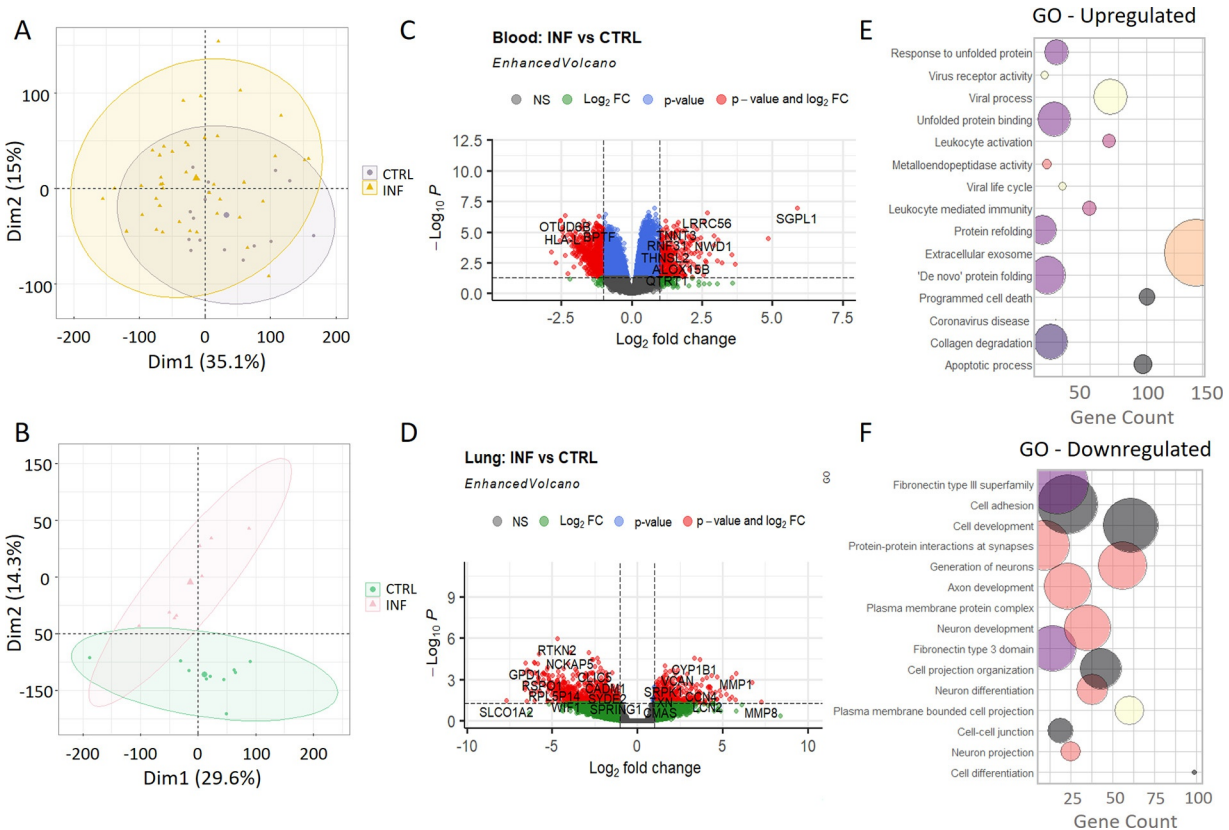

Fig. 5 Alteration of blood and lung transcriptome due to SARS-CoV-2 infection. (A) Principal component analysis (PCA) based on blood transcriptome from healthy (CTRL) and infected patients (INF); (B) PCA showing lung transcriptome-based separation of groups that developed fatal COVID-19 (INF) and non-infected individuals (CTRL); (C) Volcano plot indicating differentially expressed genes identified in blood; (D) Volcano plot indicating differentially expressed genes identified in the lung; Gene ontology (GO) analysis of upregulated (E) and downregulated (F) genes in the INF (blood and lung) groups.

processes, collagen degradation, and viral life cycle were among the upregulated enriched GO terms corroborating the findings of plasma proteomics. On the other hand, there was an enrichment associated with misfolded proteins that were not identified in plasma (Fig. 5E). Ontologies related to downregulated genes showed greater differences with respect to proteomic findings, although processes such as cell adhesion are common to both. Terms related to cell development and differentiation, especially of the brain components, are highlighted (Fig. 5F). Processes related to lipid and cholesterol metabolism were not identified.

4.4 The integration of omics data identifies clusters of proteins and genes related to key processes in infection

Weighted Gene Coexpression Network Analysis (WGCNA) is a popular method applied to identify correlation patterns between genes in samples. Integration of the 5000 common genes between lung and blood transcriptome did not show high modulus preservation between studies (Fig. 6A). However, the green (Fig. 6B) and turquoise (Fig. 6C) modules have $5 < Z < 10$, indicating moderate preservation. By selecting and working with the 400 key genes (labeled on the basis of kME) differentially expressed in the assessed groups of the turquoise and green modules, we identified enriched biological processes and pathways. The green module has genes related to leukocyte activation and proteasome activity, which was also identified in the turquoise module (Fig. 6D). However, it was also possible to verify processes linked to the virus, including proteins associated with spike viral glycoprotein maturation and deregulation of host-glycosylation (Fig. 6E), a key process during infection.

The clustering of 421 features common to lung transcriptome and plasma proteome data from healthy (CTRL and HE) and fatal (INF and FT) COVID-19 patients resulted in 8 clusters (Fig. 7A and B). The silhouette plot indicates genes possibly associated with the wrong cluster ($n = 27$) (Fig. 7C). These genes were disregarded in the enrichment analyses. Clusters 3, 4, 6, 7, and 8 show the highest correlation between the LogFC of the proteome and transcriptome (Fig. 7D). Considering only genes regulated in the same direction in both matrices, we identified enriched biological pathways and processes. Cluster 1 highlights terms associated with ECM; cluster 2, terms linked to the immune system and enzymatic activity; cluster 3, cytoskeletal activities; cluster 4, lipid metabolism; cluster 5, immune system and leukocyte activation; cluster 6,

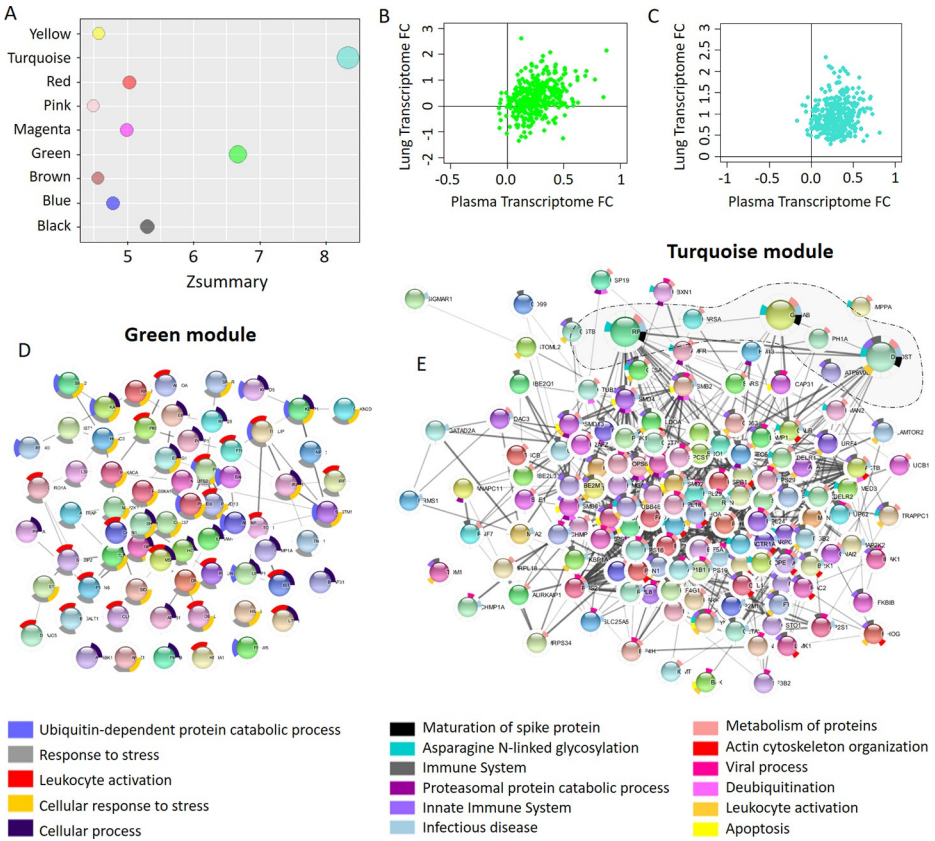

Fig. 6 Weighted gene co-expression network analysis (WGCNA). (A) *Z*-score for modules preserved between lung transcriptome and whole blood; (B–C) Scatter plot indicating the LogFC of the 400 genes comprising the green and turquoise modules ($Z > 5$), respectively, selected by the topGenesKME function; (D–E) Network indicating differentially expressed proteins from the turquoise and green modules. The donut plot indicates enriched processes and pathways (q-value < 0.05).

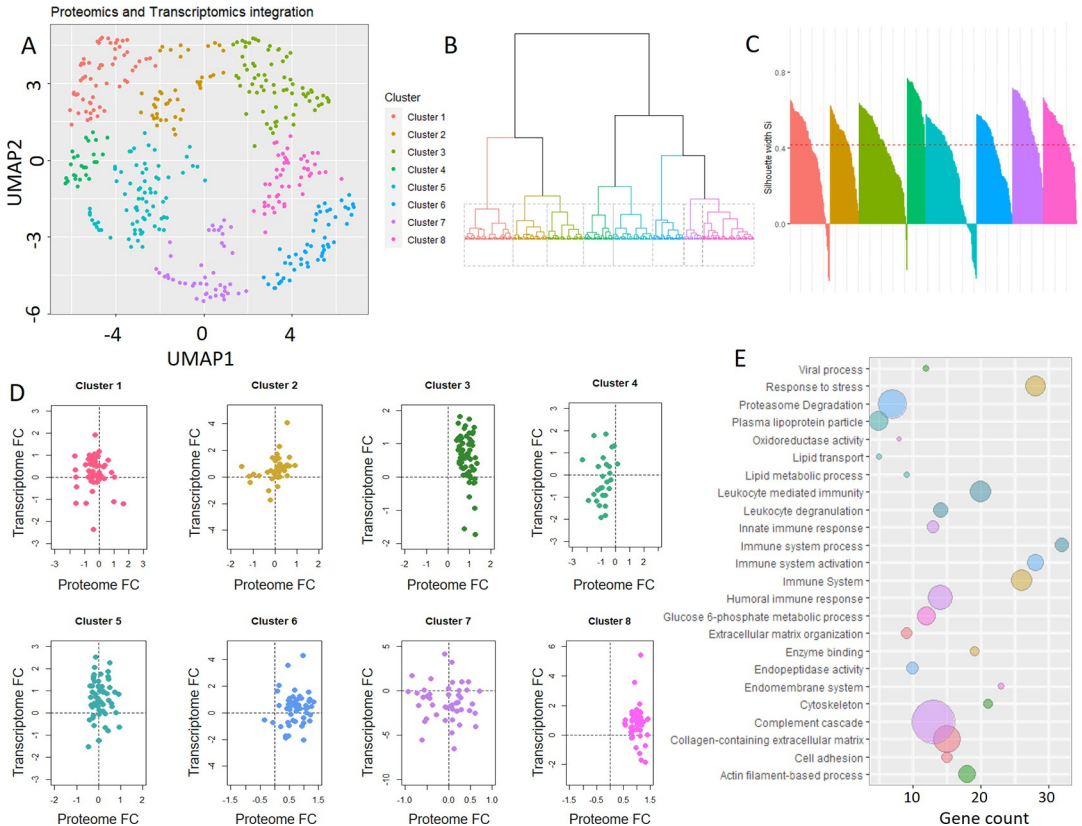

Fig. 7 Integration of lung transcriptomics and plasma proteomics data. (A) UMAP plot indicating the clustering of 421 genes/proteins identified in both evaluated datasets; (B) Dendrogram indicating the separation of 8 clusters identified; (C) Silhouette graph showing the quality of clustering applied. Negative values indicate genes that were possibly wrongly clustered. These genes were disregarded; (D) Scatter plot representing the LogFC of the proteome (x) and transcriptome (y) analysis; (E) Gene ontology and enriched pathways associated with identified clusters. The size of the bubbles indicates the -log-q-value and the color represents the associated cluster.

proteasomal degradation, immune system, and enzyme activity; cluster 7, complement system and cluster 8, glucose metabolism (Fig. 7E).

The transcripts and proteins differentially regulated with a logFC $> |1|$ in the same direction (upregulation or downregulation) in both the proteome and the transcriptome were selected for application of the ROC curve (Fig. 8). The targets with the highest specificity and sensitivity were

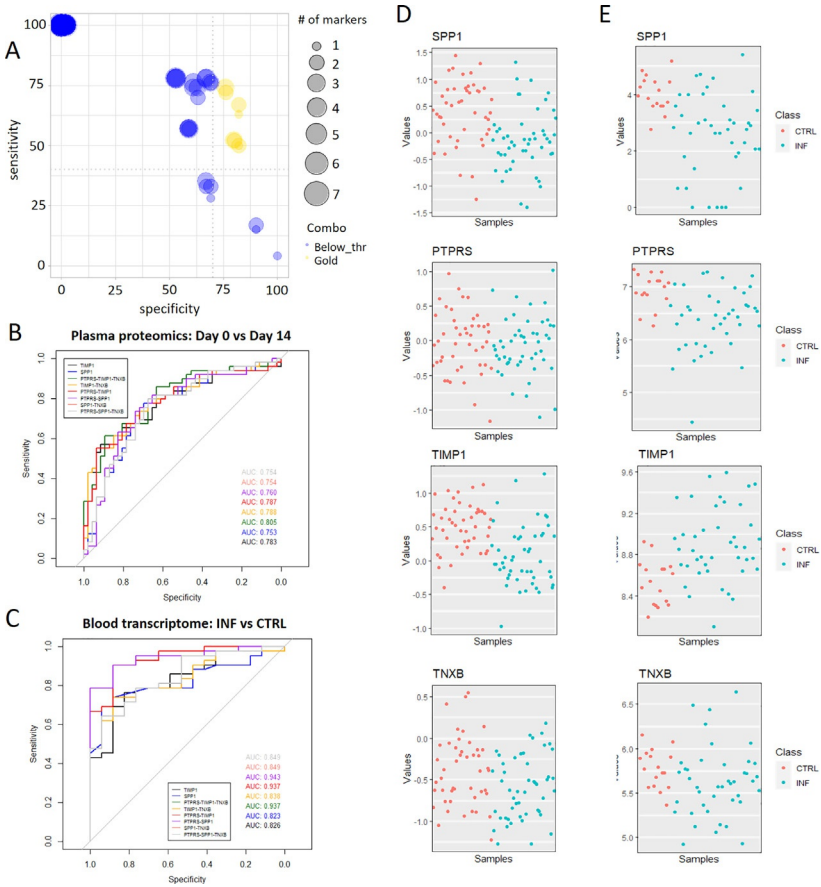

Fig. 8 ROC curves based on proteins involved in SARS-CoV-2 pathology. (A) Bubble chart indicating combinations between proteins/genes that present better results (golden combinations); (B) ROC curves of gold combinations indicating the potential for protein separation in distinguishing between groups D0 (PCR positive, day 01) and D14 (PCR negative, day 14); (C) ROC curves of gold combinations indicating the gene splitting potential in distinguishing between infected (INF) and healthy (CTRL) groups from whole blood transcriptomics analysis; (D–E) Boxplots indicating the separation of patients based on the selected protein/gene.

selected based in the integrated expression dataset (plasma proteome and lung transcriptome) (Fig. 8A). Additionally, the predictive power of these targets to stratify infected and healthy conditions was tested in another cohort (plasma proteome of D0 and D14 and whole blood transcriptome from INF and CTRL).

Three biomolecules gave the best AUC: TIMP1, a natural inhibitor of the matrix metalloproteinases involved in extracellular matrix remodeling, cell proliferation and regulated in response to cytokines; TNXB, tenascin XB, an extracellular matrix glycoprotein involved in wound healing, and PTPRS, protein tyrosine phosphatase receptor type S, the cell surface receptor that binds to glycosaminoglycans and involved in cell-cell interaction and cell growth and differentiation (Fig. 8B). The combination of the three showed an AUC greater than 0.8. At the gene level, the same combination resulted in AUC higher than 0.9 (Fig. 8C). Other combinations that showed AUC values close to 1 were PTPRS and TIMP1; and PTPRS and SPP1, a sialoprotein related to lymphocytic activation (Fig. 8D and E).

5. Discussion

To explore and understand the mechanisms of SARS-CoV-2 infection, the virus responsible for the death of more than 5 million people worldwide (https://covid19.who.int/), we present here an optimized pipeline for omics data analysis (Fig. 1), as well as extract meaningful information from high-throughput techniques. Dealing with codes in command line may not be a common expertise; however, being able to access and interpret public data is of great advantage for performing *in silico* validations and designing experiments in the wet lab to test hypotheses. We present the application of computational tools to analyze raw transcriptomics data and funnel the findings to access key disease processes. In addition, we integrated quantitative proteomics data to improve our understanding of COVID-19.

The rapid need to understand the mechanisms and impacts of SARS-CoV-2 makes omics strategies stand out as they can offer a holistic and comprehensive view of thousands of molecules (Colinge & Bennett, 2007; Haas, Muralidharan, Krogan, Kaake, & Hüttenhain, 2021). The ability to study alterations of a large number of proteins and transcripts in a single experiment is attractive (Al-Amrani, Al-Jabri, Al-Zaabi, Alshekaili, & Al-Khabori, 2021). Furthermore, accessing the modulation of molecules based on a clinical status can answer many questions about a

given disease (Amiri-Dashatan, Koushki, Abbaszadeh, Rostami-Nejad, & Rezaei-Tavirani, 2018). However, to access these countless possibilities offered by omics, it is necessary to conduct information processing to reduce noisy data and focus only on what is relevant. Computational and bioinformatic analyses complement the sample preparation and data acquisition steps allowing a comprehensive visualization of dysregulated pathways (Cannataro, 2007).

The mechanisms altered by SARS-CoV-2 during infection can be accessed through the application of shotgun proteomics (Badua, Baldo, & Medina, 2021; Rais, Fu, & Drabovich, 2021). In addition, this approach can provide the quantification of viral proteins in complex clinical samples (Bojkova et al., 2020; Lachén-Montes, Corrales, Fernández-Irigoyen, & Santamaría, 2020; Lazari et al., 2021). Reports of sequelae resulting from viral infection are increasingly frequent. Chen et al. (2021) applied a proteomic approach monitoring the health status of patients who developed COVID-19. The study showed changes in proteins related to cholesterol metabolism and heart disease. Our *in silico* analyses resulting from the application of the deft pipeline detailed here showed an alteration in processes linked to lipoproteins and cholesterol. These results are in agreement with changes in the HDL proteome associated with COVID-19 complications (Souza Junior et al., 2021). The data obtained in this manuscript show that the plasma proteome reflects the immunological status of infected patients, since pathways and proteins were identified as increased in these groups. The intra-patient comparison at different time points showed there is an activation of the immune system, especially of pathways linked to leukocytes in the first 24 h. However, these responses are diminished after 14 days. Other studies also reported the ability of the plasma and serum proteome to reflect the action of the immune system in the infection (Arthur et al., 2021; Kumar, 2021; Villar et al., 2021; Zhong et al., 2021). Among the immune system proteins that stood out were C1QA, which is crucial in the activation of the classical pathway of the complement system (Ghebrehiwet, Hosszu, Valentino, & Peerschke, 2012; Kouser et al., 2015; Nayak, Pednekar, Reid, & Kishore, 2012), and CD93, involved in the regulation of phagocytosis of apoptotic cells *in vivo*. Transcriptomics approaches have also been extensively applied to study SARS-CoV-2 infection (Butler et al., 2021; Chakraborty, Sharma, Bhattacharya, Zayed, & Lee, 2021; Islam et al., 2021; O'Donnell et al., 2021; Sun et al., 2020; Wong et al., 2021). We identified regulated pathways linked to misfolded proteins in the lung of infected patients, as previously reported (Rosa-Fernandes

et al., 2021). Moreover, we identified an increase in the enzymes involved ECM remodeling, extravasation of intracellular contents and activation of the immune system (Leeming et al., 2021; Leng et al., 2020).

Leukocyte activation was recurrent in our analyses of the integrated proteomic and transcriptomic data (Alon et al., 2021; Coradi & Vieira, 2021). Regarding our data integration approach, we saw that there is little correlation between transcripts mapped in lung and whole blood transcriptome. However, looking at moderately conserved transcripts, we found processes linked to leukocyte activation and host glycosylation. Our integrated data also mapped transcripts and proteins linked to proteasome activity in the infected groups. In fact, other groups have identified the importance of the proteasome in COVID-19 from the application of other techniques (Chatterjee et al., 2020; Longhitano et al., 2020; Wang et al., 2021).

6. Conclusions

We conducted integrated omic data analysis with focus on open tools to analyze and interpret this data type. We presented the entire pipeline for handling from raw files to differentially expressed features using the FastQC, Trim Galore!, HISAT2, samtools, featureCounts and R softwares. We also looked at useful tools to explore processes and pathways related to mapped transcripts and proteins. An important point was establishing a pipeline to integrate proteomics and transcriptomics data, since adjusting these data to make them comparable can be a big challenge. We also consistently reported applying various packages available in the bioconductor, providing the codes used. Regarding the biological findings, we evidenced increase in the cholesterol metabolism, immune system activity, ECM and proteasome degradation increased in infected patients. We also noticed a leukocyte activation profile in both proteomics and transcriptomics data. Finally, we identified a panel of proteins and transcripts that are regulated in the same direction in the lung transcriptome and plasma proteome that have a great ability to distinguish between healthy and infected groups. This panel of markers was applied to another cohort of patients and showed good results, corroborating the robustness and usefulness of the computational tools presented in this manuscript.

Acknowledgments

We are grateful for the financial support provided by the São Paulo Research Foundation (FAPESP, grants processes n° 2018/18257-1 (G.P.), 2018/15549-1 (G.P.), 2021/00140-3

(J.M.D.S.)); by the Conselho Nacional de Desenvolvimento Científico e Tecnológico ("Bolsa de Produtividade" (S.K.N.M. and G.P.)); by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (process n° 88882.463201/2019-01 (L.R.F.)).

Data and code availability

The codes used in this manuscript were made available in the Supplementary File 1 in the online version at https://doi.org/10.1016/bs.apcsb.2022.04.002. The input and output data are available in Supplementary File 2 in the online version at https://doi.org/10.1016/bs.apcsb.2022.04.002.

References

Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., & Al-Khabori, M. (2021). Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, *12*, 57–69. https://doi.org/10.4331/wjbc.v12.i5.57.

Alon, R., Sportiello, M., Kozlovski, S., Kumar, A., Reilly, E. C., Zarbock, A., et al. (2021). Leukocyte trafficking to the lungs and beyond: Lessons from influenza for COVID-19. *Nature Reviews. Immunology*, *21*, 49–64. https://doi.org/10.1038/s41577-020-00470-2.

Amiri-Dashatan, N., Koushki, M., Abbaszadeh, H.-A., Rostami-Nejad, M., & Rezaei-Tavirani, M. (2018). Proteomics applications in health: Biomarker and drug discovery and food industry. *Iranian Journal of Pharmaceutical Research*, *17*, 1523–1536.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.

Arthur, L., Esaulova, E., Mogilenko, D. A., Tsurinov, P., Burdess, S., Laha, A., et al. (2021). Cellular and plasma proteomic determinants of COVID-19 and non-COVID-19 pulmonary diseases relative to healthy aging. *Nature Aging*, *1*, 535–549. https://doi.org/10.1038/s43587-021-00067-x.

Badua, C. L. D. C., Baldo, K. A. T., & Medina, P. M. B. (2021). Genomic and proteomic mutation landscapes of SARS-CoV-2. *Journal of Medical Virology*, *93*, 1702–1721. https://doi.org/10.1002/jmv.26548.

Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., et al. (2020). Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*, *583*, 469–472. https://doi.org/10.1038/s41586-020-2332-7.

Butler, D., Mozsary, C., Meydan, C., Foox, J., Rosiene, J., Shaiber, A., et al. (2021). Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *Nature Communications*, *12*, 1660. https://doi.org/10.1038/s41467-021-21361-7.

Cannataro, M. (2007). Computational proteomics: Management and analysis of proteomics data. *Briefings in Bioinformatics*, *9*, 97–101. https://doi.org/10.1093/bib/bbn011.

Carvalho, P. C., Lima, D. B., Leprevost, F. V., Santos, M. D. M., Fischer, J. S. G., Aquino, P. F., et al. (2016). Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. *Nature Protocols*, *11*, 102–117. https://doi.org/10.1038/nprot.2015.133.

Chakraborty, C., Sharma, A. R., Bhattacharya, M., Zayed, H., & Lee, S.-S. (2021). Understanding gene expression and transcriptome profiling of COVID-19: An initiative towards the mapping of protective immunity genes against SARS-CoV-2 infection. *Frontiers in Immunology*, *12*, 724936. https://doi.org/10.3389/fimmu.2021.724936.

Chatterjee, P., Ponnapati, M., Kramme, C., Plesa, A. M., Church, G. M., & Jacobson, J. M. (2020). Targeted intracellular degradation of SARS-CoV-2 via computationally optimized peptide fusions. *Communications Biology*, *3*, 715. https://doi.org/10.1038/s42003-020-01470-7.

Chen, Y., Yao, H., Zhang, N., Wu, J., Gao, S., Guo, J., et al. (2021). Proteomic analysis identifies prolonged disturbances in pathways related to cholesterol metabolism and myocardium function in the COVID-19 recovery stage. *Journal of Proteome Research*, *20*, 3463–3474. https://doi.org/10.1021/acs.jproteome.1c00054.

Colinge, J., & Bennett, K. L. (2007). Introduction to computational proteomics. *PLoS Computational Biology*, *3*, e114. https://doi.org/10.1371/journal.pcbi.0030114.

Coradi, C., & Vieira, S. L. V. (2021). Alterações leucocitárias em pacientes com COVID-19 observadas em extensão de sangue periférico. *Research, Society and Development*, *10*, e400101119838. https://doi.org/10.33448/rsd-v10i11.19838.

Das, A., Ahmed, R., Akhtar, S., Begum, K., & Banu, S. (2021). An overview of basic molecular biology of SARS-CoV-2 and current COVID-19 prevention strategies. *Gene Reports*, *23*, 101122. https://doi.org/10.1016/j.genrep.2021.101122.

Desai, N., Neyaz, A., Szabolcs, A., Shih, A. R., Chen, J. H., Thapar, V., et al. (2020). Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nature Communications*, *11*, 6319. https://doi.org/10.1038/s41467-020-20139-7.

Dixon, B. E., Wools-Kaloustian, K. K., Fadel, W. F., Duszynski, T. J., Yiannoutsos, C., Halverson, P. K., et al. (2021). Symptoms and symptom clusters associated with SARS-CoV-2 infection in community-based populations: Results from a statewide epidemiological study. *PLoS One*, *16*, e0241875. https://doi.org/10.1371/journal.pone.0241875.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Ghebrehiwet, B., Hosszu, K. K., Valentino, A., & Peerschke, E. I. B. (2012). The C1q family of proteins: Insights into the emerging non-traditional functions. *Frontiers in Immunology*, *3*. https://doi.org/10.3389/fimmu.2012.00052.

Guangcan, S., Yong, C., Zhenlin, C., Chao, L., Shangtong, L., Hao, C., et al. (2021). How to use open-pFind in deep proteomics data analysis?—A protocol for rigorous identification and quantitation of peptides and proteins from mass spectrometry data. *Biophysics Reports*, 7, 207–226. https://doi.org/10.52601/bpr.2021.210004.

Haas, P., Muralidharan, M., Krogan, N. J., Kaake, R. M., & Hüttenhain, R. (2021). Proteomic approaches to study SARS-CoV-2 biology and COVID-19 pathology. *Journal of Proteome Research*, *20*, 1133–1152. https://doi.org/10.1021/acs.jproteome.0c00764.

Hayes, L. D., Ingram, J., & Sculthorpe, N. F. (2021). More than 100 persistent symptoms of SARS-CoV-2 (long COVID): A scoping review. *Frontiers in Medicine*, *8*, 750378. https://doi.org/10.3389/fmed.2021.750378.

Islam, A. B. M. M. K., Khan, M. A.-A.-K., Ahmed, R., Hossain, M. S., Kabir, S. M. T., Islam, M. S., et al. (2021). Transcriptome of nasopharyngeal samples from COVID-19 patients and a comparative analysis with other SARS-CoV-2 infection models reveal disparate host responses against SARS-CoV-2. *Journal of Translational Medicine*, *19*, 32. https://doi.org/10.1186/s12967-020-02695-0.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*, R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Kodama, Y., Shumway, M., Leinonen, R., & International Nucleotide Sequence Database Collaboration. (2012). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, *40*, D54–D56. https://doi.org/10.1093/nar/gkr854.

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, *14*, 513–520. https://doi.org/10.1038/nmeth.4256.

Kouser, L., Madhukaran, S. P., Shastri, A., Saraon, A., Ferluga, J., Al-Mozaini, M., et al. (2015). Emerging and novel functions of complement protein C1q. *Frontiers in Immunology*, *6*. https://doi.org/10.3389/fimmu.2015.00317.

Kumar, V. (2021). Can proteomics-based approaches further help COVID-19 prevention and therapy? *Expert Review of Proteomics*, *18*, 241–245. https://doi.org/10.1080/14789450.2021.1924684.

Kwan, P. K. W., Cross, G. B., Naftalin, C. M., Ahidjo, B. A., Mok, C. K., Fanusi, F., et al. (2021). A blood RNA transcriptome signature for COVID-19. *BMC Medical Genomics*, *14*, 155. https://doi.org/10.1186/s12920-021-01006-w.

Lachén-Montes, M., Corrales, F. J., Fernández-Irigoyen, J., & Santamaría, E. (2020). Proteomics insights into the molecular basis of SARS-CoV-2 infection: What we can learn from the human olfactory axis. *Frontiers in Microbiology*, *11*, 2101. https://doi.org/10.3389/fmicb.2020.02101.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., et al. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000 Research*, *5*. https://doi.org/10.12688/f1000research.9005.3. ISCB Comm J-1408.

Lazari, L. C., Ghilardi, F. D. R., Rosa-Fernandes, L., Assis, D. M., Nicolau, J. C., Santiago, V. F., et al. (2021). Prognostic accuracy of MALDI-TOF mass spectrometric analysis of plasma in COVID-19. *Life Science Alliance*, *4*, e202000946. https://doi.org/10.26508/lsa.202000946.

Leeming, D. J., Genovese, F., Sand, J. M. B., Rasmussen, D. G. K., Christiansen, C., Jenkins, G., et al. (2021). Can biomarkers of extracellular matrix remodelling and wound healing be used to identify high risk patients infected with SARS-CoV-2?: Lessons learned from pulmonary fibrosis. *Respiratory Research*, *22*, 38. https://doi.org/10.1186/s12931-020-01590-y.

Leng, L., Cao, R., Ma, J., Mou, D., Zhu, Y., Li, W., et al. (2020). Pathological features of COVID-19-associated lung injury: A preliminary proteomics report based on clinical samples. *Signal Transduction and Targeted Therapy*, *5*, 240. https://doi.org/10.1038/s41392-020-00355-9.

Li, Y., Hou, G., Zhou, H., Wang, Y., Tun, H. M., Zhu, A., et al. (2021). Multi-platform omics analysis reveals molecular signature for COVID-19 pathogenesis, prognosis and drug target discovery. *Signal Transduction and Targeted Therapy*, *6*, 155. https://doi.org/10.1038/s41392-021-00508-4.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

Longhitano, L., Tibullo, D., Giallongo, C., Lazzarino, G., Tartaglia, N., Galimberti, S., et al. (2020). Proteasome inhibitors as a possible therapy for SARS-CoV-2. *International Journal of Molecular Sciences*, *21*, E3622. https://doi.org/10.3390/ijms21103622.

Macedo-da-Silva, J., Rosa-Fernandes, L., Barbosa, R. H., Angeli, C. B., Carvalho, F. R., de Oliveira Vianna, R. A., et al. (2020). Serum proteomics reveals alterations in protease activity, axon guidance, and visual phototransduction pathways in infants with in utero exposure to zika virus without congenital zika syndrome. *Frontiers in Cellular and Infection Microbiology*, *10*, 577819. https://doi.org/10.3389/fcimb.2020.577819.

Mahalmani, V. M., Mahendru, D., Semwal, A., Kaur, S., Kaur, H., Sarma, P., et al. (2020). COVID-19 pandemic: A review based on current evidence. *Indian Journal of Pharmacology*, *52*, 117–129. https://doi.org/10.4103/ijp.IJP_310_20.

Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K.-M., Distler, M. G., Zelikovsky, A., et al. (2019). Systematic benchmarking of omics computational tools. *Nature Communications*, *10*, 1393. https://doi.org/10.1038/s41467-019-09406-4.

Martens, L., & Vizcaíno, J. A. (2017). A golden age for working with public proteomics data. *Trends in Biochemical Sciences*, *42*, 333–341. https://doi.org/10.1016/j.tibs.2017.01.001.

Mishra, S., Shah, M. I., Udhaya Kumar, S., Thirumal Kumar, D., Gopalakrishnan, C., Al-Subaie, A. M., et al. (2021). Network analysis of transcriptomics data for the prediction and prioritization of membrane-associated biomarkers for idiopathic pulmonary fibrosis (IPF) by bioinformatics approach. In *Advances in protein chemistry and structural biology* (pp. 241–273). Elsevier. https://doi.org/10.1016/bs.apcsb.2020.10.003.

Musich, R., Cadle-Davidson, L., & Osier, M. V. (2021). Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science*, *12*, 657240. https://doi.org/10.3389/fpls.2021.657240.

Nalbandian, A., Sehgal, K., Gupta, A., Madhavan, M. V., McGroder, C., Stevens, J. S., et al. (2021). Post-acute COVID-19 syndrome. *Nature Medicine*, *27*, 601–615. https://doi.org/10.1038/s41591-021-01283-z.

Nayak, A., Pednekar, L., Reid, K. B., & Kishore, U. (2012). Complement and non-complement activating functions of C1q: A prototypical innate immune molecule. *Innate Immunity*, *18*, 350–363. https://doi.org/10.1177/1753425910396252.

O'Donnell, K. L., Pinski, A. N., Clancy, C. S., Gourdine, T., Shifflett, K., Fletcher, P., et al. (2021). Pathogenic and transcriptomic differences of emerging SARS-CoV-2 variants in the Syrian golden hamster model. *eBioMedicine*, *73*, 103675. https://doi.org/10.1016/j.ebiom.2021.103675.

Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., et al. (2021). Large-scale multi-omic analysis of COVID-19 severity. *Cell Systems*, *12*, 23–40.e7. https://doi.org/10.1016/j.cels.2020.10.003.

Rais, Y., Fu, Z., & Drabovich, A. P. (2021). Mass spectrometry-based proteomics in basic and translational research of SARS-CoV-2 coronavirus and its emerging mutants. *Clinical Proteomics*, *18*, 19. https://doi.org/10.1186/s12014-021-09325-x.

Rosa-Fernandes, L., Barbosa, R. H., Dos Santos, M. L. B., Angeli, C. B., Silva, T. P., Melo, R. C. N., et al. (2020). Cellular imprinting proteomics assay: A novel method for detection of neural and ocular disorders applied to congenital zika virus syndrome. *Journal of Proteome Research*, *19*, 4496–4515. https://doi.org/10.1021/acs.jproteome.0c00320.

Rosa-Fernandes, L., Cugola, F. R., Russo, F. B., Kawahara, R., de Melo Freire, C. C., Leite, P. E. C., et al. (2019). Zika virus impairs neurogenesis and synaptogenesis pathways in human neural stem cells and neurons. *Frontiers in Cellular Neuroscience*, *13*, 64. https://doi.org/10.3389/fncel.2019.00064.

Rosa-Fernandes, L., Lazari, L. C., da Silva, J. M., de Morais Gomes, V., Machado, R. R. G., dos Santos, A. F., et al. (2021). SARS-CoV-2 activates ER stress and Unfolded protein response (preprint). *Biochemistry*. https://doi.org/10.1101/2021.06.21.449284.

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. K. (2020). Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant *Arabidopsis thaliana*. *International Journal of Molecular Sciences*, *21*, E1720. https://doi.org/10.3390/ijms21051720.

Shu, T., Ning, W., Wu, D., Xu, J., Han, Q., Huang, M., et al. (2020). Plasma proteomics identify biomarkers and pathogenesis of COVID-19. *Immunity*, *53*, 1108–1122.e5. https://doi.org/10.1016/j.immuni.2020.10.008.

Souza Junior, D. R., Silva, A. R. M., Rosa-Fernandes, L., Reis, L. R., Alexandria, G., Bhosale, S. D., et al. (2021). HDL proteome remodeling associates with COVID-19 severity. *Journal of Clinical Lipidology*, *15*, 796–804. https://doi.org/10.1016/j.jacl.2021.10.005.

Sun, J., Ye, F., Wu, A., Yang, R., Pan, M., Sheng, J., et al. (2020). Comparative transcriptome analysis reveals the intensive early stage responses of host cells to SARS-CoV-2 infection. *Frontiers in Microbiology*, *11*, 593857. https://doi.org/10.3389/fmicb.2020.593857.

Terracciano, R., Preianò, M., Fregola, A., Pelaia, C., Montalcini, T., & Savino, R. (2021). Mapping the SARS-CoV-2-host protein-protein interactome by affinity purification mass spectrometry and proximity-dependent biotin labeling: A rational and straightforward route to discover host-directed anti-SARS-CoV-2 therapeutics. *International Journal of Molecular Sciences*, *22*, E532. https://doi.org/10.3390/ijms22020532.

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*, 2301–2319. https://doi.org/10.1038/nprot.2016.136.

Villar, M., Urra, J. M., Rodríguez-del-Río, F. J., Artigas-Jerónimo, S., Jiménez-Collados, N., Ferreras-Colino, E., et al. (2021). Characterization by quantitative serum proteomics of immune-related prognostic biomarkers for COVID-19 symptomatology. *Frontiers in Immunology*, *12*, 730710. https://doi.org/10.3389/fimmu.2021.730710.

Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Research*, *41*, D1063–D1069. https://doi.org/10.1093/nar/gks1262.

Wang, Z., Zhao, Y., Wang, Q., Xing, Y., Feng, L., Kong, J., et al. (2021). Identification of proteasome and caspase inhibitors targeting SARS-CoV-2 Mpro. *Signal Transduction and Targeted Therapy*, *6*, 214. https://doi.org/10.1038/s41392-021-00639-8.

Wenk, M. R. (2005). The emerging field of lipidomics. *Nature Reviews. Drug Discovery*, *4*, 594–610. https://doi.org/10.1038/nrd1776.

Wong, H. S.-C., Guo, C.-L., Lin, G.-H., Lee, K.-Y., Okada, Y., & Chang, W.-C. (2021). Transcriptome network analyses in human coronavirus infections suggest a rational use of immunomodulatory drugs for COVID-19 therapy. *Genomics*, *113*, 564–575. https://doi.org/10.1016/j.ygeno.2020.12.041.

Wu, P., Chen, D., Ding, W., Wu, P., Hou, H., Bai, Y., et al. (2021). The trans-omics landscape of COVID-19. *Nature Communications*, *12*, 4543. https://doi.org/10.1038/s41467-021-24482-1.

Wu, M., Chen, Y., Xia, H., Wang, C., Tan, C. Y., Cai, X., et al. (2020). Transcriptional and proteomic insights into the host response in fatal COVID-19 cases. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 28336–28343. https://doi.org/10.1073/pnas.2018030117.

Zhong, W., Altay, O., Arif, M., Edfors, F., Doganay, L., Mardinoglu, A., et al. (2021). Next generation plasma proteome profiling of COVID-19 patients with mild to moderate symptoms. *eBioMedicine*, *74*, 103723. https://doi.org/10.1016/j.ebiom.2021.103723.

Zhu, Z., Zhang, S., Wang, P., Chen, X., Bi, J., Cheng, L., et al. (2022). A comprehensive review of the analysis and integration of omics data for SARS-CoV-2 and COVID-19. *Briefings in Bioinformatics*, *23*, bbab446. https://doi.org/10.1093/bib/bbab446.