Aref Hosseini
Freiburgstrasse 63
aref.hosseini@unibe.ch

Data Science Project

Machine Learning-Based Classification of Transcriptome Signatures of Coronavirus disease 2019 patients

Conceptual Design Report

25th of September 2024

Abstract

A biomarker is an indicator of a biological state, often in response to an intervention or the stage of a disease. In recent years, machine learning has been widely applied in biomarker discovery. Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in a global pandemic with significant morbidity and mortality worldwide, with the elderly population at particular risk for severe disease and mortality. The lack of reliable biomarkers makes it difficult to differentiate severe COVID-19 from mild patients accurately. This study introduces a machine learning (ML)-based approach for identifying mRNA signatures specific to severe COVID-19. Using next-generation sequencing (NGS) transcriptome data from biological samples of patients with severe, mild COVID-19 and control individuals, we aim to identify those with severe conditions and high risk. In the first step, we will filter out genes that show no changes in expression between these conditions to reduce noise. Then, we will apply both supervised and unsupervised machine learning approaches to identify genes with potential as classifiers for severe COVID-19.

Table of Contents

1 Project Objectives

During a pandemic like COVID-19, it is crucial to identify which patients are high-risk and need to remain in the hospital, and which patients are at lower risk and can be safely discharged. This ensures that hospital spaces and resources are available for those who need them most. While several factors can influence a physician's decision, having detailed information about the patient's condition can significantly assist in making more informed choices.

A biomarker is an indicator of a biological state, often in response to an intervention or disease progression. While biomarkers typically refer to physiological or physical phenotypes, at the molecular level, they can reflect disease-associated molecular changes and are useful in diagnosing diseases, infections, and neurological conditions, as well as in identifying therapeutic targets. In toxicological studies, biomarkers are frequently used to identify differentially expressed genes or proteins in response to toxic exposure or in chemical risk assessments. Data from various omics techniques, such as transcriptomics, proteomics, metabolomics, and epigenomics, provide valuable starting points for biomarker discovery.

Machine learning employs mathematical models to learn from data for specific tasks and has recently gained prominence in biomarker discovery. Key machine-learning techniques in this field include classification and feature selection. Among the technologies for whole transcriptome gene expression profiling, RNA sequencing (RNA-Seq) is one of the most widely used. Therefore, RNA-Seq data could be used as the input for these types of studies (1, 2). In this context, we propose a framework that utilizes machine learning to discover potential biomarkers, aiding in the identification of molecular drivers in COVID-19 patients.

Therefore, here is the primary objective of this project:

To categorize COVID-19 patients into two groups: severe patients (requiring intensive care) and mild patients (who can be discharged from the hospital).

2 Methods

Infrastructure

Raw sequence data will be obtained from the Gene Expression Omnibus (GEO), a public database repository created by the National Center for Biotechnology Information (NCBI) to store and freely distribute high-throughput gene expression data, including microarray, RNA sequencing (RNA-Seq), and other genomic datasets. GEO also contains basic clinical data related to samples. The data will then be stored locally for processing to generate an expression table for downstream analysis.

Tools and software libraries

The data analysis process can be divided into two main steps:

i. Analysis of raw data to normalized expression values:

For this step, we will convert raw read data into count values using the following Linux bash tools:

- **FastQC**: to assess the quality of the sequencing data.
- **HISAT2**: to map the reads to the reference genome.
- **FeatureCounts**: to count the number of reads overlapping with each gene, based on genome annotations (Homo_sapiens.GRCh38.94).

To normalize the values, identify differentially expressed genes, and generate visualizations, we will use the following R tools:

- **DESeq2** (Bioconductor package): for normalizing data and analyzing differentially expressed genes.
- **EnhancedVolcano**: to create volcano plots.
- **ggplot2**: for generating additional plots.

ii. Biomarker identification for each group:

For this step, Python will be the primary programming language, and we will use several libraries, including:

- **Pandas** and **NumPy**: For tasks such as data cleaning, preparation, exploration, and performing descriptive statistics. Pandas is especially effective for managing data, while NumPy excels at handling multidimensional data.
- **Plotly**: For visualizing data in the initial phases of exploration, Plotly offers interactive visualization capabilities.
- **scikit-learn**: This machine learning library will be used to train and evaluate different models on the data.

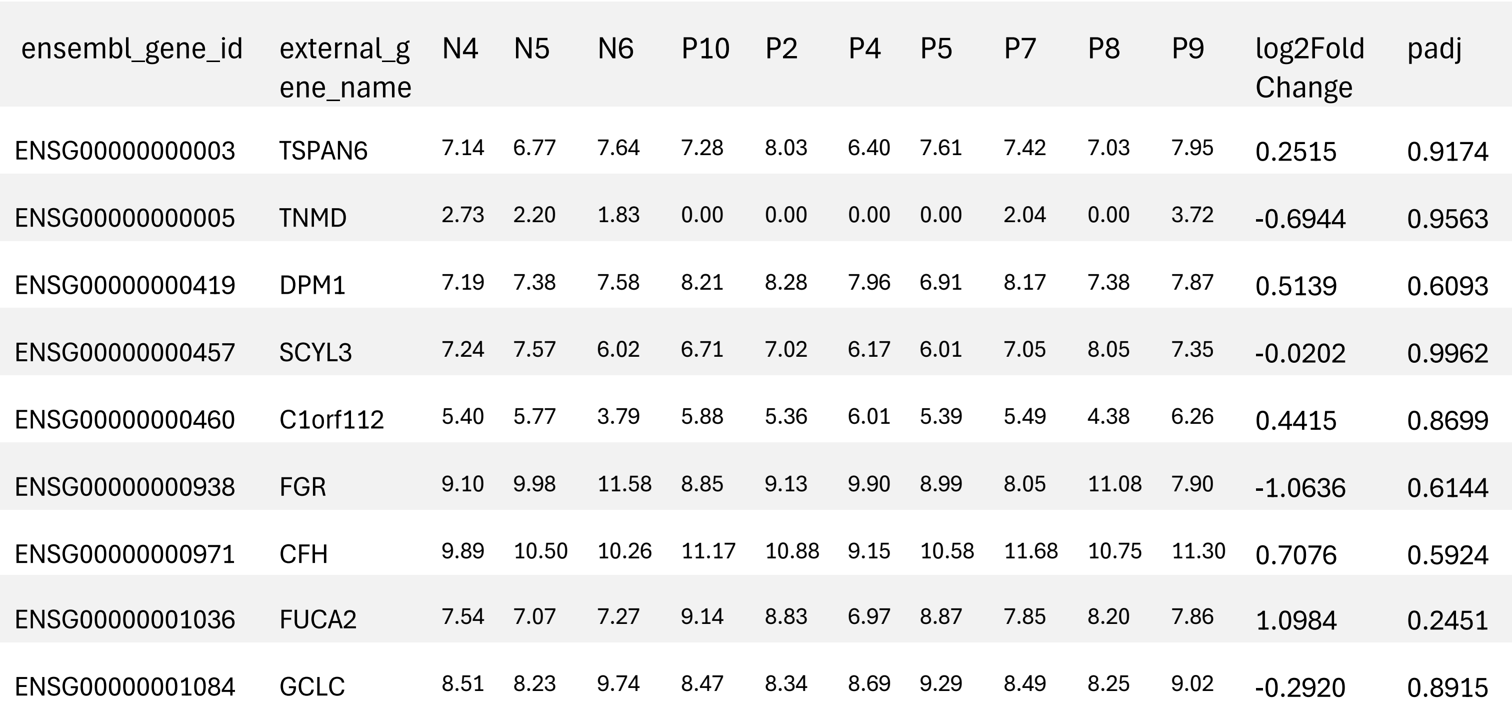
3 Data

In order to perform our analysis, we import relevant data of transcriptomic data that will be downloaded from GEO, NCBI. Since the data is publicly available, no security issues arise and no special measures need to be taken. To the time of writing this proposal, searching these keywords, (COVID-19) AND "Homo sapiens"[porgn:__txid9606], return 564 datasets. The relevant dataset with complete clinical data needs to be selected for the next step. Then the data will be downloaded through SRA database by the function of “wget” in Linux bash. The data will be converted to fastq files by using SRA Toolkit. By this, we would have the required data to be analyzed as described in the method section. Several different data sets could be used to train the models and then test the biomarker accuracy. Following you can find an example of the preprocessing of the data.

GSE208076, is a dataset that includes transcriptomics data of COVID-19 patients and control individuals. This data set has seven COVID-19 patients and three controls (3).

The output of the first part of the data analysis, providing the normalized expression values of genes in each sample, generated a table as shown in table 1.

Table 1



ensembl_gene_id	external_gene_name	N4	N5	N6	P10	P2	P4	P5	P7	P8	P9	log2FoldChange	padj
ENSG00000000003	TSPAN6	7.14	6.77	7.64	7.28	8.03	6.40	7.61	7.42	7.03	7.95	0.2515	0.9174
ENSG00000000005	TNMD	2.73	2.20	1.83	0.00	0.00	0.00	0.00	2.04	0.00	3.72	-0.6944	0.9563
ENSG00000000419	DPM1	7.19	7.38	7.58	8.21	8.28	7.96	6.91	8.17	7.38	7.87	0.5139	0.6093
ENSG00000000457	SCYL3	7.24	7.57	6.02	6.71	7.02	6.17	6.01	7.05	8.05	7.35	-0.0202	0.9962
ENSG00000000460	C1orf112	5.40	5.77	3.79	5.88	5.36	6.01	5.39	5.49	4.38	6.26	0.4415	0.8699
ENSG00000000938	FGR	9.10	9.98	11.58	8.85	9.13	9.90	8.99	8.05	11.08	7.90	-1.0636	0.6144
ENSG00000000971	CFH	9.89	10.50	10.26	11.17	10.88	9.15	10.58	11.68	10.75	11.30	0.7076	0.5924
ENSG00000001036	FUCA2	7.54	7.07	7.27	9.14	8.83	6.97	8.87	7.85	8.20	7.86	1.0984	0.2451
ENSG00000001084	GCLC	8.51	8.23	9.74	8.47	8.34	8.69	9.29	8.49	8.25	9.02	-0.2920	0.8915

4 Metadata

The metadata for the transcriptomics data in this project includes the health status of the samples. The metadata is made publicly available in the NCBI-GEO database. However, as we seek to identify a general biomarker that is conserved regardless of gender, age, or ethnicity, these variables will not be included in the analysis.

5 Data Quality

The data quality will be assessed from both technical and biological perspectives. During sequencing analysis, quality will be evaluated using the FastQC tool, with an example output shown in Figure 1. Low-quality data will be removed before proceeding to the next data processing step.

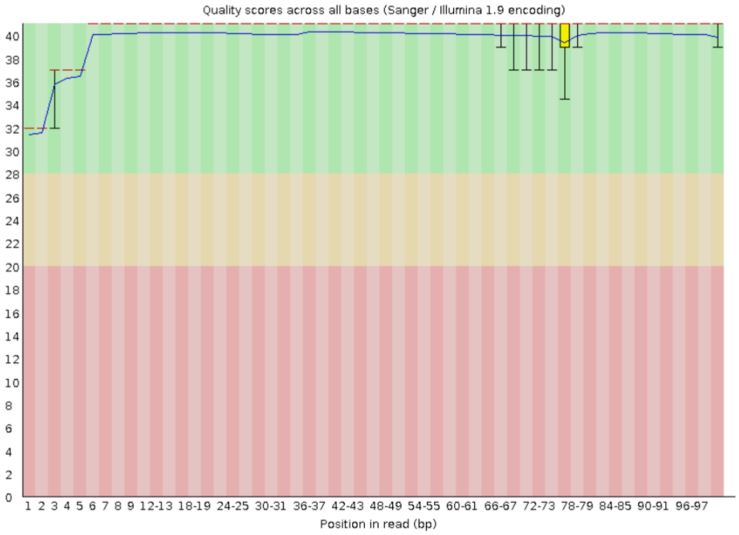


Figure 1. FastQC Quality Score Distribution Plot for RNA-Seq Data.

The plot shows the quality scores across all bases in the sequencing reads. The green region represents high-quality bases (Phred score > 30), indicating accurate base calling. The orange region corresponds to moderately reliable bases (Phred score 20–30), with some risk of errors. The red region highlights low-quality bases (Phred score < 20), where sequencing accuracy is likely compromised.

To evaluate the biological quality of the data, we will use hierarchical clustering and PCA plots to check if samples in different groups will be grouped. The datasets that the samples do not make separation between groups, will be excluded from the rest of the analysis. The PCA and heatmap plot of the GSE208076 dataset are depicted in Figure 2 a and b respectively.

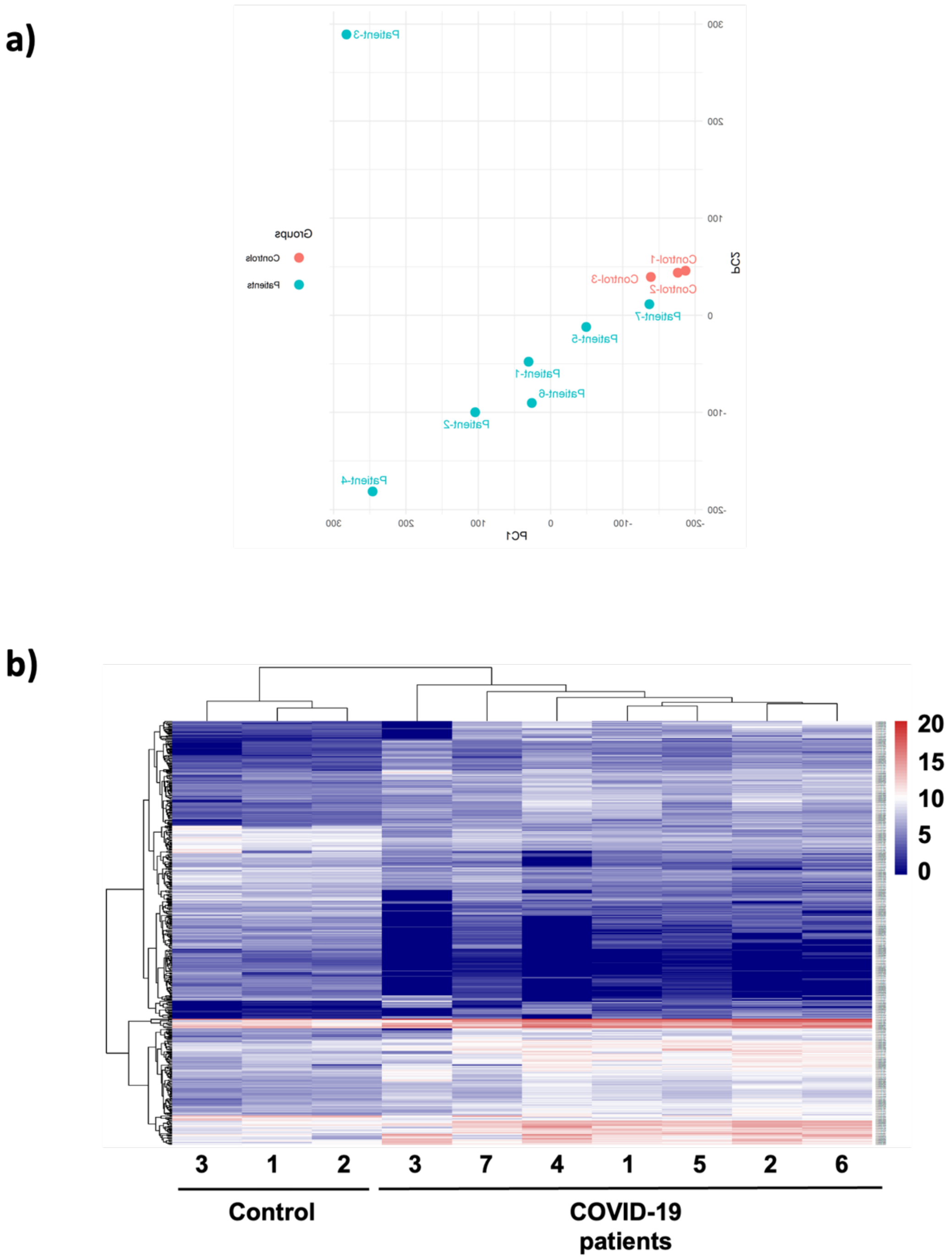


Figure 2. Principal Component Analysis (PCA) Plot and Heatmap with hierarchical clustering.

a) The PCA plot illustrates the variance in gene expression across samples. Each point represents a sample, with the proximity between points indicating their similarity based on expression profiles. The two principal components (PC1 and PC2) capture the majority of the variance, allowing visualization of sample clustering and differences. b) The heatmap shows the hierarchical clustering of differentially expressed genes across samples. Rows represent genes, and columns represent samples. Colors indicate expression levels. Dendrograms display the clustering of samples and genes based on expression similarity.

6 Data Flow

The various steps have been thoroughly detailed in the previous sections. The data flow is illustrated in Figure 3.

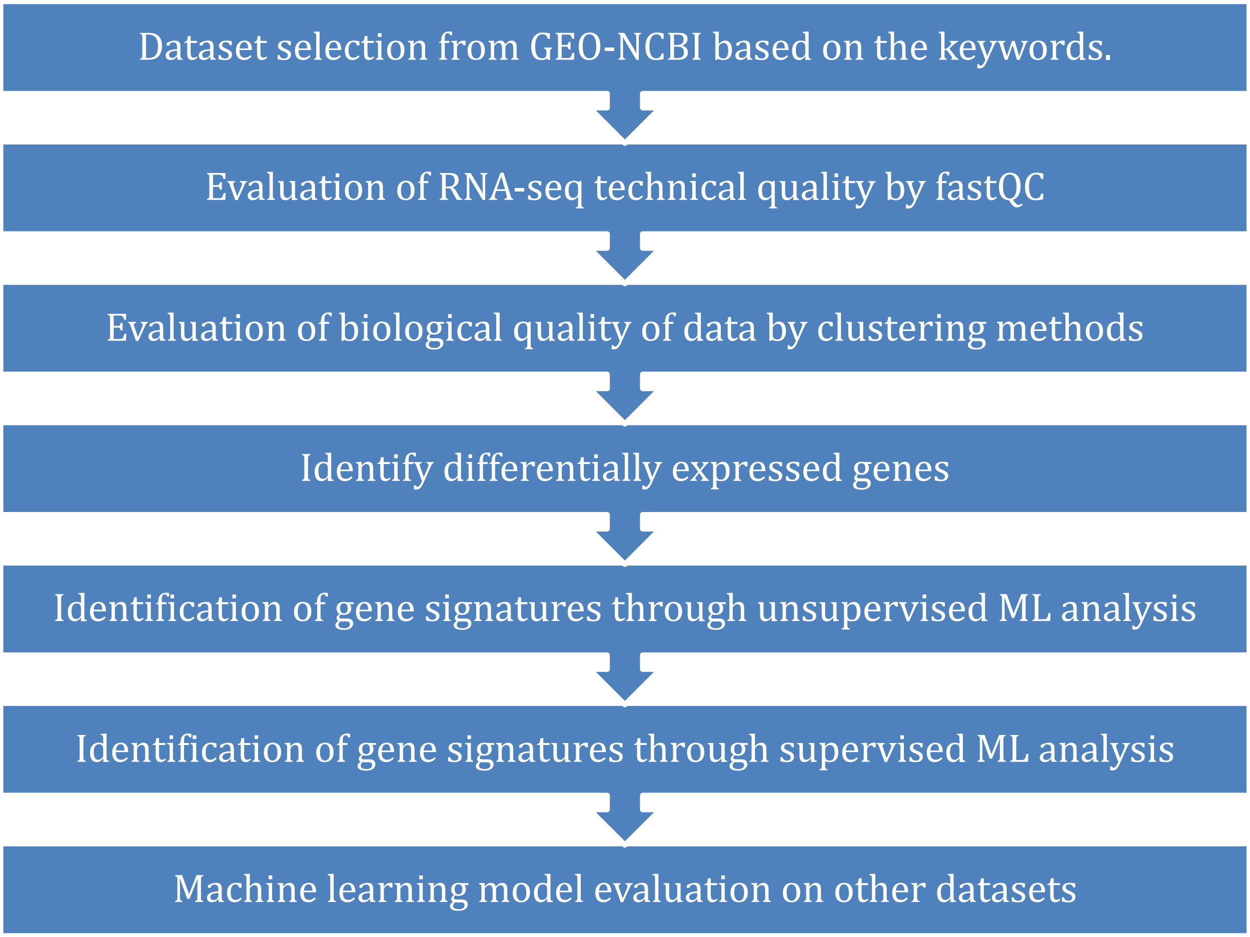

Dataset selection from GEO-NCBI based on the keywords.
Evaluation of RNA-seq technical quality by fastQC
Evaluation of biological quality of data by clustering methods
Identify differentially expressed genes
Identification of gene signatures through unsupervised ML analysis
Identification of gene signatures through supervised ML analysis
Machine learning model evaluation on other datasets

Figure 3: Flow diagram of the study.

7 Data Model

At the conceptual level, we are creating a dataset comprising gene expression profiles from individuals with COVID-19 with different severity levels and control samples.

At the logical level, we plan to utilize three different models—Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), and Random Forest Classifier (RF)—to classify individuals based on their gene expression data. The most significant genes contributing to this classification will be identified as potential biomarkers. To validate our findings, we will test our models against another available dataset.

At the physical level, the dataset is stored in a CSV file on a shared Google Drive, accessible exclusively to team members. We assume that the computations can be performed without requiring any specialized infrastructure.

8 Documentation

The results of this study have the potential to be published in a scientific journal, with the detailed script included in the supplementary materials.

9 Risks

The deposited transcriptomics data in GEO-NCBI, have various qualities. Several technical reasons could lead to low-quality input for the analysis, because of high inflammation in COVID-19 patients, the technical part is challenging and could reduce the quality of material significantly.

The other risk could be classification. If the disease did not induce huge differences in gene expression (effect size), we would not observe clear and separate groups in PCA and hierarchical clustering on the heatmap. This means that there would not be enough differentially expressed genes and good biomarkers.

10 Preliminary Studies

We have extracted differentially expressed genes between COVID-19 patients and the control group for the GSE208076 dataset. Figure 4 presents a scatter plot of the most dysregulated genes, highlighting distinct expression patterns.

To identify differentially expressed genes, we applied stringent thresholds of |log2FoldChange| > 2 and padj < 0.05 to ensure robust results. Figure 4 presents a volcano plot that visually highlights these differentially expressed genes. Using DESeq2 for differential expression analysis, we identified 272 upregulated and 202 downregulated genes in COVID-19 patients compared to control donors. These differentially expressed genes will serve as the basis for further biomarker analysis, where we will test their potential to act as reliable biomarkers for distinguishing between disease states.

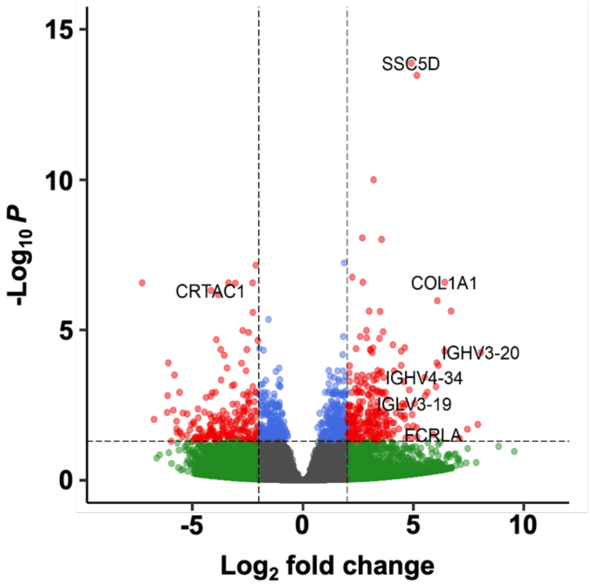


Figure 4. The volcano plot of differentially expressed genes (DEGs) in COVID–19 patients’ lungs was compared to control donors. A total of 474 DEGs were identified.

11 Conclusions

In this conceptual design report, we outlined how the reanalysis of publicly available datasets can be utilized to achieve the primary objective of our project, which is to identify specific biomarkers for each patient category. By leveraging these datasets, we aim to uncover molecular signatures that differentiate between patient groups, providing insights into disease mechanisms and aiding in the development of targeted diagnostic or therapeutic approaches. Additionally, this method can be applied to other categorizations of different groups using transcriptomic data, broadening its potential for discovering biomarkers in various diseases and conditions. This approach maximizes the value of existing data while addressing gaps in biomarker discovery across diverse patient populations.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 21.09.2024

Signature(s): Aref Hau

References and Bibliography

[1] Zhang X, Jonassen I, Goksøyr A. Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. Bioinformatics [Internet]. Brisbane (AU): Exon Publications; 2021 Mar 20. Chapter 4. (doi: 10.36255/exonpublications.bioinformatics.2021.ch4)

[2] Akshay A, Besic M, Kuhn A, Burkhard FC, Bigger-Allen A, Adam RM, Monastyrskaya K, Hashemi Gheinani A. Machine Learning-Based Classification of Transcriptome Signatures of Non-

Ulcerative Bladder Pain Syndrome. Int J Mol Sci. 2024 Jan 26;25(3):1568. (doi: 10.3390/ijms25031568.)

[3] Hosseini A, Stojkov D, Fettrelet T, Bilyy R, Yousefi S, Simon HU. Transcriptional Insights of Oxidative Stress and Extracellular Traps in Lung Tissues of Fatal COVID-19 Cases. Int J Mol Sci. 2023 Jan 31;24(3):2646. (doi: 10.3390/ijms24032646.)