
SAARLAND UNIVERSITY

Center for Bioinformatics & Saarland Informatics Campus
Master's Programme in Bioinformatics



Master's Thesis

Mulltiomics COVID-19 Analysis:

Results from the CORSAAR study in a Web App

Submitted by

Omar Laham

Supervisor

Prof. Dr. Andreas Keller

Advisor

Dr. Tobias Fehlmann

Reviewers

Prof. Dr. Andreas Keller

Prof. Dr. med. Dr. rer. nat. Robert Bals

Saarbrücken

February 2022



Advisor:

Prof. Dr. Andreas Keller
Clinical Chair for Bioinformatics
Bioinformatics Saarland Informatics Campus
Saarbrücken, Germany

Reviewer 1: Prof. Dr. Andreas Keller

Reviewer 2: Prof. Dr. med. Dr. rer. nat. Robert Bals

Saarland University
Bioinformatics Saarland Informatics Campus
Clinical Chair for Bioinformatics
Campus - Building E2.1
66123 Saarbrücken
Germany

Declarations

I hereby confirm that every part of this thesis is the result of my own work. I have also documented all the sources used to accomplish this work.

Saarbrücken, February 2022

Omar Laham

Acknowledgements

I would like to start this paper with soft thankful words to those who helped me and taught me every thing I needed to finish my thesis.

I start with thanking my supervisor Prof. Dr. Andreas Keller, who didn't save any effort to push this work to the beautiful result it ended up with. To my advisor Dr. Tobias Fehlmann, who wasn't only an advisor but a generous teacher who taught me lots in bioinformatics and was with me in every step. To everyone saw in me the doctor and the engineer and encouraged me to enroll in this master program to become an important joint between medicine and informatics in the future. To all teachers and scientists who lived and will live on this earth.

At the end, I'd like to write my warm thankful words to Prof. Dr. med. Dr. rer. nat. Robert Bals for reviewing my thesis and for leading the clinical efforts of collecting the clinical data for the CorSaar project as the head of the internal medicine department at the university hospital.

Abstract

The world got hit by COVID-19 in 2019 changing the way how the global health system and governments deal with pandemics. The rising number of infected individuals, the intensive care units full with patients with high level of disease severity and the high rate of mortality pushed everyone working in the medical field to do everything possible to help with this stunning issue.

Clinical bioinformaticians are ones of those who are at the first line challenging COVID-19 by applying bioinformatics to understand the biological changes behind COVID-19 and the biological underlying state of the patients infected by SARS-CoV-2.

Many universities and research centers around the globe are conducting very important multiomic studies on the available biological data they have from COVID-19 patients. In these studies, researchers integrate data from different omics (e.g. transcriptome, lipidome, epigenome) to uncover new insights about the disease aiming that they can come up with new guidelines that raise the survival of the patients and help them get stronger against the virus.

CORSAAR project is a project from Saarland state in Germany where specialists collected metadata and multiomics data from 49 COVID-19 patients in the Universitätsklinikum des Saarlandes.

This project provides analysis and visualization for single and multi-omics data from the CORSAAR project. This includes RNA-seq, proteome and peptidome, lipidome, metabolome, epigenome (DNA methylation) and miRNome.

By exploring the data and visualization from single and cross omic analysis, we notice how COVID-19 affects the body by creating significant modifications in multiple biological layers. Finding the most correlated features from each layer with disease severity is very useful for targeting treatment and drug industry.

By providing this project as a free open-access web app, we are aiming to leverage the ability of other researchers all around the globe to benefit from the data collected in the university hospital of Saarland and to confirm the findings they come up with while analyzing the data they obtain from other medical centers.

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Related Work	2
1.3	Thesis Statement	3
1.4	Outline	3
2	Biological Background	5
2.1	RNA-seq	5
2.2	Proteome	6
2.3	Lipidome	7
2.4	Epigenome (DNA Methylation)	8
2.5	miRNome	9
2.6	Integrative Omic Studies	10
3	Methods	12
3.1	Web App	12
3.1.1	Web server core	12
3.1.2	UI Theme	12
3.1.3	Visualization JavaScript Libraries	13
3.1.4	Data Storage	13
4	Application Overview	14
4.1	The two targets of the study	14
4.1.1	Severity Score and Mapping to WHO Score	15
4.2	Metadata	16
4.3	Feature Selection	18
4.4	Homepage	19
5	Single-Omic Analysis	22
5.1	Transcriptome (RNA-seq)	22
5.1.1	General Information	22
5.1.2	PCA (Principal Component Analysis)	23

5.1.3	Differential Analysis	24
5.1.4	Enrichment Analysis	27
5.1.5	Prime Features and Example Features	29
5.2	Proteome and Peptidome	34
5.2.1	General Information	34
5.2.2	PCA (Principal Component Analysis)	36
5.2.3	Prime Features and Example Features	37
5.3	Lipidome	45
5.3.1	General Information	45
5.3.2	PCA (Principal Component Analysis)	45
5.3.3	Differential Analysis	46
5.3.4	Prime Features and Visualization	47
5.4	Epigenome - DNA Methylation	51
5.4.1	General Information	51
5.4.2	PCA (Principal Component Analysis)	51
5.4.3	Prime Features and Example Features	52
5.5	miRNome	56
5.5.1	General Information	56
5.5.2	PCA (Principal Component Analysis)	57
5.5.3	Differential Analysis	58
5.5.4	miEAA - Enrichment Analysis	60
5.5.5	Prime Features and Example Features	64
5.6	Metabolome	68
5.6.1	General Information	68
6	Multomics Analysis	69
6.1	General Information	69
6.2	Overall Correlation Graph	69
6.3	Pairwise Analysis	71
6.3.1	Top Pairwise Correlation with Target	71
6.3.2	Correlation Heatmap for Prime Features Selected Under a Specific Target	71
6.3.3	Correlation Scatter Plot	72
7	Data Availability	75
7.1	Tables	75
7.2	Charts	76

8 Conclusion & Future Work	77
8.1 Conclusion	77
8.2 Future work	78
Bibliography	79

Chapter 1

Introduction

The world got hit by COVID-19 in 2019 changing the way how the global health system and governments deal with pandemics. This disease is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which belongs to the family coronaviridae in the order nidovirales. Four genera have been discovered for this virus: Alpha, Beta, Gamma and Delta coronavirus[1].

The world is currently mostly concerned about the Beta Corona virus genus since it's responsible for the majority of COVID-19 cases and many other serious respiratory diseases that can lead to death in humans[2]. After the revolution of genomics, scientists noticed that there must be other factors than DNA variants that are regulating the body and leading to disease. From that point, many other layers, like transcribed RNA (transcriptome), proteins (proteome), peptides (peptidome) and metabolites (metabolome) were explored excessively. The information we get from these layers, either in a single matter or in cross studies corresponds to a different omics type and is very important for the medical and pharmaceutical fields[3].

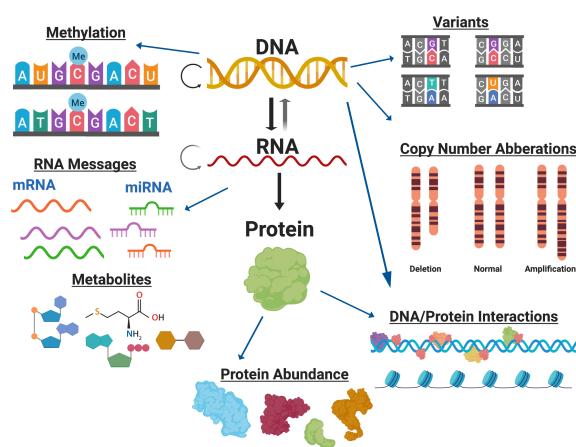


Figure 1 The multi-layer biological structure represented by multiomics. Obtained from [4].

In Figure 1, we can see the DNA (representing the base biological layer) giving raise to different types of RNA including mRNA from which proteins are then translated to do their diverse role in the body.

1.1 Motivation

The more the clinicians and researchers know about what's happening inside the bodies of humans suffering from a specific disease, the more help and protection they can provide to the humanity against it. Among these diseases, COVID-19 is not an exception. Many universities and research centers around the world are conducting very important multiomic studies on the available biological data they have from COVID-19 patients. In these studies, researchers integrate data from different omics (e.g. transcriptome, lipidome, epigome) to uncover new insights about the disease aiming that they can come up with new guidelines that raise the survival of the patients and help them get stronger against the virus.

CORSAAR project is a project from Saarland state in Germany where specialists collected metadata and multiomics data from 49 COVID-19 patients in the university hospital of Saarland.

This thesis presents a is available at [web application](#) that provides analysis and visualization for single and cross-omic analysis for the data from the CORSAAR project. This includes transcriptome, proteome and peptidome, lipidome, metabolome, epigenome (DNA methylation) and miRNome. The project is as available a free open-access resource at <https://wwwccb.uni-saarland.de/corsaar>.

It will always be very significant to unravel the biological changes behind illness to push the wheel of the pharmaceutical industry forward. More biological and medical knowledge is always needed as long as the humanity need better treatment and prevention. As we will discuss under "Future work" section later, it's very important to know the features that are mostly correlated with the disease severity if we want to use machine learning to manage pandemics.

1.2 Related Work

The medical literature is recently getting richer rapidly with new publications about COVID-19 research and projects. Some examples are [Large-Scale Multi-omic Analysis of COVID-19 Severity](#)[5], [SARS-CoV-2 in severe COVID-19 induces a TGF--dominated chronic immune response that does not target itself](#)[6] and [COVID-19 Virulence in Aged Patients Might Be Impacted by the Host Cellular MicroRNAs Abundance/Profile](#)[7] studies.

There are actually some papers that are continuously exploring topics that are very specific. E.g. in "[Hidden pandemic: COVID-19-related stress, SLC6A4 methylation, and infants' temperament at 3 months](#)"[8], the authors study the the changes in the mood of the infants of pregnant infected by COVID-19 and the relation between that and DNA-methylation.

From all studies, single and multi omics studies are extremely important because they unravel the causes behind the disease and empower the industry of vaccination and medication for better prevention and therapy.

It's true that many of the publications are enlightening different fields related to COVID-19 but they are either focusing their study on a single omic analysis or not providing their

results in a manner that makes them easily accessible e.g. as web services. Of course there are some exceptions like "[Large-Scale Multi-omic Analysis of COVID-19 Severity](#)" study[9] that provides the results as an interactive web app at <https://covid-omics.app>. What makes our project unique, is that it contains the results for a multi-omic Covid-19 analysis provided as an online open access resource that allows the user to explore single and cross-omic analysis using a very user-friendly interface. One other aspect that makes this project different is that we do not use an only one method for feature selection. Instead we use 4 different methods to build a feature selection consensus score for every feature. An exception is the DNA methylation data where we use only 3 different methods as we will discuss later.

1.3 Thesis Statement

This project presents a web application that provides analysis and visualization for metadata , single and multi-omic data from the CORSAAR project from Saarland state in Germany.

This following six omics are included in the project: transcriptome, proteome and peptidome, lipidome, metabolome, epigenome (DNA methylation) and miRNome.

As an online open access project, this web server gives the researchers around the world the ability to explore the data collected in the university hospital of Saarland and to cross check with the finding they have found in the data collected in the hospitals and medical centers in their regions. The simple interface and the clear visualization for the single and multiomic views increase the usability of the project by providing a user-friendly webserver.

For single omics, the app provides ultra fast searching, even for DNA methylation with more than 800,000 features. The features are also sortable by many fields including the feature selection score. Visualization by scatter-plots, bar-plots and box-plots makes it much easier to figure out the relationship between the feature and the severity of the disease. Annotated and interactively sorted and clustered heatmaps make them more useful for the researchers browsing the application. By conducting feature selection using the consensus of 4 different statistical methods and algorithms instead using only one like in many other studies, the users can feel more confident about how real the features are correlated with disease severity.

Finally, the "Pair-wise correlation scatter plot" is a feature provided by this project, since it allows to generate a scatter plot for any two different features from any two different omics colored according to the disease severity by admission or by the max. disease severity reached by the patients.

cross omics view is a user-friendly interface that provides simple controls to generate network analysis and to view the top correlated features from different omics.

1.4 Outline

The thesis consists of 8 chapters:

- *Chapter 2* explains the biological role played by different omics to motivate the use cases of multiomics studies including CORSAAR.
- *Chapter 3* discusses the methods, including the techniques, frameworks and libraries used in the web server.

- *Chapter 4* provides an overview of the project including the two targets of the study, metadata, feature selection and the web app homepage.
- *Chapter 5* describes the single omic analysis conducted on six different omics, starting with RNA-seq and ending with miRNome.
- *Chapter 6* describes the multiomics analysis features provided by the project.
- *Chapter 7* explain the two simple ways to get all data included in the project.
- *Chapter 8* discusses the results of the project and ideas for future work.

Chapter 2

Biological Background

This chapter explains the biological role played by different omics to motivate the use cases of multiomics studies including CORSAAR. Omics, including genome, transcriptome, proteome, epigenome and others are the different biological layers that drive the life inside an organism. They achieve this by the regulatory effect of each of them on a group of the other and therefore on the phenotype and disease.

2.1 RNA-seq

From the moment when the scientists became able to sequence a full human genome a new era for biomedical research started and scientists started working very hard to map the changes in the nucleotides to those in the phenotypes and disease. However, it was clear later that many of these changes can't be understood only according to the DNA variants. As stated above, the regulatory role played by other biological layers but the genome are very important for this purpose. The first layer where the researchers started to explore after the genome was the transcriptome (the part of the RNA that is transcribed into proteins)[10].

Many techniques had been used to study RNA-seq including hybridization-based microarrays and RNA-seq where the former had many issues, like needing to have knowledge about the sequence and therefore the inability to detect novel transcripts. On the other side, it's true that RNA-seq needs specialists from the technical field to process the data, but it saves lots of time, efforts[11] [12].

Gene expression, studied by RNA-seq, has great medical applications. For example, it can unravel the biology behind disease, support diagnosis by finding biomarkers for a wide-range of clinical phenotypes and be used in personalized medicine[12]. According to GENCODE[13] project, 244939 human transcripts are known till the date of writing this thesis.

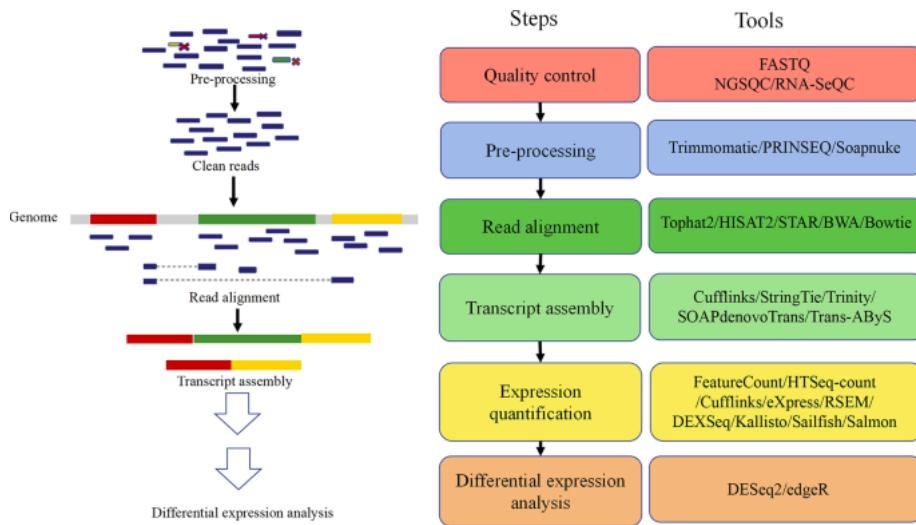


Figure 2.1 RNA-seq analysis steps and tools. Obtained from [14].

Figure 2.1 shows the subsequent steps of an RNA-seq analysis including quality control, preprocessing, read alignment, assembly, expression quantification and DEA with examples of the tools that can be used for each step.

2.2 Proteome

When talking about "Proteome", we mean the study of the group of the proteins translated from mRNA. This makes this layer the one that lays directly on the top of the transcriptome. In living organisms, proteins are the motors that drive the life inside the body and they form the signals that make cells interact with each other. One significant difference between proteome and transcriptome is that proteins are composed from combinations from twenty variable types of amino acids while transcriptions are composed from combinations from only four variable types of nucleic acids . In addition, proteins are vulnerable for changes after being translated[15].

The variability and quantity of proteins in a cell play a very important role in determining the age, characters and functions of it[15].

Proteins, affected by biological and environmental factors, play a mirror role that reflects the state of the organism in willness and disease. Adding this to what has been stated above makes conducting multi-omic studies that include proteomics as one of the major layers of a very high benefit since the researchers will be able to study the result of the interaction between proteomics and other layers [16].

Some medical applications for proteomics are their role as biological markers, diagnostics, vaccine production industry and tuning expression patterns thanks to the invaluable role they play as signals[17].

In "Single-Omic Analysis" chapter, we will mention some articles from the clinical literature where changes in the proteome (e.g. the complement elements) have been found to be highly related to disease severity and prognosis in Covid-19 patients.

A very high number of different proteins is expected to be found in the human body by time but the number of the protein entries in the Human Protein Reference Database is 30,047 at the date of writing this thesis[18].

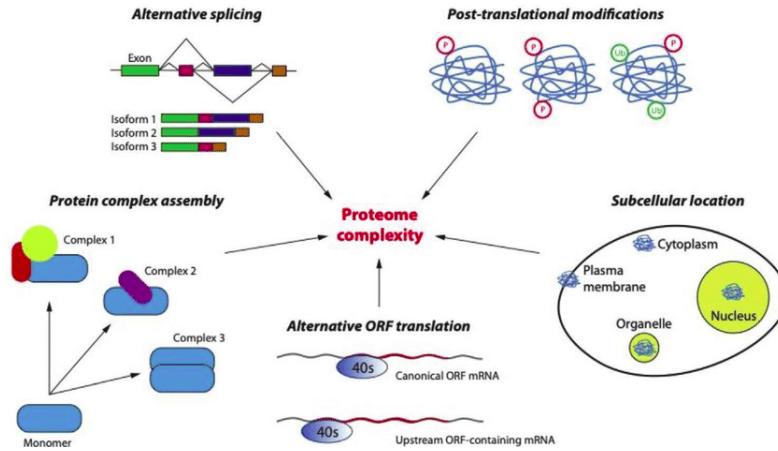


Figure 2.2 Proteins are widely-variable because of post-translational modifications, alternative splicing and other significant factors. Obtained from [19].

Figure 2.2 shows the mechanisms responsible for the high complexity of the proteome. This includes alternative splicing, post translational modifications (including miRNA suppression), diverse protein complex assemblies and different subcellular locations.

2.3 Lipidome

Changes in the levels of lipids have been observed in cancer, lymphocyte dysfunction and drug resistance[20][21].

For these observations and many others, clinicians are constantly monitoring the levels of cholesterol and triglyceride. They also tend to prescribe medication to lower these levels when elevated[22].

Moreover, researchers are exploring the results of detailed lipidomics analysis for their important role as a measurement for the efficacy of drugs and as diagnostic markers[22]. New technologies are helping lipidomics research to develop rapidly creating the opportunity to obtain more information about more lipids[22].

Many studies about lipids and Covid-19 suspect that the virus is making advantage of the formers as containers and basis to gather and combine the different components of the virus[23].

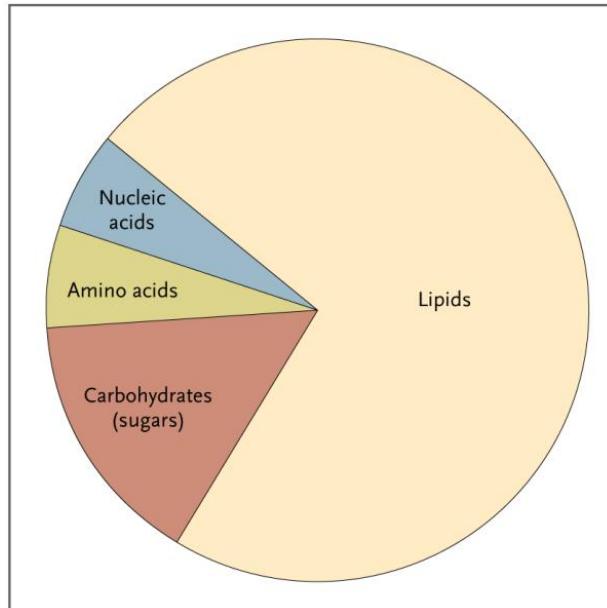


Figure 2.3 Relative distribution of biologic molecules in human plasma showing the high contribution of lipids to the regulation of our bodies. Obtained from [22].

Figure 2.3 shows how lipids contribute to more than 70% of the biological molecules in human plasma while carbohydrates, amino acids and nucleic acids are contributing for the rest.

2.4 Epigenome (DNA Methylation)

Epigenetics is the field of genomic sciences that studies the different mechanisms that alter and regulate gene expression. It focuses on the changes in the phenotype that are not related to those in the genotype in different cases of wellness and disease[24].

DNA methylation is a mostly steady layer of epigenomics that other layers lay on its top.[25]. For a specific cell type, the state of DNA methylation layer plays a significant role in determining its characteristics and is inherited through mitosis[26].

During DNA methylation in mammals, a methyl group is added to the C5 position of the nucleic acid cytosine transferring it to 5-methylcytosine. This leads to many changes including the charge of the nucleic acid and therefore the electrostatic forces between the DNA double helix and the histones. This is now a very well proved mechanism to increase or decrease the role of transcription factors and therefore the regulation of gene transcription. Methyl-binding proteins are the most important bridge between DNA methylation and another epigenomic layer called histone modification[27].

From all stated above, it's very clear how important it is to include as many as possible layers from epigenomics in multiomic studies.

In "Single-Omic Analysis" chapter, we will mention some articles from the clinical literature where changes in the DNA methylation layer have been found to be highly related to disease severity and prognosis in Covid-19 patients.

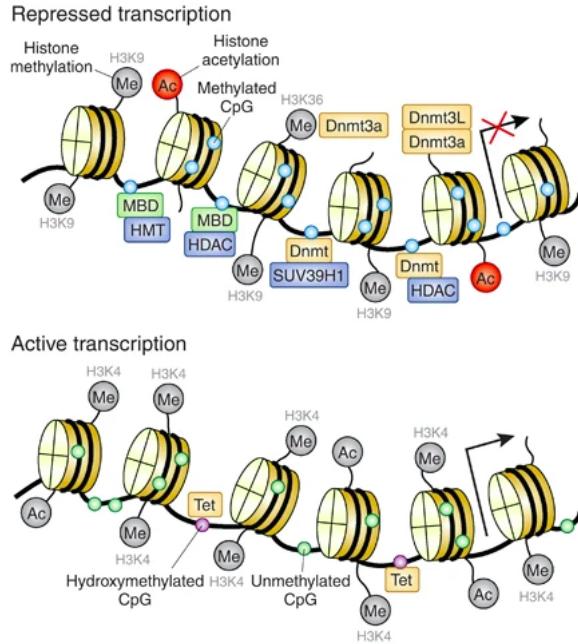


Figure 2.4 The role of DNA methylation in altering the chromatin state in different positions of the DNA double helix to regulate gene expression. Obtained from [27].

In the upper part of Figure 2.4 we notice the repressed transcription under the effect of DNA methylation. We can also notice DNA methylation writers attached to the chromatin. On the other hand, we can see the active transcription in the absence of methylation.

2.5 miRNome

While mRNA (messenger RNA) represents the coding part of the genome, microRNA represents one of the short non-coding RNAs that holds a great research interest. miRNAs play their role by applying post-transcriptional regulations. For example, miRNAs can play their regulation role through degradation of mRNA. Another common example is blocking the process of translation of mRNA[28].

Next generation sequencing and the rapidly developing bioinformatics techniques are working side by side with the traditional biologists to unravel more information about miRNAs and to make the best use of the obtained knowledge, especially in the medical field.[28].

The relationship between disease and changes in miRNome expression profiles became very clear as more studies were conducted. miRNAs can now be used as biomarkers that help with determining tissues characteristics[29] and monitoring multiple physiological changes in different types of cells [30].

Recent studies confirm that miRNAs can be used soon in the therapeutic field, e.g. as a treatment for the viral infection by affecting the viral genome[31]. For example, in Covid-19 patients, miRNAs proved their ability to block or reduce viral replication, viral binding to cellular receptors and the function of viral proteins[32].

We will come back to a brief discussion about the relationship between the miRNome and Covid-19 later in the "Single-Omic Analysis" chapter by exploring the data, searching for

the experimentally validated significantly correlated miRNAs and presenting Covid-19 articles related to them in the clinical literature.

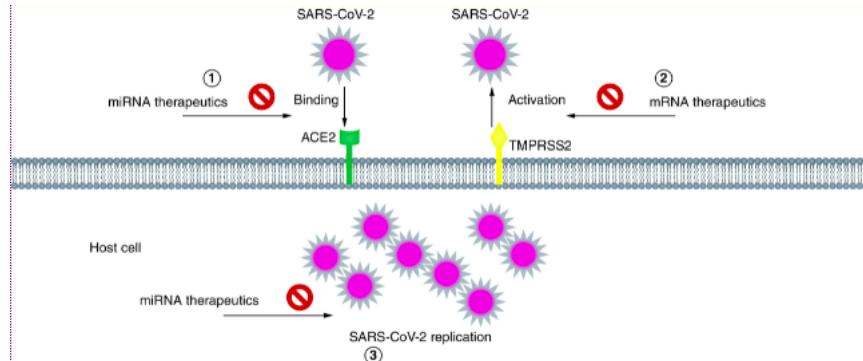


Figure 2.5 An illustration of miRNAs attacking coronavirus in different ways. Obtained from [31].

Figure 2.5 shows different ways in which miRNAs can be used to attack pathogens (in this SARS-CoV-2), e.g. It can block binding, activation and replication.

2.6 Integrative Omic Studies

"Proteomics is not an island", says the English poet John Donne. This quote was very inspiring for the authors of the article [Proteomics Is Not an Island: Multi-omics Integration Is the Key to Understanding Biological Systems](#)[16].

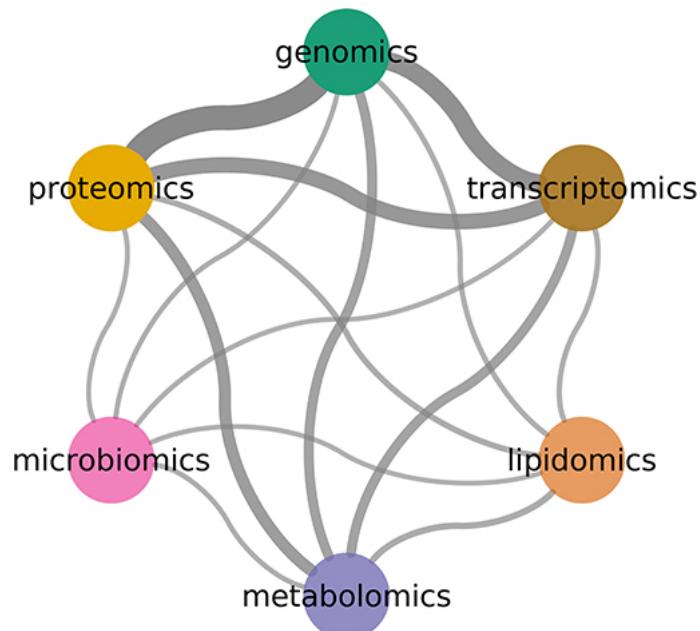


Figure 2.6 Some of the major different omics interacting with each other to regulate the organism. Obtained from [33].

Figure 2.6 illustrates the strong relationship between different omics (the nodes) and supports the motivation behind multi-omics studies. The edges between the omics represent the strength of the correlation between pairs of omics where the thicker the edge is, the higher the correlation is.

After we stated the significant role played by every single omic, it's should be clear how important it's to study the relationship between each pair or even between all of omics. This is not impossible anymore with the help of high throughput sequencing techniques and the very advanced bioinformatics algorithms available for use.[34].

So far, many large-scale multi-omic Covid-19 studies have been conducted. Some examples are: [Large-Scale Multi-omic Analysis of COVID-19 Severity \[5\]](#), [Multi-omic approach identifies a transcriptional network coupling innate immune response to proliferation in the blood of COVID-19 cancer patients \[35\]](#) and [Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19 \[36\]](#).

In some details, in [Large-Scale Multi-omic Analysis of COVID-19 Severity \[5\]](#), the authors focused on studying RNA-seq data, proteomics, metabolome, and lipidome and mapped the results to the clinical phenomena.

On the other hand, the authors in [Multi-omic approach identifies a transcriptional network coupling innate immune response to proliferation in the blood of COVID-19 cancer patients \[35\]](#) were more attracted to study the connection between transcription levels and the immunology system.

Finally, [Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19 \[36\]](#) is another paper that focused on the effect of COVID-19 on the immunology system.

In this project, we present CORSAAR web app as an online open-access integrative multiomic analysis tool for SARS-COV-2.

Chapter 3

Methods

3.1 Web App

3.1.1 Web server core

Our web app is available at <https://wwwccb.uni-saarland.de/corsaar> and it's built using a [dockerized](#) image of CookieCutter[37] with Debian Linux distribution as the OS. The full-stack web development platform used is [Django](#) (v3.2.2) where Python (v3.8.8) and R (v3.6.3) are the core back-end programming languages used. Front-end development was achieved using Jquery JS library and Django templating system. Anaconda[38] has been used as the trusted repository for all packages added to the webserver.

3.1.2 UI Theme

The web server uses the free version of [Gradient Able Bootstrap Lite](#) theme from [CodeThemes](#) (<https://codedthemes.com/>).

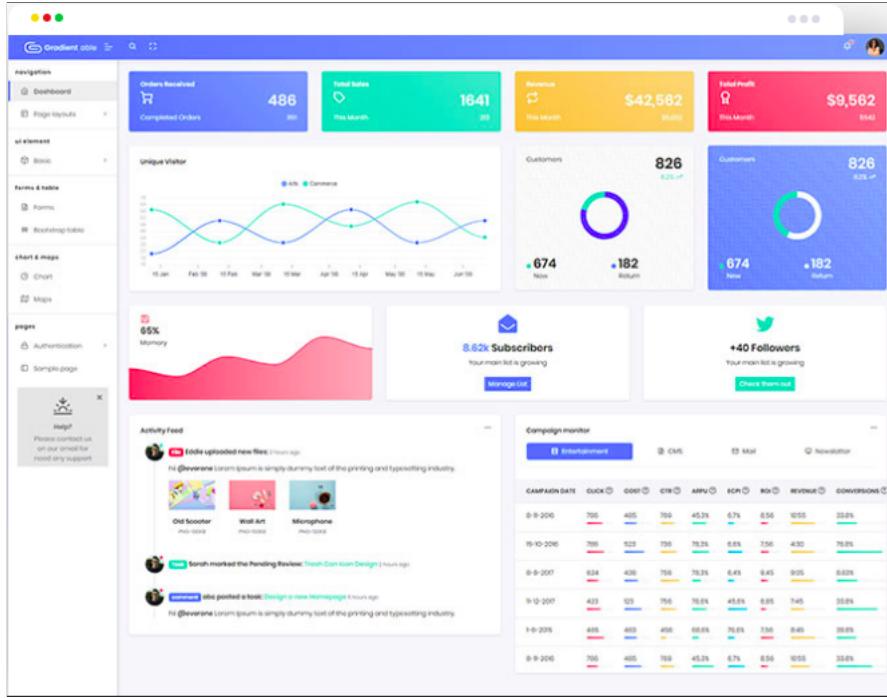


Figure 3.1.2 A screen shot for the dashboard of the Gradient Able theme from CodedThemes.

3.1.3 Visualization JavaScript Libraries

For visualization, two JavaScript libraries are used. These are: [HighCharts](#) (v9.1.0) and [ApexCharts](#) (v3.26.3).

3.1.4 Data Storage

All data is stored as .csv files. An exception is the DNA methylation data where .hdf (HDF5) files are used to achieve ultra fast search and retrieval queries. Since DNA methylation is written once and read multiple times, The decision to choose .hdf file format was made according to the information in the paper "Comparison of Array Management Library Performance - A Neuroscience Use Case"[39].

Chapter 4

Application Overview

This chapter presents some information about the key points that are common between different upcoming chapters and that are essential for every researcher who wants to be aware of how the study is conducted. This includes the targets of the study, the severity score used to describe the state of the patients and the statistical and machine learning methods used for feature selection. It will also give an idea about the charts that have been added to the homepage to highlight some information about the gender, mortality rates, status by admission and max. severity reached by the patients of interest.

4.1 The two targets of the study

The Analysis, feature selection and visualization are conducted under two targets: "Disease severity level by admission" (Severity by admission) and "Max. disease severity level reached by the patient" ("Max. severity reached"), where "Max. severity reached" could be "dead".

To switch between the two targets, simply click on the desired tab button of the target at the beginning of the page as displayed Figure 4.1.

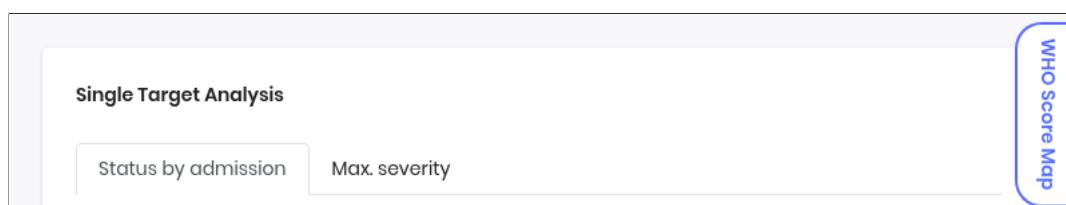


Figure 4.1 To switch between the two targets, simply click on the desired tab button of the target at the beginning of the page.

4.1.1 Severity Score and Mapping to WHO Score

In this study, we are using the severity score proposed by the University Hospital of Saarland. This score has five levels as the following: recovery (0), uncomplicated = outpatient (1), complicated = stationary (2), critical = ventilation (3) and dead (4). The following table in Figure 4.1.1.1 shows how the severity score proposed by the study is mapped to the severity score of WHO.

SEVERITY	PATIENT STATE WHO	DESCRIPTION	SCORE
Uncomplicated (outpatient)	Ambulatory mild disease	Asymptomatic, viral RNA detected	1
		Symptomatic; independent	2
		Symptomatic; assistance needed	3
Complicated (stationary)	Hospitalized: moderate disease	Hospitalized; no oxygen therapy	4
		Hospitalized; oxygen by mask or nasal prongs	5
Critical (ventilation)	Hospitalized: severe disease	Hospitalized; oxygen by NIV or high flow	6
		Intubation and mechanical ventilation, $pO_2 / FIO_2 > 150$ or $SpO_2 / FIO_2 > 200$	7
		Mechanical ventilation, $pO_2 / FIO_2 < 150$ ($SpO_2 / FIO_2 < 200$) or vasopressors	8
		Mechanical ventilation, $pO_2 / FIO_2 < 150$ ($SpO_2 / FIO_2 < 200$) and (vasopressors, dialysis, or ECMO)	9
Recovery		Recovery	0
Dead		Dead	10

Figure 4.1.1.1 Mapping to WHO severity score for COVID-19 patients.

While surfing the web app, it's always possible to display the mapping table by clicking on the floating button "WHO Score Map" to the right as displayed in Figure 4.1.1.2.



Figure 4.1.1.2 "WHO Score Map" floating button is reachable on every page of the app.

Clicking on this button will fire the mapping modal shown in Figure 4.1.1.3:

Mapping to WHO Score		
SEVERITY	PATIENT STATE WHO	DESCRIPTION
Uncomplicated (outpatient)	Ambulatory mild disease	Asymptomatic, viral RNA detected Symptomatic; independent Symptomatic; assistance needed
Complicated (stationary)	Hospitalized: moderate disease	Hospitalized; no oxygen therapy Hospitalized; oxygen by mask
Critical (ventilation)	Hospitalized: severe disease	Hospitalized; oxygen by NIV or ventilator Intubation and mechanical ventilation, $\text{PO}_2/\text{FiO}_2 > 150$ or $\text{SpO}_2/\text{FiO}_2 < 92\%$ Mechanical ventilation, $\text{PO}_2/\text{FiO}_2 < 150$ ($\text{SpO}_2/\text{FiO}_2 < 92\%$ or vasopressors) Mechanical ventilation, $\text{PO}_2/\text{FiO}_2 < 150$ ($\text{SpO}_2/\text{FiO}_2 < 92\%$ and (vasopressors, dialysis, circulatory support))
Recovery		Recovery
Dead		Dead

Figure 4.1.1.3 Clicking on "WHO Score Map" fires a modal that shows the mapping to WHO severity score.

4.2 Metadata

The case of each patient is provided with some metadata including: age, sex, status by admission, max. status reached, life/death state, the length of stay at the hospital (in days) and the length of stay at the intensive care unit (ICU) in days.

PATIENT	AGE	SEX	STATUS BY ADMISSION	MAX. SEVERITY	DEA
CSI000	54	male	critical = ventilation	critical = ventilation	aliv
CSI005	52	male	critical = ventilation	critical = ventilation	aliv
CSI006	51	male	critical = ventilation	dead	dec
CSI008	72	female	complicated = stationary	dead	dec
CSI009	52	male	complicated = stationary	complicated = stationary	aliv
CSI010	59	male	complicated = stationary	complicated = stationary	aliv
CSI011	55	female	complicated = stationary	complicated = stationary	aliv
CSI012	72	male	critical = ventilation	dead	dec
CSI014	87	male	complicated = stationary	dead	dec
CSI015	24	male	complicated = stationary	complicated = stationary	aliv

Showing 1 to 10 of 37 entries

Previous 1 2 3 4 Next

Figure 4.2.1 Main metadata table including different metadata information about each patient included in the study.

Metadata features can be numerical features or categorical features. In the first case (e.g. age), a scatter plot and a box plot chart for the values will be displayed. On the other hand, when the feature is a categorical feature (e.g. max. severity), then a barplot is used to visualize the data.

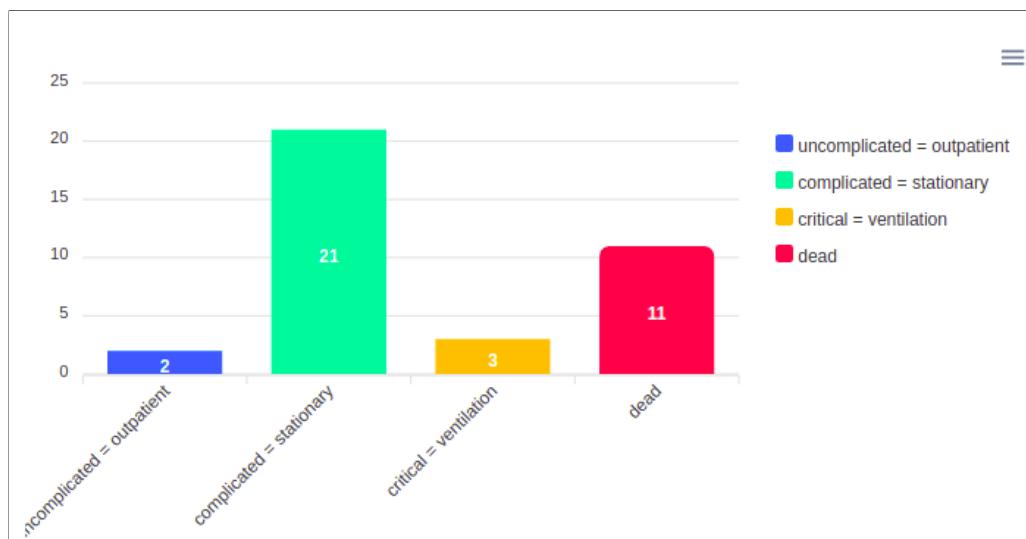


Figure 4.2.2 A bar plot is used to visualize the data of the categorical feature "Max. severity".

Figure 4.2.2 shows the numbers of the patients in every severity label bin. However, it's better to refer to the figures in the section 4.4 for a better comparison between counts between "at admission" and "max" severity levels.

4.3 Feature Selection

Since every omic has a different number of features, which can sometimes be very large (e.g. more than 800,000 feature in DNA methylation), it is very important to apply feature selection to attract researchers' attention to the features with the most significant effect on the targets of the study (severity by admission and max. severity reached).

However, full features are also available for those who are looking for data about specific features or ready to invest more time exploring more data to get better insights about COVID-19.

Every study has its one method to apply feature selection. In this study, feature selection is applied using consensus depending on the following different methods and algorithms: Spearman's correlation, Mutual information, RFE (Recursive feature elimination) and Anova. When two features have the same score of consensus, the one with the higher absolute value of Spearman's correlation has the priority.

Spearman's correlation is chosen in this study over Pearson correlation since the former is a rank-order based correlation which fits the biological data better than a linear correlation like in Pearson's. However, for RFE, the liblinear solver is used to provide the mix of feature selection supporters with a sense of linearity.

It's very easy to filter the features to display only the **Prime Features**. This is simply done by toggling the "Show only prime features?" switch in any single omic view as in Figure 4.3. A prime feature always has a star symbol in the "Prime" column in the features table.

The screenshot shows a user interface for feature selection. At the top, there are two tabs: "Status by admission" and "Max. severity". Below the tabs, a section titled "Single Feature Analysis" contains a switch labeled "Show only prime features?" which is checked. A green box below the switch contains the text: "What are Status by admission prime features? These are the features which have the highest effect on Status by admission according to the machine learning methods that we have applied." At the bottom of this section are "Copy" and "CSV" buttons. A table follows, with columns: FEATURE TYPE, PRIME, FEATURE, and FEATURE SELECTION SCORE. The table has three rows:

FEATURE TYPE	PRIME	FEATURE	FEATURE SELECTION SCORE
		<input type="text" value="Search"/>	
numerical	★	ADIPO_HUMAN (ADIPOQ)	2
numerical	★	ANXAI_HUMAN (ANXAI)	2

Figure 4.3 Filtering features using the "Show only prime features?" switch.

4.4 Homepage

The homepage includes only some statistical charts about the study. This includes a sex pie chart, a mortality pie chart, a "Sex - Death" bar chart, a status by admission pie chart and a max. severity pie chart.

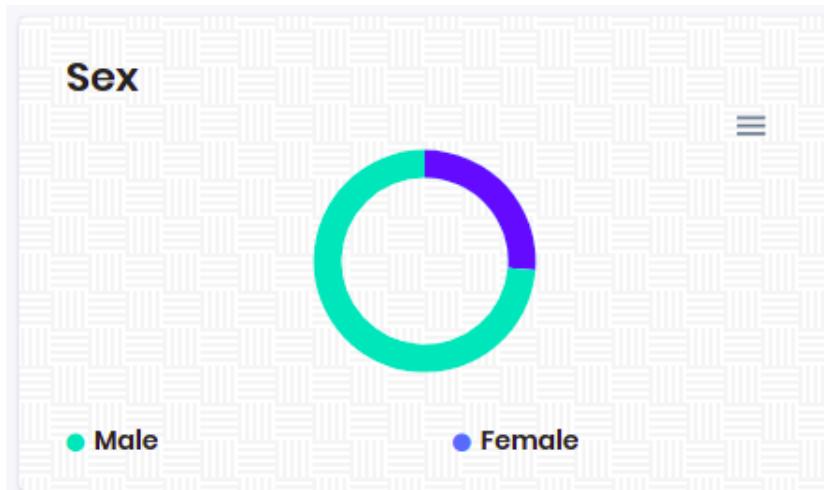


Figure 4.4.1 The sex pie chart. The study included 10 females vs 28 males.



Figure 4.4.2 The mortality pie chart. At the end of the study, 27 patients were alive while 10 patients died.

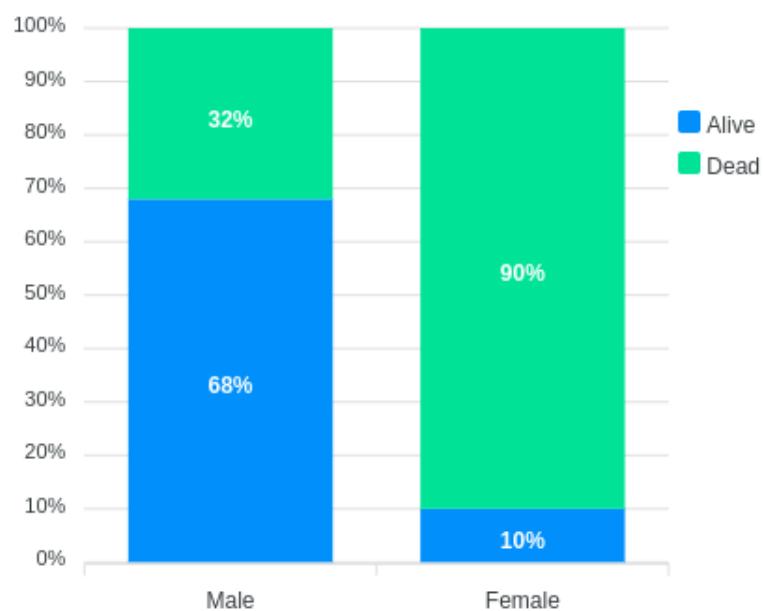


Figure 4.4.3 The "Sex - Death" bar chart. This chart shows the percentage of death and survival from both genders among the patients in this study.

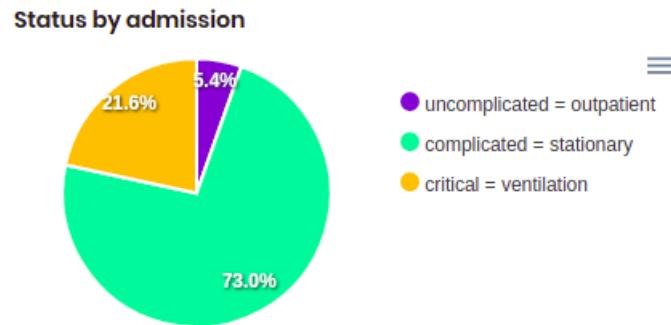


Figure 4.4.4 A pie chart that illustrates the severity levels by admission.

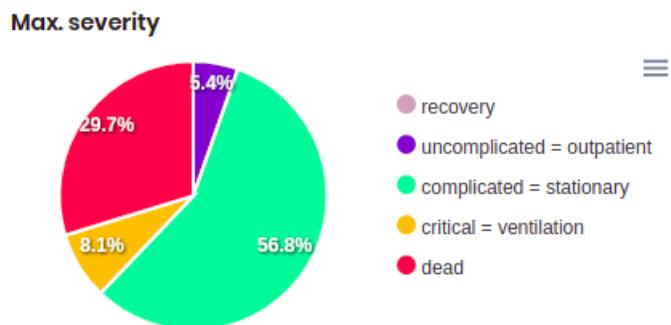


Figure 4.4.5 A pie chart that illustrates the max severity levels reached by the patients who are the subject of the study.

Chapter 5

Single-Omic Analysis

5.1 Transcriptome (RNA-seq)

5.1.1 General Information

RNA-seq is a technique that uses next-generation sequencing to study the protein-coding RNA (mRNA) which forms around 5-7% of total RNA content in mammalian cells[40]. It is one of the omics that gained researchers attention in a lot of studies since the expression values of the different genes play a very significant role in determining the state of the cell and the whole organism. Some related examples from the literature are [Transcriptomic landscape, gene signatures and regulatory profile of aging in the human brain](#)[41], [Transcriptomic landscape of blood platelets in healthy donors](#)[42] and [The Transcriptomic Landscape of Prostate Cancer Development and Progression: An Integrative Analysis](#)[43].

In this study, the transcriptome was studied by applying a bulk RNA-seq analysis on the whole blood samples taken from the patients.

The sequencing of the transcriptomics has been achieved by using the Illumina NovaSeq 6000 platform for bulk RNA-seq next generation sequencing.

After that, the raw data (.fastq.gz) went through preprocessing and analysis using the nf-core/rnaseq pipeline[44] [45]. This included FastQC for quality check, Trime Galore ! for trimming, Star Salmon for genome/transcriptome alignment and quantification, SAMTools for sort and index alignments and more[46].

For every patient, the transcriptomic data of 29,743 genes (features) were studied under the two targets of the study. For every feature, the features table contains the transcribed gene name, a column that indicates if the feature is a prime feature or not and the significance of the feature that has been assigned by the methods of feature selection.

Copy
CSV

FEATURE TYPE	PRIME	GENE	FEATURE SELECTION SCORE
Search			
numerical	★	CDC25A	3
numerical	★	S1PR5	2
numerical	★	FCRL6	2
numerical	★	LINC00035	2
numerical	★	H2AFB2	2
numerical	★	GPR84	2
numerical	★	RP4-779E11.3	2
numerical	★	GADD45G	2
numerical	★	PCP4L1	2
numerical	★	CNPY1	2

Showing 1 to 10 of 30 entries (filtered from 29,743 total entries)

Previous 1 2 3 Next

Figure 5.1.1 RNA-seq features table under target "severity by admission".

Figure 5.1.1 displays the first 30 out of 29,743 genes whose normalized expression values are stored and analyzed. The table includes the gene name, the feature selection consensus score and a column that indicates if the feature is a prime feature.

5.1.2 PCA (Principal Component Analysis)

The principal component analysis shows the feature GLT1D1 as the PC1 representing 14.8% of the variance and the feature SZT2 as the PC2 representing 10.7% of the variance. This figure shows the PCA scatter plot chart for the "severity by admission" target.

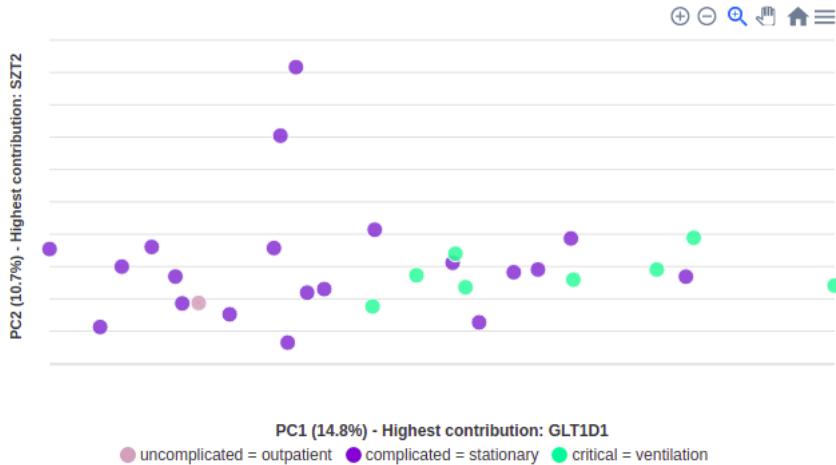


Figure 5.1.2.1 PCA scatter plot for "severity by admission" target. We notice how most of the cases labeled as "complicated" lay on the left side of the plot.

The PCA scatter plot in Figure 5.1.2.1 doesn't show any clear clustering nor leads to specific findings. However, we can find 2 outliers labeled as "complicated". On the other hand, this is the scatter plot for the PCA analysis under the "max. severity" target.

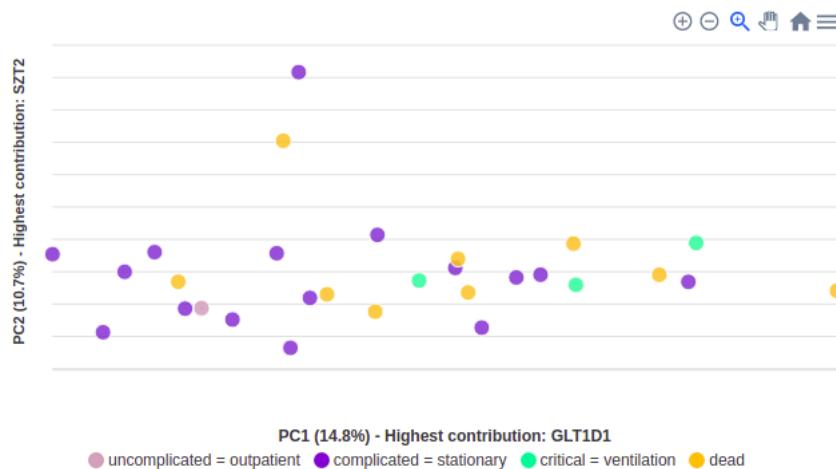


Figure 5.1.2.2 PCA scatter plot for "max. severity" target. No clear clustering is found.

The PCA scatter plot in Figure 5.1.2.2 also doesn't show any clear clustering nor leads to specific findings. We have one clear outlier related to the "complicated" group.

5.1.3 Differential Analysis

The differential analysis for RNA-seq data was conducted using the DESeq2[47] R package where the p-values adjustment method used was the FDR (Benjamini-Hochberg)

method.

For RNA-seq, an interactive clustered heatmap and volcano plots have been created.

Interactive Clustered Heatmap

This study uses Clustergrammer[48], the web-based heatmap visualization and analysis tool, to create interactive clustered heatmaps. Using this online tool, the samples can be annotated, the features can be re-ordered, the heatmap and the clusters can be resized and more. The following figure displays the DEA analysis heatmap created for RNA-seq. The rows of this heatmap represent the union of the top 30 prime features from each target of the study.



Figure 5.1.3.1 An interactive clustered heatmap created for transcriptomics using Clustergrammer online tool. The columns represent the patients and the rows represent the union of the top 30 prime features from the two targets of the study.

The rows in the heatmap in Figure 5.1.3.1 are labeled with different colors. Orange, green, purple refer to "By admission", "Max. severity" and "Both" consecutively. Relative High expression values are noticed for VNN1 and HP genes especially for patients labeled as "critical" by admission or "dead" as "max. severity".

Volcano Plots

Using the fold change values calculated by DESeq2[47] and the adjusted p-values, volcano plots are generated to compare two different levels of severity under each of the two different targets of the study. For a gene to be up / down regulated, its normalized expression value must have an adj-pval ≤ 0.05 and a fold change ≥ 1.5 or ≤ -1.5 consecutively.

According to DESeq2[47], applying fold change shrinkage gives better ranking and visualizing than filtering the data by using arbitrary gene count criteria. While generating the data for volcano plots, the shrinkage estimator "ashr"[49] is used for LFC shrinkage. This fold change shrinkage method is recommended more than the normal one by DESeq2[47] since it can filter out more noise while keeping most of the signal.

By choosing any two different levels of severity, a volcano plot is generated showing the up-regulated (in red) and down-regulated (in blue) genes.

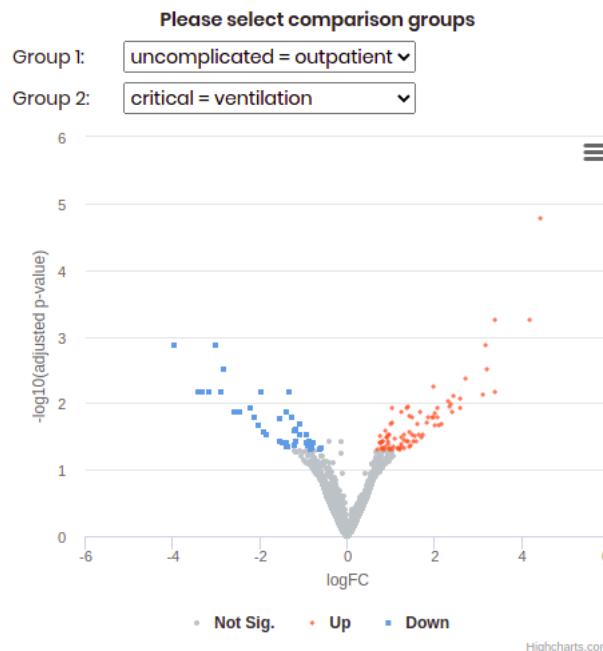
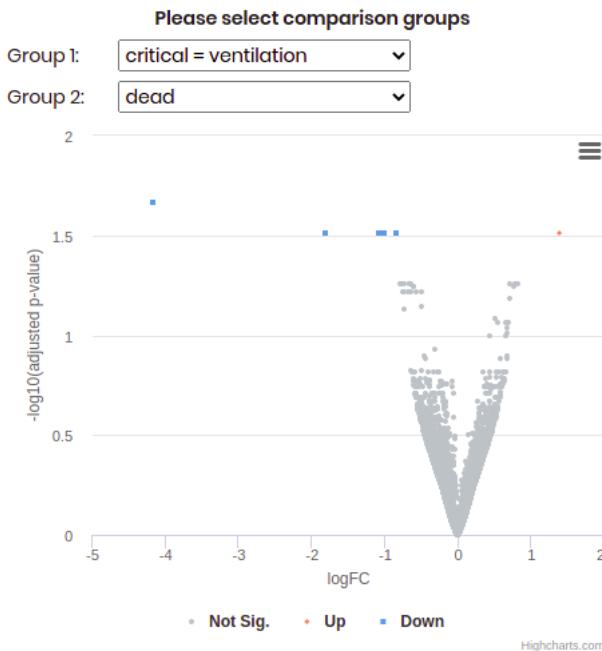


Figure 5.1.3.2 The volcano plot showing up- and down-regulated genes between "uncomplicated = outpatient" and "critical = ventilation" severity levels under the target "severity by admission".

When hovering over the dots in the plot above in the interactive web app (Figure 5.1.3.2), we find that FCRL6, GZMH and PPP2R2B are the top 3 upregulated DEGs, while FCRL1, COBLL1 and MS4A1 are the top 3 downregulated DEGs.

In the plot above, the most significantly up regulated gene is the FCRL6 gene which is a gene to which many pneumological and immunological diseases are associated, including Viral Pneumonia[50].

On the other hand, the most significantly down regulated gene is the IGJ gene which is very important for the immune system[50] and has been stated many times in the literature exploring the transcriptome of Covid-19. One example is the [SARS-CoV-2 in severe COVID-19 induces a TGF--dominated chronic immune response that does not target itself](#) article[6].



(Figure 5.1.3.3) The volcano plot showing some up-regulated genes (in blue) and one down-regulated gene (in red) in comparison between "critical = ventilation" and "dead" severity levels under the target "max. severity".

As we can see, we have only a small number of DEGs and when we use the interactive web interface, we find that GZMH is the only upregulated DEG, while MS4A1, POU2AF1 and FCRL1 are the top 3 downregulated DEGs.

5.1.4 Enrichment Analysis

We conducted GO (Gene Ontology) enrichment analysis using topGo[51] R package. From each severity level (under the two study targets) the 10 terms with the lowest Fisher's exact test p-value were picked to create the bar plot. For visualization, the negative log10 of the Fisher's exact test p-value has been used. It's important to add that only biological process terms have been added to this study.

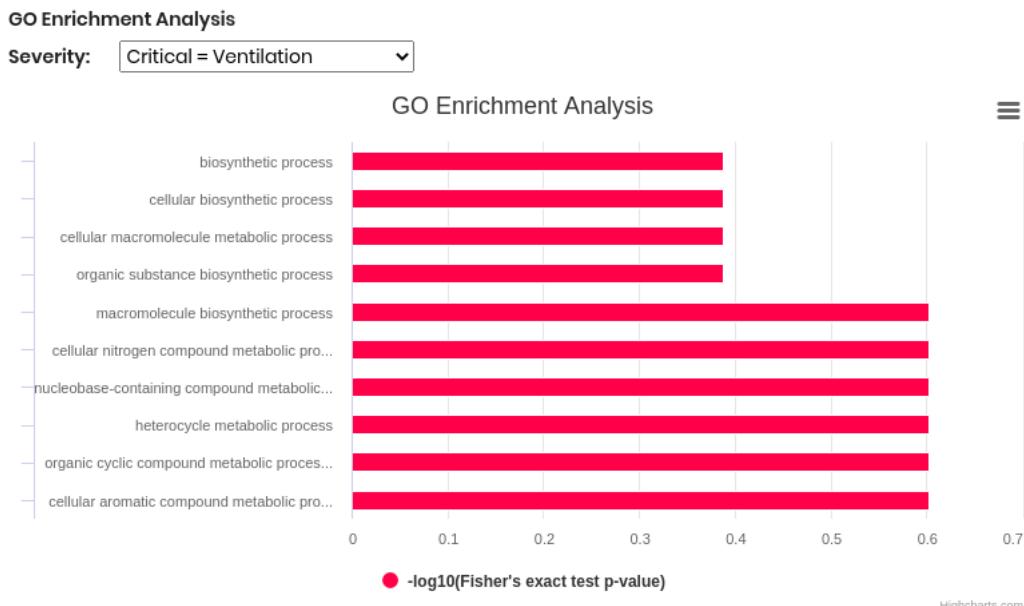


Figure 5.1.4.1 A bar plot showing the top 10 GO enriched biological process terms with their negative log₁₀ Fisher's exact test p-values for patients annotated as "critical = ventilation" as severity by admission level. These results are according to the output of the topGO[51] R package.

In Figure 5.1.4.1, we can find "biosynthetic process" in most of the top 10 enriched terms. This finding is supported by multiple articles from the clinical literature that can find on Pubmed[52]. For example, the article [Actively or passively deacidified lysosomes push -coronavirus egress](#) is talking about the role of biosynthetic secretory pathway used by SARS-CoV-2 virus to leave the cells.

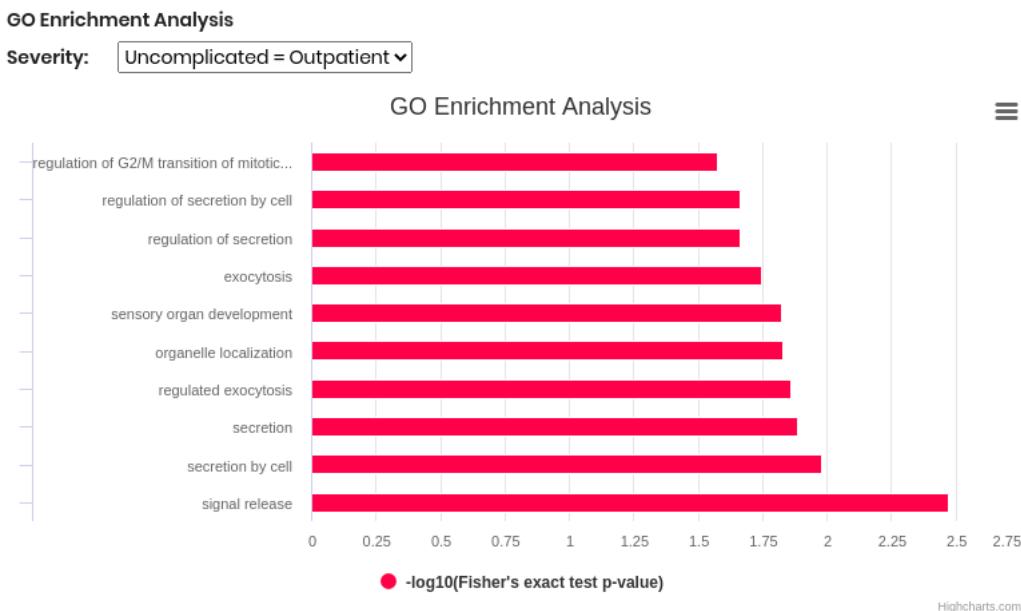


Figure 5.1.4.2 A bar plot showing the top 10 GO enriched biological process terms for patients annotated as "uncomplicated = outpatient" as severity by admission level. Most of these terms are associated to significant findings in many trusted journals with high impact factors.

In the plot above (Figure 5.1.4.2), we can find the term "secretion" in most of the top 10 enriched terms. Searching for COVID-19 and "secretion" at Pubmed[52] yields 44 results at the date of writing this thesis. These papers explore different topics including immunology, coagulation and vaccines.

5.1.5 Prime Features and Example Features

Prime Features

Keeping in mind that PCA relies on linear relationships, it gives a very important idea about the (n) components responsible for the most of the variance in the data. However, we are always interested in more features that are contributing significantly to the target of the study. This is especially important when the selected features (the ones called prime features in our app) are selected upon the contribution of different methods and algorithms for features selection.

The following 30 features have been selected by the consensus of our feature selection methods as the prime features of transcriptomics:

Feature	Consensus Score
CDC25A	3/4
GPR84	2/4
GADD45G	2/4
CNPY1	2/4
LINC00035	2/4
RP4-779E11.3	2/4
H2AFB2	2/4
PCP4L1	2/4
S1PR5	2/4
UGT2B15	2/4
FCRL6	2/4
TMEM38A	1/4
VNN1	1/4
MYBL2	1/4
GGH	1/4
SHROOM4	1/4
CCDC158	1/4
TFF3	1/4
PROM1	1/4
BEND7	1/4
FOXM1	1/4
LGR6	1/4
PPP2R2B	1/4
HP	1/4
DUSP13	1/4
ADAMTS1	1/4
ERG	1/4
RGL4	1/4
CCNB2	1/4
PGLYRP1	1/4

Table 5.1.5.1 Transcriptomics prime features under target "severity by admission".

Feature	Consensus Score
PCP4L1	2/4
FOXI3	2/4
CD164L2	2/4
LOC101928899	2/4
TMED10P1	2/4
YTHDF3-AS1	2/4
TRNAS13	2/4
DKFZP434F142	2/4
BVES	2/4
SNORA79	2/4
CECR3	2/4
TM4SF1	2/4
KCTD21-AS1	2/4
DYX1C1	1/4
C9orf84	1/4
ZNF670-ZNF695	1/4
IL19	1/4
VNN1	1/4
ACTA1	1/4
CNPY1	1/4
B4GALT2	1/4
ANGPTL4	1/4
DEPDC4	1/4
APOBEC3B	1/4
FLJ45079	1/4
TMEM38A	1/4
LOC102723402	1/4
PVRL2	1/4
MMP2	1/4
SCARNA14	1/4

Table 5.1.5.2 Transcriptomics prime features under target "max. severity".

Example Features

By Clicking on a feature of interest from the features table, a table displaying the **normalized** transcription counts will appear.

We will explore the transcription profile of CDC2FA which is a prime feature for the "severity by admission" target with a consensus score of 3/4 and explore some plots.

Feature: CDC25A		
Show 10 entries	Search:	
CASE ID (COSAAR ID)	VALUE	STATUS BY ADMISSION
CorSaar1	131092035116754	critical = ventilation
CorSaar10	22.8883271654513	critical = ventilation
CorSaar12	171008254258475	critical = ventilation
CorSaar13	2.64482199553768	complicated = stationary
CorSaar15	8.10991164627273	complicated = stationary
CorSaar16	2.70980707577899	complicated = stationary
CorSaar19	19.3216673239915	critical = ventilation
CorSaar20	2.80429344621555	complicated = stationary
CorSaar21	1.24086230425654	complicated = stationary
CorSaar23	19.8556261248977	uncomplicated = outpatient

Figure 5.1.5.1 The normalized transcription counts table for gene "CDC25A" some gene under target: severity by admission.

Figure 5.1.5.1 shows feature table for the feature CDC25A and it includes the normalized transcription counts for the gene and the severity by admission value for each patient. Since the transcription counts are numerical values (the feature is a numerical feature), a scatter plot and box plots will also be displayed.

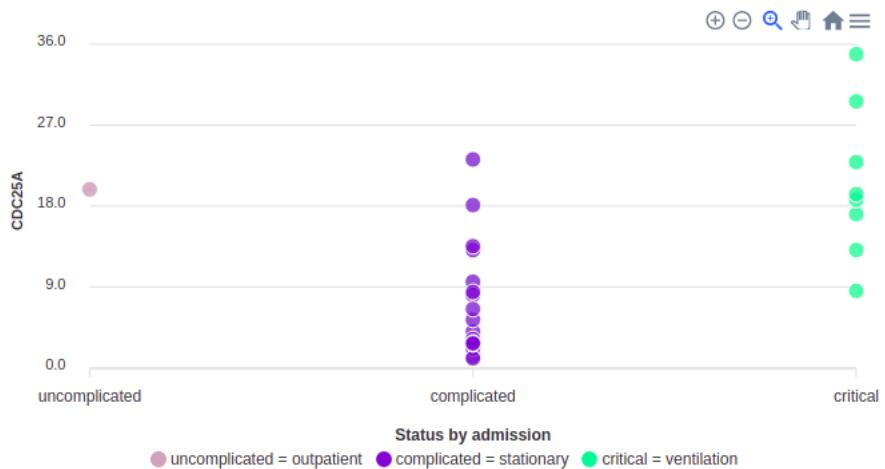


Figure 5.1.5.2 The scatter plot for the normalized transcription counts table for gene "CDC25A" some gene under target: severity by admission.

In Figure 5.1.5.2, we can clearly notice the increase in the normalized transcription values for the CDC25A gene when we compare different data points (patients) from "complicated" and "critical" severity.

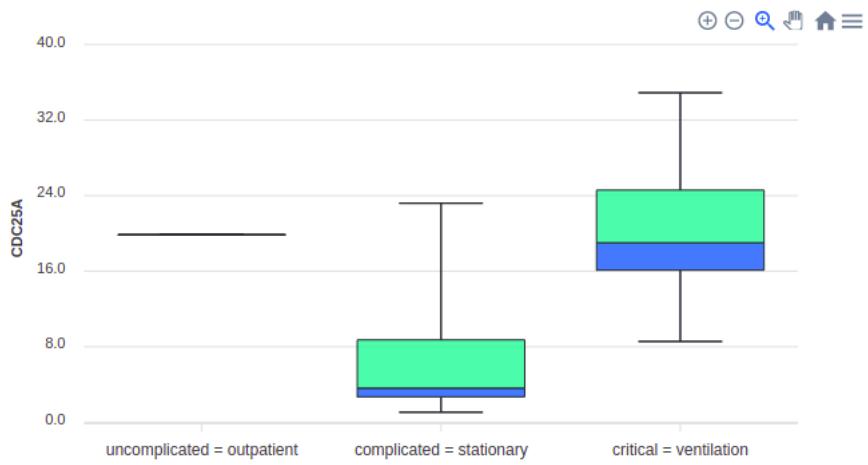


Figure 5.1.5.3 The box plot chart for the normalized transcription counts table for gene "CDC25A" some gene under target: severity by admission.

Figure 5.1.5.3 agrees with Figure 5.1.5.3 and it shows how the median and maximum normalized transcription values for the "critical" severity level under the target "severity by admission" are significantly higher than those for the "critical" severity level.

5.2 Proteome and Peptidome

5.2.1 General Information

The proteome and peptidome data was processed by PZMS (Präklinischen Zentrums für Molekulare Signalverarbeitung)[53] from plasma samples using nano LC-ESI-MS/MS (Ultimate 3000 RSLC nano system coupled to an LTQ Orbitrap Velos Pro (ThermoScientific, TF, Dreieich, Germany).

532 different proteins were studied from 19 patients included in the study. Unfortunately, it was not possible to study proteomics for all patients. The following figure, provided by Dr. Claudia Fecher-Trost from PZMS[53] - added with permission - , shows the process of sample preparation for proteome samples.

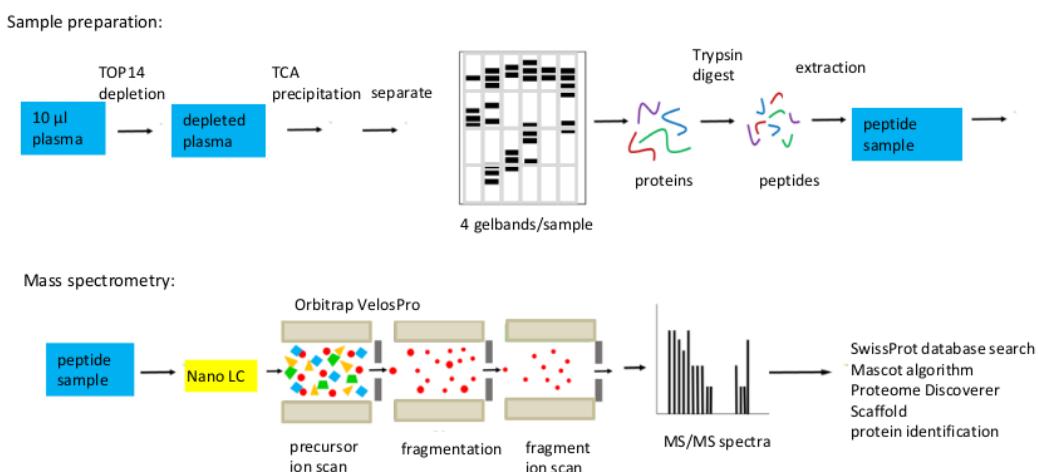


Figure 5.2.1.1 Proteomics sample preparation process. The figure is provided by Dr. Claudia Fecher-Trost from PZMS[53].

As illustrated in Figure 5.2.1.1, the process starts with sample preparation by taking 10 μl of plasma and applying TOP14 depletion to get the depleted plasma which is later precipitated using TCA and separated to form 4 gelbands sample. Then, we get the peptides from proteins through Trypsing digestion to finally extract peptide samples. The next step is passing the peptide samples to the mass spectrometry and to apply Nano Liquid Chromatography and continuing with many other steps that result in protein identification.

On the other hand, the following figure shows the process of sample preparation for peptidome samples. This illustration is also provided by Dr. Claudia Fecher-Trost from PZMS[53] and added with permission.

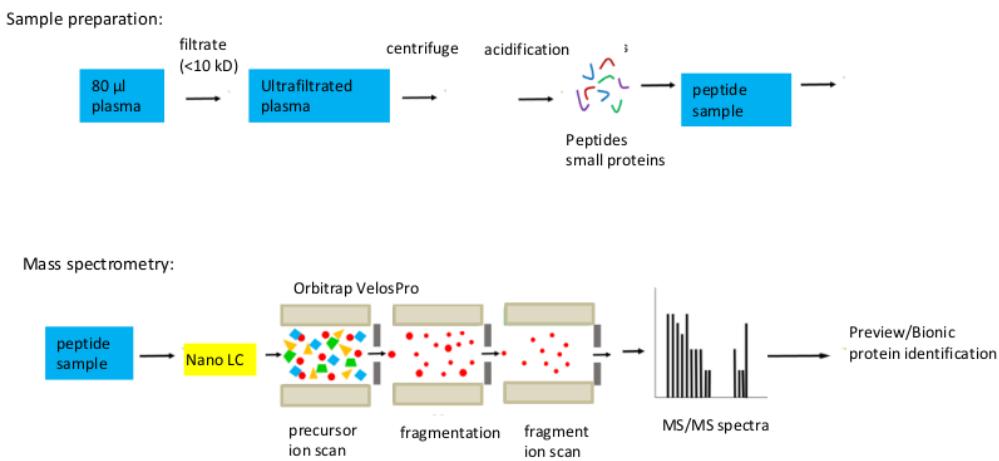


Figure 5.2.1.2 Peptidomics sample preparation process. This figure is also provided by Dr. Claudia Fecher-Trost from PZMS[53].

The process in Figure 5.2.1.2 is a little bit different from that in Fig 5.2.1. In peptidomics analysis, we start sample preparation by taking 80 μ l of plasma and filter those peptides < 10 kD. This results in the ultrafiltrated plasma that undergo centrifugation and acidification to result in the peptide samples.

When it comes to mass spectrometry, there are no differences between proteomics and peptidomics analysis except for the last step as illustrated in the figure.

The features data table provides the user with the protein name, a column that indicates if the feature is a prime feature or not and the significance of the feature that has been assigned by the methods of feature selection.

FEATURE TYPE	PRIME	FEATURE	FEATURE SELECTION SCORE
<input type="text" value="Search"/>			
numerical	★	FBLN1_HUMAN (FBLN1)	3
numerical	★	SAA4_HUMAN (SAA4)	3
numerical	★	C4BPB_HUMAN (C4BPB)	3
numerical	★	APOM_HUMAN (APOM)	3
numerical	★	APOA1_HUMAN (APOA1)	2
numerical	★	FA9_HUMAN (F9)	2
numerical	★	ADIPO_HUMAN (ADIPOQ)	2
numerical	★	ANXA1_HUMAN (ANXA1)	2
numerical	★	LIRA3_HUMAN (LILRA3)	2
numerical	★	PLTP_HUMAN (PLTP)	2

Showing 1 to 10 of 30 entries (filtered from 532 total entries)

Previous 1 2 3 Next

Figure 5.2.1.3 Proteomics features table under target "severity by admission".

Figure 5.2.1.3 displays the proteomics features table which includes multiple information about the feature (protein) under selected target including its name, its feature selection consensus score and and the prime column that indicates if the feature is a prime feature.

5.2.2 PCA (Principal Component Analysis)

The principal component analysis shows the TLN1_HUMAN (TLN1) protein as the protein with the highest contribution to the PC1 which represents 19.2% of the variance and the GSTP1_HUMAN (GSTP1) protein as the protein with the highest contribution to the PC2 which represents 15.8% of the variance. The following figure shows the PCA scatter plot chart for the "severity by admission" target. No clustering or important findings are clear in the plot.

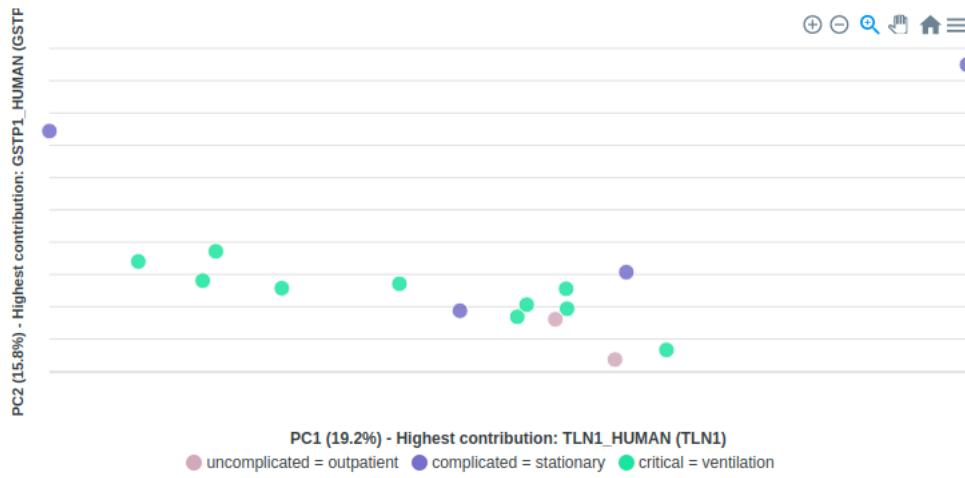


Figure 5.2.2.1 PCA scatter plot for "severity by admission" target with no clear clustering.

Figure 5.2.2.1 shows no clear clustering. However, we can notice two outliers labeled as "complicated".

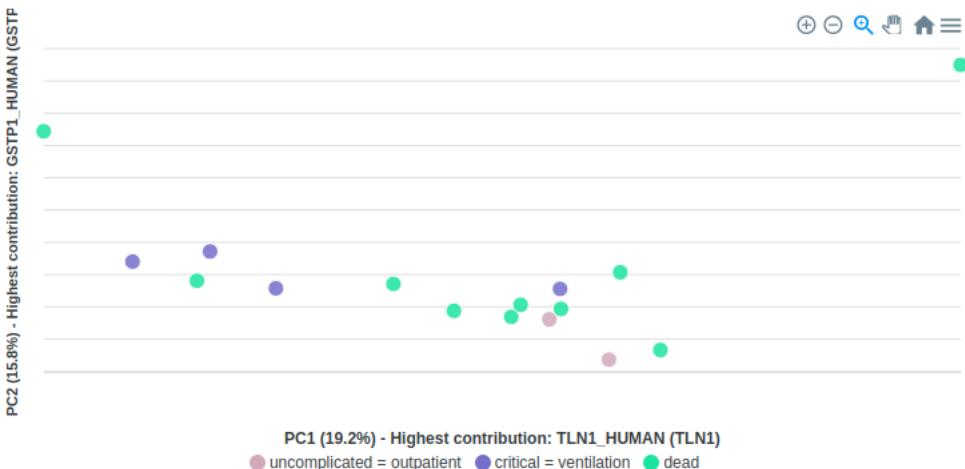


Figure 5.2.2.2 PCA scatter plot for "max. severity" target with data points showing no clear clustering.

Figure 5.2.2.2 shows no clear clustering. However, we can notice two outliers labeled as "dead".

5.2.3 Prime Features and Example Features

Prime Features

The following 30 features have been selected by the consensus of our feature selection methods as the prime features of proteomics:

Feature	Consensus Score
FBLN1_HUMAN (FBLN1)	3/4
SAA4_HUMAN (SAA4)	3/4
C4BPB_HUMAN (C4BPB)	3/4
APOM_HUMAN (APOM)	3/4
DEF1_HUMAN (DEFA1)	2/4
PLTP_HUMAN (PLTP)	2/4
FA9_HUMAN (F9)	2/4
PZP_HUMAN (PZP)	2/4
LIRA3_HUMAN (LILRA3)	2/4
ADIPO_HUMAN (ADIPOQ)	2/4
ANXA1_HUMAN (ANXA1)	2/4
GELS_HUMAN (GSN)	2/4
FCGBP_HUMAN (FCGBP)	2/4
C1QC_HUMAN (C1QC)	2/4
IGHG3_HUMAN (IGHG3)	2/4
C1QB_HUMAN (C1QB)	2/4
PGRP2_HUMAN (PGLYRP2)	2/4
SAMP_HUMAN (APCS)	2/4
APOA1_HUMAN (APOA1)	2/4
CBPN_HUMAN (CPN1)	2/4
APOF_HUMAN (APOF)	2/4
HGFL_HUMAN (MST1)	1/4
CO7_HUMAN (C7)	1/4
FBLN3_HUMAN (EFEMP1)	1/4
B2MG_HUMAN (B2M)	1/4
VASN_HUMAN (VASN)	1/4
C1R_HUMAN (C1R)	1/4
S10A8_HUMAN (S100A8)	1/4
MASP1_HUMAN (MASP1)	1/4
GSTP1_HUMAN (GSTP1)	1/4

Table 5.2.3.1 Proteomics prime features under target "severity by admission".

Feature	Consensus Score
ITIH1_HUMAN (ITIH1)	3/4
TGON2_HUMAN (TGOLN2)	3/4
MYL6_HUMAN (MYL6)	2/4
NP1L1_HUMAN (NAP1L1)	2/4
CH60_HUMAN (HSPD1)	2/4
EHD3_HUMAN (EHD3)	2/4
PTPRJ_HUMAN (PTPRJ)	2/4
TBA1B_HUMAN (TUBA1B)	2/4
PRDX2_HUMAN (PRDX2)	2/4
AMPN_HUMAN (ANPEP)	2/4
CO5_HUMAN (C5)	2/4
BIP_HUMAN (HSPA5)	2/4
TBA4A_HUMAN (TUBA4A)	2/4
RAB1B_HUMAN (RAB1B)	2/4
APOF_HUMAN (APOF)	2/4
1433G_HUMAN (YWHAG)	2/4
CD14_HUMAN (CD14)	2/4
TRY1_HUMAN (PRSS1)	2/4
FA12_HUMAN (F12)	2/4
GP1BA_HUMAN (GP1BA)	1/4
HEP2_HUMAN (SERPIND1)	1/4
GPV_HUMAN (GP5)	1/4
ATPB_HUMAN (ATP5F1B)	1/4
MYH9_HUMAN (MYH9)	1/4
MBL2_HUMAN (MBL2)	1/4
PARVB_HUMAN (PARVB)	1/4
CBG_HUMAN (SERPINA6)	1/4
EMIL1_HUMAN (EMILIN1)	1/4
RAP1B_HUMAN (RAP1B)	1/4
CALX_HUMAN (CANX)	1/4

Table 5.2.3.2 Proteomics prime features under target "max. severity".

Example Features

By clicking on a feature of interest from the features table, a table displaying multiple information about the selected protein will appear. For each of the studied patients, this table includes information about: the exclusive spectrum count, target value, peptides table button (if available), protein identification probability, exclusive unique peptide count, percentage of total spectra and the exclusive unique spectrum count. For better interpretation, you can refer to the quick guide [How to interpret results from shotgun MS analysis](#). According to this guide, "Exclusive Spectrum Count" refers to "The number of spectra associated with only a single protein group", "Exclusive Unique Peptide Count" refers to "The number of different amino acid sequences, regardless of any modification that are associated with a single protein group" and "Exclusive unique spectrum count" refers to "Number of distinct spectra associated only with a single protein group. Spectra are considered distinct when: i) they identify different sequences of amino acids or peptides; ii) they identify different charge states or a modified form of the peptide within the same identified sequences of amino acids". In all circumstances, a user can display

any of the descriptions stated above while surfing our web view by hovering over the help tooltip in the headers of the table columns.

For proteomics, we will choose the CO5_HUMAN (C5) protein which is a prime feature for the "max. severity" target. This protein is a very important complement component and is a target for many clinical studies about Covid-19. For example, a study titled [Complement C3 vs C5 inhibition in severe COVID-19: Early clinical findings reveal differential biological efficacy](#)[54] points to the significant decrease in the levels of this protein in patients with high severity levels of Covid-19. This is something we can notice clearly by viewing the plots and comparing the patients from the highest two severity level in the study.

Feature: CO5_HUMAN (C5)				
Search: <input type="text"/>				
CASE ID (COSAAR ID)	EXCLUSIVE SPECTRUM COUNT	MAX. SEVERITY	PEPTIDES	PROTEIN IDEN
CorSaar1	117	critical = ventilation	Q	100.0%
CorSaar10	112	critical = ventilation	Q	100.0%
CorSaar12	49	dead	Q	100.0%
CorSaar13	90	dead	Q	100.0%
CorSaar14	100	dead	Q	100.0%
CorSaar19	89	dead	Q	100.0%
CorSaar2	119	critical = ventilation	Q	100.0%
CorSaar20	52	dead	Q	100.0%
CorSaar23	5	uncomplicated = outpatient	Q	100.0%
CorSaar24	27	uncomplicated = outpatient	Q	100.0%

Showing 1 to 10 of 16 entries

Previous [1](#) [2](#) Next

Figure 5.2.3.1 The feature table for the protein "CO5_HUMAN (C5)" under target "severity by admission" showing multiple important information.

In Figure 5.2.3.1 we notice the feature table for the protein "CO5_HUMAN (C5)" under target "severity by admission" including the Case ID, the exclusive spectrum count, severity level under selected target, links to peptides tables, protein identification prob, exclusive unique peptide count, total spectra % and exclusive unique spectrum count. The following table explains the description of many column headers in the table above:

Column name	Description
Exclusive Spectrum Count	The number of spectra, associated only with a single protein group.
Exclusive Unique Peptide Count	The number of different amino acid sequences, regardless of any modification that are associated with a single protein group.
Exclusive Unique Spectrum Count	Number of distinct spectra associated only with a single protein group. Spectra are considered distinct when: i) they identify different sequences of amino acids or peptides; ii) they identify different charge states or a modified form of the peptide within the same identified sequences of amino acids.

Table 5.2.3.3 Brief descriptions for the columns headers used in the proteomics feature table.

Peptides Table: CorSaar1						
<small>② Click on the blue magnifier icon in the table above to display peptides details for specific case ID. You can refer to this article for more details about the features.</small>						
Show	10	entries	Search:			
PEPTIDE	OBSERVED M/Z	Z	OBSERVED (M+H)	CALC. MASS (M+H)	OFF-BY-X ERROR	MASS
R.EGVQKEDIPPADLSDQVPDTESETR.I	919.1012	3	1855.52891	1855.52904	0	-0.5
R.SEETKENEQFTVTAEGK.G	610.2902	3	1855.5256	1855.5256	0	0.2
R.SEETKENEQFTVTAEGK.G	928.4301	2	1855.52529	1855.5256	0	-1.4
R.SEETKENEQFTVTAEGK.G	928.4317	2	1855.52561	1855.5256	0	0.3
R.SEETKENEQFTVTAEGK.G	928.4319	2	1855.52566	1855.5256	0	0.5

Figure 5.2.3.2 The peptides table of the "CO3_HUMAN (C3)" protein for the CorSaar1 case (patient).

In Figure 5.2.3.2, we observe the different peptides associated with the selected protein in "CorSaar1" patient. The table includes multiple important information including Observed m/z, z, Observed (M+H), Calc.Mass (M+H), Off-By-X Error, Mass Error (PPM), Starting Position, Cleavage, |Log Prob| and the number of unique peptides. for more details about these terms, please refer to table 5.2.d.

The information used for visualization in proteomics is the "Exclusive Spectrum Count" and since it contains only numerical values a scatter plot and box plots will also be displayed after selecting the protein of interest.

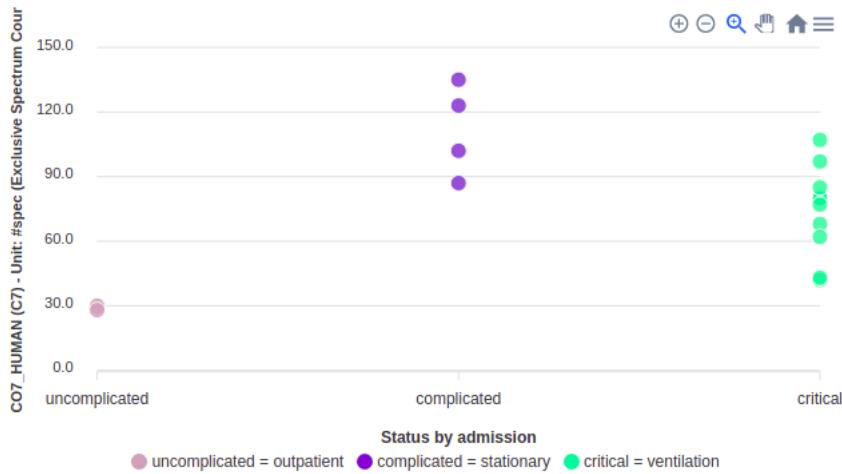


Figure 5.2.3.3 The scatter plot for the exclusive spectrum count of protein "CO7_HUMAN (C7)" under target "severity by admission".

Figure 5.2.3.3 shows how the exclusive spectrum count for "CO7_HUMAN" protein has lower values in patients labeled as "critical" in comparison with those labeled as "complicated" under the "severity by admission" target.

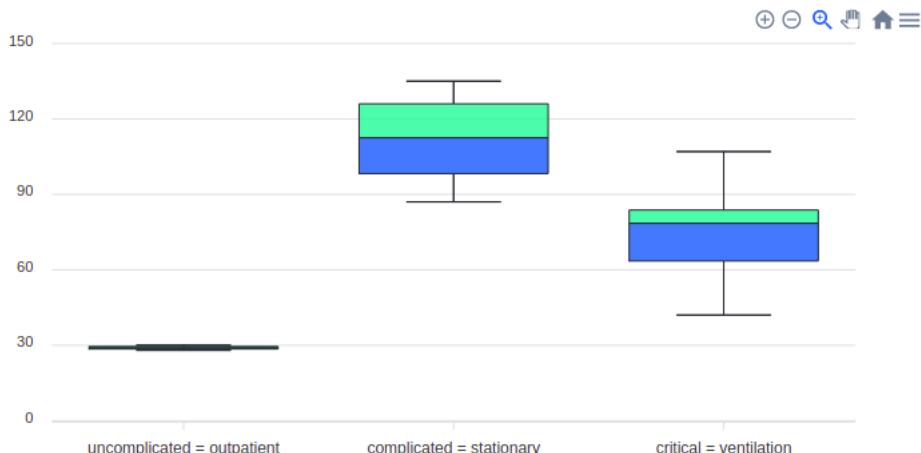


Figure 5.2.3.4 The box plot chart for the exclusive spectrum count of protein "CO7_HUMAN (C7)" under target "severity by admission".

It's clear Figure 5.2.3.4 supports the findings in Fig @@@ by comparing the medium, maximum and minimum values between "complicated" and "critical" severity levels under the same target. To have a look on the peptidome data, we will choose another protein which is CO3_HUMAN (C3) and then click on one of the "Peptides" buttons (the blue magnifier buttons) in the feature table of a protein, the peptides table for the selected protein of the case of interest will appear at the end of the page. This table provides information about the peptide, observed m/z, z, observer (m+h), calc.mass (m+h), off-by-x error, mass error (PPM), starting position, cleavage, absolute value of log prob and the number of unique peptides. For better interpretation, you can refer to the

article: [Byonic: Advanced Peptide and Protein Identification Software](#). According to this article, the following table provides brief descriptions for the column names used in the peptides table:

Column name	Description
Observed M/Z	Observed mass/charge.
Z	Charge.
Observed(M+H)	The observed (singly charged) precursor mass.
Calc.Mass(m+H)	The calculated (singly charged) precursor mass.
Off-By-X Error	"[M _{Observed} – M _{Computed}], where M _{Observed} is the observed M+H (singly charged) precursor mass and M _{Computed} is the computed M+H precursor mass, and [] means closest integer".
Mass Error(PPM)	" $10^6 \times (M_{\text{Observed}} - M_{\text{Computed}}) / (M_{\text{Computed}})$ ". The ppm mass error is computed after correcting for off-by-x errors".
Starting Point	"Position within the protein of the N-terminal residue of the peptide".
Cleavage	"Digestion specificity, where Specific means fully specific, Nragged means nonspecific at the N-terminus, Cragged (or Semi) means nonspecific at the C-terminus, and Non means nonspecific at both termini".
Log Prob	The absolute value of log base 10 of the protein p-value.
# of Unique Peptides	"Total number of PSMs (peptide-spectrum matches), discounting duplicates. (The same modification differently placed counts as a distinct PSM)".

Table 5.2.3.4 Brief descriptions for the columns headers used in the peptides table.

Peptides Table: CorSaar1

② Click on the blue magnifier icon in the table above to display peptides details for specific case ID.
You can refer to [this article](#) for more details about the features.

Show entries Search:

PEPTIDE	OBSERVED M/Z	Z	OBSERVED (M+H)	CALC. MASS (M+H)	OFF-BY-X ERROR	MASS
R.EGVQKEDIPPADLSDQVPDTESETRI	919.1012	3	2755.2891	2755.2904	0	-0.5
R.SEETKENEGFTVTAEGK.G	619.2902	3	1855.856	1855.8556	0	0.2
R.SEETKENEGFTVTAEGK.G	928.4301	2	1855.8529	1855.8556	0	-1.4
R.SEETKENEGFTVTAEGK.G	928.4317	2	1855.8561	1855.8556	0	0.3
R.SEETKENEGFTVTAEGK.G	928.4319	2	1855.8566	1855.8556	0	0.5

Showing 1 to 5 of 5 entries Previous Next

Table 5.2.3.5 The peptides table of the "CO3_HUMAN (C3)" protein for the CorSaar1 case (patient).

5.3 Lipidome

5.3.1 General Information

Starting from the moment when some clinicians noticed some relationship between medication that can affect lipid levels and between Covid-19 outcome[55], many bioinformaticians and other clinicians wanted to explore this relationship better.

Considering this as one of the significant motivations to integrate lipidome in our study, the lipidomics for 48 patients were obtained from the plasma samples and then processed by Helmholtz Institute for Pharmaceutical Research Saarland (HIPS)[56] using timsTop pro platform from BRUKER[®]. The processing started with 2857 MS buckets across 48 samples (3 replicates), 3 blank (3 replicates each) and a pooled QC sample (30 replicates). After blank removal and replicate comparison 1590 buckets were kept that met the two following conditions: not present in blanks and present in at least 2/3 replicates. After that, the mean intensity of each bucket was calculated for each sample.

5.3.2 PCA (Principal Component Analysis)

The principal component analysis shows the PC1 representing 11.1% of the variance with the highest contribution from the bucket "548.37_1.67" and the PC2 representing 9.3% of the variance with the highest contribution from the bucket "376.107_1.01". The following figures shows the PCA scatter plot chart for the two targets of the study.

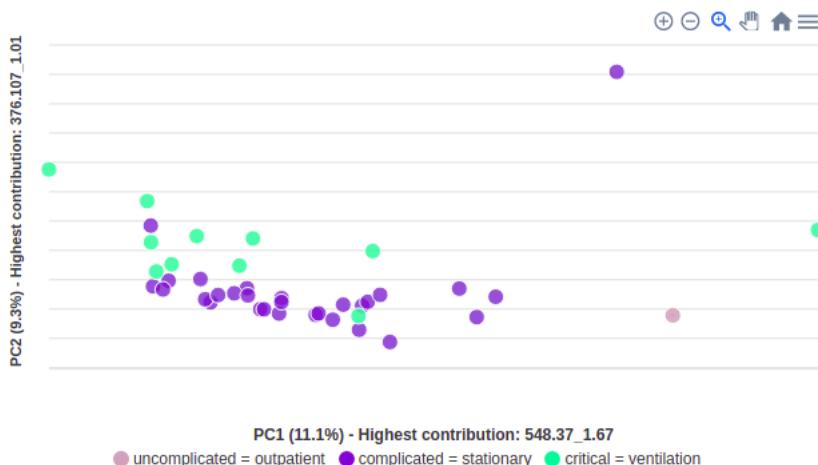


Figure 5.3.2.1 PCA scatter plot for "severity by admission" target. The majority of the points from "complicated" and "critical" severity levels lay on different levels of the y-axis.

Figure 5.3.2.1 shows nice clustering for each of the "complicated" and "critical" groups. It also shows some outliers from different severity levels groups under "severity by admission" target.

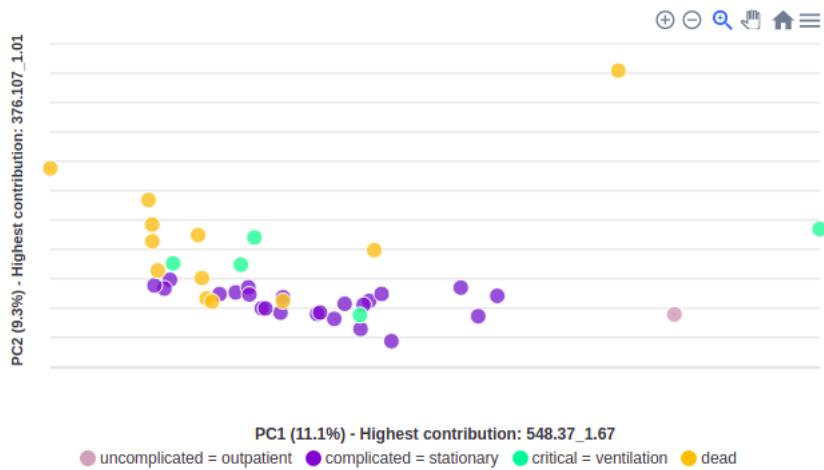


Figure 5.3.2.2 PCA scatter plot for "max. severity". It's clear how the points for the most of the patients who died at the end of the study are clustering well on this plot.

We can't see clear clustering in Figure 5.3.2.2 for the different severity level groups under this target and we have some outliers from different groups.

5.3.3 Differential Analysis

Interactive Clustered Heatmap

An annotated interactive clustered heatmap for lipidomics was created using Clustergrammer[48]. The following figure displays the DEA analysis heatmap created for lipidomics after zooming to show some features clearly.

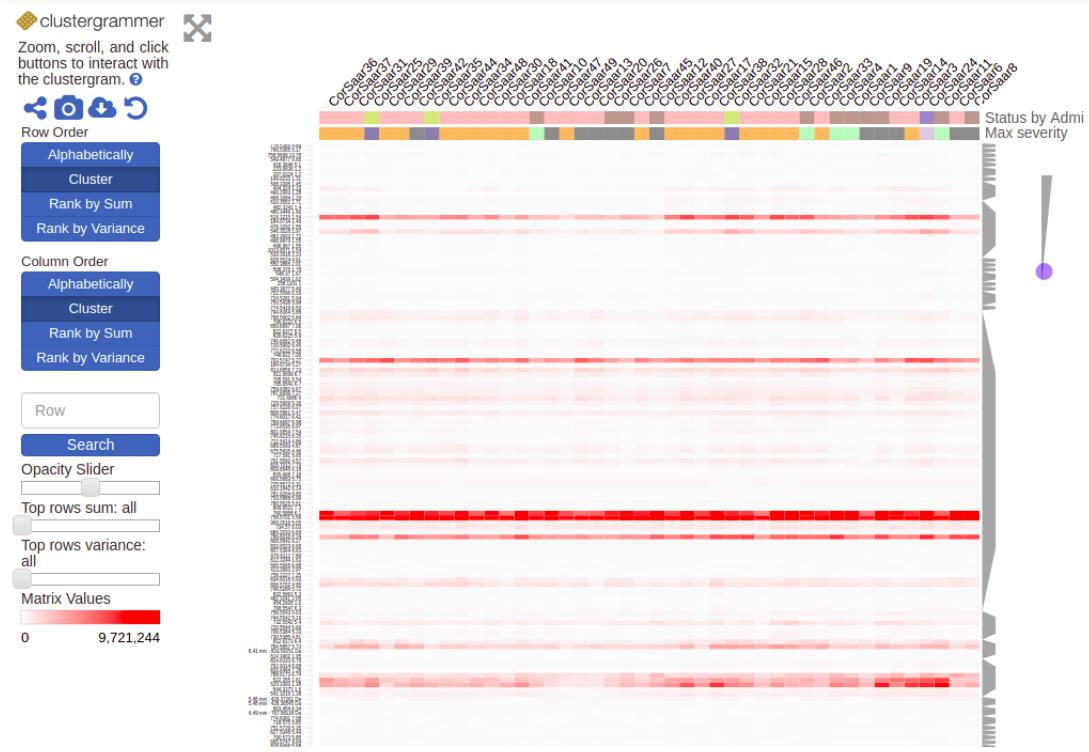


Figure 5.3.3 An annotated interactive clustered heatmap created for lipidomics using Clustergrammer online tool. Some zooming is applied so features' names become a little bit clear.

By exploring the annotation using the interactive heatmap displayed in Fig 5.3.3, we can't find a strong relationship between the mean intensity of the bucket and the target.

5.3.4 Prime Features and Visualization

Prime Features

The following 30 features have been selected by the consensus of our feature selection methods as the prime features of proteomics:

Feature (Bucket ID)	Consensus Score
315.1339_1.02	3/4
664.6034_9.51	3/4
757.6226_6.07	3/4
370.2953_1.14	3/4
785.6541_6.7	3/4
344.2795_1.1	3/4
870.5246_4.92	3/4
744.5903_6.39	3/4
900.8046_9.89	2/4
744.5892_6.39	2/4
874.7887_9.88	2/4
688.6035_9.32	2/4
846.757_9.58	2/4
876.8044_10.14	2/4
830.5684_5.69	2/4
848.7727_9.87	2/4
385.1161_1.02	2/4
283.1078_1.01	2/4
850.7885_10.15	2/4
711.5414_4.86	2/4
659.5018_7.36	2/4
765.5135_4.45	2/4
693.559_9.31	2/4
342.2639_1.05	2/4
415.3571_5.11	2/4
808.5857_5.68	2/4
811.6699_6.7	2/4
902.8205_10.15	1/4
668.6347_10.13	1/4
729.5909_5.36	1/4

Table 5.3.4.1 Lipidomics prime features under target "severity by admission".

Feature (Bucket ID)	Consensus Score
830.5684_5.69	3/4
780.5916_5.96	3/4
510.3918_2.23	3/4
766.5747_5.35	3/4
342.2639_1.05	3/4
870.5246_4.92	3/4
820.6229_6.32	2/4
688.6035_9.32	2/4
664.6034_9.51	2/4
768.5902_5.84	2/4
659.5018_7.36	2/4
344.2795_1.1	2/4
693.559_9.31	2/4
480.3446_1.66	2/4
524.3709_2.58	2/4
639.4961_6.35	2/4
794.6064_5.88	2/4
1013.6571_1.54	2/4
510.3552_1.71	2/4
281.0511_1.41	1/4
632.6343_8.4	1/4
572.3695_1.56	1/4
521.3812_4.72	1/4
796.6222_6.5	1/4
744.5903_6.39	1/4
808.5861_5.47	1/4
744.5892_6.39	1/4
757.6226_6.07	1/4
383.3156_2.75	1/4
818.6046_6.5	1/4

Table 5.3.4.2 Lipidomics prime features under target "max. severity".

Example Features

By clicking on a feature of interest from the features table, a bucket-specific table will appear. For each of the studied patients, this table includes information about the mean intensity value for the bucket and the target value for each of the studied patients. Since the mean intensity value is a numerical value, the visualization of lipidomics features includes a scatter plot and a box plot chart for each feature under the two targets of the study.

As an example, we will pick the bucket with ID 870.5246_4.92 which is a prime feature for the "max. severity" target.

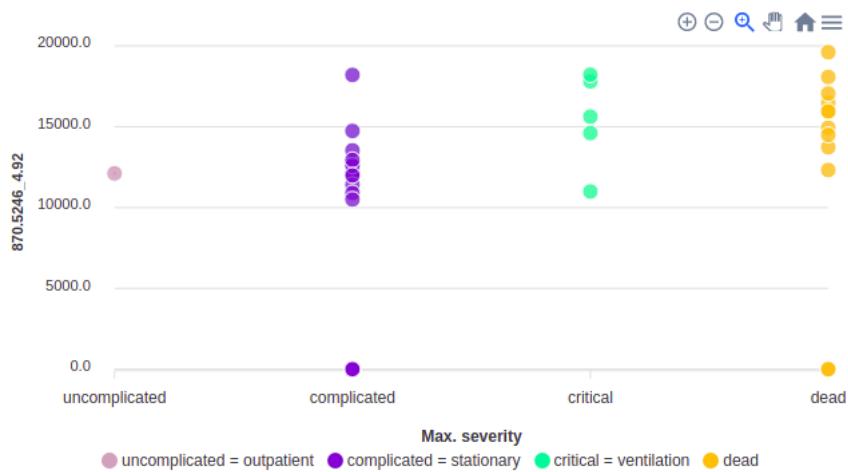


Figure 5.3.4.1 The scatter plot for the mean intensity values of the bucket with ID: "870.5246_4.92" under target "max. severity".

For the selected bucket, Figure 5.3.4.1 shows a weak positive relationship between the mean intensity values and the severity level under the selected target.

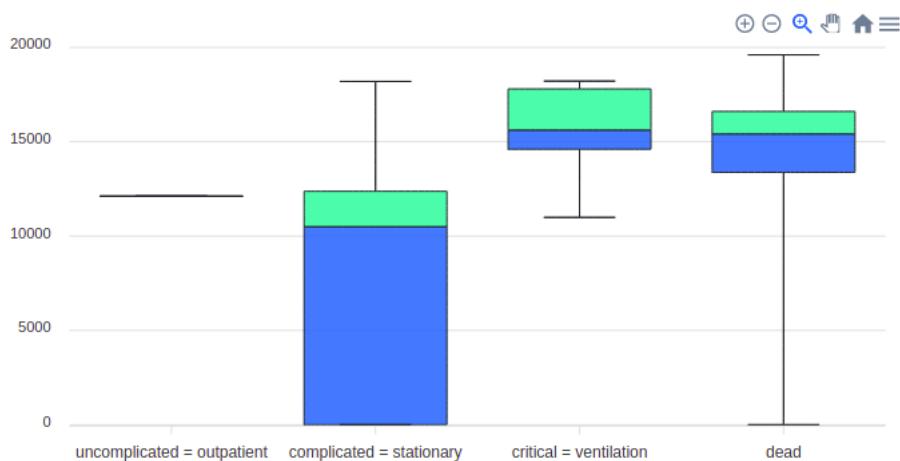


Figure 5.3.4.2 The box plot chart for the mean intensity values of the bucket with ID: "870.5246_4.92" under target "max. severity". We can notice how the median average intensity for this lipid and the severity level of the disease have a direct relationship.

Figure 5.3.4.2 shows a clear difference in the median mean intensity values for patients labeled as "complicated" and those labeled as "critical" or "dead" under the selected target. Unfortunately, the maximum and minimum don't lead us to clear observations. From what has been stated above and by exploring more features from lipidome and their relationship with the outcome of the disease, it's very important to start thinking about the way in which we can control the lipids levels in a Covid-19 patient to give him / her a better chance of living. What is the best single or combined medication that can cover more effected lipids? This question should be about the best medication for each patient when talking from the perspective of the personalized medicine.

5.4 Epigenome - DNA Methylation

5.4.1 General Information

DNA methylation is the lowest layer of epigenomics. It plays an important role in gene regulation by repressing gene transcription. This happens when a methyl group is added to Cytosine at C5 position changing the charge and the structure of the double helix of the DNA[57].

The processing of the DNA methylation layer from epigenomics was carried out by the lab of Prof. Dr. Jörn Walter at the university of Saarland. We studied 805,779 DNA methylation positions using the Infinium MethylationEPIC Kit from Illumina[®].

For each of the 805,779 features, the position ID, the chromosome number, the start position, the end position and the strand data is included.

Copy CSV							
FEATURE TYPE	PRIME	ID	CHROMOSOME	START	END	STRAND	FEATURE SELECTION SCORE
		<input type="text" value="Search"/>	<input type="button" value="All"/>	<input type="button" value="Search"/>	<input type="button" value="Search"/>	<input type="button" value="Search"/>	<input type="text" value="Search (>=)"/>
numerical	★	cg08691479	chr11	9595265	9595266	-	2
numerical	★	cg06758649	chr12	107381105	107381106	-	2
numerical	★	cg13826167	chr15	65203855	65203856	-	1
numerical	★	cg14256511	chr12	123347684	123347685	-	1
numerical	★	cg12743270	chr19	3028879	3028880	+	1
numerical	★	cg21725888	chr6	43253015	43253016	+	1
numerical	★	cg26289012	chr17	3540036	3540037	-	1
numerical	★	cg13319975	chr6	146136371	146136372	+	1
numerical	★	cg21287517	chr11	73498860	73498861	+	1
numerical	★	cg24218995	chr8	67873134	67873135	-	1

Showing 1 to 10 of 30 entries (filtered from 805,779 total entries)

Previous 1 2 3 Next

Figure 5.4.1 DNA methylation features table under the target "severity by admission".

Figure 5.4.1 shows the DNA methylation features table including many important information about every methylation position. This includes the chromosome, the start, the end, the strand, the feature selection consensus score and if the feature is a prime feature. All columns are searchable and many of them are sortable. Results for search queries are delivered in milliseconds thatnks to the HDF5 data format used to store DNA methylation data.

5.4.2 PCA (Principal Component Analysis)

The principal component analysis shows the PC1 representing 17.7% of the variance with the DNA methylation position with ID "cg14939082" as the best contributor and the PC2 representing 13.1% of the variance with the position with ID "cg12742294" as the best contributor. The following figures show the PCA scatter plot chart for both targets of the

study.

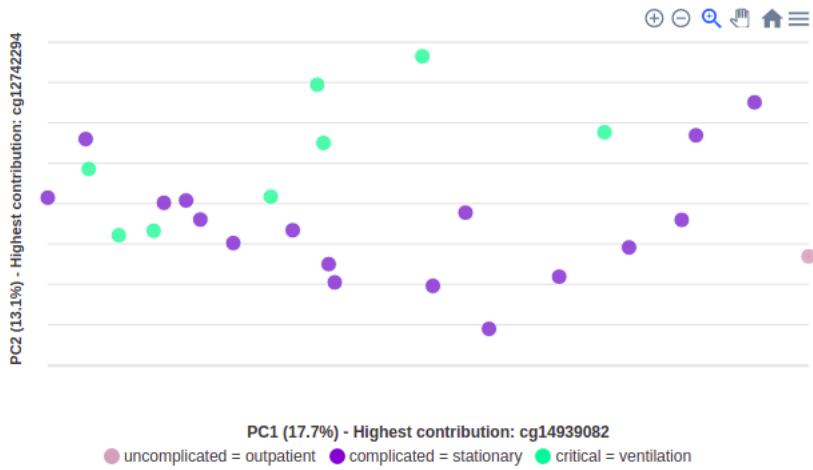


Figure 5.4.2.1 PCA scatter plot for "severity by admission".

Figure 5.4.2.1 shows no clear clustering or significant observation.

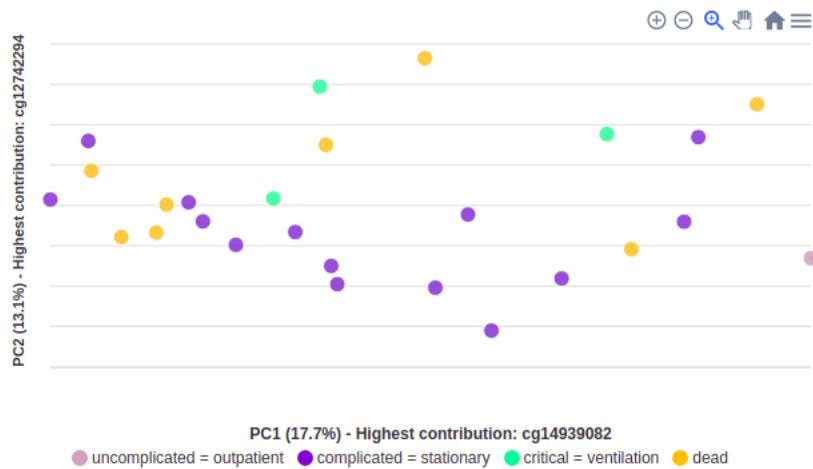


Figure 5.4.2.2 PCA scatter plot for "max. severity".

As in Figure 5.4.2.2, we notice that Fig @@@ also shows no clear clustering. It doesn't show significant observations, too.

5.4.3 Prime Features and Example Features

Prime Features

The following 30 features have been selected by the consensus of our feature selection methods as the prime features of DNA methylation. It's important to add that only three

feature selection supporters have been used for this omic. These are Spearman's correlation, ANOVA and mutual information. It was not possible to add support from recursive feature elimination (RFE) since the former took very long running on high computational resources without being able to finish the computation. This is quite predictable since DNA methylation data includes values from more than 800,000 features per patient.

Feature	Consensus Score
cg08691479	2/3
cg06758649	2/3
cg22046830	1/3
cg17895870	1/3
cg12195820	1/3
cg19377127	1/3
cg05160915	1/3
cg24039631	1/3
cg13826167	1/3
cg26289012	1/3
cg22100228	1/3
cg14256511	1/3
cg05026462	1/3
cg15500862	1/3
cg12850793	1/3
cg21725888	1/3
cg21287517	1/3
cg12598048	1/3
cg07248136	1/3
cg04174309	1/3
cg13319975	1/3
cg00602655	1/3
cg12743270	1/3
cg22630433	1/3
cg08125401	1/3
cg07359215	1/3
cg10570662	1/3
cg22696073	1/3
cg24218995	1/3
cg24413235	1/3

Table 5.4.3.1 DNA methylation prime features under target "severity by admission".

Feature	Consensus Score
cg09070855	1/3
cg23274883	1/3
cg15840554	1/3
cg05730122	1/3
cg14284257	1/3
cg14212967	1/3
cg02837122	1/3
cg09958090	1/3
cg08145231	1/3
cg22202031	1/3
cg18951734	1/3
cg23689457	1/3
cg16581179	1/3
cg26266046	1/3
cg03237071	1/3
cg10743424	1/3
cg13573892	1/3
cg06049528	1/3
cg24169995	1/3
cg15484899	1/3
cg02867373	1/3
cg02340576	1/3
cg09217522	1/3
cg02734661	1/3
cg18651727	1/3
cg08931968	1/3
cg13826167	1/3
cg14329508	1/3
cg14827499	1/3
cg21325232	1/3

Table 5.4.3.2 DNA methylation prime features under target "max. severity".

Example Features

As an example, we will pick the DNA methylation position with the ID "cg08691479" on the chromosome 11 and visualize some plots under the target "max. severity".

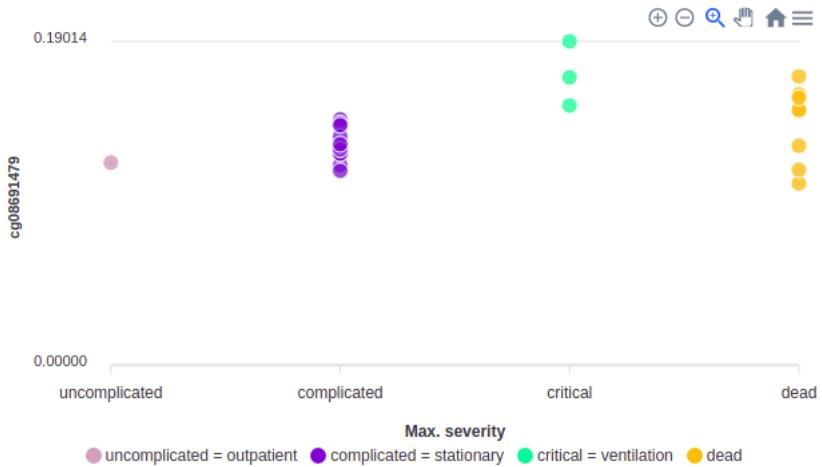


Figure 5.4.3.1 The scatter plot for the DNA methylation values of the position with ID: "cg08691479" under target "max. severity".

Figure 5.4.3.1 shows clear increase in the methylation at the selected position between "complicated" and "critical" groups under the selected target. The following boxplot could provide us with better observations.

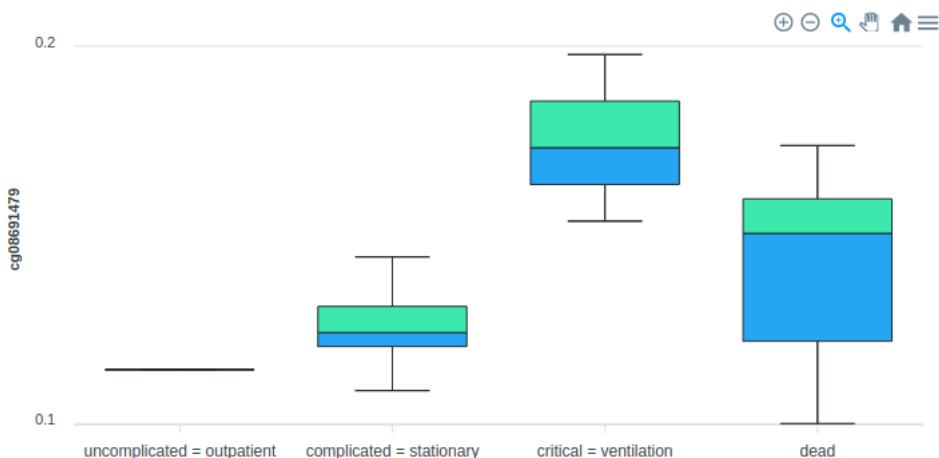


Figure 5.4.3.2 The box plot chart for the DNA methylation values of the position with ID: "cg08691479" under target "max. severity".

In both plots, we can clearly notice the strong direct relationship between the DNA methylation value in this position and the severity level of the disease by admission.

It's clear from Figure 5.4.3.2 that the methylation median value for "complicated" group at the selected position is significantly lower than the ones for "critical" and "dead" groups. It's also very clear , according to the 25% and 75% percentiles, how the values fall in a much wider range in "complicated" severity group under the selected target.

Recently, multiple articles about the relationship between Covid-19 and epigenetics (especially DNA methylation) are enriching the clinical literature.

In this context, it's important to mention the two following studies. The first is titled "[Hidden pandemic: COVID-19-related stress, SLC6A4 methylation, and infants' temperament at 3 months](#)" and it's talking about the multiple DNA methylation changes detected among pregnant women infected with SAR-CoV2 and the impact of that on the temperament of the infants at 3 months[8]. "[Magnesium treatment on methylation changes of transmembrane serine protease 2 \(TMPRSS2\)](#)" is another article that should be stated here. It emphasizes on the role of magnesium in motivating DNA methylation at some positions helping in the prevention and treatment of patients with low severity levels of Covid-19[58].

5.5 miRNome

5.5.1 General Information

The miRNome layer, as one of the most important post-transcriptional factors in gene regulation, was studied by applying a bulk miRNA analysis on the peripheral blood mononuclear cells (PBMCs) obtained from the patients.

The sequencing of the transcriptomics has been achieved using DNB-SEQ 400 / BGI-SEQ 500 platform for bulk miRNA next generation sequencing.

After that, the raw data (.fastq.gz) went through preprocessing and analysis using miRMaster2[59] webservice. The analysis of 32 samples out of 32 received samples was conducted and the results were integrated in the web app.

For every patient, the expression values of 1,434 miRNAs (features) have been studied. For every feature, the features table contains the miRNA ID, a column that indicates if the feature is a prime feature or not and the feature selection consensus score of the feature.

FEATURE TYPE	PRIME	FEATURE	FEATURE SELECTION SCORE
Search			
numerical	★	hsa-miR-371b-3p	2
numerical	★	hsa-miR-1976	2
numerical	★	hsa-miR-4632-3p	2
numerical	★	hsa-miR-125b-2-3p	2
numerical	★	hsa-miR-552-3p	2
numerical	★	hsa-miR-3678-3p	2
numerical	★	hsa-miR-552-5p	2
numerical	★	hsa-miR-335-3p	2
numerical	★	hsa-miR-4700-5p	2
numerical	★	hsa-miR-187-3p	2

Showing 1 to 10 of 30 entries (filtered from 1,434 total entries)

Previous 1 2 3 Next

Figure 5.5.1 miRNome features table under target "severity by admission".

Figure 5.5.1 displays the miRNome features table that includes the miRNA name, its feature selection consensus score and a value that indicated if it's considered as a prime feature according to feature selection.

5.5.2 PCA (Principal Component Analysis)

The principal component analysis shows the PC1 representing 18.3% of the variance with the miRNA "hsa-miR-4446-3p" as best contributor and the PC2 representing 13% of the variance with the miRNA "hsa-miR-148b-3p" as best contributor. This figure shows the PCA scatter plot chart for the two targets of the study.

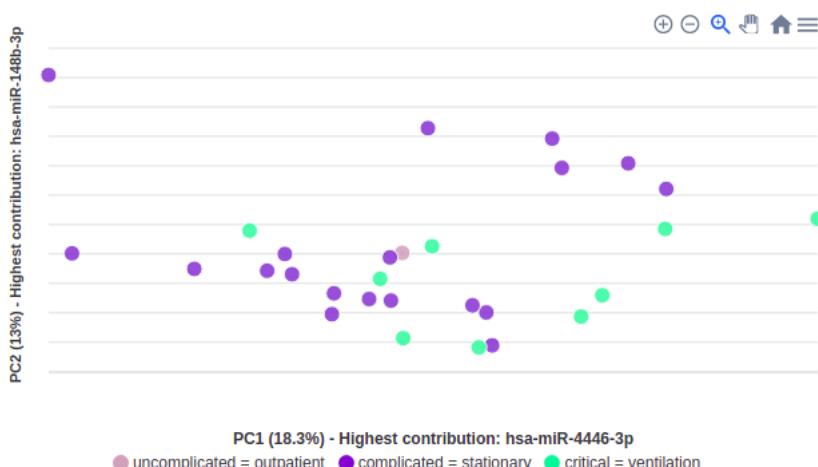


Figure 5.5.2.1 miRNA PCA scatter plot for "severity by admission" target.

There is no clear noticeable clustering in Figure 5.5.2.1. Only some outliers are observable.

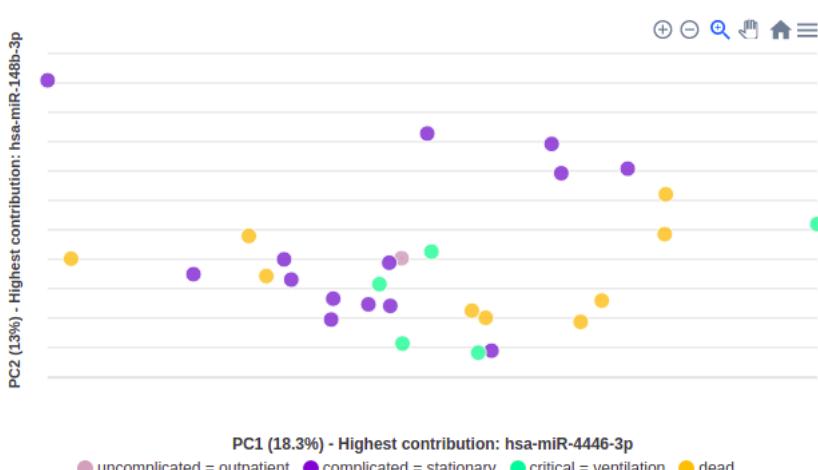


Figure 5.5.2.2 miRNA PCA scatter plot for "max. severity" target.

As in the previous figure, there is no clear noticeable clustering in Figure 5.5.2.2 and only some outliers are observable.

5.5.3 Differential Analysis

miRMaster2[59] uses the wilcoxon rank-sum test for conducting differential analysis where p-values are adjusted using the FDR (Benjamini-Hochberg) method (adj-pval <= 0.05).

As in some other omics, an interactive clustered heatmap and volcano plots have been created.

Interactive Clustered Heatmap

The following figure displays the DEA analysis heatmap created for miRNome data using Clustergrammer[48] online tool.

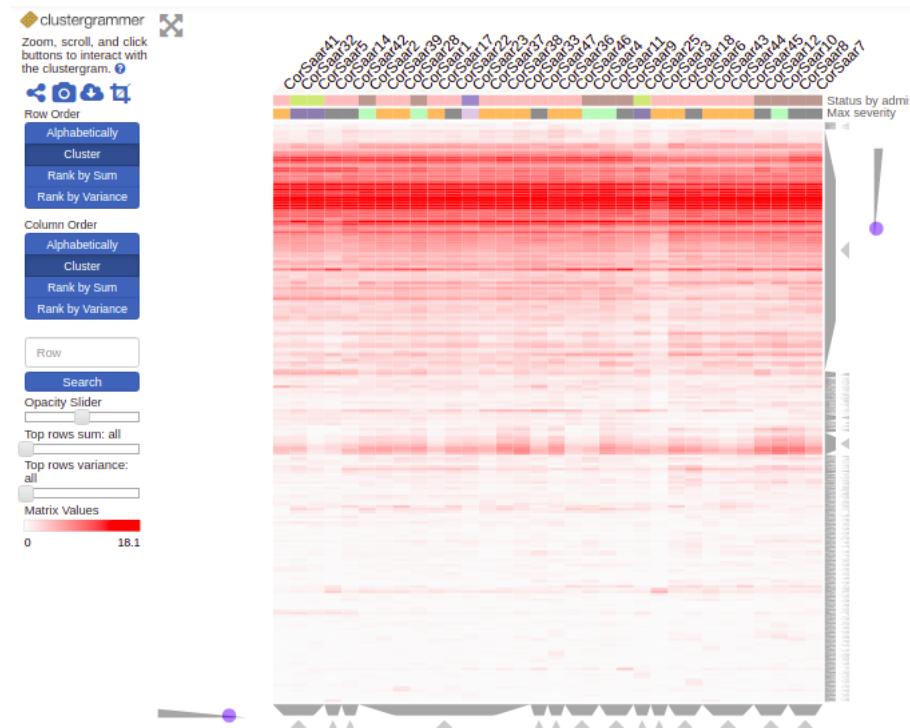


Figure 5.5.3.1 An interactive clustered heatmap created for miRNome using Clustergrammer online tool.

By exploring the interactive heatmap in Figure 5.5.3.1 using our web app, we notice that there is no strong relationship between the miRNA normalized expression values and the severity levels from any of the targets. From the figure we can notice the diverse clustering patterns of the miRNAs and the cases (patients).

Volcano Plots

using the fold change values calculated by DESeq2[47] and the adjusted p-values, volcano plots are generated to compare two different levels of severity under each of the two different targets of the study. By choosing any two different levels of severity, a volcano plot is generated showing the differentially expressed miRNAs. A miRNA is considered as differentially expressed if it has an adj-pval ≤ 0.05 and an absolute value for fold change ≥ 1.5 .

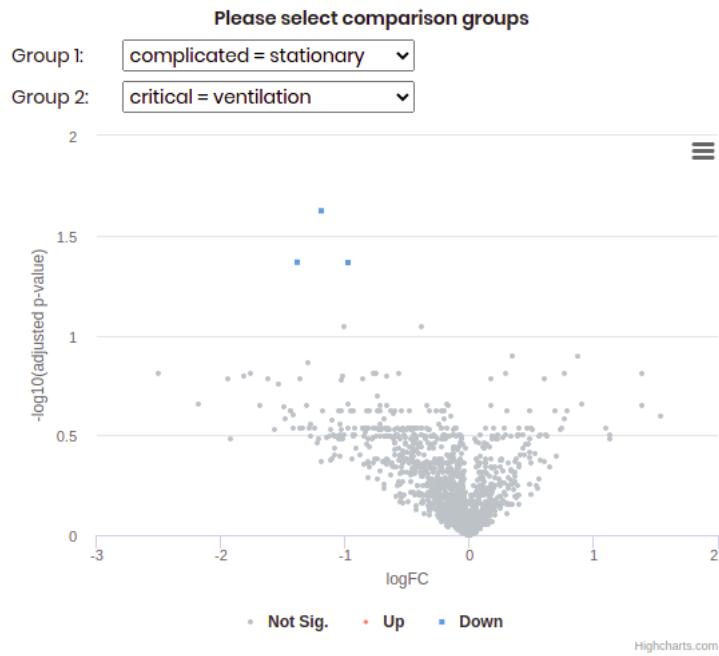


Figure 5.5.3.2 The volcano plot showing down-regulated genes between "complicated = stationary" and "critical = ventilation" severity levels under the target "severity by admission".

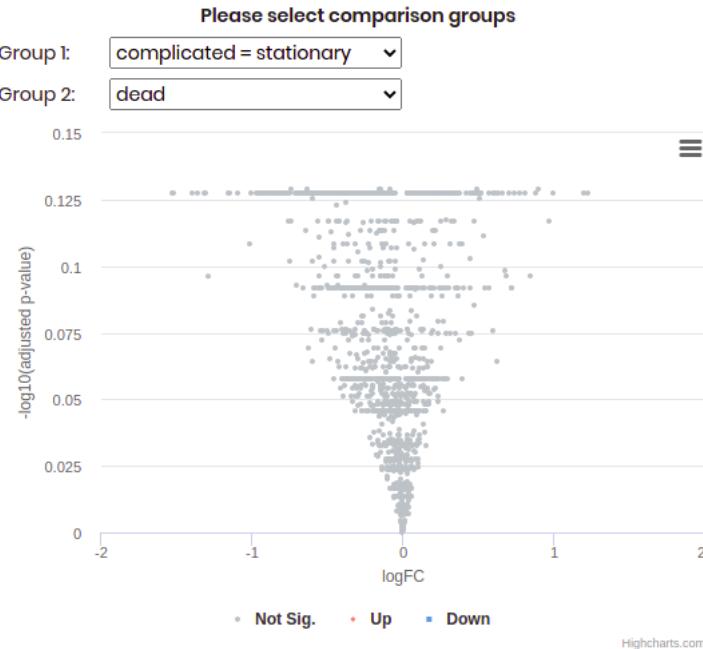


Figure 5.5.3.3 The volcano plot for "complicated = stationary" vs "dead" severity levels comparison under the target "max. severity" showing no significantly differentially expressed miRNA.

For our miRNome data, it was very hard to find differentially expressed miRNAs and that is clear in the two example plots above (Figure 5.5.3.2 and Figure 5.5.3.3). However, in the "complicated vs critical" volcano plot under the "severity by admission" target shown above, we can notice "hsa-miR-16-2-3p", "hsa-miR-126-5p" and "hsa-miR-335-3p" as significantly downregulated miRNAs.

5.5.4 miEAA - Enrichment Analysis

This study uses miEAA2[60] which is a miRNA enrichment and annotation tool to apply enrichment analysis on the miRNome data under the two targets of the study, "severity by admission" and "max. severity".

miEAA API is integrated to conduct a GSEA analysis (gene set enrichment analysis) in a very simple way. The only thing needed is choosing the target and comparison (sorting) factor, which could be mean or median, as in the following figure.

The interface has two tabs at the top: "Status by admission" and "Max. severity". Below the tabs is a section titled "miRNA Enrichment Analysis (miEAA API)". It contains two blue buttons: "Run miEAA sorting by mean (RPM)" and "Run miEAA sorting by median (RPM)".

Figure 5.5.4.1 A user can choose the comparison factor and the target under which you want to run the miEAA enrichment analysis.

After clicking on one or both of the blue buttons, a miEAA job on the miEAA server will

start and the user can start monitoring the job by clicking the "Open .." button.

Figure 5.5.4.2 When the job on miEAA's server is ready, you can open it and wait for the results.

The results include an enrichment table, a wordcloud of categories and a miRNA / precursor to category heatmap.

Assuming that we choose "Status by admission" and "Mean (RPM)" as comparison, we will get the following results after the miEAA job is done:

Category	Subcategory	Running Sum	Enrichment	P-value	P-adjusted	Q-value	Observed	miRNAs/precursors	Enrichment plot
Localization (RNALocate)	microvesicle		enriched	1.09e-133	1.20e-132	1.20e-132	852	hsa-miR-223-3p; hsa-miR-26a-5p; hsa-let-7a-5p; hsa-miR-451a; hsa-miR-150-5p; hsa-miR-16-5p; hsa-miR-191-5p; hsa-miR-21-5p; hsa-let-7f-5p; hsa-miR-142-3p;	Download
Diseases (MNDR)	renal clear cell carcinoma		enriched	5.01e-122	5.99e-119	5.99e-119	445	hsa-miR-223-3p; hsa-miR-26a-5p; hsa-let-7a-5p; hsa-miR-451a; hsa-miR-150-5p; hsa-miR-16-5p; hsa-miR-191-5p; hsa-miR-21-5p; hsa-let-7f-5p; hsa-miR-142-3p;	Download
Diseases (MNDR)	osteosarcoma		enriched	2.19e-118	1.31e-115	1.31e-115	409	hsa-miR-223-3p; hsa-miR-26a-5p;	Download

Figure 5.5.4.3. Results for the GSEA under the "severity by admission" target and using "Mean (RPM)" as the sorting criteria.

The table in Figure 5.5.4.4 shows the enriched / depleted miRNAs and precursors with important information about their category (e.g. localization, disease, Gene Ontology), subcategory(e.g. microvesicle, renal clear cell carcinoma), running sum, enrichment(enriched, depleted), p-value, adj p-value, and q-value with the ability to download the enrichment plot. When displaying the interactive inline plot, we can see the "running sum" curve in red if the miRNA is enriched or in green if it's depleted.

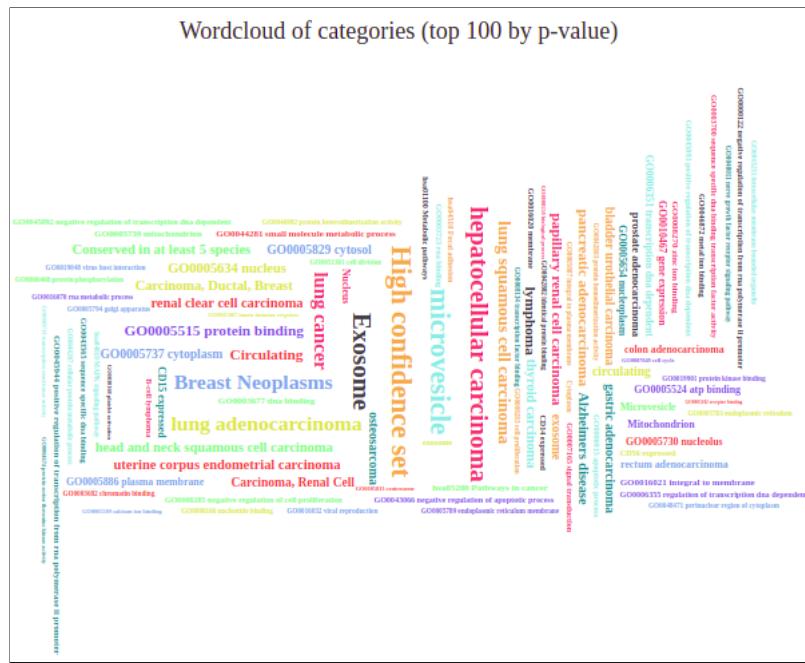


Figure 5.5.4.4. The wordcloud plot for the GSEA under the "severity by admission" target and using "Mean (RPM)" as the sorting criteria.

Figure 5.5.4.4 shows the wordcloud plot generated by miEAA for the maximum top 100 most significant subcategories.

The wordcloud reflects the most significant subcategories, with sizes reflecting the observed counts of each. The scaling can be adjusted using the provided drop-down.

This wordcloud is considered as a great search starting-point in the clinical literature on Pubmed. For example, let's have a look at the two terms: "Exosome" and "microvesicle". Searching Pubmed for articles including ["Covid-19" or "SARS-COV-2"] AND ["exosomes"] leads us to a bunch of trusted clinical articles emphasizing on the role of exosomes in the pathology and the therapy for Covid-19.

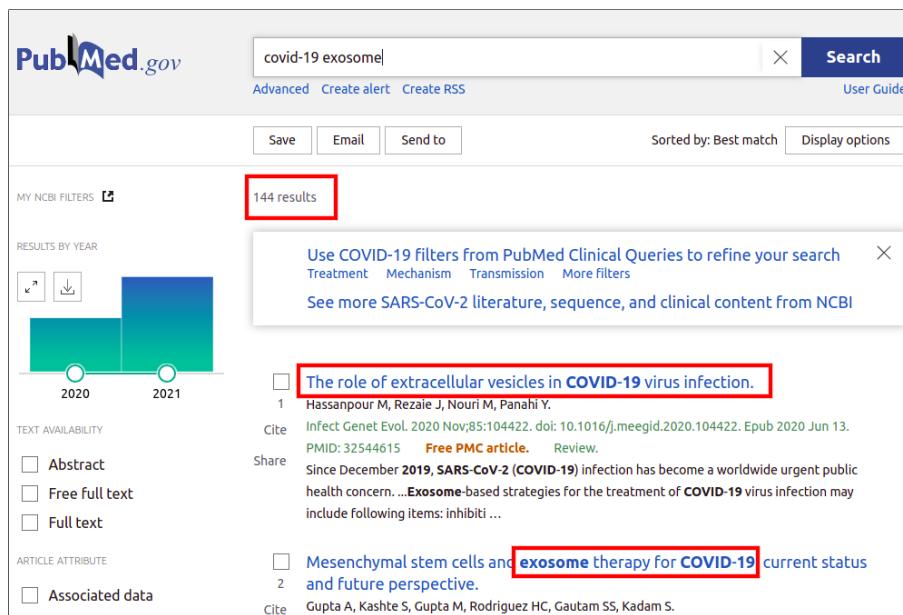


Figure 5.5.4.5 The role of exosomes in pathology and therapy for Covid-19..

Searching for ["Covid-19" or "SARS-COV-2"] AND ["microvesicles"] will also lead us to many articles confirming the significant role played by mircrovesicles in the underlying pathology of the disease.

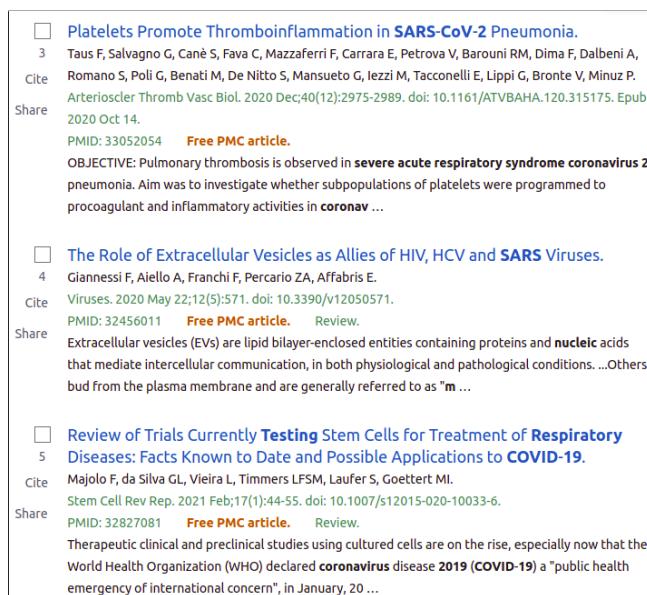


Figure 5.5.4.6 Articles related to both terms “Microvesicles” and “Covid-19” on Pubmed.

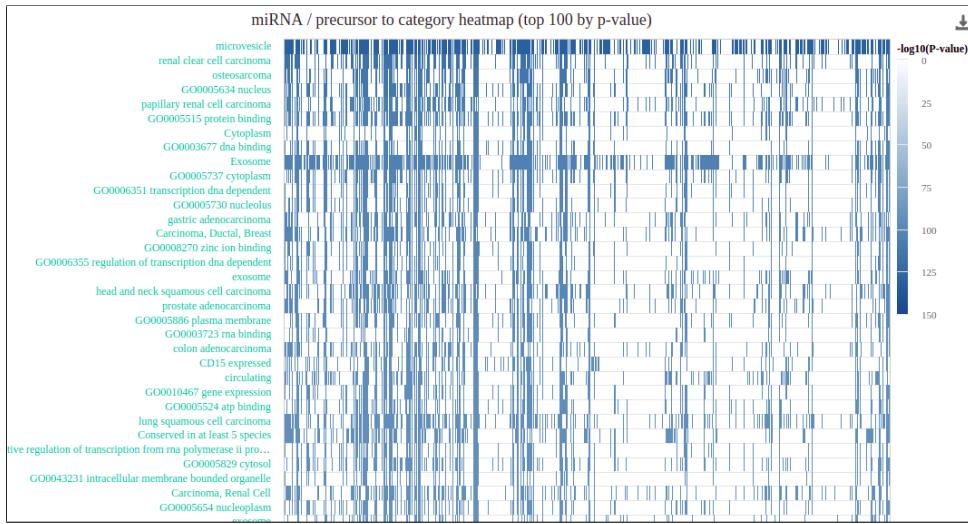


Figure 5.5.4.7 miEAA results. miRNA / precursor to category heatmap (top 100 by p-value).

The heatmap in Figure 5.5.4.7 shows the negative log base 10 p-values of the top subcategories as rows. On the other hands, the columns represents the miRNAs/precursors.

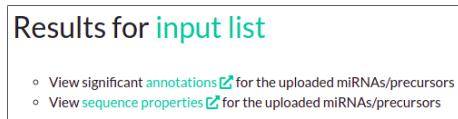


Figure 5.5.4.8 The miRNA / precursor to category heatmap plot for the GSEA under the "severity by admission" target and using "Mean (RPM)" as the sorting criteria.

A final note to add here, is that the miEAA2 analysis buttons in our app run the analysis in the non-expert mode. This means that the categories included are the categories listed in miEAA2 as non-expert categories.

5.5.5 Prime Features and Example Features

Prime Features

The following 30 features have been selected by the consensus of our feature selection methods as the prime miRNAs:

Feature	Consensus Score
hsa-miR-16-2-3p	2/4
hsa-miR-335-3p	2/4
hsa-miR-125b-2-3p	2/4
hsa-miR-1276	2/4
hsa-miR-552-3p	2/4
hsa-miR-3678-3p	2/4
hsa-miR-552-5p	2/4
hsa-miR-371b-3p	2/4
hsa-miR-154-3p	2/4
hsa-miR-133a-5p	2/4
hsa-miR-4632-3p	2/4
hsa-miR-4700-5p	2/4
hsa-miR-1291	2/4
hsa-miR-1976	2/4
hsa-miR-187-3p	2/4
hsa-miR-548a-5p	1/4
hsa-miR-126-5p	1/4
hsa-miR-4511	1/4
hsa-miR-335-5p	1/4
hsa-miR-20a-5p	1/4
hsa-miR-625-5p	1/4
hsa-miR-655-3p	1/4
hsa-miR-4455	1/4
hsa-miR-548ax	1/4
hsa-miR-1-3p	1/4
hsa-miR-4766-5p	1/4
hsa-miR-4644	1/4
hsa-miR-369-3p	1/4
hsa-miR-487a-3p	1/4
hsa-miR-5582-3p	1/4

Table 5.5.5.1 Prime miRNAs under target "severity by admission".

Feature	Consensus Score
hsa-miR-103a-1-5p	3/4
hsa-miR-664b-3p	3/4
hsa-miR-3162-5p	2/4
hsa-miR-577	2/4
hsa-miR-5683	2/4
hsa-miR-5000-3p	2/4
hsa-miR-7854-3p	2/4
hsa-miR-125b-2-3p	2/4
hsa-miR-7704	2/4
hsa-miR-11401	2/4
hsa-miR-20a-5p	1/4
hsa-miR-126-5p	1/4
hsa-miR-16-2-3p	1/4
hsa-miR-5582-3p	1/4
hsa-miR-1276	1/4
hsa-miR-17-5p	1/4
hsa-miR-4766-5p	1/4
hsa-miR-4795-3p	1/4
hsa-miR-182-3p	1/4
hsa-miR-153-5p	1/4
hsa-miR-106a-5p	1/4
hsa-miR-335-3p	1/4
hsa-miR-3678-3p	1/4
hsa-miR-6749-5p	1/4
hsa-miR-556-5p	1/4
hsa-miR-4420	1/4
hsa-miR-93-5p	1/4
hsa-miR-9899	1/4
hsa-miR-628-5p	1/4
hsa-miR-4659a-3p	1/4

Table 5.5.2 Prime miRNAs under target "max. severity".

Example Features

By Clicking on a feature of interest from the features table, a table displaying the **normalized** expression values for different patients will appear. We will choose "hsa-miR-548aj-3p" under the target "max. severity" as an example. This miRNA has 854 experimentally validated targets in miRTarBase[61].

Feature: hsa-miR-664b-3p		
Show 10 entries	Search:	
CASE ID (COSAAR ID)	VALUE	MAX. SEVERITY
CorSaar1	2.65781604441666	critical = ventilation
CorSaar10	2.59704244629224	critical = ventilation
CorSaar11	3.26189332392369	critical = ventilation
CorSaar12	2.32318269390146	dead
CorSaar14	2.50471953455923	dead
CorSaar17	2.87408281739357	complicated = stationary
CorSaar18	3.48200926295707	complicated = stationary
CorSaar2	2.84093046047216	critical = ventilation
CorSaar22	2.50025137884262	dead
CorSaar23	2.42268780857981	uncomplicated = outpatient

Showing 1 to 10 of 29 entries

Previous 1 2 3 Next

Figure 5.5.5.1 The normalized expression value table for miRNA "hsa-miR-548aj-3p" under target: "max. severity".

Figure 5.5.5.1 shows the normalized expression values for "hsa-miR-548aj-3p" under target: "max. severity" with the severity levels under the selected target. Since the expression values are numerical values (the feature is a numerical feature), a scatter plot and box plots will also be displayed.

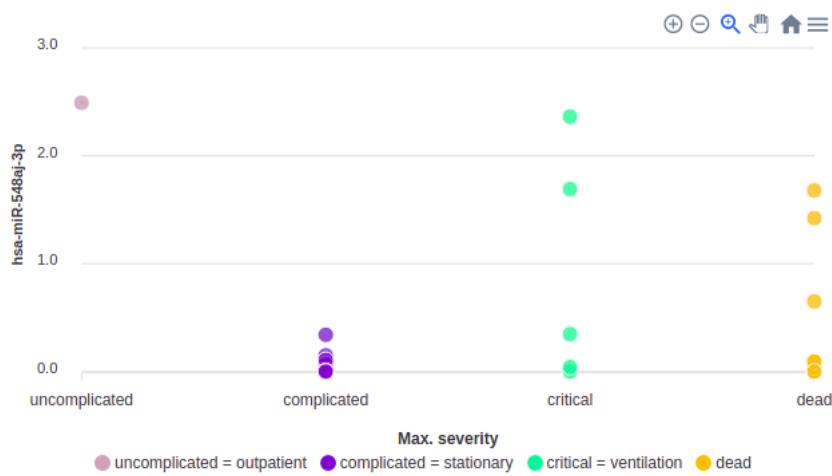


Figure 5.5.5.2 The scatter plot for the expression values of miRNA "hsa-miR-548aj-3p" under target: "max. severity".

From Figure 5.5.5.2 we can notice how the positive relationship between the severity level and the normalized miRNA expression values under the selected target. However there is one patient labeled as "uncomplicated" with a relatively high normalized expression value but that's not enough to make a strong observation or to change the observation we came up with using all other cases.

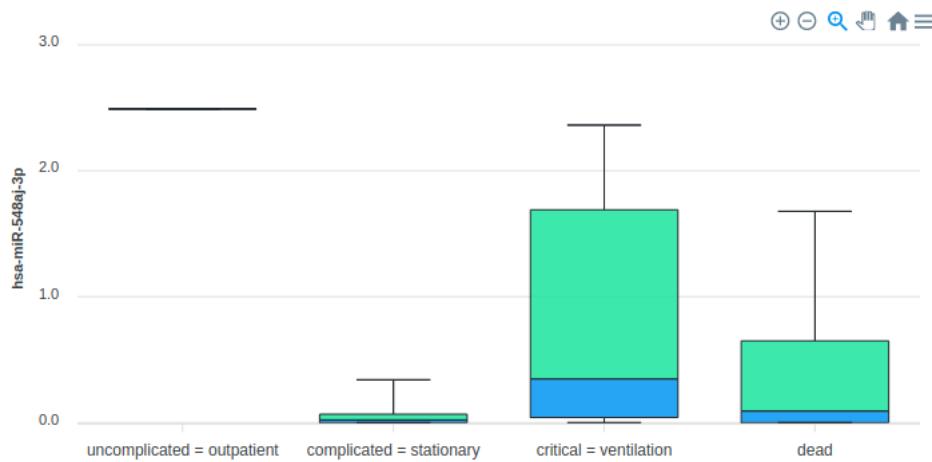


Figure 5.5.5.3 The box plot chart for the expression values of miRNA "hsa-miR-548aj-3p" under target: "max. severity".

Figure 5.5.5.3 shows the wide difference between the maximum value of normalized miRNA expression values for "complicated" patients vs those for "critical" and "dead" patient under the "max. severity target".

The figures above show variance in the values between different severity level groups. Indeed, the family to which this miRNA belongs (hsa-miR-548) has already been stated in the clinical literature as one of the miRNA families that play a very important role in the outcome of Covid-19 patients. One example is an article published in Semantic Scholar journal and titled "[COVID-19 Virulence in Aged Patients Might Be Impacted by the Host Cellular MicroRNAs Abundance/Profile](#)"[7].

5.6 Metabolome

5.6.1 General Information

The metabolome data was obtained from plasma samples using the UHPLC-HRMS/MS platform and processed using the XCMS3 data pipeline. Two different column coatings for separation of molecules with different polarities are used. These are Phenyl-Hexyl (PhenHex) and Hydrophilic Lipophilic Interaction Chromatography (HILIC).

Metabolomics are integrated in the web server at the moment including PCA charts, interactive clustered heatmaps and multiple feature visualization methods. However, they are not ready to help with any clinical value now since the features still hold virtual names (e.g. FT0001 referring to Feature_0001) and a lot of samples are not mapped to their disease severity. We aim that further future work will be conducted later on the metabolomics data that we collected and analyzed.

Chapter 6

Multiomics Analysis

6.1 General Information

For cross-omics, R^2 (R Squared) has calculated from the multiple variable Spearman's correlation for each feature from all omics against all features from other omics. A threshold of 0.65 was used for the absolute value of R^2 . An exception was DNA methylation where correlations were calculated only for the 50k features (CpGs) with the highest variance since it was very hard to compute correlation for more than 800k features against features from all other omics.

6.2 Overall Correlation Graph

This view provides a network analysis by creating a graph that includes the top correlated features between selected omics under the selected study target. For this purpose, R^2 calculated from Spearman's correlation is used.

For the two targets of the study, the values of Spearman's correlation between every pair of features from all omics was calculated by creating a Pandas data frame with 3 columns: first feature values column, second feature values column and severity score column where every row represented a patient. After that the method corr() of the data frame was called with the method parameter set to "spearman". This created the correlation data frame from which we calculated adjusted R^2 .

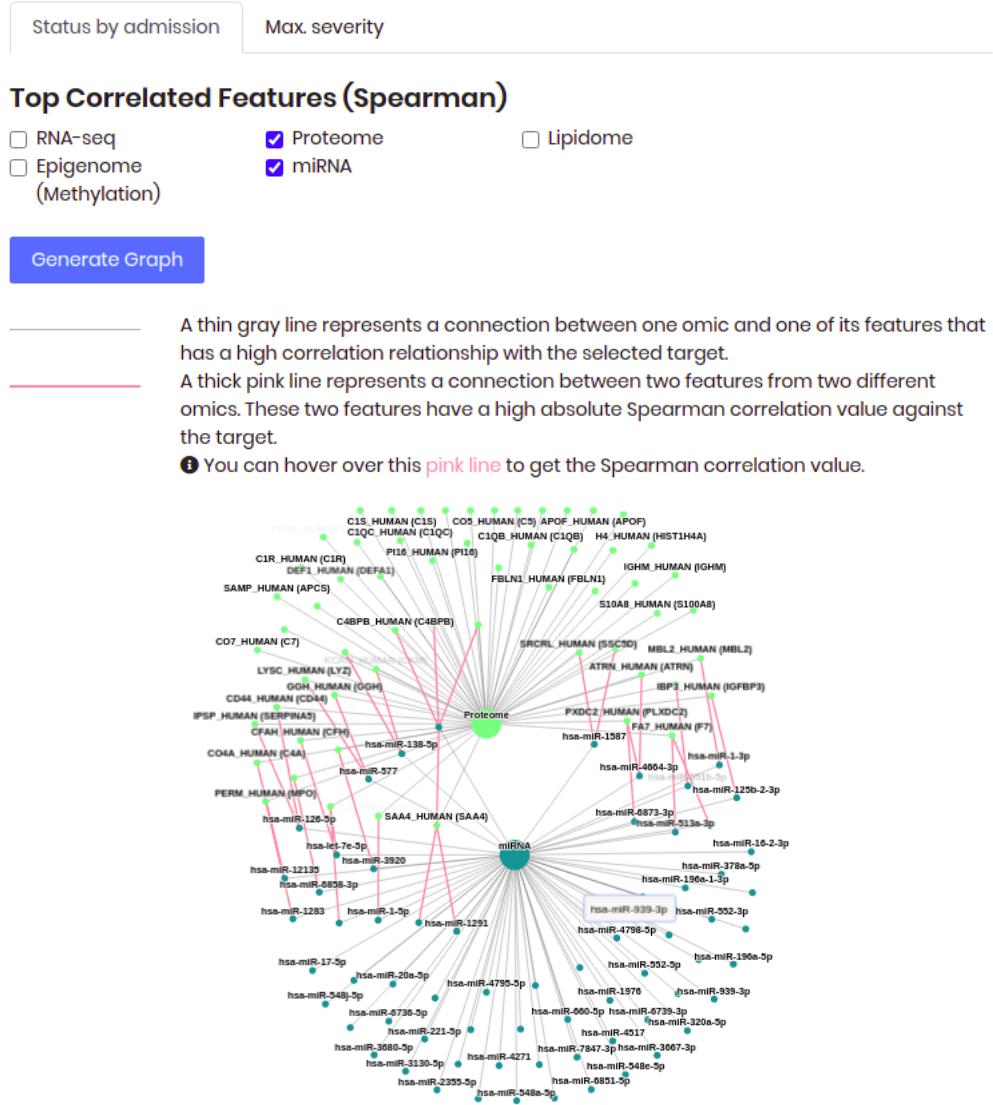


Figure 6.2 An overall correlation graph showing a network analysis that includes the top correlated features from proteomics and miRNome using Spearman's correlation under the target "severity by admission".

In Figure 6.2, a thin gray line represents a connection between one omic and one of its features that has a high correlation relationship with the selected target. On the other hand, A thick pink line represents a connection between two features from two different omics. These two features have a high R^2 value under the selected study target. Hovering over this pink line shows the R^2 value in a simple tool tip.

For example, the R^2 for the R^2 value that will appear after hovering on the thick pink line between "Perm_HUMAN (MPO)" protein and "has-miR-1226-5p" miRNA" is very high and equals 0.9919.

6.3 Pairwise Analysis

By selecting two different omics from the lists and clicking on the "Get Data" button, a "Table of Top 30 Correlated Features" and a "Correlation Heatmap for Prime Features" will appear.

6.3.1 Top Pairwise Correlation with Target

The top pairwise correlation table shows the names of the features, the value of the R^2 of the Spearman's multiple variables correlation and a link to display the scatter plot of their values in different cases (patients) for the top 30 highly correlated pair of features with a threshold of 0.65 for the R^2 value.

RNA-seq features	miRNA features	R Squared	Operations
MRC2	hsa-miR-6850-5p	0.9353603092284746	Scatter plot
BPI	hsa-miR-4697-5p	0.933332313243576	Scatter plot
EEFIG	hsa-miR-6750-5p	0.9310853431632152	Scatter plot
DISP1	hsa-miR-26a-1-3p	0.9299878892342954	Scatter plot
NKAPPI	hsa-miR-4654	0.92858600946263	Scatter plot
DUSP15	hsa-miR-541-3p	0.92797280153068	Scatter plot
NKAPPI	hsa-miR-4646-5p	0.9276148175240956	Scatter plot
FUT4	hsa-miR-3150a-3p	0.9267857186561598	Scatter plot
GBAS	hsa-miR-221-5p	0.9256138184464222	Scatter plot
PGAMI	hsa-miR-4697-5p	0.9250757643815642	Scatter plot

Showing 1 to 10 of 30 entries

Previous 1 2 3 Next

Figure 6.3.1 Top pairwise correlation table for RNA-seq and miRNome under the target "severity by admission".

6.3.2 Correlation Heatmap for Prime Features Selected Under a Specific Target

The correlation heatmap is an interactive clustered heatmap for the Spearman's correlation values of the **prime features from the two omics selected under one of the two targets of the study**. This doesn't mean that this will calculate the multiple variable correlation for the two features and the target at once. This heatmap is created using the Clustergrammer[48] online tool.

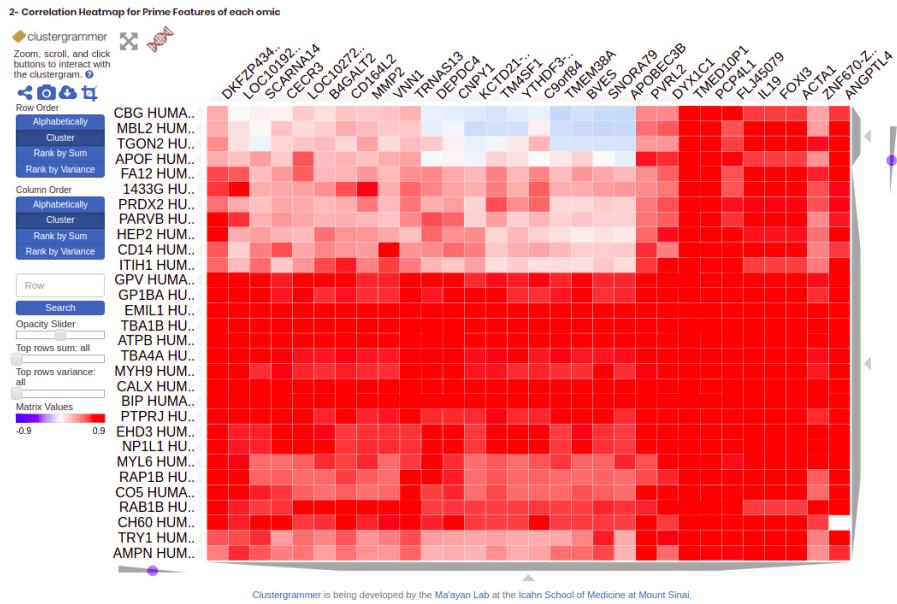


Figure 6.3.2 Correlation Heatmap for the Prime Features of RNA-seq (rows) and proteome (cols) under target "max. severity".

Figure 6.3.2 illustrates a heatmap for the prime features of RNA-seq (rows) and proteome (cols) under target "max. severity", where each cell's background color represents the mean R value of the Spearman's correlation between the two features. The darker the red the higher the R^2 value of the Spearman's correlation is. On the other hand, the darker the blue the lower the R^2 value is.

6.3.3 Correlation Scatter Plot

The final section of the pairwise analysis view allows you to select any two different features from any two different omics and displays a scatter plot in which each point represents a patient in the study. The x-axis represents the first selected omic, while the y-axis represents the other.

Feature names and their feature selection consensus score are searchable and sortable.

3- Correlation Scatter Plot

miRNA (x-axis)

Please select a feature from the table
Selected Feature: (none)

FEATURE NAME	FEATURE SELECTION SCORE
Search	2
hsa-miR-125b-2-3p	2
hsa-miR-1276	2
hsa-miR-1291	2
hsa-miR-133a-5p	2

Showing 1 to 4 of 4 entries (filtered from 1,434 total entries)

Lipidome (y-axis)

Please select a feature from the table
Selected Feature: (none)

FEATURE NAME	FEATURE SELECTION SCORE
Search	3
315.1339_1.02	3
370.2953_1.14	3
870.5246_4.92	3
744.5903_6.39	3

Showing 1 to 4 of 8 entries (filtered from 1,590 total entries)

Previous 1 2 3 4 Next

Previous 1 2 Next

Figure 6.3.3.1 Selecting features for correlation scatter plot. miRNAs and lipids are filtered to those with feature selection consensus score ≥ 2 and ≥ 3 consecutively.

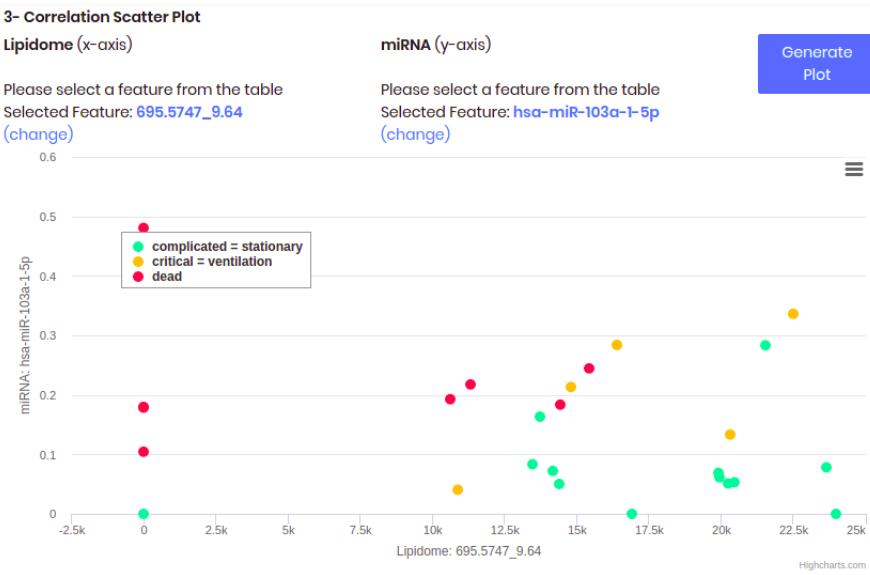


Figure 6.3.3.2 Correlation scatter plot from pairwise analysis. This chart plots the values for "695.5747_9.64" bucket from lipidome and "hsa-miR-103a-1-5p" miRNA from miRNome under the target "max. severity".

From the right half of Figure 6.3.3.2 we can observe the positive relationship between the severity level and both of the bucket mean intensity and the miRNA normalized

expression value.

The aim of creating such a feature in our app is to allow the researchers to study the relationship between any two different features from different omics in the data obtained from the patients in this study.

Chapter 7

Data Availability

The data included in this app can be categorised into tables and charts (plots).

7.1 Tables

All data tables are available for download using the links under the "Downloads" section. In this page, a user can find links to download full data matrices and consensus feature selection files.

The screenshot shows a user interface for data downloads. On the left, there's a sidebar with 'navigation' and 'useful links'. The 'useful links' section has a highlighted 'Downloads' item. The main content area is divided into three sections: 'RNA-seq', 'Proteome', and 'Peptidome'. Each section contains a list of download links, each preceded by a blue download icon.

Category	Download Links
RNA-seq	<ul style="list-style-type: none">RNA-seq full (.csv)Severity by admission prime features - consensus (.csv)Severity by admission prime features - only Spearman's correlation (.csv)Max. severity prime features - consensus (.csv)Max. severity prime features - only Spearman's correlation (.csv)
Proteome	<ul style="list-style-type: none">Proteome full (.csv)Severity by admission prime features - consensus (.csv)Severity by admission prime features - only Spearman's correlation (.csv)Max. severity prime features - consensus (.csv)Max. severity prime features - only Spearman's correlation (.csv)
Peptidome	<ul style="list-style-type: none">Peptidome full (.zip)

Figure 7.1.1 A screenshot from the downloads page.

One other option to download a data table is to use the gray buttons above that table.

Copy CSV							
FEATURE TYPE	PRIME	ID	CHROMOSOME	START	END	STRAND	FEATURE SELECTION SCORE
		<input type="text" value="Search"/>	<input type="button" value="All"/>	<input type="button" value="Search"/>	<input type="button" value="Search"/>	<input type="button" value="▼"/>	<input type="text" value="Search (>=)"/>
numerical	★	cg08691479	chr11	9595265	9595266	-	2
numerical	★	cg06758649	chr12	107381105	107381106	-	2
numerical	★	cg13826167	chr15	65203855	65203856	-	1
numerical	★	cg14256511	chr12	123347684	123347685	-	1
numerical	★	cg12743270	chr19	3028879	3028880	+	1
numerical	★	cg21725888	chr6	43253015	43253016	+	1
numerical	★	cg26289012	chr17	3540036	3540037	-	1
numerical	★	cg13319975	chr6	146136371	146136372	+	1
numerical	★	cg21287517	chr11	73498860	73498861	+	1
numerical	★	cg24218995	chr18	67873134	67873135	-	1

Showing 1 to 10 of 30 entries (filtered from 805,779 total entries)

Previous 1 2 3 Next

Figure 7.1.2 Export buttons for DNA methylation features data table.

7.2 Charts

All charts are available for download in multiple formats (.png, .pdf, .svg) using the charts export buttons.

Chapter 8

Conclusion & Future Work

8.1 Conclusion

This work, defined by the analysis and visualization of the biological data from different genomic and epigenomic layers, comes with many strength points. We will discuss the above mentioned points in this section while reserving the next section to discuss further work that can be carried out in the future by other researchers, either clinicians or bioinformaticians.

Starting with strength points, this study has two targets: "Disease severity level by admission" and "Max. disease severity level reached by the patient", which allows not only to study the relationship between different features and the severity of the disease, but also with the prognosis. By selecting prime features depending on the consensus of 4 different methods and algorithms, the top 30 selected features listed under each omic should give the researcher more confidence about the features affecting the severity and prognosis more. This is very important as a starting point for the clinical and pharmaceutical research. By using different ways of visualizing, we aimed to accelerate the process of understanding the information retrieved from the analysis. Providing enrichment analysis for RNA-seq and miRNA highlights the abnormal biological processes in Covid-19 patients helping the researchers to be in the right direction to bring these paths back to balance. Beside the network analysis and the pair-wise correlation studies carried out under cross-omics section, the pairwise correlation scatter plot comes as a great feature to study the correlation between a pair of any different features from any different omics and any of the targets of the study.

As a free open-access source for Covid-19 multiomics data, this work is a very useful tool for researchers who have already found interesting findings while working on the data they have got and want to cross check with other resources.

8.2 Future work

This work and other similar multi-omics COVID-19 studies like [Large-Scale Multi-omic Analysis of COVID-19 Severity](#) [5], [Multi-omic approach identifies a transcriptional network coupling innate immune response to proliferation in the blood of COVID-19 cancer patients](#) [35] and [Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19](#) [36] open the gate for great future work. One option is to build Covid-19 severity level prediction models by applying machine learning on the selected prime features. This becomes more reliable by getting more samples and more detailed metadata to select prime features with higher consensus and therefor building ML models with higher accuracy that can be used in the clinical practise. One other field, which is actually an extension of the idea stated above, is "Pandemic management". We are very lucky that the Covid-19 pandemic stayed under control in many countries and that vaccines for Covid-19 were developed rapidly. However, it's very essential that we develop "Pandemic Management" systems that can predict the current disease severity level with prognosis to label each case with an accurate priority level allowing the patient to be admitted in a hospital or in the ICU or not. These are not the only possible extensions for projects that are similar to CORSAAR project, but they are very essential and promising.

Bibliography

- [1] Patrick C. Woo, Yi Huang, Susanna K. Lau, and Kwok-Yung Yuen. Coronavirus genomics and bioinformatics analysis. *Viruses*, 2(8):1804–1820, 2010.
- [2] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, and et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395(10224):565–574, 2020.
- [3] Masato Akiyama. Multi-omics study for interpretation of genome-wide association study. *Journal of Human Genetics*, 66(1):3–10, 2020.
- [4] Lasseigne Lab. Lasseigne lab. <https://www.lasseigne.org/projects/multiomics/>, 2021.
- [5] Katherine A. Overmyer, Evgenia Shishkova, Ian J. Miller, Joseph Balnis, Matthew N. Bernstein, Trenton M. Peters-Clarke, Jesse G. Meyer, Qiuwen Quan, Laura K. Muehlbauer, Edna A. Trujillo, and et al. Large-scale multi-omic analysis of covid-19 severity. *Cell Systems*, 12(1), 2021.
- [6] Marta Ferreira-Gomes, Andrey Kruglov, Pawel Durek, Frederik Heinrich, Caroline Tizian, Gitta Anne Heinz, Anna Pascual-Reguant, Weijie Du, Ronja Mothes, Chaofan Fan, and et al. Sars-cov-2 in severe covid-19 induces a tgf--dominated chronic immune response that does not target itself. *Nature Communications*, 12(1), 2021.
- [7] Sadanand Fulzele, Bikash Sahay, Ibrahim Yusufu, Tae Jin Lee, Ashok Sharma, Ravindra Kolhe, and Carlos M. Isales. Covid-19 virulence in aged patients might be impacted by the host cellular micrornas abundance/profile. *Aging and Disease*, 11:509 – 522, 2020.
- [8] Livio Provenzi, Fabiana Mambretti, Marco Villa, Serena Grumi, Andrea Citterio, Emanuela Bertazzoli, Giacomo Biasucci, Lidia Decembrino, Rossana Falcone, Barbara Gardella, and et al. Hidden pandemic: Covid-19-related stress, slc6a4 methylation, and infants' temperament at 3 months. *Scientific Reports*, 11(1), 2021.
- [9] Katherine A. Overmyer, Evgenia Shishkova, Ian J. Miller, Joseph Balnis, Matthew N. Bernstein, Trenton M. Peters-Clarke, Jesse G. Meyer, Qiuwen Quan, Laura K. Muehlbauer, Edna A. Trujillo, and et al. Large-scale multi-omic analysis of covid-19 severity. *Cell Systems*, 12(1), 2021.
- [10] Barbara E Stranger and Emmanouil T Dermitzakis. From dna to rna to disease and back: The central dogma of regulatory disease variation. *Human Genomics*, 2(6):383, 2006.
- [11] Kimberly R. Kukurba and Stephen B. Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11), 2015.

- [12] Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1):22–24, 2013.
- [13] Gencode Project. Gencode project. <https://www.gencodegenes.org/>, 2021.
- [14] Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. Rna sequencing: new technologies and applications in cancer research. *Journal of Hematology Oncology*, 13(1), 2020.
- [15] Mike Tyers and Matthias Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, 2003.
- [16] Bing Zhang and Bernhard Kuster. Proteomics is not an island: Multi-omics integration is the key to understanding biological systems. *Molecular and Cellular Proteomics*, 18(8), 2019.
- [17] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. Proteomics: Technologies and their applications. *Journal of Chromatographic Science*, 55(2):182–196, 2016.
- [18] Human Protein Reference Database. Human protein reference database. <http://www.hprd.org>, 2021.
- [19] J. Wade Harper and Eric J. Bennett. Proteome complexity and the forces that drive proteome imbalance. *Nature*, 537(7620):328–338, 2016.
- [20] Allan Legaspi, Malayappa Jeevanandam, H.Fletcher Starnes, and Murray F. Brennan. Whole body lipid and energy metabolism in the cancer patient. *Metabolism*, 36(10):958–963, 1987.
- [21] David A. Foster. Phosphatidic acid signaling to mtor: Signals for the survival of human cancer cells. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1791(9):949–955, 2009.
- [22] Oswald Quehenberger and Edward A. Dennis. The human plasma lipidome. *New England Journal of Medicine*, 365(19):1812–1823, 2011.
- [23] The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-ncov and naming it sars-cov-2. *Nature Microbiology*, 5(4):536–544, 2020.
- [24] Mirang Kim and Joseph Costello. Dna methylation: An epigenetic mark of cellular memory. *Experimental and Molecular Medicine*, 49(4), 2017.
- [25] Myriam Hemberger, Wendy Dean, and Wolf Reik. Epigenetic dynamics of stem cells and cell lineage commitment: Digging waddington’s canal. *Nature Reviews Molecular Cell Biology*, 10(8):526–537, 2009.
- [26] C. H. Waddington. The epigenotype. *International Journal of Epidemiology*, 41(1):10–13, 2011.
- [27] Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2012.
- [28] Yu Li and Kris V. Kowdley. Micrornas in common human diseases. *Genomics, Proteomics Bioinformatics*, 10(5):246–253, 2012.

- [29] Mariana Lagos-Quintana, Reinhard Rauhut, Abdullah Yalcin, Jutta Meyer, Winfried Lendeckel, and Thomas Tuschl. Identification of tissue-specific micrornas from mouse. *Current Biology*, 12(9):735–739, 2002.
- [30] Lee P. Lim, Nelson C. Lau, Philip Garrett-Engele, Andrew Grimson, Janell M. Schelter, John Castle, David P. Bartel, Peter S. Linsley, and Jason M. Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–773, 2005.
- [31] Mona Fani, Milad Zandi, Saeedeh Ebrahimi, Saber Soltani, and Samaneh Abbasi. The role of mirnas in covid-19 disease. *Future Virology*, 16(4):301–306, 2021.
- [32] Negar Balmeh, Samira Mahmoudi, Niloofar Mohammadi, and Anasik Karabedian-hajiabadi. Predicted therapeutic targets for covid-19 disease by inhibiting sars-cov-2 and its related receptors. *Informatics in Medicine Unlocked*, 20:100407, 2020.
- [33] Guillermo de Anda-Jáuregui and Enrique Hernández-Lemus. Computational oncology in the multi-omics era: State of the art. *Frontiers in Oncology*, 10:423, 2020.
- [34] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14:117793221989905, 2020.
- [35] Andrea Sacconi, Claudia De Vitis, Luisa de Latouliere, Simona di Martino, Francesca De Nicola, Frauke Goeman, Carla Mottini, Francesca Paolini, Michela D’Ascanio, Alberto Ricci, and et al. Multi-omic approach identifies a transcriptional network coupling innate immune response to proliferation in the blood of covid-19 cancer patients. *Cell Death & Disease*, 12(11), 2021.
- [36] Aaron J. Wilk, Madeline J. Lee, Bei Wei, Benjamin Parks, Ruoxi Pi, Giovanny J. Martínez-Colón, Thanmayi Ranganath, Nancy Q. Zhao, Shalina Taylor, Winston Becker, and et al. Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in covid-19. *Journal of Experimental Medicine*, 218(8), 2021.
- [37] cookie cutter. Cookie cutter. <https://github.com/pydanny/cookiecutter-django/>, 2021.
- [38] Anaconda software distribution, 2021.
- [39] Donghe Kang, Oliver Rübel, Suren Byna, and Spyros Blanas. Comparison of array management library performance - a neuroscience use case. *The 2019 International Conference for High Performance Computing, Networking, Storage, and Analysis (Poster session)*.
- [40] N HASTIE. The expression of three abundance classes of messenger rna in mouse tissues. *Cell*, 9(4):761–774, 1976.
- [41] Oscar González-Velasco, Dulce Papy-García, Gael Le Douaron, José M. Sánchez-Santos, and Javier De Las Rivas. Transcriptomic landscape, gene signatures and regulatory profile of aging in the human brain. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(6):194491, 2020.
- [42] Anna Supernat, Marta Popeda, Krzysztof Pastuszak, Myron G. Best, Peter Grešner, Sjors In Veld, Bartłomiej Siek, Natalia Bednarz-Knoll, Matthew T. Rondina, Tomasz Stokowy, and et al. Transcriptomic landscape of blood platelets in healthy donors. *Scientific Reports*, 11(1), 2021.

- [43] Jacek Marzec, Helen Ross-Adams, Stefano Pirrò, Jun Wang, Yanan Zhu, Xueying Mao, Emanuela Gadaleta, Amar S. Ahmad, Bernard V. North, Solène-Florence Kammerer-Jacquet, and et al. The transcriptomic landscape of prostate cancer development and progression: An integrative analysis. *Cancers*, 13(2):345, 2021.
- [44] Harshil Patel, Phil Ewels, Alexander Peltzer, Rickard Hammarén, Olga Botvinik, Gregor Sturm, Denis Moreno, Pranathi Vemuri, silviamorins, Lorena Pantano, FriederikeHanssen, Maxime U. Garcia, nf-core bot, Chris Cheshire, rfenouil, marioeppner, Peng Zhou, Gisela Gabernet, Daniel Straub, Jose Espinosa-Carrasco, Matthias Hörtenhuber, Paolo Di Tommaso, Sven F., George Hall, Senthilkumar Panneerselvam, Denis OMeally, jun wan, Alice Mayer, Katie Lennard, and Sofia Haglund. nf-core/rnaseq: nf-core/rnaseq v3.3 - bronze bear, July 2021.
- [45] Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnse. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3):276–278, 2020.
- [46] Harshil Patel and Matthias Hörtenhuber. nf-core/rnaseq: Citations. <https://github.com/nf-core/rnaseq/blob/master/CITATIONS.md>, 2021.
- [47] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014.
- [48] Nicolas F. Fernandez, Gregory W. Gundersen, Adeeb Rahman, Mark L. Grimes, Klarisa Rikova, Peter Hornbeck, and Avi Ma’ayan. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific Data*, 4(1), 2017.
- [49] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 10 2016.
- [50] Gene Cards. Gene cards - the human genes database, 2021.
- [51] Adrian Alexa and Jorg Rahnenfuhrer. topGO: Enrichment Analysis for Gene Ontology, 2019. R package version 2.37.0.
- [52] Pubmed. Pubmed, 2021.
- [53] PZMS Praeklinisches Zentrum fuer Molekulare Signalverarbeitung. Pzms praeklinisches zentrum fuer molekulare signalverarbeitung, 2021.
- [54] Dimitrios C. Mastellos, Bruno G.P. Pires da Silva, Benedito A.L. Fonseca, Natasha P. Fonseca, Maria Auxiliadora-Martins, Sara Mastaglio, Annalisa Ruggeri, Marina Sironi, Peter Radermacher, Akrivi Chrysanthopoulou, and et al. Complement c3 vs c5 inhibition in severe covid-19: Early clinical findings reveal differential biological efficacy. *Clinical Immunology*, 220:108598, 2020.
- [55] Gianfranco Mitacchione, Marco Schiavone, Antonio Curnis, Marcello Arca, Spinello Antinori, Alessio Gasperetti, Giosuè Mascioli, Paolo Severino, Federica Sabato, Maria M. Caracciolo, and et al. Impact of prior statin use on clinical outcomes in covid-19 patients: data from tertiary referral hospitals during covid-19 pandemic in italy. *Journal of Clinical Lipidology*, 15(1):68–78, 2021.
- [56] Helmholtz Institute for Pharmaceutical Research Saarland (HIPS). Helmholtz institute for pharmaceutical research saarland (hips), 2021.

- [57] Guoping Fan Lisa D Moore, Thuc Le. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(3):23–38, 2013.
- [58] Lei Fan, Xiangzhu Zhu, Yinan Zheng, Wei Zhang, Douglas L. Seidner, Reid Ness, Harvey J. Murff, Chang Yu, Xiang Huang, Martha J. Shrubsole, Lifang Hou, and Qi Dai. Magnesium treatment on methylation changes of transmembrane serine protease 2 (tmprss2). *Nutrition*, 89:111340, 2021.
- [59] Tobias Fehlmann, Fabian Kern, Omar Laham, Christina Backes, Jeffrey Solomon, Pascal Hirsch, Carsten Volz, Rolf Müller, and Andreas Keller. miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale. *Nucleic Acids Research*, 49(W1):W397–W408, 04 2021.
- [60] Fabian Kern, Tobias Fehlmann, Jeffrey Solomon, Louisa Schwed, Nadja Grammes, Christina Backes, Kendall Van Keuren-Jensen, David Wesley Craig, Eckart Meese, and Andreas Keller. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Research*, 48(W1):W521–W528, 05 2020.
- [61] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, Hsiao-Chin Hong, Yun Tang, Yi-Gang Chen, Chen-Nan Jin, Yuan Yu, Jia-Tong Xu, Yue-Ming Li, Xiao-Xuan Cai, Zhen-Yu Zhou, Xiao-Hang Chen, Yuan-Yuan Pei, Liang Hu, Jin-Jiang Su, Shi-Dong Cui, Fei Wang, Yue-Yang Xie, Si-Yuan Ding, Meng-Fan Luo, Chih-Hung Chou, Nai-Wen Chang, Kai-Wen Chen, Yu-Hsiang Cheng, Xin-Hong Wan, Wen-Lian Hsu, Tzong-Yi Lee, Feng-Xiang Wei, and Hsien-Da Huang. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Research*, 48(D1):D148–D154, 10 2019.