Identification of metabolite associated with Covid-19 disease

Wanchai Grossrieder

Genomics and Bioinformatics Miniproject 2024

Introduction

The aim of this research is to determine the molecular mechanisms underlying SARS-CoV-2 infection and disease severity. We are interested to determine whether the conclusion of the original paper from Krishnan et al. are reproducible. We are also interested in determining whether this identified mechanism could predict the disease's severity. For that, metabolite analysis has been used to determine key metabolic pathways affected by the SARS-COV-2. We have 41 PCR positive COVID-19 Patients, mixed between severe and mild and 31 healthy controls. Plasma untargeted metabolomics was used by making an ultrahigh performance LC–tandem mass spectroscopy (UPLC-MS/MS). A total of 812 metabolites have been used in the analysis.

Methods

The metabolomic data has been cleaned. All the metabolites with more than 20% of missing data have been removed. The missing data has been replaced by the mean of the metabolite.

UMAP was used with the metabolomics data. We use default python parameters for the UMAP: number of neighbors = 15, minimal distance = 0.1, metric = Euclidean and set a random seed for reproducibility.

For the hierarchical clustering we use the average method and the Euclidean distance. We used the log2 fold change between the mild hospitalized and severe hospitalized cohorts to determine the significant metabolites and applied a Z-score to the data. We use the t-test to determine the Bonferroni adjusted p-value of the metabolites between these cohorts.

We plot a volcano plot to see which metabolites are differentially present between the two cohorts. We use the log2 fold change and the p-value. The 10 most differentially changed metabolites and the 10 metabolites with the smallest adjusted p-value have their names on the plot.

PCA was used on standardized data to determine whether the metabolites can separate the different cohorts. We used support vector machines to linearly separate groups. One PCA was done on the whole metabolomics data to discriminate between the healthy and hospitalized cohort. The second PCA was done on the mild and severe cohort to see if the metabolites can separate the two cohorts. We extracted the loadings of the metabolites on the first components of the PCA to see which metabolites explain the most variance in the data.

Discussions

We showed that plasma metabolomic profiles cluster differently on a UMAP. The figure suggests here that healthy control (the two groups) and hospitalization (the two groups) have different metabolomic profiles.

However, we noticed that inside the groups the disease severity and the presence of SARS-CoV-2 antibody does not impact the metabolomic profile. Suggesting that it is difficult to determine disease severity from the metabolomic profile (Figure 1). The figure we obtained has a different scale from the original paper but has the same general shape as the original in the paper. The UMAP result is like the original paper of Krishnan.

In our analysis, we are interested in exploring the metabolites implicated in the disease severity. For that aim, we log2 transform the metabolites expression profile and Z-transform them. We isolated the metabolites that have a significantly different expression profile between the severe and the mild conditions. We used a Bonferroni adjusted p-value to determine the 67 significant metabolites. As the paper did not tell how the data was cleaned and replaced, we have a slight reduction in the number of significant metabolites.

We used hierarchical clustering to display the metabolites that are correlated together. We can see in general that the metabolites that are differently expressed between the mild and the severe groups (Yellow vs Orange) are also the metabolites that are strongly differentially expressed between the healthy control and the hospitalized control (Green vs Yellow/Orange). Therefore, metabolites that can be used to distinguish between the two severity groups can also be used to distinguish between the healthy and disease groups. We can also see that the metabolites that are differentially expressed are mostly amino acid–related and lipids-related pathways (Figure 2).

We identified the most important metabolites differently present between the two groups of severity by making a Volcano plot of the log2 fold change (Figure 3). We used similar methodologies as the paper of origin. Surprisingly, we did not get the same values as the original paper for the p-values and the fold change. This is probably because of the Bonferroni correction, as the data cleaning was not mentioned, the number of tests we performed was different than the number of the paper due to the columns we removed. However, the top significant metabolites appear to be the same, as the paper from Krishnan et al. We also obtained that the mannose is the most significantly differentially expressed. However, its fold change is reasonably low.

After that we performed a PCA to identify whether the PCA can cluster differently the expression profile of the metabolite of the different cohort. We performed two PCA, the first one with all the data, by taking the healthy control vs hospitalized. For the second one, we used a subset of the data with only the sample of individuals with the disease. The results show that the expression profile is different between the healthy control and the hospitalized cohort. With one exception the PCA data is linearly separable. Using the similar methodology but only on the mild and severe cohort we can see that the data are not linearly separable (Figure 4 & Figure 5).

This suggests that the proteomic data is not quite different between the two cohorts. Having a strong p-value in the t-test does not mean that we can conclude that a strong difference in these metabolites presence is useful to predict the condition of the patients.

We also notice that the metabolites that are the most significant in the volcano plot are not the same as the metabolites that explain the most variance in the data (Table 1 and Table 2).

Conclusion

We identified key metabolites that are differentially present between the severe and mild cohort in Table 1. The results we obtained are different from the paper, but it led to the same conclusions. By using the UMAP and the PCA we showed that the metabolites are expressed differently between the healthy and hospitalized cohort. However, the metabolites are not quite different between the mild and severe cohort. This led us to the following conclusion: the sars-cov-2 infection affects several metabolomic pathways, especially the mannose as depicted in the paper. But these are insufficient for the prediction of the severity of the disease.

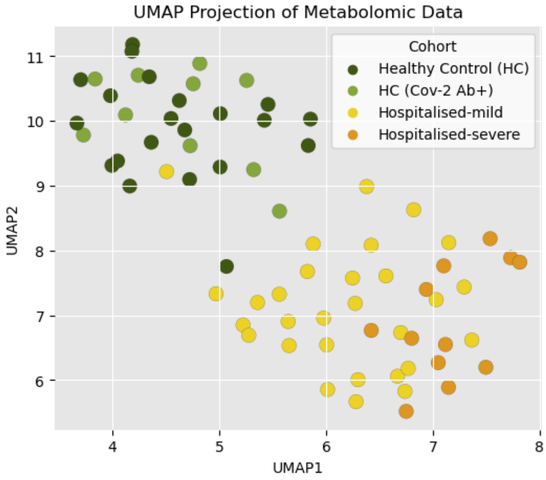


Figure 1 Sample distribution for quantitative metabolite measurements plotted in 2D space after performing dimensionality reduction using UMAP.

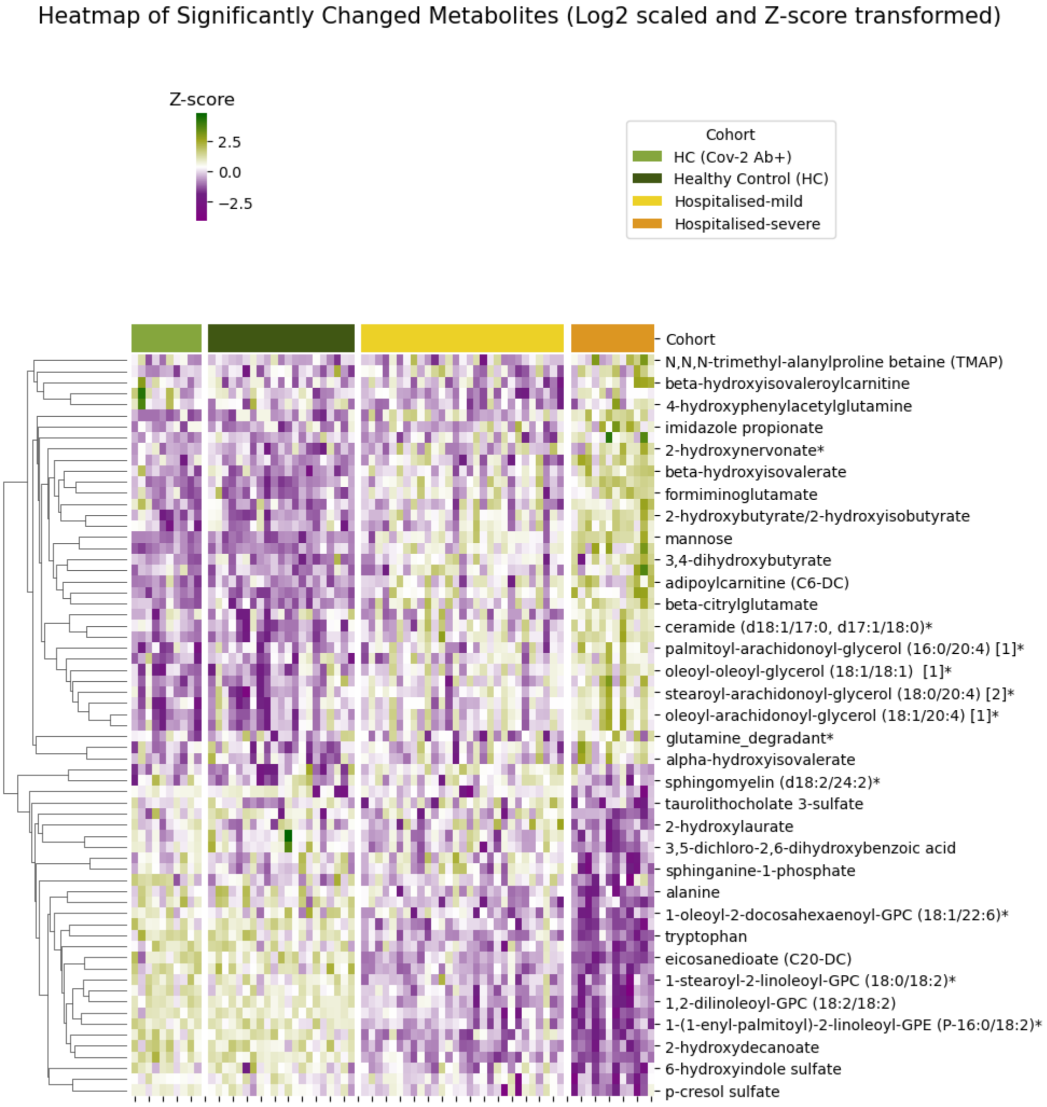


Figure 2 Heatmap of log2 scaled and Z-score transformed significantly changed metabolites between hospitalized mild and hospitalized severe groups.

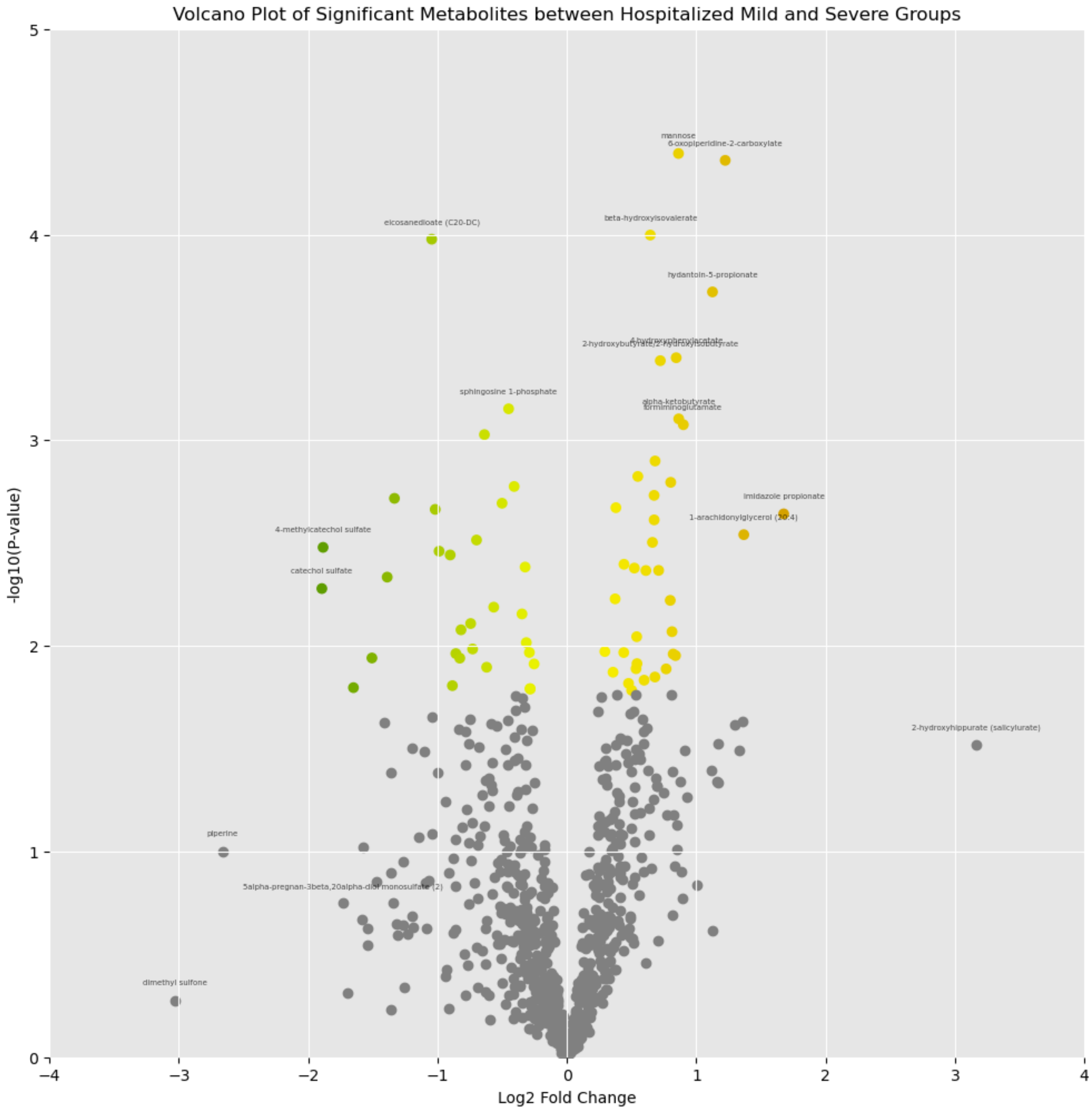
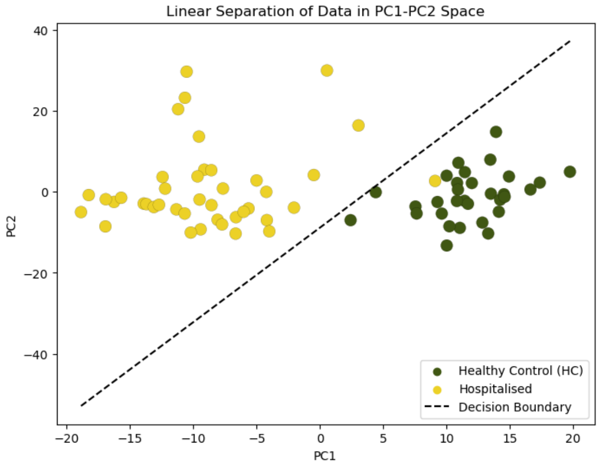


Figure 5 Volcano plot showing all the metabolites that differ significantly between hospitalized mild and hospitalized severe groups.



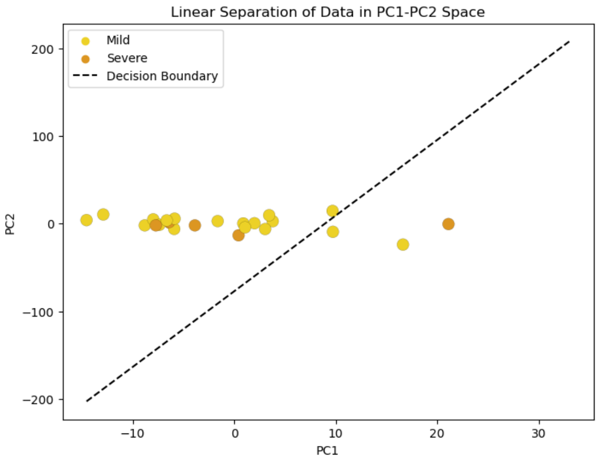


Figure 3 Principal Component Analysis of the data, we labelled the healthy control and the hospitalised. We use support vector machines to determine the decision boundary.

Figure 4 Principal Component Analysis of the data, we labelled the mild and severe hospitalised. We use support vector machines to determine the decision boundary.

Table 1 Top ten of the metabolites that have the lowest P-value in the log2 fold change between the mild and severe hospitalized cohorts.

	Metabolite	p-value	logFC
547	mannose	0.000040	0.862536
242	6-oxopiperidine-2-carboxylate	0.000044	1.223961
296	beta-hydroxyisovalerate	0.000101	0.644475
396	eicosanedioate (C20-DC)	0.000105	-1.044953
492	hydantoin-5-propionate	0.000190	1.125393
210	4-hydroxyphenylacetate	0.000397	0.844976
123	2-hydroxybutyrate/2-hydroxyisobutyrate	0.000410	0.722034
752	sphingosine 1-phosphate	0.000703	-0.450608
260	alpha-ketobutyrate	0.000787	0.864469
410	formiminoglutamate	0.000840	0.900230

Table 2 Top ten of the metabolites that have the biggest loading in the PC1. PC1 is the component that separate the most the mild vs the severe condition and therefore we are interested in the loadings that contribute the most to it.

	Metabolites	Loadings
0	1-linoleoyl-GPG (18:2)*	0.808650
1	1-palmitoyl-GPG (16:0)*	0.801397
2	N-acetylisoleucine	0.791654
3	2-stearoyl-GPE (18:0)*	0.788107
4	1-oleoyl-GPG (18:1)*	0.776658
5	dihomo-linolenoylcarnitine (C20:3n3 or 6)*	0.776348
6	1-stearoyl-GPG (18:0)	0.763511
7	palmitoyl-linolenoyl-glycerol (16:0/18:3) [2]*	0.762064
8	2-palmitoleoyl-GPC* (16:1)*	0.755737
9	ornithine	0.749007