REPLICATED RESEARCH

Replication of Large-Scale Multi-omic Analysis of COVID-19 Severity

Sriram Hathwar, Rishwanth Raghu, Emily Dale

Abstract

Background: As of December 2020, the SARS-CoV-2 virus continues to infect millions of patients worldwide. Therefore, there is a need to understand how patients from various backgrounds respond to the virus on a molecular level. We used RNA-seq and high-resolution mass spectrometry (HRMS) data collected on 128 blood samples of COVID-19 positive and COVID-19 negative patients to understand how metabolites, proteins, lipids, and transcripts are regulated in response to this viral infection.[1]

Results: Using an elastic net to perform feature selection, we determined that of the 17,287 biomolecules measured across COVID and non-COVID patients, there were 219 features deemed important in determining severity and outcome of disease. Further annotation and enrichment analysis of these features showed complement activation and neutrophil degranulation to be significantly up-regulated, while platelet function and blood coagulation were shown to be significantly down-regulated in COVID-19 positive patients.

Conclusion: The identification of enriched molecular features allows for better understanding of host biochemical pathways that the virus infects. In doing so, we construct a particular phenotype for COVID-19 patients, allowing us to suggest therapeutic points. Moreover, for a more public facing tool, we show efficacy of our machine learning algorithm to predict COVID-19 severity.

Background

The SARS-CoV-2 virus that causes the disease COVID-19 has caused roughly 1.5 million deaths worldwide as of December 2020.[2] Patient responses to the virus vary significantly, from purportedly asymptomatic state to ICU hospitalization. Several environmental factors such as age, gender, geography, and social inequalities are known to increase likelihood of contraction, and subsequently, severity of cases.[3] While symptoms such as sepsis, cough, and inflammatory response are rather significant to the naked eye,[4] the biomolecular signature of COVID-19 positive patients are lacking in the growing scientific literature.[1]

Using nucleic acid sequencing technologies and high-resolution mass spectroscopy data, the goal of this study is to take broad measurements of protein, lipid, metabolite, and transcript levels in blood samples of COVID-19 positive and COVID-19 negative patients so as to determine which biological processes are most heavily regulated by the virus. Taking measurements across multiple "omes" allows for a holistic characterization of the COVID-19 phenotype, which could inform future experiments to develop therapeutics.[1] In doing so, we hope to recapitulate known COVID-19 disease mechanisms while also illuminating new ones.

Experimental Design

From April 6, 2020 to May 1, 2020, blood samples were collected from 128 adults who reported severe respiratory issues and were admitted to the medical floor or ICU at the Albany Medical Center. Blood samples were collected at time of enrollment and were categorized based on test results. As such, there were n = 102 COVID-19 positive patients and n = 26 COVID-19 negative patients total in this study. In addition to the clinical data provided upon entry, metadata—including length of stay in hospital, ventilation status, ICU status, acute physiological assessment and chronic health evaluation (APACHE II) score, Charlson co-morbidity index, sequential organ failure assessment (SOFA) score, and specific laboratory measurements like ferritin and fibrinogen— were all collected. In essence, the goal here was to collect as much metadata to serve as the predicted variable for future regression analyses.

Notably, the COVID-19 positive group was more racially diverse than the negative group. It is widely cited that COVID-19 disproportionately affects poorer populations and people of color,[5] so the ethnic health disparities are highlighted in this study.

Additionally, average age of COVID-19 positive patients and COVID-19 negative patients were roughly similar; the mean COVID-19 positive age was 61.3 years, while the mean COVID-19 negative age was 63.8 years. While people over 18 could participate in the study, few young people contract the virus,[6] consistent with these data.

Additionally, of note, there was a slight sex difference in COVID-19 severity. In particular, more COVID-19 positive men were likely to be hospitalized in the ICU and put on a ventilator than their female counterparts (Fig. 1) despite having relatively similar age and ethnicity distribution. One can only speculate the cause of this phenomenon.

Key Methods

Once blood samples were collected, they were stored at -80°C and aliquoted into 100µL samples by biomolecule class. To determine a broad range of lipids, metabolites, proteins, and transcripts, high-throughput sequencing technologies were applied specific to each "-ome."

Specifically, for metabolomics data processing, gas chromatography- mass spectrometry (GC-MS) raw profiles were processed. Only feature groups that had at least 10 fragment ions and that were found to be triple the background were retained for further analysis. Blood sample proteins were determined through shotgun proteomics. Raw tandem MS scans were then used to quantify protein levels in the data processing stage. Liquid-chromatography mass spectro-metry (LC-MS) was used to determine the presence of various lipids. Once again, 3-fold intensity increase over the background was used as the metric for keeping lipids. Finally, transcripts were determined using Illumina RNA-seq.[1]

In terms of my work, I performed standard unsupervised methods on all these biomolecular features, including PCA using the R method prcomp() (Figure 2A) and volcano plots (Figure 2B). Then, I used an elastic net to perform feature selection; in other words, I hoped to find the most important biomolecular features. I determined that of the 17,287 initial biomolecules, only 219 were significantly correlated with COVID-19 status and hospital free days at day 45 (Figure 2C) using the regression HFD-45 ~ biomolecule abundance + age + sex.

Each of the identified biomolecules were then enriched through GO term annotation and correlated to COVID-19 status and severity, and I plotted the top 7 up-regulated and top 7 down-regulated GO terms by q-value (Figure 2D, 2E). Finally, my colleagues created an ExtraTrees ML classifier to predict COVID severity.

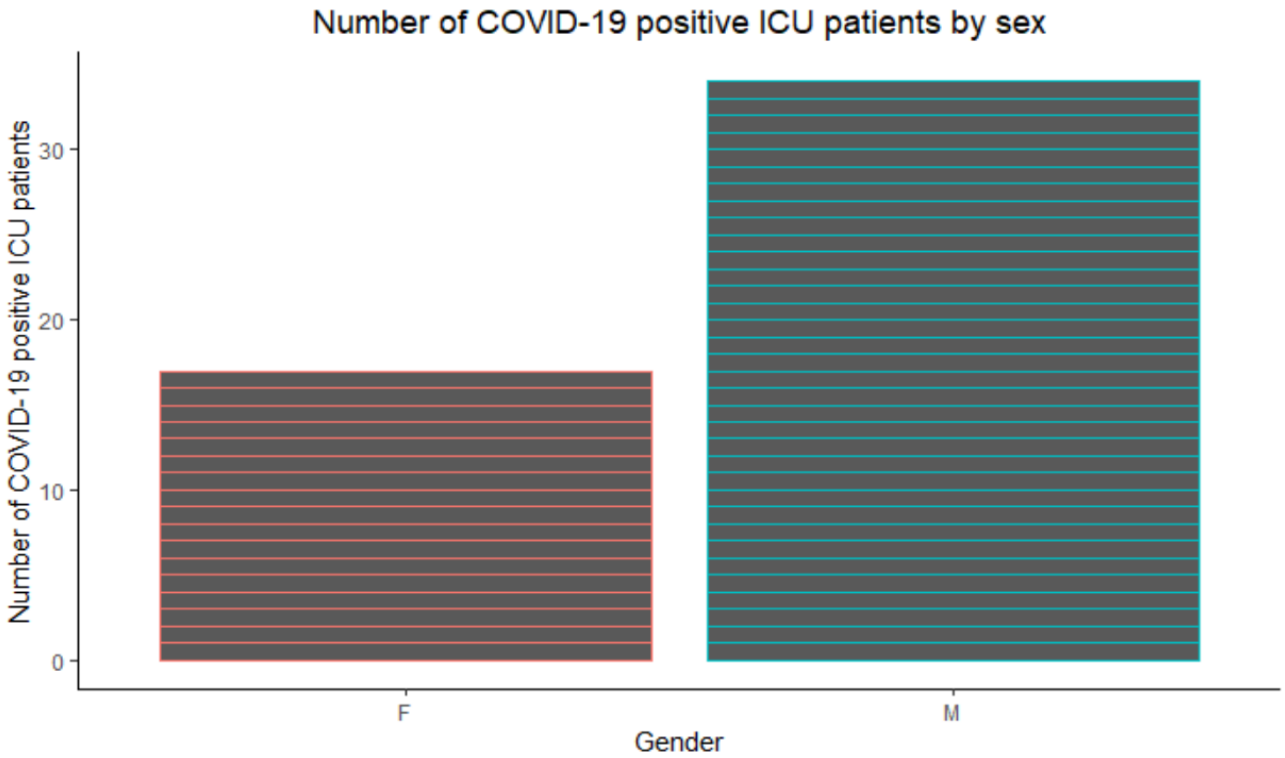
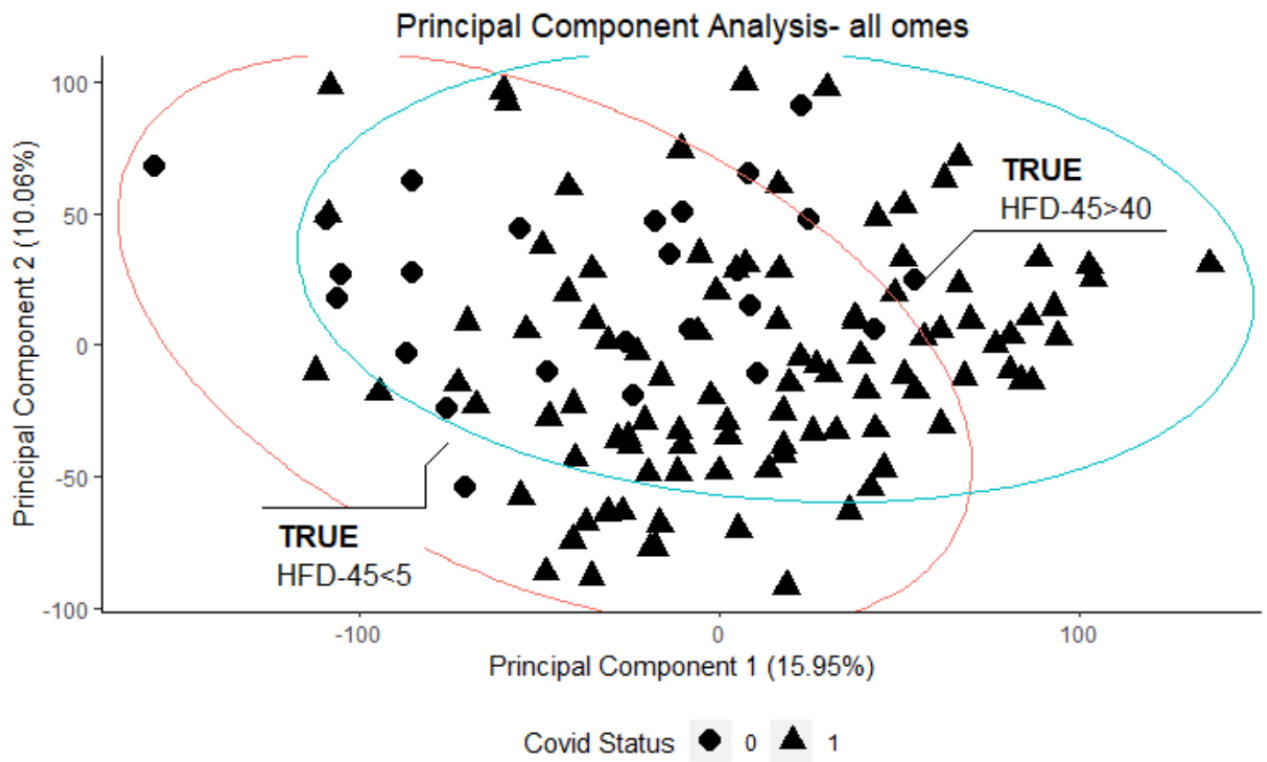


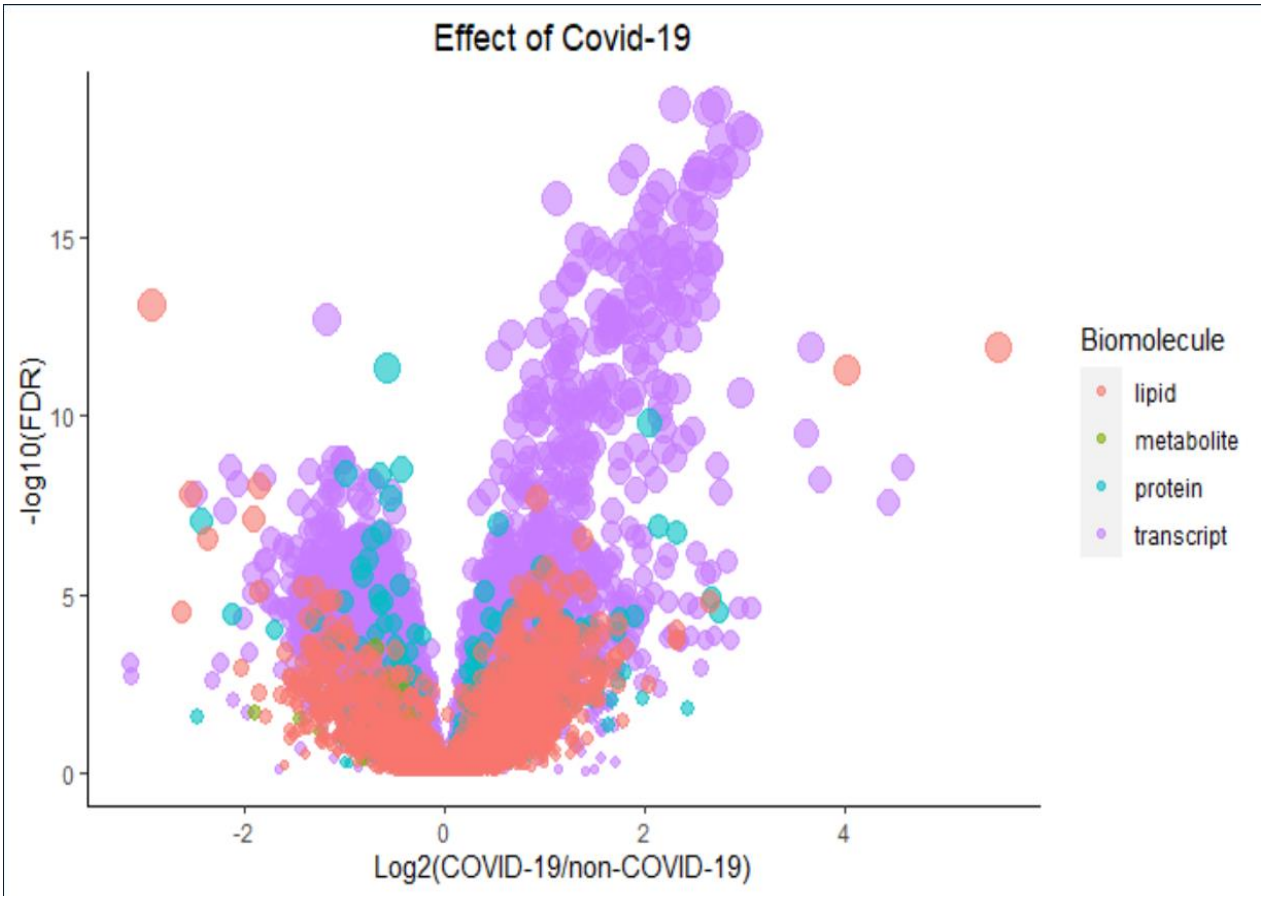
Figure 1. Sample Cohort and Experimental Design Analysis

Male COVID-19 positive patients were significantly more likely to be admitted to the ICU than their female counterparts. Produced by Sriram Hathwar.

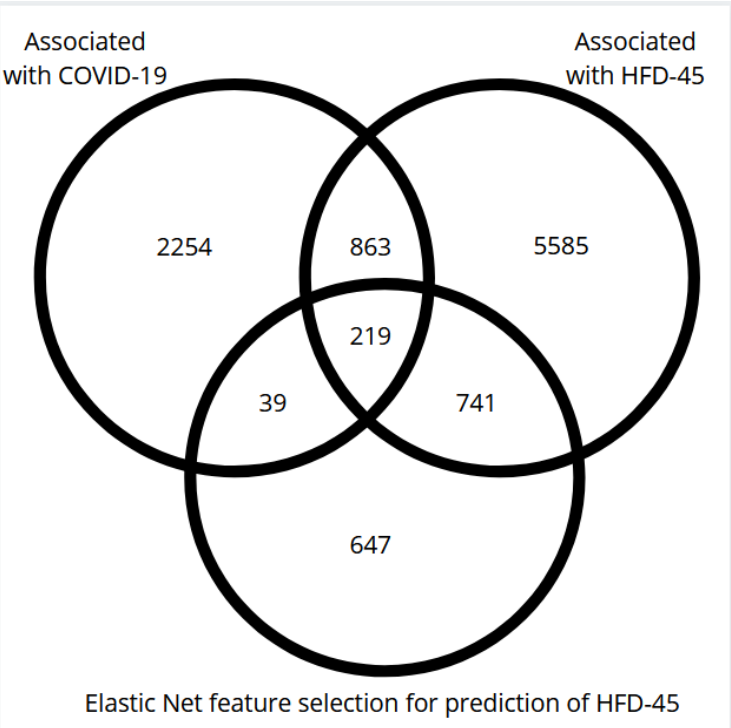
A



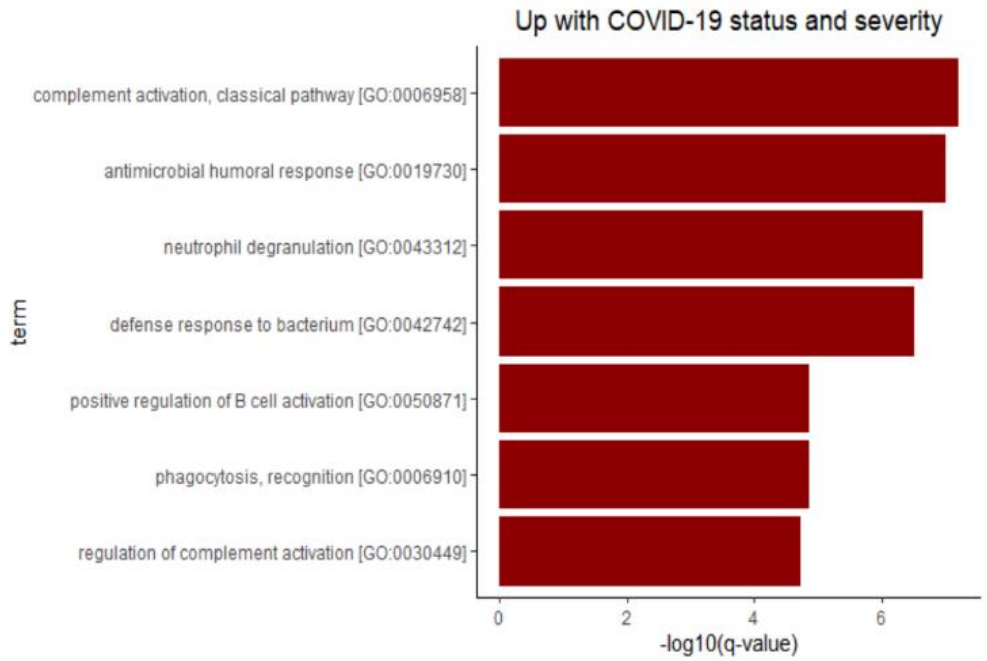
B


Effect of Covid-19
-log10(FDR)
15
10
5
0
-2
0
2
4
Log2(COVID-19/non-COVID-19)
Biomolecule
lipid
metabolite
protein
transcript

C


Associated with COVID-19
Associated with HFD-45
2254
863
5585
219
39
741
647
Elastic Net feature selection for prediction of HFD-45

D



E

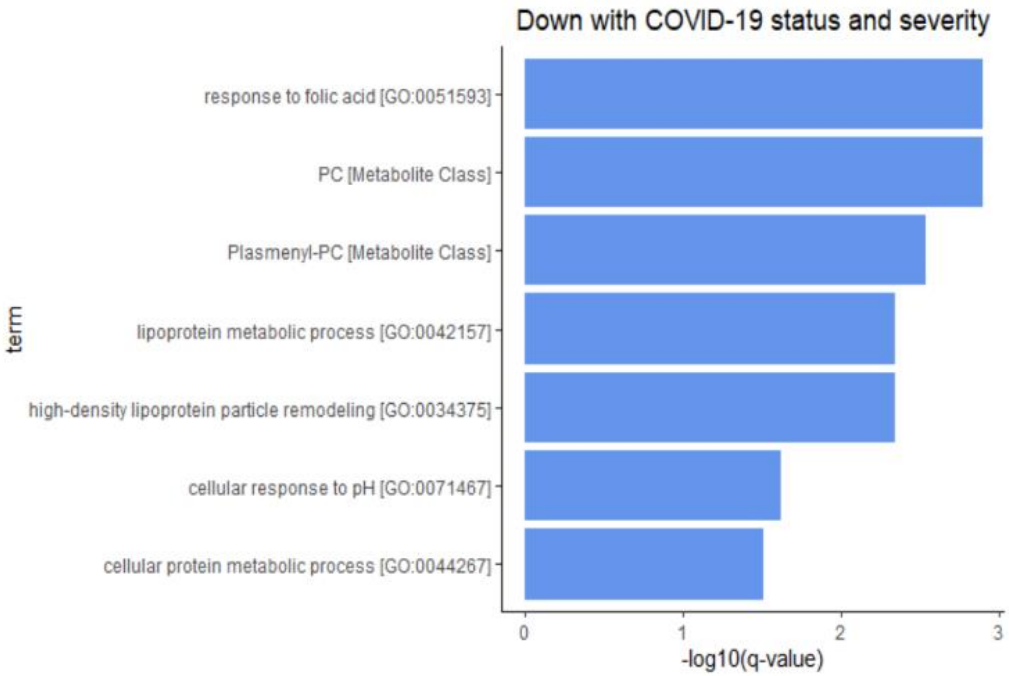


Figure 2. Strong Biomolecular Signatures Associated with COVID-19 Status and Severity

(**A**) PCA with all 17,827 biomolecules. A notable linear trend of hospital-free days at 45 days (HFD-45) appears. (**B**) Volcano plot of adjusted p value vs. $\log_2$ fold enrichment. (**C**) Results of elastic net revealed 219 biomolecules significantly associated with COVID-19 status and severity. (**D**) Top 7 GO terms up-regulated in COVID-19 patients. (**E**) Top 7 GO terms up-regulated in COVID-19 patients. Produced by Sriram Hathwar.

Results

After performing all the necessary sequencing and mass spectrometry, ultimately 17,827 biomolecules were identified across all blood samples. Specifically, 517 proteins, 12,263 transcripts, 646 lipids, 150 metabolites, and 2,786 unidentified lipids were the most abundantly measured across all samples. Such a large array of biomolecules from a diverse cohort of positive and negative patients enables effective determination of the biological processes most affected by SARS-CoV-2.

We then performed a principal component analysis (Figure 2A) and found significant grouping of patient samples based on hospital-free days at day 45 (HFD-45). Notably, HFD-45 was much lower for COVID-19 patients than non-COVID patients, attesting to the severity of the novel infection. Moreover, HFD-45 correlated rather strongly with the principal components. Additionally, as represented by the shapes on the plot, there is a subtler, yet still significant, grouping of patients by COVID-19 status. These two factors motivate a further exploration.

Subsequently, we conducted a multivariate regression based on these factors as well as known factors like age and sex to determine the features most associated with COVID-19 status and HFD-45. Using ANOVA analysis, we found significant changes in levels of proteins, lipid, and plasma of COVID-19 patient plasma. Once again, our formulas included the potentially confounding variables of age and sex. We find that upon plotting adjusted p-values (false discovery rate) versus $\log_2$ fold change of means for COVID-19 versus non-COVID-19 groups for each biomolecule, many transcripts are significantly associated with COVID-19 status, as demonstrated by the purple in the upper right corner (Figure 2B).

Interestingly, we found that the second most abundant feature in the COVID-19 positive plasma across patient samples was azithromycin (Z-Pak), a prescribed drug which a large number of patients had taken to alleviate their COVID-19 symptoms. This finding attests to the high sensitivity of mass spectrometry to detect traces of drugs.

HFD-45 serves as our proxy for COVID-19 severity, as the closer HFD-45 is to zero, the more severe the outcome. I then performed multi-variate linear regression on HFD-45 using the elastic net penalty. This is a common approach used to select only the most important features, as the algorithm sends the coefficients of unimportant features to zero. Notably, I was able to replicate the work of the original authors to find the 219 biomolecules most significantly associated with COVID-19 status and severity (Figure 2C), but only after looking at the grid space from the authors' GitHub code.[7] Because of the finickiness of hyperparameter tuning, I figured I would try to optimize using the same linear space. Nevertheless, the ability to replicate the work of the original authors illustrates the profundity of the findings, as these 219 identified features could lend insight into future therapeutics. Yet, noting above the presence of Z-Pak as a significant feature, we must be careful in determining the actual biological molecules.

Using the Uniprot database,[8] the researchers applied gene ontology (GO) term and molecular enrichment to determine differential expression between COVID-19 positive and negative patients. In my replication, I did not confirm these features in the Uniprot database, and instead just opted to use the annotated terms provided in the supplementary data. Given the 219 features, a simple database search for the corresponding proteins and transcripts should yield the same results as the original work.

Using the original researchers' GO term enrichment data, we determined the biological processes that were up-regulated in COVID-19 positive patients (Figure 2E) as well as biological processes that were down-regulated in these patients (Figure 2F). These results confirmed notable processes in the literature. In particular, complement activation, anti-microbial response, and neutrophil degranulation were the top three GO terms enriched with COVID-19 status.

Complement activation is a broad term for immunoglobulin gene recombination.[9] Because of its role in the immune system, biomolecules involved in complement activation have been posited as potential therapeutic targets. Other features that has increased enrichment between the COVID-19 positive patient and the negative counterpart include antimicrobial humoral response and neutrophil degranulation. These are both negative consequences of viral load in the body, although the mechanisms for the microbial response and antibody response are the work of future studies.

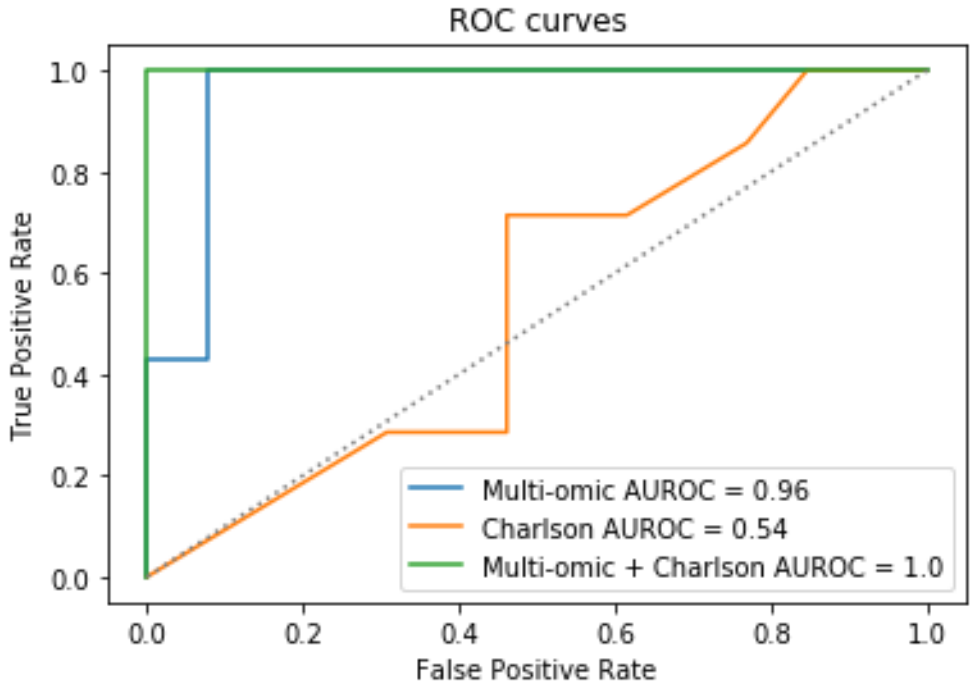
Biological processes down-regulated in COVID-19 patients were less statistically significant ($-\log_{10}$ of q-value = 3) than the up-regulated processes ($-\log_{10}$ of q-value = 7), but these were significant nevertheless. The down-regulated factors, such as decreased plasma gelsolin (pGSN), could serve as important biomarkers for COVID-19 status and disease progression. pGSN specifically is associated with acute lung injury, which corresponds to the severe pulmonary symptoms of a COVID-19 infection.[10] In fact, a recombinant form of plasma gelsolin has been administered in phase I/II trials.[11] Additionally, diminished response to folic acid is found in COVID-19 patients; folate deficiency also corresponds to other inflammatory settings, such as IBD and Celiac disease.[12]

My colleagues continued analysis based on my work to perform cross-ome correlation analysis. In doing so, Rishwanth Raghu found significant correlation between proteins and small molecules. Another round of unsupervised clustering of the data using heatmaps shows how these 219 features have high Kendall-Tau correlation coefficients. Notably in this analysis, additional medicinal metabolites were found in patient samples, such as citrate, which is used to prevent coagulation caused by COVID-19. Once again, sensitivity to determine drug metabolites is quite remarkable, but this points to the need to differentiate drug metabolites from actual biomolecules.

Furthermore, Emily Dale continued analysis on specific biological processes that are down-regulated by COVID-19. As mentioned earlier, analyzing the down-regulated metabolites is of particular use as these can often serve as biomarkers for disease prognosis and progression. Her analysis builds on the 219 features originally identified by my elastic net.

Finally, my colleagues created an ExtraTrees ML classifier using multi-omic data to determine COVID-19 severity. Granted, obtaining all omic data from a hospitalized COVID-19 patient is generally not quite useful in the real world, where the focus should be on effective bedside treatment. In this model, accuracy on test set data was determined by five-fold cross validation (Figure 6D, 6E). Identification of specific biomolecules associated with COVID-19 status and severity are relevant in determining a more facile prognosis and therapeutic. However, because we are dealing with human cases, it seems highly unlikely that all drugs that people might take to mitigate COVID-19 symptoms, such as hydroxychloroquine, would be accounted for in this model.

D



E

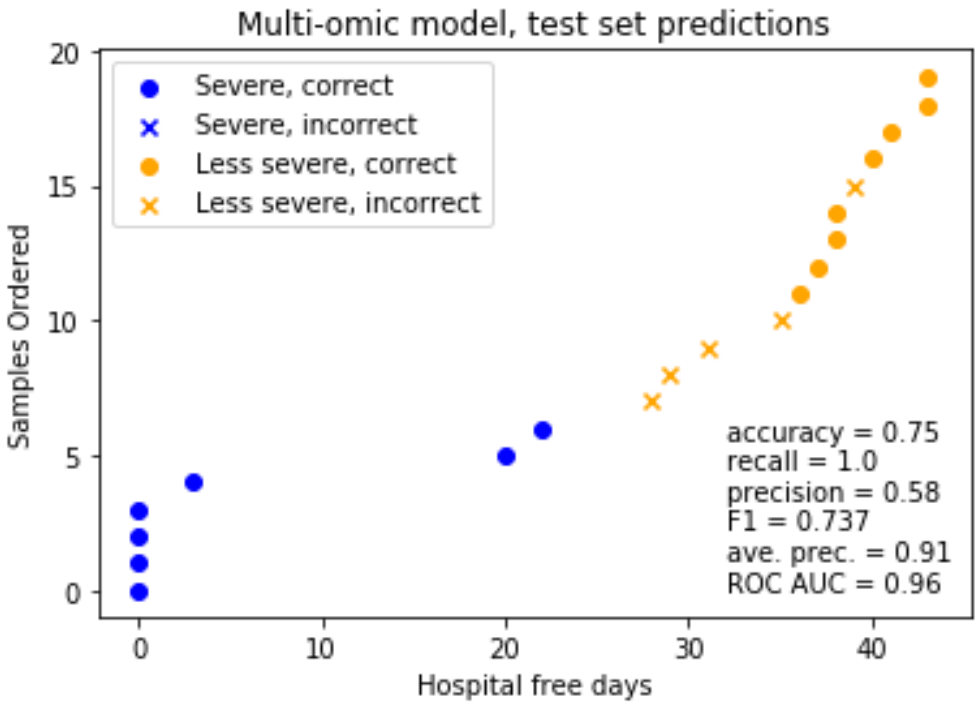


Figure 6. ExtraTrees classifier to predict COVID-19 severity based on multi-omic data

(**D**) Receiver-operator curves (ROC) for ExtraTrees classifier models trained with all biomolecule data, Charlson co-morbidity index, or both. Here, the multi-omic +_Charlson AUROC = 1.0, implying 100% accuracy in classifying positives and negatives. (**E**) Predictions of the ExtraTrees classifier on test set. Produced by Rishwanth Raghu.

Discussion

This cohort study applies cutting-edge sequencing and mass spectrometry technologies to better understand the biomolecular signatures of COVID-19 positive and COVID-19 negative patients. Using data from lipidomics, proteomics, transcriptomics, and metabolomics theoretically provides a holistic COVID-19 phenotype, as opposed to a list of external symptoms like pulmonary inflammation and anosmia.

Broadly, the identification of these 219 enriched molecular features provide a springboard for further analysis. Although some of the enriched molecular features were medicinal metabolites (Z-Pak), the transcripts, proteins, lipids, and metabolites discovered through this sequencing provide a basis for understanding the biochemical pathways the virus uses to attack its host.

One notable example of this is cited in Rishwanth Raghu's portion of the paper which shows that lower levels of the anti-coagulant citrate are highly correlated with COVID-19. This recapitulates known biology, as the inflammatory response is triggered in response to the virus in symptomatic patients, increasing blood clotting rates. Yet, citrate has been shown to mitigate the deleterious effects of neutrophil degranulation, a significantly under-enriched biological process in COVID-19 patients. Hence, in looking across a broad spectrum of biomolecules, certain therapeutics like citrate can be determined.

In terms of actual experimental design, there are a few notable shortcomings of this study. One of the most notable pitfalls is the regionality of the study. While there is a relatively diverse population in Albany, NY, the confounding variable of geography might play a role in the types of biomolecules most seen. For instance, it could be the case that certain metabolites are found in the water near Albany, which could lead to much more correlation between samples than would be expected in a statewide study, let alone a national or international study.

In addition to increasing breadth of the patient samples, this analysis would benefit a lot from a larger sample size. As the Central Limit Theorem states, the larger the sample size, the more accurate the sampled data represents the population. Because of how much patient responses vary to COVID-19, it might be useful to extend this study to other areas heavily infected, like Europe or India.

That all said, the scale of this study is still remarkable for a first look at the biomolecular processes underlying COVID-19.

As future directions go, it would be useful to perform the same multi-omic analysis on recovered patients as a follow-up. In doing so, it would shed light on the types of biomolecules that are retained in the immune and circulatory system after convalescence. Specifically, conducting the multi-omic analysis at 1 month and 3 months post-recovery could shed light on the antibodies needed for developing therapeutic sera, some of which have already been approved for EUA as of December, 2020.[13]

Another interesting note is that MIT researchers have developed an AI algorithm to listen to patients' coughs and determine whether they have COVID-19 or not. Their detection algorithm has 100% specificity and 83% sensitivity, as their algorithm can distinguish 1000 features in a cough vs. the human ear that can only distinguish roughly 30 features.[14] It would be interesting to pair into one app the MIT cough algorithm and this paper's multi-omic ExtraTrees Classifier for the average individual to both test for whether (s)he have COVID and to know how severe the disease progression will be.

Author Contributions

Sriram Hathwar wrote the entirety of this report. He wrote the code to reproduce Figure 1, which is a simple demographic display of ICU hospitalization, as well as the code for Figure 2, which focuses on the correlations between the 17,287 biomolecule abundances and COVID-19 severity. While Sriram Hathwar did not do any collecting of omic measurements, he analyzed the data and supplementary data to reproduce the figures in the original paper from scratch.

Rishwanth Raghu wrote the code to reproduce Figure 3, which focuses on cross-ome correlations, specifically between protein abundances and lipid and metabolite abundances. He also wrote the code to reproduce 6D and 6E, which provides accuracy metrics for a multi-omic ExtraTrees machine learning algorithm used to determine COVID-19 severity. His figures 6D and 6E are included above and credited to him in this report.

Emily Dale wrote the code to reproduce Figure 4, which focuses on COVID-19 related changes in GO annotated biological processes. She also wrote the code to reproduce 6A, 6B, and 6C, which is the multi-omic ExtraTrees machine learning algorithm.

All 3 authors apportioned the figures evenly and contributed accordingly.

Competing Interests

The authors declare no competing interests.

Acknowledgments

This final project was an extremely insightful and engaging exercise to understand how computational sequencing technologies and mass spectrometry can be used to understand disease mechanisms. Currently taking this class during the COVID-19 pandemic, I find it extremely rewarding, yet humbling, to determine the biomolecular signatures of SARS-CoV-2. I would first like to acknowledge the authors of this paper, Katherine A. Overmyer et. al, for performing the biological experiments and collecting the data for interesting analyses.

I would also like to thank Professor Singh, Professor Akey, and Professor White for their stellar guidance both in the classroom and in office hours during this time. In addition, I would like to thank the lab TAs Riley Skeen-Gaar, Antonio Muscarella, and Tyler Park for their prompt responses to Ed posts and classmate questions. Lastly, I would like to thank Princeton University for creating a strong academic curriculum during this time.

References

1. Overmyer, Katherine A., et al. "Large-scale multi-omic analysis of COVID-19 severity." *Cell systems* (2020).
2. "WHO Coronavirus Disease (COVID-19) Dashboard." *World Health Organization*, World Health Organization, 8 Dec. 2020, covid19.who.int/table.
3. Verity, Robert, et al. "Estimates of the severity of coronavirus disease 2019: a model-based analysis." *The Lancet infectious diseases* (2020).
4. Alhazzani, Waleed, et al. "Surviving Sepsis Campaign: guidelines on the management of critically ill adults with Coronavirus Disease 2019 (COVID-19)." *Intensive care medicine* (2020): 1-34.

5. Chowkwanyun, Merlin, and Adolph L. Reed Jr. "Racial health disparities and Covid-19—caution and context." *New England Journal of Medicine* (2020).
6. Liao, Jiaqiang, et al. "Epidemiological and clinical characteristics of COVID-19 in adolescents and young adults." *The Innovation* 1.1 (2020): 100001.
7. Miller, Ian J., et. al Charles, COVID-19 Multi-Omics, (2020), GitHub repository, https://github.com/ijmiller2/COVID-19_Multi-Omics/
8. UniProt Consortium. "UniProt: a hub for protein information." *Nucleic acids research* 43.D1 (2015): D204-D212.
9. Noris, Marina, and Giuseppe Remuzzi. "Overview of complement activation and regulation." *Seminars in nephrology*. Vol. 33. No. 6. WB Saunders, 2013.
10. Peddada, Nagesh, Amin Sagar, and Renu Garg. "Plasma gelsolin: a general prognostic marker of health." *Medical hypotheses* 78.2 (2012): 203-210.
11. Tannous, Abla, et al. "Safety and Pharmacokinetics of Recombinant Human Plasma Gelsolin in Patients Hospitalized for Nonsevere Community-Acquired Pneumonia." *Antimicrobial Agents and Chemotherapy* 64.10 (2020).
12. Dahele, Anna, and Subrata Ghosh. "Vitamin B12 deficiency in untreated celiac disease." *The American journal of gastroenterology* 96.3 (2001): 745-750.
13. Tanne, Janice Hopkins. "Covid-19: FDA approves use of convalescent plasma to treat critically ill patients." *Bmj* 368 (2020): m1256.
14. Laguarta, Jordi, Ferran Hueto, and Brian Subirana. "COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings." *IEEE Open Journal of Engineering in Medicine and Biology* (2020).