Plasma proteomics of SARS-CoV-2 infection and severity reveals impact on Alzheimer's and coronary disease pathways

Lihua Wang
Washington University School of Medicine in St. Louis
Daniel Western
Washington University School of Medicine in St. Louis
Jigyasha Timsina
Washington University School of Medicine in St. Louis
Charlie Repaci
Washington University School of Medicine in St. Louis
Joanne Norton
Washington University School of Medicine in St. Louis

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.

Recommended Citation
Wang, Lihua; Western, Daniel; Timsina, Jigyasha; Repaci, Charlie; Norton, Joanne; Kohlfeld, Pat; Budde, John; Butt, Omar H; Templeton, Alan R; O'Halloran, Jane; Presti, Rachel; Goss, Charles W; Mudd, Philip A; Ances, Beau M; Sung, Yun Ju; Cruchaga, Carlos; and et al., "Plasma proteomics of SARS-CoV-2 infection and severity reveals impact on Alzheimer's and coronary disease pathways." iScience. 26, 4. 106408 (2023).
https://digitalcommons.wustl.edu/oa_4/2420

This Open Access Publication is brought to you for free and open access by the Open Access Publications at Digital Commons@Becker. It has been accepted for inclusion in 2020-Current year OA Pubs by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors
Plasma proteomics of SARS-CoV-2 infection and severity reveals impact on Alzheimer’s and coronary disease pathways

Lihua Wang, Daniel Western, Jigyasha Timsina, ..., Bin Zhang, Yun Ju Sung, Carlos Cruchaga

Highlights
1,449 plasma proteins are altered by COVID-19 infection, ventilation, or death
Distinct prognostic models for ventilation and death after COVID-19 infection
Altered proteins link to Alzheimer’s and coronary artery disease pathways
Eight proteins are potentially causal for COVID-19 infection

Wang et al., iScience 26, 106408 April 21, 2023 @ 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.106408
Plasma proteomics of SARS-CoV-2 infection and severity reveals impact on Alzheimer’s and coronary disease pathways

Lihua Wang,1,2,13 Daniel Western,1,2,13 Jigyasha Timsina,1,2 Charlie Repaci,1,2 Won-Min Song,3 Joanne Norton,1,2 Pat Kohlfeld,1,2 John Budde,1,2 Sharlee Climer,4 Omar H. Butt,5 Daniel Jacobson,6 Michael Garvin,6 Alan R. Templeton,7 Shawn Campagna,8 Jane O’Halloran,9 Rachel Presti,9 Charles W. Goss,10 Philip A. Mudd,11 Beau M. Ances,11 Bin Zhang,3 Yun Ju Sung,1,2,10,14 and Carlos Cruchaga1,2,12,14,15,*

SUMMARY
Identification of proteins dysregulated by COVID-19 infection is critically important for better understanding of its pathophysiology, building prognostic models, and identifying new targets. Plasma proteomic profiling of 4,301 proteins was performed in two independent datasets and tested for the association for three COVID-19 outcomes (infection, ventilation, and death). We identified 1,449 proteins consistently associated in both datasets with any of these three outcomes. We subsequently created highly accurate models that distinctively predict infection, ventilation, and death. These proteins were enriched in specific biological processes including cytokine signaling, Alzheimer’s disease, and coronary artery disease. Mendelian randomization and gene network analyses identified eight causal proteins and 141 highly connected hub proteins including 35 with known drug targets. Our findings provide distinctive prognostic biomarkers for two severe COVID-19 outcomes, reveal their relationship to Alzheimer’s disease and coronary artery disease, and identify potential therapeutic targets for COVID-19 outcomes.

INTRODUCTION
The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that causes coronavirus disease 2019 (COVID-19) has proven to be an extremely contagious and deadly threat to human health in the past three years. As of January 2023, the virus has caused over 663 million COVID-19 cases worldwide, leading to 6.7 million deaths and countless other long-term health effects (https://covid19.who.int/). The clinical presentation of COVID-19 infection is highly heterogeneous, ranging from asymptomatic or mild symptoms to critical conditions (e.g., cytokine storm, cardiovascular damage, pneumonia, and death). About 20% of hospitalized COVID-19 patients become critically ill, requiring oxygen therapy and/or ventilation. More importantly, 20–40% of severe COVID-19 cases1 have an alarmingly low oxygen saturation level (~50–80%) without respiratory distress, designated as silent hypoxia.4 Compared to patients with symptomatic hypoxia, these patients are less likely to seek medical attention resulting in delayed diagnosis and treatment. This distinctive variation of severe COVID-19 manifestation is associated with worse damage within the patients.1 Even after individuals recover from the short-term acute phase of the disease, long-term post-COVID-19 symptoms, such as fatigue (28.44%), cognitive dysfunction (4.14%), and shortness of breath (22.55%), often persist for many months.4

Contrary to some studies indicating that SARS-CoV-2 was absent from brain5 and cerebrospinal fluid (CSF),6 several studies have provided strong lines of evidence that SARS-CoV-2 may directly impact brain function.6–8 COVID-19 has negative long-term impacts on the central nervous system (CNS) resulting in a variety of neurological symptoms such as altered taste and smell (observed in ~15–23% of COVID-19 patients)5,10 memory decline (“brain fog”; observed in 1 out of 3 patients),11 and neurologic manifestations (observed in ~36% of COVID-19 cases).8 This suggests that COVID-19 infection may result in brain injury and neurocognitive impairment even in individuals without pre-existing neurological conditions. In addition, SARS-CoV-2 affects the myocardium and causes myocarditis and ischemia in 8–12% of COVID-19
admitted patients. Long-term post-COVID-19 syndrome is shown to be involved with the cardiovascular system (elevated cardiac troponin T). All these findings together support the long-term impact of COVID-19 infection on CNS dysfunction and impairment of the cardiovascular system.

Proteomic studies offer a powerful strategy to identify factors affecting COVID-19 infection and severity and to better understand the influence of SARS-CoV-2 on biological processes. Similar approaches have been used for aging, cancer, kidney diseases and other viral infections, including Chikungunya virus and avian influenza virus. Several small-scale proteomic studies found that patients infected with SARS-CoV-2 had elevated pro-inflammatory cytokines including IL-6, IL-1β, IL-10, and TNF. Proteins associated with neutrophil degranulation have also been shown to be upregulated in COVID-19 patients, suggesting that targeting these pathways in conjunction with antiviral medicine may provide an effective treatment for severe cases. Although these and other proteomic studies identified proteins significantly associated with COVID-19 status, they were limited in sample size (<100 participants) or breadth of proteins measured (<200 proteins), reducing power for detecting potentially significant contributing factors to disease outcome and severity.

To understand the pathogenic events linked to COVID-19 infection and identify potential biomarkers and drug targets, we obtained high throughput proteomic signatures of COVID-19 susceptibility and severity by studying more than 4,300 plasma protein levels obtained after hospital admission but before ventilation or death. Proteins associated with infection and severity in the discovery cohort were replicated within an independent cohort. The replicated proteins were then used to create a prediction model for infection (all cases versus healthy controls), ventilation (cases requiring ventilation versus cases without ventilation support), and death (cases that died versus cases that survived) as well as to identify what biological pathways are dysregulated because of infection.

RESULTS
Study design
To identify plasma proteins associated with COVID-19 infection and severity, create prediction models and identify pathways dysregulated with infection, we generated high-throughput proteomic data from 332 COVID-19 patients and 150 controls at Washington University in St Louis (WUSTL; Figure 1). Among the COVID-19 cases, there were ~90% symptomatic patients, 93.7% were hospitalized, 46.7% with ICU admission, 24.7% on ventilation, and 19% died because of COVID-19 (82 ventilated and 63 died; 44 of the deceased had been ventilated before death). COVID-19 patients were 59 years old on average, 58.7% men and 67.8% of African American ancestry (Tables 1 and S1). The control group was age-, sex-, and race-matched. Plasma was obtained when admitted to Barnes-Jewish Hospital (BJH) and the event of ventilation and death occurred on average 6.8 and 36.2 days, respectively, after hospitalization (Figure S1). Proteomic data was generated using the SomaScan v4.1 7K panel. After rigorous quality control (QC), a total of 7,055 proteins remained. However, because the replication cohort had 4,301 proteins that were measured and passed QC, these 4,301 were used in the analyses (STAR Methods; Figure S2).

For replication, we leveraged the publicly available proteomics data from 297 COVID-19 cases (84 ventilated and 41 died; 21 of the deceased had been ventilated prior death) and 76 controls recruited at Massachusetts General Hospital (MGH, Table S1). Instead of years, only age range was available for the MGH samples. In controls, 44% were in the age group 64–79. The COVID-19 cases were younger, with 36% in the age group 49–64. The need for ventilation and death occurred on average 0.5 and 18.2 days, respectively, after hospitalization, which is earlier than in the discovery cohort (Figure S1). Proteomic data for the MGH dataset was generated using the SomaScan platform, but in the 5K panel, where 4,301 proteins passed QC (STAR Methods).

To robustly identify proteins associated with COVID-19 infection, ventilation and death, we pursued a three-stage study design (discovery, replication, and meta-analysis). We performed differential abundance analysis for each COVID-19 outcome in both the discovery and replication cohorts. We performed a fixed-effect meta-analysis, combining discovery and replication. First, we identified proteins associated with each COVID-19 outcome at Benjamini-Hochberg false discovery rate (FDR) < 0.05 in the WUSTL discovery cohort. Second, within the proteins significantly identified from discovery, we identified those proteins that also associated with COVID-19 outcomes using the MGH replication cohort at FDR < 0.05 and in the same direction as in discovery. Finally, from the significant proteins identified in replication, we identified those significantly associated with COVID-19 status, they were limited in sample size (<100 participants) or breadth of proteins measured (<200 proteins), reducing power for detecting potentially significant contributing factors to disease outcome and severity.

To understand the pathogenic events linked to COVID-19 infection and identify potential biomarkers and drug targets, we obtained high throughput proteomic signatures of COVID-19 susceptibility and severity by studying more than 4,300 plasma protein levels obtained after hospital admission but before ventilation or death. Proteins associated with infection and severity in the discovery cohort were replicated within an independent cohort. The replicated proteins were then used to create a prediction model for infection (all cases versus healthy controls), ventilation (cases requiring ventilation versus cases without ventilation support), and death (cases that died versus cases that survived) as well as to identify what biological pathways are dysregulated because of infection.

School of Medicine, St Louis, MO, USA
11These authors contributed equally
12These authors contributed equally
13Lead contact
14Correspondence: cruchagac@wustl.edu
https://doi.org/10.1016/j.isci.2023.106408

iScience 26, 106408, April 21, 2023
proteins that passed the more stringent study-wide Bonferroni corrected threshold ($p < 0.05/4,301 = 1.16 \times 10^{-5}$, Figure 1) in the meta-analysis.

Using these significant proteins assayed at patient’s admission to the hospital, we created prediction models not only for infection, but more importantly for ventilation and death, which occurred after blood draw. Prediction models were created using the proteins associated with all three outcomes, as well as ventilation and death-specific models based on the proteins unique for each outcome. As the number of proteins in each model was relatively high, we used least absolute shrinkage and selection operator (LASSO) and 5-fold cross validation to identify the minimum number of informative proteins for each model.

In addition, we leveraged the same subset of proteins and performed pathway enrichment analyses to identify the biological pathways that were disrupted because of COVID-19 infection and need for...
ventilation or death. To identify proteins that contribute to the pathological events of COVID-19 infection, we performed Mendelian randomization (MR) (STAR Methods). We also used causal co-expression network analysis for those proteins associated with COVID-19 approaches to identify proteins that are likely to be part of the pathological processes implicated in the infection, and determined if those are also druggable.

Proteomic signatures of COVID-19 infection and severity
To identify plasma proteomic signatures for COVID-19 infection (compared to healthy controls), we performed association analyses including age, sex, plate and race as covariates (no significant differences were found using surrogate variables instead of plate as shown in Figure S3). We found 3,236 proteins (1,558 upregulated and 1,678 downregulated) significant after FDR correction in the discovery cohort (Tables S2 and S3; Figure 2). In the replication cohort, 906 proteins were also associated in the same direction and passed FDR correction (Table S2). In the meta-analysis, 841 proteins (363 upregulated and 478 downregulated; Tables S2 and S3) were significant after study-wide Bonferroni correction (for the 4,301 proteins; Tables S2 and S3; Figure 2).

Among the 332 COVID-19 cases, 82 individuals needed ventilation in 6.8 ± 7.7 days after hospital admission (Figure S1). Using the same approach, we identified 2,341 proteins (1,320 upregulated, 1,021 downregulated) associated with ventilation in the discovery cohort (Tables S2 and S4; Figure 2). In the replication cohort, 906 proteins were also associated in the same direction and passed FDR correction (Table S2). In the meta-analysis, 833 proteins (363 upregulated and 478 downregulated; Tables S2 and S3) were significant after study-wide Bonferroni correction (for the 4,301 proteins; Tables S2 and S3; Figure 2). Among the 332 COVID-19 cases, 63 individuals died in 36.2 ± 41.2 days after hospital admission. We identified 2,101 proteins (1,215 upregulated and 886 downregulated) associated with death in discovery (Tables S2 and S5). Out of 2,101 proteins, 993 proteins replicated in the same direction in the replication cohort. In the meta-analysis, 833 proteins (583 upregulated and 250 downregulated) were significant at Bonferroni corrected threshold (Tables S2 and S4; Figure 2).

Finally, we found 64 proteins that were consistently associated with all three COVID-19 outcomes (infection, ventilation and death; Figure 1, Table S2). There were 59 proteins consistently upregulated including CSF-1, IL-1Ra, LST1, PKC-Z, TGFβ1, BMP10, ANGL4, and vimentin. These proteins were upregulated in COVID-19 patients (compared to healthy controls), upregulated in ventilated patients (compared to those who were not) and also upregulated in individuals who died (compared to those who survived; Figure 1). The remaining 5 proteins, MZF1, Apo A-I, MDC, MLRA and NAGPA, were consistently downregulated in COVID-19 infection status as well as disease severity.

Table 1. Demographic and clinical characteristics of the WUSTL discovery cohort

<table>
<thead>
<tr>
<th>Race</th>
<th>COVID-19 case</th>
<th>COVID-19 control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALL</td>
<td>EA</td>
</tr>
<tr>
<td>N</td>
<td>332</td>
<td>102</td>
</tr>
<tr>
<td>Male (%)</td>
<td>195 (58.7)</td>
<td>72 (70.6)</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>59.7 ± 16.5</td>
<td>61.5 ± 15.6</td>
</tr>
<tr>
<td>Asymptomatic (%)</td>
<td>33 (9.9)</td>
<td>15 (14.7)</td>
</tr>
<tr>
<td>Hospitalization (%)</td>
<td>311 (93.7)</td>
<td>98 (96.1)</td>
</tr>
<tr>
<td>ICU (%)</td>
<td>155 (46.7)</td>
<td>54 (52.9)</td>
</tr>
<tr>
<td>Ventilation (%)</td>
<td>82 (24.7)</td>
<td>27 (26.5)</td>
</tr>
<tr>
<td>Death (%)</td>
<td>63 (18.98)</td>
<td>21 (20.6)</td>
</tr>
</tbody>
</table>

ALL indicates all samples including European and African ancestry. There are 5 cases and 4 controls who had other ancestries (Asian or admixed). EA represents European ancestry. AA represents African ancestry.

The basic demographic information includes sample size (N), number and percentage of males, mean and standard deviation (SD) of age stratified by COVID-19 status and ancestry.

Clinical characteristics includes the number (percentage) of asymptomatic cases, hospitalized cases, cases admitted to the intensive care unit (ICU), cases received mechanical ventilation and died cases.

N/A denotes not applicable.

*There were 44 individuals who were both on ventilation and died; they were included in both “Ventilation (%)” and “Death (%).”
Interactive web portal to visualize our results

To enable other interested researchers to navigate our findings for other COVID-19 related research, we created an interactive web portal (https://covid.proteomics.wustl.edu/). The results for each of three COVID-19 outcomes are presented in a tabular format, from which any user can select a particular protein of interest to examine the boxplot. The volcano plots for each of three outcomes (Figure 2) can be interactively accessed via this portal.

Prediction models

Next, we leveraged the robustly identified proteins through three-stage analysis (that passed FDR correction in discovery and replication and passed the Bonferroni correction in the meta-analysis) to create prediction models. Individual proteins did not show high and consistent area under the curve (AUC; Figure S4), so we created models with multiple proteins.

General model for all three outcomes

First, we used the 64 proteins that were associated with all three outcomes to create prediction models for each of infection, ventilation and death. The AUCs for this model, that included age, were >0.95 for all outcomes and in both discovery and replication (Figure 3), with positive predictive value (PPV) > 96% and negative predictive value (NPV) > 88%. The models without age also showed similar AUC, PPV and NPV (Figure S5 and Table S6). All models were significantly better than age alone (p < 1.01 × 10^{-4}, Table S7).
The proteins included in this model (n = 64) may be too many for a clinically translatable prediction model and some of them were highly correlated (heat maps in Figure S6 and Table S8), which may lead to overfitting. For this reason, we obtained a minimum number of informative proteins using LASSO regression and 5-fold cross validation. These analyses were performed on the discovery cohort. We obtained one model containing 22 proteins (including CSF-1, ANGL4, IL-6; Table S9) and another model with 10 proteins (including NXPH1, IL-1R4, and LST1; Table S9). For predicting ventilation, the model with 22 proteins showed similar performance than the default model with 64 proteins (Table S6; Figure 3B), but better performance than the model with 10 proteins. For predicting death, all three models were similar.

Here, we identified sets of 64, 22 and 10 proteins that can be used to predict patients at risk of COVID-19 infection, ventilation, and death with high accuracy (Table S9). Although these proteins can predict patients at high risk of death and/or in need of ventilation, they cannot distinguish patients who need ventilation but survive or patients at risk of death without ventilation. For this reason, we wanted to create outcome-specific prediction models (for ventilation only and for death only).

**Ventilation-specific model**

To build a ventilation-specific prediction model, we identified proteins that were significant for ventilation-only, by excluding those significant for death: i.e., (1) significant in both discovery and replication (FDR<0.05) for ventilation; (2) significant in the meta-analysis at Bonferroni corrected threshold (p < 1.16 × 10−5) and concordant direction of effects between discovery and replication for ventilation; and (3) not significant for death in the meta-analysis (p > 0.05). There were 50 proteins (Table S10; Figure S7) that fit these criteria.

As the number of individuals in each group (patients on ventilation versus patients not on ventilation) is highly unbalanced, NPV and PPV may be biased. For this reason, NPV and PPV was computed using sub-sampling with a balanced numbers of individuals in each group and 100 bootstrapped samples (STAR Methods; Table S11). The AUCs for the model that included the 50 proteins were 1 for discovery and 0.92 for replication, with NPV >0.88 and PPV >0.93 (Figures 3B and 3D). This model was also significantly better than age alone (p = 1.47 × 10−3).

Similar to the general model, we identified a subset of 32 proteins that provided a similar performance as the model with the 50 proteins (Table S12), by using a LASSO. The model with the 32 proteins had AUC higher than 0.91 and PPV and NPV higher than 0.80 for ventilation (Figure 3D) while showing low prediction power for death (AUC <0.55). Results were similar without age in the model for the above protein sets (Figure S8). These results indicate that these 50 (and the subset of 32) proteins measured at hospital admission can predict patients in need of ventilation significantly better than age alone, with high sensitivity and specificity. Despite the sample overlap between ventilation and death outcomes, these models showed relative low prediction power for death (AUC of 0.55–0.63; Figure 3D).

**Death-specific model**

To create death-specific models we used a similar approach to that of ventilation. We identified proteins that were significant for death only by excluding those significant for ventilation: significant in discovery (FDR <0.05), replication (FDR <0.05), and the meta-analysis (Bonferroni corrected p < 1.16 × 10−5) with concordant direction of effects, for death, and not significant for ventilation in the meta-analysis (p > 0.05). This led to the identification of 5 proteins: ASB9, Endocan, jun-D, EGFLA, and glutathione S-transferase pi (Table S10; Figure S7).

The AUCs for death using these five proteins was 0.95 in both discovery and replication, significantly higher than AUC with age (p = 2.29 × 10−7 in discovery and 1.28 × 10−2 in replication) with high PPV (0.83 in discovery and 0.80 in replication) and high NPV (0.86 in discovery and 0.80 in replication). The AUCs of these five proteins was relatively low for ventilation (AUC <0.56, with ~0.6 PPV and ~0.5 NPV).

In summary, we created three prediction models: (1) A general model with 64 proteins (or subsets of 22 or 10 proteins) that predicts infection and both ventilation and death (AUC = 0.95–1 for ventilation and 1.00 for death); (2) a ventilation-specific model with 50 proteins (or a subset of 32) that mainly predicts ventilation (AUC = 0.9–1 for ventilation; 0.55–0.63 for death); (3) a death-specific model with 5 proteins that mainly predicts death (AUC = 0.95 for death; 0.52–0.56 for ventilation). By considering all three models, it is possible
to predict the most likely outcome for a specific individual (Figure 3A). If the general model predicts ‘no’, then the individual will be of no concern. When the general model predicts ‘yes’, then there will be 4 possible cases: (1) On ventilation and die; (2) on ventilation but survive; (3) not on ventilation but die; and (4) infected by COVID-19 without any worse outcome (Figure 3A).

Proteins associated with COVID-19 infection, ventilation and death are enriched in immune-related, Alzheimer’s disease and cardiovascular disease pathways

To identify the common pathways and biological processes involved in COVID-19 infection and severity, we performed enrichment analysis for the significant proteins in each outcome (841 for COVID-19 status, 833...
for ventilation and 253 for death) using clusterProfiler 4.0 R package. We found an enrichment for “genes downregulated by SARS-CoV-2 infection” a COVID-specific pathway that was created based on cell models after COVID-19 infection (GEO accession GSE147507). This pathway was enriched for the proteins associated with COVID-19 infection (FDR = 3.4 × 10^{-3}). There were many other pathways enriched for immunologic signatures including adaptive (FDR = 2.2 × 10^{-3}), innate (FDR = 1.5 × 10^{-2}), and cytokine signaling in immune system (FDR = 7.5 × 10^{-3}; Figure 4A, Table S13). However, we also identified an enrichment for pathways that are not related to the immune response to a virus infection, including Alzheimer’s disease (AD; FDR = 1.0 × 10^{-3}) and coronary artery disease (FDR = 1.1 × 10^{-2}), suggesting that COVID-19 infection may impact pathways beyond what would be expected.

**Cytokine and immune related pathways**

Among key immunopathological consequences of severe COVID-19, a cytokine storm, characterized by elevated levels of pro-inflammatory mediators, is shown to be a central feature. The proteins identified and replicated for COVID-19 infection were significantly enriched in cytokine signaling in immune system (GO:0034097; FDR = 7.5 × 10^{-2}), including IL-1Ra, IL-21 sR, and CSF-1. The innate (GO:0045824; FDR = 1.5 × 10^{-2}) and adaptive immune system (GO:0006954; FDR = 2.21 × 10^{-2}) were also significant for the COVID-19 infection analyses. Our tissue-specific enrichment analysis also supports that COVID-19 affects lung-specific proteins (p = 3.21 × 10^{-3} for COVID-19 infection; Figure 4D and Table S14).

**Alzheimer’s disease (AD) related pathways.** Central nervous system (CNS) damage can be present early in the course of COVID-19. Several proteins associated with COVID-19 outcomes in this study are part of the AD pathway (disease gene network C0276496; FDR = 1.0 × 10^{-2}) were also significant for the COVID-19 infection analyses. Our tissue-specific enrichment analysis also supports that COVID-19 affects lung-specific proteins (p = 3.21 × 10^{-3} for COVID-19 infection; Figure 4D and Table S14).

**Coronary artery disease (CAD) related pathways.** SARS-CoV-2 infection can increase the risk of cardiovascular disease. Our analyses indicate that the proteins associated with COVID-19 outcomes are enriched for coronary artery disease (CAD) related pathways (FDR = 1.0 × 10^{-2} for COVID-19 infection; Figure 4A and Table S13). Proteins that are part of this pathway include troponin T, ANGL4, FURIN, AT1B1, and TGFBI. We found significantly higher troponin T levels in COVID-19 patients (p = 6.46 × 10^{-9}; Figure 4C) and those requiring ventilation (p = 7.24 × 10^{-13}; Figure 4C). Similarly, we found significantly higher ANGL4 levels in COVID-19 infection, ventilation and death (p = 3.98 × 10^{-6}, 8.91 × 10^{-27}, 1.1 × 10^{-11}, respectively, Figure 4C) and significantly higher FURIN levels in COVID-19 infection (p = 4.68 × 10^{-23}) and ventilation (p = 6.31 × 10^{-15}). Significantly lower ATP1B1 levels were found in COVID-19 patients (p = 3.55 × 10^{-14}, Figure 4C).

**Identification of causal and potential therapeutic targets**

Pathway analyses point to specific biological processes that are altered because of COVID-19 infection and that are associated with specific outcomes, such as ventilation and death, but those analyses cannot differentiate the causal proteins from those that are because of secondary changes. To identify potential causal pathways, we performed Mendelian randomization (MR) using plasma protein quantitative trait loci (pQTL) as instrumental variables for those proteins that were associated with infection, ventilation and death, as well as performed Bayesian network analyses to identify key drivers and potential targets for known therapies.
Mendelian randomization (MR) analyses and colocalization analyses

A recent GWAS including 49,562 COVID-19 patients by host genetics initiative (HGI) identified eight loci associated with COVID-19 infection. In addition, there was a recent plasma pQTL study based on 4,907 proteins in 35,559 Icelanders. We leveraged these datasets to determine if any of the proteins identified in our study is a causal plasma biomarker for COVID-19 infection. Of the 1,449 proteins associated with any of the three outcomes, 1,437 were present in the plasma pQTL atlas. Among these, 616 proteins had either cis or trans pQTL near COVID-19 loci (within 1 Mb), and 15 of these 616 proteins had cis pQTL near the COVID-19 loci. We performed two-sample MR analyses by integrating the pQTL atlas and COVID-19 GWAS results in European ancestry (Figure 5A). Only proteins that passed FDR correction, were nominally significant in the MR-Egger (run only for trans pQTL to correct for potential pleiotropy) and had an MR estimate in the same direction as in our differential abundance analysis were considered significant.

Among the 616 proteins with either cis plus trans pQTL near COVID-19 loci (instrument SNPs in Table S15), six proteins (GOLM1, IR, PLA2R, BCAT2, ACE, and EPHA4) were identified as causal (Figure 5B and Figure 5C). We performed tissue-specific enrichment analyses via FUMA for 30 tissues in GTEx v8. We found significant enrichment for lung (FDR = 3.21 × 10^{-5}). The detailed results are in Table S14.

Figure 4. Enrichment analyses for proteins associated with COVID-19 infection, ventilation and death

(A) We performed pathway enrichment analyses via clusterProfiler using Entrez gene symbol of the robustly identified proteins (841 proteins for infection, 833 proteins for ventilation, 253 proteins for death). These proteins were enriched for 16 pathways. Two differentially abundant proteins (CSF1 and GFRA3) for 3 outcomes were enriched in COVID-19 related pathways. Proteins were significantly enriched for Alzheimer’s disease (AD) pathway (FDR = 1.0 × 10^{-2} for COVID-19 infection). Enrichment of COVID-19 infection (FDR = 1.0 × 10^{-2}) was observed for coronary artery disease. The detailed results are in Table S13.

(B) Forest plots of AD-related proteins: APP, NFL, EPHAS, and TMEM106B. MAPT and GFAP showed a nominal significance (p < 0.05).

(C) Forest plots of cardiovascular-related proteins: Troponin T (cardiac muscle), ANGL4, FURIN, AT1B1 and TGFB1.

(D) We performed tissue-specific enrichment analyses via FUMA for 30 tissues in GTEx v8. We found significant enrichment for lung (FDR = 3.21 × 10^{-5}). The detailed results are in Table S14.
Among the 15 proteins with cis-only pQTL near COVID-19 loci (instrument SNPs in Table S17), three proteins (BCAT2, Fucosyltransferase 3, and NUCB1) were identified as causal (Figure 5 Band Table S18). Combining both of these MR analyses, a total of eight unique proteins were nominated causal for COVID-19 infection. Three proteins (BCAT2, GOLM1, and NUCB1) had positive effect estimates, indicating that higher levels of these proteins increase risk for COVID-19 infection. The remaining five proteins (IR, PLA2R, ACE, EPHA4, and Fucosyltransferase 3) had negative effect estimates, indicating lower protein levels increase risk for COVID-19 infection.

BCAT2 was identified as causal for COVID-19 outcome in both MR analyses. Two SNPs were cis pQTLs for BCAT2 (rs4802507, cis pQTL = 2.57 x 10^{-09}; rs777777024, p = 2.26 x 10^{-08}), associated with COVID-19 infection (HGI GWAS for rs4802507, p = 1.37 x 10^{-06}; rs777777024, p = 6.94 x 10^{-05}), and showed strong evidence for colocalization at the BCAT2 locus (posterior probability PP.H4 = 0.99; Figure 5C and Table S19). For GOLM1 (FDR for differential protein levels for COVID-19 infection = 3.08 x 10^{-08} and MR-Egger p = 3.33 x 10^{-06}), one SNP (rs633862) also showed strong evidence for colocalization at the ABO locus with COVID-19 infection (PP.H4 = 0.99; Table S19).

**Table S16.** Among the 15 proteins with cis-only pQTL near COVID-19 loci (instrument SNPs in Table S17), three proteins (BCAT2, Fucosyltransferase 3, and NUCB1) were identified as causal (Figure 5B and Table S18). Combining both of these MR analyses, a total of eight unique proteins were nominated causal for COVID-19 infection. These proteins (BCAT2, GOLM1, and NUCB1) had positive effect estimates, indicating that higher levels of these proteins increase risk for COVID-19 infection. The remaining five proteins (IR, PLA2R, ACE, EPHA4, and Fucosyltransferase 3) had negative effect estimates, indicating lower protein levels increase risk for COVID-19 infection.

**Identification of causal and potential therapeutic targets**

Gene co-expression network analysis is commonly used to identify and prioritize potentially functional genes and clusters of genes contributing to disease. However, many of these techniques are limited by...
prior assumptions on the number of clusters or set thresholds for determining co-expression. We used MEGENA, a co-expression analysis package that avoids this limitation. MEGENA uses a Bayesian approach to create causal models to dissect genetic pathways and provide mechanisms implicated in disease. Using MEGENA, we built co-expression networks based on the robustly identified proteins associated with COVID-19 outcomes, and distinguished clusters of correlated proteins. Based on these clusters, MEGENA identified highly connected hub proteins that are potentially novel key drivers of COVID-19 infection and severity. These hubs were chosen as the focus of our analyses.

We used as input all differentially abundant proteins across three outcomes (COVID-19 infection, ventilation, and death), totaling 1,449 unique proteins. We ran MEGENA using the hub proteins identified in the combined runs to identify clustering of the hubs. For the 1,449 proteins that were significant from any of the three outcomes, we identified 141 key (hub) proteins through MEGENA (Figure 6, Table S20). These proteins were used as input to MEGENA to identify clusters, of which eight were identified; 47 proteins were not assigned to a cluster (Figure 6). Of these 141, at least 20 have links to immune function according to GeneCards. At least 15 are kinase-related genes, and evidence supports the use of kinase inhibitors in COVID-19 treatment. An additional 21 are involved in the extracellular matrix or cellular adhesion, suggesting potential importance in the virus-cell interface. We matched the identified hub proteins to pathways enriched for differentially abundant proteins (Table S20). The Gene Ontology sets containing the most hub proteins are proteolysis (31 hub proteins), protein phosphorylation (27), positive regulation of signaling (27), Blalock Alzheimer disease up (25), and Naba matrisome (25). Dysregulation of protein phosphorylation has been identified in Vero E6, supporting a link between these hub proteins and COVID-19.

Among the 141 hub proteins, we found that 35 have associated drugs in the DrugBank database and 19 of them have a drug with previous links to antiviral activity (Figure 6, Table S20). Network plots of all 1,449 proteins (Figure S9), 841 proteins associated with COVID-19 infection (Figure S10), 833 proteins with ventilation (Figure S11), and 253 proteins associated with death (Figure S12) were created using MEGENA and Cytoscape. The AUCs for each of the identified hub proteins can be found at Table S21.

**DISCUSSION**

Deep plasma proteomic profiling of COVID-19 patients can be instrumental not only for identifying the biological processes implicated in response to COVID-19, but also for discovering models that can predict the progression to future severe outcomes such as ventilation and death. Here we generated and analyzed deep proteomic data from two large and well characterized independent cohorts to identify proteins dysregulated because of COVID-19 infection and associated with a later need of ventilation and death. We used a three-stage study design (discovery, replication and meta-analysis) with stringent multiple test correction to identify a robust set of associated proteins. Our analyses identified a total of 1,449 proteins that replicated and passed Bonferroni correction. There were 841 proteins associated with COVID-19 infection, 833 proteins with ventilation and 253 proteins for death. We created an interactive web portal (https://covid.proteomics.wustl.edu/) for users to navigate these plasma proteomic signatures of COVID-19 infection and outcomes.

The number of differentially abundant proteins is around 20% of those present in both cohorts, which may look high. We used very stringent criteria, only presenting the proteins that passed multiple test correction (FDR<0.05) in two independent datasets and passed the very stringent Bonferroni correction in the meta-analyses. One possible reason behind this high number of proteins could be explained because more than 90% of our COVID-19 patients in this study were hospitalized and suffered from a severe viral infection, resulting in big changes in the blood proteome (compared to healthy controls). The number of proteins for death was smaller as fewer subjects died.

Several methods have been used to try to identify individuals with bad outcomes after COVID-19 infection. They include the Charlson comorbidity index score (CCIS), the sequential organ failure assessment (SOFA) score or the acute physiology and chronic health evaluation (APACHEII). However, all these methods showed limited performance in predicting the severe COVID-19 outcome. As a promising strategy, a recent proteomic study identified plasma prognostic biomarkers including several markers of neutrophil activation (resistin, lipocalin-2, hepatocyte growth factor, interleukin-8, and granulocyte colony-stimulating factor), which precede the onset of critical illness with high discriminatory power for detecting critical illness. However, this study only tested 78 immunologic function related proteins to distinguish
In this study, we created three prediction models. The first model is for generally predicting patients with any of COVID-19 outcomes: infection, ventilation and death. This model can differentiate individuals that have an ongoing COVID-19 infection (AUC>0.97), and within the COVID-19 positive individuals those that will need ventilation (AUC > 0.95) or that will die (AUC = 1; Figure 3A). Through machine learning, we identified 10 proteins that showed similar predictive power as those 64 that could be implemented in a clinical test. As this model does not distinguish between ventilation and death, we created two additional prediction models: one for ventilation and another for death. The ventilation-specific model can accurately distinguish high-risk COVID-19 patients that will require mechanical ventilation. The death-specific prediction model demonstrated significantly high AUC (0.95; Figure 3D) to identify patients that will die. These models specific for ventilation or death showed some prediction performance between them, as expected, because there is overlap in the individuals that need ventilation and later die. By considering all three models, one can obtain the most likely outcome for an individual. These three sets of prediction models can be used as prognostic biomarkers to guide the allocation of medical resources to COVID-19 cases at high risk for severe illness. Because this study was limited by relatively small sample size for COVID-19 patients in severe outcomes, our prediction models require additional validation in independent cohorts.

The proteins associated with COVID-19 outcomes can be essential to identify the dysfunctional biological processes in pathways and to nominate potential causal proteins implicated in COVID-19. Our robustly identified proteins showed enrichment for genes down regulated by SARS-CoV-2 infection. This
represents a positive control, indicating that our methodology is able to capture relevant biological processes. We also identified many pathways related to the immune system (adaptive, FDR = 2.21 × 10^{-2}; innate, FDR = 1.50 × 10^{-2}; cytokine, FDR = 7.47 × 10^{-2}) that included IL-1Ra, IL-21sR, IFN-a/b R1, IFNα5, TGFβ1, and CSF-1. IL-1Ra modulates IL-1 related immune and inflammatory responses of acute infection. Anakinra, an FDA approved immunosuppressive drug for rheumatoid arthritis and deficiency of IL-1A, demonstrated improved overall survival and invasive ventilation-free survival for acute respiratory distress syndrome patients associated with COVID-19. TGFβ1 is shown to suppress IFN response in alveolar macrophages, and the TGF-β dominated immune response from COVID-19 was observed in single cell transcriptome levels from plasmablasts. Higher CSF-1 levels are reported in serum of COVID-19 patients and proposed to be a potential target for COVID-19. More importantly, we identified several other pathways that can help to understand some of the reported long-term effects of COVID-19 infection, including brain-fog/dementia and cardiovascular problems. Several studies suggest that COVID-19 can lead to Alzheimer’s-like dementia or other forms of neurocognitive impairment in persons free of cognitive impairment before infection. Our enrichment results for brain specific genes and AD related pathways (FDR = 1.0 × 10^{-2} for infection-associated proteins) support CNS involvement in COVID-19 infection. Our analyses identified proteins that are part of AD pathogenesis including APP, TMEM106B, EPHA5, CR1, CD33 or MAPT. APP is a well-known AD gene that is linked to autosomal dominant AD. Genetic variants in TMEM106B are associated with risk for Parkinson disease (PD), frontotemporal dementia (FTD), and neuronal survival. Recent studies also suggest that TMEM106B is a proviral host factor for SARS-CoV-2. TMEM106B is a component of the lysosome. The SARS-CoV-2 virus hijacks the lysosome to exit the cells via ORF3a and nsp6. Lower plasma TMEM106B protein levels reflecting disruption of normal lysosome function were found in COVID-19 patients (p = 1.51 × 10^{-14}). We also found that well-known biomarkers for AD and neurodegeneration, including NFL and GFAP, were also altered because of infection or associated with ventilation and death. This finding has important repercussions on the interpretation of the biomarkers moving forward. Individuals that have a recent COVID-19 infection may have altered AD-related biomarkers that may affect their diagnosis. Therefore, there is a need to perform further studies of AD-related biomarkers in the context of COVID-19 infection and potentially identify specific biomarker cut-offs for individuals with or without COVID-19 infection. Our results are in line with recent studies indicating a mechanistic overlap between AD and COVID-19. Our data further indicate that AD and general neurodegenerative biomarkers are altered in COVID-19 patients and could also be used to identify individuals at risk of having long-term impact on memory loss, brain-fog and other brain-related outcomes. In the absence of effective treatments, a mechanistic understanding of how patients with AD become particularly vulnerable targets of SARS-CoV2 and how COVID-19 leads to cognitive impairment and AD in cognitively healthy individuals is critical to guide evidence-based healthcare management and development of targeted intervention.

In addition to the CNS impairment, COVID-19 cases may also lead to an increased risk for coronary artery disease (CAD) by increasing myocardial injury from viral infection. This is supported by the significantly enriched CAD related pathways and heart specific genes from our dysregulated proteins including Troponin T, ANGL4, FURIN, and ATP1B1. Troponin T is a protein integral to cardiac muscle and a known plasma biomarker for myocardial infarction. Loss-of-function variants of ANGL4 are shown to significantly reduce risk of CAD. Genetic variants that increase the expression of FURIN has been associated with CAD, and ATP1B1 is a causal gene for CAD. Owing to the complexity of the brain and heart, the damage of these two organs may be permanent and can cause post-acute COVID-19 syndrome (PACS) such as tiredness, dyspnea, chest pain, brain fog, and headache. Thus, our identified dysregulated proteins belonging to the CAD related pathway might be used as prognostic tool to identify the cases at high risk for CAD and PACS in general, although additional follow-up studies are needed.

By performing Mendelian randomization (MR), we nominated three proteins (BCAT2, GOLM1, and NUCB1) as causal risk proteins, as well as five proteins (IR, PLA2R, ACE, EPHA4, and Fucosyltransferase 3) as causal protective proteins for COVID-19 infection. Except for GOLM1, and ACE known causal risk factor for COVID-19 infection, the remaining seven proteins are detected as causal biomarkers for the first time. The presence of COVID-19-related observations or an organ dysfunction for these proteins is in line with our findings. For instance, the Capture Hi-C Omnibus Gene Score indicates that BCAT2 is a key plasma biomarker for COVID-19. ACE is a critical enzyme converting Ang I into Ang II predominantly in the lungs. Dysregulated ACE/ACE2 balance underpins the rapid progression of COVID-19. FUT3 is the gene that
encodes Fucosyltransferase 3 (Lewis blood group) and its homozygous nonfunctional genotypes were reported to be protective against severe COVID-19 cases. Eph receptors might serve as the entry receptor for SARS-CoV-2 in the central nervous system, contributing to COVID-19-associated neurodegenerative disease development. IR is a member of the insulin/IGF signaling pathway that crosstalks with SARS-CoV-2. PLA2R is the major causal autoantigen for membranous nephropathy. NUCB1, also called CALNUC, is a Golgi calcium binding protein that binds to beta-amyloid precursor protein, implicating it in pathogenesis of AD. Agents targeting these eight proteins including BCAT2 and GOLM1 may be able to reduce the risk for COVID-19 infection as well as COVID-19-related cognitive and metabolic dysfunction. Gaziano et al. identified six genes (CCR1, CCR5, IL10RB, IFNAR2, PDE4A and ACE2) as causal repurposing candidates for COVID-19 using MR analyses with cis eQTLs of 49 tissues in GTEx. They were not replicated in our analyses. This failed replication might be because of eQTLs not translating to pQTLs used in our MR analyses. It is known that mRNA levels do not necessarily correlate well with protein levels. Our previous study also demonstrated a low level of overlap between expression and protein QTLs. The limitation of our MR analyses is that the causal inference for COVID-19 infection cannot be generalized to other ethnic groups because of lacking pQTL resources for non-European ancestries. Further investigation of the roles of these proteins in other ethnic groups is needed.

To identify key-drivers in the proteins network that could be also part of the causal pathway and potentially druggable, we performed an additional Bayesian approach. MR is restricted to those proteins in a strong pQTL and might miss many potential therapeutic targets. Gene network analyses identified 141 key proteins, of which 35 have known drug targets, some of them already being tested in several clinical trials. One drug, Fostamatinib, was found to target 5 hub proteins (ABL1, RIPK2, PIM1, PLK-1, and PAK7). Fostamatinib is a tyrosine kinase inhibitor usually used to treat chronic immune thrombocytopenia. Studies support it for treatment of acute respiratory distress syndrome in COVID-19 and a stage 2 clinical trial found it safe and showed evidence of better outcomes for COVID-19 patients compared to placebo, suggesting its standard use for care. In addition, Prio (encoded by PRNP) and MK12 (encoded by MAPK12) are both targeted by tetracyclines. These antibiotics have been used to treat numerous types of infection, including pneumonia, cholera, and syphilis. Tetracyclines are also reported to protect against HIV. They may be able to act through multiple pathways, including inhibition of RNA replication or downregulation of the NF-κB pathway, to decrease severity of COVID-19. Additional studies will be needed to confirm that the proposed targets and FDA compounds have an impact on COVID-19 outcomes.

Obesity and diabetes can lead to dysregulated immune response, resulting in more severe and prolonged lung pathology. In particular, obesity or higher body mass index (BMI) is recently shown to increase severe COVID-19 outcome by 1.8-fold. Because obesity and BMI information was not collected in our study, we were not able to consider them as covariates. It is possible that our COVID-19 cases have higher BMI than controls, and some differential abundant proteins we identified might be related to BMI. Nilsson et al. reported 25 proteins positively correlated and four proteins negatively correlated with BMI. Among these 25 positively correlated proteins, we found four proteins (IL18, TGFβ1, CSF1, and IL18R1) with significantly higher level in COVID-19 cases, five proteins (CSF1, IL6, TGFβ1, IL10RB, and IL18) with higher levels in patients that needed ventilation, and five proteins (CSF1, HGF, TGFβ1, IL10RB, and IL6) with higher levels in those that died in our study. Among the four negatively correlated proteins, CCL11 was significantly downregulated in both COVID-19 cases and ventilated patients. As these inflammatory protein biomarkers significantly correlate with BMI, we cannot rule out the possibility that these proteins we identified might be related to BMI and unrelated to COVID-19 outcomes.

In summary, we performed deep proteomics profiling of COVID-19 patients, identifying dysregulated proteins associated with severe outcomes. We created an interactive web portal (https://covid.proteomics.wustl.edu/) for use by the scientific community. We created prediction models for ventilation and death, presented AD and CAD related changes with COVID-19 infection, and identified seven novel potential druggable targets. Our findings show the promise of proteomic studies contributing to the understanding of COVID-19 biology and pathophysiology.

Limitations of the study
Despite the strength and the significant findings reported, our study has several limitations. First, because more than 90% of our samples are hospitalized, the accuracy of our prediction models will need to be
further evaluated in other independent studies. Second, owing to the lack of SARS-CoV-2 variant information such as the Delta and Omicron variants and other important factors associated with infection and outcomes (BMI, smoking, diabetes), future studies analyzing the protein dysregulation and the biological pathways impacted by these different SARS-CoV-2 variants and the influence of other biological factors will be needed. Third, we used the SomaScan platform in this study. We and others have validated the performance of SomaScan technology, showing that the SomaScan assay is as accurate as classic enzyme linked immunosorbent assay (ELISA), mass spectrometry and Olink proximity extension assay. To translate these findings to clinical tests, these measures will need to be validated using the clinical laboratory improvement amendments (CLIA) or U.S. food and drug administration (FDA) approved approaches.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Sample collection
  - Plasma proteomic data
  - Statistical analyses
  - Replication strategy
  - Meta-analyses
  - Prediction model
  - Pathway enrichment analyses and tissue-specificity enrichment analyses
  - Mendelian randomization (MR) analyses
  - Colocalization analyses
  - Hub and druggable proteins

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106408.

ACKNOWLEDGMENTS
We thank all the participants and their families, as well as the many involved institutions and their staff. We acknowledge SomaLogic Operating Co., Inc. as the provider of the proteomic data measured using the modified aptamer-based SomaScan Assay. SomaScan, SOMAmer and SomaSignal are trademarks of SomaLogic Operating Co., Inc.

This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics and Informatics Center (NGI: https://neurogenomics.wustl.edu/) and the Departments of Neurology and Psychiatry at Washington University School of Medicine.

This work was supported by grants from the National Institutes of Health (RF1AG074007 (YJS), R01AG044546 (CC), P01AG003991(CC, JCM), RF1AG053303 (CC, SC), RF1AG058501 (CC), and U01AG058922 (CC), Alzheimer’s Association Research Grant 925002 (SC), the Chan Zuckerberg Initiative (CZI), and the Alzheimer’s Association Zenith Fellows Award (ZEN-22-848604, awarded to CC).

The recruitment and clinical characterization of research participants at Washington University were supported by NIH P30AG066444 (JCM), P01AG03991 (JCM), and P01AG026276 (JCM).

This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics and Informatics Center (NGI: https://neurogenomics.wustl.edu/) and the Departments of Neurology and Psychiatry at Washington University School of Medicine.
This study utilized samples obtained from the Washington University School of Medicine’s COVID-19 biorepository, which is supported by: the Barnes-Jewish Hospital Foundation; the Siteman Cancer Center grant P30 CA091842 from the National Cancer Institute of the National Institutes of Health; and the Washington University Institute of Clinical and Translational Sciences grant UL1TR002345 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the view of the NIH.

**AUTHOR CONTRIBUTIONS**


S.C., D.J., M.G., A.R.T., S.C. and B.Z. were involved on study design and interpretation. Supervision: C.C. and Y.J.S.; Writing – original draft: L.W., D.W., Y.J.S., and C.C.; Writing – review and editing: All authors reviewed, edited and approved the manuscript.

**DECLARATION OF INTERESTS**

C.C. has received research support from: Biogen, EISAI, Alector, GSK, an anonymous foundation and Paragon. The funders of the study had no role in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. C.C. is a member of the advisory board of Vivid Genomics, and Circular Genomics and own stocks.

**INCLUSION AND DIVERSITY**

We support inclusive, diverse, and equitable conduct of research.

Received: September 30, 2022
Revised: January 21, 2023
Accepted: March 10, 2023
Published: March 14, 2023

**REFERENCES**


## STAR★METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deposited data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SomaScan v4.1 7K proteomic data</td>
<td>This study</td>
<td>NIAGADS dataset#: ng00128 <a href="https://dss.niagads.org/datasets/ng00128/">https://dss.niagads.org/datasets/ng00128/</a></td>
</tr>
<tr>
<td>SomaScan 5K proteomic data</td>
<td>Massachusetts General Hospital</td>
<td>Mendeley Data: <a href="https://data.mendeley.com/datasets/nf853r8xsj/1">https://data.mendeley.com/datasets/nf853r8xsj/1</a></td>
</tr>
<tr>
<td>round 5 COVID-19 GWAS results</td>
<td>COVID-19 host genetics initiative</td>
<td><a href="https://www.covid19hg.org/">https://www.covid19hg.org/</a></td>
</tr>
<tr>
<td>GWAS summary statistics for 4907 aptamers</td>
<td>the Icelandic Cancer Project and various genetic programs at deCODE genetics, Reykjavík, Iceland</td>
<td><a href="https://www.decode.com/summarydata/">https://www.decode.com/summarydata/</a></td>
</tr>
</tbody>
</table>

### Software and algorithms

<table>
<thead>
<tr>
<th>Software and algorithms</th>
<th>Source</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 4.2.2</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/">https://cran.r-project.org/</a></td>
</tr>
<tr>
<td>R 3.5.2</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/">https://cran.r-project.org/</a></td>
</tr>
<tr>
<td>VennDiagram</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/web/packages/VennDiagram/">https://cran.r-project.org/web/packages/VennDiagram/</a></td>
</tr>
<tr>
<td>ggplot2</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/web/packages/ggplot2/index.html">https://cran.r-project.org/web/packages/ggplot2/index.html</a></td>
</tr>
<tr>
<td>SomaDataIO 1.8.0</td>
<td>Open-source</td>
<td><a href="https://github.com/SomaLogic/SomaDataIO">https://github.com/SomaLogic/SomaDataIO</a></td>
</tr>
<tr>
<td>caret 6.0-93</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/web/packages/caret/index.html">https://cran.r-project.org/web/packages/caret/index.html</a></td>
</tr>
<tr>
<td>pROC 1.18.0</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/web/packages/pROC/">https://cran.r-project.org/web/packages/pROC/</a></td>
</tr>
<tr>
<td>clusterProfiler 4.0</td>
<td>Open-source</td>
<td><a href="https://github.com/YuLab-SMU/clusterProfiler">https://github.com/YuLab-SMU/clusterProfiler</a></td>
</tr>
<tr>
<td>FUMA GWAS</td>
<td>Open-source</td>
<td><a href="https://fuma.ctglab.nl/">https://fuma.ctglab.nl/</a></td>
</tr>
<tr>
<td>TwoSampleMR 0.5.6</td>
<td>Open-source</td>
<td><a href="https://github.com/MRCIEU/TwoSampleMR">https://github.com/MRCIEU/TwoSampleMR</a></td>
</tr>
<tr>
<td>coloc 5.1.1</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/web/packages/coloc/index.html">https://cran.r-project.org/web/packages/coloc/index.html</a></td>
</tr>
<tr>
<td>MEGENA</td>
<td>Open-source</td>
<td><a href="https://cran.r-project.org/web/packages/MEGENA/index.html">https://cran.r-project.org/web/packages/MEGENA/index.html</a></td>
</tr>
<tr>
<td>Cytoscape</td>
<td>Open-source</td>
<td><a href="https://cytoscape.org/">https://cytoscape.org/</a></td>
</tr>
<tr>
<td>UniProt</td>
<td>Open-source</td>
<td><a href="https://www.uniprot.org/">https://www.uniprot.org/</a></td>
</tr>
<tr>
<td>DrugBank</td>
<td>Open-source</td>
<td><a href="https://go.drugbank.com/">https://go.drugbank.com/</a></td>
</tr>
</tbody>
</table>

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Carlos Cruchaga (cruchagac@wustl.edu).
Materials availability
This study did not generate new materials.

Data and code availability
Proteomic data from the Knight ADRC participants are available at the NIAGADS and can be accessed at https://www.niagads.org/Knight ADRC-collection (NIAGADS dataset#: ng00128);

The summary results using these data are also available to the scientific community through a public web browser: https://covid.proteomics.wustl.edu/. All our analyses used open-source software that have been included in the key resources table.

METHOD DETAILS
Sample collection
We recruited 332 COVID-19 patients with positive SARS-CoV-2 test from March 26, 2020 to September 30, 2020 at Barnes Jewish Hospital, St. Louis Children’s Hospital or affiliated Barnes Jewish Hospital testing sites in St. Louis. We obtained the ventilation status and death status from the hospital records. Using plasma samples collected prior to November 2019, we selected 150 controls by matching on sex, race and age. We obtained the informed consent for enrollment in the study from 482 participants. The study was approved by institution review board at Washington University School of Medicine in St. Louis.

Plasma proteomic data
Peripheral blood was collected, and plasma was isolated by centrifuge and stored at –80°C. The proteomic data in plasma was measured using SomaScan v4.1 7K, a multiplexed, single-stranded DNA aptamer-based platform80 from SomaLogic (Boulder, CO). Three different dilution sets (1:5, 1:200, 1:20,000) were used for plasma samples. Instead of physical units, the readout in relative fluorescent units (RFU) was used to report the protein concentration targeted by 7,584 modified aptamers. To mitigate technical variation introduced by the readout, pipetting errors, inherent sample variation, initial data standardization at the sample and aptamer level including hybridization control normalization, median signal normalization and inter-plate calibration was performed by SomaLogic with control aptamers (positive and negative controls) and calibrator samples.81 Hybridization control normalization on each plate was used to correct for systematic variability in hybridization. Median signal normalization was performed within the same dilution group to remove the bias due to pipetting variation, variation in reagent concentrations, and assay timing among others. Finally inter-plate Calibration based on calibrator samples was performed separately on each aptamer to remove plate bias.

We further processed the proteomics data using SomaDataIO (v1.8.0)80 and Biobase (v2.42.0)82 and removed the aptamer outliers and sample outliers by using the following approaches68 (Figure S2): 1) Aptamer outliers had greater than 15% samples with concentration below 2 standard deviation above the average RFU level of the dilution buffer. 2) Aptamer outliers had greater than 0.5 of maximum difference between aptamer calibration factor and median of plate calibration factor. 3) Aptamer outliers had median coefficient of variation (CV) more than 0.15. 4) Aptamer outliers had greater than 15% of samples whose log10 transformed RFU level fell outside of either end of 1.5-fold of interquartile range (IQR). 5) Sample outliers had more than 15% aptamers falling outside of either end of 1.5-fold of IQR. 6) Aptamer outliers shared by ~80% of sample outliers.

All 7,584 aptamers were annotated with Universal Protein Resource (UniProt)83 identifiers and Entrez Gene symbols. After QC, 7,055 aptamers and 482 individuals remained and were used for differential abundance analyses.

Statistical analyses
To identify the plasma biomarkers associated with COVID-19 infection and severity in the discovery cohort, protein levels were first log10 transformed to approximate the normal distribution. Regression analysis was then performed using lm function in R (R version 3.5.2 (2018-12-20)) with:

\[
\log_{10} (\text{protein abundance level}) = \text{outcome} + \text{age} + \text{sex} + \text{ethnicity (self-reported)} + \text{plate}
\]
where an outcome corresponds to each of three COVID-19 outcomes (infection, ventilation and death) and plate denotes eight different plates used by SomaScan. Volcano plots were created with the ggplot2 package and Venn diagrams were created with R VennDiagram package.

The proteomic data for 482 individuals were assayed in 8 different plates. Even though the intraplate and inter-plate variation of aptamer RFU levels has been normalized and the plate was used as covariate in our analyses to control for batch effects, there may be unobserved hidden factors which correlate with the outcomes and biased our estimate. Thus, we estimated these hidden factors from the matrix of 7,055 proteins for 482 individuals using surrogate variable analysis (SVA) implemented in R sva package. Since sva estimation requires no missing data, we first imputed the missing data points by random sampling of non-missing data points for each protein separately by COVID-19 cases and controls. Then number of surrogate variables was estimated using a permutation procedure originally proposed by Buja and Eyuboglu84 ("BE" method) and asymptotic approach proposed by Leek85 ("LEEK" method). BE method yielded 29 surrogate variables and LEEK method yielded 3 surrogate variables for our proteomic data. Regression analysis was performed using lm function in R for 3 outcomes with 1) log10 (protein abundance level) = outcome + age + sex + ethnicity (self-reported) + 29 BE surrogate variables; 2) log10 (protein abundance level) = outcome + age + sex + ethnicity (self-reported) + 3 LEEK surrogate variables. Finally we compared the effect estimates from these two regression models with the main model shown above and calculated the Pearson correlation coefficient of the effect size among the three models (Figure S3).

Replication strategy
To replicate the proteins associated with COVID-19 infection and severity, we searched publicly available COVID-19 resource and downloaded plasma proteomic data from Massachusetts General Hospital (MGH) COVID-19 cohort.24 They recruited 384 individuals (306 cases and 78 controls) and assayed 5,284 aptamers using SomaScan at the first time point. As this proteomic data included neither calibrators nor negative/positive controls, we performed only 2 rounds of IQR strategy to identify the aptamer outliers and individual outliers. After QC, 373 individuals (297 cases and 76 controls) and 5,258 aptamers remained. Among these aptamers, 4,301 proteins were present in both the discovery and replication data and were used for replication analyses using the following model:

\[ \text{log10 (protein abundance level)} = \text{outcome} + \text{age-range} \]

Sex, race and plate information were not available in this public data and were not included in the covariates. In addition, instead of actual age, age range (1 = 20–34, 2 = 36–49, 3 = 50–64, 4 = 65–79 and 5 = 80+) was provided. We used COVID-19 ("COVID" column), ventilation (ventilated within 28 days) and death (death within 28 days) as each of three outcomes.

Meta-analyses
For 4,301 proteins present in both discovery and replication cohort, we performed meta-analysis to identify proteins significant in both discovery and replication samples and boost the statistical power of our analyses. We meta-analyzed the differential abundance analysis results from discovery and replication cohorts using the inverse variance weighed fixed-effect model ("SCHEME STDERR") implemented in METAL.86 The Bonferroni corrected threshold (1.16 \( \times \) 10\(^{-6}\)) was used in the meta-analysis to correct for multiple testing.

Prediction model
Least absolute shrinkage and selection operator (LASSO) regression performs both variable selection and regularization, which forces the coefficients of several weak variables as zero and excludes them in the model. For each of the 3 outcomes, we used five-fold cross validation and Lasso regression in caret R package and selected important features from proteins associated with COVID-19 (n = 841), ventilation (n = 833) and death (n = 253). In addition to these 3 sets of protein, we also evaluated the performance of prediction model for COVID-19 and severity using the proteins commonly present and consistent across all 3 outcomes.

In addition, to construct a ventilation specific prediction model, we selected 50 proteins using the following criteria: 1) they were significant in discovery and replication for ventilation (FDR<0.05), 2) significant in the combined (discovery and replication) meta-analysis at Bonferroni corrected threshold (\( p < 0.05/4; 301 = 1.16 \times 10^{-5}\)) for ventilation, 3) concordant direction of effects between discovery and replication for
ventilation, and 4) not significant in the meta-analysis for death (p >0.05). From these proteins, Lasso regression further selected 32 and evaluated as a separate prediction model for ventilation.

Similarly, for death specific prediction model, 5 death specific proteins (ASB9, Endocan, jun-D, EGFLA, and Glutathione S-transferase Pi) were selected using the following criteria: 1) they were significant in discovery and replication for death (FDR<0.05), 2) significant in the combined (discovery and replication) meta-analysis at Bonferroni corrected threshold (p < 1.16 x 10^-5) for death, 3) concordant direction of effects between discovery and replication for death, and 4) not significant in the meta-analysis for ventilation (p >0.05).

ROC curve and AUC statistics were generated using pROC package. To evaluate the performance of these proteins, the positive predictive value (PPV) and negative predictive value (NPV) were estimated based on Youden’s J statistic optimal cut-off using pROC coords function with an optimal cut-off of ~0.5. The COVID-19 samples who were ventilated and died (44 in discovery and 21 in replication) were removed in the evaluation process. For ventilation and death prediction, both discovery and replication samples were unbalanced and included more controls than cases. To compute NPV and PPV, we first selected the age, gender and ancestry matched controls for each of ventilated cases and each of died cases separately in discovery cohort and age matched controls for each of ventilated cases and each of died cases separately in replication cohort. Then bootstrap sampling of these case-control balanced subsamples was repeated 100 times separately for ventilation and death to increase the estimation accuracy. The PPV and NPV was calculated based on the average of PPV and NPV across 100 bootstrap samples.

Pathway enrichment analyses and tissue-specificity enrichment analyses

We performed pathway enrichment analyses using clusterProfiler 4.0 R package and tissue enrichment analysis using Gene2Func in FUMA GWAS, separately for proteins associated with COVID-19 infection (n = 841), ventilation (n = 833), death (n = 253). For both of these analyses, we used the genes that codify for the proteins included in the SomaScan panel as the background list. Both of these approaches applied over-representation analyses by calculating the amount of input genes that belong to the certain pathways using Fisher’s exact test. In addition, we downloaded one “COVID-19 related gene sets 2021” known as “434 genes down-regulated by SARS-CoV-2 in ferret trachea from GSE147507” (https://maayanlab.cloud/covid19/genesets/512) and performed hypergeometric test with phyper function in R. The significance was defined as Benjamini-Hochberg FDR <0.05. The ggplot2 R package was used to create the dot chart and tile plots for the significant enriched pathways and significant enriched tissues.

Mendelian randomization (MR) analyses

We downloaded the round 5 COVID-19 GWAS results in European population performed by COVID-19 Host Genetics Initiative (COVID19_HGI_C2_ALL_eur_leave_23andme_20210107.txt.gz) and GWAS results of 4,907 plasma proteins in 35,559 Icelanders. Among the 1,449 proteins associated with COVID-19 outcomes, 1,437 were included in GWAS results of 4,907 plasma proteins, 616 proteins had either cis or trans pQTL near the COVID-19 loci (within 1 Mb). Among these 616 proteins, 15 proteins had cis pQTL near the COVID-19 loci. Since genetic variants were randomly inherited from parents at birth which validate their occurrence before any biological traits and diseases, this has been applied in Mendelian randomization (MR) analyses. MR uses genetic variants of exposures as a natural experiment to investigate the causal relations between modifiable exposures and disease. Two-sample MR implemented in R package TwoSampleMR version 0.5.6 was performed to identify the causal plasma biomarkers using each of 616 proteins as exposure and COVID-19 infection as outcome. To select instrument variables (IV) for each of 616 proteins in cis or trans pQTL (Table S15) and 15 proteins in cis pQTL (Table S17), non-palindromic SNPs with association p< 1.5 x 10^-6 for each assessed protein were selected and independent genetic variants was identified using clump_kb = 10000 and clump_r2 = 0.001. Then these independent variants were extracted from outcome (COVID-19 infection) GWAS results. Before MR analyses, we performed data harmonization including fixing the strands and removing variants with non-matching alleles. One critical assumption for TwoSampleMR is no horizontal pleiotropy. We tested for pleiotropic effects by Egger regression intercept (MR-Egger) to check and correct for possible pleiotropy. For the 616 proteins with cis or trans pQTL, causal proteins were identified as: 1) FDR <0.05 with inverse variance weighted (IVW) method if more than one IV (or Wald ratio if one IV); 2) MR-Egger p < 0.05; and 3) consistent direction of effect estimates between MR and differential abundance analyses. For the 15 proteins with cis pQTL, causal...
proteins were identified as \( p < 0.05 \) with IVW if more than one IV (or Wald ratio if one IV) and consistent direction of effect estimates between MR and differential abundance analyses.

**Colocalization analyses**

We performed Bayesian colocalization analyses using coloc.abf function in coloc R package (version 5.1.1).\(^8^9\) Using the default priors as \( P_1 = \frac{1}{3 \times 10^{-4}} \), \( P_2 = \frac{1}{3 \times 10^{-4}} \) and \( P_{12} = \frac{1}{3 \times 10^{-5}} \), the posterior probability of colocalization was estimated with PP.H4 > 0.8 which indicates both proteins and disease shared the same causal variants.

**Hub and druggable proteins**

To identify proteins potentially driving infection or severity, we utilized multi-scale embedded gene co-expression network analysis (MEGENA), a co-expression network analysis technique, to identify the most connected “hub” proteins.\(^3^0\) Because these hub proteins share connections with many others, they are predicted to contribute to disease status or severity. The MEGENA pipeline was used to analyze a set combining the differentially abundant proteins from all three analyses (1,449). Proteins with both increased and decreased abundance were included for each set. A matrix of protein measurements and samples was used as the input data. MEGENA consists of three main steps; first, a fast planar filtered network is constructed. To do this, MEGENA first calculates abundance similarity between each protein pair. We utilized Pearson correlation to measure similarity. We performed false-discovery rate filtering of the protein abundance similarities at a \( p \)-value threshold of 0.05, done by calculating pairwise \( p \)-values with Fisher’s Z-transformation. We computed correlations based on protein levels from COVID-19 positive individuals only. To build the planar filtered network of protein connections, MEGENA starts from a network of protein nodes with no connections. It then iteratively orders protein pairs by their co-expression and tests for planarity. If the pair passes planarity, an edge between the proteins is added. This process is repeated until the number of edges is equal to three times number of vertices (–2).

Next, MEGENA performs multi-scale clustering analysis on the finalized planar filtered network. This is done through optimization of the local clustering structure through local path index. Because MEGENA calculates clusters and identifies hubs using a permutation-based approach, we performed the analysis ten times for each of the four sets of proteins and pooled the hub proteins identified by each run. This produced four lists of hub proteins, each identified in one of the sets of proteins. Each of the ten runs performed 100 permutations. Clusters were limited to a size of number of nodes/2 on the high end and ten on the low end. Bonferroni-corrected \( p < 0.05 \) is used to identify hub proteins.

For each of the hub proteins, we characterized them based on their presence in enriched pathways identified through our earlier analysis, including pathways associated with COVID-19, immune response, and cell surface proteins among others. Additionally, each protein name was used as a query into UniProt,\(^5^3\) which provides information on drugs associated with each protein curated by DrugBank.\(^9^0\) These drugs may be useful for repurposing for COVID-19 treatment or prevention. To visualize the correlations between the proteins for each analysis, we utilized the RCy3\(^9^1\) R package to export the igraph object produced by MEGENA into Cytoscape.\(^9^2\) Using RCy3, we were able to customize the appearance of each of the nodes based on their status as a hub protein, the cluster they were a part of, and the number of significant correlations for each protein.

We also performed MEGENA analysis using a matrix consisting only of the hub proteins identified from the combined analysis. This was performed only to identify clustering of the hub proteins; we did not investigate the “hubs of the hubs”, although by default MEGENA identifies them. We used the same criteria for identification of clusters, with a minimum of ten nodes per cluster and a maximum of \( n/2 \). This analysis was performed once with 100 permutations. We customized nodes by their association with drugs in DrugBank, their identification as hub proteins in each of the four analyses, and their classification in each cluster identified by MEGENA.

We produced a network plot for each of the four analyses and for the hub proteins specifically identified in the combined analysis of all differentially abundant proteins. Using RCy3, we imported the igraph object produced by MEGENA based on the correlation calculations. We customized the appearance of each of the nodes based on their status as a hub protein, the cluster they were a part of, and the number of significant correlations for each protein. First, the lowest-level cluster identified by MEGENA determines the
color of each node/protein. For example, a protein may be part of a large cluster three, as well as subclusters 25 and 79. As cluster 79 is the most specific cluster, the protein is colored corresponding to that cluster. These colors were manually matched to every cluster by supplying random RGB color codes. Node size was changed based on the number of edges containing that node, with larger nodes corresponding to higher connectedness. Finally, node shape was determined by hub protein status; diamonds represent hubs, while non-hubs are circular.